Development of an Advanced Molecular Profiling Pipeline for Human Population Screening

Thesis submitted by

Matthew Reese Lewis

For the degree of Doctor of Philosophy (PhD) of Imperial College London

and the Diploma of Imperial College London (DIC)

Supervisors: Professor Elaine Holmes, Dr. Olivier Cloarec,

Dr. Elizabeth Want, & Professor John Shockcor

Computational Systems Medicine

Department of Surgery and Cancer

Faculty of Medicine

Imperial College London

2014

Abstract

The interaction between a human's genes and their environment is dynamic, producing phenotypes that are subject to variance among individuals and across time. Metabolic interpretation of phenotypes, including the elucidation of underlying biochemical causes and effects for physiological or pathological processes, allows for the potential discovery of biomarkers and diagnostics which are important in understanding human health and disease. The study of large cohorts has been pursued in hopes of gaining sufficient statistical power to observe subtle biochemical processes relevant to human phenotypes. In order to minimise the effects of analytical variance in metabolic profiling and maximise extractable information, it is necessary to develop a refined analytical approach to large scale metabolic profiling that allows for efficient and high quality collection of data, facilitating analysis on a scale appropriate for molecular epidemiology applications. The analytical methods used for the multidimensional separation and detection of metabolic content from complex biofluids must be made fit for this purpose, deriving data with unprecedented reproducibility for direct comparison of metabolic profiles across thousands of individuals. Furthermore, computational methods must be established for collating this data into a form that is suitable for analysis and interpretation without compromising the quality achieved in the raw data. These developments together constitute a pipeline for large scale analysis, the components of which are explored and refined herein with a common thread of improving laboratory efficiency and measurement precision. Complimentary chromatographic methods are developed and implemented in the separation of human urine samples, and further mated to separation and detection by mass spectrometry to provide information rich metabolic maps. This system is optimised to derive precision from sustained analysis, with emphasis on minimisation of sample batching thereby allowing the development of metabolite collation tools that leverage the chromatographic reproducibility. Finally, the challenge of metabolite identification in molecular profiling is conceptually addressed in a manner that does not preclude the further reinvention of the analytical approaches established within this thesis. In summary, the thesis offers a novel and practical analytical pipeline suitable for achieving high quality population phenotyping and metabolome wide association studies.

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Copyright Declaration

Acknowledgments

It is my great pleasure to have this opportunity to formally acknowledge and thank the many people who played a part in my work and development throughout these past years at Imperial College London.

I wish to thank my supervisors, Professor Elaine Holmes, Dr. Olivier Cloarec, Dr. Elizabeth Want, and Professor John Shockcor for their advice and support during my studies, as well as for their time and efforts in reviewing this thesis: John, without you, none of this would be possible. It was your support and advocacy that opened the doors to life on the other side of the world. It was your unique perspective that saw a fit for me at Imperial College London. I'm most grateful for your unique view of the world, as well as for your honesty and criticism. You have influenced the course of my life profoundly – I hope I've made you proud. Liz, you've laid the foundation on which the chromatographic developments contained herein have been built. I acknowledge wholeheartedly that without your fundamental contributions to the field of small molecule profiling, much of this work would not have had the opportunity to be developed, and we would not be achieving the analyses that we are capable of today. You're the most petite of giants – and I am sincerely lucky to have had your shoulders to stand on. Olivier, you bought me my first beer in England, and we've been friends ever since. Somehow, you've made Kathy and I feel at home in a country that is not your own, which speaks volumes for your unique character. As fortunate and humbled as I am to know someone of your intellect, I'm happier still just to know you. Finally, to Elaine – I can only wish for everyone to be so fortunate in having someone who believes in them as you have believed in me. You have supported me at every turn, through successes and missteps alike. For your patience and understanding, for your honesty and your shelter, for your friendship - I am forever in your debt.

I would like to thank three additional people for their scientific and personal mentorship. First, to Professor Zoltan Takats, for your inspiring creativity and helpful guidance. It has been a pleasure working with and learning from you. Second, to Mrs. Julia Anderson, for your wisdom and rational interpretation of all situations, no matter how convoluted. Without your vigilance we would all be lost! And finally to Professor Jeremy Nicholson – the fount of opportunity – for both building the house and

keeping the roof up over our heads. Your precedent and your vision have given us everything we need to succeed. I am conscious of that gift daily.

I would like to thank the talented scientific staff of the MRC-NIHR National Phenome Centre – putting all of this into motion with you in the pursuit of ever-larger datasets has been immensely rewarding. Much of the development and testing throughout this thesis would not have come to fruition without your assistance. My specific thanks and gratitude to the "original five" - Ms. Katie Chappell, Mr. Mark David, Dr. Dave Berry, Mrs. Ada Armstrong, and Dr. Verena Horneffer-van der Sluis - for your assistance and important contributions to the pursuit of refined UPLC-MS reference mixture recipes, long term reference generation, assessment of MS source configurations, and UPLC-MS system stability testing.

