STRATHCLYDE INSTITUTE OF PHARMACY & BIOMEDICAL SCIENCES

University of Strathclyde Science

Metabolomic Modelling and Applications in Inflammatory Bowel Diseases

By

Adel Ibrahim S. Alghamdi

A Thesis Submitted in Fulfillment of the Requirements for the Award of Degree of Doctor of Philosophy in Strathclyde Institute of Pharmacy and Biomedical Sciences at the University of Strathclyde

2019

Declaration

‘This thesis is the result of the author’s original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.’

Signature: _____

Date: _____

Acknowledgements

First and foremost, I wish to express my sincerest gratitude to Allah for providing me with health, commitment and patience to complete my PhD study. I am deeply grateful to my principal supervisor, Dr David G Watson, for his immense support, continuous encouragement and motivation throughout my PhD study. He has provided invaluable guidance, advice and support throughout the duration of this study with his patience and knowledge for which I will always be truly grateful.

Many thanks must go to Dr Nicholas Rattray for his support and help in statistical modelling. My thanks also extend to our collaborator Dr Konstantinos Gerasimidis, University of Glasgow, and his group for their help in providing samples, advice and valuable insights. My sincere thanks also go to all of Dave's research group, PhD student and co-workers in the laboratory for their cooperation, assistance and help. I wish to thank the Saudi government represented in Al Baha University for sponsoring my PhD study.

Last but definitely not least, I specially dedicate this work to the soul of my father who always prayed for me to achieve it. I would like to offer my sincere gratitude to my family (mother, brothers and sisters) for their emotional support and prayers. I am grateful to my wife Dr Nasim and my kids Hala and Hassan for their patience, support and love. I hope this makes you proud.

Table of contents

Table of Contents

Declaration II
Acknowledgements III
Table of contents IV
List of Tables VIII
List of figures X
List of abbreviations XIII
Papers published and posters presented XVI
General abstract XVII
1 General Introduction 2
1.1 Inflammatory bowel disease (IBD) 2
1.2 Epidemiology 3
1.3 Cost effect of IBD 5
1.4 Pathology 6
1.4.1 General characteristics of IBD 6
1.4.2 Crohn's disease 7
1.4.3 Ulcerative colitis 7
1.5 Inflammatory Bowel Disease (IBD) Therapy 7
1.5.1 Medical therapy 7
1.5.2 Nutritional therapy 11
1.6 Metabolomics 12
1.6.1 Approaches to metabolome analysis 14
1.7 Metabolomics applications in IBD 16
1.8 Analytical techniques 17
1.9 Liquid chromatograph mass spectrometry (LC-MS) 17
1.10 Liquid chromatography (LC) 18
1.10.1 Mass spectrometry (MS) 22
1.11 Data extraction and metabolites identification using Mzmatch and IDEOM 24
1.12 Analysis of metabolomics data 26
1.12.1 Univariate data analysis 26
1.12.2 Multivariate data analysis 26
1.13 Data pre-treatment 27

1.13.1 Transformation 27
1.13.2 Scaling 30
1.14 Data visualisation 31
1.14.1 Unsupervised Techniques 31
1.14.2 Supervised Techniques 32
1.15 Model validation 34
1.15.1 Model parameters 34
1.15.2 Permutation test 35
1.15.3 Cross validated ANOVA (CV-ANOVA) 37
1.16 Variable importance in the projection (VIP) 37
2 Materials and Methods 39
2.1 Solvents and chemicals 39
2.2 LC-MS Analysis 39
2.2.1 Mobile phase solutions for ZIC-pHILIC chromatography 39
2.2.2 Mobile Phase for ACE C4 Chromatography 40
2.2.3 HPLC setup 41
2.2.4 Orbitrap Exactive MS setup 42
2.3 Metabolomic profiling 43
2.3.1 Statistical softwares used 43
2.3.2 Data Pre-processing and Modelling 43
2.3.3 Model Validation 45
2.3.4 Data Filtration 46
2.3.5 Ranking, Grouping and Confirmation of Significant Metabolites 46
2.3.6 Data bases used for identification 48
3 Data Pre-treatment and Statistical Model Selection 50
3.1 Abstract 50
3.2 Introduction 52
3.2.1 Metabolomic data 52
3.2.2 Missing data imputation: 52
3.2.3 Mean-centering 53
3.2.4 Scaling 54
3.2.5 Transformation 56
3.2.6 Aims 57
3.3 Materials and method 58

3.3.1 Data 58
3.3.2 LC-MS Analysis 58
3.3.3 Data pre-processing 59
3.3.4 Data pre-treatment 59
3.3.5 Software tools 62
3.3.6 Models validation 62
3.4 Results 64
3.4.1 Nonlinear Iterative Partial Least Squares algorithm (NIPALS) 64
3.4.2 K-Nearest Neighbours algorithm (KNN) 68
3.4.3 Small value (minimum intensity/2) 71
3.4.4 Mean and median 74
3.5 Discussion 77
3.5.1 Missing data imputation 77
3.5.2 Scaling and transformation 79
4 Untargeted Metabolomics of Paediatric Crohn's Disease Patients Against Healthy Controls 83
4.1 Abstract 83
4.2 Introduction 85
4.3 Aim of the study 87
4.4 Materials and methods 87
4.4.1 Solvents and chemical 87
4.4.2 Samples and Sample Preparation 87
4.4.3 LC-MS analysis 90
4.4.4 LC-MS/MS analysis 91
4.4.5 Data analysis 91
4.5 Results 92
4.5.1 Pooled samples 92
4.5.2 Data Visualisation 93
4.6 Discussion 107
5 Metabolomics Discrimination between Crohn's Disease and Ulcerative Colitis 114
5.1 Abstract 114
5.2 Introduction 116
5.3 Metabolomics discriminations between CD and UC 117
5.4 Aim of the study 121
5.5 Materials and methods 122

5.5.1 Solvents and chemicals 122
5.5.2 Sample preparation 122
5.6 LC-MS analysis 123
5.6.1 Mobile phase solutions for ZIC-pHILIC chromatography 123
5.6.2 HPLC setup 123
5.6.3 Orbitrap Exactive MS setup 123
5.6.4 Data analysis 123
5.6.5 Group comparisons 123
5.7 Results 124
5.7.1 Data visualization 124
5.7.2 Model selection 126
5.7.3 Biomarker identification 128
5.8 Discussion 158
5.8.1 Common putative metabolites shared between CD and UC 158
5.8.2 Specific putative metabolites that discriminate CD from HC 161
5.8.3 Specific putative metabolites that discriminate UC from HC 162
5.8.4 Comparison between CD and UC 164
6 Summary and future works 167
6.1 Data pre-treatment and statistical model selection 168
6.2 Untargeted Metabolomics of Paediatric Crohn's Disease Patients Against Healthy Controls 169
6.3 Metabolomics Discrimination between Crohn's Disease and Ulcerative Colitis . 170
6.4 Future works 171
7 References 172
8 Appendix 188
8.1 Appendices Chapter 3 188
8.2 Appendices for chapter 4 191
8.3 Appendixes chapter 5 214

List of Tables

Table 2.1 Gradient elution programme applied for ZICpHILIC in LC-MS analysis..... 40

Table 2.2 Gradient elution programme applied for ACE C4 column in LC-MS analysis 41

Table 3.1 The pre-treatment methods and equations. 62

Table 3.2 Model parameters after NIPALS algorithm zero imputation. 65

Table 3.3 Model parameters after KNN algorithm zero imputation. 69

Table 3.4 Model parameters after Min/2 zero imputation. 72

Table 3.5 Model parameters after mean imputation. 75

Table 3.6 Model parameters after median imputation. 76

Table 4.1 Subject data for HCs and patients. 89

Table 4.2 Samples numbers and groups of paediatric CD before, during, and after EEN and HCs. 89

Table 4.3 An overview of all the orthogonal partial least square-discriminant analysis (OPLS-DA) parameters and their validity. 94

Table 4.4 List of metabolites that were significantly different in the pre-EEN treatment group (PA) compared to the HCs (HC), based on an OPLS-DA model..... 96

Table 4.5 The relative abundance of long chain fatty acids in the faecal extracts based on analysis of a ZICp HILIC column..... 96

Table 4.6 Details of characterization of the eight marker compounds shown in Table 4.4 obtained in positive (+) or negative (−) ion mode. 104

Table 4.7 Concentration of fatty acids in each sample (µg/g of dry faeces). P-value based on Log base2. 106

Table 5.1 Summarization of previous metabolomic studies using faecal extract samples. 119

Table 5.2 Subject data for HCs and patients. 122

Table 5.3 Misclassification table showing the proportion of correctly classified observations in IBD patients and HC using OPLS-DA model 128

Table 5.4: The number of significant metabolites according to pathways. 130

Table 5.5 Putative biomarkers and their pathways that discriminate CD from HC samples 139

Table 5.6 Top 10 putative metabolites that discriminate CD from HCs samples based on VIP ranking. 141

Table 5.7 Putative biomarkers and their pathways that discriminate UC from HC samples 146

Table 5.8 Top 10 putative metabolites that discriminate UC from HC samples based on VIP ranking. 148

Table 5.9 Putative biomarkers and their pathways that discriminate CD from UC samples 151

Table 5.10 Top 10 putative metabolites that discriminate CD from UC samples based on VIP ranking. 155

List of figures

Figure 1.1 An overview of the four major "omics" fields, from genomics to metabolomics..... 13

Figure 1.2 A schematic diagram to illustrate the components of an HPLC system ... 19

Figure 1.3 A schematic diagram to show the main components of a mass spectrometer..... 23

Figure 1.5 Normal probability plot of residuals. 29

Figure 1.6 Score plot of Orthogonal Partial Least Square Discriminant Analysis (OPLS-DA)..... 34

Figure 1.7 Permutations test..... 36

Figure 2.1 Flow chart for sample and data analysis..... 48

Figure 3.1 Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) score plot for NIPALS zero imputed data set after applying UV and Par scaling..... 66

Figure 3.2 Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) score plot for NIPALS zero imputed data set after applying Log_{10} transformation to UV and Par scaled data. 67

Figure 3.3 Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) score plot for KNN zero imputed data set after applying UV and Par scaling. 70

Figure 3.4 Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) score plot for Min/2 zero imputed data set. The significant models were presented after applying different pre-treatment methods. 73

Figure 4.1 2D Scores plot of the principal components analysis (PCA) for the quality control (QC) samples (blue) and all samples (grey) based on 606 putative metabolites 93

Figure 4.2. Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) score plot of pre-EEN samples (PA) against HCs (HC). 100

Figure 4.3 Log_2 of the fold-change in the eight differentiated metabolites in the CD groups (before, during, and after EEN treatment) compared with the group of HCs. 102

Figure 5.1 Factors affecting IBD pathogenesis. 117

Figure 5.2 2D PCA score plot for QC (pooled) faecal water samples. 124

Figure 5.3 2D PCA score plots for CD (Blue-CD) and UC (Grey-UC) against HCs (purple-HC). 126

Figure 5.4 OPLS-DA score plots for CD (Blue-CD) and UC (Grey-UC) against HCs (purple-HC) faecal extract samples and its permutation test (999 times). 127

Figure 5.5 Venn diagram Venn-Diagrams of the screened metabolites. 129

Figure 5.6 Metabolic pathways are altered in CD (CD) compared to HC (HC) subjects. 131

Figure 5.7 Metabolic pathways are altered in UC (UC) compared to HC (HC) subjects. 132

Figure 5.8 Common Fatty acids shared between CD and UC in relative to HCs. 133

Figure 5.9 Common Diacylglycerols shared between CD and UC in relative to HC. 134

Figure 5.10 Common Glycerophospholipids shared between CD and UC in relative to HCs. 135

Figure 5.11 OPLS-DA score plots for CD (CD) against HC (HC) samples. 136

Figure 5.12 OPLS-DA score plots for CD (CD) against HC (HC) samples after outliers' exclusion. 138

Figure 5.13 OPLS-DA score plots for CD (CD) against HC (HC) samples based on VIP ranking. 143

Figure 5.14 OPLS-DA score plots for UC (UC) against HC (HC) samples. 145

Figure 5.15 OPLS-DA score plots for UC (UC) against HC (HC) samples based on VIP ranking. 150

Figure 5.16 OPLS-DA score plots for CD (CD) against UC (UC) samples. 154

Figure 5.17 OPLS-DA score plots for CD (CD) against UC (UC) samples based on VIP ranking. 156

Figure 5.18 Heat map showing the top 10 putative metabolites based on VIP ranking that are significantly discriminate CD from UC. 158

List of abbreviations

6-MP	6-mercaptopurine
ACG	American College of Gastroenterology
ANOVA	Analysis of variance
APCI	Atmospheric Pressure Chemical Ionisation
ASA	Aminosalicylic acid
AUROCC	Area Under the ROC Curve
BCG	British Society of Gastroenterology
CD	Crohn's Disease
CV	Cross validation
DAD	Diode Array Detectors
DNA	Deoxyribonucleic Acid
EEN	Exclusive Enteral Nutrition
EI	Electron Impact
ELSD	Evaporative Light Scattering Detector
ESI	Electrospray Ionisation
FT-ICR	Fourier Transform Ion Cyclotron Resonance
GC	Gas chromatography
GIT	Gastrointestinal Tract
HC	Healthy Controls
HILIC	Hydrophilic Interaction Liquid Chromatography
HPLC	High Performance Liquid Chromatography

IBD	Inflammatory Bowel Disease
IL-12	Interleukin-12
IL-4	Interleukin-4
IL-5	Interleukin-5
IT	Ion Traps
KNN	k-nearest neighbors algorithm
LC	Liquid chromatography
LC-MS	Liquid chromatography-mass spectrometry
m/z	Mass-to-Charge ratio
MS	Mass Spectrometry
MVA	Multivariate analysis
NF-κβ	nuclear factor kappa β
NIPALS	Nonlinear Iterative Partial Least Squares algorithm
NMR	Nuclear Magnetic Resonance spectroscopy
NPC	Normal phase chromatography
OPLS-DA	Orthogonal partial least squares - discriminant analysis
Par	Pareto scaling
PCA	Principal Component Analysis
PLS-DA	Partial least squares-discriminant analysis
Q	Quadrupoles
Q^2	Goodness of Prediction
QSRR	Quantitative Structure Retention Relationships

R^2	Goodness of Fit
RCT	Randomised Controlled Trial
RNA	Ribonucleic Acid
ROC	Receiver Operator Characteristic
RPC	Reversed phase chromatography
RSD	Relative Standard Deviation
RT	Retention time
SIMCA	Soft Independent Modeling of Class Analogy
SIMCA	Soft-Independent Modelling of Class Analogy
SPSS	Statistical Package for Social Scientists
TNF	Tumour Necrosis Factor
TOF	Time-of-flight
UC	Ulcerative Colitis
UV	Unit Variance scaling
UVA	Univariate analysis
VIP	Variable importance in the projection
ZIC	Zwitterionic

Papers published and posters presented

Papers

1. Alghamdi, A. et al. (2018) 'Untargeted Metabolomics of Extracts from Faecal Samples Demonstrates Distinct Differences between Paediatric Crohn's Disease Patients and Healthy Controls but No Significant Changes Resulting from Exclusive Enteral Nutrition Treatment', Metabolites, 8(4), p. 82. doi: 10.3390/metabo8040082.
2. Svolos, V. et al. (2019) 'Treatment of Active Crohn's Disease With an Ordinary Food-based Diet That Replicates Exclusive Enteral Nutrition', Gastroenterology, 156(5), pp. 1354-1367.e6. doi: 10.1053/j.gastro.2018.12.002.

Poster

1. Adel Alghamdi, Vaios Svolos, Konstantinos Gerasimidis and David Watson. Untargeted Metabolomics Screening for Crohn's Disease Biomarkers in Paediatric Patients Using Liquid Chromatography/Mass Spectrometry Method. Metabomeeting 2017, Birmingham.

Oral presentation

1. Adel Alghamdi, Vaios Svolos, Konstantinos Gerasimidis and David Watson. Untargeted Metabolomics Screening for Crohn's Disease Biomarkers in Paediatric Patients Using Liquid Chromatography/Mass Spectrometry Method. Scottish metabolomic network 2017, Glasgow.

General abstract

Background: Metabolomics experiments typically produce high dimensional data and its handling is an extremely important step in data pre-treatment. Metabolomics is an indispensable research tool for the identification and tracking of biomarkers in biological systems. In a typical metabolomics study, complex extracts or body fluids are analysed and compared by various methods to generate metabolic fingerprints. Crohn's Disease (CD) and ulcerative colitis (UC) are major components of Inflammatory Bowel Disease (IBD), a multifactorial disorder most likely resulting from altered immune response to commensal or pathogenic gut microbes under the influence of environmental factors, such as diet. Exclusive Enteral Nutrition (EEN) is the most common treatment for paediatric CD in the UK and the rest of Europe. Non-invasive metabolomics approaches could be used to diagnose and differentiate between related diseases. This could enhance disease control, management and patient compliance. It is known that gut microbiota may discriminate IBD subtypes from each other, therefore, metabolomics of faecal extracts was used to examine metabolites in faeces many of which result from the activity of gut microbiota and thus to differentiate between IBD subtypes and healthy controls as well as within IBD subtypes.

Methodology: This study investigated the effect of pre-treatment strategies on data set derived from LC-MS based metabolomics experiments. Different methods of imputing missing values were examined in conjugation with various scaling and transformation methods. SIMCA-P 14 was used to evaluate the model parameters for each pre-treatment method.

In this thesis, metabolomics was employed in various studies to assess metabolite biomarkers associated with healthy controls and IBD diseases. All the studies employed liquid chromatography-mass spectrometry (LC-MS) on an Orbitrap Exactive mass analyser, and using ZIC-pHILIC or/and C18 analytical columns. Data was acquired using XCalibur software and metabolite identification was ascertained based on accurate mass detection, retention time comparisons with authentic external standards, and database searching. The acquired data was analysed using both unsupervised (PCA-X) and supervised (OPLS-DA) models in SIMCA in order to determine discriminating metabolite biomarkers responsible for the observed clustering patterns.

Results: Compared to the various imputation methods used in this study, NIPALS algorithm along with suitable transformation and scaling was significantly better according to the model parameter evaluation Pareto (Par) as scaling and Log transformation were better able to explain the data. The OPLS-DA model was able to discriminate the CD samples from the controls at different time points after the commencement of treatment .The models were not able to differentiate the CD samples from one another at the different time points during treatment with exclusive enteral nutrition. The metabolites identified in the CD samples which varied between CD samples and controls included tyrosine, an ornithine isomer, arachidonic acid, eicosatrienoic acid, docosatetraenoic acid, a sphingomyelin, a ceramide, and dimethylsphinganine. Similarly, the OPLS-DA model was able to discriminate the CD samples from the UC. Based on VIP values, the top 10 metabolites were used in the OPLS-DA model, and there was a clear separation between CD and UC with *p* CV-ANOVA = 5.30541e-007.

Conclusion: The SIMCA-P 14's (NIPALS) default logarithm was the only imputation methodology that generated a valid model according to valid criteria (R^2-Q^2<0.3). This was in conjugation with Pareto scaling in conjugation with Log transformation were the best data pre-treatment methodology. Despite successful treatment, underlying differences remained in the metabolome of the CD patients. Untargeted metabolomics analysis was also performed to classify the faecal extracts from patients suffering from different inflammatory bowel diseases to evaluate the use of this technique as a diagnostic tool and categorize specific metabolites in the faecal extract of participants with specific types of inflammation.

Chapter 1:

General Introduction

1 General Introduction

1.1 Inflammatory bowel disease (IBD)

In UK, around 150,000 citizens are affected by inflammatory bowel disease (IBD), which is a chronic, incapacitating condition that impacts the patients' gastrointestinal tract (Loftus, 2004). For more than 100 years, it has been recognized as individual disease entities. In 1761, Morgangni introduced the concept of intestinal inflammation, which is now known to be Crohn's disease (CD) (Lockhart-Mummer and Morson, 1964), however it took until 1932 for terminal ileitis to be recognised. Subsequently, Lockhart-Mummery and Morson identified granulomatous colitis, which they described as impacting both the large and small bowel (Daiss, Scheurlen and Malchow, 1989). Hence, this determined phenotypic distinction from ulcerative colitis. It is typically accepted that the first pathological account of simple ulcerative colitis (UC) was proposed by Wilkes in 1859, and further expanded upon in 1875 in collaboration with Moxon (Lockhart-Mummer and Morson, 1964).

IBD is a chronic inflammation which may affect any parts of the gastrointestinal tract (GIT) and is characterized by two main diseases, CD and UC (Cosnes *et al.*, 2011). It is diagnosed based on the type and the location of inflammation. The pathogenesis of IBD involves dysregulation or current activation of the mucosal immune system caused by intestinal microbial imbalance (microbiota dysbiosis) (Dolan and Chang, 2017) . In addition to that, patient genetic factors strongly may also play a significant

role. However, the specific pathophysiology and etiology of IBD are not fully clarified. The manner for development of both diseases and their pathogenesis is influenced by environmental factors, diet, smoking habits and microbial factors (Ananthakrishnan, 2015). Currently, there is an emerging consensus hypothesis proposing that the initiating or maintaining of the disease comes from variable factors such as microbial dysbiosis or differences in the microbial environment (Sheehan, Moran and Shanahan, 2015). Therefore, the compositional variation may be indicated in metabolic activities of the gut microbiota which in turn lead to alteration in the metabolites.

1.2 Epidemiology

The incidence of IBD refers to the occurrence of new cases in a specific population over a particular time. IBD literature usually expresses this as cases per 100,000 people per annum. The prevalence of IBD is the number of IBD cases at any one time in a specific population. This is generally expressed as the rate per 100,000 of the population. There is a substantial body of published literature that documents the global incidence and prevalence of Crohn's disease (CD) and ulcerative colitis (UC). From these articles, it is evident that geographic region has a strong bearing on the incidence rates.

In North America, the UC and CD incidence rates range from 8.8-23.14 and 6.3-23.8 cases per 100,000 of the population per year respectively, whilst the prevalence ranges from 139.8-286.3 (UC) and 96.3-318.5 (CD) cases per 100,000 of the population (Ng *et al.*, 2017). By generalising these figures for the 365 million people

(estimated combined population) in the US and Canada in 2019, there are between 32,000 and 85,000 new UC diagnoses annually. Similarly, the figures for new CD diagnoses per year are between 23,000 and 87,000 cases (Ng *et al.*, 2017).

According to the multicentre European Collaborative study on Inflammatory Bowel Disease, the combined incidence rates for UC and CD are 8.7-11.8 and 3.9-7.0 cases per 100,000 of the population per year respectively (Shivananda *et al.*, 1996). Based on this, it can be concluded that there are approximately 50,000-68,000 new cases of UC and 23,000-41,000 new cases of CD diagnosed each year in Europe. The study also investigated the north-south gradient and determined that the rates of IBD were 40-80% higher in Northern Europe (Shivananda *et al.*, 1996). According to another multicentre European Collaborative study on Inflammatory Bowel Disease, the combined median range of incidence in Western Europe rates for CD and UC are 0-10.7 (median 6.5) and 2.9-31.5 (median 10.8) , and in Eastern Europe rate for CD and UC are 04-11.5 (median 3.1) and 2.4-10.3 (median 4.1) cases per 100,000 of the population per year respectively (Burisch *et al.*, 2014).

In the past, there have been few incidences of IBD on other continents, however in recent years, UC cases have begun to increase in areas such as Japan (Ng, Wong and Ng, 2016), Northern India (Kedia and Ahuja, 2017), and South America (Kotze *et al.*, 2020). Conversely, incidences of CD remain low. There are several commonalities in the IBD incidences' temporal and geographic trends. There continues to be low IBD incidence rates in developing countries. This could be accurate, in that there are simply few cases, or it could be attributed to a lack of diagnostic ability or mistaking

infectious causes of diarrhoea. However, evidence from epidemiology figures has shown that as developing countries become increasingly westernised or industrialised, changes to diet and environment occur, and cases of UC materialise first, followed by CD (Loftus, 2004).

1.3 Cost effect of IBD

The UK NHS spends approximately £3000 and £6000 per year to treat any patient with UC and CD, respectively (Ghosh and Premchand, 2015). Obviously, this is a massive expense to the health system, but this is necessary to manage the incidence and prevalence of UC and CD and the chronic complexion of IBD. On average, taking diagnosis, management and treatment into account, each patient costs £3,000 per year (Luces and Bodger, 2006). A study conducted in Liverpool attempted to determine the actual cost of treating patients at a secondary case level by investigating a single centre university hospital that provides care to a population of 330,000 (Bassi *et al.*, 2004). Over six months, 307 cases of ulcerative (or indeterminate) colitis and 172 cases of CD were diagnosed, and the relevant demographic and clinical data were collected. The average cost per patient for six months was found to be £1652 for CD and £1256 for colitis. When the dormant cases of IBD were evaluated, it was determined that relapse costs for cases where the patient was not admitted to hospital doubled or tripled, whilst costs for patients who were admitted increased 20-fold. These findings indicate that whilst only a minority of cases required hospitalisation, these cases comprised 50% of the total direct IBD patient costs.

It can be concluded that a more in-depth comprehension of the aetiology and the pathogenesis of IBD resulting in an improved level of medical management of patients would have a dual effect of significantly enhancing patients' quality of life and decreasing the cost to the NHS of IBD.

1.4 Pathology

1.4.1 General characteristics of IBD

IBD is a disease characterized by relapsing and remitting episodes of diarrhea and bloody diarrhea, which can be further subdivided into CD and UC based on unique clinical, endoscopic, and pathologic features. In addition to that, IBD patients may present with abdominal pain, weight loss and other symptoms, such as low grade fever (Yamada *et al.*, 2015). Prior to rendering a diagnosis of IBD, careful inspection of macroscopic and microscopic features are required to rule out other disease processes. To that end, three key histologic features must be closely evaluated: chronic injury, disease distribution, and disease activity. Chronic injury to gastrointestinal mucosa is established following months to years of repeated damage rather than days to weeks. As such, the characteristic histologic features of chronic injury must be present in order to render a diagnosis of inflammatory bowel disease. Next, the extent of active disease, or the degree of inflammation, is assessed to determine the severity of disease at the time of biopsy. Finally, to sub-classify IBD into either Crohn's disease or ulcerative colitis, the distribution of both chronic and active inflammation must be taken into consideration (Baumgart, 2017).

1.4.2 Crohn's disease

CD is characterised by patchy transmural inflammation that can affect any part of the gastrointestinal tract from the mouth to the anus, but most commonly affects the ileocaecal region. There is a tendency to develop inflammatory or fibrotic strictures as well as fistulae, both internal and perianal. The presentation of the disease can be rather heterogeneous due to the variety of locations that can be affected. Typically, symptoms can include weight loss, abdominal pain, diarrhoea, vomiting and systemic upset.

1.4.3 Ulcerative colitis

UC is characterised by a transmucosal inflammation that is continuous from the rectum to the extent of the disease, but classically only affects the colon, although the terminal ileum can be affected by a backwash ileitis, and in very rare cases a more diffuse small bowel inflammation can result. The predominant symptom tends to be bloody diarrhoea (Gajendran *et al.*, 2019).

1.5 Inflammatory Bowel Disease (IBD) Therapy

1.5.1 Medical therapy

Inflammatory bowel disease (IBD) is characterised by acute cases of relapse and remission. It is because of this that the foundations of IBD therapy entail the instigation of remission followed by the use of medical treatments to maintain the disease status as a means of preventing the need for surgery. Surgery is employed, however, for the extraction of diseased bowel regions, the removal of intestinal blockage, restrictions or fistulations or the control of intra-abdominal sepsis.

Proctocolectomy can be used as treatment for ulcerative colitis (UC), however, Crohn's disease (CD) cannot be treated and thus requires the maintenance of the state of remission using drugs. There are various medical treatments extensively employed including corticosteroids, 5-aminosalicylic acid (5-ASA) medication, immunosuppressive agents, antibiotics and biological therapies (Talley *et al.*, 2011).

1.5.1.1 Corticosteroids

Almost all elements of the immune response are constrained by corticosteroids (Talley *et al.*, 2011). Corticosteroids interact with the glucocorticoid receptors located in the nuclei of cells. Through this interaction they prevent adhesion molecule expression and the transport of inflammatory cells to their target cells and tissues which includes the intestines. Their action also reduces the expression of cytokines released in response to inflammation and simultaneously initiates apoptosis of active lymphocytes (Goulding, 2004). In 1954, the first ever randomised controlled trial (RCT) was described and involved the use of cortisone for UC (Truelove and Witts, 1954) and ever since serious IBD reactions have been treated with corticosteroids. These are not employed for maintenance of a condition, rather they are used for a short period to initiate remission. Their severe side effects are significant and include a greater chance of developing infections (Toruner *et al.*, 2008) and the manifestation of psychiatric conditions in the short-term. However, when such steroids are used long-term they are also detrimental as they can cause diabetes mellitus, poorer bone mineral density and other more severe side effects (Irving *et al.*, 2007).

1.5.1.2 5-Aminosalicylic acid

These medications are thought to function as anti-inflammatory drugs as they perform their action by the inhibition of nuclear factor kappa β (NF-κβ) and chemoattractant leukotrienes, whilst also changing the metabolism of prostaglandins (Desreumaux and Ghosh, 2006). Negative side effects are uncommon as 5-ASA formulations have low bioavailability (Kane *et al.*, 2003). Nevertheless, significant adverse events can occur such as pancreatitis, interstitial nephritis, pericarditis, hepatitis and pneumonitis (Gisbert, González-Lama and Maté, 2007).

1.5.1.3 Immunosuppressive agents

Several drugs are employed in the management of IBD as inducers of remission and for its maintenance and these include analogues of thiopurine e.g. 6-mercaptopurine (6-MP) and azathioprine (6-MP's pro-drug), methotrexate and the calcineurin inhibitors; specifically ciclosporin and tacrolimus (Talley *et al.*, 2011). Side effects associated with the analogues of thiopurine include allergic responses, hepatitis, nausea, and severe pancreatitis, and malignancy, suppression of bone marrow and a higher chance of infection. Regarding methotrexate: pneumonitis, stomatitis, infection, hepatotoxicity, myelosuppression and alopecia through malignancy, all are noted side effects. Whilst renal toxicity is the primary side effect linked to the calcineurin inhibitors. Nevertheless, others may manifest including hirsutism, headaches, high blood pressure, infection, paraesthesia and seizures (Aberra and Lichtenstein, 2005). Various immunomodulatory features are attributed to thalidomide use. These include the inhibition of tumour necrosis factor (TNF), interleukin-12 (IL-12) and interferon-γ (IFN), the stimulation of interleukin-4 (IL-4)

and interleukin-5 (IL-5), disrupting expression of integrin, prevention of angiogenesis and the reduction of numbers of circulating helper T-cells. Thalidomide has also been employed as IBD therapy in the past. Due to its harmful effect on foetus growth, thalidomide was used with high restrictions to treat erythema nodules of leprosy, multiple myeloma, HIV and cancer (Vargesson, 2015). This, however, was changed following the publication of a systematic review that indicated that evidence was lacking regarding its benefits in inducing or maintaining remission in either UC or CD patients (Yang *et al.*, 2015).

1.5.1.4 Antibiotics

The aetiology of CD also includes the effects of various bacteria such as Mycobacterium (Feller *et al.*, 2007), Escherichia coli and Listeria (Palmer *et al.*, 2007). Research has demonstrated that the creation of an ileostomy leads to a change in the path of the faecal stream which in turn decreases repeat cases of colonic CD (Janowitz, Croen and Sachar, 1998). Antibiotic use can lead to initiation of remission in active forms of UC and they can also stop relapse in cases of quiescent CD (Talley *et al.*, 2011). One limitation in this study, however, was in the extensive range of antibiotics examined, thus preventing the recommendation of one antibiotic over the others.

The role of antibiotics therapy in CD has been described whilst taking into consideration the guidelines that have been published for IBD by a number of associations. Guidelines for the treatment of IBD were published by the British Society of Gastroenterology (BSG), in 2011 (Mowat *et al.*, 2011). They give details of the important role antibiotics have in treating secondary complications in CD, quoting bacterial overgrowth and abscesses. They further state that there might be a specific

use for ciprofloxacin and metronidazole in CD therapy. In clinical practice, antibiotics are prescribed generally for the treatment of Crohn's disease (CD). However, controlled trials have not substantively established their efficacy in the setting of luminal Crohn's disease, according to the American College of Gastroenterology (ACG) (2009). Suitable antibiotic therapy or drainage are necessary for infection or abscesses. Metronidazole should be used to treat non-suppurative perianal complications of CD, either alone or in combination with ciprofloxacin. Continuous therapy is required to avoid recurrent drainage (Lichtenstein *et al.*, 2018).

Regarding to UC treatment guidelines, according to the BSG guidelines, the use of antibiotics in UC as disease-modifying therapy is not proven and, therefore, is not supported (Mowat *et al.*, 2011). The ACG published guidelines for the treatment in 2010. There was no reference to antibiotic treatment for mild to moderate disease (Kornbluth and Sachar, 2010). Controlled antibiotic trials displayed no therapeutic benefits when intravenous steroids were combined with antibiotics for the treatment of severe colitis in the absence of proven infection. Normally, broad-spectrum antibiotics are prescribed for patients with signs of toxicity, or for those who even with maximal medical therapy, develop more severe symptoms, as a part of the protocol that is outline treatment regimens for severe colitis.

1.5.2 Nutritional therapy

Following Crohn's disease diagnosis, patients are administered Exclusive enteral nutrition (EEN) which is a nutrient approach to induce remission and optimise nutrition. It requires the administration of a diet formulation that is liquid based for

a period of time as the only intervention that can lead to remission (Critch *et al.*, 2012). EEN formulations could be elemental, semi-elemental or polymeric. It is recommended for children with luminal disease, including those with colonic involvement (Navas López *et al.*, 2014). EEN has no side effects but it correlates with high rates of mucosal healing. It is however, not curative but similar to other currently available therapies for CD.

Although EEN has been used for many years, the mechanism of this therapy has just been understood in the last few years. Three primary components are implicated in the actions of EEN: changes in the intestinal microflora (LEACH *et al.*, 2008), barrier function enhancement (Nahidi *et al.*, 2013) and direct anti-inflammatory effects (de Jong, Leach and Day, 2007). The relationship between these actions and the triggers for these changes are yet known.

1.6 Metabolomics

The omics approaches have enabled great progress in biological sciences in the last twenty years. Genomics have been used to identify genes, transcriptomics is used to study the production of Ribonucleic acids (RNAs) from genes while proteomics is used to investigate whether an RNA is converted to protein and any modifications taking place in the protein after conversion or translation, and finally metabolomics is used to monitor any changes in metabolites due to protein expression (

*Figure **1.1***).

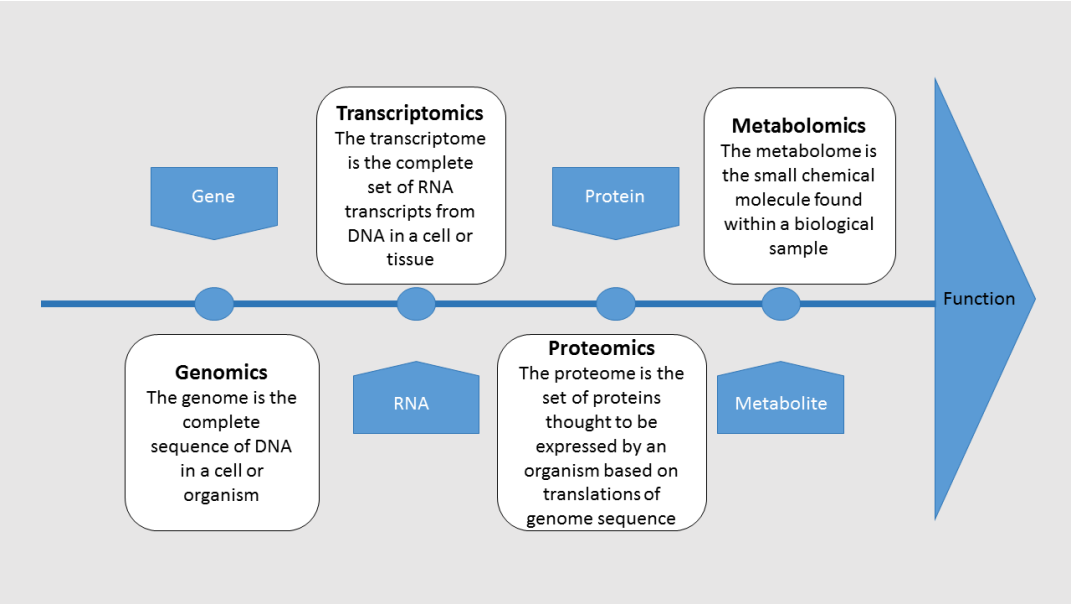


Figure 1.1 An overview of the four major "omics" fields, from genomics to metabolomics

Metabolomics has come to be recognised as the "un-biased comprehensive identification and quantification of the entire metabolome under a given set of analytical conditions with high selectivity and sensitivity (Dunn, Bailey and Johnson, 2005). It is derived from "metabolism" which itself originates from the Greek word *metabole´* meaning "change". A precise definition of metabolomics may not be possible but a common theme is usually used in its definition such as the study of low molecular weight molecules found in cells and organisms and which participate in the metabolic functions necessary for the functions of the cell such as growth and maintenance (Harrigan and Goodacre, 2012). Metabolites are usually the end

products of gene and protein expression in cells depending on the interaction of the cells with the immediate environment (Fiehn, 2002).

Metabolomics complements transcriptomics and proteomics but has certain advantages such as identifying the output products of gene expression and providing a better description of the biological system, compared to transcriptomics and proteomics, which are sensitive to post-translational modifications and other activities in the regulatory pathway. Metabolomics involves lower cost than any other omic approaches. This is because it has a higher capability for analysing a greater number of samples in a single study (Broadhurst and Kell, 2007). Apart from metabolomics, no other analytical methodology or platform is able to detect, quantify and identify high number of metabolites in a given sample. Analytical samples that can be used for metabolomics study include microorganisms, tissue, cell culture, and biological fluids such as serum (Kolho *et al.*, 2017), urine (Martin *et al.*, 2017) and faecal samples (Marchesi *et al.*, 2007; Svolos *et al.*, 2019).

1.6.1 Approaches to metabolome analysis

Metabolomics profiling can be used in the study of metabolic perturbations associated with various disease conditions or treatment using targeted, semi-targeted or untargeted metabolomics techniques (Dunn, 2013). These approaches differ in their quantitative (whether absolute or relative), and qualitative (level of experimental precision and accuracy), sample complexity in terms of the number of metabolites involved, and the objective of the study.

1.6.1.1 Targeted metabolomics:

The fully targeted metabolomics method requires predefined or predetermined metabolites before sample analysis and data acquisition. It is precise, accurate and selective for the targeted metabolite(s), hence the targeted approach is more quantitative than qualitative. The targeted method utilises data initially generated from an untargeted or semi-targeted study to justify using a robust technique requiring pure chemical standards. The identification of significant biomarkers using targeted metabolomics is useful in providing conclusions regarding their biological relevance in the initial hypothesis that was based on untargeted methods.

1.6.1.2 Semi-targeted metabolomics

Semi-targeted metabolomics is similar to fully targeted metabolomics. Semi-targeted metabolomics however, requires the elucidation and confirmation of detected metabolites in the sample while a fully targeted approach makes use of the identity of certain metabolite(s) prior to sample analysis. As in fully targeted metabolomics, the semi targeted method also requires high precision, accuracy and selectivity for the targeted analytes and hence this method is also quantitative.

1.6.1.3 Untargeted metabolomics

Untargeted metabolomics profiling is used for a general screening and detection of metabolites. It is commonly performed for the detection of a wide range of chemical classes and infer as much as possible, broad a picture of metabolism and metabolic processes (Kell and Oliver, 2004). Samples are analysed and data processed using chemometric or statistical tools. The results may lead to a hypotheses based on the

significant metabolites or observations. This method yields a lot of metabolite data and some unidentifiable metabolites. On the other hand, some identified metabolites cannot be confirmed due to absence of standards or cost.

Metabolomics studies have only been made possible now due to advances in analytical techniques and informatics tools. These advanced and modern analytical methods have enabled rapid analysis of complex mixtures and the vast amounts of data generated can be analysed and modelled using various software and online based tools. These technologies have been applied in plant, environmental and mammalian systems with the aim of identifying novel biomarkers and understanding possible biological mechanisms resulting from different treatments or genetic modifications (Dunn, 2013) . The strategy of untargeted metabolic profiling is generally advantageous as a *de novo* knowledge of the metabolites present is not required.

1.7 Metabolomics applications in IBD

Different studies involving metabolomics analysis have been applied in the classification and diagnosis of IBD (Kolho *et al.*, 2016; Soubieres and Poullis, 2016). Most studies concentrated on recognising CD or UC disease fingerprints when compared to a healthy control (HC) (Bjerrum *et al.*, 2015; Kolho *et al.*, 2017). Other studies have applied metabolite profiling to discriminate between IBD subtypes and to classify the disease-related metabolites (Marchesi *et al.*, 2007; Jansson *et al.*, 2009). A few studies tried to use metabolite differences to ascertain disease severity or the location of the disease (Gerasimidis *et al.*, 2011). There are

presently a limited number of studies that have tracked treatment outcomes in IBD subtypes (Gaifem *et al.*, 2018; Ning *et al.*, 2019; Serena and Fasano, 2019; Svolos *et al.*, 2019).

1.8 Analytical techniques

The two major analytical techniques used in metabolomics are Nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS). They are used both for general profiling, un-targeted or targeted metabolomics (Alonso, Marsal and Julià, 2015). These techniques have been improved to meet the requirements for the different objectives of a study. The early development of metabolomics relied on NMR, which has the disadvantage of low resolution and the narrow range of metabolites that can be identified. Mass spectrometry has increasingly become the more common tool (van Ginneken *et al.*, 2007). A mass spectrum sorts ions depending on mass to charge ratio (m/z) to give an idea of the composition of a sample. The robustness and high sensitivity of mass spectrometry has made it an important method to detect and quantify metabolites in a variety of samples. The introduction of the LTQ Orbitrap innovation has provided the foremost technology possessing high and consistent mass accuracy along with the fast scanning required for compatibility with chromatographic systems such as HPLC (Makarov *et al.*, 2006; Kamleh *et al.*, 2008)

1.9 Liquid chromatograph mass spectrometry (LC-MS)

Hyphenated techniques such as gas chromatography (GC), liquid chromatography (LC) combined with mass spectrometry have played a major role in the progress of

metabolomics studies due to their high sensitivity and specificity. These combined methods are able to detect and quantify, under optimal conditions, the majority of metabolites predicted for simple organisms. MS records the molecular mass of compounds and their daughter ions resulting from fragmentation under the MS conditions. It involves the ionization of compounds and fragmentation of the ions produced into smaller units whose m/z values are quantified by a detector. The structure and chemical nature of a compound influences its ionization and fragmentation hence a unique spectrum of mass fragments is generated for the compound and can be used for its identification. Liquid chromatography–mass spectrometry (LC-MS) is the most commonly used technique in metabolomics. It is able to analyse high molecular weight compounds (>600 Da) such as phospholipids, glycosides and sugars.

1.10 Liquid chromatography (LC)

Generally chromatographic techniques are used for the separation of compounds in a mixture based on differential affinities of the analytes for the mobile and the stationary phases (Steehler, 2009).The advantages of using chromatography in combination with mass spectrometry are in decreasing ion suppression effects and the ability to differentiate between isomers. The modern High Performance Liquid Chromatography (HPLC) system consists of solvent reservoirs, an online degasser, a pump, an autosampler, column, and a suitable detector (Figure 1.2). Since the separation would be affected by mobile and stationary phases, the mobile phase

should have a suitable modifier or buffer and the column should have an appropriate stationary phase. There are various types of detectors that can be used with a HPLC and these could be ultra violet, diode array detectors (DAD), evaporative light scattering detector (ELSD) and mass spectrometer (MS) (Harris, 2010). Each of these detectors are associated with certain strengths and limitations. For example, ultra violet detectors are unable to detect compounds lacking chromophores while ELSD has limited quantitative capacity. The mass spectrometer in combination with a HPLC offers the most powerful and reliable analytical platform for metabolomics studies (Steehler, 2009).

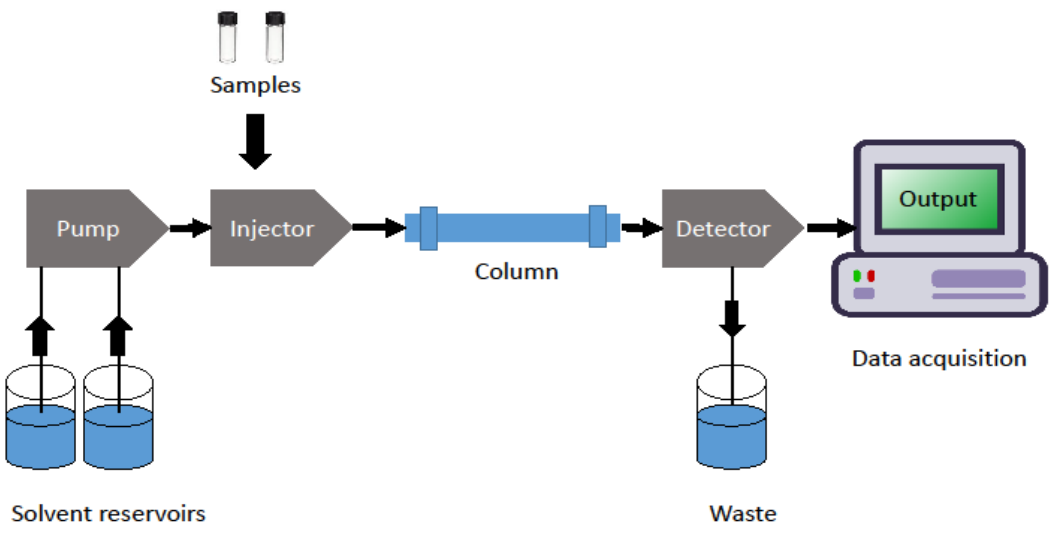


Figure 1.2 A schematic diagram to illustrate the components of an HPLC system

There are three major types of liquid chromatographic technique; reversed phase (RP) chromatography, normal phase (NP) chromatography and hydrophilic interaction liquid chromatography (HILIC) based on the type of the stationary phase used. A reversed phase technique is most widely employed in liquid chromatographic separation. The columns are usually silica-based or monolithic and they can be derivatised with different ligands to make them suitable for RP, NP or HILIC chromatographic separations. Reversed phase chromatography is a suitable technique for the analysis of metabolomic samples especially lipophilic compounds, because they are eluted in order of their lipophilicity, hence it is very suitable for most drugs in biological systems. The stationary phase in RP is hydrophobic such as an octadecdyl (C18) column. A hydrophilic mobile phase in which water is mixed with a miscible organic solvent such as methanol, ethanol, acetonitrile, or tetrahydrofuran is used (Harris, 2010) . The disadvantage of RP chromatography, however, is the presence of ion suppression and interference caused by phospholipids which are strongly retained during the chromatographic run. This problem can be overcome by washing the column with a high level of organic solvent following each run. In addition, RPC is not suitable for highly polar metabolites that have little retention in the RPC column as they may elute at the void volume and are thus not subjected to chromatographic separation. Hydrophilic interaction liquid chromatography (HILIC) which does not have the problems of RP chromatography has been introduced and it is gaining acceptance in metabolomics studies.

The HILIC mechanism of retention depends on a water surface layer (pseudo–stationary phase) associated with a zwitterionic or polar surface coating on the

column particles. Therefore, it gives higher retention of hydrophilic metabolites and low retention for hydrophobic ones. The use of ZIC-pHILIC, a hydrophilic interaction liquid chromatography methodology is capable of adequately separating a wide range of polar compounds (Gika, Wilson and Theodoridis, 2014). This chromatography technique was at first used in the separation of polar analytes including amino acids and polar drugs rather than using RP-HPLC which is made up of a low-aqueous/high-organic mobile phase (Hemström and Irgum, 2006).

HILIC retains compounds using partitioning between an organic mobile phase and a hydrophilic stationary phase while the elution is driven by increasing the water content in the mobile phase. The surface layer of water on the stationary phase is considered a pseudo-stationary phase and it is this in combination with a polar surface or zwitterionic groups on the column that enables retention in HILIC mode. The chief aspect of the zwitterionic coating found in HILIC columns is that its net charge is neutral but that it also has the ability to separate molecules that are both positively and negatively charged through the interaction of the individual charged groups with the analytes. In opposition to RPC, the use of HILIC chromatography is increasing in popularity every day and this is attributed to its highly efficacious separation efficiency which is due to favourable mass transfer through the organic mobile phase which is characterised by a low level of viscosity. Moreover, HILIC is more suitable for use with LC-MS as it improves the ionisation efficacy which shows a greater degree of efficiency in the low viscosity mobile phase. The previous years have seen a greater need for the examination of biochemicals that are polar in nature

such as oligosaccharides and proteins and this has resulted in the improvement of the HILIC methodology (McCalley, 2017).

Traditionally, the common practice in metabolomics involves the use of two chromatographic methodologies including both RP and HILIC to achieve separation of polar and non-polar molecules in any one sample. In doing this, though, more issues arise through sample preparation (Hemström and Irgum, 2006). Moreover, the requirement that samples need to be processed on two individual columns continuously means that both the analysis time and the complexity of the data are increased.

1.10.1 Mass spectrometry (MS)

The mass spectrometer has three main components: an ionisation chamber, a mass analyser, and a detector (Figure 1.3). The ion source is used to produce ions in the gas phase. The ionisation processes use various techniques and they include electron impact (EI) and chemical ionisation (CI) which are carried out under vacuum, while electrospray ionisation (ESI) and atmospheric pressure chemical ionisation (APCI) which can be carried out at atmospheric pressure (Kraj, Desiderio and Nibbering, 2008). The ions produced by the ion source are accelerated through a region of electric and/or magnetic fields so that only those ions with mass-to-charge (m/z) in a given range can reach the analyser and be detected. Modern mass spectrometer systems are suitable for metabolomics studies since they use soft ionisation techniques such as APCI or ESI which result mainly in molecular ions without fragmentation, allowing the compounds to be identified based on information on

databases, constructed specifically using accurate mass data of the common metabolites (Watson and Sparkman, 2007).

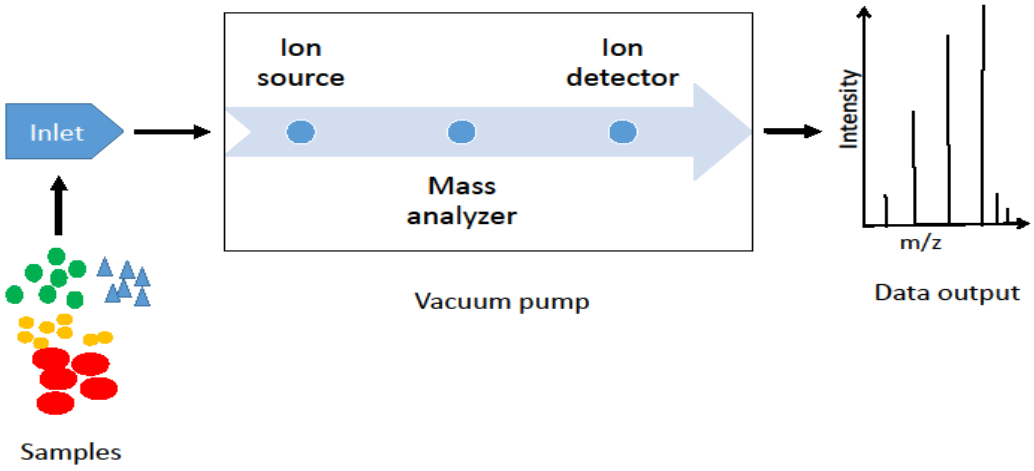


Figure 1.3 A schematic diagram to show the main components of a mass spectrometer.

The second main component of a MS instrument is the analyser, where the ions are separated based on their m/z ratio. Different mass analysers are currently used in MS systems to separate the ions in time or space; these include quadrupoles (Q), ion traps (IT), time-of-flight (TOF), Fourier transform ion cyclotron resonance (FT-ICR), and the Orbitrap analyser (Pitt, 2009) . The main differences between these mass analysers arise from their resolving power, accuracy, sensitivity, dynamic range and fragmentation capabilities for MS^n studies. Recent developments have led to hybrid

mass systems that combine the strengths of various mass analysers so that a single mass spectrometer can have various capabilities based on the ion separation techniques being employed in the hybrid system. Examples of such hybridised mass analysers include triple-Q, Q-IT, TOF-TOF, Q-TOF, IT-Orbitraps, LTQ-Orbitraps and Q-Exactives (Michalski *et al.*, 2011; Gallien *et al.*, 2012).

In the third part of a mass spectrometer which is the detector, the mass-to-charge ratio (m/z) of the detected ions and their abundances are measured. In the Orbitrap, for example, detection is based on image current of the ions in the mass analyser (Eliuk and Makarov, 2015).

1.11 Data extraction and metabolites identification using Mzmatch and IDEOM

LC-MS data are usually processed in a 3D-matrix comprising of m/z, retention time and intensity. There are several software for processing these data and deconvolving them into a matrix of detected peaks, metabolite identification (ID), with the peak response for the metabolites determined. The software should be able to align retention times and accurate masses that drift due to the order of injections. MzMatch is an open resource and platform software that is used to process raw LC-MS data. The data are processed with conversion from instrument-specific data format to XCMS Centwave for peak picking. MzMatch is used for noise filtering, peak detection and alignment, then identification is carried out by IDEOM. MzMatch is applied in R statistical language (Scheltema *et al.*, 2011). In a single experiment MzMatch can analyze over a hundred LC-MS data files with many groups of

experiments and can compare more than two sets. Database retention time is updated in each experiment by RT calculator using the Quantitative Structure Retention Relationships (QSRR) approach to predict retention times based on known retention times of standards and the physicochemical nature of the interactions of analyte with columns that determine degree of retention (Creek *et al.*, 2012). QSRR is a technique capable of improving the identification of a compound by predicting its retention time when analyzed by liquid chromatography. It is aims to predict the retention for solutes by identifying the most important structural descriptors relevant to the retention behaviour of the solute coupled with an understanding of the molecular mechanism of separation operating in a given chromatographic system (Goryński *et al.*, 2013).

A Microsoft Excel template IDEOM is used for automated data processing of high resolution LC-MS data obtained from untargeted metabolomics studies (Creek *et al.*, 2012). Under the IDEOM platform, extensive noise filtration is carried out and the standards are matched with the sample metabolites. It is sometimes necessary to update the retention times in the database with a list of retention times from standard runs in each experiment; this list is created using ToxID^{TM} (which is an automated compound identification tool that dramatically simplifies processing of LC-MS data and identifies compounds according to retention times and elemental composition. The retention time calculator also uses physiochemical properties (depending on the functional group and chemical nature of the compounds) in the data base sheet to predict retention times based on a multiple linear regression model with the authentic standards.

1.12 Analysis of metabolomics data

1.12.1 Univariate data analysis

Univariate analysis (UVA) is considered the simplest form of dealing with metabolomics data. It considers only one variable at a time and it can be inferential or descriptive, but does not deal with relationships such as regression analysis. Univariate data analysis includes tests to compare different sets of samples such as ratio, t-test and Analysis of variance (ANOVA). In metabolomics, tens or hundreds of variables are produced, and one of the objectives is to examine the relationship between the metabolomics change and the intervention. Therefore, metabolomics data analysis needs significant testing and collection of a huge number of variables to reduce false positives (Saccenti *et al.*, 2014).

1.12.2 Multivariate data analysis

Biological systems and individuals are well described by genomics, proteomics and metabolomics. Relating the large quantity of data on many different individuals to their current (and possibly even future) phenotype is a task not well suited to classical multivariate statistics.

The datasets generated by metabolomics techniques very often violate the requirements for classical multivariate analysis (MVA) such as multiple regression, samples (N) must be greater than variables (K), the K variables should be noise-free and uncorrelated and the X-matrix should be complete without any missing values. For MVA, K can be much larger than N, the K variables can be multicollinear and the

X- matrix noisy and incomplete i.e with missing values. However, another statistical approach exists as an alternative to classical statistical treatments (that was developed in the early part of this century by Hermann Wold and colleagues) that can overcome these problems . This approach, called multivariate analysis (MVA), has the potential to revolutionise medical diagnostics in a broad range of diseases. It opens up the possibility of expert systems that can diagnose the presence of several diseases simultaneously, and even make predictions about any diseases that an individual is likely to suffer in the future (Worley and Powers, 2013).

There are two steps involved in multivariate analysis followed by univariate analysis (Kirwan *et al.*, 2012). The multivariate step consists of two phases. The first is the use of pattern recognition by using unsupervised techniques in order to get an overview of the data and to ensure that it does not contain outliers. The second step is biomarker identification followed by model validation to ensure predictive ability. Data should be pre-processed prior to starting the steps of data visualisation and biomarker identification.