To my incredible colleagues, present and past – I have been so fourtinate and humbled to work among you from FLORINASH (Dr. Richard Barton and Dr. Marc Dumas) to surgical mass spectrometry (Dr. Reza Mirnezami and Dr. James Kinross, Dr. Steve Pringle and Professor Mike Morris), and in CSM (Dr. Anthony Dona, Dr. Toby Athersuch, Dr. James Ellis, Dr. Anisha Wijeyesekera, Dr. Volker Behrends, and Dr. Jia Li). To Dr. Jake Pearce I extend my deepest gratitude. Since the start of my time at Imperial, it has been my good fortune to work with such an immensely bright and unwaveringly logical colleague and friend. Of all the work in the past few years, I am most proud of the things we have accomplished together. Finally, I have you to thank for encouraging me to put pen to paper and outline this thesis in the first place.

Throughout this work, nothing has served me so well as perspective. For this, I carry with me the words of Albert Camus, that "There is but one truly serious philosophical problem and that is suicide. Judging whether life is or is not worth living amounts to answering the fundamental question of philosophy. All the rest – whether or not the world has three dimensions, whether the mind has nine or twelve categories – comes afterwards. These are games; one must first answer." With love and sincerity, to the London Family – Esti, Panos, Perrine, Dina, Paul, Claire, Lea, Renaud, Florian, Alex, Evie, Anas, Gabriel, Michael, Ioanna, Maria, Magali, and Abdullah - beyond the games of science, your friendship and humanity have made life here worth living. A very special thanks to my old sport Dr. Renaud

MESTDAGH and the lovely Dr. Olesea ROMAN, for your encouragement, positivity, and shelter while writing, I am eternally grateful.

To my family at home – I've missed you these years, but it is because of you that I have had the opportunities and the education that eventually led me to where I am. I am thankful for that – for all that you've taught me, and everything that you've made me.

And finally – most importantly, to my wife Kathy, for your love and support, your sanity, and being up for this whole adventure in the first place.

Table of Contents

Abstract....3

Statement of Originality....4

Copyright Declaration....5

Acknowledgments....6

Table of Contents....9

List of Tables....14

List of Figures....16

General Introduction....21

Objectives....23

Thesis Structure....24

Chapter 1: Introduction....25

1.1 Metabonomics and applications in population phenotyping....25

1.2 Urinalaysis....27

Chapter 2: Analytical strategies....29

2.1 Introduction....29

2.2 Fundamental principles of column-based liquid chromatography....30

2.3 Methods for assessing chromatographic performance....36

2.3.1 Defining chromatographic peak width....36

2.3.2 Measurement of chromatographic efficiency....37

2.3.3 Measurement of chromatographic resolution and peak capacity....38

2.4 LC detection by mass spectrometry....40

2.4.1 Ionisation....41

2.4.2 Mass spectrometry for LC-based molecular profiling 42

2.4.3 Orthogonal separations in LC-MS applications 46

2.6 Data review and pre-processing 48

2.6.1 Manual data review and illustrations 48

2.6.2 Automated pre-processing of LC-MS data 53

2.7 Multivariate data analysis 57

2.7.1 Principal components analysis (PCA) 58

2.7.2 Partial least squares (PLS) and orthogonal projection to latent structures (OPLS) analyses 60

2.8 Metabolite identification 63

2.8.1 The elemental composition approach to elucidating a molecular formula 63

2.8.2 Use of mass and spectral databases for metabolite/biomarker assignment 64

2.8.3 Validation to authentic standards for metabolite identification 64

2.8.4 Method-specific LC-MS retention time databases 65

Chapter 3: Development of coupled liquid chromatography and mass spectrometric methods for the profiling of human urine from large patient cohorts. 67

3.1 Introduction 67

3.2 Specific Objectives 72

3.3 Reagents and biofluids for method development and system optimisation. 72

3.3.1 Development of a pooled urine sample for use as a representative matrix 73

3.3.2 Development of chemical reference mixtures 73

3.4 Adaptation of chromatographic separations 78

3.4.1 Urine analyte hydrophobicity and retention in a reversed-phase system 79

3.4.2 Optimisation of reversed-phase gradient elution conditions 85

3.4.3 Assessment of reversed-phase peak capacity 96

3.4.4 Adaptation of HILIC for complementary retention and separation of small polar analytes in urine 102

3.4.5 Assessment of separation complementarity 108

3.5 Maximisation of analytical batch size through optimisation of the LC-MS platform. 117

3.5.1 MS sensitivity as a currency for longitudinal precision 117

3.5.2 Source optimisation for sensitivity and minimal impact of sample on MS inlet 118

3.5.3 Testing the limits of UPLC-MS system stability for an optimised configuration 121

3.6 Optimisation of sample preparation batch size. 129

3.6.1 Assessment of urine stability 130

3.6.2 Simulation of analytical cycles in a large profiling experiment conducted within a model working environment 133

3.7 Method finalisation 138

3.8 Application to large-scale molecular profiling 141

3.8.1 Chromatographic precision 143

3.8.2 Intensity precision 147

3.9 Conclusions 148

Chapter 4: LC-MS feature grouping suitable for real-time application in large-scale profiling 151

4.1 Introduction 151

4.2 Specific Objectives of Method Development 154

4.3 Existing tools and strategies for feature alignment and grouping 155

4.4 Definitions 157

4.5 Experimental LC-MS dataset 158

4.5.1 LC-MS data acquisition 158

4.5.2 Expected and observed patterns chromatographic retention deviation 159

4.5.3 Feature clusters and cluster migration 164

4.5.4 Pairwise comparison 169

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup) 170

4.6.1 Intra-sample matching to define feature clusters 172

4.6.2 Inter-sample matching 178

4.6.3 Feature communities and resolution of ambiguous matching 179

4.6.4 Reporting links between features 187

4.7 Comparison to existing techniques 190

4.7.1 Creation of a synthetic dataset 190

4.7.2 Means of assessing grouping method performance. 194

4.7.3 Results 196

4.8 Discussion and conclusions 199

4.8.1 Chromatographic retention time variance and grouping method performance 199

4.8.2 Potential for application in real-time 200

4.8.3 Application to datasets with increased biological variance 202

Chapter 5: In-solution databases to facilitate rapid metabolite identification in metabolic profiling studies 204

Objectives 204

5.1 Introduction 204

5.2 Methods 207

5.2.1 Preparation of individual chemical standards and standards mixtures 207

5.2.2 Acquisition of reference spectra 207

5.2.3 Generation and interpretation of a spectral library 209

5.2.4 Development and implementation of a deconvolution script for the assignment of mixed standards 210

5.2.5 UPLC-MS profiling of human urine from a bariatric surgery subject cohort and standards mixtures 213

5.2.6 Pre-processing, multivariate statistical analysis, and biomarker discovery 214

5.3 Results 215

5.3.1 Collection of reference spectra for a physical properties database 215

5.3.2 Implementation and testing of automated mixture deconvolution and annotation 219

5.3.3 Evaluation of the master standards mixture in comparison to a biological sample 222

5.3.4 Use of a standards mixture for molecular identification in UPLC-MS profiling 223

5.4 Discussion 230

Chapter 6: General discussion and future work 232

References 235

References 236

Appendices 245

Appendix 1 245

Appendix 2 248

Appendix 3 250

Appendix 4 274

Appendix 5 281

Appendix 6: Publications (2012-2014) 298

List of Tables

Table 2-1. User defined parameters for LC-MS feature extraction using centWave in XCMS. 55

Table 2-2. The data matrix product of feature extraction and grouping, showing the feature intensity for six distinct features across nine hypothetical samples. 56

Table 3-1. List of 57 high priority standards assessed for inclusion in the development of RPC and HILIC test mixtures. 75

Table 3-2. Chemical reference standards composition of the RPC SSTM and HILIC SSTM. 77

Table 3-3. Chromatographic gradient of the reference method showing the duration and mobile phase composition of each step. 80

Table 3-4. CentWave parameters used for feature extraction within XCMS. 81

Table 3-5. Comparison of the chromatographic column volumes used in the reference and optimised RPC methods. 94

Table 3-6. Chromatographic gradient of the optimised RPC method, showing the programmed gradient times and mobile phase composition (A = H2O + 0.1% formic acid; B = acetonitrile + 0.1% formic acid). 96

Table 3-7. XCMS parameters for the extraction of features from urine LTR analyses using both the reference and optimised chromatographic RPC methods. 99

Table 3-8. Chromatographic gradient of the HILIC reference method showing the duration and mobile phase composition of each step. 103

Table 3-9: Chromatographic gradient of the optimised HILIC method, showing the programmed gradient times and mobile phase composition. 106

Table 3-10: Comparison of chromatographic column volumes used in the reference and optimised HILIC methods 106

Table 3-11: Chromatographic method used for fractionation of the concentrated urine LTR sample. Programmed gradient times and mobile phase composition (A = H2O + 0.1% formic acid; B = methanol + 0.1% formic acid) are shown. 111

Table 3-12: XCMS parameters for the extraction of features from urine fractions analysed by both RPC and HILIC methods. 113

Table 3-13: Sample composition and distribution within the 96-well format. 121

Table 3-14: XCMS parameters for the extraction of features from urine analysed by LC-MS using the optimised RPC method. 123

Table 3-15: Finalised RPC (left) and HILIC (right) chromatographic methods after extending to 14.65 minutes (for a 15 minute injection-to-injection cycle with 0.35 minute inter-analysis delay) and distribution of excess time to key method steps. 141

Table 3-16: Detector gain voltages applied during the first and second batches of data acquisition for each instrument and method type. 143

Table 4-1. XCMS parameters for the extraction of features from urine analyses by LC-MS using the optimised RPC method developed in Chapter 3 159

Table 4-2: The feature matching (link) ledger, abbreviated, for hypothetical feature sets from three samples. 188

Table 4-3. Volume of feature groups produced by each grouping method on each dataset (A-D) expressed in terms of percentage of the true number of groups (12,756). 197

Table 4-4. Tabulation of observed feature groups (of 12756 true groups) created by applying three feature grouping methods to four synthetic datasets. 198

Table 4-5. Precision and recall values for the nearest and ROgroup methods of feature grouping across datasets A-D. The maximum values per dataset are highlighted in green. 199