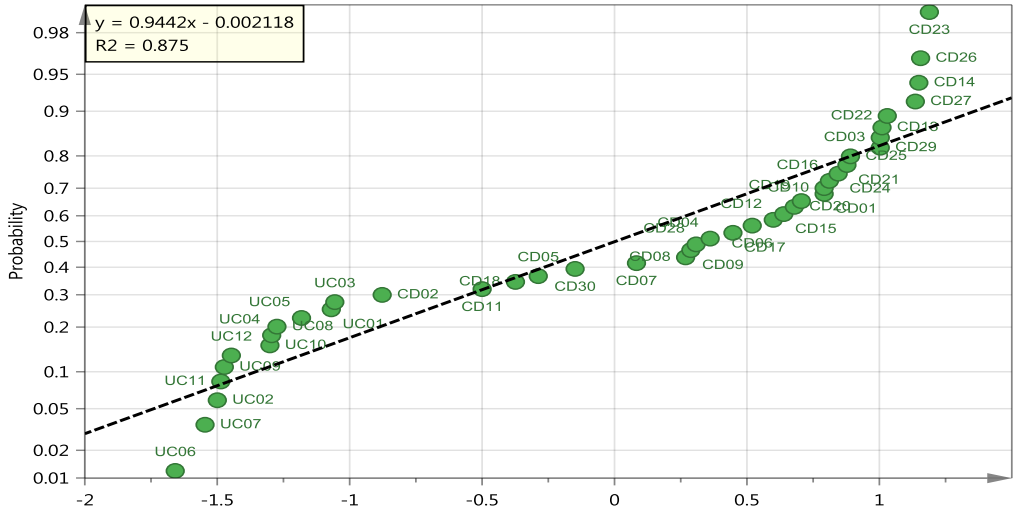
1.13 Data pre-treatment

1.13.1 Transformation

In some data, the variables can deviate from a normal distribution. In such cases there is a need to ensure the data can move towards normality. This is termed transformation. In other words, through transformation the residuals are increased in normality which in turn aids in the exclusion of outliers (Eriksson *et al.*, 2013). There

are various types of transformations including Log_2, Log_{10}, power, inverse, etc. Their use is reliant on the data that needs to be transformed. An assessment of this can be performed through normal probability plots as indicated in (Figure 1.4) below (Eriksson *et al.*, 2013). In this figure it is clear the observations in plot B are situated on the straight line with R2 =0.97 after log_2 transformation compared to the untransformed observations in plot A which deviate from the straight line with R2 =0.87.

A



B

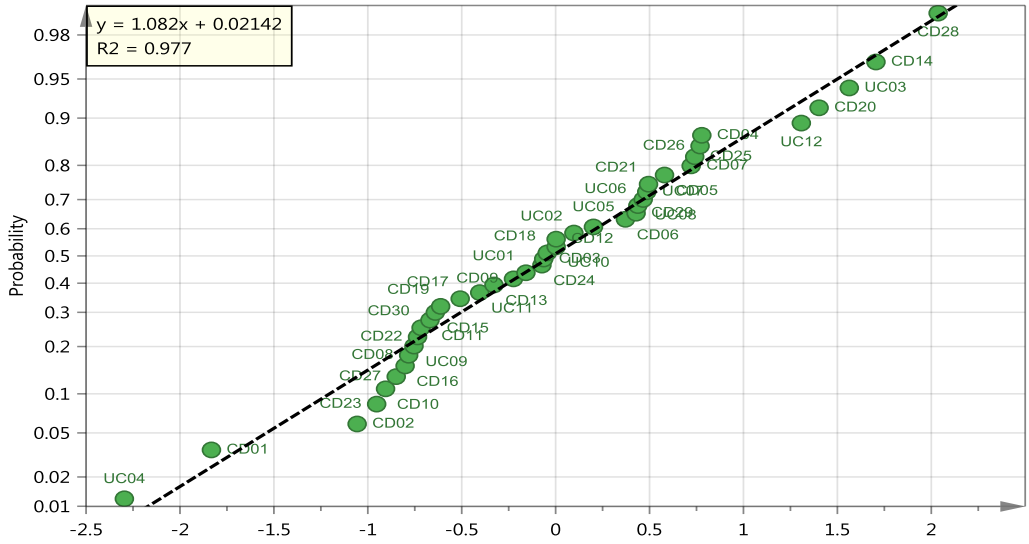


Figure 1.4 Normal probability plot of residuals.

The above is an illustration of the normalised residuals as seen on a double Log scale on the y-axis against a standard deviation which is observed in the x-axis. The outliers include any data that is greater or smaller than -4 or +4 standard deviations. Both plots include a regression line for indication of data normality with (A) indicating the data prior to transforming the variables and (B) the normalised variables which was achieved through Log_2 transformation

1.13.2 Scaling

The focal point of statistical analysis is the lower weight given to low intensity metabolites in comparison with high intensity metabolites. That said, the smaller metabolites should not be disregarded as they may hold significant biological roles. Such issues are normally resolved in metabolomics using a method termed scaling (Xi *et al.*, 2014). To perform scaling various methods may be performed: [1] through mean centering which involves attaining each variables' average and then subtracting this from the intensity of each row's variable; [2] through univariate scaling where the mean and the standard deviation of the features are calculated. The variables are first mean-centered. Thereafter each number in the mean-centered values is divided by the standard deviation. When assessing variables with different units, scaling is the most suitable method although it may result in higher noise variables. Such detrimental effects can be reduced once again through [3] Pareto scaling as suggested by Xi *et al.*, (2014) . In this latter method each variable's mean centered value is divided by the square root of the standard deviation intensity (Karaman, 2017). This methodology has increased use in tackling spectroscopic data (Xi *et al.*, 2014). [4] Block weighting, is another scaling technique where a variable in each row is multiplied by $1/(k_{block})^{1/2}$, where k_{block} = number of variables in that block.

Univariate scaling together with block weighting are the most commonly used methodologies to evaluate cases where the variable's units are varied and where the case is taken over by a block of large values overshadowing blocks with smaller values as occurs in the case of height (m) and systolic blood pressure (mmHg) (Eriksson *et al.*, 2013).

1.14 Data visualisation

1.14.1 Unsupervised Techniques

The amount of data produced from LC-MS can be high in the case of biological data sets (e.g. metabolomics) and this is attributed to the nature of this data. As is the case in this study, where there is a need to identify associations between the different variables, it is clear that the bigger the data generated, then the greater the degree of complexity and challenge posed in generating the necessary findings. It would be almost impractical to perform a thorough assessment and evaluation of the information without suitable statistical packages. Thus, it is crucial that the statistical techniques employed are appropriate for the task in hand so that there is greater opportunity for establishing possible parallels or variations existing between the different data samples. This would be achieved by decreasing the data dimensionality of the input space so that the number of dimensions dealt with is small.

The samples analysed would have to be grouped into sets with similar elements. This would help in an understanding of the underlying issue. To accomplish this, statistical methodologies used would include Cluster Analysis and Principal Component Analysis (PCA). Although the samples in a set will possess similar elements, these elements will differ from the samples in the other sets. No data will be known for any set prior to the analysis and no assumptions will be made when allocating a sample into a group. Thus the pattern recognition technique employed would be one of unsupervised analysis. Such a technique aims to decrease the complexity of the data

and to then display the patterns or clusters identified in a graphical manner (Worley and Powers, 2013).

One example of unsupervised clustering is PCA. This methodology seeks to examine how clustering of variables occurs without reference to the set a sample relates to(Kirwan *et al.*, 2012). This is described as the chief methodology employed by researchers for minimisation of data for the attainment of beneficial findings (Yamamoto *et al.*, 2009). The methodology involves the collation of variables that are associated into a small number of their underlying variables (components). The greater the association identified between the samples then the lower the number of components required where component numbers are less than that of the observation—this will be performed by avoiding loss of a high amount of the total variation present in the data. Regarding metabolomic data analysis, the initial stage tends to involve PCA (Kirwan *et al.*, 2012). This allows the data to be observed and for outliers to be identified.

1.14.2 Supervised Techniques

Despite establishing an outline of the data sets analysis, PCA does not establish associations of the phenotype to disease status, for example as would occur in associating any individual with the quantified variables. It is possible to use Partial least squares-discriminant analysis (PLS-DA) which is a PCA analysis that is carried out on the Y-matrix and therefore on the observations/samples. This establishes a low latent variable number and then allows the formation of a set of latent variables using the X-matrix. Thus, through use of descriptors/variables/metabolites. This would aid

in the understanding of the greatest level of variance in the Y-matrix derived latent variables.

A further addition to the PLS-DA model can be found in Orthogonal partial least squares - discriminant analysis (OPLS-DA). This has the added benefit of being able to distinguish variation in X that can be associated with Y (horizontally). Such variation is referred to as predictive variation. This methodology can also distinguish X variation that is not associated to Y (orthogonal) as illustrated in (Figure 1.5). OPLS-DA was described as the strongest methodology for the assessment of between groups variation (Kirwan *et al.*, 2012). It can establish dependable biomarkers with powerful correlation with between groups separation (Trygg, Holmes and Lundstedt, 2007). It can also link disruptions in metabolic pathways to diseases (Goodacre, 2007) and therefore it can enhance our comprehension of the pathophysiological state and possible treatment targets for later development.

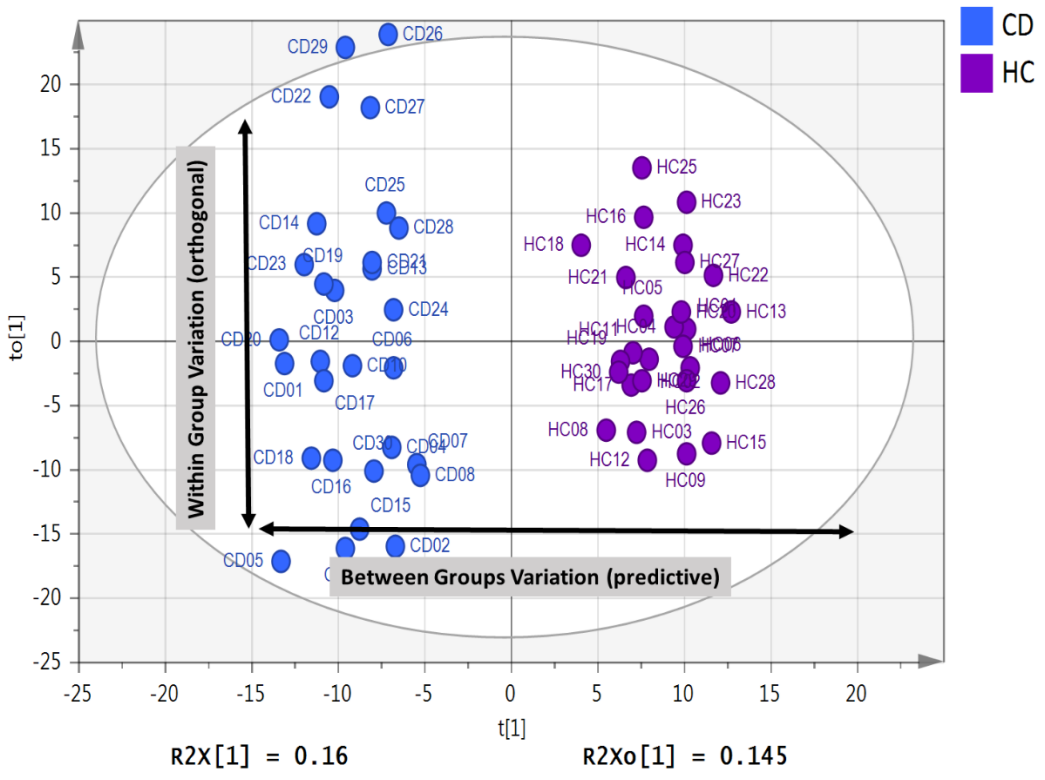


Figure 1.5 Score plot of Orthogonal Partial Least Square Discriminant Analysis (OPLS-DA)

The OPLS-DA score plot illustrated demonstrates faecal extract samples for Crohn's patients (CD) (blue) or healthy control subjects (HC) (purple). It is evident that a horizontal separation is formed between the two groups (t-predictive). Moreover, a greater within group variability (to, orthogonal) is observed in those in CD group compared to HC group. This model enhances understanding of 14.5% of the between groups (predicative) variation and 16 % of the within group (orthogonal) variation.

1.15 Model validation

1.15.1 Model parameters

The strongest tools for validation of an applied model used during an analysis are R^2 and Q^2. R^2 allows the goodness of fit to be described quantitatively by associating between the observations (y) and the variables (x) through quantification of the portion of observations that are clarified by the variation in the variables. One

particular problem that arises with this parameter is the possibility of it being established near the value 'one' which is its maximal value and this is most likely as the component number is raised. Such a result would lead to over-fitting of the data as a high variable number would be contrasted against a low number of observations which would lead to findings that are more positive than is realistically the case. To manage such a situation the prediction parameter Q^2 is required to be accurate. A cross validation (CV) is conducted to attain Q^2 (Kirwan *et al.*, 2012). In doing so a predefined number of observations should be eliminated and a model produced re-fitting re-run. This procedure should be performed on all the data until the point that all data have been eliminated on a single occasion (Eriksson *et al.*, 2013). Next, a comparison should be made between R^2 and the mean value of the refitted model's Q^2. This would demonstrate that any chance occurrence is identified more easily. For the purpose of cross validation SIMCA P software - by default - leaves 1/7th of the data out. An observed vs predicted plot is employed to examine the efficiency of CV, by which the R^2 of the regression line should be improved (Triba *et al.*, 2015).

1.15.2 Permutation test

Permutation tests are employed to determine whether the way that the observations were grouped in both the designed sets has greater significance than that which could be achieved through other random groupings in any two random classes (Westerhuis *et al.*, 2008; Worley and Powers, 2012). This test involves a comparison of the R^2 and Q^2 parameters derived from the initial model and their comparison to those derived from the permuted model. Repetition of this procedure would lead to the production of new quality parameters. All parameters derived from the

permutation should signify lower values compared to the original ones. Moreover, for the predictive model, the regression line should cut through the zero line (horizontal) as observed in (Figure 1.6) (Eriksson *et al.*, 2013).

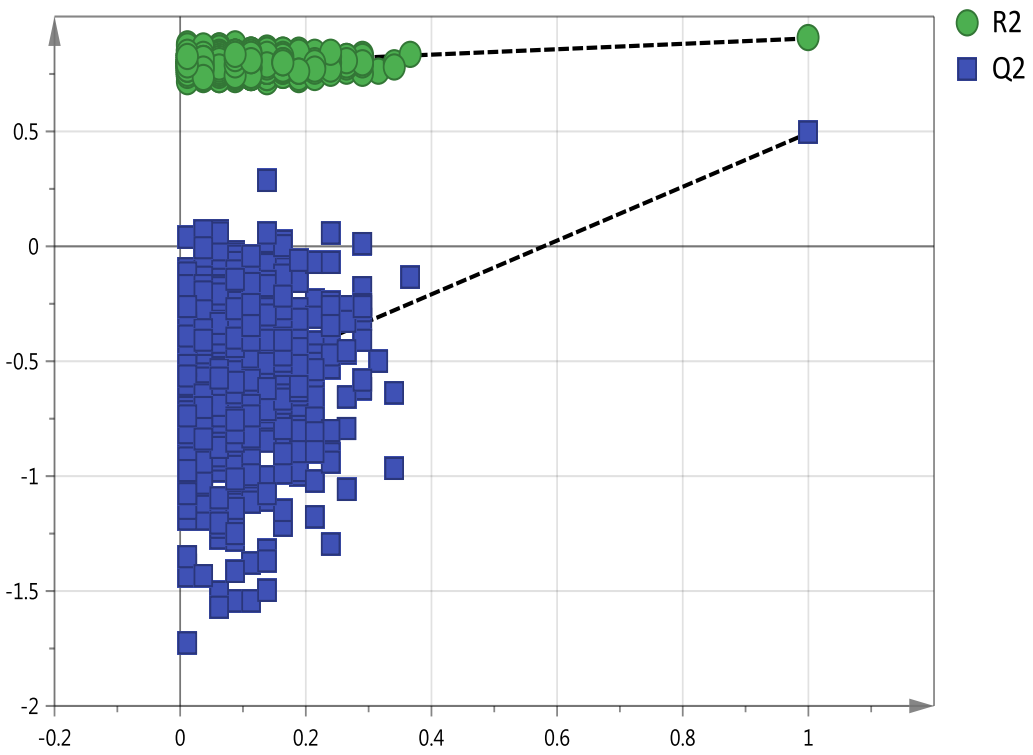


Figure 1.6 Permutations test

The plot shows, the vertical axis gives the R^2Y and Q^2Y -values of each model. The horizontal axis represents the correlation coefficient between the original Y, which has correlation 1.0 with itself, and the permuted Y. If the supervised model has valid predictive ability, the R^2Y and Q^2Y of the real model are always larger than the corresponding values of the models fitted to the permuted responses

Performance of ANOVA on the cross validated residuals (CV-ANOVA) is performed as a means of establishing the significance of the predicted variation for the supervised model. Following validation of the model's predictive ability, the researcher should

seek to establish the model's degree of precision in distinguishing between observations on the basis of their metabolic profile. This should then be detailed through employment of an area under the receiver operating characteristic (ROC) curve.

1.15.3 Cross validated ANOVA (CV-ANOVA)

Supervised model validity as well assessed using cross validated ANOVA (CV-ANOVA) which tests the variation predicted by the model against H_0 hypothesis of equal cross validated predictive residuals around the mean (Triba *et al.*, 2015). It is a diagnostic tool for assessing the reliability of the model. The advantages of using the CV-residuals are that no extra calculations are needed and that this procedure secures reasonably independent data

1.16 Variable importance in the projection (VIP)

The contribution of each metabolite in a given model is examined by considering the variable importance in the projection (VIP). This parameter estimates and ranks the importance of each variable (metabolite) in the projection and it is often used for variable selection during metabolomics (Chong and Jun, 2005) . Metabolites are generally considered to have a high contribution in the model if VIP > 1 (Eriksson *et al.*, 2013; Zhang *et al.*, 2016). VIP, provides a value for each metabolite in terms of its contribution to the difference between groups (VIPpred) and its contribution to the within group variability (VIPortho). Metabolite with high VIPpred and low VIPortho values is sensitive and specific.

Chapter 2:

Materials and Methods

2 Materials and Methods

2.1 Solvents and chemicals

Chemicals and Solvents were High-performance liquid chromatography (HPLC) grade. Acetonitrile (ACN) was purchased from Fisher Scientific (Loughborough, UK), and HPLC grade water was produced by a Direct-Q3 UltrapureWater System (Millipore, Watford, UK). AnalaR-grade formic acid (98%) was obtained from BDH-Merck (Poole, UK). Authentic stock standard metabolites (Sigma-Aldrich, Poole, UK) were prepared as previously described (Zhang *et al.*, 2014) . Each metabolite standard was prepared at 1 mg/ml with HPLC grade methanol and water (1:1, v/v) as the stock solution and stored at -20°C. 100 µl was taken from each stock solution, about 58 metabolites were mixed and then the solution was made up to 10 ml with acetonitrile. Consequently, the final concentration for each metabolite standard was 10 µg/ml and 348 metabolite standards were distributed into six mixed metabolite standard solutions. In order to avoid identity confusion, isomers were distributed into different standard solutions and in-source fragments were also carefully verified since they could be mistaken for another metabolite.

2.2 LC-MS Analysis

2.2.1 Mobile phase solutions for ZIC-pHILIC chromatography

Mobile phase solvents were freshly prepared and stored at room temperature for up to 48 h. Mobile phase A: ammonium carbonate buffer (20 mM, pH 9.2) was prepared by the addition of 1.92 g of ammonium carbonate to 800 mL of HPLC-grade water, followed by an adjustment to pH 9.2 with ammonia solution and then filled to a

volume of 1 L. Mobile phase B: HPLC-grade acetonitrile only. The samples were eluted from a ZICpHILIC column (150 × 4.6 mm, 5 µm particle size) fitted with a ZICpHILIC guard column supplied by Hichrom Ltd. (Reading, UK) with a mobile phase consisting of 20 mM ammonium carbonate in HPLC-grade water (solvent A) and acetonitrile (solvent B), at a flow rate of 0.3 mL/min. The elution gradient was an A:B ratio of 20:80 at 0 min, 80:20 at 30 min, 92:8 at 30 min 92:8 at 35 min, 20:80 at 36 min, and 20:80 at 45 min (Table **2.1**).

Table 2.1 Gradient elution programme applied for ZICpHILIC in LC-MS analysis

Time (min)	Mobile phase A%	Mobile phase B %	Flow rate (ml/min)
0	20	80	0.3
30	80	20	0.3
31	92	8	0.3
36	92	8	0.3
37	20	80	0.3
45	20	80	0.3

2.2.2 Mobile Phase for ACE C4 Chromatography

An ACE C4 column was used to estimate the levels of unsaturated fatty acids. The mobile phase for the elution of the ACE C4 column consisted of 1 mM acetic acid in water (A) and 1 mM acetic acid in acetonitrile (B) at a flow rate of 0.4 mL/min. The elution gradient was as follows: A:B ratio 60:40 at 0 min, 0:100 at 30 min, 0:100 at 36 min, 60:40 at 37 min, and 60:40 at 41 min (Table **2.2**).

Table 2.2 Gradient elution programme applied for ACE C4 column in LC-MS analysis

Time (min)	Mobile phase A%	Mobile phase B %	Flow rate (ml/min)
0	60	40	0.4
30	0	100	0.4
36	0	100	0.4
37	60	40	0.4
41	60	40	0.4

2.2.3 HPLC setup

The HPLC was fitted with the appropriate mobile phase components. The auto-sampler needle and sample syringe were flushed with the syringe wash solution (Methanol: Water, 1:1). Initially, the system was flushed with 100% of mobile phases B followed by 100% of mobile phase A at a flow at 5 ml/min for 5 min in both mobile phases. The drain valve was then closed and the outlet tube was disconnected from the mass spectrometer. After that, the selected HPLC column was conditioned with 50% of mobile phase B at a flow rate of 0.3 ml/min for 10 min. The operating pump pressure was continuously monitored to ensure that it was below 2,000 p.s.i. Chromatographic separations were performed on both ZIC-pHILIC and ACE C4 column columns by applying two separate linear gradient elutions over 30 min (excluding re-equilibration, as shown in (Table **2.1**) and (Table **2.2**) using the mobile phases described in sections 2.2.1 and 2.2.2 above respectively at a flow rate of 0.3ml/min. While on the instrument, samples were kept on a vial tray which was set to a constant temperature of 4˚C to avoid any possible degradation of samples.

2.2.4 Orbitrap Exactive MS setup

LC-MS was performed with an Accela HPLC pump connected to an Exactive (Orbitrap) mass spectrometer from (Thermo Fisher Scientific, Bremen, Germany). The quality of data acquired from an instrument has an implication on the accuracy of the deductions that can be made from a study as a whole. In this experiment, the quality of data was ascertained by using standard mixtures run with each set of samples to assess parameters such as peak width, height, retention time, and chromatographic resolution. The relative standard deviations (RSDs) of these parameters were checked to ensure that they did not vary by more than 20% for each of the standards. The retention time shifts in the data obtained at the beginning and at the end of a given sequence was expected not to be more than 0.3 min. When this condition was violated, the HPLC system was checked for any leaks before the use of a new column was considered. If any defects were found in the instrument, analysis was postponed until the system was serviced. Instrument sensitivity was assessed weekly and any residues in the ion source chamber were removed to maintain enhanced sensitivity. This was done by sonicating the sample cone and the ion transfer capillaries in a 50:50 (vol/vol) methanol/water solution for 15 min.

The Thermo Calmix standard solutions were used to tune and calibrate the MS in based on the manufacturer's specifications. The signals of acetonitrile dimer (2xACN+H) m/z 83.0604 and m/z 195.03765 for caffeine were used as lock masses for positive ion electrospray ionization (PIESI) mode and m/z 91.0037 (2 x formate-H) was used as a lock mass for negative ion electrospray ionization (NIESI) mode, during each analytical run. The MS accuracy was tested using standard analytes with

intensities between 104 and 107 as calibrants. The calibrant peaks were checked to make sure that the mass deviations were less than 3 p.p.m, otherwise the instrument was recalibrated to correct the mass errors.

The electrospray ionisation (ESI) interface was operated in both positive and negative modes. The spray voltage was 4.5 kV for the positive mode and -4.0 kV for negative mode, while the ion transfer capillary temperature was 275 °C. Full scan data was obtained in the mass-to-charge range of m/z 75 to m/z 1200 for both ionisation modes. The MS system was fully calibrated prior to running the samples according to the manufacturer's guidelines. The nitrogen sheath and auxiliary gas flow rates were maintained at 50 and 17 arbitrary units. The resulting data was acquired using the XCalibur 2.1.0 software package (Thermo Fisher Scientific, Bremen, Germany).

2.3 Metabolomic profiling

2.3.1 Statistical softwares used

All data analysis, including data visualisation, biomarker identification, diagnostics and validation, was implemented using SIMCA software v.14 (Umetrics AB, Umeå, Sweden) for multivariate analysis (Zhang *et al.*, 2016). Metaboanalyst 4.0 (www.metaboanalyst.ca) (Chong *et al.*, 2018) and Minitab Statistics software package version 18 (State College, PA: Minitab, Inc.) were used for univariate analysis.

2.3.2 Data Pre-processing and Modelling

The data was extracted by using MZ Match software (version 1, http://mzmatch.sourceforge.net/) (Scheltema *et al.*, 2011), and the identification of putative metabolites was made via the macro-enabled Excel file, IDEOM

(http://mzmatch.sourceforge.net/ideom.html) (Creek *et al.*, 2012). The lists of the metabolites obtained from these searches were then manually evaluated by considering the quality of their peaks and their retention time match with the standard metabolite mixtures run in the same sequence. All reported metabolites were within 3 ppm of their exact masses. The Excel sheet output provided from Mzmatch was pre-processed to improve data quality.

The RSD ((standard deviation/mean) × 100)) for each of the metabolites was calculated using quality control (QC) samples, and the metabolites were excluded from the analysis if the RSD was > 30%. Metabolites were also excluded if the missing values were more than 20% in the biological samples. The remaining metabolites were then transformed using log base 2 to reduce data skewing and improve data normality (van den Berg *et al.*, 2006). The multivariate analysis and data mining were carried out using SIMCA-P software v.14.1 (MKS Umetrics AB, Umeå Sweden). The data were Pareto scaled, which divided each metabolite intensity by the square root of its standard deviation (Shaffer, 2002).

Then, unsupervised principal components analysis (PCA) was used to evaluate the QC samples and exclude technical errors. After the data was transformed and Pareto scaled, the groups were defined, and a supervised OPLS-DA model was applied to all metabolites. In this model, the variation was divided into two analyses. The first was a prediction variation, which is the correlated variation between X and Y. This variation represents the inter group variation. The second analysis was an orthogonal variation which is orthogonal to the first analysis and the uncorrelated variation

between X and Y. This variation analysis represents the intra group variation (Blasco *et al.*, 2015).

2.3.3 Model Validation

The next step was to evaluate the separation between the groups and to start the group comparisons. The model parameters cumulated the amount of variation in matrix X R^2X (cum), R^2, and Q^2, and a permutation test was examined to evaluate the model's validity. The significant differences in the model were assessed by calculating the *p*-values from the cross-validation analysis of variance (CV-ANOVA). A p value of 0.05 was used as the significant value. The difference between R^2 and Q^2 ($R^2 - Q^2$) was calculated to reduce the possibility of overfitting in the supervised model (Eriksson *et al.*, 2013). If $R^2 - Q^2 > 0.3$, the model would be considered over-fitted and therefore invalidated. The significance of the model was also evaluated using a permutation test (Worley and Powers, 2012). The same procedure was repeated in this study 999 times (the maximum threshold in the SIMCA-P software version 14.1), and the parameters were compared to the original data parameters. The model was considered valid if the Q^2 regression line crossed the zero line or if no Q^2 value from the permutated data set was more than the Q^2 from the original data set. The significance of the group separation was assessed by using the *p*-value provided from the CV-ANOVA (Eriksson, Trygg and Wold, 2008; Wheelock and Wheelock, 2013). SIMCA-P produced this test based on a cross-validated model.

2.3.4 Data Filtration

In this study, several steps were applied to exclude metabolites with unreliable data points. The first filtering step was the *p*-value provided from the Student's t-test. Metabolites with *p*-values > 0.05 were excluded from the list. The remaining metabolites were filtered using jack-knifing uncertainties. This filter evaluates the precision of each metabolite by estimating the prediction error rate after cross-validation. It can be provided by calculating the 95% confidence intervals (95% CI) from the supervised model (Efron and Gong, 1983). Metabolites which registered zero within the 95% CI were excluded from the list. The significant metabolites were transferred to Metaboanalyst (http://www.metaboanalyst.ca/) to compute the corrected *p*-value (q-value) and the area under the curve. The *p*-value was corrected using the Benjamini & Hochberg False Discovery Rate, and metabolites with q-values > 0.05 were excluded (Benjamini *et al.*, 2001). Area under the curves were tested for each of the significant metabolites, and the metabolites with areas < 0.7 were considered poor classifiers and excluded from the model (Eriksson, Trygg and Wold, 2008). A rough classification for areas under the curve is as follows: 0.9–1.0 = excellent classifier; 0.8–0.9 = good classifier; 0.7–0.8 = fair classifier; 0.6–0.7 = poor classifier; and 0.5–0.6 = failed classifier.

2.3.5 Ranking, Grouping and Confirmation of Significant Metabolites

The significant metabolites that passed the filtration steps were ranked by variable importance in the projection (VIP) values. VIP measures the contribution of each significant variable in the observed metabolomic change in a given model compared

to that of the rest of the variables (Eriksson *et al.*, 2013). Metabolites with a VIP total > 1 were considered to have high contribution levels to the model (Xia *et al.*, 2013). In addition, confidence intervals on the VIP column plot should be positive (Zhang *et al.*, 2017). The VIP values were divided into VIP predicted (VIPpred) and VIP orthogonal (VIPortho), where VIPpred represents the contribution of a metabolite to the difference between groups compared to the other metabolites, and VIPortho represents the contribution of a metabolite to the difference within groups compared to the other metabolites. The ratio of VIPpred/VIPortho was used in addition to the total VIP to evaluate metabolite contributions. Where VIPortho is > VIPpred a metabolite is not relevant as a biomarker. The overall workflow of the study is summarised in (Figure 2.1). The main work flow was divided into five main steps, starting with sample analysis and data generation (blue), followed by data pre-processing and modelling (orange), model validation (yellow), metabolite filtering (purple) and finally ranking, grouping and conformation of the significant metabolites (green). Q-Q tests were conducted in Excel.

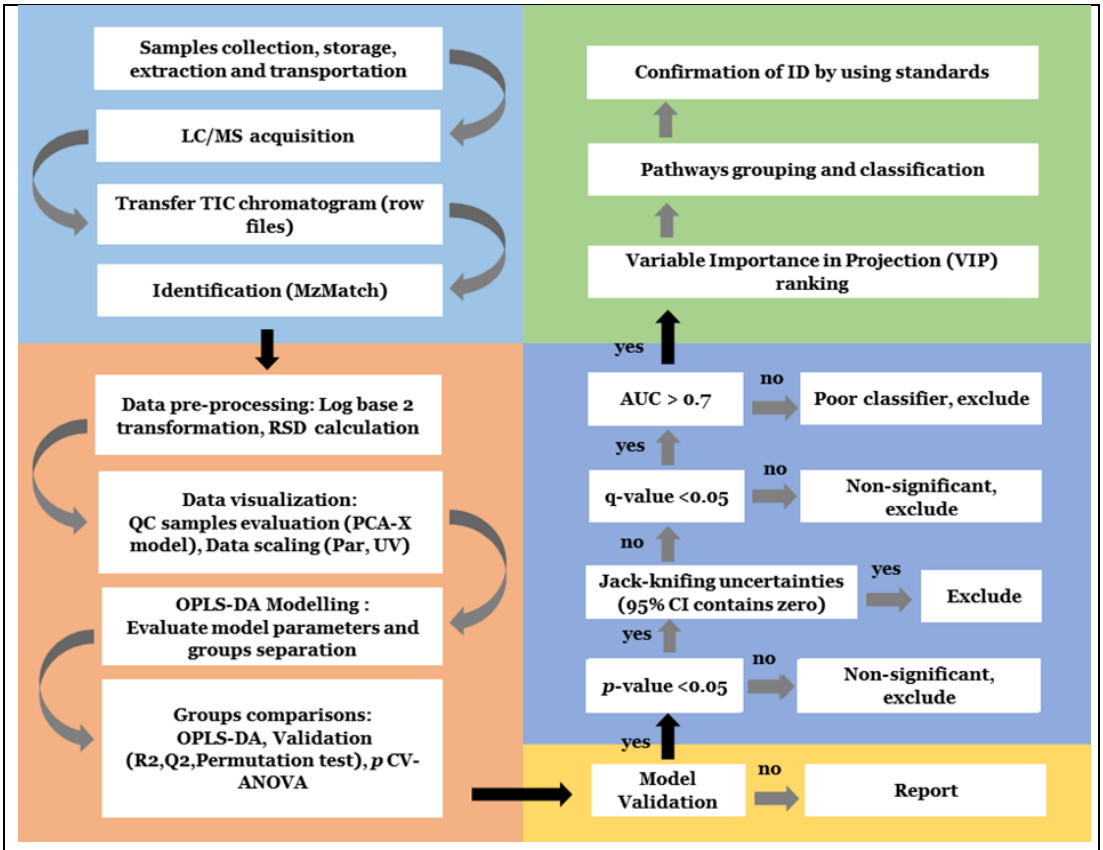


Figure 2.1 Flow chart for sample and data analysis

2.3.6 Data bases used for identification

In this thesis, different web sites and databases were used:

- KEGG Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/)
- HMDB Human Metabolome Data Base (http://www.hmdb.ca/)
- LIPID MAPS (http://www.lipidmaps.org/)

Chapter 3:

Data Pre-treatment and Statistical Model Selection

3 Data Pre-treatment and Statistical Model Selection

3.1 Abstract

Background and aim: Metabolomics significantly contributes to understanding in biology, diseases, and drugs research. Moreover, as metabolomics experiments typically produce high dimensional data, it is important to select the right bioinformatic tools to ensure efficient and high throughput data processing for removing systematic bias and exploring biologically significant findings. Missing value imputation may affect all of the following steps as well as the procedure of the data treatment and analysis. Its handling is an extremely important step in data pre-treatment, particularly in LC-MS metabolomic studies. This chapter aims to describe and select the best methods for pre-treatment of metabolomics data obtained after using liquid chromatography mass spectrometry (LC-MS).

Methodology: This study investigated the effect of pre-treatment strategies on data sets obtained from LC-MS based metabolomics experiment on 40 football players' urine samples (n=80). The data matrix consists of 2056 metabolites (i.e., 164480 values), and there were 43576 (approximately 26%) missing values. Different missing value were examined in conjugation with various scaling and transformation methods. SIMCA-P 14 was used to evaluate the model parameters for each pre-treatment method.

Results: Compared to the various imputation methods used in this study, NIPALS algorithm was significantly better according to the model parameter evaluation along with suitable transformation and scaling. Pareto (Par) scaling is better able to explain

the data than Unit variance (UV) with R^2X (Cum)=0.51. The result clearly shows that Log transformation was always at the top of each table based on R^2 (goodness of fit) considering all missing data.

Conclusion: Each step of data pre-treatment can significantly impact the model parameters, such as goodness of fit and goodness of prediction. In addition, it is necessary to further examine the data structure before conducting untargeted metabolomic analysis. The SIMCA-P 14's (NIPALS) default logarithm was the only imputation methodology that generated a valid model according to valid criteria (R^2-Q^2<0.3). This was in conjugation with Par or UV scaling with Par offering slight advantage over UV scaling. Log transformation was the best transformation methodology.

3.2 Introduction

3.2.1 Metabolomic data

Metabolomics analysis in clinical applications is often performed by either NMR or LC-MS (Yin and Xu, 2014). Metabolomics significantly impacts biology, diseases, and drugs research. The considerable diversity in metabolites' intensity and chemical properties in clinical samples creates the need for using various analytical techniques to survey the entire metabolome. Numerous methods have been developed to extract, detect, identify, and quantify the metabolome (Kim and Verpoorte, 2010). Moreover, as metabolomics experiments typically produce high dimensional data, it is important to use sophisticated bioinformatic tools to ensure efficient and high throughput data processing for removing systematic bias and exploring biologically significant findings (Atherton *et al.*, 2006). Thorough examination of the data handling may reduce the possibility of the final finding being misjudged and incorrectly interpreted (Hendriks *et al.*, 2005). Both multivariate statistical analysis and data visualization are critical in extracting relevant information and interpreting the results of metabolomics experiments.

3.2.2 Missing data imputation:

Missing value imputation may affect all of the following steps as well as the procedure of the data treatment and analysis. As stated by Sumner et al. 2007 missing values handling is an extremely important step in data pre-treatment, particularly in LC-MS metabolomic studies. In metabolomic studies, missing values typically remain as they are or they are replaced by Not a Number (NaN) which cannot be

distinguished from the real zero values of metabolite intensity rather than an analysis error (Gromski *et al.*, 2014).There are three major reasons for zero values appearing in output data, as stated by Di Guida et al. (2016). First, although a metabolite is identified in one sample, it is not included in the other samples in any concentration. Second, although a metabolite is included in one sample, its concentration is less than the detection limit of the analytical method. Third, although a metabolite is included in one sample in a higher concentration than the detection limit of the analytical method, it is not identified and reported by the data processing software.

Moreover, logarithmic error during deconvolution and identification may lead to an increase in missing values appearance in the output data sheet (Jenkins *et al.*, 2013).

3.2.3 Mean-centering

Mean-centring is commonly conducted for centring the data distribution in the multidimensional space's origin (Boccard, Veuthey and Rudaz, 2010). It should be noted that every row in a typical data matrix which is implemented to conduct multivariate statistical analysis stands for a different sample and that the metabolite intensities are arranged in particular columns. To execute mean-centring, every peak's (column) mean value is subtracted from the each sample's (row) corresponding value for a particular peak. Centring by subtracting the average (X-mean) is the SIMCA default(Eriksson *et al.*, 2013). It converts all the intensities into variations around zero rather than around the metabolite intensities' mean.

Mean-centring intends to eliminate offset from the data while focusing on the biological variation and on the data samples' similarities/dissimilarities. Further,

metabolites with high intensities tend to result in high values in the data table while displaying significant differences among samples as against the metabolites with low intensities (Boccard, Veuthey and Rudaz, 2010). Moreover, mass spectrometry platforms efficiently quantify both low intensity and highly intensity metabolites. Because PCA as well as OPLS-DA emphasises the maximum variance, it is not sufficient to only centre the data for determining biomarkers as the metabolites with high intensity significantly contribute to the model. Hence, it is possible to mask metabolites that are biologically important while having low abundance, and the statistical analysis results can become biased. Thus, centring is implemented as part of the pareto and unit variance scaling methodology.

3.2.4 Scaling

3.2.4.1 Unit Variance (UV) scaling or called autoscaling

In the case of no prior information regarding the variables' importance, it is suggested that all variables should be autoscaled to unit variance (UV), which can be considered the same as ensuring all variable axes are of the same length, thus giving equal importance to all variables. All metabolites, following autoscaling, have a standard deviation of one, because of which the data is evaluated as per correlations (van den Berg *et al.*, 2006).

UV is a well-known scaling method that metabolomics uses. The UV method involves calculating the mean as well as the standard deviation of the features. First, the intensities are mean-centered, and then the standard deviation divides every number in the mean-centered values (mean-centered/SD). It helps in ensuring that all

variables are given equal weight. Hence, metabolites that have low as well as high intensities can equally add to the multivariate model. On the other hand, it should be noted that a limitation of UV scaling is that intensities that are noisy and uninformative are also given equal importance as the interesting variables. In addition, there will be an increase in the metabolite measurement errors as they are impacted more. It is thus important to ensure that there is good quality of variables in the data table; that is, intensity with low repeatability/linearity or noisy intensity are removed. In case of considering NMR data analysis, according to Karaman, (2017), UV scaling is a more suitable option after noisy and outlying/contaminant regions are removed from the spectra.

3.2.4.2 Pareto scaling

Pareto scaling can be considered to be similar to UV scaling. However, as per Eriksson *et al.*(2013), the square root of the standard deviations is divided into every element in the mean-centered features in Pareto scaling (mean-centered/sqrt(SD)). Moreover, Pareto scaling can be regarded as a middle ground for mean-centering and auto-scaling as there is less dominance Pareto scaling by metabolites that have high intensities than the corresponding mean-centred ones. However, the Pareto scaled data are placed near the mean-centred data, and the limitations of only implementing mean-centring is also applicable to Pareto scaling. Hence, it is possible that multivariate analysis continues to be inclined to focus on metabolites having high abundance. Compared to unit variance, Pareto scaling is smoother and is able to enhance the significance of low intensity compounds while not increasing the noise considerably (Yi *et al.*, 2016).

3.2.5 Transformation

As metabolomics studies tend to be concerned with relative changes in metabolite levels, typically, a Log or other suitable transformation is applied before performing higher order statistical analysis. This is because a Log transformation helps eliminate heteroscedasticity from the data and ensure data normality (van den Berg *et al.*, 2006).

3.2.5.1 Log transformation

Log transformation is easily performed in SIMCA-P 14 using the transformation function. It also helps reduce the effect of large peaks in data analysis, especially in data sets that include outlying observations, and make the data more normally distributed (Feng *et al.*, 2014; Xi *et al.*, 2014). However, the disadvantage of Log transformation is that it cannot deal with zero or negative values. Therefore, missing values imputation should be done prior to Log transformation.

The general form Log (x, base) computes logarithms with any desired based. Metabolomics studies often use the Log base 2 and Log base 10 transformation. SIMCA-P 14, however, does not include Log base 2 option and this must be calculated by using another package such as Excel, after which it is transferred to SIMCA-P 14 software for completing the analysis. The Log transformation generally makes a skewed distribution more normal in shape for applying statistical tests with an underlying assumption of normality.

3.2.5.2 Power transformation

Power transformation calculates the square root of each element in the feature and replaces the original data (Eriksson *et al.*, 2013; Tugizimana *et al.*, 2016). Although it does not convert the multiplicative noise into additive noise, its effects are similar to those of Log transformation. This function can be provided directly using SIMCA-P 14 from the transformation option.

3.2.6 Aims

- This chapter aims to describe methods for pre-treatment of metabolomics data obtained after using liquid chromatography mass spectrometry (LC-MS).
- SIMCA-P 14 is used to evaluate the data pre-treatment methods regarding the model parameters and validity.
- The best pre-treatment strategy for LC-MS data is selected.

3.3 Materials and method

3.3.1 Data

Excel data sheet was provided from a previous project (generated by another PhD student in the group). Moreover, the data was generated from analysing 40 football players' urine samples. There were 80 samples that were divided into two groups: pre-training samples and post-training samples. The data matrix consists of 2056 metabolites (i.e., 164480 values), and there were 43576 (approximately 26%) missing values.

3.3.2 LC-MS Analysis

3.3.2.1 Solvents and chemical

Chemicals and solvents are described in section 2.1

3.3.2.2 Samples preparation:

Urine samples were obtained from the freezer and they were thawed out and then mixed thoroughly. Then, 0.2 ml was transferred from each urine sample into an Eppendorf and each sample was mixed with 0.8 ml of acetonitrile in the Eppendorf tube. Next, every Eppendorf tube was centrifuged until two layers of precipitate and supernatant were formed. The supernatant was removed from each Eppendorf tube and transferred to a HPLC vial for analysis using the ZIC-PHILIC metabolomics method.

3.3.2.3 HPLC setup

The HPLC setup and conditions are described in section 2.2.3.

3.3.3 Data pre-processing

Data pre-processing and extraction are described in section 2.3.2.

3.3.4 Data pre-treatment

3.3.4.1 Missing data imputation

During the SIMCA-P multivariate modelling, the data was attributed through the default imputation algorithm Nonlinear Iterative Partial Least Squares algorithm (NIPALS) or were allocated prior to using K-nearest neighbours algorithm (KNN), mean, median, or small value Min/2. Moreover, SIMCA-P 14 software was used for evaluating five missing value imputation methods. Following are the imputation methods:

3.3.4.1.1 Nonlinear Iterative Partial Least Squares algorithm (NIPALS)

SIMCA-P by default deals with missing values using adjusted NIPALS algorithm. The software provides an alert if the missing values exceed 50% for each metabolites (Eriksson *et al.*, 2013; Tugizimana *et al.*, 2016). Essentially, when calculating the principal components or latent variable, this method establishes the missing values' residuals to zero or replaces the missing value with their minimum distance projections on the loading and score vector's present estimate (Pedreschi *et al.*, 2008).

3.3.4.1.2 K-Nearest Neighbours algorithm (KNN)

The K-Nearest Neighbours algorithm (KNN) approach was applied to the data using R software. This algorithm involves using the Euclidean distance method for determining the missing value's nearest neighbours. Then, the missing value

imputation is conducted by replacing the missing value by the average of the identified neighbours (Hrydziuszko and Viant, 2012). In this study, the value of parameter K (number of nearest neighbours) is 10.

3.3.4.1.3 Small value (minimum intensity/2)

The minimum intensity (Min) of each metabolite was identified and divided by 2. Then, the missing values were replaced by the Min/2 value for each metabolite. This calculation and procedure were applied to the data set using Excel.

3.3.4.1.4 Mean

In this imputation method, the missing values were replaced by the average of each metabolite for each metabolite using Excel.

3.3.4.1.5 Median

In this imputation method, the missing values were replaced by the median of each metabolite using Excel.

3.3.4.2 Scaling

In this chapter, as shown in Table **3.1**, three scaling methods were evaluated: mean centering (center), unit variance (UV) and Pareto (Par).

3.3.4.2.1 Mean centering (Center)

The mean centering was applied to the data set by subtracting the intensity value (X) of each metabolite from the average (X – mean). SIMCA-P 14 considered mean centring as one of the scaling methods available.

3.3.4.2.2 Unit variance (UV)

In SIMCA-P 14, UV scaling was conducted after the variables were mean centred. Then, the mean centred values are multiplied by the inverse standard deviation and scaled to 'Unit Variance', such that the base weight is computed as 1/SD, with SD being the standard deviation of the variable.

3.3.4.2.3 Pareto (Par)

In SIMCA-P 14, Par scaling was conducted after variables were mean centred. Then, the mean centred values were multiplied by the square root of the standard deviation and scaled to 'Pareto', such that the base weight is computed as 1/sqrt(SD), with SD being the standard deviation of variable.

3.3.4.3 Transformation

In this study, as shown in Table **3.1**, two transformations were evaluated: Log_{10} and power transformation.

3.3.4.3.1 Log_{10}

In SIMCA-P 14, all values were transformed to a logarithmic scale using the Log function. The Log base 10 was the default in the software.

3.3.4.3.2 Power

The power transformation was applied to all values (X) using power 2 (C3 = 2), as described in (Table **3.1**). SIMCA-P 14 includes power transformation as one of the available transformation methods.

Table 3.1 The pre-treatment methods and equations.

Pre-treatment	Methods	Equation
$Mean\ centering$	$Centering$	$(X - mean)$
$Scaling$	$Unit\ Variance\ (UV) or\ autoscaling$	$\frac{X - mean}{SD}$
	$Pareto$	$(X - mean)/\sqrt{SD})$
$Transformation$	$\log_{10}(X)$	$\log_{10}(C1 * X + C2)\ where\ C1 = 1\ and\ C2 = 0$
	$power$	$(C1 * X + C2)^{C3} where\ C1 = 1, C2 = 0\ and\ C3 = 2$

3.3.5 Software tools

The data set was scaled and transformed using SIMCA-P 14 software. Moreover, all models and graphs were generated using the SIMCA-P 14 software. KNN zero imputation was conducted using the R package. The Min/2, mean, and median zero imputation were executed using Microsoft Excel, after which the data set was transferred to SIMCA-P 14 for conducting the following steps.

3.3.6 Models validation

After the data sets' missing values were imputed, scaled, and transformed, multivariate analysis was conducted. In this process, unsupervised Principle Component model (PCA) and supervised Orthogonal Partial Least Squares-Discriminant model (OPLS-DA) were applied to the data set. At the end of the modelling, the model parameters were evaluated for determining the best pre-treatment method.

The unsupervised model (PCA) was used to visualise samples clustering and separation during the analysis as per the metabolomic changes, regardless of its

grouping. Then, a supervised model's (OPLS-DA) quality was assessed using R^2 (the goodness of fit) and Q^2 (the goodness of prediction), as well as *p CV*-ANOVA (the *p*-value of the model) from seven-fold cross-validation procedures, as given below:

1. R^2X (cum): the cumulated amount of variation in matrix
2. R^2 : The fraction of the original data explained by the model, such that if R^2 = 1.0, the model can explain 100% of the data. Thus, the model's goodness of fit is assessed by this parameter.
3. Q^2 : The fraction of the original data predicted by a cross validated model, such that if Q^2 = 1.0, the model is able to predict 100% of the data. Thus, the model's goodness of prediction can be measured by this parameter.
4. *p* CV-ANOVA: The seven-fold cross-validation procedures form the basis for the model's *p*-value that indicates the models' extent of significance. The p value being > 0.05 suggested a model's insignificance.

First, in terms of the OPLS-DA model's validity criteria, the permutation parameters were arranged according to the higher R^2 value and differences between R^2 and Q^2 below 0.3 (low to high). If R^2-Q^2 > 0.3, then the model had poor robustness and indicated model overfitting. Second, the model's significant differences were assessed by calculating the cross-validation analysis of variance's (CV-ANOVA) *p*-values. A p value of < 0.05 was considered a significant value. Finally, the model's significance was also evaluated using a permutation test. If the Q^2 regression line went beyond the zero line or if no Q^2 value from the permutated data set was more than the Q^2 from the original data set, then the model was valid. The results of implementing various pre-treatment methods for the data set were present as per

the missing data imputation methods. Each imputation involved different scaling and transformation and a table of models parameters was provided for every imputation method. Visually, more dense clusters are preferable; and greater distance between clustered groups is also a preferred outcome.

3.4 Results

3.4.1 Nonlinear Iterative Partial Least Squares algorithm (NIPALS)

After applying different scaling and transformation to NIPALS imputed data set, seven methods were observed as showing significant *p* CV-ANOVA value (<0.05). A further, six methods passed the validity criteria, as shown in (Table **3.2**). In method 44, the OPLS-DA model was insignificant after no transformation or scaling was applied. From method 59 and 36, the application of transformation without scaling didn't fit the data to the model (Model didn't fit). It is evident that UV and Par scaling significantly impacted the data fitting and model parameters (method 4 and 2), as shown in (Figure 3.1). The value of $R^2X(Cum)$ was reduced after applying UV scaling ($R^2X(Cum)=0.37$) compared to after Par scaling method ($R^2X(Cum)=0.51$) was applied. In comparison, centre scaling did not allow fitting of an OPLS-DA model. When Log_{10} and centre scaling (method 8) were applied, the model was invalid as the value of R^2-Q^2 exceed the threshold value (0.3).

Table 3.2 Model parameters after NIPALS algorithm zero imputation.

The methods arranged are based on the highest OPLS-DA R^2 values. Valid model should include R^2-Q^2 < 0.3 and p CV-ANOVA < 0.05 and pass the permutation test.

Method number	Data pre-treatment			Models								
				PCA		OPLS-DA						
	Missing data imputation	Transformation	Scaling	R^2X	Q^2	R^2X (Cum)	R^2	Q^2	*p* CV-ANOVA	R^2-Q^2	Permutation (999 times)	Valid
Method 8	NIPALS	Log_{10}	Centre	0.71	0.46	0.52	0.89	0.57	$2.74E^{-07}$	0.32	Yes	No
Method 3	NIPALS	Log_{10}	Par	0.70	0.48	0.53	0.80	0.63	$1.98E^{-09}$	0.17	Yes	Yes
Method 5	NIPALS	Log_{10}	UV	0.71	0.50	0.51	0.79	0.62	$4.78E^{-09}$	0.17	Yes	Yes
Method 4	NIPALS	-	UV	0.69	0.41	0.37	0.76	0.59	$4.41E^{-09}$	0.17	Yes	Yes
Method 2	NIPALS	-	Par	0.65	0.39	0.51	0.69	0.56	$2.12E^{-08}$	0.13	Yes	Yes
Method 1	NIPALS	Power	UV	0.65	0.05	0.23	0.65	0.45	$1.17E^{-07}$	0.2	Yes	Yes
Method 44	NIPALS	-	-	0.98	0.66	0.95	0.20	0.04	N.S	N.A	No	No
Method 59	NIPALS	Log_{10}	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 56	NIPALS	-	Centre	0.91	0.48	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 35	NIPALS	Power	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 58	NIPALS	Power	Centre	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 60	NIPALS	Power	Par	0.92	0.52	D.F	D.F	D.F	N.A	N.A	N.A	N.A

N.S = insignificant, NA = Not applicable, D.F = model didn't fit

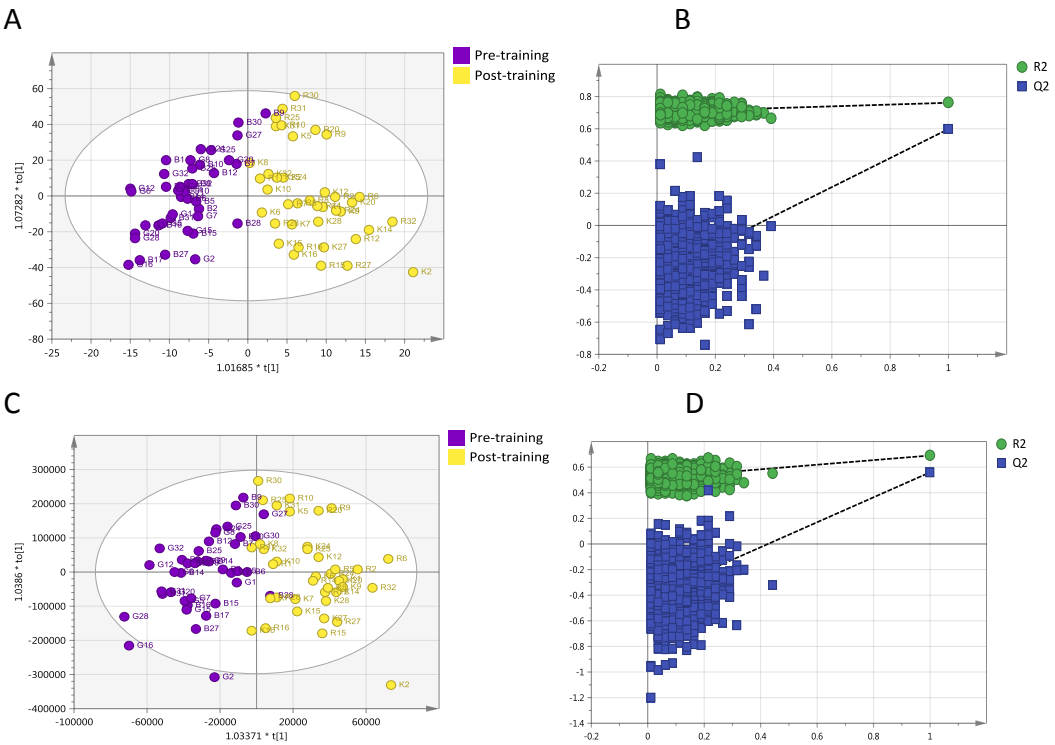


Figure 3.1 Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) score plot for NIPALS zero imputed data set after applying UV and Par scaling.

(A) *The OPLS-DA model based on method 4. The model consists of one predictive x-score component, component t[1], and two orthogonal x-score components, to [1-2]. The t [1] component explains 3% of the predictive variation in x, while the to [1-2] components explain 34% of the orthogonal variation in x, R^2X (cum) = 0.37. The goodness of fit (R^2) = 0.76, the goodness of prediction (Q^2) = 0.59, and the p value are associated with the cross-validation analysis of variance (CV-ANOVA) = $4.41E^{-09}$. (B) The permutation test for UV scaled data. (C) The OPLS-DA model is based on method 2. The model includes one predictive x-score component, component t[1], and two orthogonal x-score components, to [1-2]. The t [1] component explains 4 % of the predictive variation in x, while the to [1-2] components explain 47% of the orthogonal variation in x, R^2X (cum) = 0.51. The goodness of fit (R^2) = 0.69, the goodness of prediction (Q^2) = 0.56, and the p value are associated with the cross-validation analysis of variance (CV-ANOVA) = $2.12E^{-08}$. (D) The permutation test for Par scaled data.*

In NIPALS, the imputed data it appears to be the combination of Log_{10} as transformation and Par (method 3) and UV (method 5) as scaling methods give better OPLS-DA parameters with a small advantage of Par compared to UV scaling. As shown in (Figure 3.2 C), within group variation after Par scaling is lower than within group variation after UV scaling. Further, there is a slight improvement in goodness of fit

(R^2) and goodness of prediction (Q^2) after Par scaling compared to after UV scaling. Regarding the significant model's *p* CV-ANOVA, NIPALS imputation were more significant than for the other imputation methods.