Table 5-1. A rapid elution program for the high throughput generation of reference spectra from chemical standards. 208

Table 5-2. XCMS centWave parameters for feature detection. 215

Table 5-3. Number of standards (from 55 total) with observed monoisotopic parent mass, fragment mass(es), adduct mass(es), and adduct fragment mass(es). 218

List of Figures

Figure 2-1. A cartoon depiction of chromatography. 31

Figure 2-2. Chromatographic dispersion illustrated in a van Deemter plot. 34

Figure 2-3. A U/HPLC pump system for column chromatography. 35

Figure 2-4. Measurement of peak width at the peak base (A) and at half of the peak height (B)..... 37

Figure 2-5. Measurements required for the calculation of chromatographic resolution. 38

Figure 2-6. Ideal distribution of chromatographic peaks illustrating the maximum number of peaks that can be fitted into a chromatogram with a resolution of one. 39

Figure 2-7. Histogram (1 Da. bins) of compounds in the Human Metabolome Database v2.5 (Wishart et al., 2009) with the mass range 1-1200 Da. 41

Figure 2-8. A two-dimensional screen capture of interactive three-dimensional data obtained by LC-MS analysis of a single urine sample. 49

Figure 2-9. A two-dimensional projection of three-dimensional data obtained by UPLC-MS analysis of a single urine sample..... 50

Figure 2-10. TIC visualisation of a human urine sample analysed by UPLC-MS. 51

Figure 2-11. Comparison between TIC and BPI visualisations for the same human urine sample analysed by UPLC-MS. 52

Figure 2-12. EIC of a selected m/z, with inset mass spectra. 53

Figure 2-13. PCA scores plot of an analytical reproducibility intervention study. 59

Figure 2-14. OPLS-DA loadings “S-plot” (A) and variable line plots of selected features (B). 62

Figure 3-1. A schematic of the basic structure of large-scale UPLC-MS profiling analysis 70

Figure 3-2: Strategy for selection of chemical standards for reference testing..... 74

Figure 3-3. Venn diagram illustrating overlap of chemicals used in standardising LC-MS instrument and chromatographic performance across three research groups. 78

Figure 3-4. TIC traces of a representative urine separation by the method of Want and colleagues (purple trace) and by isocratic elution at initial conditions (green trace). 80

Figure 3-5. Distribution (smoothed count) of chromatographic peaks detected (N = 5103) using the centWave method in XCMS. 82

Figure 3-6. Column efficiency calculated using the average values of three replicate injections of the standards mixture, run on three independent columns of each length (100mm and 150mm)..... 84

Figure 3-7. Chromatographic separation of the urine LTR using a 2.1 x 150mm Acquity HSS T3 reversed-phase column and three distinct methods. 85

Figure 3-8. Assessment of feature density within a linear separation of the urine LTR. 87

Figure 3-9. The relationship between UPLC system pressure and mobile phase flow rate. 90

Figure 3-10. Chromatographic peak area of selected SSTM reference standards decreases with respect to increased mobile phase flow rate. 92

Figure 3-11. System pressure traces of the optimised (green) and reference (red) methods. 93

Figure 3-12. Cytidine retention as a function of column equilibration time at initial chromatographic conditions (99:1 water-to- acetonitrile with 0.1% formic acid added). 95

Figure 3-13. Selected SSTM reference standards were extracted from a single analysis by the reference method (top) and optimised method (bottom). 98

Figure 3-14. Density of detected features in relation to chromatographic retention time using the reference method. 100

Figure 3-15. EIC of a metabolite cluster (m/z=153.058 +/- 0.1 Da) spanning the area of chromatographic distortion caused by one minute of isocratic elution at initial conditions. 101

Figure 3-16. Overlay of the feature density distribution between the reference method (grey) and optimised method (blue), illustrating the approximately equivalent footprint of chromatographic elution between the reference and optimised methods..... 102

Figure 3-17. HILIC separation of the HILIC SSTM. 108

Figure 3-18: Two-dimensional plots of orthogonal LC-MS urine separations. 109

Figure 3-19: Illustration of a selected fraction (fraction 14, shown in green) analysed by RPC (top) and HILIC (bottom) chromatography. 112

Figure 3-20. Heat map representation of the density of detected features in 0.25 minute retention time bins (x-axis) for each fraction (y-axis) by both RPC and HILIC optimised methods. 114

Figure 3-21: The molecular contents of representative fractions 17 (top) and 49 (bottom) derived from RPC separation of LTR urine using methanol as the strong eluent are mildly dispersed by the analytical RPC method utilising acetonitrile. 116

Figure 3-22: Illustration of an electrospray probe angle in relation to the MS inlet cone. 120

Figure 3-23: Normalised observed signal (integrated peak area) of selected reference chemicals from the RPC SSTM analysed by RPC using electrospray ionisation with varying probe angle relative to the MS source cone. 120

Figure 3-24: The total number of features detected in the 16 SR samples of each plate of 96 sample analyses, across 9 sequential plates in total. 124

Figure 3-25: The plate-wise distribution of CV observed for featured extracted from each plate's 16 SR samples. 125

Figure 3-26: PCA scores plot showing the distribution of SR samples (coloured by plate number) across principal components 1 and 2, accounting for 69.1% and 6.4% of the total dataset variance, respectively. 126

Figure 3-27. Variable line plots illustrating the intensity of two selected features as measured across all SR samples in each of 9 plates, shown in ascending order (left to right). 128

Figure 3-28. EICs of feature m/z = 358.259 @ 6.4 min from the first SR samples analysed in plates 1 through 4 (bottom to top). 129

Figure 3-29: Illustration of feature behaviour related to run order and sample age. 131

Figure 3-30: A graphical illustration of a 24 hour period (00:00 to 24:00) annotated with the assumed components of the working day. 134

Figure 3-31: Illustration of sample batch reload times, assuming analysis is initiated at 3pm for analytical cycle times between 2 and 32 minutes. 136

Figure 3-32: Simulation revealing the maximum age accumulated by a sample during the continuous analysis of 20 96-well plates for analytical cycle times between 2 and 32 minutes. 137

Figure 3-33: Within the HILIC analysis of LTR urine (TIC shown in red, top), m/z=170.093 (EIC shown in purple, bottom) is the latest eluting species observed. 140

Figure 3-34: TIC traces from the first (top) and last (200th, bottom) HILIC+ urine LTR analyses in the set of 25 randomised 96-well plates, approximately 2400 injections apart, showing similarity. 144

Figure 3-35: TIC traces from the 33rd (top) and last (200th, bottom) RPC+ urine LTR analyses in the set of 25 randomised 96-well plates, approximately 2016 injections apart, showing remarkable similarity. 145

Figure 3-36: TIC traces from the 8th (top) and last (200th, bottom) RPC- urine LTR analyses in the set of 25 randomised 96-well plates, approximately 2300 injections apart, showing remarkable similarity. 146

Figure 4-1. LOESS retention time deviation curves for 96 sequential urine analyses from plate 1 of the test sample set. 160

Figure 4-2. Panel A illustrates two possible patterns of analyte retention drift due to system instability. 162

Figure 4-3. Heterogeneity in chromatographic feature drift. 163

Figure 4-4. A representative feature cluster (m/z = 310.2015 +/- 0.0025) from the test set SR urine analyses shown as an EIC and as featured detected by XCMS centWave. 165

Figure 4-5. A cluster of features (m/z = 314.23) exhibiting the same retention migration behavior (A) with respect to increasing run order. 166

Figure 4-6. A cluster (m/z = 314.23) exhibiting retention migration behavior with respect to increasing run order is illustrated (A) along with cartoon representations of feature density and binning approaches to feature grouping (B) and the proposed pairwise approach to grouping in run order (C). 168

Figure 4-7. Pairwise agreement in feature retention between sequential quality control samples from the early, mid, and late portions of the 864-sample analysis. 170

Figure 4-8. For a set of samples, iterative rounds of pairwise matching (purple) are conducted between feature sets from sequential sample analyses. 172

Figure 4-9. A single feature cluster detected in a representative sample of human urine. 174

Figure 4-10. Spatial orientation of large clusters in a representative reversed-phase LC-MS dataset. . 175

Figure 4-11. Representative large clusters of 10 or more features from early, mid, and late retention times (top to bottom, respectively.) 177

Figure 4-12. Tolerance windows for inter (blue) and intra (green) sample feature matching. Red dotted lines denote matches. 178

Figure 4-13. Community boundary definition by iterative collation of clustered and matching features. 180

Figure 4-14. Feature community #627, illustrated as EIC peaks and as detected features. EICs of m/z = 310.2015 +/- 0.01Da in the first (red) and last (green) SR analyses from plate 1 of the test set, representing the template and candidate for matching, respectively (top). 182

Figure 4-15. A heatmap of the retention time differences (in seconds) among all candidate and template features in community #627. 183

Figure 4-16. Pattern matching of feature clusters within community #627. 184

Figure 4-17. A heatmap of the community-optimised retention time differences (in seconds) among all candidate and template features in community #627. 185

Figure 4-18. A masked heatmap of the community-optimised retention time differences (in seconds) among all candidate and template features in community #627. 186

Figure 4-19. A heatmap of the m/z value differences among all candidate and template features in community #627. 187

Figure 4-20. Two dimensional plot of feature groups created using the ROgroup method on the first 16 SR urine analyses from the test dataset. 190

Figure 4-21. The programmed logarithmic drift in retention time for a given feature as a function of run order was applied to a feature from the dataset, along with random noise, to produce a realistic retention profile across the sample set. 193

Figure 5-1. Envisioned workflow for the use of an in-solution database (mixed standards) and known physical properties (e.g mass spectra) of the mixture components (physical properties database) for the *de novo* generation of an empirical database. 206

Figure 5-2. Visual algorithm of the deconvolution workflow. 212

Figure 5-3. Chromatographic separation of a representative standard from potentially obscuring salt. 216

Figure 5-4. Interpretation of spectral data from four interleaved DDA MS/MS acquisitions for the UPLC-MS analysis of cytidine 5'-monophosphate. 217

Figure 5-5. Selected results from the feature matching procedure between expected spectral features from the physical properties database and the UPLC-MS data acquired by profiling of the subset standards mixture. 220