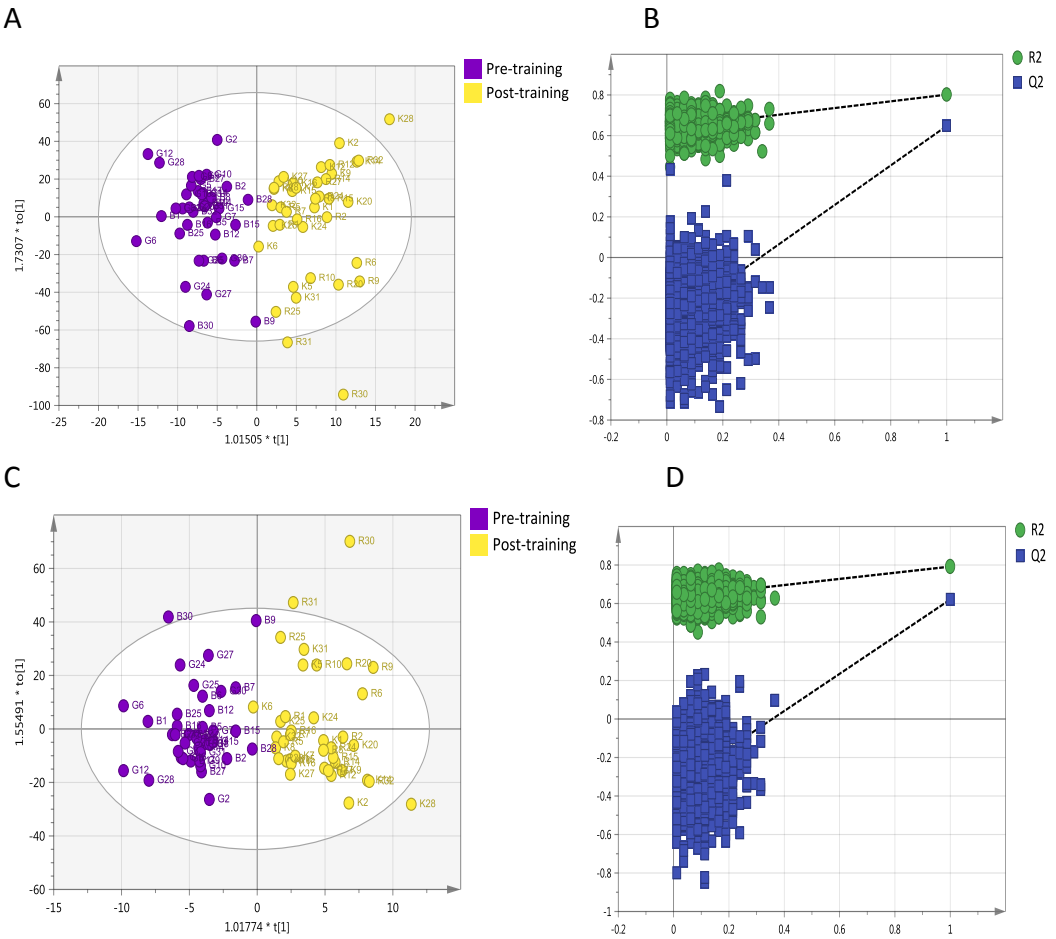


Figure 3.2 Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) score plot for NIPALS zero imputed data set after applying Log_{10} transformation to UV and Par scaled data.

(A) The OPLS-DA model based on method 5. The model includes one predictive x-score component, component t[1], and two orthogonal x-score components, to [1-2]. The t [1] component explains 3% of the predictive variation in x, while the to [1-2] components explain 48% of the orthogonal variation in x, R^2X (cum) = 0.51. The goodness of fit (R^2) = 0.79, the goodness of prediction (Q^2) = 0.62, and the p value associated with the cross-validation analysis of variance (CV-ANOVA) = $4.78E^{-09}$. (B) The permutation test for Log_{10} transformed data with UV scaling. (C) The OPLS-DA model based on method 5. The model includes one predictive x-score component, component t [1], and two orthogonal x-score components, to [1-2]. The t [1] component explains 3 % of the predictive variation in x, while the to [1-2] components explain 50% of the orthogonal variation in x, R^2X (cum) = 0.53. The goodness of fit (R^2) = 0.80, the goodness of prediction (Q^2) = 0.63, and the p value associated with the cross-validation analysis of variance (CV-ANOVA) = $1.98E^{-09}$. (D) The permutation test for Log_{10} transformed data with Par scaling.

3.4.2 K-Nearest Neighbours algorithm (KNN)

No valid method could be produced following KNN imputation of missing variables. Validity criteria after various pre-treatment methods were applied, as shown in (Table **3.3**). The model parameters clearly indicate that KNN can improve the goodness of fit $R^2 > 0.9$. The goodness of prediction, however, had lower value at $Q^2 < 0.5$ for all significant models. Regarding scaling methodology, Par scaling showed lower R^2-Q^2 value (method 10) than UV scaling (method 15). On the other hand, visually, UV scaling has better group separation as it has high R^2 value (goodness of fit) compared to Par scaling $R^2 = 0.72$, as shown in (Figure *3.3*).

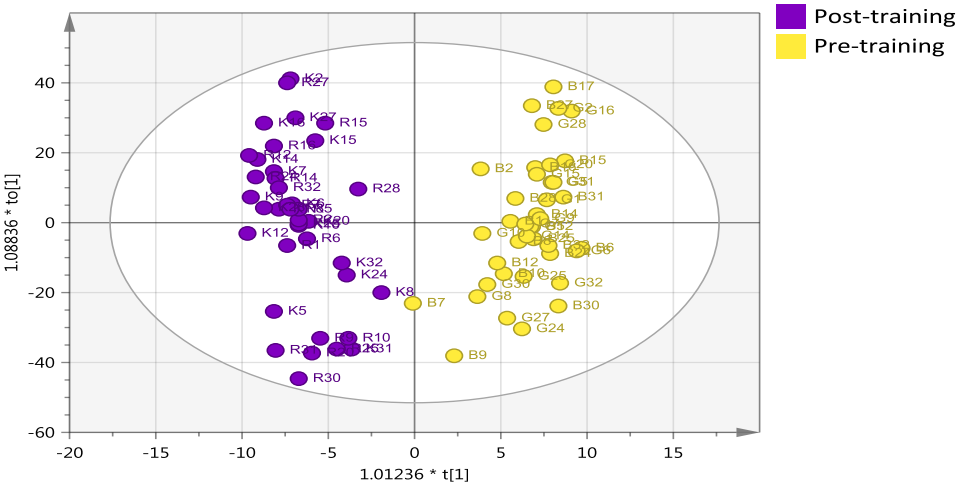
Table 3.3 Model parameters after KNN algorithm zero imputation.

The methods arranged based on the highest OPLS-DA R^2 values. Valid model should have R^2-Q^2 < 0.3 and p CV-ANOVA < 0.05 and pass the permutation test.

Method number	Data pre-treatment			Models								
				PCA		OPLS-DA						
	Missing data imputation	Transformation	Scaling	R^2X	Q^2	R^2X (Cum)	R^2	Q^2	*p* CV-ANOVA	R^2-Q^2	Permutation (999 times)	Valid
Method 21	KNN	Log_{10}	UV	0.56	0.28	0.44	0.97	0.42	$2.80E^{-05}$	0.55	Yes	No
Method 24	KNN	Log_{10}	Par	0.57	0.28	0.42	0.96	0.37	$2.80E^{-04}$	0.59	Yes	No
Method 15	KNN	-	UV	0.54	0.19	0.30	0.92	0.44	$1.32E^{-06}$	0.48	Yes	No
Method 27	KNN	Log_{10}	Center	0.58	0.25	0.37	0.92	0.28	$2.04E^{-03}$	0.64	Yes	No
Method 17	KNN	Power	UV	0.53	0.02	0.24	0.91	0.38	$2.62E^{-05}$	0.53	Yes	No
Method 10	KNN	-	Par	0.65	0.25	0.40	0.72	0.35	$1.97E^{-05}$	0.37	Yes	No
Method 31	KNN	-	-	0.98	0.58	0.98	0.54	0.23	N.S	N.A	N.A	No
Method 38	KNN	-	Center	0.89	0.33	0.85	0.51	0.17	N.S	N.A	N.A	No
Method 33	KNN	Power	Par	0.65	0.25	0.87	0.48	0.10	N.S	N.A	N.A	No
Method 41	KNN	Log_{10}	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 43	KNN	Power	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 45	KNN	Power	Center	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A

N.S = insignificant, NA = Not applicable, D.F = model didn't fit

A



B

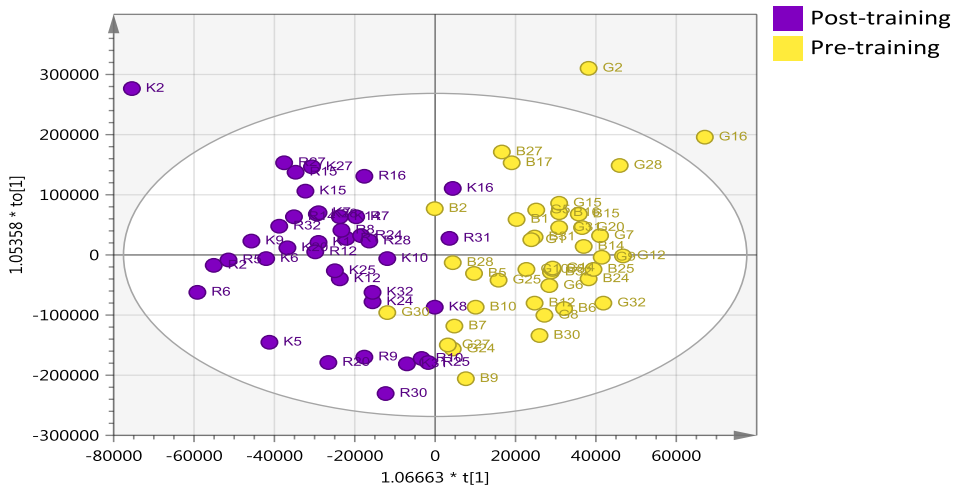


Figure 3.3 Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) score plot for KNN zero imputed data set after applying UV and Par scaling.

(A) The OPLS-DA model based on method 15. The model includes one predictive x-score component, component t [1], and three orthogonal x-score components, to [1-3]. The t [1] component explains 2 % of the predictive variation in x, while the to [1-3] components explain 28% of the orthogonal variation in x, R^2X (cum) = 0.30. The goodness of fit (R^2) = 0.92, the goodness of prediction (Q^2) = 0.44, and the p value associated with the cross-validation analysis of variance (CV-ANOVA) = $1.32E^{-06}$. (B) The OPLS-DA model based on method 10. The model includes one predictive x-score component, component t [1], and two orthogonal x-score components, to [1-2]. The t [1] component explains 2.5 % of the predictive variation in x, while the to [1-2] components explain 37.5% of the orthogonal variation in x, R^2X (cum) = 0.40. The goodness of fit (R^2) = 0.72, the goodness of prediction (Q^2) = 0.35, and the p value associated with the cross-validation analysis of variance (CV-ANOVA) = $1.97E^{-05}$.

3.4.3 Small value (minimum intensity/2)

Similar to KNN method, no valid methods were observed that passed the validity criteria for Min/2 imputed data after the various pre-treatment methods were applied, as shown in (Table **3.4**). Moreover, only three methods which are UV (method 23), Power/UV (method 14), and Par (method 19) had significant CV-ANOVA *p*-value. These three methods are presented in (Figure *3.4*). This figure shows that the UV scaling improved the model separation compared to Par scaling. In terms to model parameters, the majority of the models presented here have low R^2X value.

Table 3.4 Model parameters after Min/2 zero imputation.

The methods arranged based on the highest OPLS-DA R^2 values. Valid model should have R^2-Q^2 < 0.3 and p CV-ANOVA < 0.05 and pass the permutation test.

Method number	Data pre-treatment			Models								
				PCA		OPLS-DA						
	Missing data imputation	Transformation	Scaling	R^2X	Q^2	R^2X (Cum)	R^2	Q^2	*p* CV-ANOVA	R^2-Q^2	Permutation (999 times)	Valid
Method 32	Min/2	Log_{10}	Par	0.14	0.05	0.14	1.00	0.00	N.S	N.A	N.A	No
Method 39	Min/2	Log_{10}	Center	0.15	0.06	0.12	0.99	-0.02	N.S	N.A	N.A	No
Method 37	Min/2	Log_{10}	UV	0.14	0.04	0.12	0.98	0.00	N.S	N.A	N.A	No
Method 23	Min/2	-	UV	0.38	0.09	0.23	0.94	0.38	$2.78E^{-05}$	0.56	Yes	No
Method 14	Min/2	Power	UV	0.49	0.05	0.20	0.82	0.35	$1.97E^{-05}$	0.47	Yes	No
Method 19	Min/2	-	Par	0.50	0.12	0.32	0.77	0.25	$1.79E^{-03}$	0.53	Yes	No
Method 42	Min/2	-	-	0.99	0.56	0.93	0.25	0.04	N.S	N.A	N.A	No
Method 40	Min/2	Power	Par	0.82	0.27	0.74	0.02	-0.02	N.S	N.A	N.A	No
Method 28	Min/2	Log_{10}	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 36	Min/2	-	Center	0.84	0.25	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 30	Min/2	Power	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 57	Min/2	Power	Center	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A

N.S = insignificant, NA = Not applicable, D.F = model didn't fit

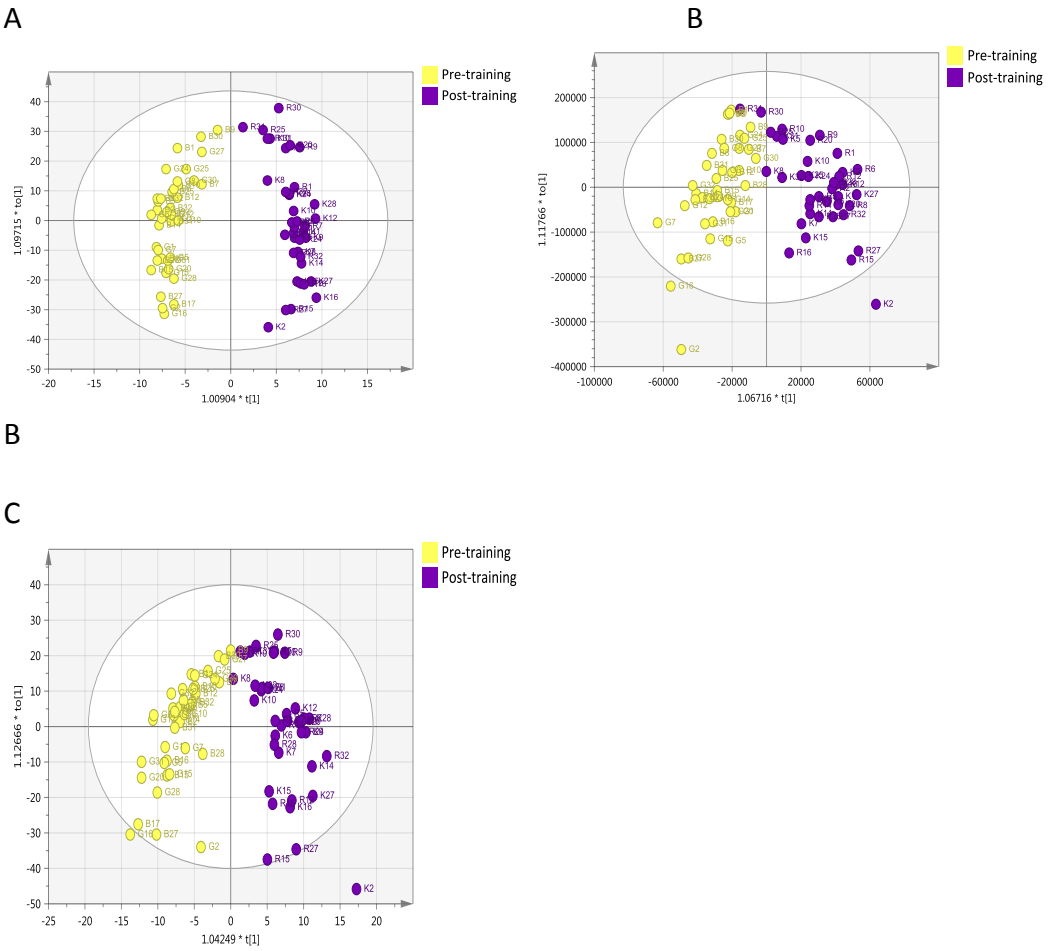


Figure 3.4 Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) score plot for Min/2 zero imputed data set. The significant models were presented after applying different pre-treatment methods.

(A) The OPLS-DA model based on method 23. The model includes one predictive x-score component, component t[1], and three orthogonal x-score components, to [1-3]. The t [1] component explains 2.3 % of the predictive variation in x, while the to [1-3] components explain 20.7% of the orthogonal variation in x, R^2X (cum) = 0.23. The goodness of fit (R^2) = 0.94, the goodness of prediction (Q^2) = 0.38, and the p value associated with the cross-validation analysis of variance (CV-ANOVA) = $2.78E^{-05}$. (B) The OPLS-DA model based on method 19. The model includes one predictive x-score component, component t[1], and two orthogonal x-score components, to [1-2]. The t [1] component explains 2.6 % of the predictive variation in x, while the to [1-2] components explain 29.4% of the orthogonal variation in x, R^2X (cum) = 0.32. The goodness of fit (R^2) = 0.77, the goodness of prediction (Q^2) = 0.25, and the p value associated with the cross-validation analysis of variance (CV-ANOVA) = $1.79E^{-03}$. (C) The OPLS-DA model based on method 14. The model includes one predictive x-score component, component t [1], and three orthogonal x-score components, to [1-3]. The t [1] component explains 3% of the predictive variation in x, while the to [1-2] components explain 17% of the orthogonal variation in x, R^2X (cum) = 0.20. The goodness of fit (R^2) = 0.82, the goodness of prediction (Q^2) = 0.35, and the p value associated with the cross-validation analysis of variance (CV-ANOVA) = $1.97E^{-05}$.

3.4.4 Mean and median

It was observed that mean and median zero imputation had similar model parameters, as shown in (Table **3.5**) and (Table **3.6**). Both imputation methodologies involved using Log_{10} transformation and a scaling method for improving the parameters. Similar to previous imputation methods, transformation without any scaling method was not applicable for the model.

Table 3.5 Model parameters after mean imputation.

The methods arranged based on the highest OPLS-DA R^2 values. Valid model should have R^2-Q^2 < 0.3 and p CV-ANOVA < 0.05 and pass the permutation test.

Method number	Data pre-treatment			Models								
				PCA		OPLS-DA						
	Missing data imputation	Transformation	Scaling	R^2X	Q^2	R^2X (Cum)	R^2	Q^2	*p* CV-ANOVA	R^2-Q^2	Permutation (999 times)	Valid
Method 16	Mean	Log_{10}	UV	0.57	0.29	0.46	0.97	0.46	$4.02E^{-06}$	0.51	Yes	No
Method 20	Mean	Log_{10}	Par	0.56	0.28	0.43	0.97	0.42	$2.26E^{-05}$	0.55	Yes	No
Method 25	Mean	Log_{10}	Center	0.58	0.24	0.40	0.97	0.37	$2.54E^{-04}$	0.60	Yes	No
Method 11	Mean	-	Par	0.67	0.26	0.45	0.84	0.46	$3.50E^{-07}$	0.38	Yes	No
Method 13	Mean	-	UV	0.55	0.20	0.28	0.83	0.41	$6.50E^{-07}$	0.41	Yes	No
Method 6	Mean	Power	UV	0.39	0.06	0.18	0.65	0.31	$2.73E^{-06}$	0.34	Yes	No
Method 48	Mean	Log_{10}	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 50	Mean	-	-	0.98	0.62	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 46	Mean	-	Center	0.89	0.38	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 47	Mean	Power	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 49	Mean	Power	Center	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 51	Mean	Power	Par	0.90	0.36	D.F	D.F	D.F	N.A	N.A	N.A	N.A

N.S = insignificant, NA = Not applicable, D.F = model didn’t fi

Table 3.6 Model parameters after median imputation.

The methods arranged based on the highest OPLS-DA R^2 values. Valid model should have R^2-Q^2 < 0.3 and p CV-ANOVA < 0.05 and pass the permutation test.

Method number	Data pre-treatment			Models								
				PCA		OPLS-DA						
	Missing data imputation	Transformation	Scaling	R^2X	Q^2	R^2X (Cum)	R^2	Q^2	*p* CV-ANOVA	R^2-Q^2	Permutation (999 times)	Valid
Method 26	Median	Log_{10}	Center	0.58	0.30	0.44	0.99	0.38	$6.95E^{-04}$	0.60	Yes	No
Method 22	Median	Log_{10}	Par	0.59	0.30	0.45	0.97	0.41	$4.66E^{-05}$	0.56	Yes	No
Method 18	Median	Log_{10}	UV	0.59	0.33	0.46	0.96	0.44	$9.15E^{-06}$	0.53	Yes	No
Method 9	Median	-	Par	0.67	0.24	0.45	0.84	0.47	$2.28E^{-07}$	0.37	Yes	No
Method 12	Median	-	UV	0.55	0.19	0.28	0.82	0.41	$6.37E^{-07}$	0.41	Yes	No
Method 7	Median	Power	UV	0.39	0.05	0.18	0.64	0.34	$2.98E^{-06}$	0.30	Yes	No
Method 29	Median	Log_{10}	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 53	Median	-	-	0.98	0.61	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 55	Median	-	Center	0.89	0.38	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 52	Median	Power	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 34	Median	Power	Center	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 54	Median	Power	Par	0.90	0.36	D.F	D.F	D.F	N.A	N.A	N.A	N.A

N.S = insignificant, NA = Not applicable, D.F = model didn't fit

3.5 Discussion

The aim of chemometric analysis of high dimensional metabolomics data is to generate as parsimonious a model as possible (Ohno, Karagiannis and Taniguchi, 2014). There are different opinions and choices for selecting the suitable methods as the nature of the data differs in different studies and methods. The aim is to select the best pre-treatment method for decreasing statistical and modelling errors, and thus, the biomarkers can present the real biological variation between samples. Multivariate analysis typically requires pre-treatment over the univariate analysis. In addition, clinical/human samples may have a high level of variation, thus resulting in the necessity for data pre-treatment. Pre-treatment is important in the present study as the data is generated from human clinical studies (Xiao *et al.*, 2014).

3.5.1 Missing data imputation

The most important step after data extraction and identification is in handling the missing data that tends to be abnormally distributed over the data set (Sumner *et al.*, 2007). In the metabolomics work flow, a metabolite with 50% or more missing values can exclude from the analysis, and SIMCA-P software can do this in the beginning of the analysis. When the number of missing values in an observation or a variable exceeds the specified threshold, SIMCA displays a warning message and then these observations or variable can be deleted from the analysis. Threshold of missing values for observations in percent. The threshold applies to both the work set and the prediction set. In fact, no specific percentage has been agreed upon for the missing value per metabolite (Goodacre *et al.*, 2007). The present study did not exclude

metabolites that exceeded the 50% criteria as the effect of all the data set's missing data on the model parameters was to be assessed. The missing data in metabolomic data can be filled in a different way. One way is with the manual method of inspecting the spectrum in the raw file for each metabolite and determining the intensity concerning the metabolite in each sample. However, this is time-consuming and may be inaccurate and can only be used for specific metabolites (Huan and Li, 2015). On the other hand, using a computerized algorithm such as KNN or NIPALS can facilitate the imputation with the standard method for all missing value. In NIPALS algorithm imputation, the majority of the studies based on SIMCA-P software used this approach unless another imputation method was mentioned.

Compared to the various imputation methods used in this study, NIPALS algorithm was significantly better according to the model parameter evaluation along with suitable transformation and scaling (Appendix, Table S.1. 1). It is not possible for all imputation methods to model the data without transformation and/or scaling. This indicates that pre-treatment significantly impacts the data set. Moreover, the logarithm available by default in SIMCA-P (NIPALS) was the only method compared to all imputation methods that could generate valid models (Table **3.2**). This method is widely used in chemometrics missing data imputation (Grung and Manne, 1998). This may be because NIPALS was initially developed and used to extract the principle component (Nelson, Taylor and MacGregor, 1996).

The second imputation method was the KNN algorithm. The advantage that KNN offers is that it is able to deal with large variables containing missing values

(Troyanskaya *et al.*, 2001). Compared to the various imputation methods, it has been argued that KNN is better than the mean, median and zero imputation method (Gromski *et al.*, 2014), and the random forest (RF) imputation method is the only method that is more advantageous than KNN in model variation presentation. This result provided after UV scaling methodology was applied to the same data set and analysis was conducted using R version 2.15.0 and 3.1.0. Univariate as well as multivariate analysis of metabolomic study described KNN imputation as the best method for zero imputation and normalization compared to eight other imputation methods (Hrydziuszko and Viant, 2012). The present study used the R computing environment and noted that KNN imputation was better than mean, median and Min/2 imputation approaches. Further, none of the previous studies compare NIPALS with KNN imputation. The goodness of fit (R^2) was reported as high (≈1) in KNN imputation using capillary electrophoresis-mass spectrometry (CE-MS) data (Armitage *et al.*, 2015). In addition, the present study concluded that KNN offers more advantages compared to median and min/2 imputation methods. Overall, the result shows that the model's goodness of fit was improved along with using transformation and scaling (Par and UV).

3.5.2 Scaling and transformation

Scaling had a clear impact on the modelled data and indicated the extent to which this data pre-treatment process can change the data fitting and prediction. The most important and beneficial role of scaling is of making the variable more comparable in size (Di Guida *et al.*, 2016). UV scaling is described as the most reliable scaling method (van den Berg *et al.*, 2006; Goodacre *et al.*, 2007). However, considering the result

presented in (Table **3.2**), Par scaling is better able to explain the data than Unit variance (UV) with R^2X (Cum)=0.51, thus suggesting that Par scaling can improve the model's ability to explain the data more than UV scaling. This result is similar to that observed by (Tugizimana *et al.*, 2016) who modelled different data sets using different pre-treatment methodologies using SIMCA-P software and noted that using Par scaling without transformation produced a srtonger model than UV. They also observed that using Par or UV along with Log transformation in OPLS-DA modelling had a similar effect according to the provided diagnostic parameters. This was similar to the present study's results with noting slightly more advantage of Par scaling along with Log transformation.

Data transformation is always applied to high dimensional data for correcting its heteroscedasticity and skewness; that is, making the data more normally distributed (Kvalheim, Brakstad and Llang, 1994). The only drawback of transformation is that Log and power transformation can result in a pseudo-scaling effect on the data as it decreases large values in the data set relatively more compared to small values (van den Berg *et al.*, 2006). Therefore, to reduce the impact of the pseudo-scaling effect, the scaling method should be applied after data transformation. Further, Log transformation is unable to deal with zero values, thus indicating the importance of missing value imputation process in LC-MS data.

The result clearly shows that Log transformation was always at the top of each table based on R^2 (goodness of fit) considering all missing data. Therefore, Log transformation offers advantages compared to power transformation. The method

that used Log transformation and Par scaling (method 3) had a valid R^2-Q^2 value (R^2-Q^2=0.17) after using NIPALS imputation method, as shown in (Table **3.2**). This method also had higher R^2X (cum) compared to UV scaling method. Moreover, van Berg et al. (2006) noted that there were no differences between Log and power transformation. They replaced zero values with a small value of one to allow Log transformation to work and prevent an infinity value when the transformed value was equal to zero (Yi *et al.*, 2016).

Chapter 4:

Untargeted Metabolomics of Paediatric Crohn's Disease Patients Against Healthy Controls

4 Untargeted Metabolomics of Paediatric Crohn's Disease Patients Against Healthy Controls

4.1 Abstract

Background and aim: Crohn's Disease (CD) is a component of Inflammatory Bowel Disease (IBD), a multifactorial disorder likely resulting from altered immune responses to commensal or pathogenic gut microbes under the influence of an environmental factor, including diet. Exclusive Enteral Nutrition (EEN) is the most common treatment for paediatric CD in the UK and the rest of Europe. Metabolomics is an indispensable research tool for the identification and tracking of biomarkers in biological systems and fluids. In the current study, metabolomic profiling based on LC-MS data was used to identify significantly differentiated metabolites in the faecal samples of children with CD before, during and after EEN treatment

Methodology: Metabolomic profiling using high resolution mass spectrometry with hydrophilic interaction chromatography was applied to the analysis of 11 faecal extracts from eleven healthy control (HC) children and to 43 faecal extracts from eleven children undergoing exclusive enteral nutrition for the treatment of active Crohn's Disease (CD) at timepoints before, during (15, 30, and 60 days), and after treatment. Differences between the control and CD samples were identified at each timepoint.

Results: An orthogonal partial least square-discriminant analysis (OPLS-DA) model identified eight metabolites that were normally distributed according to Q-Q plots. The OPLS-DA model was able to discriminate the CD samples from the controls at

every timepoint, but the model was not able to differentiate the CD samples from one another at the different timepoints during treatment with exclusive enteral nutrition. The differentiated metabolites identified in the CD samples included tyrosine, an ornithine isomer, arachidonic acid, eicosatrienoic acid, docosatetraenoic acid, a sphingomyelin, a ceramide, and dimethylsphinganine.

Conclusion: Despite successful treatment, underlying differences remained in the metabolome of the CD patients. These differences dominated the separation of the samples when multivariate methods were applied.

4.2 Introduction

Crohn's Disease (CD) is a component of Inflammatory Bowel Disease (IBD), a multifactorial disorder likely resulting from altered immune responses to commensal or pathogenic gut microbes under the influence of an environmental factor, including diet (Jansson *et al.*, 2009). Children and adolescents represent 15 to 20% of all CD cases, in whom the disease presents more extensively and severely (Day and Lopez, 2015). The disease has distinct stages: onset, severity, progression, remission, and relapse. A dysbiotic gut microbiota is thought to play a role in the disease pathogenesis. Correlations between CD and diet are believed to be equally important, but the specific molecular interactions remain unclear. Therefore, knowledge of a defined metabolomic fingerprint in CD could be useful for diagnosis, treatment, detection of disease pathogenesis, and prediction of disease progression.

Exclusive Enteral Nutrition (EEN) is the most common treatment for paediatric CD in the UK and the rest of Europe (Navas López *et al.*, 2014). EEN is a liquid-only diet comprised of a proprietary nutritional feed that is administered to CD patients for up to eight weeks. EEN induces clinical remission in approximately 80% of cases (Cameron *et al.*, 2013) and results in mucosal healing more often than treatment with high doses of oral steroids (Borrelli *et al.*, 2006). Two mechanisms have been suggested for the effectiveness of EEN treatment. The first relates to changes in the gut microbiota composition and metabolism (Gerasimidis *et al.*, 2014; Quince *et al.*, 2015). The second involves exclusion of dietary triggers of the disease, such as food

emulsifiers and preservatives (Gatti *et al.*, 2017). However, the exact mechanism of EEN treatment has not been fully elucidated and requires further investigation.

Metabolomics is an indispensable research tool for the identification and tracking of biomarkers in biological systems and fluids. This holistic approach provides the broadest array of functional information in systems biology (Leiss *et al.*, 2011). An unbiased, data-driven method, metabolomics presents a novel means of interrogating biological systems that could lead to new hypotheses and biological knowledge. In a typical metabolomics study, complex extracts or body fluids are analysed and compared by various methods to generate metabolic fingerprints (Zhou *et al.*, 2012). The primary metabolomic techniques are either based on nuclear magnetic resonance (NMR) (Leiss *et al.*, 2011) or mass spectrometry (MS) (Zhou *et al.*, 2012). When MS is applied, it is often used in combination with gas chromatography (GC-MS) or liquid chromatography (LC-MS). Due to the wide structural and chemical diversity of metabolites, a single analytical method may not provide a complete index of all the metabolites present in an organism at the time the sample was obtained (Dunn, 2008). Consequently, a combination of methods is preferred for metabolomic studies. The recorded dataset is processed and compared to a range of metabolic fingerprints using multivariate data analysis (MVDA). This analysis can reveal features in the dataset that could be linked to biomarkers for differential diagnosis and monitoring of treatment (Boccard, Veuthey and Rudaz, 2010). There have been a number of previous studies which have applied metabolomics profiling in IBD without any firm agreement with regard to the biomarkers indicative of the disease (Jansson *et al.*, 2009; Schicho *et al.*, 2012;

Bjerrum *et al.*, 2015; De Preter *et al.*, 2015; Kolho *et al.*, 2016). Few studies have applied LC-MS to the analysis of faecal extracts and the majority of studies have used NMR or GC-MS for the analysis (De Preter, 2015; Karu *et al.*, 2018). There are also no studies in children with CD during treatment with EEN in comparison with healthy controls (HC). Comparing differences between HCs and CD patients over the course of treatment offers the opportunity to unravel factors implicated in disease pathogenesis and the mechanism of EEN action.

4.3 Aim of the study

In the current study, metabolomic profiling based on LC-MS data was used to identify significantly differentiated metabolites in the faecal samples of children with CD before, during and after EEN treatment. The relative abundances of these identified metabolites were examined and compared to the metabolomic profiles of HCs.

4.4 Materials and methods

4.4.1 Solvents and chemical

Chemical and solvents were previously described previously in section 2.1. The quantification of fatty acids was performed using commercial standards: arachidonic acid, (CAS number 506-32-1, Sigma-Aldrich, Poole, UK) and Cis-8, 11, 14-Eicosatrienoic acid (CAS number 1783-84-2, Sigma-Aldrich, Poole, UK). All other standards were obtained from Sigma Aldrich, Poole, UK.

4.4.2 Samples and Sample Preparation

Serial faecal samples were collected during exclusive enteral nutrition (n = 54) from 11 CD children (4 females, age mean (SD): 11.5 (2.4)) (Table **4.2**). A single spot sample

was collected for comparative purposes from 11 age and gender matched HCs (4 females, age mean (SD): 10.2 (2.3)) with no familiar history of IBD. From the 11 children with CD, 7 were newly diagnosed, treatment naïve and four received a repeat course of EEN (all within a year of diagnosis) due to disease relapse. All patients completed a 7–8 weeks course of exclusive enteral nutrition using Modulen IBD (Nestle, Vevey, Switzerland). Four patients (2 newly diagnosed and 2 patients on relapse) were on concomitant treatment with azathioprine and three on 5-aminosalicylates. No patient had received antibiotics within 3 months prior to recruitment. At treatment initiation, the mean (SEM) BMI z-score was -1.61 (0.27) (BMI 13.8 ± 1.4) with 7 out of 11 (64%) patients classified as undernourished (BMI < 2nd centile). Following 4- and 8-week treatment on EEN, the baseline BMI z-score significantly (both $p < 0.001$) increased by 1.6 (0.38) (BMI 15.7 ± 1.3) and 1.7 (0.35) SD (BMI 16.2 ± 1.5) respectively (this data is summarised in Table **4.1**). Seven patients had a BMI z-score below the 2nd centile at treatment initiation, all patients had active disease (Paediatric Disease Activity Index (PCDAI) > 10 units). At treatment completion, 7 patients entered in clinical remission (PCDAI < 10 units); 3 others had a significant improvement in clinical disease activity but did not enter clinical remission (PCDAI > 10 units) and one patient did not respond to treatment and oral steroid was initiated following EEN cessation at 8 weeks.

Table 4.1 Subject data for HCs and patients.

na = not recorded, nr = not relevant. PCDAI = Paediatric Disease Activity Index

Subjects	CD Patients	Healthy Controls
Sex	4 F 7 M	4 F 7 M
Age	11.5 ± 2.4	10.2 ± 2.3
BMI at Enrolment (kg/m2)	13.8 ± 1.4	na
Weight (kg) at Enrolment (kg/m2)	28.9 ± 6.0	na
BMI Z Score at Enrolment (kg/m2)	−1.61 ± 0.27	na
BMI (kg/m2) at 4 Weeks	15.7 ± 1.3	nr
Weight (kg) at 4 Weeks	30.8 ± 6.3	nr
BMI Z Score Increase 4 Weeks	1.6 ± 0.38	nr
BMI (kg/m2) at 8 Weeks	16.2 ± 1.5	nr
Weight (kg) at 8 Weeks	33.3 ± 5.2	nr
BMI Z Score Increase 8 Weeks	1.7 ± 0.35	nr
Treatment Naïve	7	nr
Previously Treated	4	nr
PCDAI at Start	11 > 10	nr
PCDAI at End	7 < 10	nr

Table 4.2 Samples numbers and groups of paediatric CD before, during, and after EEN and HCs.

Group ID	Description	n
PA	CD children pre-EEN treatment	11
PB	CD children 15 days of EEN treatment	10
PC	CD children 30 days of EEN treatment	11
PD	CD children 60 days of EEN treatment	11
PE	CD children back to normal diet	11
HC	Healthy children control	11

From the children with CD, samples were collected starting either before EEN initiation or the first sample passed after EEN initiation to a maximum of five days after EEN initiation (PA). Follow up samples were collected during treatment at 15 days after EEN initiation (PB), 30 days after EEN initiation (PC), and 60 days after EEN initiation (PD). A final sample (PE) was collected two to four months post treatment

after the patients had resumed their free diet. Faecal calprotectin (FC, mg/kg) was raised in all patients prior to EEN initiation (median, IQR: 2262, 2089:2582) and significantly decreased after 30 [FC change (SEM) from treatment initiation at 15 days: −483 (211), p = 0.123; at 30 days: −679 (204), p = 0.012; at 60 days: −1002 (211), p < 0.001]. 4 out of the 11 patients had a FC below 150 mg/kg at the end of EEN. FC concentration returned to pre-treatment levels within 2–4 months of food reintroduction (median, IQR, min-max: 2248, 1969–2431, 1632–2495).

All samples were freeze dried then extracted immediately with chloroform/methanol/water (1:3:1 v/v). The extracts were stored at −80 ◦C until analysis by LC-MS. Calprotectin values were determined as described previously (Gerasimidis *et al.*, 2011) and are shown in Table S3. Samples were randomized to avoid inter-batch differences. Pooled samples (n = 5) were prepared from a combination of all samples and intermittently injected throughout the sequence. The samples were randomised and analysed in batches of 13 faecal extracts with one pooled sample in between batches in LC-MS analysis. This study received ethics approval by the Yorkhill Research Ethics Committee (05/S0708/66). Both carers and patients provided written consent.

4.4.3 LC-MS analysis

4.4.3.1 Mobile phase solutions for ZIC-pHILIC chromatography

The mobile phases for ZIC-pHILIC analysis and its preparations were describe previously in section 2.2.1

4.4.3.2 Mobile Phase for ACE C4 Chromatography

The mobile phases for ACE C4 analysis and its preparations were described previously in section 2.2.2

4.4.3.3 HPLC setup

The HPLC setup and conditions were described previously in section 2.2.3

4.4.3.4 Orbitrap Exactive MS setup

The experiment conditions and procedures were described previously in section 2.2.4

4.4.4 LC-MS/MS analysis

Additional experiments were carried out on an Orbitrap Fusion connected with a ZICpHILIC column using the conditions described above in section 2.2.1 . The nitrogen sheath and auxiliary gas flow rates were maintained at 40 and 5 arbitrary units. ESI interface was operated at a positive mode at 4.3 kV, the ion transfer capillary temperature was 325 ◦C. MS^2 and MS^3 spectra were obtained using a collision energy of 30 V. For data dependent MSn experiments the inclusion list consisted of the ions at m/z 133.097, 328.32, 564.53, and 813.68.

4.4.5 Data analysis

The data analysis details and methods were reported above in section 2.3

4.5 Results

4.5.1 Pooled samples

The initial screening detected 606 putatively identified metabolites. The pooled samples (n = 5) were clustered, indicating that no technical errors occurred during the analysis (Figure 4.1). Metabolites were identified to Metabolomics Standard Initiative (MSI) levels 1 or 2, and matching was carried out against authentic standards (Sumner *et al.*, 2007) where available. The details of our standard mixtures were provided in a previous publication (Howe *et al.*, 2018). To quantify the precision of the determinations, the relative standard deviation (RSD) was calculated between the five pooled samples based on the total intensities in each sample, resulting in an RSD of 14%. Using the percentage RSD criteria, metabolites with an RSD > 30% were excluded, accounting for 230 compounds. The remaining 376 metabolites were retained in the study, and the analysis continued as described above in section 2.3.

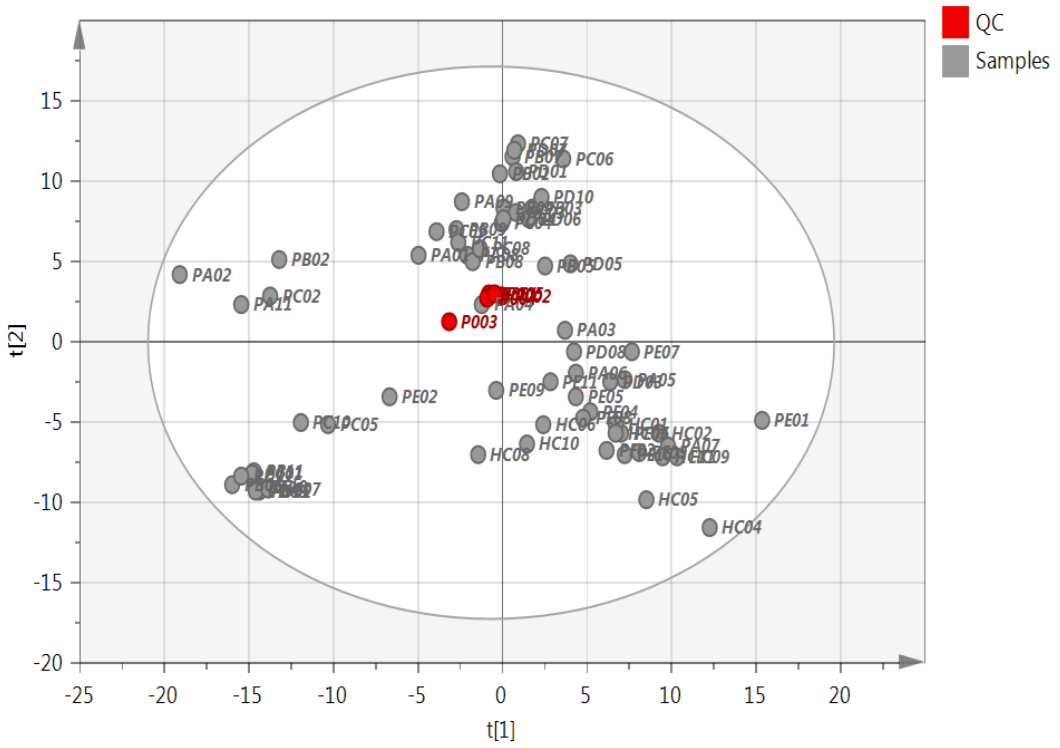


Figure 4.1 2D Scores plot of the principal components analysis (PCA) for the quality control (QC) samples (red) and all samples (grey) based on 606 putative metabolites

4.5.2 Data Visualisation

As shown in Appendix (Figure S4. 1), Log_2 transformation improved data clustering and separation. The samples from the children in the HC (HC) group were clearly separated from the CD patient groups. During EEN treatment, the serially collected samples PB, PC, and PD clustered together on the left half of the ellipse. After the CD patients completed EEN treatment and returned to their free habitual diet (samples PE), the samples appeared between the pre-treatment and HC groups. There was a clear separation between the pre-treatment (PA) and the HC groups.

An orthogonal partial least square-discriminant analysis (OPLS-DA) model was constructed, and the validation process was carried and the data are shown in (Table **4.3**) for models based on 376 metabolites. The only PA group valid models were PA vs. HC and PA vs. PC. Both models produced a goodness of prediction (Q^2) > 0.5, and the differences between the goodness of fit (R^2) and Q^2 were less than 0.3. However, as shown in (Table **4.3**) the HC group in comparison with all groups produced valid and significant models.

Table 4.3 An overview of all the orthogonal partial least square-discriminant analysis (OPLS-DA) parameters and their validity.

The p CV-ANOVA column denotes the p value associated with the cross-validation analysis of variance (CV-ANOVA). (HC) HC children, (PA) CD children pre-EEN treatment, (PB) CD children 15 days post-EEN treatment, (PC) CD children 30 days post-EEN treatment, (PD) CD children 60 days post-EEN treatment, (PE) CD children back to a free diet, (R^2X (cum)) the cumulated amount of variation in matrix X, (R^2) the goodness of fit, (Q^2) the goodness of prediction

Model	R^2X (Cum)	R^2	Q^2	Permutation (999 times)	R^2-Q^2	Valid	*p* CV-ANOVA	significance
PA vs HC	**0.63**	**0.95**	**0.71**	**yes**	**0.24**	**yes**	**1.83E^{-03}**	**yes**
PA vs PB	0.60	0.88	0.51	yes	0.37	no	1.37E^{-01}	no
PA vs PC	**0.65**	**0.88**	**0.66**	**yes**	**0.22**	**yes**	**1.00E^{-02}**	**yes**
PA vs PD	0.67	0.89	0.43	yes	0.46	no	2.42E^{-01}	no
PA vs PE	0.47	0.67	0.33	yes	0.34	no	1.56E^{-01}	no
HC vs PB	0.68	0.99	0.91	yes	0.08	yes	2.03E^{-06}	yes
HC vs PC	0.67	0.99	0.91	yes	0.08	yes	4.81E^{-07}	yes
HC vs PD	0.72	0.99	0.86	yes	0.13	yes	6.69E^{-04}	yes
HC vs PE	0.54	0.99	0.72	yes	0.27	yes	1.19E^{-02}	yes
PB vs PC	0.68	0.97	0.08	yes	0.89	no	9.97E^{-01}	no
PB vs PD	0.61	0.76	0.12	yes	0.64	no	7.16E^{-01}	no
PB vs PE	0.63	0.99	0.93	yes	0.06	yes	3.31E^{-07}	yes
PC vs PD	0.58	0.68	0.24	yes	0.44	no	2.98E-01	no

PC vs PE	0.60	0.98	0.89	yes	0.09	yes	$1.90E^{-07}$	yes
PD vs PE	0.57	0.84	0.69	yes	0.15	yes	$3.43E^{-04}$	yes

By applying the methodology described in (Figure 2.1), eight differentiated metabolites were identified in the PA and HC samples (Table 4.4) There was a clear separation between these groups (Figure 4.2, a). The final model remained valid after data analysis, even for the short list of metabolites as shown in (Figure 4.2).

Table 4.4 List of metabolites that were significantly different in the pre-EEN treatment group (PA) compared to the HCs (HC), based on an OPLS-DA model.

All marker compounds were normally distributed according to a Q-Q test

Putative metabolite	*Pathway*	*(PA/HC)*	*p-value*	*q-value*	*AUC*	*VIP total*	*VIP (pred/ortho)*
Ornithine isomer	unknown	0.15	$7.82E^{-03}$	$2.67E^{-02}$	0.84	1.85	4.28
Dimethylsphingenine	Sphingoid bases	6.54	$2.03E^{-02}$	$3.92E^{-02}$	0.75	1.81	2.82
Tyrosine	Tyrosine metabolism	0.37	$2.64E^{-02}$	$4.98E^{-02}$	0.83	1.63	1.84
SM (d18:1/24:1)	Ceramide phosphocholines (sphingomyelins)	14.52	$3.28E^{-03}$	$2.67E^{-02}$	0.87	1.26	1.08
Eicosatrienoic acid	Biosynthesis of unsaturated fatty acids	16.18	$3.48E^{-04}$	$4.67E^{-03}$	0.88	1.07	1.53
Docosatetraenoic acid	Biosynthesis of unsaturated fatty acids	20.32	$9.11E^{-04}$	$6.15E^{-03}$	0.92	1.02	1.01
Arachidonic acid	Fatty Acids and Conjugates	18.05	$4.79E^{-03}$	$1.94E^{-02}$	0.88	0.99	2.91
Cer (d18:1/18:1)	Ceramides	11.88	$5.31E^{-05}$	$1.08E^{-03}$	0.94	0.89	1.21

Table 4.5 The relative abundance of long chain fatty acids in the faecal extracts based on analysis of a ZICp HILIC column.

(*HC) HC children, (PA) CD children pre-EEN treatment, (PB) CD children 15 days post-EEN treatment, (PC) CD children 30 days post-EEN treatment, (PD) CD children 60 days post-EEN treatment, (PE) CD children back to a free diet*

Mass	**RT**	**Putative metabolite**	**p value HCPA**	**PA/HC**	**p value HCPB**	**PB/HC**	**p value HCPC**	**PC/HC**	**p value HCPD**	**PD/HC**	**p value HCPE**	**PE/HC**
254.2246	3.6	Hexadecenoic acid	0.005	2.334	0.004	3.028	0.001	3.219	0.027	2.287	0.035	1.623
256.2401	3.6	Hexadecanoic acid	0.839	1.038	0.860	1.038	0.957	1.011	0.522	0.866	0.249	1.212
258.1829	3.6	Tetradecanedioic acid	0.016	0.362	0.010	0.332	0.007	0.300	0.068	0.507	0.637	1.200
258.2198	3.5	Hydroxypentadecanoic acid	0.984	1.005	0.713	1.143	0.401	1.379	0.266	1.448	0.716	1.095
260.1988	3.3	Dihydroxytetradecanoic acid	0.891	0.935	0.051	0.240	0.037	0.180	0.034	0.161	0.241	1.512

266.1882	3.4	Hydroxyhexadecatrienoic acid	0.537	0.839	0.007	0.391	0.015	0.459	0.029	0.506	0.192	1.387
268.2036	3.1	Hydroxyhexadecadienoic acid	0.016	0.376	0.001	0.118	0.001	0.113	0.002	0.176	0.552	0.852
270.2195	3.5	Hydroxyhexadecenoic acid	0.940	1.024	0.174	0.702	0.846	1.081	0.902	0.963	0.095	1.472
272.2351	3.5	Hydroxyhexadecanoic acid	0.905	0.947	0.228	0.578	0.389	0.690	0.538	0.785	0.449	1.287
278.2245	3.5	Octadecatrienoic acid	0.004	0.115	0.005	0.143	0.004	0.135	0.004	0.128	0.014	0.293
280.2401	3.6	Octadecadienoic acid	0.212	0.513	0.220	0.517	0.211	0.508	0.098	0.342	0.195	0.500
282.2559	3.6	Octadecenoic acid	0.826	1.093	0.374	0.666	0.839	1.102	0.407	0.688	0.065	1.871
284.2713	3.5	Octadecanoic acid	0.399	0.756	0.022	0.400	0.020	0.394	0.018	0.374	0.928	0.976
288.23	3.2	Dihydroxyhexadecanoic acid	0.810	0.893	0.047	0.252	0.048	0.256	0.052	0.269	0.296	1.436
296.2349	3.5	Hydroxyoctadecadienenoic acid	0.004	0.285	0.004	0.271	0.004	0.269	0.004	0.279	0.235	0.700
298.2506	3.6	Hydroxyoctadecenoic acid	0.797	1.105	0.509	0.784	0.921	1.038	0.950	1.023	0.019	2.064
304.2401	3.5	Eicosatetraenoic acid	0.088	18.052	0.100	4.904	0.049	2.915	0.228	1.968	0.008	3.448
306.2558	3.5	Eicosatrienoic acid	0.002	16.182	0.008	13.854	0.014	9.671	0.049	8.821	0.003	8.751
308.2715	3.5	Eicosadienoic acid	0.028	8.716	0.041	6.338	0.017	6.119	0.008	5.658	0.000	3.709
310.2145	3.5	Dihydroxyoctadecatrienoic acid	0.042	0.593	0.691	0.890	0.368	0.783	0.017	0.496	0.676	1.147
310.2871	3.5	Eicosenoic acid	0.045	1.793	0.308	1.315	0.422	1.181	0.772	1.076	0.020	1.588
312.2301	3.6	Dihydroxyoctadecadienoic acid	0.871	1.053	0.868	1.053	0.860	0.947	0.191	0.647	0.350	1.244

312.2663	3.5	Hydroxynonadecenoic acid	0.232	1.695	0.013	2.360	0.018	2.560	0.045	2.148	0.733	1.149
312.3028	3.5	Eicosanoic acid	0.851	1.071	0.203	1.595	0.082	2.166	0.304	1.614	0.766	1.085
330.2405	3.7	Trihydroxyoctadecenoic acid	0.175	0.508	0.373	1.590	0.646	1.220	0.878	0.938	0.583	1.261
332.2716	3.5	Docosatetraenoic acid	0.006	20.326	0.004	9.694	0.003	6.946	0.006	6.532	<0.001	8.116
334.2144	3.7	Dihydroxyeicosapentaenoic acid	0.964	1.024	0.938	1.043	0.661	0.775	0.605	0.743	0.233	0.414
334.2871	3.5	Docosatrienoic acid	0.127	1.701	0.491	1.457	0.856	1.098	0.405	1.491	0.201	1.664
336.3029	3.5	Docosadienoic acid	0.738	1.226	0.444	0.705	0.037	0.341	0.116	0.488	0.880	0.939
338.3186	3.5	Docosenoic acid	0.119	1.665	0.768	0.935	0.192	0.759	0.185	0.728	0.039	1.459
340.334	3.4	Docosanoic acid	0.610	1.264	0.017	0.320	0.019	0.342	0.026	0.356	0.262	1.570
342.2769	3.4	Eicosanedioic acid	0.021	0.311	0.061	0.454	0.069	0.468	0.088	0.504	0.331	0.718
346.2353	3.9	Tetrahydroxyoctadecenoic acid	0.046	0.447	0.301	0.691	0.168	0.618	0.244	0.670	0.978	0.992
352.3341	3.4	Tricosenoic acid	0.164	1.402	0.006	2.492	0.005	2.596	0.127	1.788	0.088	1.392
354.2408	3.7	Trihydroxyeicosatetraenoic acid	0.799	0.938	0.254	0.697	0.078	0.598	0.239	0.730	0.097	1.433
354.3134	3.4	Hydroxydocosenoic acid	0.209	0.563	0.022	0.361	0.036	0.419	0.088	0.529	0.602	1.160
354.3498	3.4	Tricosanoic acid	0.999	1.000	0.011	0.351	0.019	0.427	0.019	0.408	0.511	1.225
356.329	3.4	Hydroxydocosanoic acid	0.494	0.713	0.011	0.239	0.013	0.270	0.021	0.325	0.461	0.763
364.3342	3.4	Tetracosadienoic acid	0.037	4.794	0.214	2.076	0.651	1.220	0.312	2.154	0.048	3.661
370.2358	3.8	Tetrahydroxyeicosatrienoic acid	0.456	0.840	0.044	0.532	0.012	0.464	0.039	0.536	0.075	1.544

372.2509	3.8	Tetrahydroxyeicosadienenoic acid	0.086	0.574	0.030	0.537	0.004	0.402	0.016	0.497	0.080	1.459
382.2719	3.6	Dihydroxydocosatrienoic acid	0.039	0.268	0.101	0.379	0.041	0.278	0.140	0.459	0.131	0.485
382.3447	3.3	Hydroxy tetracosanoic acid	0.556	0.702	0.090	0.282	0.084	0.270	0.121	0.349	0.835	1.103

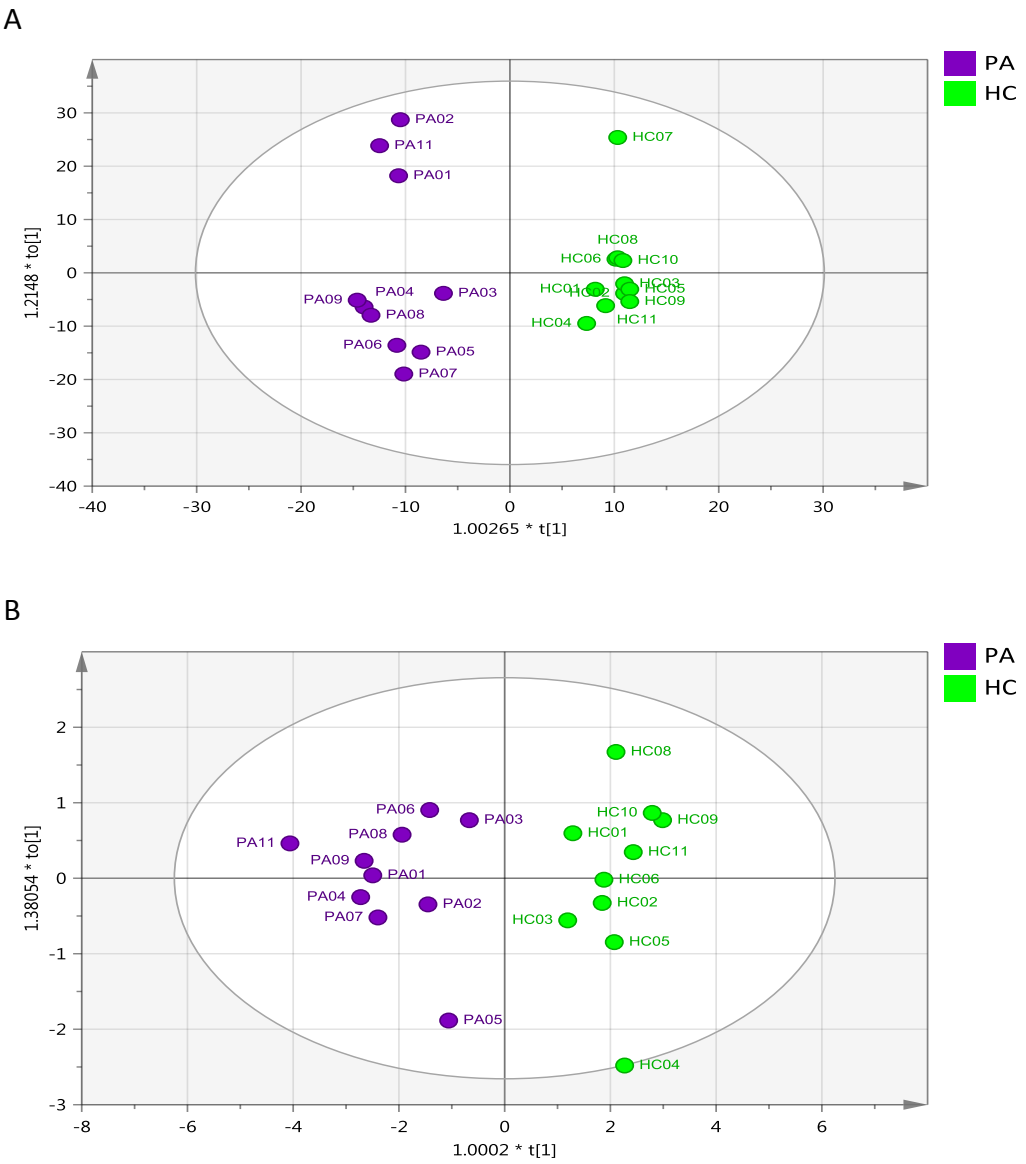


***Figure 4.2*. Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) score plot of pre-EEN samples (PA) against HCs (HC)**.