Figure 5-6. Retention reference of N-acetyl-cysteine. 221

Figure 5-7. Base-peak intensity (BPI) chromatograms of a synthetic standards mixture (top) and a composite sample of human urine (bottom) from a pre vs. post bariatric surgery study. 223

Figure 5-8. A PCA scores plot of principal components 1 vs. 2 of urine samples from a cohort of subjects before and after surgical intervention. 225

Figure 5-9. OPLS-DA loadings plot of the feature-set from pre- and post-bariatric surgery subjects, and manual interpretation of selected features elevated in post-surgery patient samples. 226

Abbreviations

BPI	Base peak intensity (chromatogram)
C18	Octadecyl carbon chain
CID	Collision-induced dissociation
EIC	Extracted ion chromatogram
ESI	Electrospray ionisation
FWHM	Full width at half maximum
GWAS	Genome-wide association study
HILIC	Hydrophilic interaction chromatography
HPLC	High performance liquid chromatography
HSS	High strength silica
LC-MS	Liquid chromatography-mass spectrometry
LTR	Long term reference
m/z	Mass-to-charge (number) ratio
MR	Method reference
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MWAS	Metabolome-wide association study
NMR	Nuclear magnetic resonance
NPC	MRC-NIHR National Phenome Centre
OPLS	Orthogonal projection to latent structures
OPLS-DA	Orthogonal projection to latent structures discriminant analysis
PC	Principal component
PCA	Principal components analysis
PEG	Polyethylene glycol
PLS	Partial least squares
ppm	Parts per million
Q	Quadrupole
QC	Quality control
Q-ToF	Quadrupole time of flight
RPC	Reversed-phase chromatography
RT	Retention time
SSTM	System suitability test mixture
TIC	Total ion chromatogram
ToF	Time of flight
UPLC	Ultra performance liquid chromatography

General Introduction

The work in this thesis addresses the optimisation of ultra performance liquid chromatography mass spectrometry (UPLC-MS) systems and methods as well as the development of data processing and metabolite identificaion workflows required by the scope and scale imposed by recent initiatives in screening of large populations. For a general introduction to liquid chromatography, as an underpinning of all chapters contained herein, the reader is directed to the excellent work of Snyder, Kirkland, and Dolan: Introduction to Modern Liquid Chromatography, Third Edition (Snyder et al., 2010).

This research is dedicated to facilitating the molecular profiling of human urine samples with the goal of developing a comprehensive image of a wide range of human phenotypes. The chemical density and rich diversity of urine as a key human biofluid provides an analytical and informatics challenge which must be met in the context of modern high throughput analysis in order to realize the goal of population screening of hundreds and thousands of individuals. Such a project requires a dedicated and fit-for-purpose pipeline for sample analysis and data processing. To construct this pipeline, high performing analytical methods are advantageously deployed across a large instrument resource to support the volume of analysis demanded by molecular epidemiology, to harness the power of large population studies, and to support metabolome-wide association studies. The efficient operation of such a platform is critical to maximise both the volume and quality of output. For this reason, much of the focus herein is on constraint-guided development, whereby the laboratory working environment is considered alongside sample integrity and analytical performance in order to produce a holistic approach to population phenotyping. The datasets generated are consequentially characterised by high precision despite the large-scale of application, yet remain simple to execute, rapid to collect, and efficient.

Furthermore, the development of data processing strategies that work in conjunction with high throughput analysis play a large part in supporting the overall efficiency achieved within the laboratory. The scale of the data produced by high resolution profiling can be staggering, often precluding rapid assessment of a large dataset post-acquisition. However, the development of real time pre-processing

of complex profiling data described herein lays a foundation for real-time multivariate analysis, which could in turn be utilised in monitoring of data quality, control of data acquisition, and in targeting signals of potential interest for advanced analysis as they appear. The desire for these abilities and improvements to molecular profiling warrants steadfast effort at the development of underlying and enabling methodology.

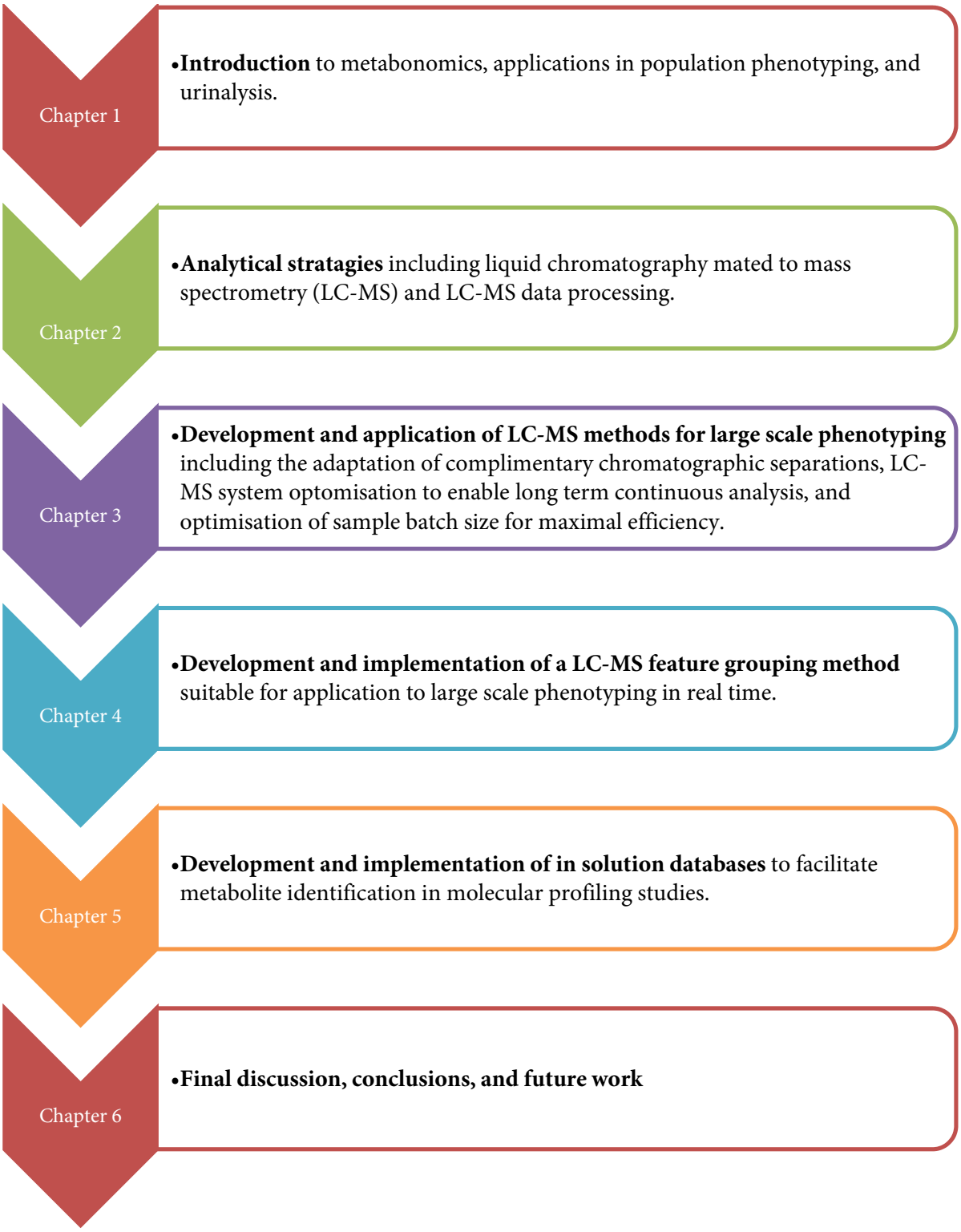
Objectives

The overall aim of the work comprising this thesis is to deliver an advanced analytical UPLC-MS and informatics platform capable of supporting the phenotypic characterisation of large populations and sample banks in order to extract metabolically relevant information relating to disease risk and prevalence.

The specific objectives of this project are:

- To develop advanced methods for molecular profiling of human urine by UPLC-MS which are fit for the purpose of application to large population cohorts
- To characterise the analytical variance and molecular coverage of those methods
- To demonstrate the performance and applicability of those methods in the context of human phenotype analysis.
- To develop and implement a method of feature grouping across samples that accommodates the observed analytical variance in large sample set profiling, and is suitable for real-time application.
- To demonstrate an approach to confident metabolite identification by multiplexing reference materials in complex mixtures that are deconvolved on a per-experiment basis to form *de novo* databases.

Thesis Structure



Chapter 1: Introduction

1.1 Metabonomics and applications in population phenotyping

The field of measuring small molecules in biofluids and tissues with respect to disease or treatment states, commonly known as metabonomics, is a modern approach to biochemical assessment of human health and metabolic status. Definitively summarised as "the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification" (Nicholson et al., 1999), metabonomics is a concept underpinned by the analytical technology and data processing methods required to both capture and interpret detailed information across complex systems (Lindon et al., 2004, Patti et al., 2012b, Kaddurah-Daouk et al., 2008). The analytical challenge of the metabonomic approach is to produce measurements with both appropriately broad coverage and sufficient resolution for differentiating hundreds or thousands of individual metabolites. Biofluids such as urine, blood, and faecal water are rich with chemical imprint of metabolism, as are solid tissues and tissue extracts (Coen et al., 2008, Beckonert et al., 2010, Want et al., 2010, Monleon et al., 2009, Waldram et al., 2009, Keun and Athersuch, 2011). The challenges of deconvolving the chemical complexity of these common matrices and biologically meaningful interpretation of the resulting data continue to drive the development of advanced instrumentation, analytical methodologies, and automated data processing workflows.