(**A**) The OPLS-DA model based on 376 metabolites. The model consists of one predictive x score component, component t[1], and one orthogonal x-score component to [1]. The t [1] component explains 19.1% of the predictive variation in x, while the to[1] component explains 11.6% of the orthogonal variation in x, R^2X (cum) = 0.307. The goodness of fit (R^2) = 0.96, the goodness of prediction (Q^2) = 0.848 and the p value associated with the cross-validation analysis of variance (CV-ANOVA) = $5.41E^{-06}$. The PA10 sample was excluded as outlier. (**B**) The OPLS-DA model based on 8 differentiated metabolites. The model consists of one predictive x-score component, component t [1], and one orthogonal x-score component, to [1]. The t [1] component explains 67.7% of the predictive variation in x, while the to[1] component explains 11.7% of the orthogonal variation in x, R^2X (cum) = 0.307. R^2 = 0.837, Q^2 = 0.57 and p CV-ANOVA = $1.47E^{-02}$. The PA10 and HC07 samples were excluded as outliers. Both models were based on Log base 2 variables that were Pareto scaled

The levels of an ornithine isomer and tyrosine were significantly lower in the PA samples than the HC group (Log_2 (PA/HC) = −2.74 and −1.43 for the ornithine isomer and tyrosine, respectively). The remaining metabolites were found in a higher abundance in children with active CD at the sampling points compared to the HC group (Figure 4.3). The eight marker compounds remained largely significantly lower or higher than the controls although some of the metabolites moved closer to the control levels, with the effect being most marked for arachidonic acid and ceramide. The retention times of four of the marker compounds could be matched against available standards. Thus, four of the compounds were only identified to MSI level 2 (De Preter, 2015).

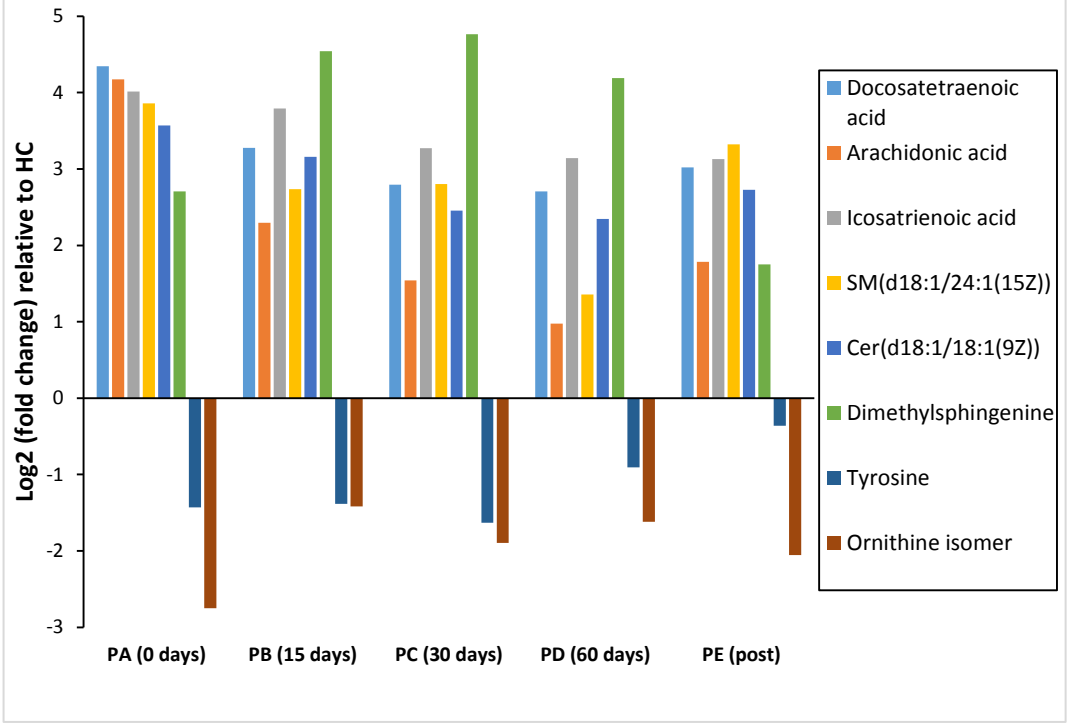


Figure 4.3 Log$_2$ of the fold-change in the eight differentiated metabolites in the CD groups (before, during, and after EEN treatment) compared with the group of HCs.

(HC) HC children, (PA) CD children pre-EEN treatment, (PB) CD children 15 days during EEN treatment, (PC) CD children 30 days during EEN treatment, (PD) CD children 60 days during EEN treatment, (PE) CD children back to a free diet

MSn fragmentation was carried out using an Orbitrap Fusion for these compounds with mixed success. The details of the characterization of the compounds are given in (Table **4.6**). Quite definitive identification of the C20 sphinganine and the C18 sphingosine was achieved. Clear and logical fragments were obtained for the isomer of ornithine although the MS^2 was weak, however it would be difficult to propose a definitive structure based on these. The ceramide yielded abundant fragments it was not possible to make sense of these. There was one fragment at 264.2 in low resolution in MS^3 mode which was the same as a fragment obtained for the C18

sphingosine which is associated with the C18 sphinganine core of the molecule. Correlation plots for the marker compounds against the values obtained for calprotectin for the samples did not reveal any strong correlation between the peak areas for the marker metabolites and the calprotectin values.

Table 4.6 Details of characterization of the eight marker compounds shown in Table 4.4 obtained in positive (+) or negative (–) ion mode.

MS^n fragments obtained at 30 V collision energy for three of the marker compounds shown in Table 4.4 obtained by using an Orbitrap Fusion mass spectrometer at 50000 resolution in MS^2 mode and low resolution in MS^3 mode. Chromatography carried out on ZICpHILIC or and ACE C4 column (C4).

m/z	Rt min	Elemental Composition	Putative name	Deviation ppm	MS^2/MS^3	Comments
133.0971 (+)	87	C5H13O2N2	Ornithine isomer	−0.332	MS^2 115.085 (C5H11ON2), 98.060 (C5H8ON), 69.033 (C4H5NO)	Nearest alternative composition C3H11N5O (+9.8 ppm) Despite the MS^2 fragments making sense (Appendix, Figure S4. 3), it is difficult to propose a definitive structure.
328.3211 (+)	3.4	C20H42O2N	C20 sphinganine	+0.499	MS^2 311.2943 (C20H39O2),310.30951 (C20H42ON) 228.1957(C13H26O2N) 188.1644 (C10H22O2N)	Proposed fragmentation scheme shown in (Figure S4. 4, Appendix). (Spectrum Figure S4. 5).
813.6851 (+)	3.3	C47H94N2PO6	Ceramide d18:1 24:1	−1.145	MS^2 795.61, 553.53 MS^3 (7956) 777.3, 614.6, 495.22, 264.1	This marker remains unidentified since it is not possible to relate the fragments to the proposed structure. Nearest alternative composition. Nearest match C51H91NO6 (1.5 ppm). MS^2 and MS^3 spectra (Figure S4. 6,Appendix) and (Figure S4. 7,Appendix)
182.0810 (+)	13.5	C9H12NO3	Tyrosine	−0.810	-	Matches retention time of standard. Nearest alternative composition C7H10NO2 (+7.9 ppm)
305.2484 (−)	19.5 C4	C20H33O2	Eicosatrienoic acid	−0.438	-	Matches retention time of standard. Nearest alternative composition C18H31ON3 (+3.9 ppm)
329.2484 (−)	191 C4	C22H33O2	Docosapentaenoic acid	−0.406	-	No standard available but logically the retention time falls close to eicosatrienoic acid because number of hydrogens is the same.

303.2329 (−)	18.5 C4	C20H31O2	Arachidonic acid	−0.045	-	Matches retention time of standard. Nearest alternative composition C18H29ON3 (+4.3 ppm)
564.5361 (+)	3.1	C36H70NO3	Octadecenoylsphinganine	+19.8	MS^2 546.5239(C36H68NO2) 528.5128 (C36H66NO) 282.2782 (C18H38NO) 264.2680 (C18H36N)	Proposed fragmentation scheme shown in (Figure S4. 8, Appendix). MS2 spectrum (Figure S4. 9, Appendix).

The Appendix Table S4. 1 shows a complete list of significant metabolites in ascending molecular weight, indicating where the retention time of the metabolite was matched to that of a standard as well as the p values and ratios obtained for the comparison of the HC group against the pre-treatment samples. The fatty acids identified by the ZICpHILIC screen were not strongly retained on the column.

To confirm their identity, two marker fatty acids, arachidonic acid and eicosatrienoic acid, were matched against the retention times of their corresponding standards on a C4 reversed phase column. A quantitative estimate of the fatty acids in the samples was performed by preparing calibration curves in the range 0.1 µg to 16 µg/ml and estimating the fatty acid content in the faecal extracts for the HC and pre-treatment samples based on the calibration lines. Table **4.7** reports the levels of the fatty acids in the HC and pre-EEN treatment samples in µg/g.

Table 4.7 Concentration of fatty acids in each sample (µg/g of dry faeces). P-value based on Log base2.

Heathy Controls			Crohn's Disease		
Sample	**Arachidonic Acid**	**Cis-8, 11, 14-Eicosatrienoic acid**	**Sample**	**Arachidonic Acid**	**Cis-8, 11, 14-Eicosatrienoic acid**
HC01	47.6	112.4	PA01	4406	3854.4
HC02	13.6	10	PA02	4365.2	1671.6
HC03	25.6	19.6	PA03	432.4	492.4
HC04	7.6	3.6	PA04	1644.8	9462.4
HC05	15.2	2	PA05	92	514.4
HC06	63.6	210.4	PA06	3510	3600.4
HC07	86.8	136.4	PA07	98	196
HC08	144.8	127.6	PA08	44.8	904.8
HC09	12.8	5.6	PA09	27.6	181.6
HC10	13.2	16.4	PA10	14	50
HC11	34.8	60	PA11	4262.8	1029.2
Mean	42.4	64	Mean	1718	1996
SD	42.4	71.6	SD	1985.2	2806.8

SEM	12.8	21.6	SEM	598.4	846.4
			p-value	0.019	0.046

4.6 Discussion

In this study, several amino acids and amino acid metabolites were present at significantly higher levels in the pre-EEN treatment samples of the CD patients in comparison with controls. These observations are generally in line with Kolho et al. who found elevations of the following metabolites in faecal samples from CD patients: aspartate, glycine, tryptophan, carnosine, allantoin, citrulline, serine, threonine, ornithine, creatine, asparagine, choline, kynurenine, histidine, taurine, phenylalanine, alanine, and metanephrine (Kolho *et al.*, 2017). In their study, these elevated metabolites could be used to discriminate between the CD patients and the HCs. Jannson et al. found that tyrosine and its metabolites as well as phenylalanine and tryptophan were significantly higher in CD patients (Jansson *et al.*, 2009). In the current study, tyrosine levels were significantly lower in the pre-EEN treatment versus the HC samples and remained either lower or significantly lower throughout the treatment and post-treatment samples (Table S4. 1 and Figure S4. 3, Appendix).

In another study, Bjerrum et al. found that leucine, isoleucine, valine, lysine, alanine, tyrosine, phenylalanine, and glycine were all present at high levels in faecal extracts from CD patients compared to HCs (Bjerrum *et al.*, 2015). The study by Bjerrum is in agreement with our results except for the tyrosine levels, which were consistently lower. Schicho et al. (2012) reported increased levels of methionine, lysine, glycine, arginine, and proline and decreased levels of valine, tyrosine, and serine in faecal

extracts from CD patients. Schicho et al.'s findings regarding tyrosine levels reflect our tyrosine results, but we found that valine and serine were either consistently higher than the controls or no different from the controls.

In the current study, we used a rigorous selection procedure to determine important markers that could discriminate between HC and pre-treatment CD samples and then followed these markers during the course of EEN. Given the small set of patients, it was not possible to assume that the peak areas obtained for the metabolites were normally distributed, even after logarithmic transformation. Although p values have been reported in previous studies using similarly small sample sets, we could not be certain that a null hypothesis could be rejected without conducting a Q-Q test. For example, in the current study, taurine is significantly higher in most of the treated and untreated patient samples in comparison with the control (Table S4. 1, Appendix), and it is tempting to conclude that taurine is an important disease marker, given its anti-inflammatory effects (Schuller-Levis and Park, 2003). However, the Q-Q test indicated that taurine was not normally distributed and appears to be normally distributed in two groups (Figure S4. 2, Appendix); thus, its p values could not be reported. The same was true for acetyl choline, which was significantly higher in all the patient samples but did not pass the Q-Q test returning a low R^2 value (Figure S4. 2, Appendix)

Q-Q tests are time consuming to perform, and it is not possible to carry these out for large numbers of markers. Multivariate statistics using the SIMCA-P software (14.1) was applied to solve this problem. The multivariate models produced by the SIMCA-

P software do not assume a normal distribution of marker compounds. In the model shown in Figure 5, the non-parametric jack-knife test (Efron and Gong, 1983) was used to select reliable markers, reducing the marker list to eight. A Q-Q test was then applied to these markers to check for normal distribution. Six out of the eight markers were normally distributed with the ceramide (SM (d18:1/24:1)) having too many missing values to give normal distribution (Figure S4. 2, Appendix).

Large differences were identified in the levels of these marker compounds between the HC and the CD patients. Only two marker compounds were reduced in the CD patients, tyrosine and an ornithine isomer. Tyrosine has previously been reported as a CD marker that was increased in faecal extracts from CD patients (Kolho *et al.*, 2017) and decreased in the plasma from CD patients. In our study, the low tyrosine levels were not significantly changed after EEN treatment in comparison with the HC group. Several tyrosine metabolites were also present in low amounts in the CD patients, including dopamine, noradrenaline, metanephrine, normetanephrine, adrenaline, and DOPA. Catecholamines are normally at very low concentrations in plasma, but the levels excreted in urine are generally much higher. There is no substantial literature on the levels of catecholamines in faeces. Further research is needed on this issue, as it was not possible to validate the identities of these putatively identified markers when their retention times were compared with authentic standards.

The other marker compound that was found at reduced levels in our analysis of CD patients, with an average intensity of 0.15 compared to that in the HCs, was an ornithine isomer. Since ornithine has two basic centres, it runs very late in our HILIC

method, while the marker compound ran much earlier than the ornithine standard. Two ornithine isomers were present in our database; one of these would have been expected to elute late from the column since it is a diamine, but N4-acetyl-N4-hydroxy-1-aminopropane would be expected to elute early. This ornithine isomer is found as a biosynthetic intermediate in the synthesis of siderophores in Rhizobia bacteria, but whether similar pathways might exist in the microbiome bacteria is not known (Fabiano and O'Brian, 2012).

Dietary omega 6 fatty acids that include arachidonic acid and eicosatrienoic acid may be implicated in IBD (Musso, Gambino and Cassader, 2010). In our study, the levels of arachidonic acid, eicosatrienoic acid, and docosatetraenoic acid were much higher in the CD patients compared to the HC group (Table **4.4**) and (Table **4.7**) and remained high both pre- and post-EEN treatment. These fatty acids cannot have derived from the enteral nutrition formula since their levels were higher in both the PA and PE samples compared with the HC group. In addition, Table **4.5** indicates that elevation does not occur for most of the fatty acids evaluated in this study. The greatest accumulations were seen for three C20 polyunsaturated acids and a C22 polyunsaturated acid. In contrast, there was not much difference in the levels of C16 and C18 acids between the CD patients at all the time points and the HC group. These results suggest that CD pathogenesis or progression might be related to the metabolism or absorption of this fatty acid class and replicate findings of other groups that demonstrate higher levels too (Uchiyama *et al.*, 2013). Although these fatty acids are not strongly indicative of the effectiveness of treatment. It can be seen from the data in (Table **4.6**) that while the fatty acid marker compounds are much higher in

the CD group than in the HC group, there is a wide variation of levels within the CD group, this might give an indication of the severity of the disease but since the calprotectin measurements, as mentioned above, did not correlate with the levels of the fatty acids in the samples there is no means of confirming this.

Omega 6 fatty acids have been shown to be pro-inflammatory in a mouse model (Kaliannan *et al.*, 2015); those pro-inflammatory effects were suppressed in transgenic mice that were capable of converting omega 6 to omega 3 fatty acids. Omega 3 fatty acids have been shown to promote the formation of intestinal alkaline phosphatase, which breaks down the potent pro-inflammatory lipopolysaccharides produced by Escherichia coli, which may, in turn, drive CD inflammation. In our study, EEN treatment had some impact on the levels of these fatty acids, but they still remained higher in the CD patients than in the controls throughout all phases of treatment.

The role of sphingomyelins and ceramides in CD has been investigated, with variable findings (Angulo *et al.*, 2011; Baur *et al.*, 2011; Sewell *et al.*, 2012). In the current study, three of the elevated markers in the CD patients were in the sphingolipid category. The sphingolipid levels were not greatly affected by EEN treatment. A previous study observed that probiotic bacteria in a mouse IBD model produced a neutral sphingomyelinase that could convert sphingomyelin into ceramides, promoting apoptosis of mucosal immune cells leading to improved homeostasis and reduced inflammation (Uchiyama *et al.*, 2013). This theory would explain the elevated sphingomyelin levels in the current study, but it does not conform to the

elevated levels of pro-apoptotic ceramides found in the CD samples (Table S4. 3, Appendix). Of note, Sewell et al. found no differences in the ceramide composition of macrophages taken from CD patients compared to a control group. The ceramides monitored in that study corresponded largely to those shown in Appendix Table S4. 2 (Sewell *et al.*, 2012).

The partial elucidation of the structures of the marker compounds for which matching standards were not available was carried out and is summarised in Table **4.6**. Confidence in the identity of two of the sphingolipids is high and comprehensive fragmentation schemes are shown in Appendix Figure S4. 4 and Figure S4. 8. However, complete elucidation of the structure of the ceramide so far eludes us.

Chapter 5:

Metabolomics Discrimination between Crohn's Disease and Ulcerative Colitis

5 Metabolomics Discrimination between Crohn's Disease and Ulcerative Colitis

5.1 Abstract

Background and aim: Crohn's Disease (CD) and Ulcerative Colitis (UC) are considered as the two major subtypes of disorder diagnosed under inflammatory bowel disease (IBD). A non-invasive metabolomics study of biomarkers might explain the differences between CD and UC by using non-invasive samples. In this study, an untargeted metabolomics approach was applied to discriminate between CD and UC.

Methodology: Metabolomic profiling using high resolution mass spectrometry with hydrophilic interaction chromatography was applied to 62 faecal extracts from 30 healthy control (HC) children, 30 Crohn's disease (CD) and 12 Ulcerative colitis children. An orthogonal partial least square-discriminant analysis (OPLS-DA) model was applied to output data using SIMCA-P 14.

Results: The OPLS-DA model was able to discriminate the CD samples from the UC and HC. There were 127 common significant metabolites shared between CD and UC. Most of these metabolites showed the same pattern in both diseases in comparison with HC subjects. Based on VIP values, the top 10 metabolites were used in the OPLS-DA model, and there was a clear separation between CD and UC with *p* CV-ANOVA = 5.30541e-007.

Conclusion: Untargeted metabolomics analysis was performed to classify the faecal extracts from patients suffering from inflammatory bowel diseases to evaluate the use of this technique as a diagnostic tool and categorize specific metabolites in the faecal extract of participants with specific types of inflammation.

5.2 Introduction

Crohn's Disease (CD) and Ulcerative Colitis (UC) are considered as the two major subtypes of disorder diagnosed under inflammatory bowel disease (IBD). These disorders may be affected by environmental and genetic factors (Dupaul-Chicoine, Dagenais and Saleh, 2013). Moreover, immune system modulation (MacDonald, 2011) and gut microbiota (Le Gall *et al.*, 2011) may have significant roles in IBD pathogenesis (Figure 5.1). Therefore, metabolites in the gut are most likely to be changed. Metabolomics approaches could be used to diagnose and differentiate between related diseases. However, there is no clear and specific metabolomics methodology to discriminate IBD disorders. This is because IBD has different stages of prognosis and requires a combination of clinical examinations such as radiology, endoscopy, histology and biochemical tests (Tontini *et al.*, 2015). In addition to disease diagnosis applications, personalized medicine is one of the recent areas where metabolomics can be applied (Woodcock, 2007).

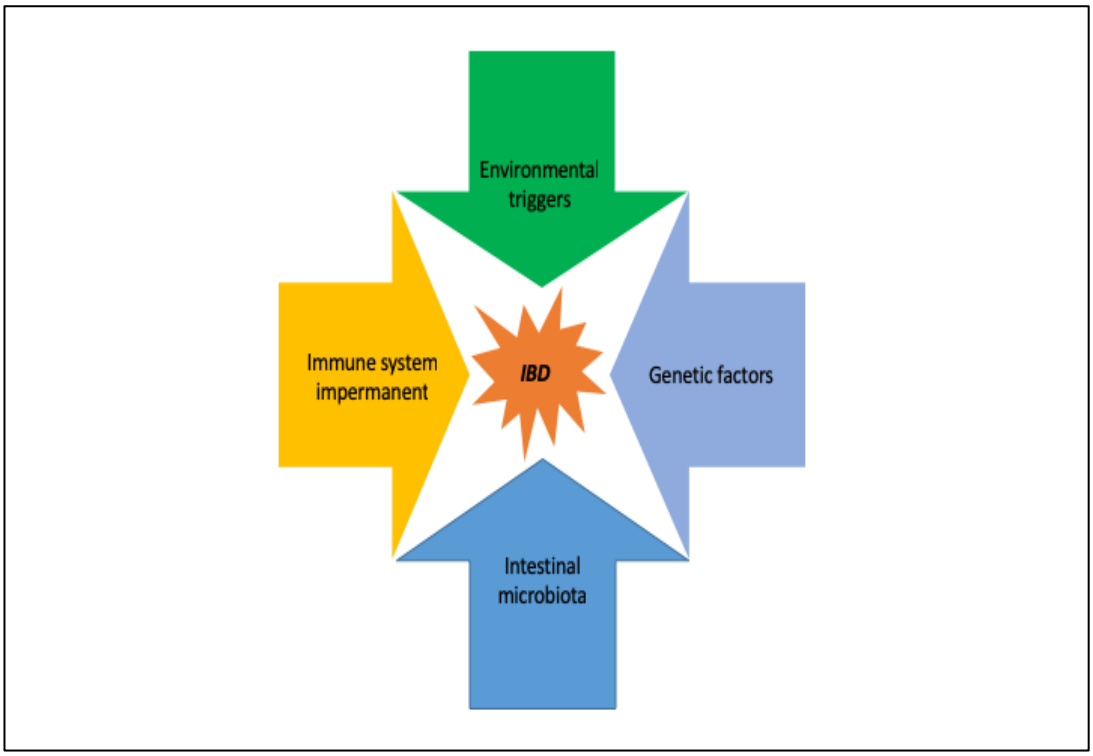


Figure 5.1 Factors affecting IBD pathogenesis.

Although IBD could be diagnosed using multidisciplinary clinical examinations, the techniques have disadvantages when compared with metabolomics approaches. The disadvantages include time, cost and their invasive nature. In addition they leave about 10% of IBD patients with unclassified diagnosis of IBD (Geboes *et al.*, 2008). Specific and speedy diagnosis may play significant roles in patient compliance and quality of life as well as in the treatment of the disease (Ricart *et al.*, 2008).

5.3 Metabolomics discriminations between CD and UC

Non-invasive diagnostic methods are needed to differentiate between CD and UC. This could enhance disease control, management and patient compliance. A metabolomics study of biomarkers might explain the differences between CD and UC by using non-invasive samples. It is known that gut microbiota may discriminate IBD

subtypes from each other (Sartor, 2008; Le Gall *et al.*, 2011), therefore, metabolomics of faecal extracts was used to examine gut microbiota and differentiate between IBD subtypes and HCs as well as within IBD subtypes. Several studies have discussed and examined metabolomics biomarkers to differentiate IBD subtypes (Table **5.1**)

Table 5.1 Summarization of previous metabolomic studies using faecal extract samples.

		CD vs HC		UC vs HC		CD vs UC		Reference
Samples	Analytical method	Increased biomarkers	Decreased biomarkers	Increased biomarkers	Decreased biomarkers	Increased biomarkers	Decreased biomarkers	Reference
CD (n= 10), UC (n= 10), and HC (n= 13)	^{1}H NMR Spectroscopy.	Alanine Isoleucine Leucine Lysine Valine	Acetate Butyrate Methylamine Trimethylamine	Glutamate Lysine	Butyrate Methylamine Trimethylamine	Alanine Glycerol Isoleucine Leucine Lysine Valine	Acetate	(Marchesi *et al.*, 2007)
CD (10 twin pairs), HC (7 twin pairs)	Ion cyclotron resonance-Fourier transform mass spectrometry (ICR-FT/MS	Dopaquinine (Tyrosine metabolism) (Bile acid metabolism) Trihydroxy-6b-cholanate Taurocholate Chenodeoxyglyco-cholate linoleic acid Palmitic acid Arachidonic acid	prostaglandin F2a					(Jansson *et al.*, 2009)
UC (n = 13 ; 31 Samples), IBS (n = 10; 21 samples), HC (n = 22 ; 72 Samples)	^{1}H NMR Spectroscopy			Taurine Cadaverine bile acid Choline glucose	2-methylbutyrate Decreased level of SCFA in IBD			(Le Gall *et al.*, 2011)

CD (n=44), UC (n=48), HC (n=21)	1H NMR Spectroscopy	Isoleucine Leucine Valine Lysine Alanine Tyrosine Phenylalenine Glycine	Butyrate Propionate	Isoleucine Leucine Valine Lysine Alanine Glycine	Butyrate Propionate	**Poor model for active CD vs active UC**	Aspartic acid glutamate for inactive CD vs inactive UC. Poor model for active CD vs active UC	(Bjerrum *et al.*, 2015)
CD (n=16), UC (n=14), HC (n=29)	Ultra-pressure liquid chromatography tandem mass spectrometry [UPLC-MS/MS]	Glycine Tryptophan Carnosine Allantoin Citrulline Serine Ornithine Creatinine Glyceraldehyde Choline Kynurenine Phenylalanine Alanine Normetanephrin	Aspartate Threonine Asparagine Cytosine Histidine Taurine	Aspartate Glycine Tryptophan Carnosine Allantoin Citrulline Serine Threonine Ornithine Creatinine Asparagine Glyceraldehyde Choline Kynurenine Histidine Taurine Phenylalanine Alanine Normetanephrine	Cytosine	Pyridoxine 4-Pyridoxate Orotate Kynurenate	Asparagine Aspartate 5-Hydroxytrypt Guanosine Serine Glutamine Arginine Threonine G-Glutamylcyst Glycine Alanine Methionine Ornithine Tyrosine Taurine Histidine	(Kolho *et al.*, 2017)

5.4 Aim of the study

- Untargeted metabolomics profiling of CD and UC in paediatric patients using LC-MS as an analytical tool and multivariate analyses as a statistical tool.
- Identification of significant metabolites that are shared between CD and UC.
- Identifying the significant metabolites that discriminate CD from UC based on comparison with HC.
- Identification of the significant metabolites that discriminate CD from UC based on direct comparison (CD vs UC).

5.5 Materials and methods

5.5.1 Solvents and chemicals

Chemicals and solvents used are given in section 2.1.

5.5.2 Sample preparation

Samples collection and extraction were carried out by Dr. Konstantinos Gerasimidis and his group at university of Glasgow. Samples were freeze dried and extracted immediately with chloroform/methanol/water (1:3:1 v/v). The extracts were stored at −80 ◦C until analysis by LC-MS. Samples were randomized to avoid inter-batch differences. Pooled samples (n = 9) were prepared from a combination of all samples and intermittently injected throughout the sequence. The samples were randomised and analysed in batches of seven with one pooled sample in between batches in the LC-MS analysis. The participants' details are described in Table **5.2**.

Table 5.2 Subject data for HCs and patients.

NA = not recorded, BMI= Body Mass Index

Groups	**CD**	**UC**	**HC**
Number of participants	30	12	30
Age (mean$_{(year)}$ ± SD)	11.56 ± 3.22	13.16 ± 2.61	9.62 ± 3.17
Sex: **Male** **female**	 24 6	 NA NA	 NA NA
Faecal Calprotectin (mean$_{(mg/kg)}$ ± SD)	1327.80 ± 437.05	1479.66± 610.85	57.10± 121.71
Weight (Z-score) (mean ± SEM)	-0.46 ± 0.23	-0.36 ± 0.63	0.96 ± 0.27
Height (Z-score) (mean ± SEM)	-0.19 ± 0.16	-0.25 ± 0.35	0.90 ± 0.35
BMI (Z-Score) (mean ± SEM)	-0.79 ± 0.28	-0.57 ± 0.70	0.50 ± 5.03

5.6 LC-MS analysis

5.6.1 Mobile phase solutions for ZIC-pHILIC chromatography

The mobile phases for ZIC-pHILIC analysis and their preparation are described in section 2.2.1

5.6.2 HPLC setup

The HPLC setup and conditions are described in section 2.2.3

5.6.3 Orbitrap Exactive MS setup

The experimental conditions and procedures are described in section 2.2.4

5.6.4 Data analysis

Data analysis and methods are described in section 2.3

5.6.5 Group comparisons

Since the study consisted of three major groups, there were three comparisons as follows:

A. CD vs HC
B. UC vs HC
C. CD vs UC

Significant metabolites were classified based on their pathways and common metabolites in CD and UC versus HC were separated from the specific metabolites identified in each group.

5.7 Results

5.7.1 Data visualization

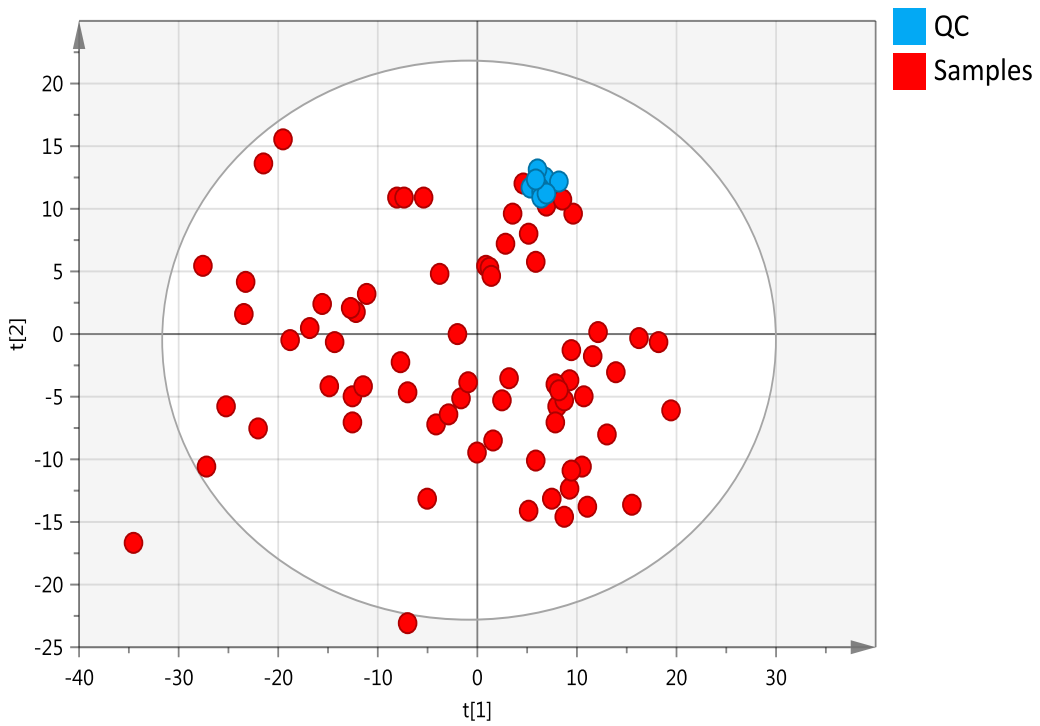


Figure 5.2 2D PCA score plot for QC (pooled) faecal water samples.

The plot shows the clustering of pooled samples (light blue) compared to the rest of samples (red). The model was generated based on normal values and Pareto scaling method.

Metabolomics analysis of 72 faecal extract samples was accomplished using LC-MS. After preparing pooled samples by taking 10 µL from each sample and transferring into single HPLC vials, the batch runs were carried out to assess the time effect on the instrument. The LC-MS instrument was programmed to inject one pooled sample after every seven faecal samples. Therefore, nine pooled samples were analysed during each experiment (Figure *5.2*).

The relative standard deviation (RSD) was calculated with the nine pooled samples based on total intensities in each pooled sample to quantify the precision of the measurements. The RSD value between all pooled samples was 10.9%. This clearly indicates that any metabolomics differences between samples cannot be due to instrumental factors only. The RSD values for each putative biomarker were calculated based on the total intensities of the biomarker in the pooled samples. The highest RSD value was observed for 16, 16-dimethyl-PGE1 (65.4%) and the lowest RSD value was for 1α, 3α-Dihydroxy-5β-cholan-24-oic Acid (3.9%). 62 putative biomarkers were excluded from the analysis since they had RSD values more than 30 in the pooled samples. The total number of putatively identified metabolites remaining to complete the data analysis was 619.

5.7.2 Model selection

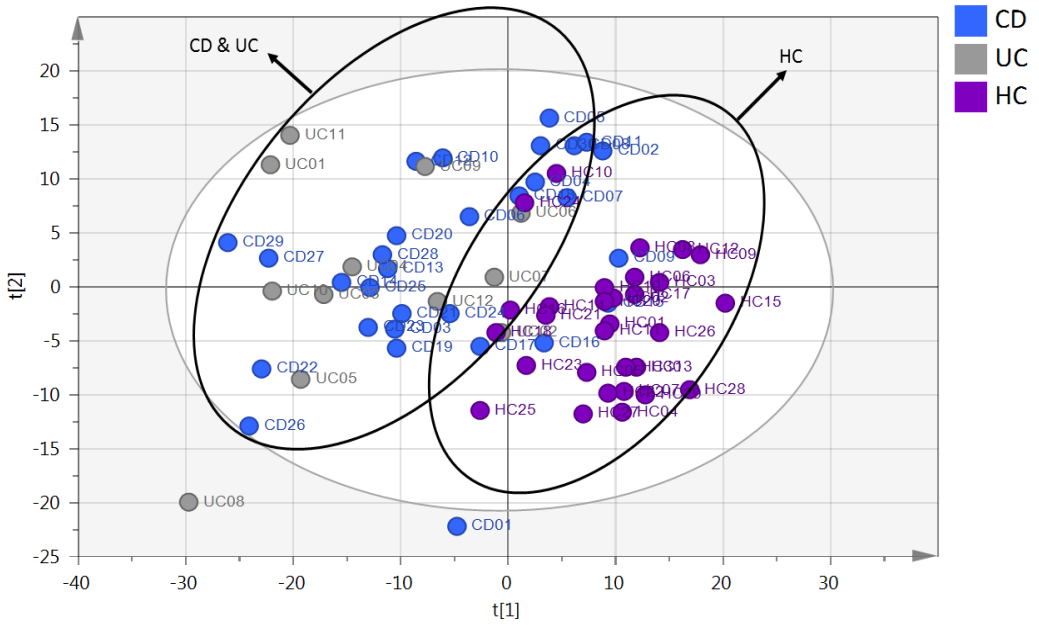


Figure 5.3 2D PCA score plots for CD (Blue-CD) and UC (Grey-UC) against HCs (purple-HC).

$R^2X(Cum)=0.643$, $Q^2=0.382$.

The unsupervised (PCA) model showed there is no clear separation between disease groups especially the CD and UC group (Figure 5.3). HC (purple samples) were well separated from disease groups. This model was able to describe 64.3% of the metabolomics changes from all samples. Despite that, the model was considered an invalid model since Q^2 was less than 0.4 (Worley and Powers, 2012). For this reason, supervised (OPLS-DA) model was used for further analysis.

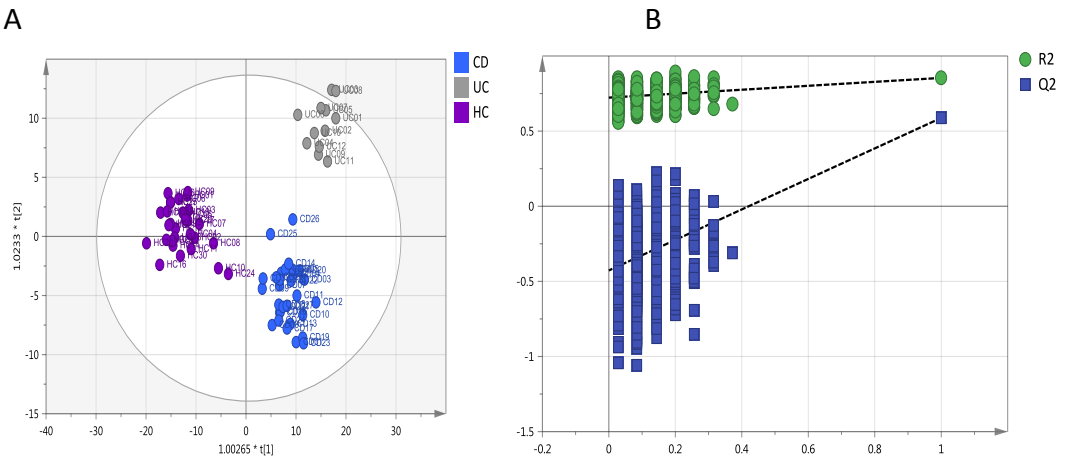


Figure 5.4 OPLS-DA score plots for CD (Blue-CD) and UC (Grey-UC) against HCs (purple-HC) faecal extract samples and its permutation test (999 times).

(A) The figure shows distribution of 72 samples based on readings of 619 putative biomarkers The model consists of two predictive x-score components; component t[1] and four orthogonal x-score components to[1-4]. t[1] explains 21.6% of the predictive variation in x, to[1-4] explains 31% of the orthogonal variation in x. R^2X (Cum) = 0.526, R^2Y(Cum)= 1, R^2 (Cum) =0.891, Q^2 (Cum)=0.652. (B) Permutation test for all samples based on readings of 619 putative biomarkers. The model will be valid if all blue Q^2 values to the left are lower than the original points to the right or the blue regression line of the Q^2 points intersects the vertical axis (on the left) at, or below zero. The model was generated based on Log base2 values and Pareto scaling method.

OPLSDA model showed clear discrimination between all studied groups (Figure *5.4*, A). The model explained 52.6% (R^2X (Cum) = 0.526) of the metabolite's changes. Around 89% of between subject variability was explained by variability in the metabolites. Based on 999 random permutation tests (Figure *5.4*, B) calculated with the default SIMCA cross-validation, this model has valid predictive ability in comparison with the permutated Q^2 values.

Using the misclassification table (Table 5.3) for discriminant analysis model (showing the percentage of correct classification), it shows how accurately the selected model

classifies the observations into known groups. All samples in the OPLS-DA model were classified to the correct group.

Table 5.3 Misclassification table showing the proportion of correctly classified observations in IBD patients and HC using OPLS-DA model

	Members	Correct	CD	UC	HC	No class
CD	30	100%	30	0	0	0
UC	12	100%	0	12	0	0
HC	30	100%	0	0	30	0
Total	72	100%	30	12	30	0
Fisher's prob.	1.10E-11					

5.7.3 Biomarker identification

The model shown in (Figure *5.4*) was divided into three comparisons. The first comparison was CD samples against HCs (CD vs HC) and the second comparison was UC against HCs (UC vs HC). Finally, both diseases were compared to each other (CD vs UC). The Venn diagram (Figure 5.5) shows a comparative relationship between the groups in the metabolites from the two disease groups against HC. As illustrated in the diagram, there were common and significant metabolites from the two comparisons. From the total number of putative metabolites (619 metabolites), there were 127 (≈ 20%) significant metabolites common between CD and UC vs HC comparisons. Moreover, the numbers of specific putative metabolites were 34 (5.5%) and 47 (7.6%) for CD and UC, respectively.

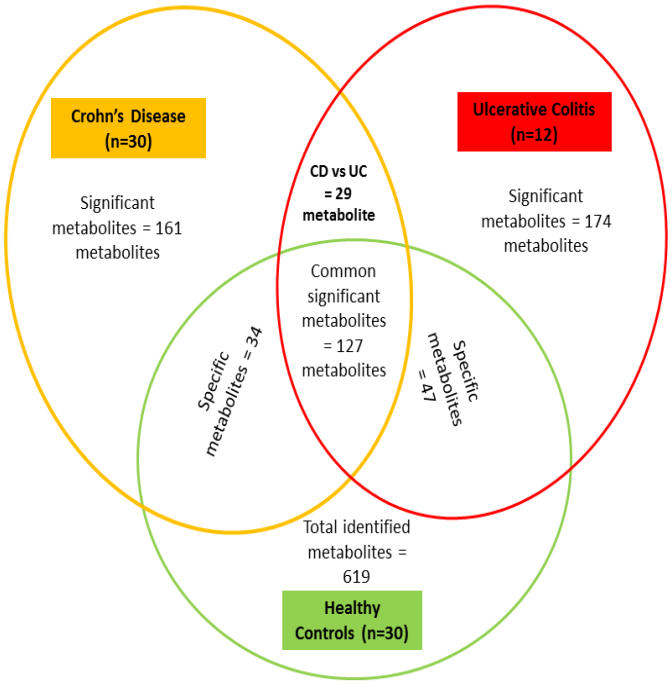


Figure 5.5 Venn diagram Venn-Diagrams of the screened metabolites.

The statistical significant metabolites were FDR < 0.05 and AUC >0.6.

5.7.3.1 Common putative metabolites between CD and UC compared to HCs

There were 127 common significant metabolites between CD and UC. These metabolites were classified into 27 pathways as shown in (Table 5.4). The top 10 pathways that are altered by CD (Figure 5.6) and by UC (Figure 5.7) include fatty acids and their conjugates, glycerolipid metabolism, diacylglycerols, arginine and proline metabolism. These groups showed the same pattern in both diseases in comparison with HC subjects. Arginine and proline pathways differed in one metabolite identified as 5-Guanidino-2-oxopentanoate. This metabolite increased in CD and decreased in UC.

Table 5.4: The number of significant metabolites according to pathways.

The –ve sign indicate decreased metabolites and the +ve sign indicates increased metabolites in related to HC subjects

		CD vs HC		UC vs HC	
Pathway	**No. of sig. metabolites**	**Increased metabolites**	**Decreased metabolites**	**Increased metabolites**	**Decreased metabolites**
Arginine and proline metabolism	6	+5	-1	+4	-2
Bile acid biosynthesis	5	+1	-4	+1	-4
Biosynthesis of steroids	1	+1	0	+1	0
Biosynthesis of unsaturated fatty acids	2	+1	-1	+1	-1
biotin biosynthesis II	1	+1	0	+1	0
Clavulanic acid biosynthesis	3	+2	-1	+1	-2
Diacylglycerols	8	0	-8	0	-8
Fatty Acids and Conjugates	14	2	-12	2	-12
Fatty alcohols	1	+1	0	+1	
Fatty aldehydes	2	+1	-1	+1	-1
Glycerolipid metabolism	11	+9	-2	+9	-2
glycine betaine biosynthesis III	3	+1	-2	+1	-2
Histidine metabolism	1	0	-1	0	-1
Isoprenoids	1	0	-1	0	-1
Monoradylglycerols	1	0	-1	0	-1
Octadecanoids	1	0	-1	0	-1
Oxygenated hydrocarbons	1	0	-1	0	-1
Porphyrin and chlorophyll metabolism	1	0	-1	0	-1
Pyrimidine metabolism	1	0	-1	0	-1

Secosteroids	3	0	-3	0	-3
Spermine and spermidine degradation	1	+1	0	+1	0
Sphingoid bases	1	0	-1	0	-1
Steroid conjugates	1	0	-1	0	-1
Sterols	11	0	-11	0	-11
linamarin degradation	1	0	-1	0	-1
Tryptophan metabolism	1	0	-1	0	-1
Miscellaneous	45	+7	-38	+5	-40

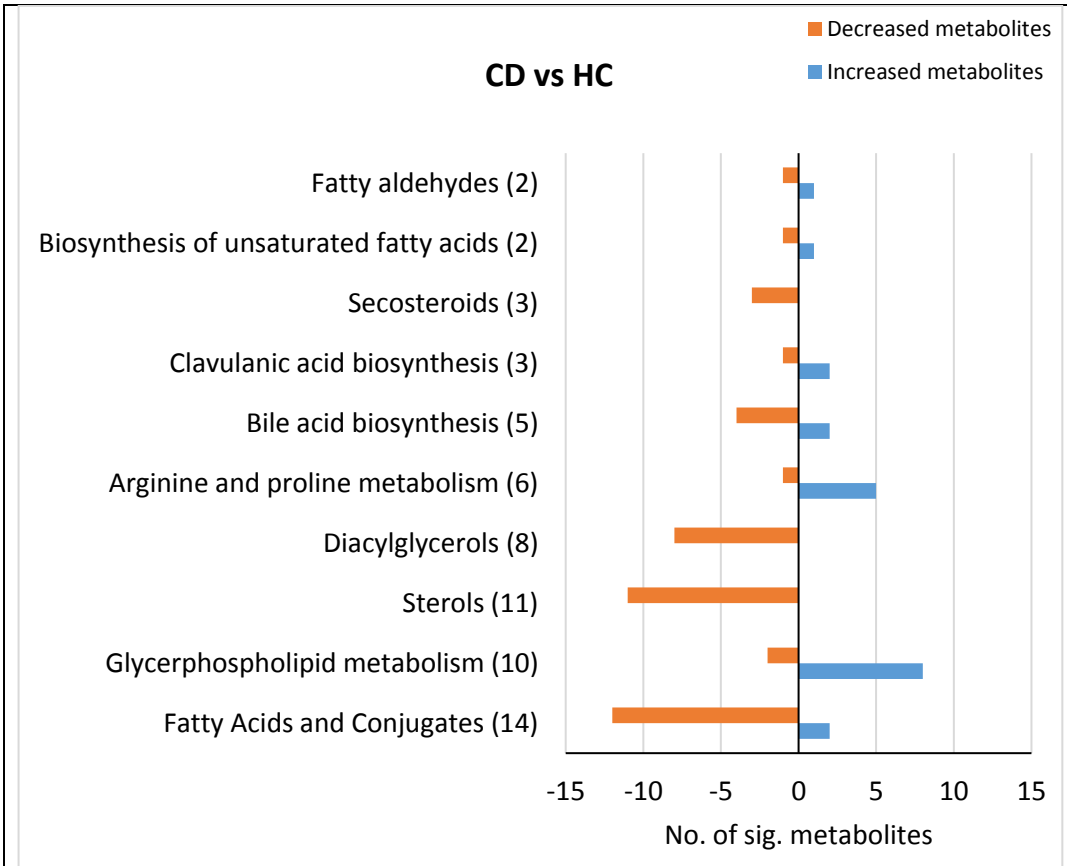


Figure 5.6 Metabolic pathways are altered in CD (CD) compared to HC (HC) subjects.

Metabolite levels were either increased (blue) or decreased (orange).

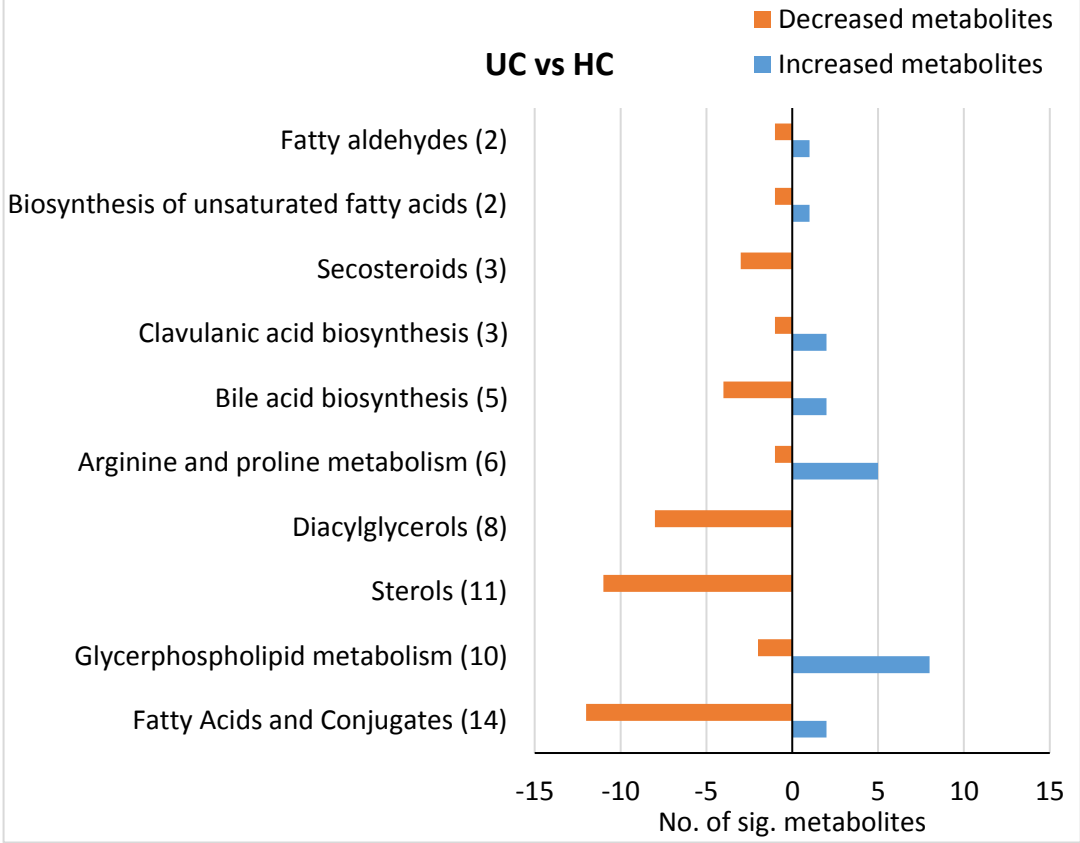


Figure 5.7 Metabolic pathways are altered in UC (UC) compared to HC (HC) subjects.

Metabolite levels were either increased (blue) or decreased (orange)

In fatty acid and conjugate metabolism (Figure 5.8), 14 metabolites were classified as common fatty acids shared between CD and UC. All of these fatty acids appeared with negative correlation relative to HC except for two fatty acids. The positive correlated metabolites were arachidonic acid and docosatetraenoic acid. The negative correlation showed as higher in CD comparing to UC.

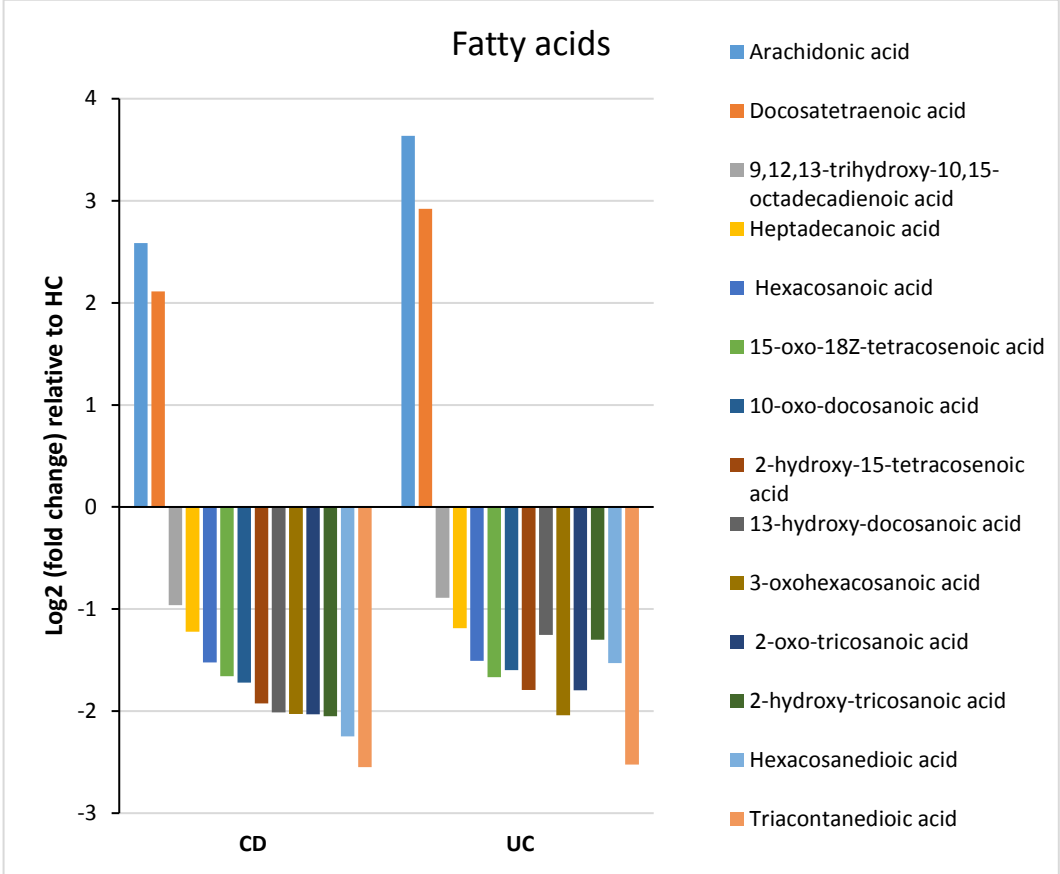


Figure 5.8 Common Fatty acids shared between CD and UC in relative to HCs.

Y-axis represent Log base 2 of metabolite intensity.

Metabolites related to diacylglycerols were found with negative correlation relative to HC in both diseases (Figure 5.9). The diacylglycerols ratios had lower values in UC in comparison to CD.

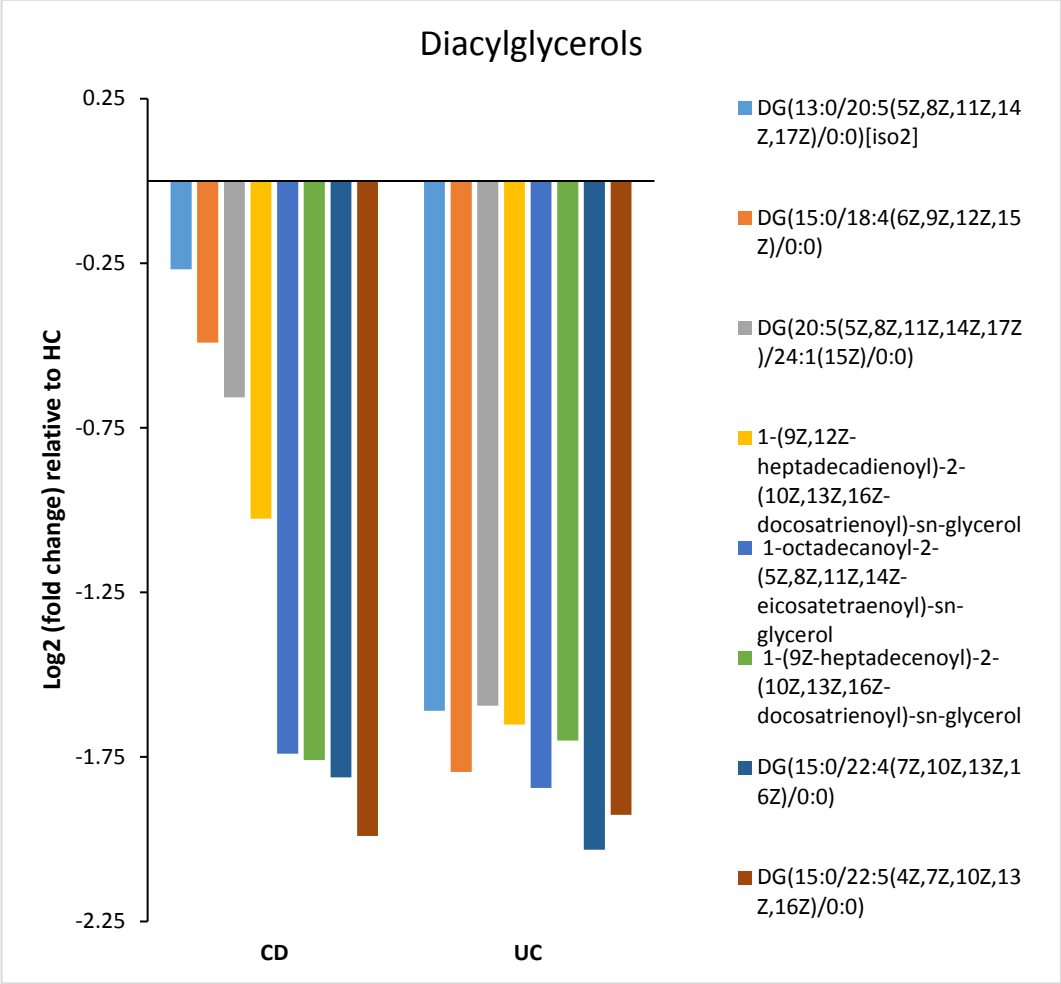


Figure 5.9 Common Diacylglycerols shared between CD and UC in relative to HC.

Y-axis represent Log base 2 of metabolite intensity.

From the bar graph (Figure 5.10), diacyl glycerophospholipids have a clear positive correlation with both diseases in comparison to HC. In both diseases, eight metabolites out of ten showed increase and only two metabolites had negative Log ratios. It also appeared that glycerophospholipids in UC had a greater Log ratio in comparison to CD.

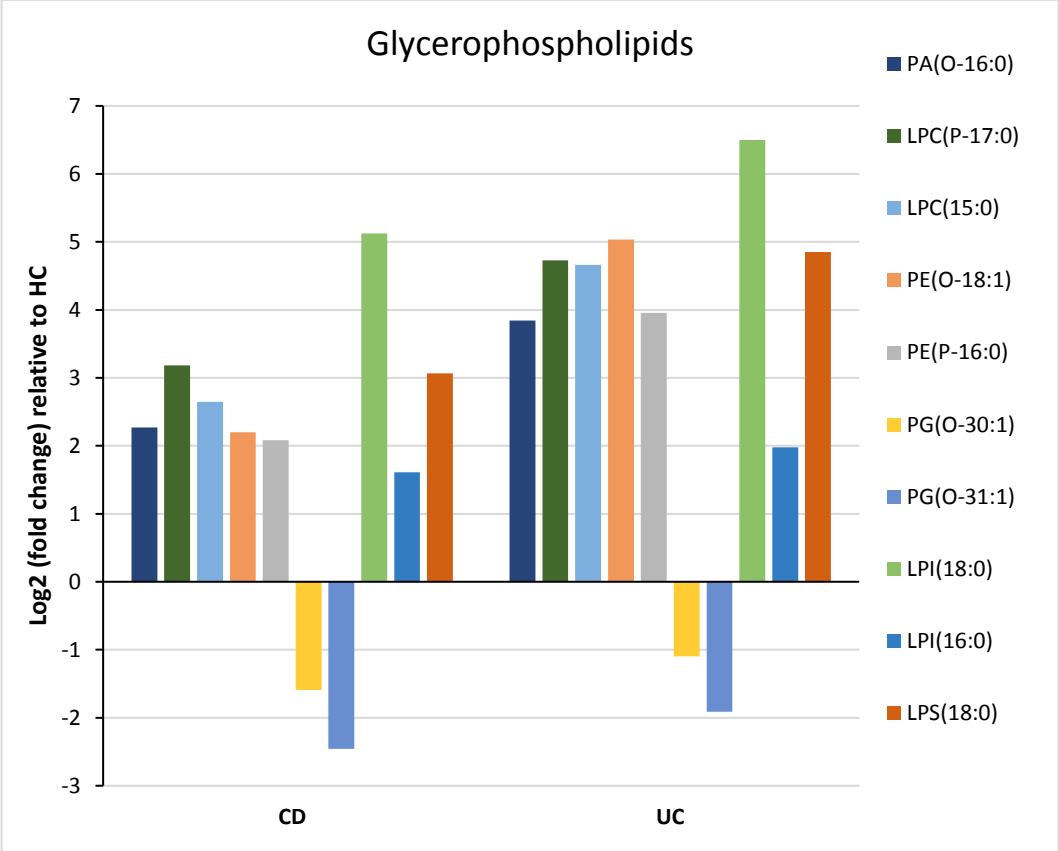


Figure 5.10 Common Glycerophospholipids shared between CD and UC in relative to HCs.

Y-axis represent Log base 2 of metabolite intensity

5.7.3.2 Specific putative metabolites that discriminate CD from HC

Comparing the two groups, there were 34 specific putative metabolites that discriminate Crohn's disease samples from HC samples. The supervised model (OPLS-DA) (Figure 5.11, A) showed an incomplete separation between groups since there were few outlier samples (Figure 5.11, B). The outliers' samples (HC10, HC24 and CD01) were excluded from the model.

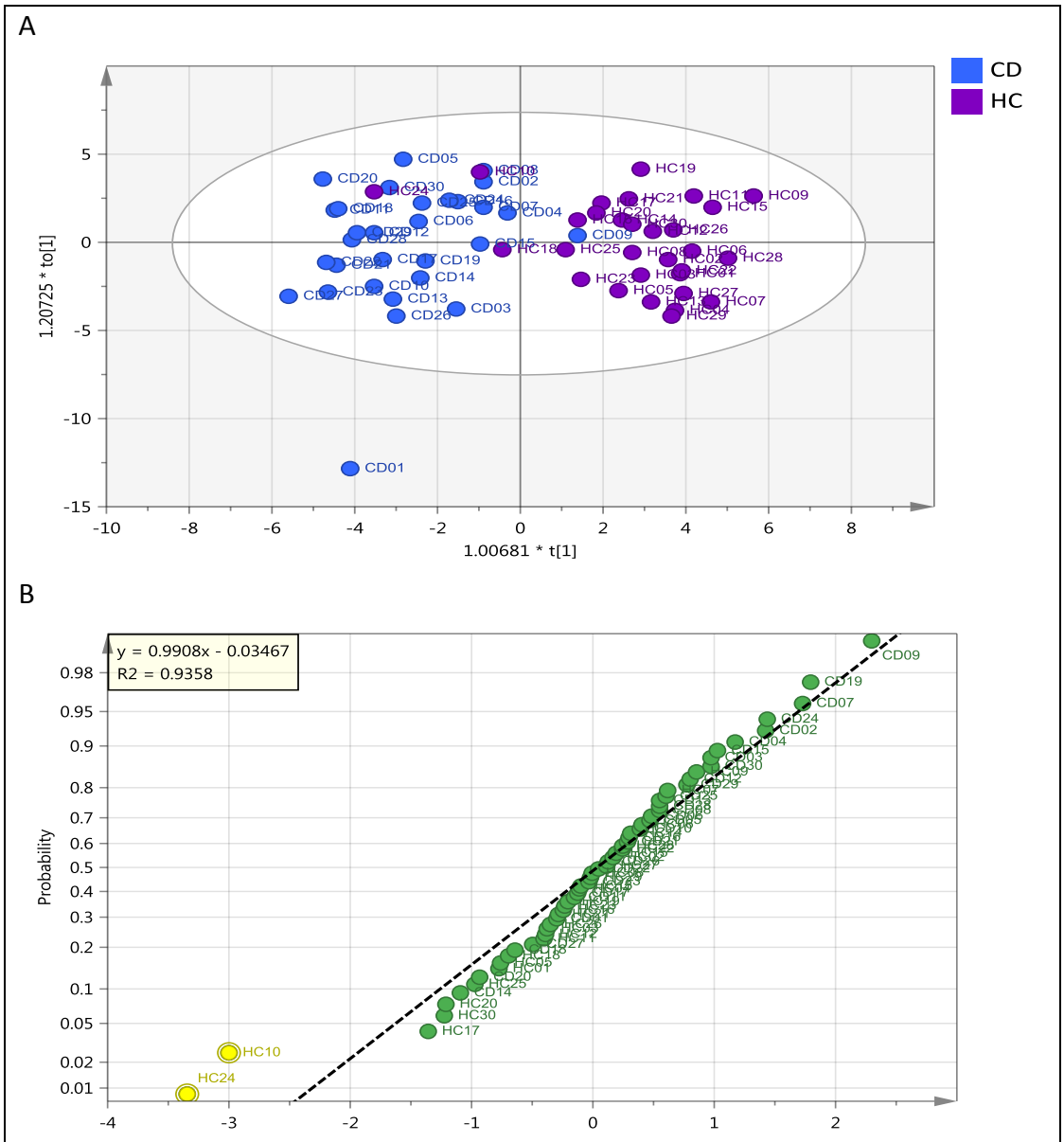


Figure 5.11 OPLS-DA score plots for CD (CD) against HC (HC) samples.

(A)The figure shows distribution of 60 samples based on readings of 34 putative biomarkers. The model consists of one predictive x-score component; component t[1] and two orthogonal x-score components to[1-2]. t[1] explaining 24.8% of the predictive variation in x, to[1-2] explaining 30.3% of the orthogonal variation in x. R^2X (Cum) = 0.511, R^2(Cum)=0.757, Q^2(Cum)=0.617 and p CV-ANOVA = $1.10967e^{-011}$. (B) Normal probability plot of residuals. The plot displaying the residuals standardized on a double Log scale on y-axis vs standard deviation on the x-axis. Outlier variables (yellow) displayed outside -4 to 4 (CD01) standard deviation interval or appeared away from the regression line (HC10 and HC24. The regression line asses the normality. All values were Log base2.

However, after removing the outliers, the model showed clear separation and was able to discriminate diseased from HCs samples (Figure 5.12, A). This model was built based on 57 samples since there were 3 samples (HC10, HC24 and CD01) excluded based on Normal probability plot of the residuals. The *p* CV-ANOVA= $1.21864e^{-010}$ indicates that the separation was significant between the two groups. The model parameter R^2X (Cum)= 0.636 demonstrates that 63% of metabolomics changes were explained by the model while the parameter Q^2 (Cum)= 0.617 demonstrated that 61.7% of the metabolomics variation was predicted by the model (goodness of prediction). Between subject variation was explained by the parameter (R^2=0.895, goodness of fit) means 89.5% of this variation can be explained by the model. The R^2 of the regression line of Normal probability plot of residuals was improved from 0.93 in in (Figure 5.11,B) to 0.99 (Figure 5.12, B) which reflects that the normality of the residuals was improved as well in this model.

The observed versus predicted plot (Figure S5. 1A, Appendix**)** examined the validity of the orthogonal components in the model (R^2= 0.89). In addition to that, permutation test of the model (Figure S5. 1B, Appendix) shows that this model was valid since the permutated Q^2 values to the left are lower than the original Q^2 points to the right. The regression line of the Q^2 points intersects the vertical axis (on the left) at, or below zero. In addition, the all new permutated R^2 values to the left allocated below the original value of R^2 to the right indicates the validity of the original model.

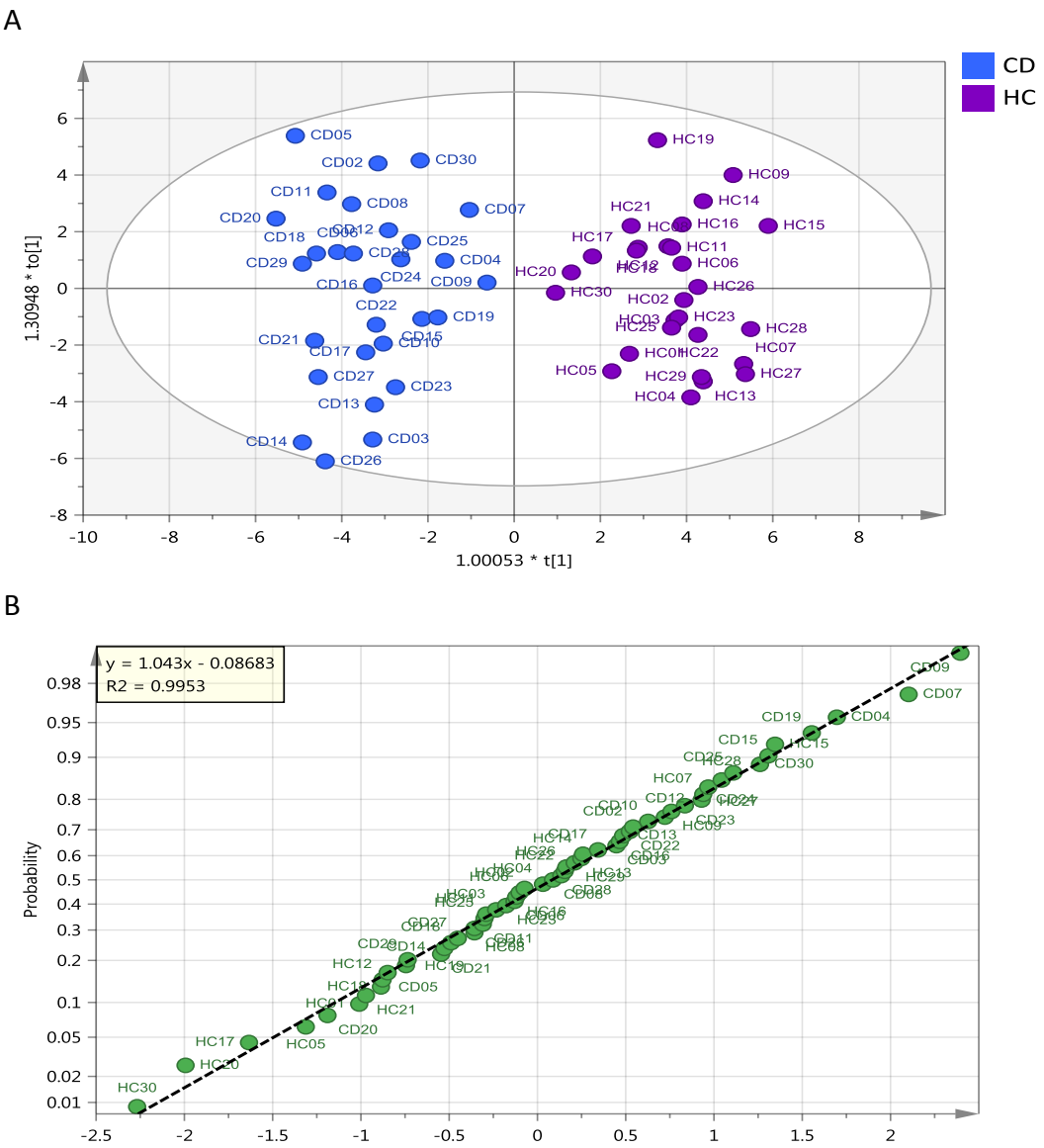


Figure 5.12 OPLS-DA score plots for CD (CD) against HC (HC) samples after outliers' exclusion.

(A) The figure shows distribution of 57 samples based on readings of 34 putative biomarkers. The model consists of one predictive x-score component; component t[1] and four orthogonal x-score components to[1-4]. t[1] explaining 28.3% of the predictive variation in x, to[1-4] explaining 35.3% of the orthogonal variation in x. $R^2X(Cum)= 0.636$, $R^2(Cum) =0.896$, $Q^2(Cum)=0.755$ and p CV-ANOVA = $1.21864e^{-010}$. (B) Normal probability plot of residuals. The plot displaying the residuals standardized on a double Log scale on y-axis vs standard deviation on the x-axis. Outlier variables displayed outside -4 to 4 standard deviation interval or appeared away from the regression line. The regression line assesses the normality. All values were Log base2. Outliers were excluded from model (HC10, HC24 and CD01).

The 34 metabolites shown in (Table **5.5**) represent the specific metabolites that discriminate CD samples from HC. All included metabolites had FDR < 0.05, AUC > 0.6 and 95% confidence intervals (95% C.I) not containing zero.

Table 5.5 Putative biomarkers and their pathways that discriminate CD from HC samples

ID	Mass	RT	Putative metabolite	Log (CD/ HC)	VIP total	VIP pred/o rth	FDR	AUC
Acidic glycosphingolipids								
206	795.515	3.75	(3'-sulfo)Galbeta-Cer(d18:1/2-OH-16:0)	2.166	1.396	1.376	$8.87E^{-04}$	0.826
Alanine and aspartate metabolism								
209	89.048	15.12	L-Alanine	0.912	0.801	0.656	$1.43E^{-02}$	0.77
Aminosugars metabolism								
217	221.09	13.47	N-Acetyl-D-glycosamine	1.151	0.932	0.881	$5.88E^{-03}$	0.736
219	424.168	16.17	Chitobiose	1.167	0.764	2.781	$8.26E^{-03}$	0.766
Arginine and proline metabolism								
227	131.106	26.13	N-Carbamoylputrescine	1.264	1.214	1.455	$1.76E^{-03}$	0.749
Ascorbate and aldarate metabolism								
243	136.037	11.03	2,3,4-trihydroxy-butanoic acid	1.241	1.134	0.947	$6.47E^{-04}$	0.761
Bile acid biosynthesis__Taurine and hypotaurine metabolism								
277	515.292	4.67	Taurocholate	4.558	1.632	0.713	$4.92E^{-03}$	0.701
Biosynthesis of unsaturated fatty acids								
296	336.303	3.86	13Z,16Z-docosadienoic acid	-0.639	0.939	0.658	$2.03E^{-02}$	0.683
Diacylglycerols								
335	652.507	3.76	DG39:7	-1.438	1.196	0.736	$1.96E^{-02}$	0.746
Eicosanoids								
343	334.214	4.85	Dihydroxy-eicosapentaenoic acid	-2.127	1.465	1.279	$1.81E^{-03}$	0.721
Fatty Acids and Conjugates								
374	332.257	4.35	trihydroxy-octadecanoic acid	-1.438	1.326	1.705	$1.14E^{-02}$	0.739
Glycerophosphoglycerols								
441	732.53	3.68	PG 34:2	-0.870	0.913	0.826	$9.98E^{-03}$	0.717
444	122.048	10.39	Nicotinamide	3.451	1.433	2.026	$2.08E^{-05}$	0.837
Glycine, serine and threonine metabolism								
467	105.043	16.21	L-Serine	1.016	0.728	0.855	$1.47E^{-02}$	0.737

468	119.058	14.83	L-Threonine	1.483	1.082	0.704	$5.68E^{-03}$	0.747
Lysine biosynthesis								
518	190.095	25.23	Diaminoheptanedioate	1.035	1.086	1.567	$7.82E^{-05}$	0.79
Monoacylglycerols								
534	440.386	3.78	MG 24:1	-2.265	1.437	1.943	$1.49E^{-04}$	0.844
Tryptophan metabolism__Phenylalanine, tyrosine and tryptophan biosynthesis__Benzoxazinone biosynthesis								
611	117.058	10.13	Indole	1.264	1.131	1.517	$4.60E^{-05}$	0.809
Tyrosine metabolism								
614	110.037	7.93	p-Benzenediol	-1.685	0.898	0.875	$5.55E^{-04}$	0.783
Valine, leucine and isoleucine degradation								
617	131.094	10.76	L-Leucine	1.026	0.882	0.777	$5.30E^{-03}$	0.78
618	131.094	11.45	L-Isoleucine	1.866	0.942	1.318	$4.41E^{-03}$	0.752
Miscellaneous								
4	99.069	7.3	N-Methyl pyrrolidinone	3.050	1.414	1.794	$1.09E^{-03}$	0.74
5	99.069	6.31	Piperidinone	2.928	1.427	2.167	$7.30E^{-04}$	0.772
8	101.048	11.53	Aminobutanolide	0.933	0.776	1.035	$9.58E^{-03}$	0.718
34	130.074	15.09	Casein K	0.865	0.799	0.522	$3.67E^{-02}$	0.72
79	189.111	14.72	L-Homocitrulline	-2.102	1.077	1.497	$4.41E^{-04}$	0.804
82	198.064	14.33	Amino-hydroxyoxo-pyridinyl propanoate	-1.723	0.736	0.691	$1.14E^{-02}$	0.71
86	203.094	4.96	Shihunine	-0.687	0.91	0.821	$1.24E^{-02}$	0.713
136	309.106	13.59	O-Acetylneuraminic acid	1.836	1.349	1.031	$2.77E^{-02}$	0.764
149	331.141	11.47	Ambelline	-2.474	1.26	1.012	$2.32E^{-06}$	0.906
162	408.287	7.7	Cholic acid	4.265	1.675	0.608	$2.52E^{-02}$	0.704
167	425.35	4.4	Oleoylcarnitine	3.167	1.519	1.671	$1.49E^{-03}$	0.853
169	427.365	4.37	Stearoylcarnitine	1.604	1.402	1.846	$1.67E^{-02}$	0.816
184	472.249	4.63	Chenodeoxycholic acid sulfate	4.775	1.732	0.544	$2.85E^{-02}$	0.712

Based on the Log ratios (CD/HC), putative metabolites with high and low Log ratios were used to discriminate the two groups and ten putative metabolites clearly separated the two groups (Table **5.6**). Chenodeoxycholic acid sulfate, Cholic acid and Taurocholate showed the highest VIP values (> 1.6) and highest Log_2 ratio (>4.5) appeared at the top of the list which may reflect the importance of the bile acid metabolites to discriminate CD from HC. Similarly, the acylcarnitine metabolites (Oleoylcarnitine, Stearoylcarnitine) were increased in Crohn's samples in comparison

to HC. The only negatively correlated metabolites were dihydroxyeicosapentaenoic acid and MG(0:0/24:1(15Z)/0:0 which decreased in CD.

Table 5.6 Top 10 putative metabolites that discriminate CD from HCs samples based on VIP ranking.

Mass	RT	Putative metabolite	Log_2 (CD/HC)	VIP total	FDR	AUC
472.249	4.63	Chenodeoxycholic acid sulfate	4.775	1.732	$2.85E^{-02}$	0.712
408.287	7.7	Cholic acid	4.265	1.675	$2.52E^{-02}$	0.704
515.292	4.67	Taurocholate	4.558	1.632	$4.92E^{-03}$	0.701
425.35	4.4	Oleoylcarnitine	3.167	1.519	$1.49E^{-03}$	0.853
334.214	4.85	dihydroxyeicosapentaenoic acid	-2.127	1.465	$1.81E^{-03}$	0.721
440.386	3.78	MG(0:0/24:1(15Z)/0:0)	-2.265	1.437	$1.49E^{-04}$	0.844
122.048	10.39	Nicotinamide	3.451	1.433	$2.08E^{-05}$	0.837
99.069	6.31	2-Piperidinone	2.928	1.427	$7.30E^{-04}$	0.772
99.069	7.3	N-Methyl-2-pyrrolidinone	3.05	1.414	$1.09E^{-03}$	0.74
427.365	4.37	Stearoylcarnitine	1.604	1.402	$1.67E^{-02}$	0.816

The final OPLS-DA model in (Figure 5.13) shows the distribution of 56 samples based on the top ten metabolites after VIP ranking (Table 5.6). There were four samples excluded from the model (HC10, HC24, CD09 and CD25). The separation between CD and HC was significant since *p* CV-ANOVA = $9.24586e^{-011}$. The model parameter R^2X (Cum) was improved from 0.63 in the previous model to 0.746 and demonstrates that 74.6% of metabolomics change is explained by the model. Variation between subjects was explained by the parameter (R^2 = 0.779, goodness of fit) and implies 78% of the variability between the subjects was explained by the variability in the metabolites. Moreover, the goodness of prediction (Q^2 (Cum)) increased to 69.7%. This parameter demonstrates that around 70% of the metabolomics variation was predicted by the model. The R^2 of the regression line of normal probability plot of

residuals was 0.96 (Figure 5.13, B) which reflects the normality of the residuals even after the reduction of the metabolites from 34 to 10. The ten including metabolites had excellent classifying ability with area under the ROC curve (AUROCC=1) of Crohn's samples compared to HC (Figure S5. 2A, Appendix). The model validated by permutation test (Figure S5. 2B, Appendix) shows that this model was valid since the permutated Q^2 values to the left are lower than the original Q^2 points on the right. The regression line of the Q^2 points intersects the vertical axis (on the left) at, or below zero. In addition, the all new permutated R^2 values to the left were below the original value of R^2 to the right which indicates the validity of the original model. The observed versus predicted plot (Figure S5. 2C, Appendix) examined the validity of the orthogonal components in the model ($R^2 = 0.78$).

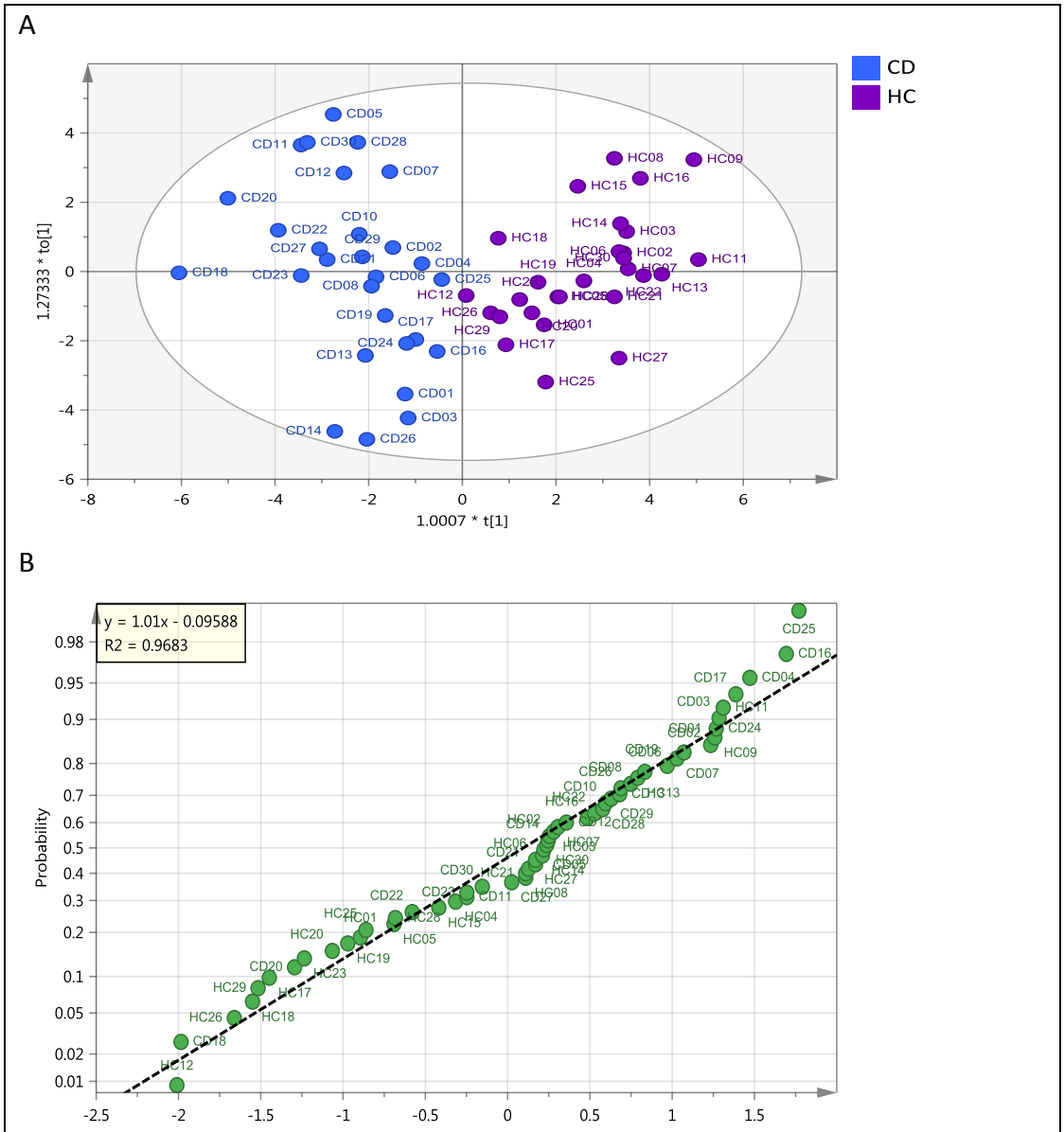


Figure 5.13 OPLS-DA score plots for CD (CD) against HC (HC) samples based on VIP ranking.

(A) The figure shows distribution of 56 samples based on readings of 10 putative biomarkers. The model consists of one predictive x-score component; component t[1] and two orthogonal x-score components to[1-2]. t[1] explaining 37% of the predictive variation in x, to[1-2] explaining 37.6% of the orthogonal variation in x. $R^2X(Cum)$= 0.746, $R^2(Cum)$ =0.779, $Q^2(Cum)$=0.697 and p CV-ANOVA = 7.97791e-013. (B) Normal probability plot of residuals. The plot displaying the residuals standardized on a double Log scale on y-axis vs standard deviation on the x-axis. Outlier variables displayed outside -4 to 4 standard deviation interval or appeared away from the regression line. The regression line assesses the normality. All values were Log base2. Five outliers were excluded from model (HC10, HC24, CD09 and CD25).

5.7.3.3 Specific putative metabolites that discriminate UC from HC

The comparison of UC with HC samples suggests that there were 47 specific putative metabolites that could discriminate UC from HC. As shown in the supervised model (OPLS-DA), (Figure 5.14), there was no outlier and the model was built based on 12 UC and 30 HC samples. The discrimination between these groups was significant with *p* CV-ANOVA = $1.0839e^{-010}$. Clearly, the UC samples were widely distributed in the left side of the ellipse and indicating between subject variability.

From OPLS-DA model parameters, R^2X (Cum) = 0.56, shows that the model explained 56% of metabolic change. Between subject variation was explained by the parameter (R^2=0.91, goodness of fit) and implies 91% of the variability between the subjects was explained by the variability in the metabolites. The model parameter Q (cum) = 0.815 (the goodness of prediction) demonstrates that around 81.5% of the metabolomics variation was predicted by the model. The R^2 of the regression line of normal probability plot of residuals was 0.98 (Figure 5.14, B) and all points lying in a straight line between -4 and +4 shows the normality of the residuals. The validity of the orthogonal component of the model was examined using observed versus predicted plots (Figure S5. 3A, Appendix) and showed R^2 = 0.91. The model was also validated by a permutation test (Figure S5. 3B, Appendix) and this showed that this model was valid since the permutated Q^2 values to the left are lower than the original Q^2 points to the right. The regression line of the Q^2 points intersects the vertical axis (on the left) at, or below zero. In addition, the all new permutated R^2 values to the left were

below the original values of R^2 to the right and indicates the validity of the original model.

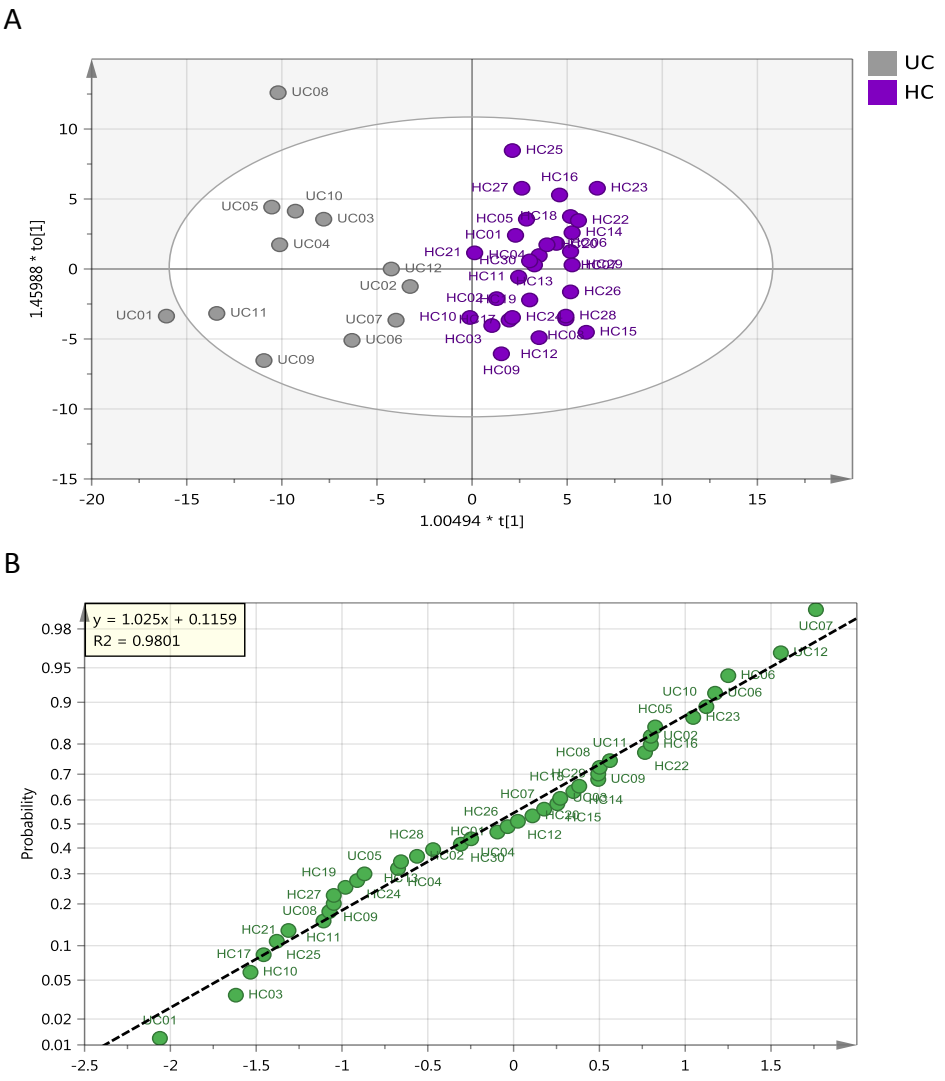


Figure 5.14 OPLS-DA score plots for UC (UC) against HC (HC) samples.

(A) The figure shows distribution 42 samples based on readings of 68 putative biomarkers. The model consists of one predictive x-score component; component t[1] and two orthogonal x-score components to[1-2]. t[1] explaining 22.8% of the predictive variation in x, to[1-2] explaining 34.8% of the orthogonal variation in x. $R^2X(Cum)$= 0.56, $R^2(Cum)$ =0.91, $Q^2(Cum)$=0.815 and p CV-ANOVA = $1.0839e^{-010}$. (B) Normal probability plot of residuals. The plot displaying the residuals standardized on a double Log scale on y-axis vs standard deviation on the x-axis. Outlier variables displayed outside -4 to 4 standard deviation interval or appeared away from the regression line. The regression line assesses the normality. All values were Log base2. No outliers were excluded.

The 47 metabolites shown in (Table **5.7**) represent the specific metabolites that discriminate CD samples from HC. All of the included metabolites had FDR < 0.05, AUC > 0.6 and 95% confidence intervals (95% C.I) not containing zero.

Table 5.7 Putative biomarkers and their pathways that discriminate UC from HC samples

ID	Mass	RT	Putative metabolite	Log (UC/HC)	VIP total	VIP pred/ orth	FDR	AUC
Alpha Linolenic Acid and Linoleic Acid Metabolism								
212	358.287	4	Tetracosapentaenoic acid (24:5n-6)	-2.786	1.492	1.98	$4.53E^{-05}$	0.9
213	210.125	4.69	(-)-Jasmonic acid	-1.276	1.119	0.99	$4.08E^{-05}$	0.897
216	292.204	4.09	12-OPDA	-0.980	0.873	0.655	$7.49E^{-03}$	0.761
Arginine and proline metabolism								
241	103.063	15.65	4-Aminobutanoate	-1.556	0.833	1.421	$8.44E^{-03}$	0.758
beta-Alanine metabolism								
263	217.142	16.51	beta-Alanyl-L-lysine	2.484	1.219	0.68	$1.31E^{-03}$	0.718
Bile acid biosynthesis								
269	450.335	4.5	Trihydroxy-cholestanoate	-1.932	1.408	1.054	$3.69E^{-05}$	0.891
274	436.355	4.47	Cholestantetrol	-1.561	1.395	1.213	$1.74E^{-06}$	0.918
280	392.293	4.55	Dihydroxycholanoic Acid	-1.515	1.346	0.657	$1.20E^{-02}$	0.755
285	466.329	4.49	Tetrahydroxycholestanoic acid	-2.591	1.615	1.597	$1.02E^{-05}$	0.948
Ceramide phosphoinositols								
306	797.541	3.97	N-(hexadecanoyl)-4R-hydroxysphinganine-1-phospho-(1'-myo-inositol)	-1.056	1.031	1.295	$3.54E^{-03}$	0.812
Daiacylglycerols								
321	522.428	3.79	DG29:2	-2.023	1.175	1.102	$4.61E^{-02}$	0.785
327	576.476	3.76	DG33:3	-2.272	1.618	0.987	$9.06E^{-03}$	0.806
328	578.492	3.74	DG33:2	-2.315	1.587	1.013	$2.47E^{-04}$	0.818
329	580.508	3.73	DG33:1	-1.333	1.461	0.819	$3.14E^{-04}$	0.794
330	582.523	3.72	DG33:0	-0.326	1.13	0.853	$3.74E^{-03}$	0.809
331	602.492	3.77	DG35:4	-2.171	1.304	1.743	$2.13E^{-02}$	0.9
Eicosanoids								
346	352.225	4.94	9S,11R-epidioxy-15S-hydroxy-5Z,13E-prostadienoic acid	-1.515	1.283	1.555	$1.30E^{-04}$	0.927
Fatty Acids and Conjugates								
355	202.12	4.18	Decanedioic acid	-0.924	1.071	0.867	$2.10E^{-04}$	0.788

362	230.152	4.48	Dodecanedioic acid	-0.462	0.927	0.723	$4.01E^{-03}$	0.709
371	298.288	3.89	Nonadecanoicacid	-0.687	0.931	1.059	$1.36E^{-03}$	0.748
Glycerolipid metabolism								
411	410.243	4.3	LPA16:0	3.442	1.449	0.898	$2.27E^{-03}$	0.842
412	436.259	4.29	LPA18:1	3.772	1.188	0.756	$3.64E^{-03}$	0.791
Glycerophosphates								
415	410.243	7.69	LGP 16:0	5.215	1.805	1.19	$8.63E^{-05}$	0.912
Glycerophosphocholines								
421	451.306	4.58	LPC14:1	3.056	0.972	0.699	$8.90E^{-04}$	0.836
422	465.322	4.41	LPE 18:1	4.692	1.938	1.815	$1.56E^{-04}$	0.948
Glycine, serine and threonine metabolism								
456	103.063	13.68	N,N-Dimethylglycine	1.299	0.931	1.407	$2.47E^{-04}$	0.758
Indole and ipecac alkaloid biosynthesis								
507	376.137	8.31	Riboflavin	-0.727	0.74	1.985	$5.36E^{-03}$	0.758
Linoleic acid metabolism								
515	312.23	4.08	Dihydroxy octadecadienoic acid	-0.664	0.857	0.703	$3.57E^{-03}$	0.724
516	280.24	3.95	Linoleate	-1.214	1.026	0.705	$5.56E^{-03}$	0.758
Lysine biosynthesis								
520	203.079	10.22	N2-Acetyl-L-aminoadipate	-3.224	1.314	0.685	$1.96E^{-03}$	0.839
Lysine degradation								
523	159.089	13.18	5-Acetamidopentanoate	-1.565	1.178	2.224	$8.67E^{-05}$	0.861
Porphyrin and chlorophyll metabolism								
554	592.327	4.57	I-Urobilinogen	-1.781	1.212	0.95	$8.43E^{-05}$	0.845
Pyrimidine metabolism								
565	244.07	10.05	Uridine	1.622	0.805	1.797	$2.67E^{-02}$	0.755
Secosteroids								
573	454.345	4.14	cholestapentaene-triol	-3.837	1.184	1.741	$5.22E^{-03}$	0.797
Steroid conjugates								
582	482.307	3.86	hydroxycholesterol sulfate	-1.211	1.006	0.896	$5.29E^{-03}$	0.752
Triacylglycerols								
605	740.594	3.71	DG 44:5	-1.938	1.275	1.309	$5.81E^{-05}$	0.939
miscellaneous								
38	132.069	7.22	Indoleamine	0.735	0.857	1.238	$5.56E^{-03}$	0.782
41	132.078	4.82	Ethyl hydroxybutanoate	1.621	0.985	2.022	$3.05E^{-04}$	0.797
105	230.152	7.7	Diisopropyl adipate	-0.921	1.07	1.058	$1.36E^{-04}$	0.821
115	243.183	4.52	N-Undecanoylglycine	1.826	1.11	5.16	$8.30E^{-03}$	0.924
128	286.251	4.01	Hydroxy-palmitic acid methyl ester	-0.706	1.129	1.159	$3.59E^{-02}$	0.797
137	314.246	4.12	Dihydroxyoctadecenoic acid	-0.927	1.149	0.812	$1.65E^{-04}$	0.833

141	318.121	4.77	diaminomethyl(nitrophenoxypropyloxypyrimidine	-0.924	0.904	1.507	$3.27E^{-03}$	0.812
154	378.277	4.74	2-Arachidonoylglycerol	-1.297	1.077	0.994	$6.44E^{-04}$	0.845
178	452.35	4.81	5b-Cholestane-3a-7a-12a-23R-25-pentol	-1.434	1.222	1.047	$1.31E^{-04}$	0.867
180	454.329	4.12	27-Norcholestanehexol	-1.415	1.181	0.801	$4.52E^{-02}$	0.786
202	466.311	3.79	Cholesterolsulfate	-1.446	1.361	1.379	$1.74E^{-06}$	0.945

Based on VIP top 10 metabolites from (Table 5.8), as shown in the supervised model (OPLS-DA), (Figure 5.15), there was significant clear separation between UC and HC with *p* CV-ANOVA = $3.85675e^{-011}$. Because of the low number of samples in UC group, exclusion of extra samples was avoided to preserve sample number. However, the remaining samples were still within the range of the normal probability plot of Residuals +/-4. The only sample excluded from the model was UC02.

Table 5.8 Top 10 putative metabolites that discriminate UC from HC samples based on VIP ranking.

Mass	RT	Putative metabolite	Log (UC/HC)	VIP total	FDR	AUC
465.322	4.41	LPE 18:1	4.692	1.938	$1.56E^{-04}$	0.948
410.243	7.69	LPA 16:0	5.215	1.805	$8.63E^{-05}$	0.912
576.476	3.76	DG33:3	-2.272	1.618	$9.06E^{-03}$	0.806
466.329	4.49	Tetrahydroxycholestanoic acid	-2.591	1.615	$1.02E^{-05}$	0.948
578.492	3.74	DG 33:2	-2.315	1.587	$2.47E^{-04}$	0.818
358.287	4	Tetracosapentaenoic acid	-2.786	1.492	$4.53E^{-05}$	0.9
580.508	3.73	DG33:1	-1.333	1.461	$3.14E^{-04}$	0.794
410.243	4.3	LPA 16:0	3.442	1.449	$2.27E^{-03}$	0.842
450.335	4.5	Trihydroxcholestanoate	-1.932	1.408	$3.69E^{-05}$	0.891
436.355	4.47	Cholestane-tetrol	-1.561	1.395	$1.74E^{-06}$	0.918

The model parameter R^2X (Cum) was improved from 0.56 in the previous model to 0.794 demonstrating that 79.4 % of metabolomics change was explained by the final

model. Between subject variation was explained by the parameter (R^2 = 0.791, goodness of fit) means 79% of the variability between the subjects was explained by the variability in the metabolites. Moreover, the goodness of prediction (Q^2 (Cum)) decreased to 77.5%. This parameter demonstrates that around 77% of the metabolomics variation was predicted by the model. The R^2 of the regression line of normal probability plot of residuals was 0.81, (Figure 5.15, B), which reflects the normality of the residuals after the reduction of the metabolites from 70 to 10. The ten included metabolites have excellent classifying ability (AUROCC = 1) of UC samples compared to HC (Figure S5. 4A, Appendix). The model validated by a permutation test (Figure S5. 4B, Appendix) shows that this model was valid since the permutated Q^2 values to the left are lower than the original Q^2 points to the right. The regression line of the Q^2 points intersects the vertical axis (on the left) at, or below zero. In addition, the all new permutated R^2 values to the left were below the original value of R^2 to the right and indicates the validity of the original model. The observed versus predicted plot (Figure S5. 4C, Appendix) examined the validity of the orthogonal components in the model (R^2 = 0.79).

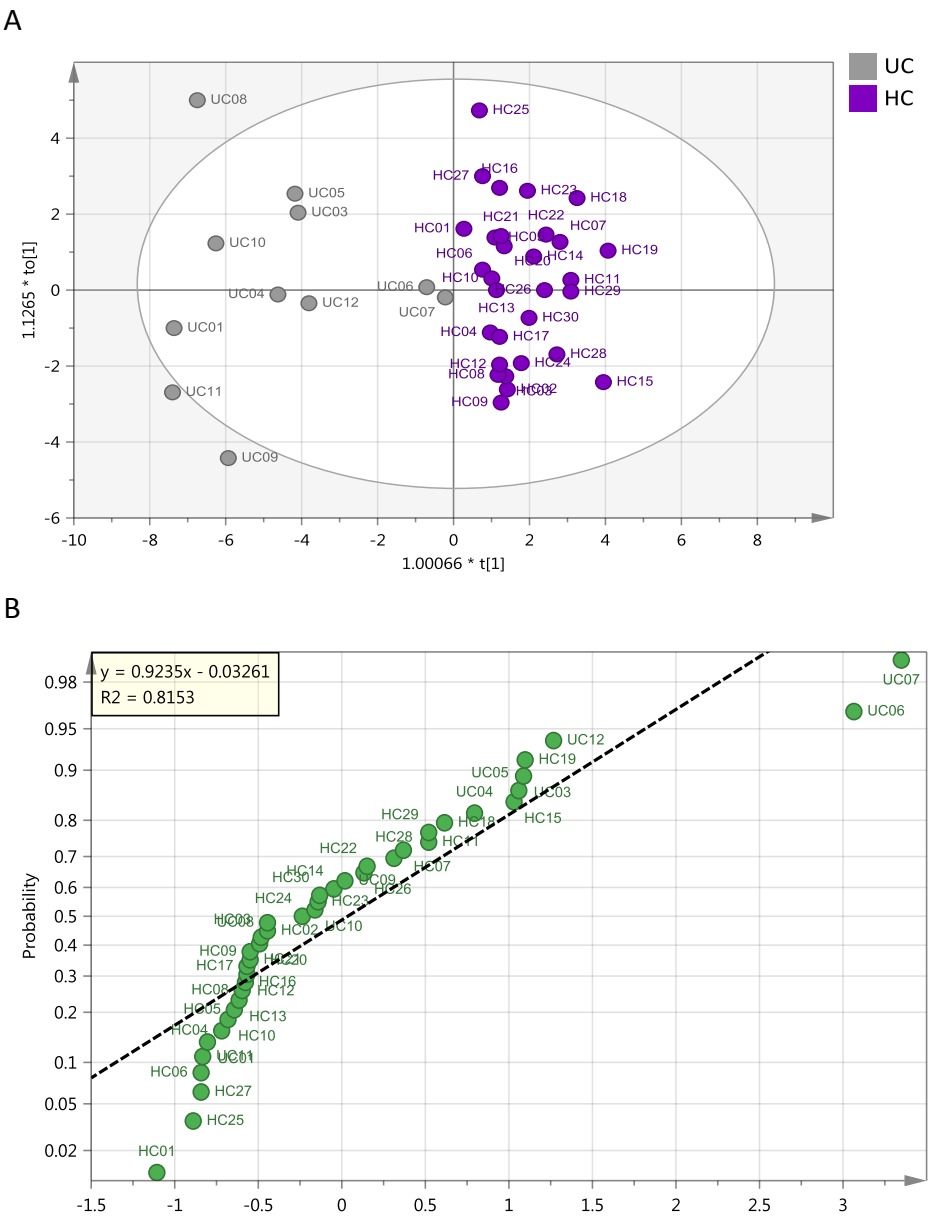


Figure 5.15 OPLS-DA score plots for UC (UC) against HC (HC) samples based on VIP ranking.

(A) The figure shows distribution 41 samples based on readings of 10 putative biomarkers. The model consists of one predictive x-score component; component t[1] and one orthogonal x-score components to[1]. t[1] explaining 54.8% of the predictive variation in x, to[1] explaining 24.9% of the orthogonal variation in x. $R^2X(Cum)= 0.794$, $R^2(Cum) =0.791$, $Q^2(Cum)=0.775$ and p CV-ANOVA = $3.85675e^{-011}$. (B) Normal probability plot of residuals. The plot displaying the residuals standardized on a double Log scale on y-axis vs standard deviation on the x-axis. Outlier variables displayed outside -4 to 4 standard deviation interval or appeared away from the regression line. The regression line assesses the normality. All values were Log base2. One sample (UC02) was excluded as an outlier.

5.7.3.4 Comparison between CD and UC.

The last comparison in this study was to differentiate between the two disease groups i.e. CD versus UC. After the filtration criteria of the metabolites, the number of significant metabolites that discriminated the groups were 29, (Table 5.9). The final OPLS-DA model was built based on 41 samples since sample UC26 was excluded from the model (Figure 5.16, A).

Table 5.9 Putative biomarkers and their pathways that discriminate CD from UC samples

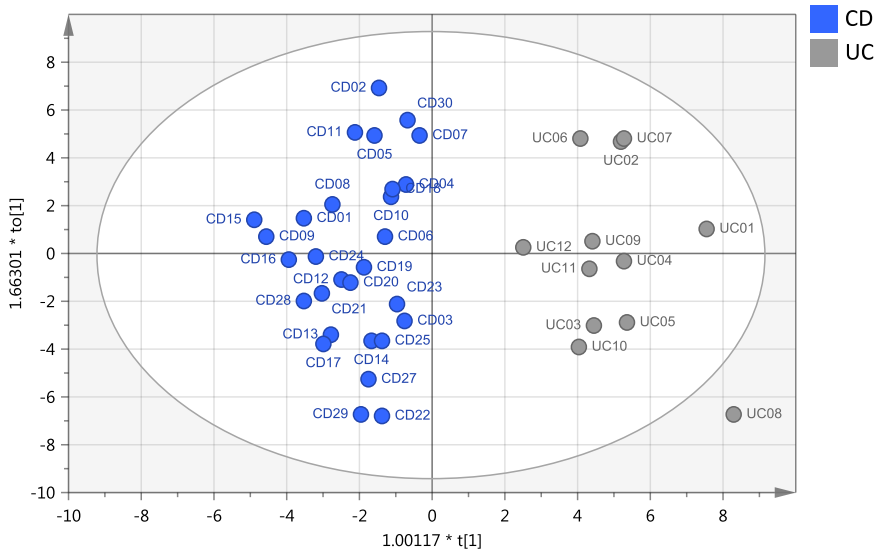
ID	Mass	RT	Putative metabolite	Log (CD/UC)	VIP total	VIP pred/ orth	FDR	AUC
Acidic glycosphingolipids								
206	795.515	3.75	(3'-sulfo) Galbeta-Cer (d18:1/2-OH-16:0)	1.224	1.31	1.82	$4.63E^{-02}$	0.74
Arginine and proline metabolism								
227	131.106	26.13	N-Carbamoylputrescine	1.407	1.375	2.298	$4.19E^{-02}$	0.74
beta-Alanine metabolism								
263	217.142	16.51	beta-Alanyl-L-lysine	-1.737	1.528	1.196	$4.63E^{-02}$	0.72
Bile acid biosynthesis								
267	434.34	4.22	Trihydroxycholesterol	1.471	1.489	1.356	$3.03E^{-02}$	0.76
269	450.335	4.5	Trihydroxy-cholestanoate	2.061	1.609	1.424	$2.98E^{-02}$	0.76
274	436.355	4.47	Cholestane-tetrol	1.297	1.732	1.575	$2.46E^{-02}$	0.78
285	466.329	4.49	Tetrahydroxy-cholestanoic acid	2.986	2.023	1.601	$2.98E^{-02}$	0.81
Diacylglycerols								
325	572.444	3.8	DG(13:0/20:5(5Z,8Z,11Z,14Z,17Z)/0:0)	1.397	1.612	1.392	$3.50E^{-02}$	0.73
326	574.46	3.78	DG(15:0/18:4(6Z,9Z,12Z,15Z)/0:0)	1.295	1.621	1.557	$3.03E^{-02}$	0.72
327	576.476	3.76	DG(15:0/18:3(6Z,9Z,12Z)/0:0)	1.963	1.927	1.952	$3.50E^{-02}$	0.76
328	578.492	3.74	DG(15:0/18:2(9Z,12Z)/0:0)	1.715	1.802	1.516	$3.10E^{-02}$	0.74
Fatty alcohols								
394	240.245	7.68	11E-Hexadecen-1-ol	-2.071	1.132	1.713	$2.46E^{-02}$	0.77
Fatty esters								

402	222.162	3.77	7E,9E,11-Dodecatrienyl acetate	0.839	1.295	2.048	$2.55E^{-02}$	0.74
404	560.481	3.7	Mayolene-18	1.463	1.607	1.482	$3.31E^{-02}$	0.73
Glycerophosphocholines								
422	465.322	4.41	1-(1Z-pentadecenyl)-sn-glycero-3-phosphocholine	-1.801	1.673	1.63	$4.08E^{-02}$	0.79
424	493.353	4.32	1-(1Z-heptadecenyl)-sn-glycero-3-phosphocholine	-1.535	1.406	1.719	$3.34E^{-02}$	0.79
Glycerophosphoethanolamines								
432	465.321	7.71	LPE 18:1	-2.806	2.156	2.016	$1.62E^{-02}$	0.84
Glycerophosphoglycerols								
436	498.297	3.85	lPG17:0	-0.521	0.879	1.816	$4.08E^{-02}$	0.73
Lysine biosynthesis								
518	190.095	25.23	LL-2,6-Diaminoheptanedioate	0.477	0.985	1.604	$4.08E^{-02}$	0.7
520	203.079	10.22	N2-Acetyl-L-aminoadipate	2.316	1.256	2.293	$2.98E^{-02}$	0.73
Steroid conjugates								
582	482.307	3.86	26-hydroxycholesterol 3-sulfate	1.297	1.332	1.555	$4.59E^{-02}$	0.71
Sterols								
595	476.35	3.99	11-acetoxy-3beta,6alpha-dihydroxy-9,11-seco-5alpha-cholest-7-en-9-one.	2.048	1.661	1.417	$3.64E^{-02}$	0.76
Miscellaneous								
38	132.069	7.22	Indoleamine	-0.542	1.052	1.685	$3.48E^{-02}$	0.76
41	132.078	4.82	Ethyl (R)-3-hydroxybutanoate	-0.986	1.145	2.3	$3.34E^{-02}$	0.78
115	243.183	4.52	N-Undecanoylglycine	-2.012	1.995	6.179	$1.49E^{-02}$	0.94
135	300.267	4.01	(R)-2-Hydroxystearate	1.258	1.586	1.846	$2.98E^{-02}$	0.78
175	438.334	4.94	Norcholestane-pentol	1.655	1.466	1.603	$2.98E^{-02}$	0.76
178	452.35	4.81	Cholestane-pentol	1.356	1.393	1.386	$3.48E^{-02}$	0.74
202	466.311	3.79	Cholesterol sulphate	1.324	1.448	1.543	$3.03E^{-02}$	0.74

The model parameter R^2X (Cum) was improved from 0.56 in the previous model to 0.624 and demonstrates that around 62% of metabolomics change was explained by the model. Between subject variation was explained by the parameter (R^2 = 0.867, goodness of fit) means around 86% of the variability between the subjects was

explained by the variability in the metabolites. The goodness of prediction (Q^2 (Cum) = 0.699) demonstrates that around 70% of the metabolomics variation was predicted by the model. The R^2 of the regression line of normal probability plot of residuals was 0.98, Figure 5.16B, which reflects the normality of the residuals. The model was validated by a permutation test (Figure S5. 5A, Appendix) shows that this model was valid since the permutated Q^2 values to the left are lower than the original Q^2 points on the right. The regression line of the Q^2 points intersects the vertical axis (on the left) at, or below zero. In addition, the all new permutated R^2 values to the left were below the original value of R^2 to the right which indicates the validity of the original model. The observed versus predicted plot (Figure S5. 5B, Appendix) examined the validity of the orthogonal components in the model (R^2 = 0.79).

A



B

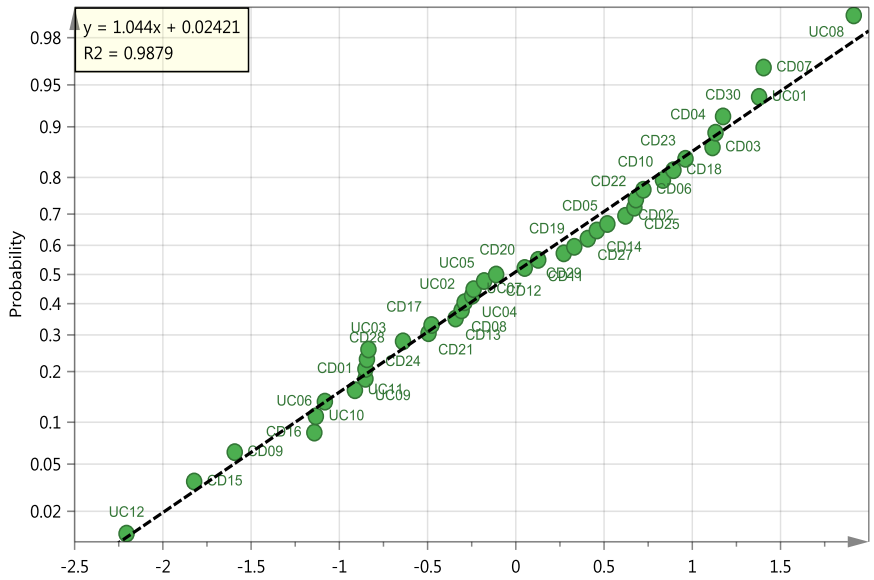


Figure 5.16 OPLS-DA score plots for CD (CD) against UC (UC) samples.

(A) The figure shows distribution of 41 samples based on readings of 35 putative biomarkers. The model consists of one predictive x-score component; component t[1] and three orthogonal x-score components to[1-3]. t[1] explaining 23.3% of the predictive variation in x, to[1-3] explaining39.6% of the orthogonal variation in x. R^2X = 0.629, R^2= 0.867, Q^2=0.699 and p CV-ANOVA = 1.7659e-006. (B) Normal probability plot of residuals. The plot displaying the residuals standardized on a double Log scale on y-axis vs standard deviation on the x-axis. Outlier variables displayed outside -4 to 4 standard deviation interval or appeared away from the regression line. The regression line assesses the normality. All values were Log base2. CD26 was excluded from model as outlier

Based on VIP, the top 10 metabolites from (Table 5.10), were used in the OPLS-DA model shown in (Figure 5.17), and there was a clear separation between CD and UC with *p* CV-ANOVA = 5.30541e-007. The only sample excluded from the model was CD26. All metabolites had VIP total more than 1.5 and LPE 18:1 had the maximum value of 2.156. One metabolite (N-Undecanoylglycine) appeared as an excellent classifier AUC = 0.94. However, seven metabolites could also be fair classifiers (AUC = 0.7-0.8).

Table 5.10 Top 10 putative metabolites that discriminate CD from UC samples based on VIP ranking.

Mass	RT	Putative metabolite	Log (CD/UC)	VIP total	FDR	AUC
465.321	7.71	LPE 18:1	-2.806	2.156	$1.62E^{-02}$	0.84
466.329	4.49	Tetrahydroxycholestanoic acid	2.986	2.023	$2.98E^{-02}$	0.81
243.183	4.52	N-Undecanoylglycine	-2.012	1.995	$1.49E^{-02}$	0.94
576.476	3.76	DG33:3	1.963	1.927	$3.50E^{-02}$	0.76
578.492	3.74	DG33:2	1.715	1.802	$3.10E^{-02}$	0.74
436.355	4.47	Cholestane-tetrol	1.297	1.732	$2.46E^{-02}$	0.78
465.322	4.41	LPE 18:1	-1.801	1.673	$4.08E^{-02}$	0.79
574.46	3.78	DG33:4	1.295	1.621	$3.03E^{-02}$	0.72
572.444	3.8	DG33:5	1.397	1.612	$3.50E^{-02}$	0.73
450.335	4.5	Trihydroxycholestanoate	2.061	1.609	$2.98E^{-02}$	0.76

In comparison between the final model in (Figure 5.17) and the initial model in (Figure 5.16), the model parameter R^2X (Cum) was improved from 0.62 in the previous model to 0.85 and demonstrates that 85 % of metabolomics change was explained by the final model. Between subject variation was explained by the parameter (R^2 = 0.79, goodness of fit) and implies 79% of the variability between the

subjects was explained by the variability in the metabolites. Moreover, the goodness of prediction (Q^2 (Cum)) increased by to 73% and demonstrates that around 73% of the metabolomics variation was predicted by the model. The R^2 of the regression line of normal probability plot of residuals was 0.94, (Figure 5.17, B), which reflect the normality of the residuals even after the reduction of the metabolites from 29 to 10.

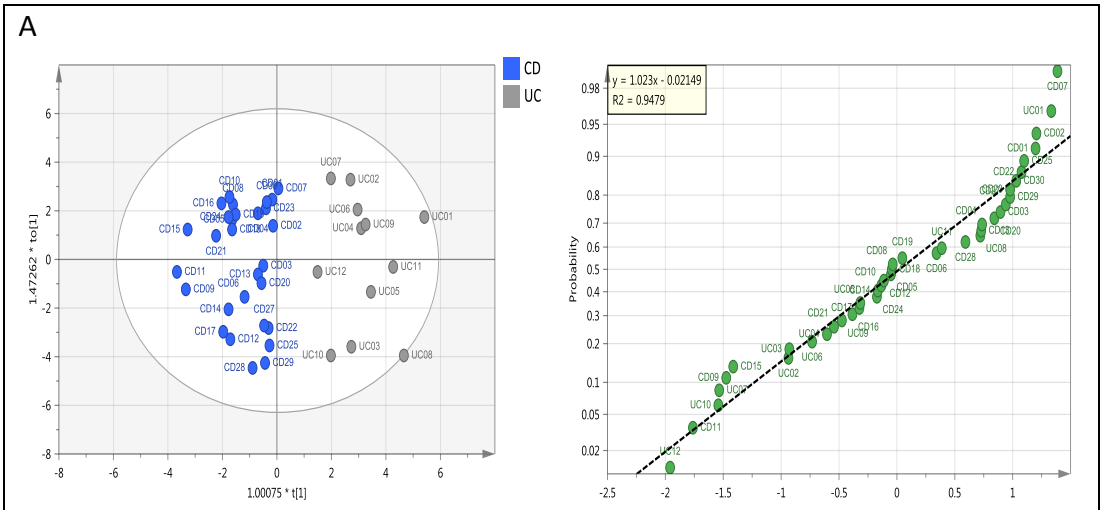


Figure 5.17 OPLS-DA score plots for CD (CD) against UC (UC) samples based on VIP ranking.

(A) The figure shows distribution of 41 samples based on readings of 10 putative biomarkers. The model consists of one predictive x-score component; component t[1] and three orthogonal x-score components to[1-3]. t[1] explaining 28.6% of the predictive variation in x, to[1-3] explaining 57.2% of the orthogonal variation in x. R^2X = 0.856, R^2= 0.795, Q^2=0.737 and p CV-ANOVA = 5.30541e-007. (B) Normal probability plot of residuals. The plot displaying the residuals standardized on a double Log scale on y-axis vs standard deviation on the x-axis. Outlier variables displayed outside -4 to 4 standard deviation interval or appeared away from the regression line. The regression line assesses the normality. All values were Log base2. CD26 was excluded from model as outlier

The included ten metabolites had excellent classifying ability (AUROCC = 1) of UC samples compared to HC (Figure S5. 6A, Appendix). The model was validated by a permutation test (Figure S5. 6B, Appendix) and it shows that this model was valid

since the permutated Q^2 values to the left are lower than the original Q^2 points on the right. The regression line of the Q^2 points intersects the vertical axis (on the left) at, or below zero. In addition, the all new permutated R^2 values to the left were below the original value of R^2 to the right which indicates the validity of the original model. The observed versus predicted plot (Figure S5. 6C, Appendix) examined the validity of the orthogonal components in the model (R^2= 0.795).

The sample differences based on the top ten putative metabolites are shown in the heat map (Figure 5.18). A clear reduction in diacylglycerols and cholesterol metabolites was observed in the UC samples in comparison to the CD samples. Also there was an increase in the other metabolites such as LPE 18:1 and undecanoyl glycine in comparison to CD samples.

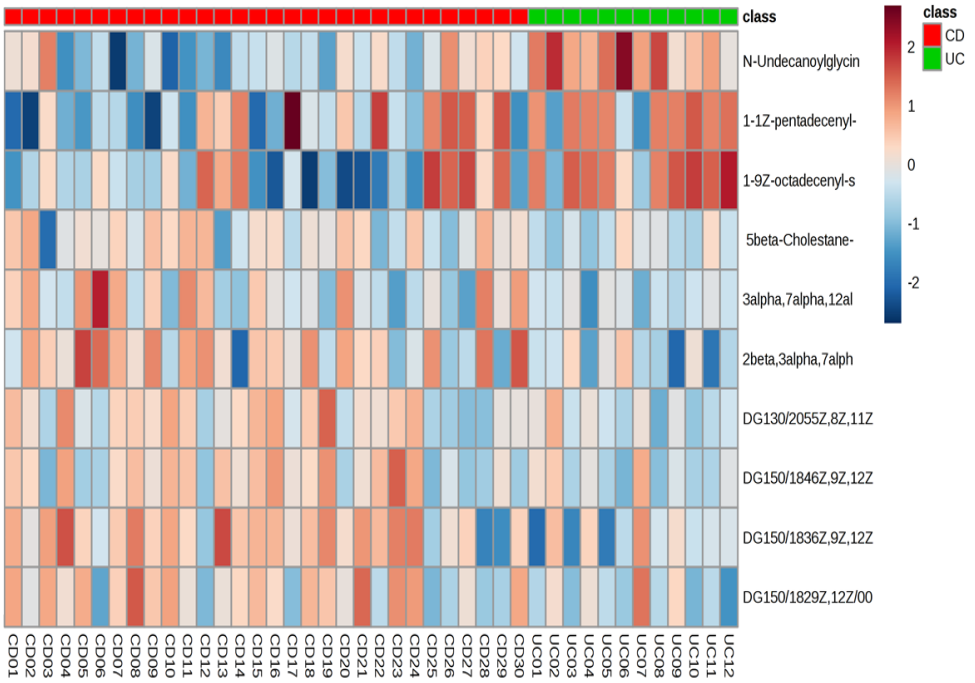


Figure 5.18 Heat map showing the top 10 putative metabolites based on VIP ranking that are significantly discriminate CD from UC.

The top row of the plot represents sample groups; CD (red) and UC (green). The different colour shades represent Log base 2 of intensities of each metabolite (rows) in each observation (column). Figure generated by Metaboanalyst.

5.8 Discussion

5.8.1 Common putative metabolites shared between CD and UC

In CD and UC, fatty acids appear to play a significant role in the disease pathology and development. There is a wide range of fatty acids that are essential factors in epithelial cell function that could regulate the inflammatory status (Donnet-Hughes, Schiffrin and Turini, 2001). Genetic factors might have a major effect on fatty acid production and synthesis. In a genetic study using real time polymerase reaction (RT-PCR), there was significant decrease in fatty acid synthase expression in ileum and

colon of UC patients (Heimerl *et al.*, 2006). The lack of this enzyme may explain the effects on fatty acids and their metabolites in our study.

Regarding to arachidonic acid which is classified as a polyunsaturated Omega-6 fatty acid, it was described as precursor of inflammatory mediators (Schmitz and Ecker, 2008). Polyunsaturated fatty acids have an important roles in inflammation processes and protection since there were high number of lipid modulators derived from these group of fatty acids (Marion-Letellier, Savoye and Ghosh, 2015). The level of arachidonic acid appeared higher in CD in comparison to HC which is consistent with a previous study (Jansson *et al.*, 2009). They analysed faecal extracts provided from 15 twin pairs in comparison with seven healthy twin pairs using Ion Cyclotron Resonance Fourier Transform Mass Spectrometry (ICR-FT/MS). The same trend also appeared in UC disease using colonic mucosa biopsies using gas chromatography–mass spectrometry and liquid chromatography–mass spectrometry (Pearl *et al.*, 2014). In addition to that, they suggested that the severity of inflammation was correlated positively with arachidonic acid level. Arachidonic acid level was found to be associated with its eicosanoid derivatives such as prostaglandins and leukotrienes and they are a key element in intestinal inflammation (Pacheco, Hillier and Smith, 1987).

Similarly, to arachidonic acid, the current study showed that Docosatetraenoic acid (adrenic acid) has positive correlation in both diseases in comparison to HCs samples. This polyunsaturated Omega-6 fatty acid described as one of the metabolites of arachidonic acid metabolites via the elongase enzyme (Wall *et al.*, 2010). This

metabolism process occurs by adding 2 carbon atoms to arachidonic acid and elongating it to docosatetraenoic acid. We found that docosatetraenoic acid was still increased in CD and UC with Log ratio > 2 in relation to HC. In consistency with our results, highly intake of Omega-6 polyunsaturated fatty acids was associated with increased risk of CD and UC (Hou, Abraham and El-Serag, 2011).

The current study found that certain diacylglycerols were reduced in CD and UC in relative to HCs. In addition to that, in UC only, 2-arachidonoylglycerol level appeared with low ratio (Log ratio=-1.2, p=0.002) in comparison with HC. The level of 2-arachidonoylglycerol in CD was slightly decreased (Log ratio = -0.13) with no significant difference (p = 0.131). One of the important functions of diacylglycerols is as substrates of diacylglycerol lipase to generate 2-arachidonoylglycerol (Ambrose and Simmons, 2018). 2-arachidonoylglycerol is described as one of the lipid mediators of the endocannabinoid system which binds and activates the cannabinoid receptor. This hydrolysis is considered as part of the initial step in the pathway to synthesis of arachidonic acid since 2-arachidonoylglycerol can be degraded by monoacylglycerol lipase to produce arachidonic acid (Schicho and Storr, 2014). The expression of both enzymes has been observed to increase in IBD patients (Marquéz *et al.*, 2009; Di Sabatino *et al.*, 2012). This may explain the depletion of diacylglycerols and 2-arachidonoylglycerol in disease samples in contrast with HC and at the same time increase in arachidonic acid levels. In IBD, the endocannabinoid system plays an important function in gut homeostasis and inflammation management (Hasenoehrl *et al.*, 2016).

5.8.2 Specific putative metabolites that discriminate CD from HC

This part of the study focused on the metabolites that were detected with significant difference between CD and HC. The variation in human microbiota between the study participants may have a significant impact on the metabolomic changes and complicates the correlation between specific metabolites and disease. Multivariate analysis can help to select the best correlated metabolites.

In our study, it was observed that bile acid/ cholesterol metabolites may have an important role in CD since we had three of these metabolites are listed in the top ten VIP metabolites (Table 5.6). This result is similar to a previous report; Jansson et al from the analysis faecal extract samples from ten twins with Crohn's disease against seven healthy twins using Ion Cyclotron Resonance Fourier Transform Mass Spectrometry (ICR-FT/MS) (Jansson *et al.*, 2009). It was observed that bile acid metabolites such as Taurocholate and Chenodeoxyglycocholate had high levels in Crohn's twins in comparison to HC. This result may confirm the malabsorption of bile acid metabolites as one of the characteristics of CD. Since the bile acids are formed from cholesterol in the liver (Färkkilä and Miettinen, 1990) and then exported via the bile duct to the gut. A study of serum cholesterol levels can be used to determine the malabsorption of bile acids from the intestines in CD. Serum levels of cholesterol were studied (Bláha *et al.*, 2009) to assess the level of serum and plasma cholesterols in patients with active CD . They found that low levels of total cholesterol, LDL- and HDL-cholesterol were associated with active CD patients (n = 24) in comparison to HC (n = 100) serum samples. The low levels of cholesterol in the blood may be due to the

malabsorption of bile acids from the intestine. Therefore, high levels of bile acids in the intestine may cause diarrhoea in CD (Martínez-Augustin and Medina, 2008).

In the current study, long chain acylcarnitines (Oleoylcarnitine and Stearoylcarnitine) were found to be elevated in CD faecal extracts. Acylcarnitines are described as important biomarkers in mitochondrial disorders of fatty acid beta-oxidation (Costa *et al.*, 2000). They are produced by carnitine- fatty acid conjugation for transport into the mitochondria for beta-oxidation. The relation between CD activity and fatty acid oxidation has been studied using blood samples (Al-Jaouni *et al.*, 2000) and it was observed that there were positive correlations between impaired fatty acid oxidation and disease activity. The increase in acylcarnitines among CD samples compared to HC samples in our study could be considered as a supportive evidence for the impairment of fatty acid oxidation in CD

5.8.3 Specific putative metabolites that discriminate UC from HC

The present comparison was designed to determine the specific features that are able to differentiate UC disease from HC. In this comparison we are found, (based on the top VIP metabolites (Table 5.8), that 1-(1Z-pentadecenyl)-sn-glycero-3-phosphocholine was increased in UC. It is classified as one of the glycerophosphocholines that is derived from choline and stored in cytosol. This increase may be due to the altered membrane choline phospholipids metabolism (MCPM) by inflammation.

Another important finding was that diacylglycerol levels were depleted in UC samples in comparison to HC. This study lists three diacylglycerols as the top VIP metabolites

which can discriminate UC from HC. The levels of these metabolites were negatively correlated in UC. A previous study showed that the expression of both diacylglycerol lipase and monoacylglycerol lipase were positively correlated in colon mucosal biopsies of patients with UC (Marquéz *et al.*, 2009) . In addition, our results also showed that the levels of 2-arachidonoylglycerol and arachidonic acid were elevated in UC samples. This result may support the depletion of diacylglycerols due to the over expression of lipase enzymes.

The other pathway that is affected by the disease is the alpha Linolenic acid and Linoleic acid metabolism. From this pathway we were able to detect three significant metabolites: Tetracosapentaenoic acid (24:5n-6), Jasmonic acid and 12-OPDA. All of these metabolites have low levels in UC in comparison to HC samples. Tetracosapentaenoic acid is an intermediate of alpha Linolenic acid metabolism(Williard *et al.*, 2001). The depletion of alpha Linolenic acid pathway metabolites may indicate that, in UC, alpha Linolenic acid may play a significant role in disease prevention and treatment. It is one of the essential n-3 polyunsaturated fatty acids that is required for cell membrane functionality and regulation of body function such as brain function and inflammatory responses (Wall *et al.*, 2010).

Bile acid metabolism in this study gave a clear differentiation between diseased and healthy faecal samples. From our results, four primary bile acid metabolites decreased in UC in comparison to HC. These metabolites were required for cholic acid biosynthesis. Three of these metabolites (Table 5.8) (2-beta,3-alpha,7-alpha,12-alpha-Tetrahydroxy-5-beta-cholestan-26-oic acid , 3-alpha,7-alpha,12-alpha-

trihydroxy-5-beta-cholestanoate and 5-beta-cholestane-3-alpha,7-alpha,12-alpha-26-tetrol) were classified as members of the top ten metabolites based on VIP ranking. The level of cholic acid in UC was increased but with an insignificant *p*-value (Log ratio = 1.6, p = 0.32). A previous study on 161 plasma samples using HPLC-MS, reported that the pool of bile acids in UC patients had no significant alteration (Gnewuch *et al.*, 2009). In addition, UC diarrhoea has been inferred to be due to the increase of the water and electrolytes in the stool and suggests that diarrhoea may not be related to bile acid alteration (Miettinen, 1971). On the other hand, in another study bile acids appeared with positive correlation in UC in comparison with HC (Le Gall *et al.*, 2011). In this study, Le Gall et al analysed 13 diseased faecal extracts against 22 HC samples using high resolution 1H NMR spectroscopy. The level of bile acids was positively correlated to the level of taurine in UC. The increased level of taurine may be due to the bacterial de-conjugation of bile acids (Ridlon, Kang and Hylemon, 2006). Our results were similar, however, taurine appeared in high levels in both UC and CD and therefore, may not be a suitable metabolite to discriminate the diseases. The low number of samples and the high variation in UC samples may affect the results in comparison to CD samples.

5.8.4 Comparison between CD and UC

In the present study, metabolomics analysis was performed to classify the faecal extracts from patients suffering from inflammatory bowel diseases to evaluate the use of this technique as a diagnostic tool and categorize specific metabolites in the faeces of participants with specific types of inflammation. The higher levels of

Diacylglycerols, such as DG (15:0/18:3(6Z,9Z,12Z)/0:0) and DG (15:0/18:2(9Z,12Z)/0:0), was a noticeable feature of Crohn's patients when compared with UC. The increase of Diacylglycerols was observed even in the common metabolites between the two diseases in comparison to HC (Figure 5.9). This study was able to classified four Diacylglycerols in the top ten VIP list of metabolites that can discriminate CD from UC. This suggests that the metabolomics effects caused by inflammation were more marked in CD than in UC patients. Diacylglycerols as lipid modulators could bind to cytokines and initiate the inflammation pathway (Johnson, Justin Milner and Makowski, 2012). Another study on faecal extracts suggest the same trend of metabolomics changes based of the higher levels of diacylglycerols in CD over UC (Marchesi *et al.*, 2007).

Similar to the trend for Diacylglycerols trend in CD, bile acid metabolites were increased in comparison to UC. As discussed previously, in previous work there was found to be a malabsorption of bile acids in CD patients that reflected the low level of bile acids in plasma (Färkkilä and Miettinen, 1990). This imbalance in bile acid levels may induce diarrhoea in CD (Martínez-Augustin and Medina, 2008). Moreover, bile acid malabsorption in CD may be caused by the reduction of the apical sodium dependent bile acid transporter expression in ileal biopsies from patients with CD (Jung, 2004).

Chapter 6:

Summary and future works

6 Summary and future works

From the beginning, this thesis was designed to investigate the application of metabolomics in human health and disease by assessing biomarkers associated with inflammatory bowel diseases. All the studies reported in this thesis employed a LC-MS analytical principle based on the Orbitrap Exactive mass analyser, and using HILIC or/and reversed phase RP analytical columns. The LC-MS employed XCalibur software through which the system functionality was controlled. LC-MS has the advantage of accurate mass detection which provides capacity for direct metabolite identification even in the absence of chromatographic resolution. Metabolite identification was based on retention times of the samples relative to authentic reference standards injected at specified intervals into the system in the same sequence. In addition, all the studies employed both unsupervised (PCA-X) and supervised (OPLS-DA) models in SIMCA in order to determine discriminating metabolite biomarkers responsible for the observed natural clustering patterns and supervised separations in OPLS-DA. Model parameters (R^2, Q^2 and *p* CV-ANOVA) were used to evaluate the validity of each model.

6.1 Data pre-treatment and statistical model selection

In LC-MS metabolomics data output is a high dimensional data that may require extra care prior to univariate or multivariate analysis. There are different ways an algorithm can be applied to this data; however, the data type as well as the aim of the study may affect the selection of the data analysis pipeline. Each step of data pre-treatment can significantly impact the model parameters, such as goodness of fit and goodness of prediction. In addition, it is necessary to further examine the data structure before conducting untargeted metabolomic analysis. The drawback of this study is that it is based on a single data set and may be useful if applied to different data sets resulting from different samples such as cell extract, natural products, and plasma.

Considering missing data imputation methods and based on the present study's data (urine extract from soccer players) and modelling exercise using SIMCA-P 14, the SIMCA-P 14's (NIPALS) default logarithm was the only imputation methodology that generated a valid model according to validation criteria (R^2-Q^2<0.3). This was in conjugation with Par or UV scaling with Par offering slight advantage over UV scaling. It is evident that this imputation method was better than KNN, Min/2, mean and median imputation as it included significant and valid models. However, other imputation can be compared to this method such as RF algorithm which is also commonly used in metabolomic analysis.

6.2 Untargeted Metabolomics of Paediatric Crohn's Disease Patients Against Healthy Controls

Several metabolomic differences were found in the faecal metabolome of paediatric patients with CD compared to the HC group. Thus, multivariate statistical methods were used to refine the marker list. An OPLS-DA model was able to separate all the CD groups throughout treatment and post-treatment from the HC group. However, it was not possible to obtain a valid model separating the CD groups throughout the different phases of treatment apart from between PA and PC. The eight markers which separated the CD groups from the HC groups were all normally distributed according to Q-Q tests.

Large elevations in omega 6 fatty acids were observed in the CD patients in comparison with the HC group, conforming to previous work that highlighted these compounds as being pro-inflammatory in the gut. The results of this study indicate that major metabolic differences remained between the HC and the CD groups even after apparently successful treatment; these metabolic differences could be clearly separated using multivariate statistical methods. The BMI values for the control group were not recorded and could impact on the results although this would seem more likely to occur for the metabolome of plasma rather than the faecal metabolome which is much more related to the activity of the microbiome.

6.3 Metabolomics Discrimination between Crohn's Disease and Ulcerative Colitis

CD and UC are considered as the two major subtypes of disorder diagnosed under IBD. These diseases described as multifactorial disorders. The gut microbiota may have significant role in IBD. Specific and speedy diagnosis may play significant roles in patient compliance and quality of life as well as in the treatment of the disease. In addition to that Non-invasive diagnostic methods are needed to differentiate between CD and UC.

In this study, untargeted metabolomics using LC-MS as analytical technique was employed to identify the significant metabolites that discriminate CD from UC based on direct comparison. In addition to that, during this comparison there were significant metabolites that were shared between CD and UC. Based on 70 faecal extract samples provided from paediatric participants, there was a clear separation between disease samples (CD and UC) and healthy controls samples (HC). This result provided after OPLS-DA model was applied.

The higher levels of Diacylglycerols was a noticeable feature of Crohn's patients when compared with UC. The significant metabolites then were ranked passed on VIP and provided a list that could discriminate CD from UC. This study was able to classify four Diacylglycerols in the top ten VIP list of metabolites that can discriminate CD from UC. Similarly, bile acids were higher in CD in comparison to UC.

6.4 Future works

The following can be proposed as studies which can be used to extend the current work.

- Carry out fuller characterisation of some of the putatively identified marker compounds observed to discriminate HC, UC and CD. This would require re-running on state of the art Orbitrap equipment since some of the marker compounds are at low levels and in order to get good quality MS^n spectra high sensitivity would be required.
- Carry out analysis of a larger set of clinical samples in order to validate the existing markers.
- Set up quantitative targeted assays for promising marker compounds by using tandem mass spectrometry enabling rapid through put of large numbers of samples.

7 References

Aberra, F. N. and Lichtenstein, G. R. (2005) 'Review article: monitoring of immunomodulators in inflammatory bowel disease', *Alimentary Pharmacology and Therapeutics*, 21(4), pp. 307–319. doi: 10.1111/j.1365-2036.2005.02343.x.

Al-Jaouni, R. *et al.* (2000) 'Energy metabolism and substrate oxidation in patients with Crohn's Disease', *Nutrition*, 16(3), pp. 173–178. doi: 10.1016/S0899-9007(99)00281-6.

Alonso, A., Marsal, S. and Julià, A. (2015) 'Analytical methods in untargeted metabolomics: state of the art in 2015', *Frontiers in bioengineering and biotechnology*, 3(March), p. 23. doi: 10.3389/fbioe.2015.00023.

Ambrose, T. and Simmons, A. (2018) 'Cannabis, cannabinoids and the endocannabinoid system – is there therapeutic potential for inflammatory bowel disease?', *Journal of Crohn's and Colitis*, pp. 1–11. doi: 10.1093/ecco-jcc/jjy185.

Ananthakrishnan, A. N. (2015) 'Epidemiology and risk factors for IBD', *Nature Reviews Gastroenterology and Hepatology*. Nature Publishing Group, 12(4), pp. 205–217. doi: 10.1038/nrgastro.2015.34.

Angulo, S. *et al.* (2011) 'Probiotic sonicates selectively induce mucosal immune cells apoptosis through ceramide generation via neutral sphingomyelinase', *PLoS ONE*, 6(3), pp. 1–12. doi: 10.1371/journal.pone.0016953.

Armitage, E. G. *et al.* (2015) 'Missing value imputation strategies for metabolomics data', *Electrophoresis*, 36(24), pp. 3050–3060. doi: 10.1002/elps.201500352.

Atherton, H. J. *et al.* (2006) 'A combined 1H-NMR spectroscopy- and mass spectrometry-based metabolomic study of the PPAR-α null mutant mouse defines profound systemic changes in metabolism linked to the metabolic syndrome', *Physiological Genomics*, 27(2), pp. 178–186. doi: 10.1152/physiolgenomics.00060.2006.

Bassi, A. *et al.* (2004) 'Cost of illness of inflammatory bowel disease in the UK: a single centre retrospective study', *Gut*, 53(10), pp. 1471 LP – 1478. doi: 10.1136/gut.2004.041616.

Baumgart, D. C. (2017) *Crohn's Disease and Ulcerative Colitis*. Edited by D. C. Baumgart. Cham: Springer International Publishing. doi: 10.1007/978-3-319-33703-6.

Baur, P. *et al.* (2011) 'Metabolic Phenotyping of the Crohn's Disease-like IBD Etiopathology in the TNF ΔARE/WT Mouse Model', *Journal of Proteome Research*, 10(12), pp. 5523–5535. doi: 10.1021/pr2007973.

Benjamini, Y. *et al.* (2001) 'Controlling the false discovery rate in behavior genetics

research.', *Behavioural brain research*, 125(1–2), pp. 279–84. doi: 10.2307/2346101.

van den Berg, R. a *et al.* (2006) 'Centering, scaling, and transformations: improving the biological information content of metabolomics data.', *BMC genomics*, 7, p. 142. doi: 10.1186/1471-2164-7-142.

Bjerrum, J. T. *et al.* (2015) 'Metabonomics of human fecal extracts characterize ulcerative colitis, Crohn's disease and healthy individuals', *Metabolomics*, 11, pp. 122–133. doi: 10.1007/s11306-014-0677-3.

Bláha, V. *et al.* (2009) 'Cholesterol metabolism in active Crohn's diseaseCholesterin-Stoffwechsel bei aktivem Morbus Crohn', *Wiener klinische Wochenschrift*, 121(7–8), pp. 270–275. doi: 10.1007/s00508-009-1150-6.

Blasco, H. *et al.* (2015) 'Comparative analysis of targeted metabolomics: Dominance-based rough set approach versus orthogonal partial least square-discriminant analysis', *Journal of Biomedical Informatics*, 53, pp. 291–299. doi: 10.1016/j.jbi.2014.12.001.

Boccard, J., Veuthey, J.-L. and Rudaz, S. (2010) 'Knowledge discovery in metabolomics: An overview of MS data handling', *Journal of Separation Science*, 33(3), pp. 290–304. doi: 10.1002/jssc.200900609.

Borrelli, O. *et al.* (2006) 'Polymeric Diet Alone Versus Corticosteroids in the Treatment of Active Pediatric Crohn's Disease: A Randomized Controlled Open-Label Trial', *Clinical Gastroenterology and Hepatology*. Elsevier, 4(6), pp. 744–753. doi: 10.1016/j.cgh.2006.03.010.

Broadhurst, D. I. and Kell, D. B. (2007) 'Statistical strategies for avoiding false discoveries in metabolomics and related experiments', 2(4). doi: 10.1007/s11306-006-0037-z.

Burisch, J. *et al.* (2014) 'East–West gradient in the incidence of inflammatory bowel disease in Europe: the ECCO-EpiCom inception cohort', *Gut*, 63(4), pp. 588 LP – 597. doi: 10.1136/gutjnl-2013-304636.

Cameron, F. L. *et al.* (2013) 'Clinical progress in the two years following a course of exclusive enteral nutrition in 109 paediatric patients with Crohn's disease', *Alimentary Pharmacology and Therapeutics*, 37(6), pp. 622–629. doi: 10.1111/apt.12230.

Chong, I.-G. and Jun, C.-H. (2005) 'Performance of some variable selection methods when multicollinearity is present', *Chemometrics and Intelligent Laboratory Systems*, 78(1–2), pp. 103–112. doi: 10.1016/j.chemolab.2004.12.011.

Chong, J. *et al.* (2018) 'MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis', *Nucleic Acids Research*. Oxford University Press, 46(W1), pp. W486–W494. doi: 10.1093/nar/gky310.

Cosnes, J. *et al.* (2011) 'Epidemiology and natural history of inflammatory bowel

diseases.', *Gastroenterology*. United States, 140(6), pp. 1785–1794. doi: 10.1053/j.gastro.2011.01.055.

Costa, C. G. *et al.* (2000) 'Quantitative analysis of urinary acylglycines for the diagnosis of β-oxidation defects using GC-NCI-MS', *Journal of Pharmaceutical and Biomedical Analysis*, 21(6), pp. 1215–1224. doi: 10.1016/S0731-7085(99)00235-6.

Creek, D. J. *et al.* (2012) 'IDEOM: an Excel interface for analysis of LC-MS-based metabolomics data', *Bioinformatics*, 28(7), pp. 1048–1049. doi: 10.1093/bioinformatics/bts069.

Critch, J. *et al.* (2012) 'Use of Enteral Nutrition for the Control of Intestinal Inflammation in Pediatric Crohn Disease', *Journal of Pediatric Gastroenterology and Nutrition*, 54(2), pp. 298–305. doi: 10.1097/MPG.0b013e318235b397.

Daiss, W., Scheurlen, M. and Malchow, H. (1989) 'Epidemiology of Inflammatory Bowel Disease in the County of Tübingen (West Germany)', *Scandinavian Journal of Gastroenterology*. England, 24(sup170), pp. 39–43. doi: 10.3109/00365528909091349.

Day, A. S. and Lopez, R. N. (2015) 'Exclusive enteral nutrition in children with crohn's disease', *World Journal of Gastroenterology*, 21(22), pp. 6809–6816. doi: 10.3748/wjg.v21.i22.6809.

Desreumaux, P. and Ghosh, S. (2006) 'Review article: Mode of action and delivery of 5-aminosalicylic acid - New evidence', *Alimentary Pharmacology and Therapeutics*, 24(SUPPL. 1), pp. 2–9. doi: 10.1111/j.1365-2036.2006.03069.x.

Dolan, K. T. and Chang, E. B. (2017) 'Diet, gut microbes, and the pathogenesis of inflammatory bowel diseases', *Molecular Nutrition and Food Research*, 61(1), pp. 1–13. doi: 10.1002/mnfr.201600129.

Donnet-Hughes, A., Schiffrin, E. J. and Turini, M. E. (2001) 'The intestinal mucosa as a target for dietary polyunsaturated fatty acids', *Lipids*. Springer, 36(9), pp. 1043–1052. doi: 10.1007/s11745-001-0815-4.

Dunn, W. B. (2008) 'Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes', *Physical Biology*, 5(1), p. 011001. doi: 10.1088/1478-3975/5/1/011001.

Dunn, W. B. (2013) 'Diabetes - the Role of Metabolomics in the Discovery of New Mechanisms and Novel Biomarkers', pp. 25–32. doi: 10.1007/s12170-012-0282-9.

Dunn, W. B., Bailey, N. J. C. and Johnson, H. E. (2005) 'Measuring the metabolome: current analytical technologies', *The Analyst*. England, 130(5), p. 606. doi: 10.1039/b418288j.

Dupaul-Chicoine, J., Dagenais, M. and Saleh, M. (2013) 'Crosstalk between the intestinal microbiota and the innate immune system in intestinal homeostasis and inflammatory bowel disease.', *Inflammatory bowel diseases*. England, 19(10), pp.

2227–2237. doi: 10.1097/MIB.0b013e31828dcac7.

Efron, B. and Gong, G. (1983) 'A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation', *The American Statistician*, 37(1), pp. 36–48. doi: 10.1080/00031305.1983.10483087.

Eliuk, S. and Makarov, A. (2015) 'Evolution of Orbitrap Mass Spectrometry Instrumentation', *Annual Review of Analytical Chemistry*, 8(1), pp. 61–80. doi: 10.1146/annurev-anchem-071114-040325.

Eriksson, L. *et al.* (2013) *Multi-and megavariate data analysis basic principles and applications*. Umetrics Academy.

Eriksson, L., Trygg, J. and Wold, S. (2008) 'CV-ANOVA for significance testing of PLS and OPLS® models', *Journal of Chemometrics*, 22(11–12), pp. 594–600. doi: 10.1002/cem.1187.

Fabiano, E. and O'Brian, M. R. (2012) 'Mechanisms and Regulation of Iron Homeostasis in the Rhizobia', in *Iron Uptake and Homeostasis in Microorganisms*. Caister Academic Press Haverhill, Suffolk, UK, pp. 41–86. doi: 10.1007/978-94-007-5267-2_3.

Färkkilä, M. and Miettinen, T. A. (1990) 'Lipid Metabolism in Bile Acid Malabsorption', *Annals of Medicine*. England, 22(1), pp. 5–13. doi: 10.3109/07853899009147233.

Feller, M. *et al.* (2007) 'Mycobacterium avium subspecies paratuberculosis and Crohn's disease: a systematic review and meta-analysis', *The Lancet Infectious Diseases*, 7(9), pp. 607–613. doi: 10.1016/S1473-3099(07)70211-6.

Feng, C. *et al.* (2014) 'Log-transformation and its implications for data analysis.', *Shanghai archives of psychiatry*, 26(2), pp. 105–9. doi: 10.3969/j.issn.1002-0829.2014.02.009.

Fiehn, O. (2002) 'Metabolomics — the link between genotypes and phenotypes', in *Functional Genomics*. Dordrecht: Springer Netherlands, pp. 155–171. doi: 10.1007/978-94-010-0448-0_11.

Gaifem, J. *et al.* (2018) 'L-Threonine Supplementation During Colitis Onset Delays Disease Recovery.', *Frontiers in physiology*. Switzerland, 9, p. 1247. doi: 10.3389/fphys.2018.01247.

Gajendran, M. *et al.* (2019) 'A comprehensive review and update on ulcerative colitis.', *Disease-a-month : DM*. United States, 65(12), p. 100851. doi: 10.1016/j.disamonth.2019.02.004.

Le Gall, G. *et al.* (2011) 'Metabolomics of fecal extracts detects altered metabolic activity of gut microbiota in ulcerative colitis and irritable bowel syndrome', *Journal of Proteome Research*, 10(9), pp. 4208–4218. doi: 10.1021/pr2003598.

Gallien, S. *et al.* (2012) 'Targeted Proteomic Quantification on Quadrupole-Orbitrap Mass Spectrometer', *Molecular & Cellular Proteomics*, 11(12), pp. 1709–1723. doi:

10.1074/mcp.O112.019802.

Gatti, S. *et al.* (2017) 'Effects of the exclusive enteral nutrition on the microbiota profile of patients with crohn's disease: A systematic review', *Nutrients*, 9(8), pp. 3–5. doi: 10.3390/nu9080832.

Geboes, K. *et al.* (2008) 'Indeterminate colitis: a review of the concept--what's in a name?', *Inflammatory bowel diseases*. England, 14(6), pp. 850–857. doi: 10.1002/ibd.20361.

Gerasimidis, K. *et al.* (2011) 'Serial fecal calprotectin changes in children with Crohn's disease on treatment with exclusive enteral nutrition: associations with disease activity, treatment response, and prediction of a clinical relapse.', *Journal of clinical gastroenterology*. United States, 45(3), pp. 234–239. doi: 10.1097/MCG.0b013e3181f39af5.

Gerasimidis, K. *et al.* (2014) 'Decline in presumptively protective gut bacterial species and metabolites are paradoxically associated with disease improvement in pediatric Crohn's disease during enteral nutrition.', *Inflammatory bowel diseases*, 20(5), pp. 861–871. doi: 10.1097/MIB.0000000000000023.

Ghosh, N. and Premchand, P. (2015) 'A UK cost of care model for inflammatory bowel disease', *Frontline gastroenterology*. 2015/02/24. BMJ Publishing Group, 6(3), pp. 169–174. doi: 10.1136/flgastro-2014-100514.

Gika, H. G., Wilson, I. D. and Theodoridis, G. A. (2014) 'LC–MS-based holistic metabolic profiling. Problems, limitations, advantages, and future perspectives', *Journal of Chromatography B*. Elsevier, 966, pp. 1–6. doi: 10.1016/j.jchromb.2014.01.054.

van Ginneken, V. *et al.* (2007) 'Metabolomics (liver and blood profiling) in a mouse model in response to fasting: A study of hepatic steatosis', *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1771(10), pp. 1263–1270. doi: 10.1016/j.bbalip.2007.07.007.

Gisbert, J. P., González-Lama, Y. and Maté, J. (2007) '5-Aminosalicylates and renal function in inflammatory bowel disease', *Inflammatory Bowel Diseases*, 13(5), pp. 629–638. doi: 10.1002/ibd.20099.

Gnewuch, C. *et al.* (2009) 'Serum bile acid profiling reflects enterohepatic detoxification state and intestinal barrier function in inflammatory bowel disease', *World Journal of Gastroenterology*, 15(25), pp. 3134–3141. doi: 10.3748/wjg.15.3134.

Goodacre, R. (2007) 'Metabolomics of a Superorganism', *The Journal of Nutrition*, 137(1), pp. 259S-266S. doi: 10.1093/jn/137.1.259s.

Goodacre, R. *et al.* (2007) 'Proposed minimum reporting standards for data analysis in metabolomics', *Metabolomics*, 3(3), pp. 231–241. doi: 10.1007/s11306-007-0081-3.

Goryński, K. *et al.* (2013) 'Quantitative structure–retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds', *Analytica Chimica Acta*, 797, pp. 13–19. doi: 10.1016/j.aca.2013.08.025.

Goulding, N. J. (2004) 'The molecular complexity of glucocorticoid actions in inflammation — a four-ring circus', *Current Opinion in Pharmacology*, 4(6), pp. 629–636. doi: 10.1016/j.coph.2004.06.009.

Gromski, P. *et al.* (2014) 'Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data', *Metabolites*, 4(2), pp. 433–452. doi: 10.3390/metabo4020433.

Grung, B. and Manne, R. (1998) 'Missing values in principal component analysis', *Chemometrics and Intelligent Laboratory Systems*, 42(1–2), pp. 125–139. doi: 10.1016/S0169-7439(98)00031-8.

Di Guida, R. *et al.* (2016) 'Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling', *Metabolomics*. Springer US, 12(5), pp. 1–14. doi: 10.1007/s11306-016-1030-9.

Harrigan, G. G. and Goodacre, R. (2012) *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*. Springer Science & Business Media.

Harris, D. c (2010) *Quantitative Chemical Analysis*, *New York*. doi: 10.1016/j.micron.2011.01.004.

Hasenoehrl, C. *et al.* (2016) 'The gastrointestinal tract - a central organ of cannabinoid signaling in health and disease', *Neurogastroenterology & Motility*, 28(12), pp. 1765–1780. doi: 10.1111/nmo.12931.

Heimerl, S. *et al.* (2006) 'Alterations in intestinal fatty acid metabolism in inflammatory bowel disease', *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1762(3), pp. 341–350. doi: 10.1016/j.bbadis.2005.12.006.

Hemström, P. and Irgum, K. (2006) 'Hydrophilic interaction chromatography', *Journal of separation science*, 29, pp. 1784–1821. doi: 10.1002/jssc.200600199.

Hendriks, M. M. W. B. *et al.* (2005) 'Preprocessing and exploratory analysis of chromatographic profiles of plant extracts', *Analytica Chimica Acta*, 545(1), pp. 53–64. doi: 10.1016/j.aca.2005.04.026.

Hou, J. K., Abraham, B. and El-Serag, H. (2011) 'Dietary intake and risk of developing inflammatory bowel disease: A systematic review of the literature', *American Journal of Gastroenterology*. Nature Publishing Group, 106(4), pp. 563–573. doi: 10.1038/ajg.2011.44.

Howe, C. *et al.* (2018) 'Untargeted Metabolomics Profiling of an 80.5 km Simulated Treadmill Ultramarathon', *Metabolites*, 8(1), p. 14. doi: 10.3390/metabo8010014.

Hrydziuszko, O. and Viant, M. R. (2012) 'Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline', *Metabolomics*, 8, pp. 161–174. doi: 10.1007/s11306-011-0366-4.

Huan, T. and Li, L. (2015) 'Counting missing values in a metabolite-intensity data set for measuring the analytical performance of a metabolomics platform', *Analytical Chemistry*, 87(2), pp. 1306–1313. doi: 10.1021/ac5039994.

Irving, P. M. *et al.* (2007) 'Review article: appropriate use of corticosteroids in Crohn's disease', *Alimentary Pharmacology & Therapeutics*, 26(3), pp. 313–329. doi: 10.1111/j.1365-2036.2007.03379.x.

Janowitz, H. D., Croen, E. C. and Sachar, D. B. (1998) 'The Role of the Fecal Stream in Crohn's Disease: An Historical and Analytic Review', *Inflammatory Bowel Diseases*, 4(1), pp. 29–39. doi: 10.1097/00054725-199802000-00006.

Jansson, J. *et al.* (2009) 'Metabolomics reveals metabolic biomarkers of Crohn's disease', *PLoS ONE*, 4(7). doi: 10.1371/journal.pone.0006386.

Jenkins, S. *et al.* (2013) 'Global LC/MS Metabolomics Profiling of Calcium Stressed and Immunosuppressant Drug Treated Saccharomyces cerevisiae', *Metabolites*, 3(4), pp. 1102–1117. doi: 10.3390/metabo3041102.

Johnson, A. R., Justin Milner, J. and Makowski, L. (2012) 'The inflammation highway: metabolism accelerates inflammatory traffic in obesity', *Immunological Reviews*, 249(1), pp. 218–238. doi: 10.1111/j.1600-065X.2012.01151.x.

de Jong, N. S. H., Leach, S. T. and Day, A. S. (2007) 'Polymeric formula has direct anti-inflammatory effects on enterocytes in an in vitro model of intestinal inflammation.', *Digestive diseases and sciences*. United States, 52(9), pp. 2029–2036. doi: 10.1007/s10620-006-9449-x.

Jung, D. (2004) 'Human ileal bile acid transporter gene ASBT (SLC10A2) is transactivated by the glucocorticoid receptor', *Gut*, 53(1), pp. 78–84. doi: 10.1136/gut.53.1.78.

Kaliannan, K. *et al.* (2015) 'A host-microbiome interaction mediates the opposing effects of omega-6 and omega-3 fatty acids on metabolic endotoxemia', *Scientific Reports*. Nature Publishing Group, 5(February), pp. 1–17. doi: 10.1038/srep11276.

Kamleh, A. *et al.* (2008) 'Metabolomic profiling using Orbitrap Fourier transform mass spectrometry with hydrophilic interaction chromatography: a method with wide applicability to analysis of biomolecules', *Rapid Communications in Mass Spectrometry*, 22(12), pp. 1912–1918. doi: 10.1002/rcm.3564.

Kane, S. V. *et al.* (2003) 'Fecal lactoferrin is a sensitive and specific marker in identifying intestinal inflammation', *The American Journal of Gastroenterology*, 98(6), pp. 1309–1314. doi: 10.1111/j.1572-0241.2003.07458.x.

Karaman, I. (2017) 'Preprocessing and Pretreatment of Metabolomics Data for

Statistical Analysis', in Sussulini, A. (ed.) *Metabolomics: From Fundamentals to Clinical Applications*. Cham: Springer International Publishing, pp. 145–161. doi: 10.1007/978-3-319-47656-8_6.

Karu, N. *et al.* (2018) 'A review on human fecal metabolomics: Methods, applications and the human fecal metabolome database', *Analytica Chimica Acta*. Elsevier Ltd, 1030, pp. 1–24. doi: 10.1016/j.aca.2018.05.031.

Kedia, S. and Ahuja, V. (2017) 'Epidemiology of Inflammatory Bowel Disease in India: The Great Shift East', *Inflammatory Intestinal Diseases*, 2(2), pp. 102–115. doi: 10.1159/000465522.

Kell, D. B. and Oliver, S. G. (2004) 'Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era.', *BioEssays : news and reviews in molecular, cellular and developmental biology*. United States, 26(1), pp. 99–105. doi: 10.1002/bies.10385.

Kim, H. K. and Verpoorte, R. (2010) 'Sample preparation for plant metabolomics', *Phytochemical Analysis*, 21(1), pp. 4–13. doi: 10.1002/pca.1188.

Kirwan, G. M. *et al.* (2012) 'Building multivariate systems biology models', *Analytical Chemistry*, 84(16), pp. 7064–7071. doi: 10.1021/ac301269r.

Kolho, K. L. *et al.* (2016) 'Faecal and serum metabolomics in paediatric inflammatory bowel disease', *Journal of Crohn's & colitis*, pp. 1–14. doi: jjw158 [pii].

Kolho, K. L. *et al.* (2017) 'Faecal and Serum Metabolomics in Paediatric Inflammatory Bowel Disease', *Journal of Crohn's & colitis*, 11(3), pp. 321–334. doi: 10.1093/ecco-jcc/jjw158.

Kornbluth, A. and Sachar, D. B. (2010) 'Ulcerative colitis practice guidelines in adults: American College Of Gastroenterology, Practice Parameters Committee.', *The American journal of gastroenterology*. United States, 105(3), pp. 501–23; quiz 524. doi: 10.1038/ajg.2009.727.

Kotze, P. G. *et al.* (2020) 'Progression of Inflammatory Bowel Diseases Throughout Latin America and the Caribbean: A Systematic Review', *Clinical Gastroenterology and Hepatology*. Elsevier, Inc, 18(2), pp. 304–312. doi: 10.1016/j.cgh.2019.06.030.

Kraj, A., Desiderio, D. M. and Nibbering, N. M. (2008) *Mass spectrometry: instrumentation, interpretation, and applications*. John Wiley & Sons. doi: 10.1021/jasms.8b03616.

Kvalheim, O. M., Brakstad, F. and Llang, Y. zeng (1994) 'Preprocessing of Analytical Profiles in the Presence of Homoscedastic or Heteroscedastlc Noise', *Analytical Chemistry*, 66(1), pp. 43–51. doi: 10.1021/ac00073a010.

LEACH, S. T. *et al.* (2008) 'Sustained modulation of intestinal bacteria by exclusive enteral nutrition used to treat children with Crohn's disease', *Alimentary Pharmacology & Therapeutics*. England, 28(6), pp. 724–733. doi: 10.1111/j.1365-

2036.2008.03796.x.

Leiss, K. A. *et al.* (2011) 'An overview of NMR-based metabolomics to identify secondary plant compounds involved in host plant resistance', *Phytochemistry Reviews*, 10(2), pp. 205–216. doi: 10.1007/s11101-010-9175-z.

Lichtenstein, G. R. *et al.* (2018) 'ACG Clinical Guideline: Management of Crohn's Disease in Adults', *American Journal of Gastroenterology*, 113(4), pp. 481–517. doi: 10.1038/ajg.2018.27.

Lockhart-Mummer, H. E. and Morson, B. C. (1964) 'Crohn's Disease of The Large Intestine', *Gut*, 5(6), pp. 493–509. doi: 10.1136/gut.5.6.493.

Loftus, E. V (2004) 'Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences', *Gastroenterology*. United States, 126(6), pp. 1504–1517. doi: 10.1053/j.gastro.2004.01.063.

Luces, C. and Bodger, K. (2006) 'Economic burden of inflammatory bowel disease: a UK perspective.', *Expert review of pharmacoeconomics & outcomes research*. England, 6(4), pp. 471–482. doi: 10.1586/14737167.6.4.471.

MacDonald, T. T. (2011) 'New cytokine targets in inflammatory bowel disease', *Gastroenterology and Hepatology*, 7(7), pp. 474–476.

Makarov, A. *et al.* (2006) 'Dynamic range of mass accuracy in LTQ orbitrap hybrid mass spectrometer', *Journal of the American Society for Mass Spectrometry*, 17(7), pp. 977–982. doi: 10.1016/j.jasms.2006.03.006.

Marchesi, J. R. *et al.* (2007) 'Rapid and noninvasive metabonomic characterization of inflammatory bowel disease', *Journal of Proteome Research*, 6(2), pp. 546–551. doi: 10.1021/pr060470d.

Marion-Letellier, R., Savoye, G. and Ghosh, S. (2015) 'Polyunsaturated fatty acids and inflammation', *IUBMB Life*, 67(9), pp. 659–667. doi: 10.1002/iub.1428.

Marquéz, L. *et al.* (2009) 'Ulcerative colitis induces changes on the expression of the endocannabinoid system in the human colonic tissue', *PLoS ONE*, 4(9). doi: 10.1371/journal.pone.0006893.

Martin, F. *et al.* (2017) 'Urinary metabolic insights into host-gut microbial interactions in healthy and IBD children', 23(20), pp. 3643–3654. doi: 10.3748/wjg.v23.i20.3643.

Martínez-Augustin, O. and Medina, F. S. De (2008) 'Intestinal bile acid physiology and pathophysiology', *World Journal of Gastroenterology*, 14(37), p. 5630. doi: 10.3748/wjg.14.5630.

McCalley, D. V. (2017) 'Understanding and manipulating the separation in hydrophilic interaction liquid chromatography', *Journal of Chromatography A*, 1523, pp. 49–71. doi: 10.1016/j.chroma.2017.06.026.

Michalski, A. *et al.* (2011) 'Mass Spectrometry-based Proteomics Using Q Exactive, a

High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer', *Molecular & Cellular Proteomics*, 10(9), p. M111.011015. doi: 10.1074/mcp.M111.011015.

Miettinen, T. A. (1971) 'The role of bile salts in diarrhoea of patients with ulcerative colitis', *Gut*, 12(8), pp. 632–635. doi: 10.1136/gut.12.8.632.

Mowat, C. *et al.* (2011) 'Guidelines for the management of inflammatory bowel disease in adults', *Gut*, 60(5), pp. 571–607. doi: 10.1136/gut.2010.224154.

Musso, G., Gambino, R. and Cassader, M. (2010) 'Interactions Between Gut Microbiota and Host Metabolism Predisposing to Obesity and Diabetes', *Annual Review of Medicine*, 62(1), pp. 361–380. doi: 10.1146/annurev-med-012510-175505.

Nahidi, L. *et al.* (2013) 'Inflammatory bowel disease therapies and gut function in a colitis mouse model', *BioMed research international*. 2013/08/06. Hindawi Publishing Corporation, 2013, p. 909613. doi: 10.1155/2013/909613.

Navas López, V. M. *et al.* (2014) 'Consensus guidelines of ECCO/ESPGHAN on the medical management of pediatric Crohn's disease', *Journal of Crohn's and Colitis*, 8(10), pp. 1179–1207. doi: 10.1016/j.crohns.2014.04.005.

Nelson, P. R. C., Taylor, P. A. and MacGregor, J. F. (1996) 'Missing data methods in PCA and PLS: Score calculations with incomplete observations', *Chemometrics and Intelligent Laboratory Systems*. Elsevier, 35(1), pp. 45–65. doi: 10.1016/S0169-7439(96)00007-X.

Ng, S. C. *et al.* (2017) 'Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies', *The Lancet*. Elsevier, 390(10114), pp. 2769–2778. doi: 10.1016/S0140-6736(17)32448-0.

Ng, W. K., Wong, S. H. and Ng, S. C. (2016) 'Changing epidemiological trends of inflammatory bowel disease in Asia', *Intestinal Research*, 14(2), pp. 111–119. doi: 10.5217/ir.2016.14.2.111.

Ning, L. *et al.* (2019) 'Quantitative Proteomic Analysis Reveals the Deregulation of Nicotinamide Adenine Dinucleotide Metabolism and CD38 in Inflammatory Bowel Disease.', *BioMed research international*. United States, 2019, p. 3950628. doi: 10.1155/2019/3950628.

Ohno, M., Karagiannis, P. and Taniguchi, Y. (2014) 'Protein expression analyses at the single cell level.', *Molecules (Basel, Switzerland)*. Switzerland, 19(9), pp. 13932–13947. doi: 10.3390/molecules190913932.

Pacheco, S., Hillier, K. and Smith, C. (1987) 'Increased arachidonic acid levels in phospholipids of human colonic mucosa in inflammatory bowel disease', *Clinical Science*. England, 73(4), pp. 361–364. doi: 10.1042/cs0730361.

Palmer, C. *et al.* (2007) 'Development of the human infant intestinal microbiota', *PLoS Biology*, 5(7), pp. 1556–1573. doi: 10.1371/journal.pbio.0050177.

Pearl, D. S. *et al.* (2014) 'Altered colonic mucosal availability of n-3 and n-6

polyunsaturated fatty acids in ulcerative colitis and the relationship to disease activity', *Journal of Crohn's and Colitis*. European Crohn's and Colitis Organisation, 8(1), pp. 70–79. doi: 10.1016/j.crohns.2013.03.013.

Pedreschi, R. *et al.* (2008) 'Treatment of missing values for multivariate statistical analysis of gel-based proteomics data', *Proteomics*, 8(7), pp. 1371–1383. doi: 10.1002/pmic.200700975.

Pitt, J. J. (2009) 'Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry.', *The Clinical biochemist. Reviews*, 30(1), pp. 19–34.

De Preter, V. *et al.* (2015) 'Faecal metabolite profiling identifies medium-chain fatty acids as discriminating compounds in IBD', *Gut*, 64(3), pp. 447–458. doi: 10.1136/gutjnl-2013-306423.

De Preter, V. (2015) 'Metabolomics in the Clinical Diagnosis of Inflammatory Bowel Disease', *Digestive Diseases*, 33(suppl 1), pp. 2–10. doi: 10.1159/000437033.

Quince, C. *et al.* (2015) 'Extensive Modulation of the Fecal Metagenome in Children With Crohn's Disease During Exclusive Enteral Nutrition.', *The American journal of gastroenterology*. Nature Publishing Group, 110(12), pp. 1718–29; quiz 1730. doi: 10.1038/ajg.2015.357.

Ricart, E. *et al.* (2008) 'Are we giving biologics too late? The case for early versus late use', *World Journal of Gastroenterology*, 14(36), pp. 5523–5527. doi: 10.3748/wjg.14.5523.

Ridlon, J. M., Kang, D.-J. and Hylemon, P. B. (2006) 'Bile salt biotransformations by human intestinal bacteria', *Journal of Lipid Research*, 47(2), pp. 241–259. doi: 10.1194/jlr.R500013-JLR200.

Di Sabatino, A. *et al.* (2012) 'The function of tissue transglutaminase in celiac disease', *Autoimmunity Reviews*. Elsevier B.V., 11(10), pp. 746–753. doi: 10.1016/j.autrev.2012.01.007.

Saccenti, E. *et al.* (2014) 'Reflections on univariate and multivariate analysis of metabolomics data', *Metabolomics*, 10(3), pp. 361–374. doi: 10.1007/s11306-013-0598-6.

Sartor, R. B. (2008) 'Microbial Influences in Inflammatory Bowel Diseases', *Gastroenterology*, 134(2), pp. 577–594. doi: https://doi.org/10.1053/j.gastro.2007.11.059.

Scheltema, R. A. *et al.* (2011) 'PeakML/mzMatch: A file format, Java library, R library, and tool-chain for mass spectrometry data analysis', *Analytical Chemistry*, 83(7), pp. 2786–2793. doi: 10.1021/ac2000994.

Schicho, R. *et al.* (2012) 'Quantitative metabolomic profiling of serum, plasma, and urine by 1H NMR spectroscopy discriminates between patients with inflammatory

bowel disease and healthy individuals', *Journal of Proteome Research*, 11(6), pp. 3344–3357. doi: dx.doi.org/10.1021/pr300139q.

Schicho, R. and Storr, M. (2014) 'IBD: Patients with IBD find symptom relief in the Cannabis field', *Nature Reviews Gastroenterology and Hepatology*. Nature Publishing Group, 11(3), pp. 142–143. doi: 10.1038/nrgastro.2013.245.

Schmitz, G. and Ecker, J. (2008) 'The opposing effects of n-3 and n-6 fatty acids.', *Progress in lipid research*. England, 47(2), pp. 147–155. doi: 10.1016/j.plipres.2007.12.004.

Schuller-Levis, G. B. and Park, E. (2003) 'Taurine: New implications for an old amino acid', *FEMS Microbiology Letters*, 226(2), pp. 195–202. doi: 10.1016/S0378-1097(03)00611-6.

Serena, G. and Fasano, A. (2019) 'Use of Probiotics to Prevent Celiac Disease and IBD in Pediatrics.', *Advances in experimental medicine and biology*. United States, 1125, pp. 69–81. doi: 10.1007/5584_2018_317.

Sewell, G. W. *et al.* (2012) 'Lipidomic profiling in Crohn's disease: Abnormalities in phosphatidylinositols, with preservation of ceramide, phosphatidylcholine and phosphatidylserine composition', *International Journal of Biochemistry and Cell Biology*. Elsevier Ltd, 44(11), pp. 1839–1846. doi: 10.1016/j.biocel.2012.06.016.

Shaffer, R. E. (2002) 'Multi-and Megavariate Data Analysis. Principles and Applications, I. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, Umetrics Academy, Umeå, 2001, ISBN 91-973730-1-X, 533pp.', *Journal of Chemometrics*. Wiley Online Library, 16(5), pp. 261–262.

Sheehan, D., Moran, C. and Shanahan, F. (2015) 'The microbiota in inflammatory bowel disease', *Journal of Gastroenterology*, 50(5), pp. 495–507. doi: 10.1007/s00535-015-1064-1.

Shivananda, S. *et al.* (1996) 'Incidence of inflammatory bowel disease across Europe: is there a difference between north and south? Results of the European Collaborative Study on Inflammatory Bowel Disease (EC-IBD).', *Gut*, 39(5), pp. 690–697. doi: 10.1136/gut.39.5.690.

Soubieres, A. A. and Poullis, A. (2016) 'Emerging Biomarkers for the Diagnosis and Monitoring of Inflammatory Bowel Diseases.', *Inflammatory bowel diseases*. England, 22(8), pp. 2016–2022. doi: 10.1097/MIB.0000000000000836.

Steehler, J. K. (2009) 'Introduction to Mass Spectrometry: Instrumentation, Applications, and Strategies for Data Interpretation, 4th Edition (by J. Throck Watson and O. David Sparkman)', *Journal of Chemical Education*, 86(7), p. 810. doi: 10.1021/ed086p810.1.

Sumner, L. W. *et al.* (2007) 'Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)', *Metabolomics*, 3(3), pp. 211–221. doi: 10.1007/s11306-007-0082-2.

Svolos, V. *et al.* (2019) 'Treatment of Active Crohn's Disease With an Ordinary Food-based Diet That Replicates Exclusive Enteral Nutrition', *Gastroenterology*, 156(5), pp. 1354-1367.e6. doi: 10.1053/j.gastro.2018.12.002.

Talley, N. J. *et al.* (2011) 'An evidence-based systematic review on medical therapies for inflammatory bowel disease.', *The American journal of gastroenterology*. United States, 106 Suppl, pp. S2-25; quiz S26. doi: 10.1038/ajg.2011.58.

Tontini, G. E. *et al.* (2015) 'Differential diagnosis in inflammatory bowel disease colitis: State of the art and future perspectives', *World Journal of Gastroenterology*, 21(1), pp. 21–46. doi: 10.3748/wjg.v21.i1.21.

Toruner, M. *et al.* (2008) 'Risk Factors for Opportunistic Infections in Patients With Inflammatory Bowel Disease', *Gastroenterology*, 134(4), pp. 929–936. doi: 10.1053/j.gastro.2008.01.012.

Triba, M. N. *et al.* (2015) 'PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters', *Mol. BioSyst.* Royal Society of Chemistry, 11(1), pp. 13–19. doi: 10.1039/C4MB00414K.

Troyanskaya, O. *et al.* (2001) 'Missing value estimation methods for DNA microarrays', *Bioinformatics*, 17(6), pp. 520–525. doi: 10.1093/bioinformatics/17.6.520.

Truelove, S. C. and Witts, L. J. (1954) 'Cortisone in Ulcerative Colitis', *BMJ*, 2(4884), pp. 375–378. doi: 10.1136/bmj.2.4884.375.

Trygg, J., Holmes, E. and Lundstedt, T. (2007) 'Chemometrics in metabonomics', *Journal of Proteome Research*, 6(2), pp. 469–479. doi: 10.1021/pr060594q.

Tugizimana, F. *et al.* (2016) 'A conversation on data mining strategies in LC-MS untargeted metabolomics: Pre-processing and pre-treatment steps', *Metabolites*, 6(4), pp. 1–18. doi: 10.3390/metabo6040040.

Uchiyama, K. *et al.* (2013) 'The fatty acid profile of the erythrocyte membrane in initial-onset inflammatory bowel disease patients', *Digestive Diseases and Sciences*, 58(5), pp. 1235–1243. doi: 10.1007/s10620-012-2508-6.

Vargesson, N. (2015) 'Thalidomide-induced teratogenesis: history and mechanisms', *Birth defects research. Part C, Embryo today : reviews*. 2015/06/04. John Wiley and Sons Inc., 105(2), pp. 140–156. doi: 10.1002/bdrc.21096.

Wall, R. *et al.* (2010) 'Fatty acids from fish: The anti-inflammatory potential of long-chain omega-3 fatty acids', *Nutrition Reviews*, 68(5), pp. 280–289. doi: 10.1111/j.1753-4887.2010.00287.x.

Watson, J. T. and Sparkman, O. D. (2007) *Introduction to mass spectrometry: instrumentation, applications, and strategies for data interpretation*. John Wiley & Sons.

Westerhuis, J. A. *et al.* (2008) 'Assessment of PLSDA cross validation', *Metabolomics*,

4(1), pp. 81–89. doi: 10.1007/s11306-007-0099-6.

Wheelock, Å. M. and Wheelock, C. E. (2013) 'Trials and tribulations of 'omics data analysis: Assessing quality of SIMCA-based multivariate models using examples from pulmonary medicine', *Molecular BioSystems*, 9(11), pp. 2589–2596. doi: 10.1039/c3mb70194h.

Williard, D. E. *et al.* (2001) 'Docosahexaenoic acid synthesis from n-3 polyunsaturated fatty acids in differentiated rat brain astrocytes.', *Journal of lipid research*, 42(9), pp. 1368–76.

Woodcock, J. (2007) 'The prospects for "personalized medicine" in drug development and drug therapy', *Clinical Pharmacology and Therapeutics*, 81(2), pp. 164–169. doi: 10.1038/sj.clpt.6100063.

Worley, B. and Powers, R. (2012) 'Multivariate Analysis in Metabolomics', *Current Metabolomics*, 1(1), pp. 92–107. doi: 10.2174/2213235X11301010092.

Worley, B. and Powers, R. (2013) 'Multivariate Analysis in Metabolomics.', *Current Metabolomics*, 1(1), pp. 92–107. doi: 10.2174/2213235X11301010092.

Xi, B. *et al.* (2014) 'Statistical Analysis and Modeling of Mass Spectrometry-Based Metabolomics Data', in *Molecular Analysis and Genome Discovery: Second Edition*, pp. 333–353. doi: 10.1007/978-1-4939-1258-2_22.

Xia, J. *et al.* (2013) 'Translational biomarker discovery in clinical metabolomics: An introductory tutorial', *Metabolomics*, 9(2), pp. 280–299. doi: 10.1007/s11306-012-0482-9.

Xiao, Q. *et al.* (2014) 'Sources of variability in metabolite measurements from urinary samples', *PLoS ONE*, 9(5). doi: 10.1371/journal.pone.0095749.

Yamada, S. *et al.* (2015) 'Ulcerative Colitis Presented as Fever and Bloody Diarrhea at Initiation of Dialysis in an Elderly Patient with End-Stage Kidney Disease', *Case reports in medicine*. 2015/11/04. Hindawi Publishing Corporation, 2015, p. 725205. doi: 10.1155/2015/725205.

Yamamoto, H. *et al.* (2009) 'Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables', *Chemometrics and Intelligent Laboratory Systems*. Elsevier B.V., 98(2), pp. 136–142. doi: 10.1016/j.chemolab.2009.05.006.

Yang, C. *et al.* (2015) 'Systematic review: Thalidomide and thalidomide analogues for treatment of inflammatory bowel disease', *Alimentary Pharmacology and Therapeutics*, 41(11), pp. 1079–1093. doi: 10.1111/apt.13181.

Yi, L. *et al.* (2016) 'Chemometric methods in data processing of mass spectrometry-based metabolomics: A review', *Analytica Chimica Acta*. Elsevier Ltd, 914, pp. 17–34. doi: 10.1016/j.aca.2016.02.001.

Yin, P. and Xu, G. (2014) 'Current state-of-the-art of nontargeted metabolomics based

on liquid chromatography-mass spectrometry with special emphasis in clinical applications', *Journal of Chromatography A*. Elsevier B.V., 1374, pp. 1–13. doi: 10.1016/j.chroma.2014.11.050.

Zhang, P. *et al.* (2017) 'Renal medulla is more sensitive to cisplatin than cortex revealed by untargeted mass spectrometry-based metabolomics in rats', *Scientific Reports*. Nature Publishing Group, 7(March), pp. 1–14. doi: 10.1038/srep44804.

Zhang, R. *et al.* (2014) 'Evaluation of mobile phase characteristics on three zwitterionic columns in hydrophilic interaction liquid chromatography mode for liquid chromatography-high resolution mass spectrometry based untargeted metabolite profiling of Leishmania parasites', *Journal of Chromatography A*. Elsevier B.V., 1362, pp. 168–179. doi: 10.1016/j.chroma.2014.08.039.

Zhang, R. *et al.* (2016) 'Metabolomic Profiling of Post-Mortem Brain Reveals Changes in Amino Acid and Glucose Metabolism in Mental Illness Compared with Controls', *Computational and Structural Biotechnology Journal*. The Authors, 14, pp. 106–116. doi: 10.1016/j.csbj.2016.02.003.

Zhou, B. *et al.* (2012) 'LC-MS-based metabolomics.', *Molecular bioSystems*, 8(2), pp. 470–81. doi: 10.1039/c1mb05350g.

Appendix

8 Appendix

8.1 Appendices Chapter 3

Table S.1. 1 Date pre-treatment methods and models parameters

Method number	Data pre-treatment			Models								
				PCA		OPLS-DA						
	Missing data imputation	Transformation	Scaling	R^2X	Q^2	R^2X (Cum)	R^2	Q^2	*p* CV-ANOVA	R^2-Q^2	Permutation (999 times)	Valid
Method 2	NIPALS	-	Par	0.65	0.39	0.51	0.69	0.56	$2.12E^{-08}$	0.13	Yes	Yes
Method 3	NIPALS	Log_{10}	Par	0.7	0.48	0.53	0.8	0.63	$1.98E^{-09}$	0.17	Yes	Yes
Method 5	NIPALS	Log_{10}	UV	0.71	0.5	0.51	0.79	0.62	$4.78E^{-09}$	0.17	Yes	Yes
Method 4	NIPALS	-	UV	0.69	0.41	0.37	0.76	0.59	$4.41E^{-09}$	0.17	Yes	Yes
Method 1	NIPALS	Power	UV	0.65	0.05	0.23	0.65	0.45	$1.17E^{-07}$	0.2	Yes	Yes
Method 7	Median	Power	UV	0.39	0.05	0.18	0.64	0.34	$2.98E^{-06}$	0.3	Yes	No
Method 8	NIPALS	Log_{10}	Center	0.71	0.46	0.52	0.89	0.57	$2.74E^{-07}$	0.32	Yes	No
Method 6	Mean	Power	UV	0.39	0.06	0.18	0.65	0.31	$2.73E^{-06}$	0.34	Yes	No
Method 10	KNN	-	Par	0.65	0.25	0.4	0.72	0.35	$1.97E^{-05}$	0.37	Yes	No
Method 9	Median	-	Par	0.67	0.24	0.45	0.84	0.47	$2.28E^{-07}$	0.37	Yes	No
Method 11	Mean	-	Par	0.67	0.26	0.45	0.84	0.46	$3.50E^{-07}$	0.38	Yes	No
Method 13	Mean	-	UV	0.55	0.2	0.28	0.83	0.41	$6.50E^{-07}$	0.41	Yes	No
Method 12	Median	-	UV	0.55	0.19	0.28	0.82	0.41	$6.37E^{-07}$	0.41	Yes	No

Method 14	Min/2	Power	UV	0.49	-0.05	0.2	0.82	0.35	$1.97E^{-05}$	0.47	Yes	No
Method 15	KNN	-	UV	0.54	0.19	0.3	0.92	0.44	$1.32E^{-06}$	0.48	Yes	No
Method 16	Mean	Log_{10}	UV	0.57	0.29	0.46	0.97	0.46	$4.02E^{-06}$	0.51	Yes	No
Method 17	KNN	Power	UV	0.53	0.02	0.24	0.91	0.38	$2.62E^{-05}$	0.53	Yes	No
Method 19	Min/2	-	Par	0.5	0.12	0.32	0.77	0.25	$1.79E^{-03}$	0.53	Yes	No
Method 18	Median	Log_{10}	UV	0.59	0.33	0.46	0.96	0.44	$9.15E^{-06}$	0.53	Yes	No
Method 21	KNN	Log_{10}	UV	0.56	0.28	0.44	0.97	0.42	$2.80E^{-05}$	0.55	Yes	No
Method 20	Mean	Log_{10}	Par	0.56	0.28	0.43	0.97	0.42	$2.26E^{-05}$	0.55	Yes	No
Method 23	Min/2	-	UV	0.38	0.09	0.23	0.94	0.38	$2.78E^{-05}$	0.56	Yes	No
Method 22	Median	Log_{10}	Par	0.59	0.3	0.45	0.97	0.41	$4.66E^{-05}$	0.56	Yes	No
Method 24	KNN	Log_{10}	Par	0.57	0.28	0.42	0.96	0.37	$2.80E^{-04}$	0.59	Yes	No
Method 25	Mean	Log_{10}	Center	0.58	0.24	0.4	0.97	0.37	$2.54E^{-04}$	0.6	Yes	No
Method 26	Median	Log_{10}	Center	0.58	0.3	0.44	0.99	0.38	$6.95E^{-04}$	0.6	Yes	No
Method 27	KNN	Log_{10}	Center	0.58	0.25	0.37	0.92	0.28	$2.04E^{-03}$	0.64	Yes	No
Method 44	NIPALS	-	-	0.98	0.66	0.95	0.2	0.04	N.S	N.A	No	No
Method 59	NIPALS	Log_{10}	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 56	NIPALS	-	Center	0.91	0.48	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 35	NIPALS	Power	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 58	NIPALS	Power	Center	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 60	NIPALS	Power	Par	0.92	0.52	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 31	KNN	-	-	0.98	0.58	0.98	0.54	0.23	N.S	N.A	N.A	No
Method 38	KNN	-	Center	0.89	0.33	0.85	0.51	0.17	N.S	N.A	N.A	No
Method 33	KNN	Power	Par	0.65	0.25	0.87	0.48	0.1	N.S	N.A	N.A	No
Method 41	KNN	Log_{10}	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 43	KNN	Power	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A

Method 45	KNN	Power	Center	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 32	Min/2	Log_{10}	Par	0.14	0.05	0.14	1	0	N.S	N.A	N.A	No
Method 39	Min/2	Log_{10}	Center	0.15	0.06	0.12	0.99	-0.02	N.S	N.A	N.A	No
Method 37	Min/2	Log_{10}	UV	0.14	0.04	0.12	0.98	0	N.S	N.A	N.A	No
Method 42	Min/2	-	-	0.99	0.56	0.93	0.25	0.04	N.S	N.A	N.A	No
Method 40	Min/2	Power	Par	0.82	0.27	0.74	0.02	-0.02	N.S	N.A	N.A	No
Method 28	Min/2	Log_{10}	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 36	Min/2	-	Center	0.84	0.25	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 30	Min/2	Power	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 57	Min/2	Power	Center	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 48	Mean	Log_{10}	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 50	Mean	-	-	0.98	0.62	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 46	Mean	-	Center	0.89	0.38	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 47	Mean	Power	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 49	Mean	Power	Center	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 51	Mean	Power	Par	0.9	0.36	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 29	Median	Log_{10}	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 53	Median	-	-	0.98	0.61	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 55	Median	-	Center	0.89	0.38	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 52	Median	Power	-	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 34	Median	Power	Center	D.F	D.F	D.F	D.F	D.F	N.A	N.A	N.A	N.A
Method 54	Median	Power	Par	0.9	0.36	D.F	D.F	D.F	N.A	N.A	N.A	N.A

8.2 Appendices for chapter 4

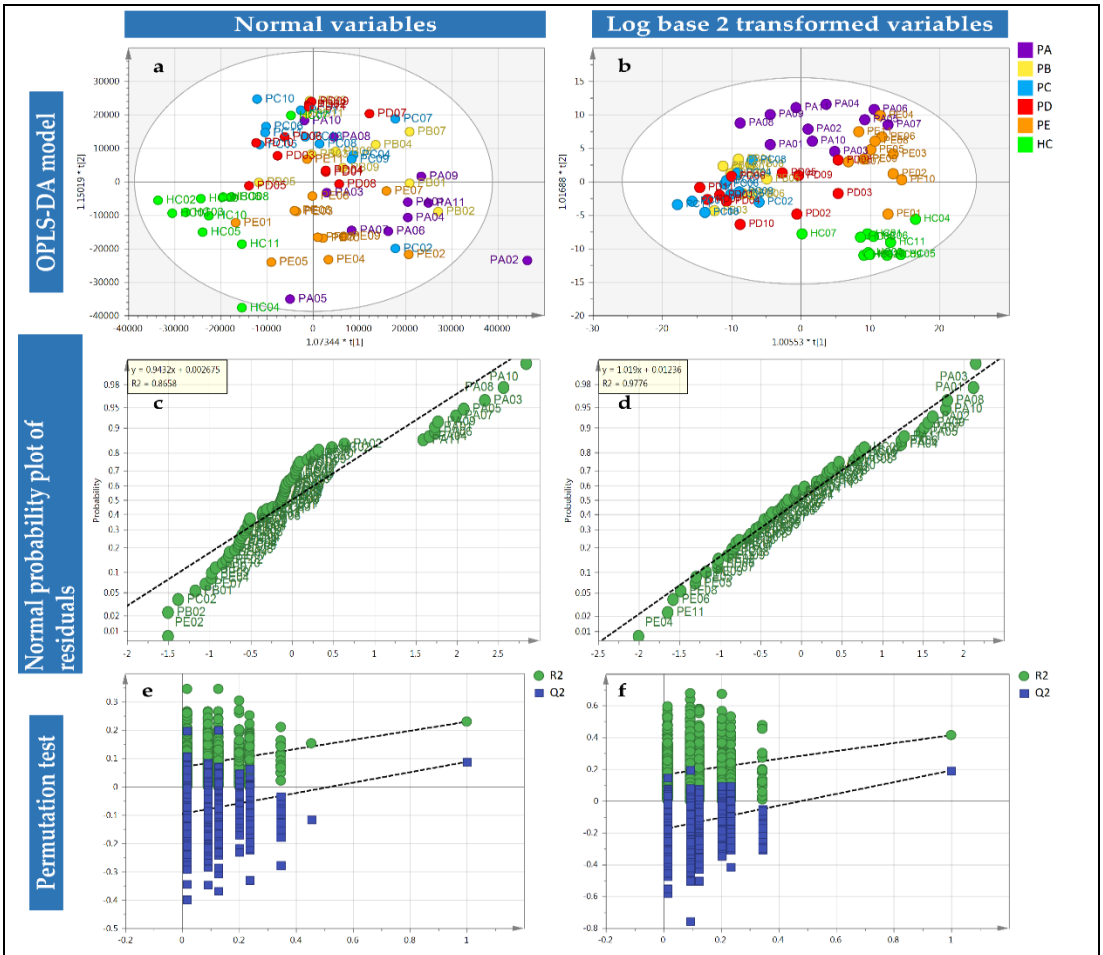
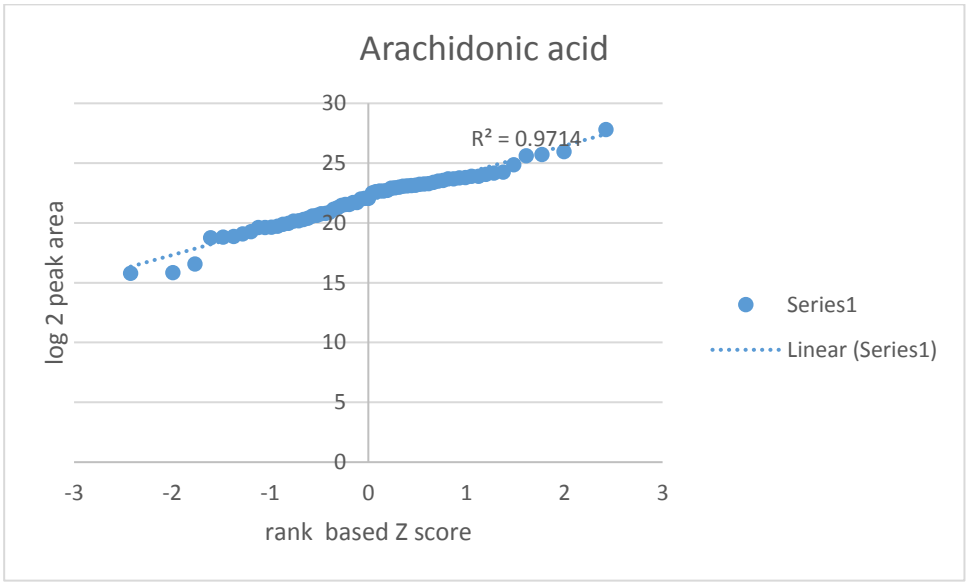
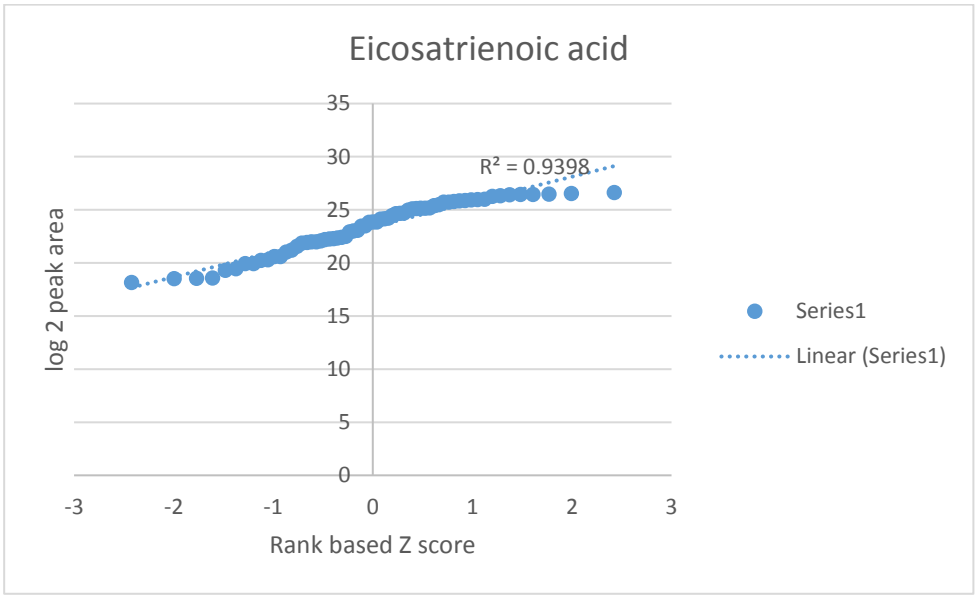


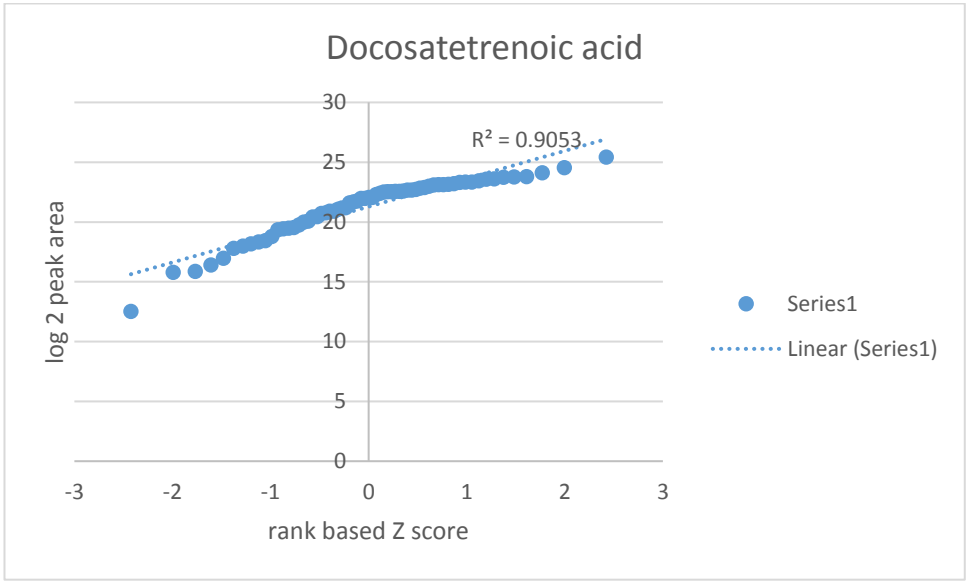
***Figure S4. 1 The effect of Log base 2 transformed variables on the model separation and validation**:*

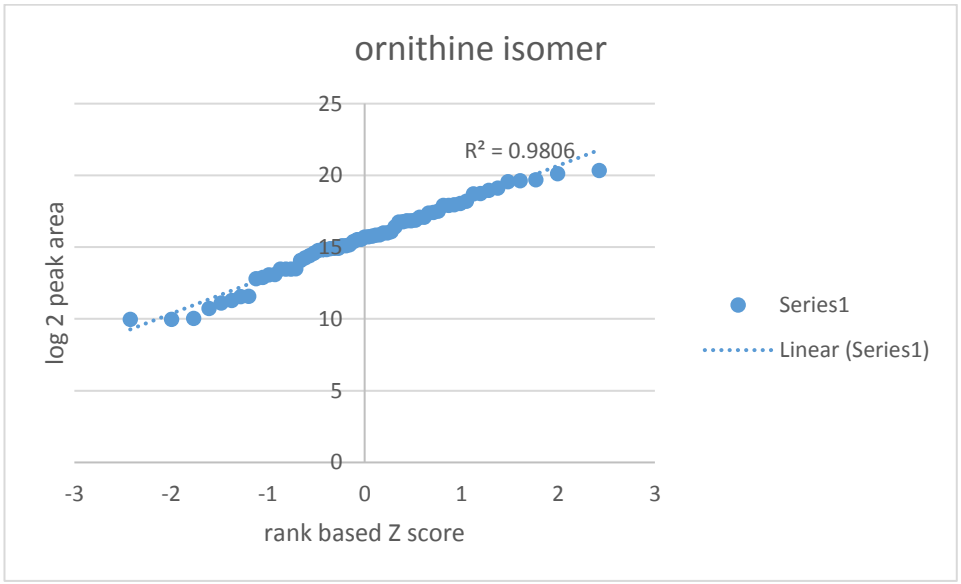
(a) The 2D Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) model score plot for the normal variables; and (b) a plot of the Log₂ transformed variables. All the variables were Pareto scaled, and the OPLS-DA model was based on 376 putative metabolites. Sample groups: (HC) HC children, (PA) CD children pre-EEN treatment, (PB) CD children 15 days during EEN treatment, (PC) CD children 30 days during EEN treatment, (PD) CD children 60 days during EEN treatment and (PE) CD children back to their habitual free diet. The data was further analysed with: (c) a normal probability plot of the residuals for the normal variables; and (d) a plot of the Log₂ transformed variables. Plot c and d display the residuals standardized on a double Log scale along the y-axis versus the standard deviation on the x-axis. Outlier variables are displayed outside -4 to 4 standard deviation intervals. The regression line assessed the normality of the residuals. A permutation test (999 times) was run for the OPLS-DA model based on all the samples: (e) the plot of the normal variables; and (f) the plot of the Log₂

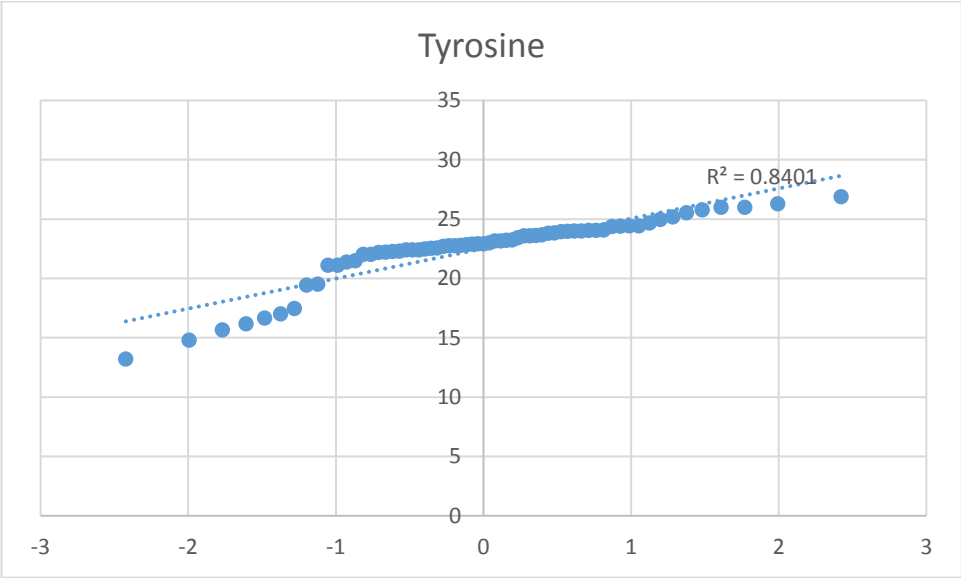
Figure S4. 2 Q-Q Plots for the marker compounds and some compounds reported in table Table S4. 1 as significant and confirming normal distribution in some cases but not others

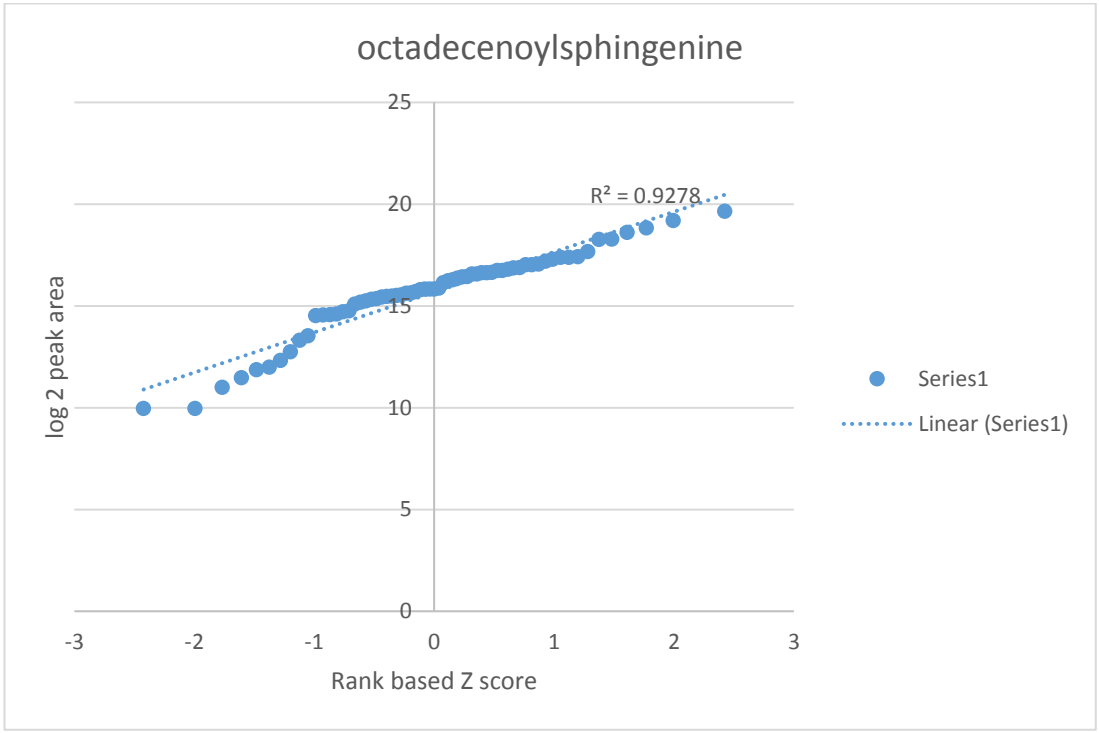


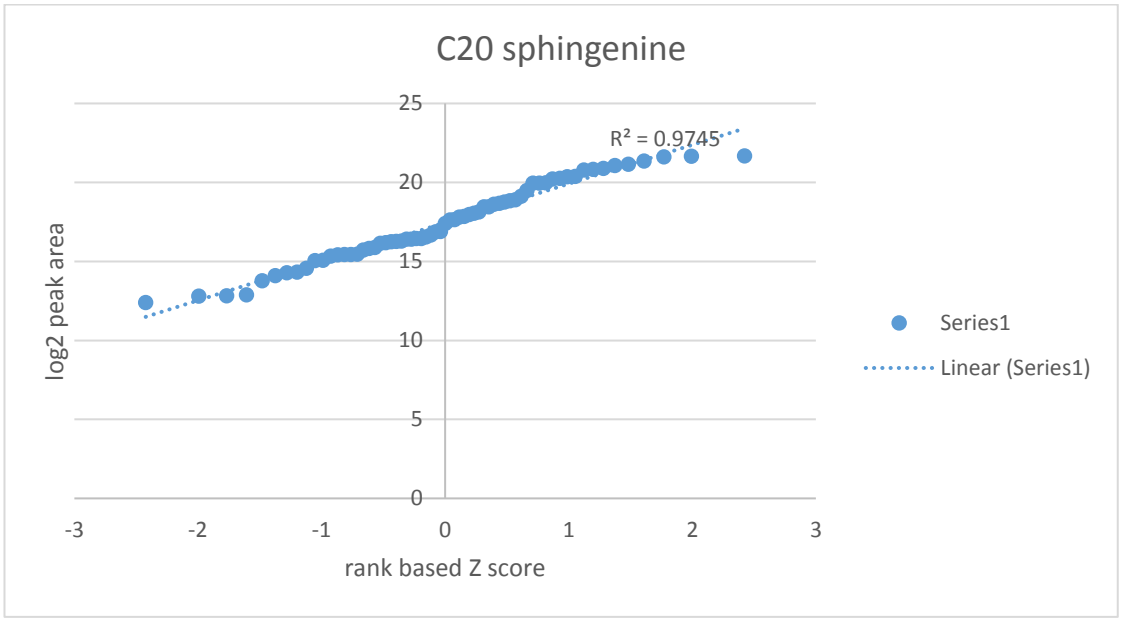


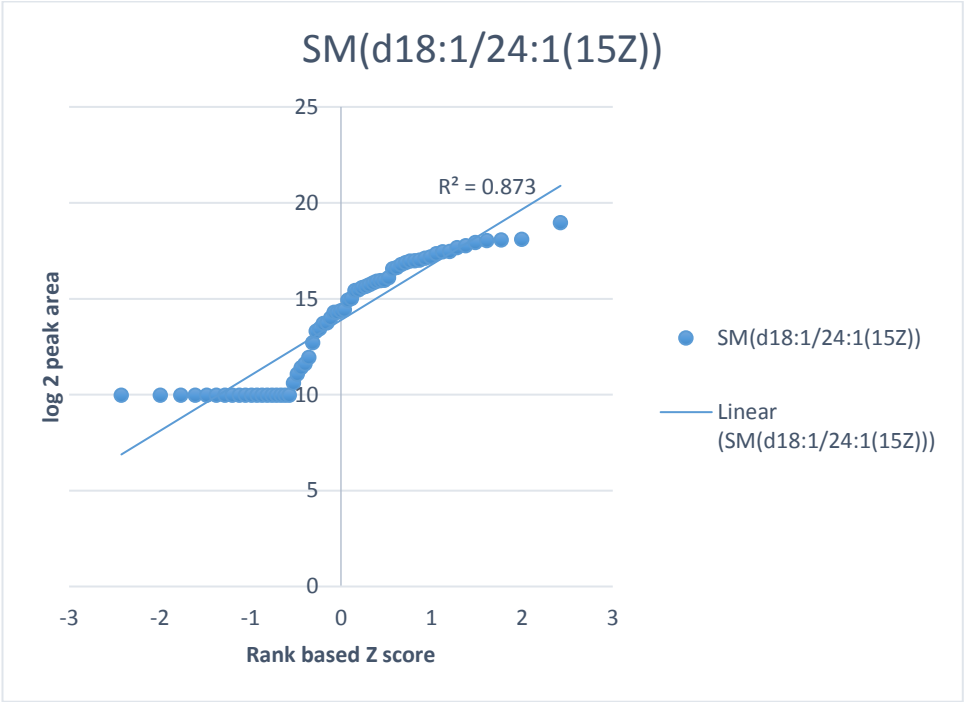

Docosatetrenoic acid
R² = 0.9053
log 2 peak area
rank based Z score
Series1
Linear (Series1)


ornithine isomer
R² = 0.9806
log 2 peak area
rank based Z score
Series1
Linear (Series1)


Tyrosine
35
30
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
R² = 0.8401


octadecenoylsphingenine
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
R² = 0.9278
log 2 peak area
Rank based Z score
Series1
Linear (Series1)


C20 sphingenine
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
R² = 0.9745
log2 peak area
rank based Z score
Series1
Linear (Series1)


SM(d18:1/24:1(15Z))
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
R² = 0.873
log 2 peak area
Rank based Z score
SM(d18:1/24:1(15Z))
Linear (SM(d18:1/24:1(15Z)))

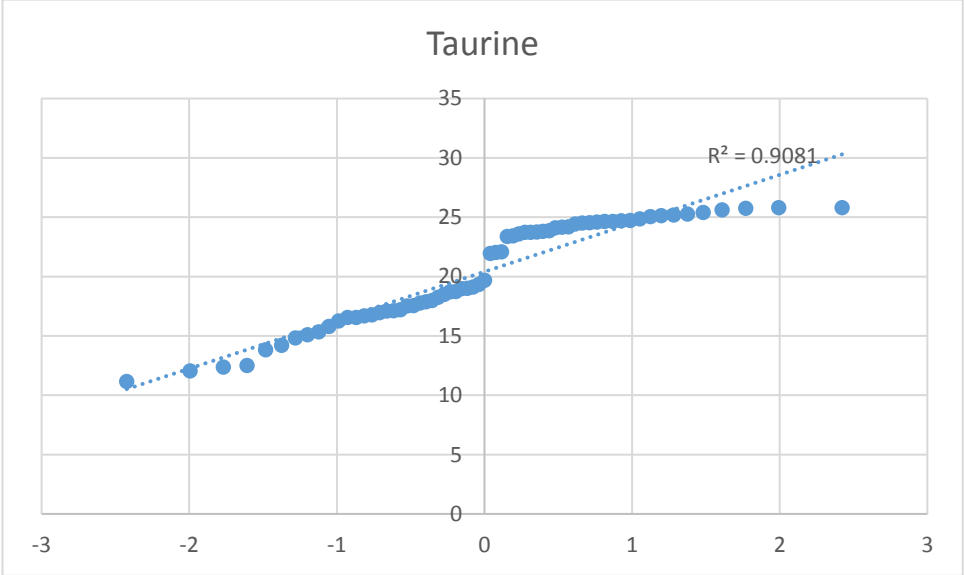

Taurine
35
30
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
R² = 0.9081

Table S4. 1 Small polar marker compounds for CD vs HC. * Corresponds to the retention time of a standard.

Mass	RT	Putative metabolite	P value HC/PA	PA/HC	P value HC/PB	PB/PC	P value HC/PC	PC/HC	p value HC/PD	PD/HC	p value HE/PC	PE/HC
75.03215	16.2	*Glycine	0.111	3.885	0.399	1.804	0.755	1.234	0.565	1.171	0.415	1.268
88.01596	8.3	*Pyruvate	0.694	0.727	0.319	0.338	0.211	0.169	0.217	0.183	0.951	0.954
88.05236	6.6	Butanoic acid	0.356	1.880	0.853	0.901	0.161	0.422	0.151	0.415	0.406	1.694
89.04762	15.3	*Alanine	0.089	3.464	0.212	2.170	0.769	1.176	0.640	0.884	0.091	1.646
89.04774	15.9	*beta-Alanine	0.028	5.995	0.140	4.787	0.273	2.699	0.362	2.325	0.097	5.351
92.04734	10.7	Glycerol	0.112	0.191	0.096	0.151	0.070	0.066	0.078	0.093	0.107	0.182
97.96741	8.8	Sulfate	0.459	1.412	0.436	1.343	0.442	1.389	0.750	0.900	0.053	2.214
103.0633	14.3	N,N-Dimethylglycine	0.029	4.109	0.096	4.191	0.102	2.577	0.057	2.938	0.015	4.437
103.0633	12.8	*3-Amino-isobutanoate	0.006	3.129	0.246	1.390	0.457	1.218	0.615	1.163	0.038	1.549
103.0634	16.1	*3-Amino-isobutanoate	0.527	0.739	0.463	0.629	0.046	0.185	0.044	0.180	0.158	2.253
104.011	8.0	Hydroxypyruvate	0.324	12.871	0.030	11.386	0.035	7.553	0.044	6.936	0.146	6.110
104.0474	7.6	4-Hydroxybutanoic acid	0.291	15.200	0.069	4.222	0.326	1.723	0.610	0.788	0.038	2.717
105.0427	16.3	*Serine	0.187	7.201	0.199	2.962	0.285	0.731	0.944	0.980	0.173	1.618
109.0528	10.5	2-Aminophenol	0.950	1.027	0.001	0.312	0.001	0.298	0.976	1.009	0.009	2.053
109.0528	7.7	4-Hydroxyaniline	0.258	124.161	0.372	12.058	0.236	27.380	0.309	52.264	0.268	82.550
111.032	8.0	Pyrrole-2-carboxylate	0.244	0.438	0.149	0.304	0.126	0.256	0.792	0.829	0.520	1.311
111.032	9.8	Pyrrole-2-carboxylate	0.069	0.214	0.044	0.115	0.041	0.096	0.985	0.982	0.698	0.785
112.0273	8.2	Uracil	0.185	3.869	0.012	6.376	0.086	5.519	0.019	3.647	0.396	1.470
112.0273	8.8	Uracil	0.765	1.124	0.041	0.456	0.001	0.215	0.018	0.434	0.622	0.854
113.0477	10.8	1-Pyrroline-2-carboxylate	0.087	1.753	0.015	3.122	0.012	2.828	0.024	3.286	0.209	1.183
113.0478	16.4	(S)-1-Pyrroline-5-carboxylate	0.016	2.333	0.159	1.678	0.988	1.004	0.225	1.452	0.027	1.732
114.0318	15.0	2-Hydroxy-2,4-pentadienoate	0.174	2.606	0.856	0.904	0.111	0.375	0.208	0.509	0.154	1.720

114.043	9.4	5,6-Dihydrouracil	0.407	1.275	0.839	1.079	0.335	0.755	0.785	0.925	0.152	1.379
115.0269	8.1	Maleamate	0.800	0.930	0.625	0.873	0.350	0.741	0.401	0.779	0.082	2.041
115.0269	9.5	Maleamate	0.207	1.716	0.101	2.092	0.200	1.613	0.315	1.801	0.006	2.565
115.0269	10.7	Maleamate	0.142	13.497	0.108	3.270	0.091	2.270	0.322	3.582	0.024	3.793
115.0633	13.3	*Proline	0.134	2.731	0.231	2.617	0.491	1.457	0.773	1.115	0.521	1.202
115.0633	9.0	3-acetamidopropanal	0.133	3.748	0.829	0.908	0.799	0.878	0.604	0.788	0.024	3.095
115.0634	16.5	Proline	0.138	4.628	0.253	4.269	0.399	1.928	0.363	1.319	0.008	1.673
117.0426	8.0	L-2-Amino-3-oxobutanoic acid	0.114	3.607	0.045	2.314	0.479	1.191	0.081	1.679	0.049	2.198
117.0579	10.7	Indole	0.017	4.495	0.094	3.169	0.149	1.859	0.076	1.954	0.055	2.180
117.0579	6.2	Indole	0.330	1.478	0.877	0.942	0.603	1.231	0.835	1.074	0.143	1.806
117.0789	13.0	[FA amino(5:0)] 2S-amino-pentanoic acid	0.027	2.649	0.225	1.626	0.982	1.008	0.367	1.292	0.129	1.564
117.0789	16.3	5-Aminopentanoate	0.017	3.414	0.089	2.193	0.242	1.663	0.373	1.401	0.017	3.043
117.0789	12.4	* Valine	0.047	2.205	0.425	1.345	0.584	0.833	0.736	1.097	0.647	1.136
117.079	11.9	* Betaine	0.037	0.382	0.002	0.084	0.002	0.055	0.005	0.185	0.454	1.324
118.063	7.8	5-Hydroxypentanoate	0.966	1.029	0.282	2.119	0.896	0.902	0.574	0.663	0.438	0.573
118.063	5.2	formyl 3-hydroxy-butanoate	0.122	2.609	0.161	1.951	0.529	1.434	0.866	1.096	0.756	1.148
119.0583	14.9	*Threonine	0.165	8.056	0.177	3.783	0.585	1.413	0.407	1.346	0.066	1.736
122.0368	7.8	Benzoate	0.146	2.946	0.189	6.681	0.142	3.795	0.346	2.525	0.097	4.791
122.048	7.9	*Nicotinamide	0.057	5.263	0.165	5.236	0.264	2.131	0.331	2.433	0.110	1.731
123.032	7.8	*Nicotinate	0.826	0.858	0.547	1.515	0.396	0.444	0.520	0.576	0.772	1.203
123.0321	13.6	Nitrobenzene	0.823	1.085	0.215	3.501	0.116	2.650	0.296	1.875	0.153	2.265
125.0146	15.0	*Taurine	0.002	10.929	0.052	5.090	0.109	3.540	0.544	1.678	0.049	4.579
125.0589	10.6	5-Methylcytosine	0.207	7.537	0.023	4.033	0.123	3.947	0.046	4.868	0.009	4.949
126.0429	8.3	Thymine	0.034	8.118	0.110	2.911	0.571	1.296	0.362	1.578	0.030	3.857

126.0429	7.8	Thymine	0.274	1.780	0.284	0.644	0.010	0.276	0.021	0.349	0.613	0.828
127.0633	14.0	2,3,4,5-Tetrahydropyridine-2-carboxylate	0.010	2.675	0.008	2.518	0.091	1.722	0.041	2.215	0.013	2.433
128.0585	9.6	gamma-Amino-gamma-cyanobutanoate	0.039	2.431	0.188	3.074	0.258	3.088	0.180	2.639	0.007	2.125
128.0585	15.4	5,6-Dihydrothymine	0.255	18.902	0.299	4.904	0.150	3.277	0.324	1.670	0.427	1.469
129.0426	8.3	Oxoproline	0.758	1.158	0.800	1.205	0.967	1.028	0.325	0.755	0.226	0.695
129.0789	12.8	N4-Acetylaminobutanal	0.827	0.933	0.034	0.435	0.071	0.495	0.108	0.552	0.724	0.907
129.079	11.6	L-Pipecolate	0.492	0.750	0.005	0.142	0.004	0.130	0.039	0.344	0.341	1.416
129.0791	12.0	N4-Acetylaminobutanal	0.933	1.032	0.030	0.331	0.041	0.348	0.077	0.414	0.222	1.731
130.063	4.6	4-Methyl-2-oxopentanoate	0.338	4.148	0.212	2.370	0.692	0.827	0.606	1.245	0.746	1.188
131.0582	9.5	N-Acetyl-beta-alanine	0.147	3.701	0.156	1.622	0.646	1.150	0.735	1.109	0.099	1.568
131.0582	14.9	N-Acetyl-beta-alanine	0.525	1.348	0.179	0.494	0.018	0.392	0.000	0.175	0.177	2.154
131.0694	15.4	*Creatine	0.081	62.697	0.276	11.737	0.165	30.663	0.141	1.776	0.319	23.801
131.0945	11.8	*Leucine	0.056	3.832	0.146	2.322	0.479	1.413	0.109	1.726	0.094	1.686
131.0946	11.4	*Isoleucine	0.068	4.854	0.131	3.337	0.484	1.730	0.153	1.647	0.126	2.247
132.0423	8.1	2-Acetolactate	0.535	1.349	0.960	1.027	0.273	0.587	0.552	0.765	0.191	1.670
132.0533	11.9	N-Carbamoylsarcosine	0.149	9.585	0.215	6.575	0.626	1.555	0.975	1.012	0.035	1.937
132.0534	8.4	3-Ureidopropionate	0.330	6.095	0.397	1.646	0.341	0.688	0.763	1.132	0.490	1.219
132.0535	15.6	*Asparagine	0.253	33.837	0.303	8.977	0.328	15.753	0.268	2.507	0.231	2.931
132.0786	4.1	hydroxy-isocaproic acid	0.265	2.789	0.507	1.682	0.531	1.795	0.646	0.800	0.706	0.867
132.0898	8.7	N4-acetyl-N4-hydroxy-1-aminopropane	0.010	0.122	0.114	0.374	0.038	0.269	0.053	0.326	0.021	0.241
133.0375	9.4	2-hydroxysuccinamate	0.577	1.319	0.588	1.343	0.386	0.730	0.976	1.010	0.037	1.957
133.0738	16.1	N-hydroxyvaline	0.635	1.212	0.691	1.268	0.862	1.097	0.621	1.350	0.010	2.364

134.0215	8.3	3-Dehydro-L-threonate	0.590	1.432	0.130	4.219	0.494	1.631	0.494	1.471	0.079	3.753
135.0545	9.5	*Adenine	0.162	0.419	0.619	0.742	0.290	0.510	0.120	0.362	0.658	0.791
135.0545	10.0	Adenine	0.124	0.328	0.041	0.138	0.035	0.106	0.041	0.136	0.097	0.316
136.0385	10.4	*Hypoxanthine	0.635	1.187	0.008	0.300	0.007	0.288	0.004	0.217	0.489	0.809
136.0525	4.2	4-Hydroxyphenylacetaldehyde	0.480	0.804	0.395	1.383	0.825	0.933	0.734	0.895	0.651	1.131
137.084	7.7	Tyramine	0.128	1.796	0.007	3.217	0.131	2.323	0.086	2.073	0.114	1.452
138.043	8.0	*Urocanate	0.025	4.703	0.050	5.256	0.203	3.748	0.090	2.842	0.080	2.386
139.0745	8.9	L-Histidinal	0.272	0.516	0.272	0.549	0.076	0.272	0.882	0.915	0.968	1.030
140.0586	8.0	Methylimidazoleacetic acid	0.188	1.916	0.622	1.398	0.063	0.371	0.448	0.659	0.241	1.786
140.9829	13.3	Carbamoyl phosphate	0.363	11.829	0.280	33.627	0.184	42.039	0.080	12.353	0.217	4.610
140.9829	15.8	Carbamoyl phosphate	0.368	3.869	0.275	12.379	0.099	16.090	0.147	4.818	0.216	2.623
140.9829	21.6	Carbamoyl phosphate	0.108	5.341	0.094	6.723	0.018	9.973	0.104	5.267	0.047	5.467
142.0742	12.5	Ectoine	0.606	0.586	0.497	0.455	0.477	0.425	0.522	0.478	0.482	1.625
142.0743	13.3	Ectoine	0.536	0.587	0.434	0.472	0.495	0.540	0.505	0.546	0.497	1.497
142.0743	14.4	Ectoine	0.021	3.553	0.115	1.920	0.807	0.930	0.340	1.401	0.334	1.324
143.0946	10.7	Stachydrine	0.315	0.203	0.258	0.098	0.284	0.145	0.343	0.246	0.467	0.417
145.0739	8.6	[FA oxo,amino(6:0)] 3-oxo-5S-amino-hexanoic acid	0.016	0.373	0.001	0.132	0.001	0.094	0.008	0.306	0.828	0.950
145.0739	8.1	6-Amino-2-oxohexanoate	0.354	0.675	0.448	0.670	0.035	0.316	0.089	0.453	0.161	0.571
145.0739	13.6	4-Acetamidobutanoate	0.708	0.720	0.135	0.025	0.135	0.026	0.148	0.059	0.587	1.426
145.0851	15.9	4-Guanidinobutanoate	0.844	0.927	0.396	0.698	0.500	0.745	0.610	0.783	0.186	2.267
145.1102	13.9	*Acetylcholine	0.003	2.616	0.017	2.420	0.170	1.610	0.042	2.327	0.014	2.508
146.0691	16.1	*Glutamine	0.317	2.982	0.547	1.392	0.018	0.462	0.370	0.766	0.653	1.173
146.0691	11.3	Glutamine isomer	0.070	6.627	0.126	5.514	0.386	1.720	0.206	1.771	0.080	1.645
146.0692	8.0	3-Ureidoisobutyrate	0.092	3.565	0.156	7.445	0.157	6.204	0.506	1.355	0.521	1.270

146.0692	15.5	Glutamine	0.078	20.231	0.078	14.337	0.152	7.519	0.004	5.379	0.000	5.790
147.0321	6.0	Indole-5,6-quinone	0.051	3.762	0.075	4.271	0.014	3.337	0.100	2.219	0.127	1.513
147.0532	9.2	*O-Acetylserine	0.245	1.597	0.520	1.306	0.854	1.078	0.519	0.826	0.344	1.238
147.0895	9.8	N-hydroxyisoleucine	0.198	1.888	0.150	1.755	0.207	1.839	0.039	2.806	0.064	2.133
147.0895	8.5	N-hydroxyisoleucine	0.006	0.307	0.001	0.206	0.002	0.263	0.011	0.422	0.004	0.337
148.0372	8.1	D-Arabinono-1,4-lactone	0.311	0.598	0.717	0.851	0.500	0.736	0.939	0.970	0.329	1.572
148.0734	9.0	(R)-2,3-Dihydroxy-3-methylpentanoate	0.925	0.969	0.243	0.732	0.274	0.733	0.830	0.937	0.052	1.607
148.0736	7.8	3R-methyl-3,5-dihydroxy-pentanoic acid	0.663	1.255	0.483	1.466	0.706	0.826	0.547	0.722	0.348	1.795
150.0527	12.2	Ribose or isomer	0.688	0.854	0.057	0.486	0.003	0.195	0.004	0.231	0.365	0.731
150.0527	15.3	Xylose or isomer	0.410	1.785	0.558	0.732	0.055	0.293	0.096	0.397	0.364	1.376
150.0527	13.7	Arabinose or isomer	0.141	0.391	0.037	0.142	0.030	0.096	0.045	0.178	0.628	0.770
151.0633	13.4	N-Methylanthranilate	0.304	0.769	0.009	0.464	0.001	0.370	0.065	0.625	0.860	0.970
151.0633	9.3	Paracetamol	0.115	0.556	0.001	0.212	0.000	0.152	0.004	0.321	0.979	1.007
152.0473	7.8	4-Hydroxyphenylacetate	0.139	5.531	0.248	20.262	0.102	7.771	0.265	5.935	0.148	11.515
152.0684	13.3	Xylitol or isomer	0.301	1.969	0.765	1.364	0.792	1.288	0.263	0.557	0.750	0.868
153.0426	13.3	Hydroxymethylpyridinecarboxylate	0.082	49.071	0.293	50.708	0.071	283.970	0.284	591.855	0.093	119.934
153.0426	7.9	Hydroxyanthranilate	0.213	901.060	0.326	1065.093	0.262	778.568	0.148	951.672	0.150	766.549
153.0789	10.4	Dopamine	0.022	0.421	0.001	0.145	0.000	0.116	0.041	0.469	0.935	0.982
154.0266	8.2	2,5-Dihydroxybenzoate	0.085	0.180	0.046	0.033	0.045	0.031	0.050	0.054	0.780	0.838
154.0378	12.2	Imidazol-5-yl-pyruvate	0.245	1.888	0.302	0.650	0.056	0.424	0.057	0.442	0.607	1.521
155.0695	15.8	Histidine	0.218	17.100	0.259	9.043	0.370	4.893	0.986	1.006	0.115	5.046
156.0535	8.0	Imidazolonepropanoate	0.279	1.428	0.509	0.828	0.300	0.727	0.743	1.152	0.355	2.587
158.0942	4.9	oxo-octanoic acid	0.317	0.715	0.113	2.127	0.035	2.378	0.045	2.414	0.197	0.702

159.0895	13.4	3-Dehydrocarnitine	0.599	0.580	0.476	0.428	0.402	0.328	0.561	0.534	0.816	0.809
161.0477	8.6	4,8-Dihydroxyquinoline	0.407	1.413	0.207	1.510	0.763	1.099	0.577	1.171	0.919	1.042
161.0687	15.1	Aminoadipate	0.037	3.022	0.798	1.128	0.867	1.099	0.182	2.429	0.028	5.189
161.0688	9.6	O-Acetylhomoserine	0.278	2.059	0.525	1.622	0.490	1.423	0.289	1.676	0.028	2.159
161.0688	11.5	N-Methyl-L-glutamate	0.005	2.658	0.611	1.208	0.560	1.366	0.887	1.049	0.005	2.558
161.1051	13.8	*Carnitine	0.030	6.274	0.025	6.175	0.009	8.092	0.097	4.771	0.307	2.121
161.1052	12.5	Carnitine isomer	0.012	0.348	0.001	0.220	0.002	0.249	0.034	0.490	0.175	1.935
162.1003	15.2	N6-Hydroxy-L-lysine	0.085	2.061	0.007	4.250	0.022	3.611	0.048	4.617	0.774	0.898
163.0667	10.6	homomethionine	0.631	0.814	0.413	0.705	0.263	0.612	0.903	1.051	0.150	1.604
164.0685	11.5	Rhamnose or isomer	0.264	1.893	0.198	0.630	0.478	0.719	0.095	0.516	0.140	1.958
164.0686	7.9	Rhamnose or isomer	0.922	0.945	0.613	0.727	0.792	0.854	0.838	1.180	0.525	1.430
165.046	13.7	L-Methionine S-oxide	0.218	2.495	0.635	1.280	0.445	0.762	0.991	0.996	0.319	1.584
165.079	10.7	*Phenylalanine	0.038	8.696	0.141	6.321	0.237	2.429	0.069	2.000	0.113	2.620
166.049	9.9	Methylxanthine	0.372	0.715	0.021	0.324	0.012	0.264	0.182	0.582	0.110	0.565
166.0492	8.3	Methylxanthine	0.693	0.722	0.270	0.258	0.319	0.324	0.487	0.521	0.902	1.091
166.063	5.1	Phenyllactate	0.395	2.677	0.794	0.789	0.325	0.359	0.459	0.489	0.585	1.533
167.0582	11.5	Methoxyanthranilate	0.055	0.527	0.456	0.804	0.680	0.885	0.606	1.166	0.175	0.736
167.0583	5.2	Isopyridoxal	0.985	0.991	0.569	0.694	0.492	0.700	0.598	1.376	0.127	2.343
167.0583	8.0	Pyridoxal	0.007	0.420	0.002	0.296	0.001	0.220	0.490	0.757	0.082	1.936
169.0739	12.4	Noradrenaline	0.030	0.494	0.099	0.623	0.066	0.585	0.993	1.003	0.563	0.874
172.0484	9.4	Hydantoin-5-propionate	0.472	0.746	0.882	1.059	0.840	1.088	0.914	0.960	0.628	0.855
172.0484	8.0	Hydantoin-5-propionate	0.021	3.464	0.010	3.986	0.016	3.104	0.062	2.591	0.005	2.246
173.0801	13.4	Guanidinoxopentanoate	0.008	0.175	0.003	0.057	0.002	0.000	0.658	0.741	0.478	1.366
174.0641	9.4	N-Formimino-L-glutamate	0.188	2.706	0.718	1.169	0.833	1.094	0.277	1.696	0.029	2.693

174.0892	5.2	Suberic acid	0.453	0.705	0.104	0.445	0.100	0.442	0.243	0.608	0.804	0.896
175.048	8.0	aminooxohexanedioic acid	0.261	0.577	0.344	0.670	0.251	0.568	0.573	0.771	0.724	1.172
175.0633	4.6	N-Acetylindoxyl	0.336	5.600	0.389	3.940	0.197	2.928	0.345	13.593	0.401	2.048
175.0633	5.5	*Indoleacetate	0.350	4.157	0.468	2.560	0.394	1.797	0.372	5.354	0.588	1.499
175.0956	16.5	*Citrulline	0.023	2.647	0.141	2.309	0.621	1.166	0.190	1.498	0.089	1.734
179.0582	7.9	Hippurate	0.505	1.333	0.581	1.342	0.906	1.053	0.829	1.090	0.013	2.104
179.0793	17.0	Glucosamine or isomer	0.410	1.566	0.975	0.979	0.118	0.397	0.394	0.649	0.169	1.808
179.0793	15.7	Glucosamine or isomer	0.910	0.954	0.023	0.236	0.013	0.146	0.114	0.469	0.009	2.477
180.0634	15.2	Glucose or isomer	0.214	2.583	0.910	0.933	0.104	0.311	0.170	0.428	0.263	1.570
180.0898	13.9	Hydroxykynurenamine	0.018	0.457	0.003	0.310	0.002	0.296	0.064	0.549	0.208	1.341
181.0739	13.5	Hydroxyphenylpyruvate	0.034	6.924	0.067	4.412	0.129	2.404	0.036	2.271	0.028	2.524
181.0739	12.3	*Tyrosine	0.006	0.372	0.007	0.383	0.003	0.323	0.041	0.533	0.287	0.778
181.0739	11.5	3-Amino-3-(4-hydroxyphenyl)propanoate	0.028	0.459	0.016	0.403	0.005	0.301	0.044	0.510	0.296	0.772
182.058	7.8	Hydroxyphenyllactate or isomer	0.931	0.953	0.083	0.160	0.099	0.200	0.079	0.146	0.582	0.724
182.058	5.2	Hydroxyphenyllactate or isomer	0.292	0.576	0.055	0.212	0.048	0.183	0.074	0.275	0.474	0.721
182.0792	14.4	Mannitol	0.341	45.164	0.295	1.912	0.799	1.088	0.990	1.005	0.577	1.175
183.0533	8.0	4-Pyridoxate	0.027	0.466	0.001	0.292	0.001	0.284	0.037	0.493	0.231	0.734
183.0896	11.3	Adrenaline	0.031	0.409	0.001	0.141	0.001	0.102	0.017	0.354	0.727	0.912
183.0896	7.8	Normetanephrine isomer	0.038	0.565	0.020	0.496	0.004	0.348	0.054	0.552	0.419	1.612
185.1052	12.4	Ecgonine	0.003	4.144	0.495	1.218	0.109	0.660	0.763	1.148	0.022	2.301
188.1161	14.4	N2-Acetyl-L-lysine	0.017	5.761	0.028	3.506	0.094	3.559	0.385	1.444	0.076	2.484
188.1161	8.5	N6-Acetyl-L-lysine	0.410	1.746	0.226	4.018	0.531	0.788	0.193	1.807	0.496	1.281
190.0953	18.8	Diaminoheptanedioate	0.084	3.783	0.608	0.805	0.030	0.370	0.199	0.595	0.112	1.890
191.0583	7.7	5-Hydroxyindoleacetate	0.061	0.214	0.043	0.151	0.033	0.095	0.043	0.149	0.249	0.512

191.0584	5.1	5,6-Dihydroxy-3-methyl-2-oxo-1,2-dihydroquinoline	0.194	0.415	0.050	0.115	0.058	0.150	0.076	0.212	0.169	0.399
194.0425	8.5	2-Dehydro-D-gluconate	0.125	0.118	0.103	0.058	0.109	0.075	0.098	0.044	0.209	0.288
194.0426	10.1	3-Dehydro-L-gulonate	0.356	4.012	0.783	0.814	0.856	0.862	0.107	0.148	0.220	0.364
194.079	13.3	1-O-Methyl-myo-inositol	0.385	2.369	0.355	10.283	0.320	15.995	0.016	0.491	0.300	5.207
194.079	9.5	3-O-Methyl-myo-inositol	0.105	4.870	0.060	8.014	0.038	8.579	0.064	8.651	0.069	2.535
195.0531	7.8	Dopaquinone	0.314	280.389	0.338	122.268	0.324	107.048	0.262	43.774	0.111	298.712
195.0531	8.5	2-Carboxy-2,3-dihydro-5,6-dihydroxyindole	0.278	96.984	0.279	18.066	0.288	50.601	0.275	126.763	0.121	876.012
195.0757	10.2	2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine	0.122	3.752	0.026	4.974	0.042	5.373	0.029	6.869	0.083	4.376
197.0688	7.7	N-Hydroxy-L-tyrosine	0.261	4.924	0.460	2.454	0.335	2.954	0.317	2.571	0.063	6.335
197.0688	12.7	DOPA	0.023	0.190	0.008	0.025	0.009	0.043	0.016	0.138	0.785	0.889
197.1052	8.0	Metanephrine	0.010	0.469	0.263	0.741	0.009	0.457	0.088	0.636	0.468	0.857
200.1048	5.0	[FA (10:1/2:0)] 2E-Decenedioic acid	0.407	0.727	0.076	0.465	0.104	0.478	0.439	0.756	0.140	1.532
200.1048	7.8	[FA (10:1/2:0)] 4Z-Decenedioic acid	0.643	0.856	0.219	0.636	0.231	0.609	0.760	0.883	0.021	2.037
200.1776	3.7	[FA methyl(11:0)] 10-methyl-undecanoic acid	0.551	0.500	0.737	0.719	0.812	0.802	0.730	0.703	0.498	0.427
202.1206	5.1	[FA (10:0/2:0)] Decanedioic acid	0.136	0.374	0.237	0.501	0.153	0.394	0.381	0.628	0.267	0.546
203.0794	7.9	N2-Acetyl-L-aminoadipate	0.669	0.605	0.262	0.080	0.252	0.060	0.372	3.932	0.217	3.746
203.1158	11.5	*O-Acetylcarnitine	0.073	52.011	0.120	19.002	0.045	33.596	0.002	4.976	0.176	6.821
203.1158	8.5	O-Acetylcarnitine	0.680	0.812	0.187	0.408	0.399	0.649	0.762	0.848	0.395	1.606
203.1158	8.1	O-Acetylcarnitine	0.131	0.428	0.116	0.401	0.922	0.931	0.210	0.524	0.555	0.788

204.111	18.6	N6-Acetyl-N6-hydroxy-L-lysine	0.064	3.718	0.129	3.394	0.411	1.948	0.941	0.948	0.072	4.847
205.0739	9.0	Indolelactate	0.003	0.230	0.000	0.013	0.000	0.000	0.001	0.182	0.173	0.673
205.1314	14.1	Pantothenol isomer	0.317	0.677	0.013	0.307	0.002	0.114	0.222	0.581	0.229	2.067
207.0896	7.6	N-Acetyl-D-phenylalanine	0.006	0.454	0.001	0.300	0.003	0.337	0.012	0.459	0.856	0.966
207.0896	11.2	N-Acetyl-L-phenylalanine	0.169	0.300	0.195	0.298	0.067	0.100	0.567	0.660	0.601	1.328
208.0848	15.0	Formyl-5-hydroxykynurenamine	0.129	0.464	0.025	0.202	0.026	0.210	0.521	0.715	0.170	1.931
211.048	8.0	5-(2'-Formylethyl)-4,6-dihydroxypicolinate	0.362	2.345	0.555	2.276	0.540	2.028	0.327	2.517	0.047	6.222
212.1411	3.3	[FA oxo(12:1)] 12-oxo-10E-dodecenoic acid	0.055	0.504	0.010	0.392	0.015	0.420	0.013	0.402	0.541	0.857
212.1413	3.8	[FA oxo(12:1)] 12-oxo-10E-dodecenoic acid	0.450	0.854	0.610	1.177	0.867	1.047	0.640	0.891	0.463	1.201
213.0637	7.9	N,N-Dihydroxy-L-tyrosine	0.059	0.313	0.031	0.207	0.026	0.175	0.427	4.227	0.359	3.279
214.1318	8.6	Dethiobiotin	0.096	4.427	0.156	10.913	0.038	3.369	0.079	3.275	0.076	1.666
215.0559	16.2	*Phosphoethanolamine	0.097	3.973	0.158	3.481	0.385	1.978	0.492	1.448	0.156	1.951
215.1158	5.1	2-Amino-9,10-epoxy-8-oxodecanoic acid	0.431	0.841	0.075	0.615	0.064	0.614	0.073	0.612	0.476	0.858
216.1723	3.7	12-Hydroxydodecanoic acid	0.177	1.814	0.011	4.999	0.028	6.143	0.027	5.282	0.160	1.989
217.1063	11.2	N-Acetyl-L-citrulline	0.079	3.933	0.271	6.082	0.195	2.017	0.734	1.157	0.184	1.585
217.1313	8.3	O-Propanoylcarnitine	0.306	0.590	0.258	0.547	0.174	0.466	0.479	0.719	0.632	1.270
217.1426	25.9	beta-Alanyl-L-lysine	0.084	5.993	0.304	3.061	0.934	0.961	0.607	1.465	0.961	0.981
217.1427	23.7	beta-Alanyl-L-lysine	0.094	3.902	0.232	4.856	0.951	1.022	0.337	1.734	0.436	1.286
218.1267	17.8	N2-(D-1-Carboxyethyl)-L-lysine	0.247	1.742	0.682	0.837	0.153	0.562	0.784	1.131	0.061	3.402
218.1267	13.8	N2-(D-1-Carboxyethyl)-L-lysine	0.235	0.481	0.136	0.359	0.071	0.222	0.243	0.503	0.965	1.021
219.1107	5.1	*Pantothenate	0.113	3.740	0.056	4.736	0.160	2.929	0.298	3.325	0.289	1.908
220.0847	11.4	5-Hydroxytryptophan	0.047	2.547	0.959	0.968	0.025	0.242	0.064	0.379	0.174	1.750

220.0848	8.6	5-Hydroxy-L-tryptophan	0.239	10.959	0.267	6.559	0.348	2.495	0.265	1.697	0.102	1.986
220.0849	8.0	5-Hydroxy-L-tryptophan	0.230	8.736	0.224	6.993	0.148	1.637	0.140	1.601	0.107	1.489
221.09	13.6	N-Acetyl-D-mannosamine	0.234	4.554	0.876	1.060	0.261	0.647	0.953	0.976	0.724	0.888
221.09	12.2	N-Acetyl-D-glucosamine	0.195	4.261	0.440	1.390	0.626	0.838	0.698	1.174	0.675	0.869
224.0798	8.1	3-Hydroxy-L-kynurenine	0.239	4.959	0.486	3.032	0.490	2.370	0.248	3.666	0.005	7.211
226.0953	8.2	Porphobilinogen	0.476	0.701	0.168	0.438	0.099	0.333	0.090	0.316	0.582	0.779
226.1065	12.1	Carnosine	0.036	5.863	0.253	2.820	0.507	1.313	0.344	0.706	0.223	1.367
228.0748	8.4	Deoxyuridine	0.479	0.724	0.796	1.191	0.004	0.225	0.116	0.524	0.708	1.387
230.1517	3.8	Dodecanedioic acid	0.004	0.144	0.005	0.184	0.004	0.139	0.010	0.268	0.316	0.623
231.147	8.9	O-Butanoylcarnitine	0.191	6.748	0.289	4.241	0.076	6.583	0.605	1.254	0.354	12.591
232.1059	9.4	N6-Acetyl-LL-2,6-diaminoheptanedioate	0.022	6.557	0.303	5.125	0.213	3.713	0.979	1.009	0.132	2.595
232.1059	15.5	N2-Succinyl-L-ornithine	0.185	33.008	0.006	0.017	0.013	0.126	0.766	1.390	0.079	16.161
232.1212	9.1	Melatonin	0.225	7.174	0.612	0.623	0.717	1.567	0.654	1.555	0.385	2.195
236.0797	12.4	L-Formylkynurenine	0.006	5.766	0.010	4.817	0.100	2.414	0.214	1.906	0.009	3.924
240.122	13.2	Homocarnosine	0.207	0.572	0.702	0.840	0.001	0.258	0.619	0.833	0.259	1.736
240.1222	16.9	Homocarnosine	0.547	2.117	0.057	0.042	0.055	0.030	0.108	0.198	0.640	1.369
240.1222	14.1	beta-Alanyl-N(pi)-methyl-L-histidine	0.264	0.618	0.011	0.361	0.010	0.353	0.265	0.694	0.206	1.830
240.1725	3.1	oxoTetradecenoic acid	0.021	0.379	0.003	0.168	0.003	0.131	0.006	0.225	0.592	0.860
241.1175	16.7	Tetrahydrobiopterin	0.069	0.220	0.060	0.190	0.054	0.166	0.559	2.107	0.744	1.204
242.0904	7.7	Thymidine	0.168	0.528	0.696	1.301	0.038	0.323	0.203	0.528	0.841	1.140
243.0856	12.4	Cytidine	0.339	0.563	0.231	0.472	0.056	0.200	0.064	0.222	0.182	0.442
244.0694	12.2	Pseudouridine	0.240	1.963	0.380	0.685	0.039	0.377	0.037	0.386	0.732	1.270
244.0694	10.1	Uridine	0.492	0.791	0.243	1.828	0.433	0.769	0.554	0.817	0.989	0.997

245.1489	26.6	beta-Alanyl-L-arginine	0.037	13.169	0.106	6.450	0.141	2.238	0.133	3.359	0.010	4.960
248.116	11.9	6-Hydroxymelatonin	0.029	8.813	0.227	3.635	0.341	1.989	0.663	1.372	0.040	3.219

Table S4. 2 Sphingosine metabolism for CD vs HC

Mass	RT	Putative metabolite	p value PC/PA	PA/PC	p value PC/PB	PB/PC	p value HC/PC	PC/HC	p value PC/PD	PD/PC	p value PE/PC	PE/PC
295.2512	4.4	Sphingatrienine	0.985	1.006	0.286	0.737	0.850	0.947	0.670	0.881	0.628	1.123
315.2774	3.6	Dehydrophytosphingosine	0.391	1.507	0.798	1.108	0.460	1.360	0.536	1.294	0.016	2.881
315.2774	4.2	Hydroxysphingenine	0.881	1.171	0.406	0.222	0.486	0.349	0.428	0.258	0.494	2.300
327.3137	4.2	N,N-Dimethylsphing-4-enine	0.006	6.526	0.013	23.310	0.007	27.210	0.038	18.263	0.020	3.368
465.3454	3.8	LysoSM(18:1)	0.756	0.901	0.931	0.976	0.669	1.124	0.918	0.968	0.419	0.792
467.3612	3.8	Sphinganinephosphocholine	0.195	0.569	0.375	0.719	0.933	1.034	0.637	0.840	0.075	0.472
481.4494	4.1	Dodecanoylsphingenine	0.024	ND	0.040	ND	0.025	ND	0.243	ND	0.017	ND
509.4814	4.1	Tetradecanoylsphingenine	0.008	8.294	0.021	6.755	0.022	6.433	0.220	3.380	0.002	5.397
537.512	4.1	Hexadecanoylsphingenine	0.005	20.235	0.028	12.052	0.047	6.116	0.141	3.591	0.000	8.375
555.5225	4.1	hydroxyhexadecanoylsphingenine	0.570	1.284	0.028	2.729	0.037	2.700	0.164	2.205	0.935	1.030
563.5277	4.0	Octadecenoylsphingenine	0.003	11.861	0.030	7.152	0.054	4.992	0.320	3.699	0.001	6.616
565.5436	4.0	Octadecanoylsphingenine	0.026	3.912	0.214	1.581	0.704	1.133	0.880	1.066	0.030	4.746
569.538	4.1	Cer(d18:0/h17:0)	0.373	0.658	0.259	0.604	0.299	0.635	0.964	1.020	0.663	1.169
647.6217	4.1	Tetracosenoylsphingenine	0.002	27.212	0.027	10.818	0.024	5.612	0.123	3.305	0.002	19.136
702.5675	4.4	Hexadecanoylsphingeninephosphocholine	0.002	13.250	0.031	5.520	0.052	4.231	0.391	1.789	0.002	8.834
812.677	4.3	SM(d18:1/24:1(15Z))	0.006	20.757	0.030	7.319	0.066	5.979	0.432	2.195	0.004	14.297

Table S4. 3 Calprotectin values. ND = not determined

Subject	Calprotectin wet	Calprotectin dry
PA01	2272.3	15056
PA03	1130.4	6161
PA04	2438.7	8088
PA05	2581.7	7876
PA06	2076.3	6022
PA07	2102.2	5751
PA08	2187.9	9334
PA09	2262.3	7371
PA11	3114.22	19147
PB01	1841.6	6335
PB02	2390.1	19899
PB03	47.7	211
PB04	2221.6	7056
PB05	2341.1	7695
PB06	1704.8	6785
PB07	1999.9	6760
PB08	2056.2	9720
PB09	1808	7659
PB11	2216.346	7803
PC01	1673.7	6676
PC02	2324.6	20520
PC03	5.8	29
PC04	2000.5	5878
PC05	2394.3	7967
PC06	1459.5	3909
PC07	1535.8	4985
PC08	2076.7	9321
PC09	1797.6	8315
PC10	296.393	713
PC11	1803.814	7743
PD01	1685.8	7400
PD02	2563.7	20312
PD03	88.4	373
PD04	1723.5	6171
PD05	2460.9	8214
PD06	39.1	139
PD07	718.1	2338

PD08	2298.3	7376
PD09	2055.8	11449
PD10	77.121	183
PD11	106.118	444
PE01	2085.3	6701
PE02	1052.7	4221
PE03	2054.6	12284
PE04	2327.7	9084
PE05	2495	7276
PE06	1712.1	4816
PE07	1632.2	7821
PE08	2169.5	7948
PE09	2470.2	11241
PE10	2418.17	6754
PE11	2355.175	9293
HC01	ND	ND
HC02	8	25
HC03	101.3	300
HC05	3.7	11
HC06	ND	ND
HC07	ND	ND
HC08	8.5	20
HC10	5.7	19
HC11	8.3	23

Characterisation of unknown markers

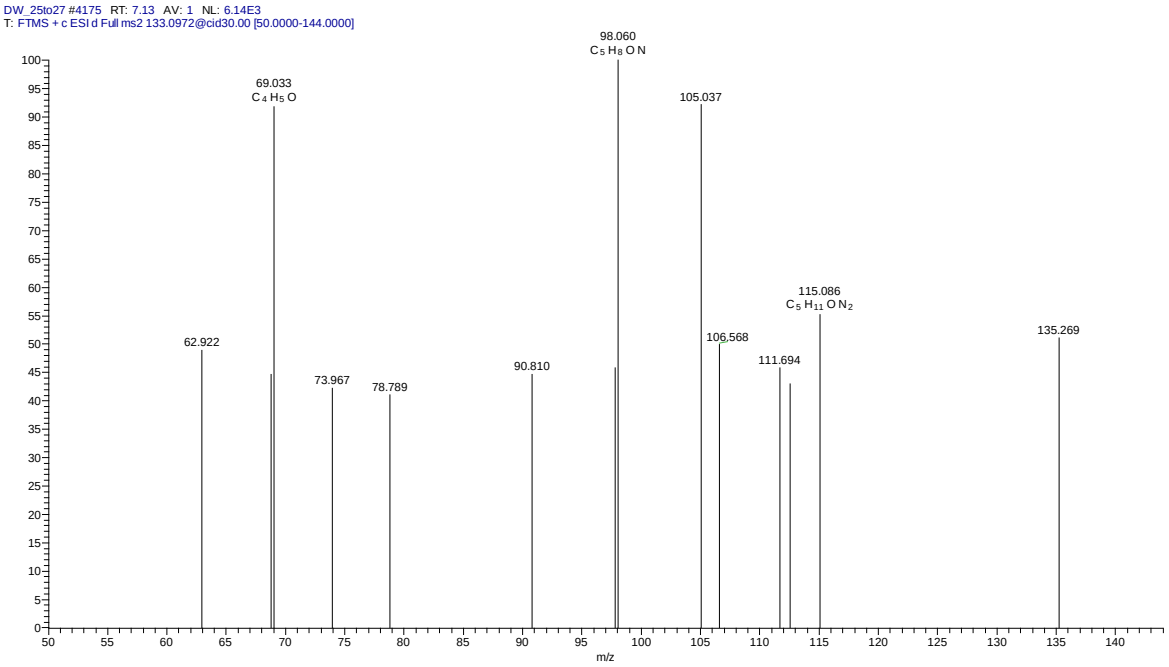


Figure S4. 3 MS² fragments of ornithine isomer

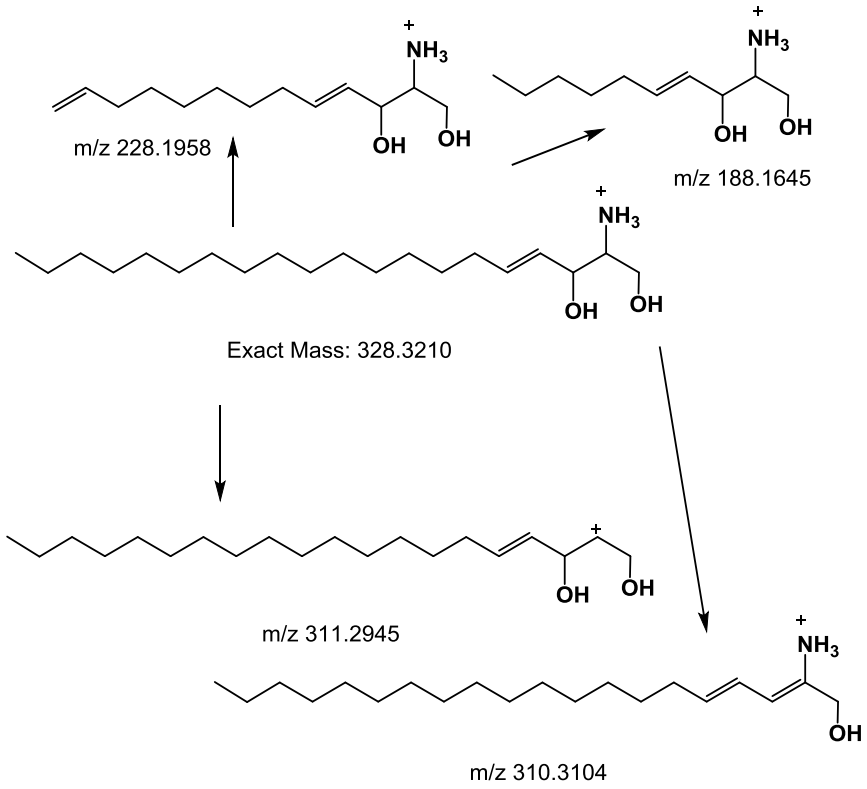


Figure S4. 4 Proposed fragmentation of C20 sphinganine

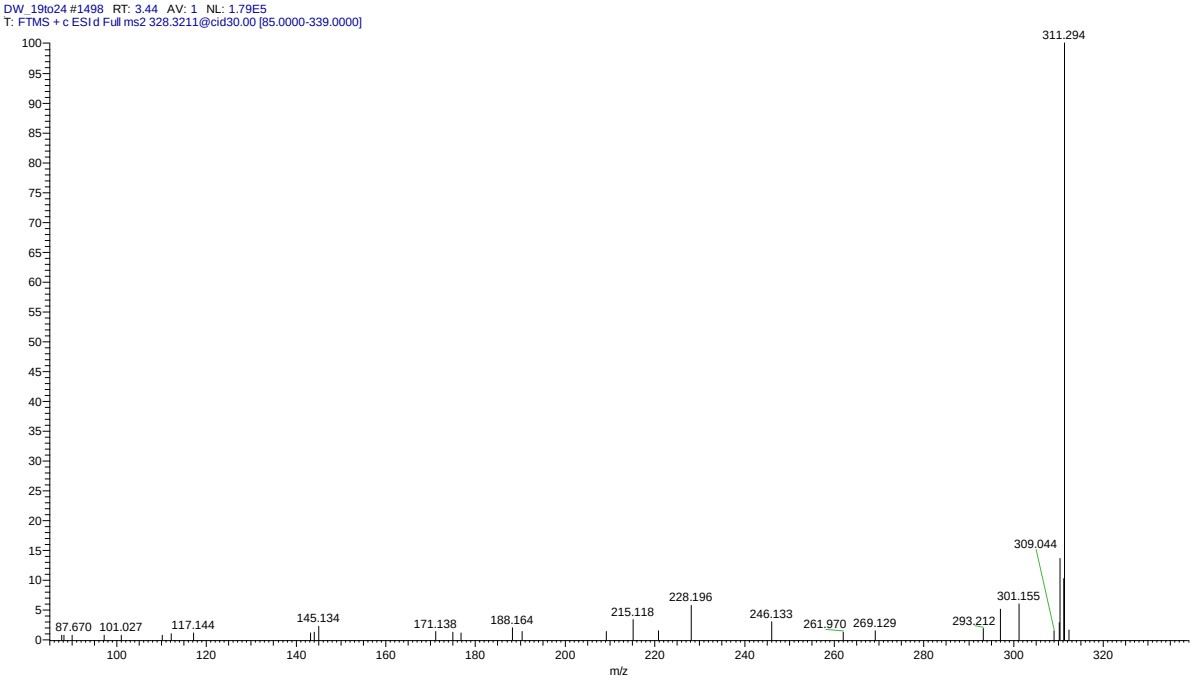


Figure S4. 5 MS^2 spectrum of C20 sphinganine

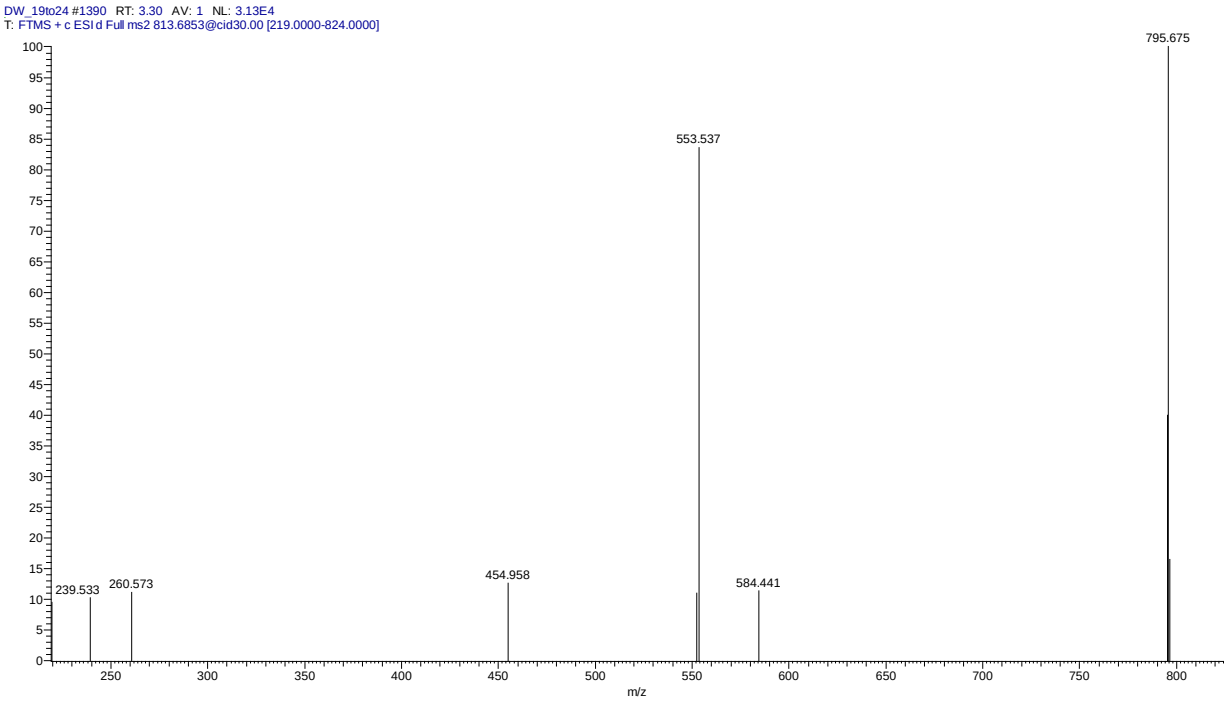


Figure S4. 6 MS² spectrum of Ceramide d18:1 24:1

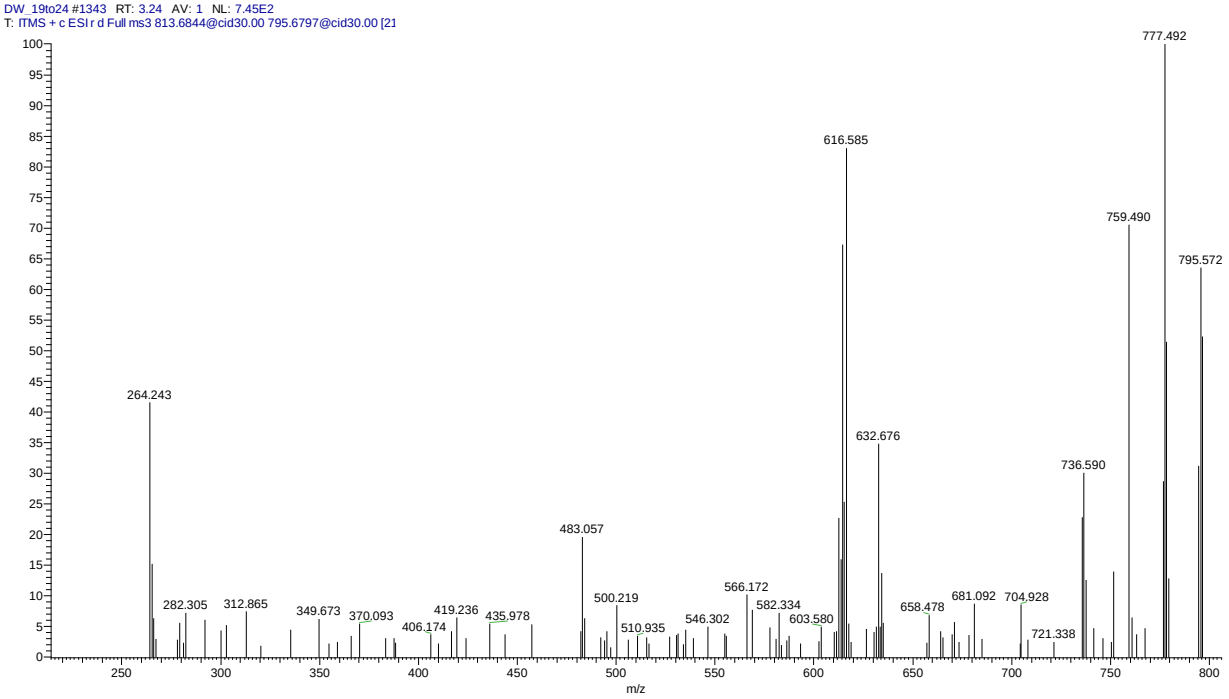


Figure S4. 7 MS³ spectrum of Ceramide d18:1 24:1 (795.5 ion).

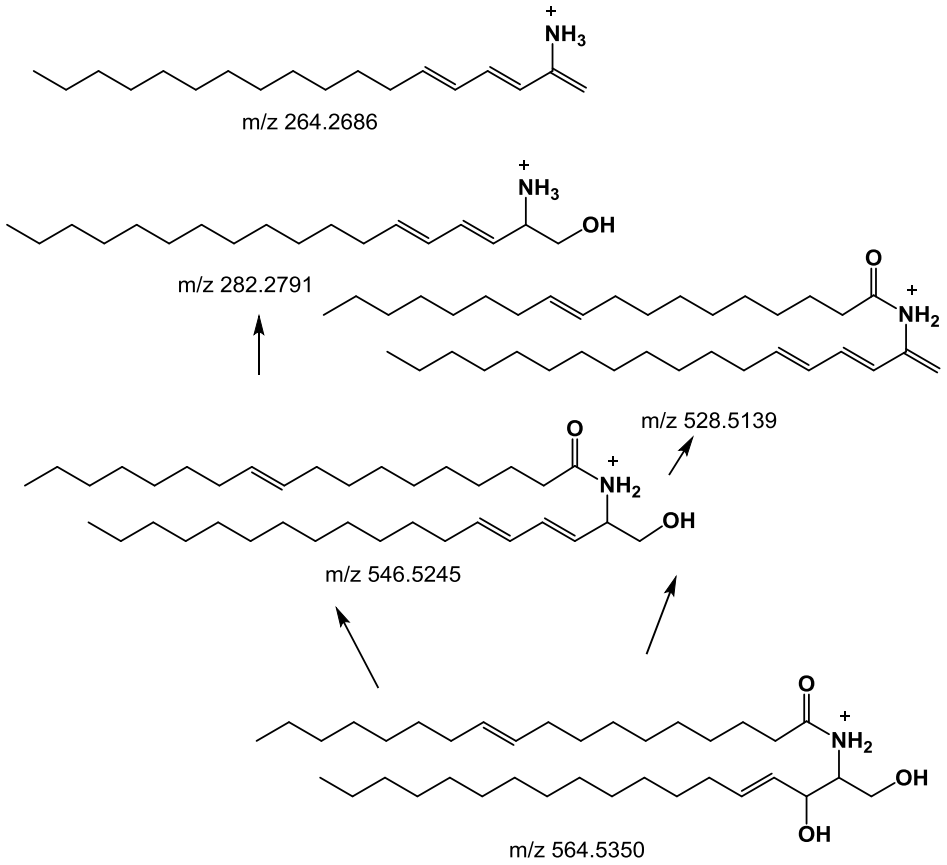


Figure S4. 8 Proposed fragmentation of octadecenoylsphingenine (MS^2 spectrum shown in figure S4.9)

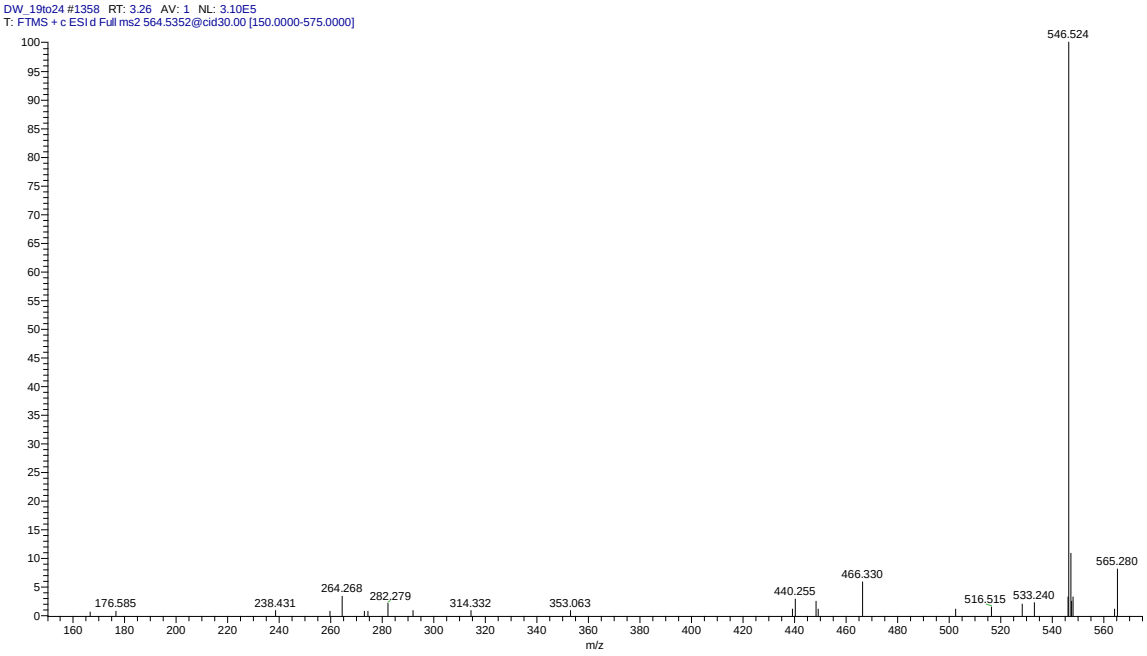


Figure S4. 9 MS^2 spectrum of octadecyl sphinganine

8.3 Appendixes chapter 5

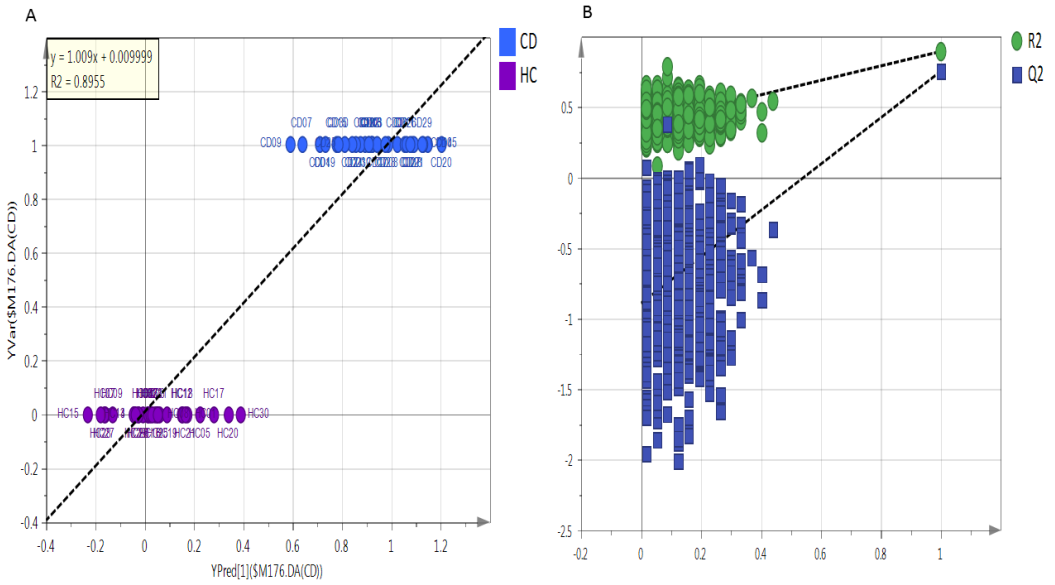


Figure S5. 1 The model of CD versus HCs based on 34 putative metabolites. *(A) The plot displays the observed (y-axis) versus predicted (x-axis) values of the selected Y-variable of the model. The R^2 of the regression line indicates the goodness of Fit = 0.89, with a good model all the observations will fall close to this 45 degree of the line, and with less good models the observations are scattered around the regression line. (B) Permutations test. The plot shows, for a selected Y-variables, on the vertical axis the values of R^2 and Q^2 for the original model (far to the right) and of the Y-permuted models further to the left. The horizontal axis shows the correlation between the permuted Y-vectors and the original Y-vector for the selected Y. The original Y has the correlation 1.0.*

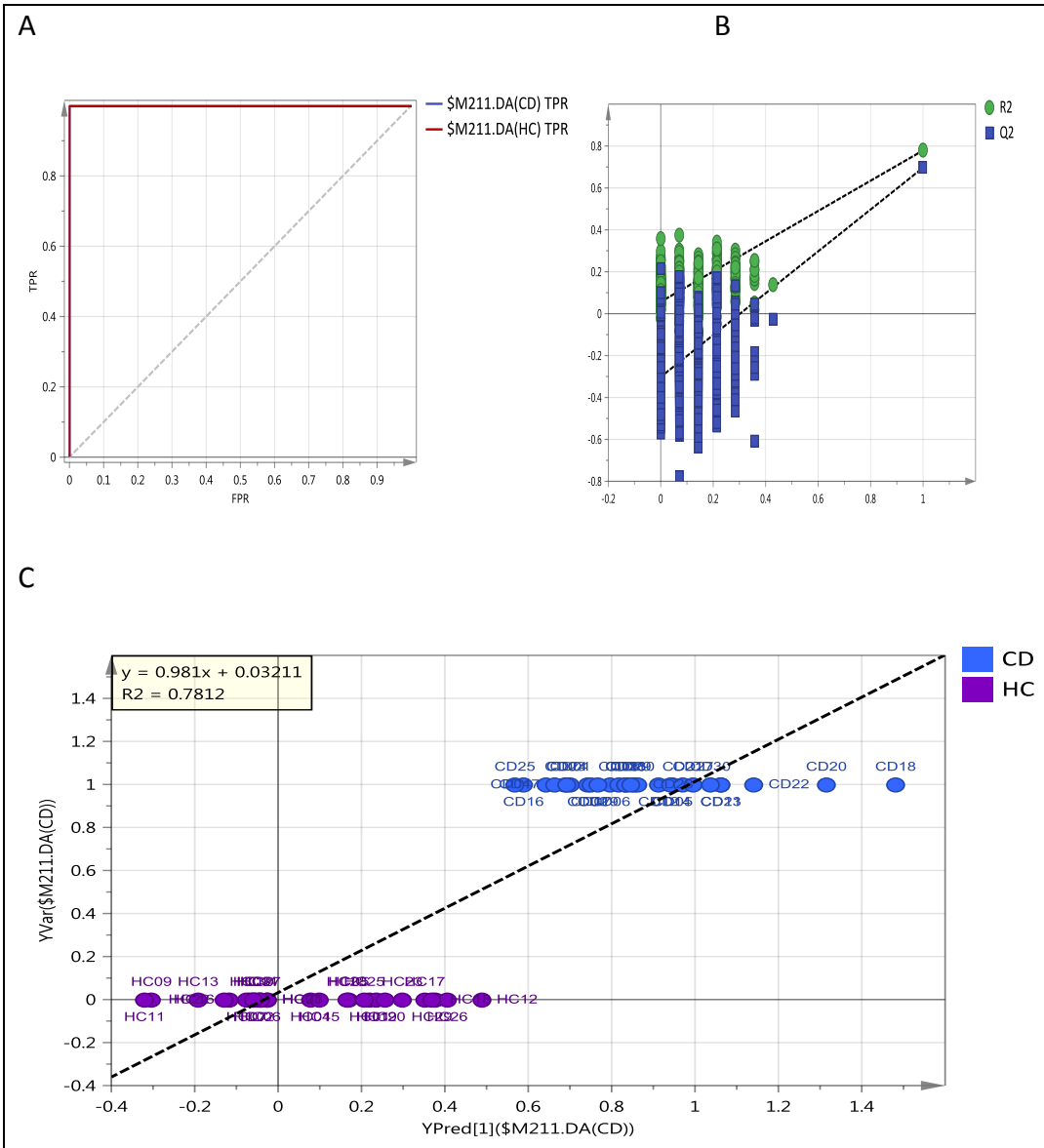


Figure S5. 2 CD versus HCs model validation based on 10 putative metabolites. *(A) Receiver Operating Characteristics (ROC) curve shows sensitivity (true positive rate (TPR)) on the y-axis versus (false positive rate (FPR = 1 - Specificity)) on the x-axis. The area under ROC curve (AUROCC) =1 for each group. (B) Permutations test. The plot shows, for a selected Y-variables, on the vertical axis the values of R^2 and Q^2 for the original model (far to the right) and of the Y-permuted models further to the left. The horizontal axis shows the correlation between the permuted Y-vectors and the original Y-vector for the selected Y. The original Y has the correlation 1.0. (C) The plot displays the observed (y-axis) versus predicted (x-axis) values of the selected Y-variable of the model. The R^2 of the regression line indicates the goodness of Fit = 0.78, with a good model all the observations will fall close to this 45 degree of the line, and with less good models the observations are scattered around the regression line*

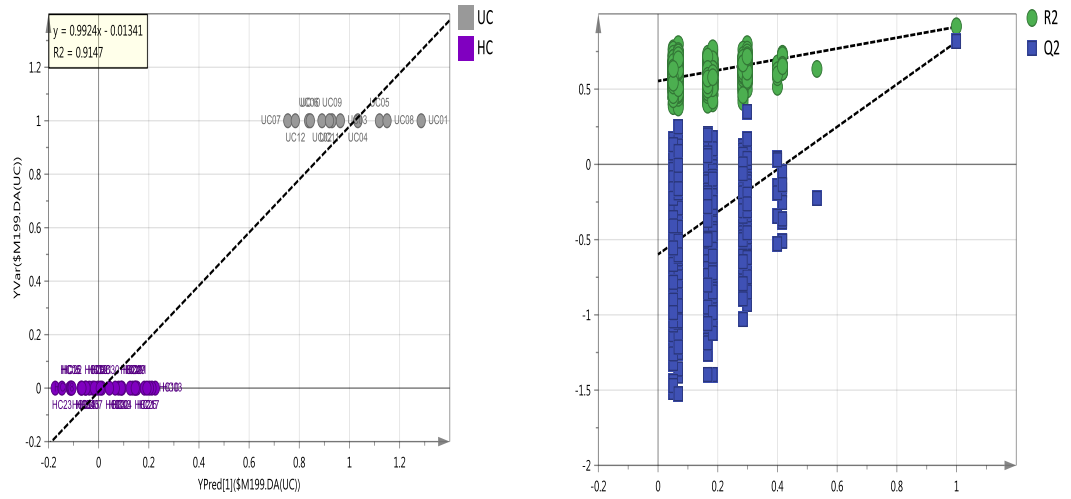


Figure S5. 3 The model of CD versus HCs based on 34 putative metabolites. *(A) The plot displays the observed (y-axis) versus predicted (x-axis) values of the selected Y-variable of the model. The R^2 of the regression line indicates the goodness of Fit = 0.91.5, with a good model all the observations will fall close to this 45 degree of the line, and with less good models the observations are scattered around the regression line. (B) Permutations test. The plot shows, for a selected Y-variables, on the vertical axis the values of R^2 and Q^2 for the original model (far to the right) and of the Y-permuted models further to the left. The horizontal axis shows the correlation between the permuted Y-vectors and the original Y-vector for the selected Y. The original Y has the correlation 1.0.*

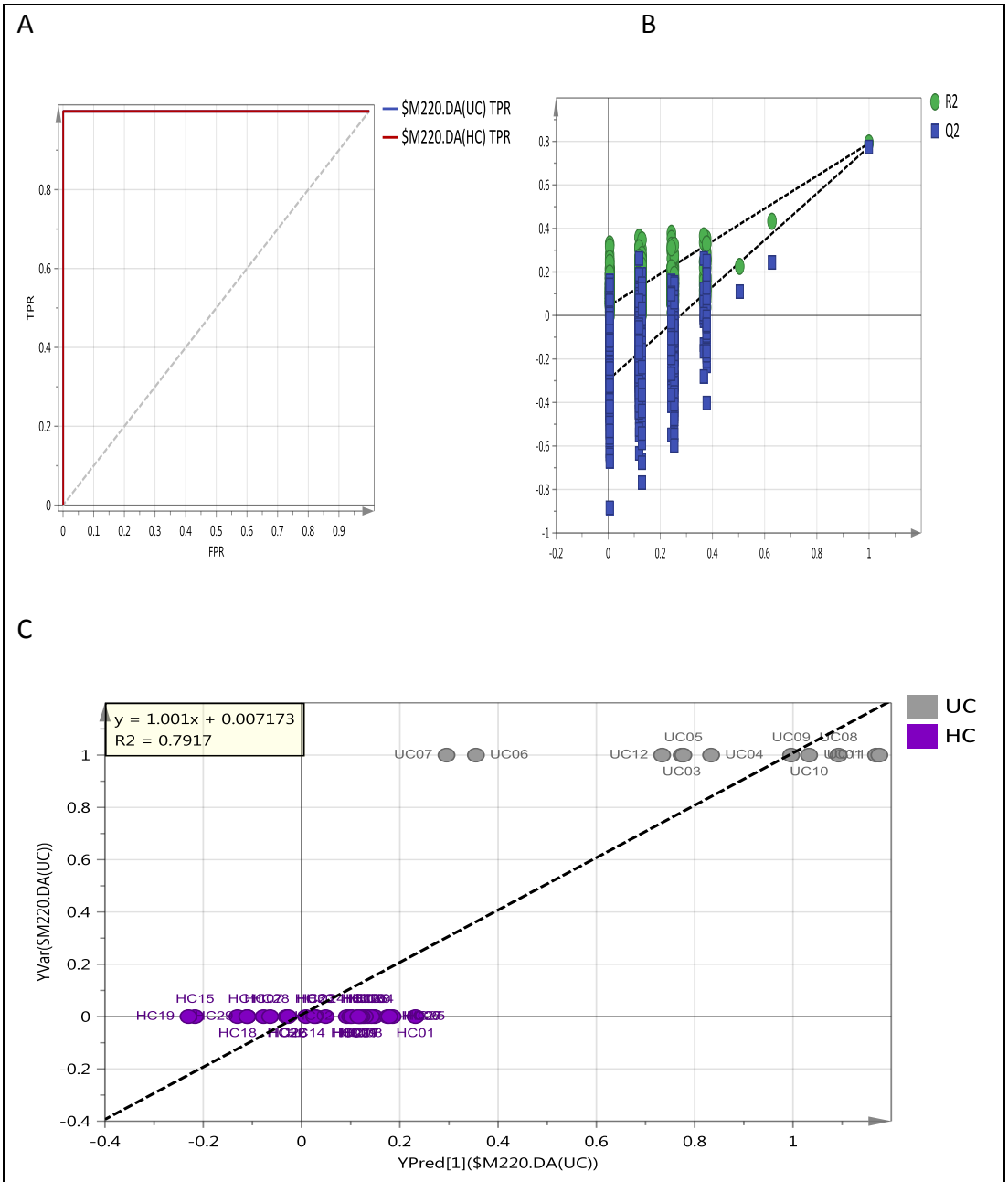


Figure S5. 4 UC versus HCs model validation based on 10 putative metabolites. *(A) Receiver Operating Characteristics (ROC) curve shows sensitivity (true positive rate (TPR)) on the y-axis versus (false positive rate (FPR = 1 - Specificity)) on the x-axis. The area under ROC curve (AUROCC) =1 for each group. (B) Permutations test. The plot shows, for a selected Y-variables, on the vertical axis the values of R^2 and Q^2 for the original model (far to the right) and of the Y-permuted models further to the left. The horizontal axis shows the correlation between the permuted Y-vectors and the original Y-vector for the selected Y. The original Y has the correlation 1.0. (C) The plot displays the observed (y-axis) versus predicted (x-axis) values of the selected Y-variable of the model. The R^2 of the regression line indicates the goodness of Fit = 0.79, with a good model all the observations will fall close to this 45 degree of the line, and with less good models the observations are scattered around the regression line*

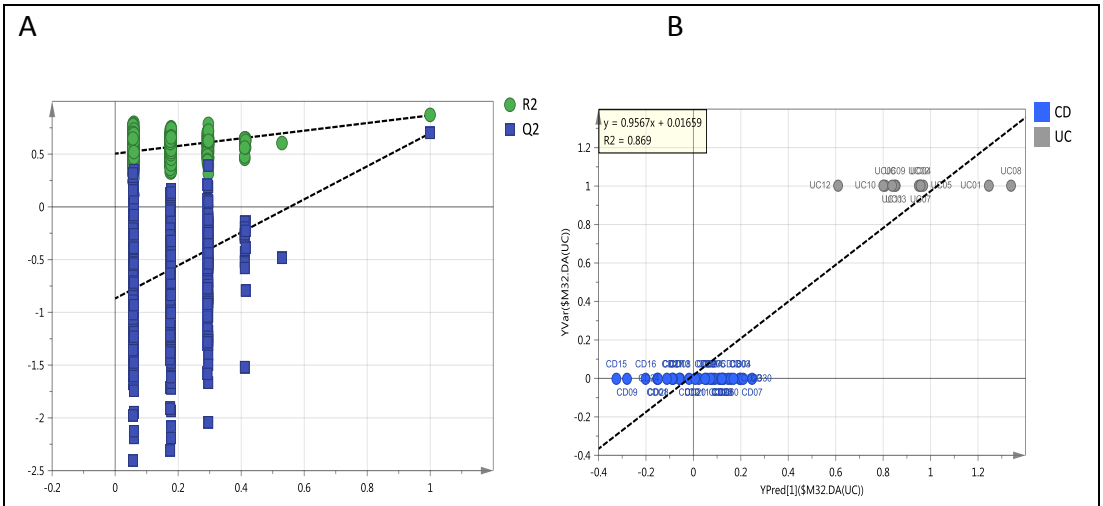


Figure S5. 5 The model of CD versus HCs based on 34 putative metabolites. *(A) The plot displays the observed (y-axis) versus predicted (x-axis) values of the selected Y-variable of the model. The R^2 of the regression line indicates the goodness of Fit = 0.956, with a good model all the observations will fall close to this 45 degree of the line, and with less good models the observations are scattered around the regression line. (B) Permutations test. The plot shows, for a selected Y-variables, on the vertical axis the values of R^2 and Q^2 for the original model (far to the right) and of the Y-permuted models further to the left. The horizontal axis shows the correlation between the permuted Y-vectors and the original Y-vector for the selected Y. The original Y has the correlation 1.0.*

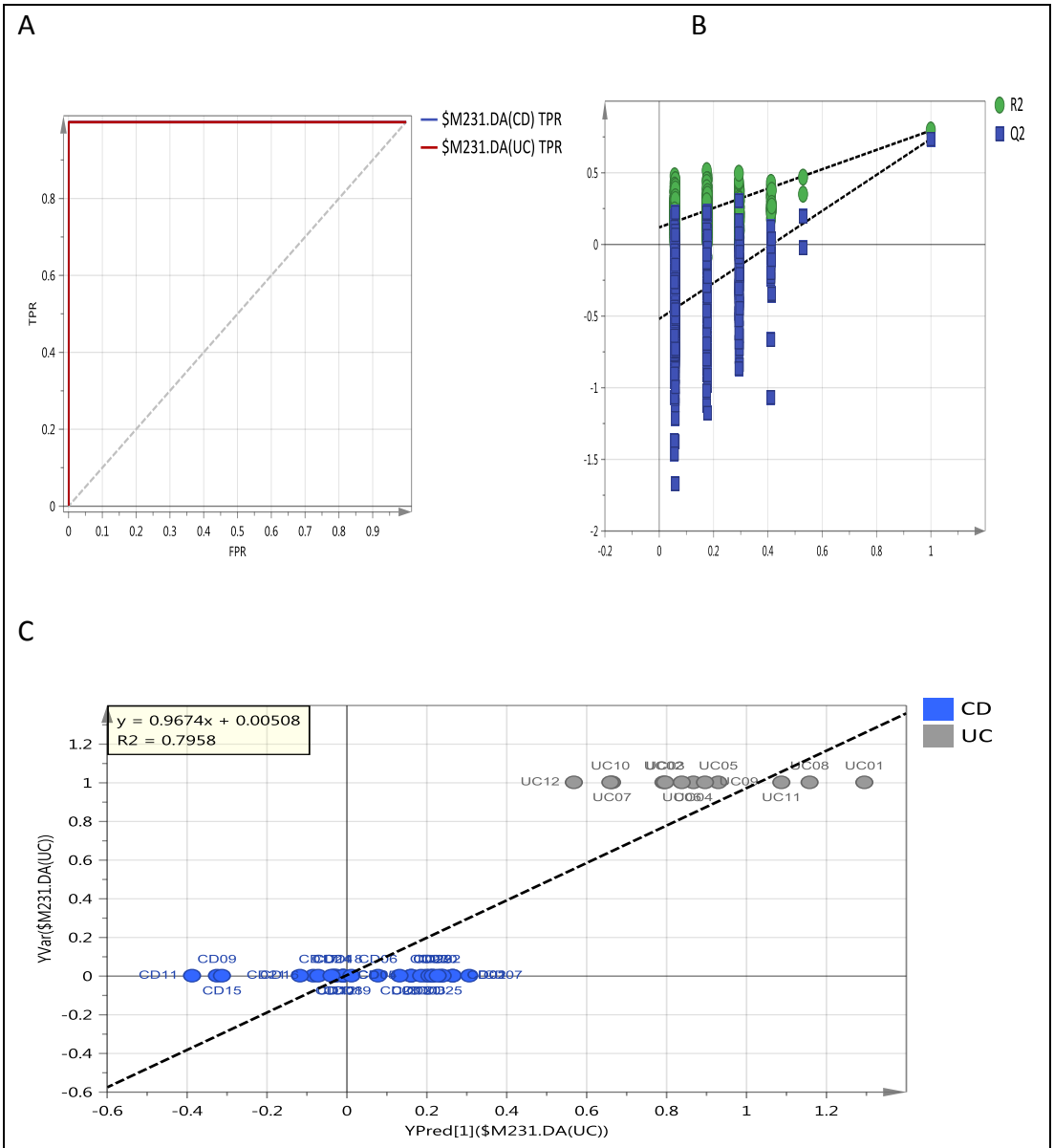


Figure S5. 6 CD versus UC model validation based on 10 putative metabolites. *(A) Receiver Operating Characteristics (ROC) curve shows sensitivity (true positive rate (TPR)) on the y-axis versus (false positive rate (FPR = 1 - Specificity)) on the x-axis. The area under ROC curve (AUROCC) =1 for each group. (B) Permutations test. The plot shows, for a selected Y-variables, on the vertical axis the values of R^2 and Q^2 for the original model (far to the right) and of the Y-permuted models further to the left. The horizontal axis shows the correlation between the permuted Y-vectors and the original Y-vector for the selected Y. The original Y has the correlation 1.0. (C) The plot displays the observed (y-axis) versus predicted (x-axis) values of the selected Y-variable of the model. The R^2 of the regression line indicates the goodness of Fit = 0.795, with a good model all the observations will fall close to this 45 degree of the line, and with less good models the observations are scattered around the regression line.*