Metabolite profiling analyses, unlike conventional clinical assays (*e.g.* for blood creatinine or urinary glucose), are not intended to be selective and are therefore applicable in the simultaneous measurement of both expected and unexpected metabolites that are not precluded by the choice of instrumentation or analytical method utilized (Lindon and Nicholson, 2008). This emphasis on inclusion makes profiling approaches essential in the pursuit of achieving comprehensive analytical coverage of the human metabolic phenome, encompassing supraorganismal metabolites from human and associated microbial action on environmental substrates, xenobiotics, and environmental contaminants. Metabolic profiles are therefore a key reflection of human individuality (Assfalg et al., 2008, Bernini et al., 2009), and have been shown to be highly variable as a consequence of the complex interactions of a

person's genetically coded metabolic machinery with their environment (Krug et al., 2012, Dallmann et al., 2012). However, the resulting phenotypes may appear more homogeneous among groups of people with similar genetics and/or environment than among those with differing background or exposures (Holmes et al., 2008, Yousri et al., 2014).

In recent years, the search for phenotypic relationships among groups of individuals has taken the form of molecular epidemiology, whereby subtle metabolic effects may be observed thanks to the statistical power afforded by population-level sample collection and analysis (Menni et al., 2013, Nicholson et al., 2011). When paired with broad metabolite profiling analytical chemistry, these large-scale analyses are able to generate unprecedented power in metabonomic comparisons and ultimately phenotype elucidation (Tzoulaki et al., 2014). The demand for metabolic profiling of biofluids from large subject cohorts is therefore increasing as epidemiologists turn to metabonomics as a maturing science capable of providing broad phenotypic insight with metabolome-wide association studies (MWAS) providing a channel for placing the analogous genome-wide association studies (GWAS) in context.

To meet this need, there exists a more fundamental requirement for high quality analytical data (Bictash et al., 2010). Nuclear magnetic resonance (NMR) spectroscopy has long been a favoured analytical platform for the generation of metabolic profiles with high precision, facilitating comparisons among individuals or groups of individuals within populations (Larive et al., 2014, Dona et al., 2014, Nicholson et al., 1984). However the technique is limited in terms of its ability to discern individual molecules from complex mixtures with high sensitivity and chemical specificity, driving the development and application of complementary instrument platforms. Liquid chromatography mated to mass spectrometry (LC-MS) has since emerged as a viable alternative approach for biofluid analysis, boasting high resolution multi-dimensional separations and sensitive detection across a broad range of chemical species (Want et al., 2010, Dunn et al., 2011). Yet, the LC-MS platform is not renowned for absolute precision, owing in part to the complexity of the hyphenated system involving the distinct processes of high pressure liquid separation followed by analyte desorption and finally mass spectrometric manipulation and detection. High coefficients of variation from LC-MS measurements have been reported when attempting analysis of samples from cohorts over 1000 patients (Swann et al.,

2013), indicating the difficulty in achieving stable metabolic signatures in large-scale analysis with this otherwise powerful platform. Yet, the allure of epidemiological-scale datasets that are high quality and comprehensive continues to drive the development of LC-MS approaches for large-scale biofluid characterisation (Zelena et al., 2009, Broeckling et al., 2013) as well as the development of informatics approaches required to combat seemingly inevitable analytical imprecision (e.g. sample batch effects) (Wang et al., 2013, Vaughan et al., 2012, Nezami Ranjbar et al., 2013). This is the state-of-the-art which must be advanced for the successful realisation of population phenotyping.

1.2 Urinalaysis

Within this thesis, urine is used exclusively as a representative biofluid for the design and testing of analytical and data processing methods. Its selection is both practical and strategic, as urine is rich with information related to human phenotypes as well as being non-invasively collected and therefore commonly available in molecular epidemiology studies. Urinalaysis for the assessment of health and detection of disease is a long established if not ancient practice (Bolodeoku and Donaldson, 1996, Ahmed, 2002) with the diagnosis of diabetes cited as the first of all laboratory tests (Haber, 1988). Sir Archibald Garrod aptly demonstrated the value of metabolic information captured in the urine in his landmark descriptions of alkaptonuria (Garrod, 1902) and later of cystinuria and pentosuria (Garrod, 1909, Garrod, 1923), founding the idea of chemical individuality and historically linking metabolic disorders to heredity. While basic urinalysis continues to play an important role in clinical diagnosis of disease, the application of advanced analytical technologies capable of profile analysis have breathed new life into this otherwise established science (Law et al., 2014, Morell-Garcia et al., 2014, Ng et al., 2012). Application of these technologies to larger sample sets can facilitate the exploration of a wide variety of human phenotypes without necessitating *a priori* hypothesis, although prior knowledge can be used to optimise the analysis. These approaches are being successfully applied to develop a deeper metabolic, diagnostic, and prognostic understanding of many diseases (Zhang et al., 2013, Maitre et al., 2014, Luan et al., 2014, Austdal et al., 2014).

Yet, global urinalysis by modern means remains an analytical challenge. Human blood is kept homeostatic at the expense of the urine, which exists to accept and rid the body of the undesirable

products of metabolic flux. The endogenous contents of urine are therefore subject to large changes in concentration, as well as the variable presence or absence of exogenous substances through the interaction of the subject with his or her environment. The result is an ever changing chemical matrix which we are virtually assured an incomplete knowledge of, despite centuries of characterisation as well as recent intensive efforts (Bouatra et al., 2013). Broad scope molecular profiling methods are therefore critical to urinary analysis in population scale research where a large number of phenotypes are represented but may be partly obscured by a high degree of human individuality resulting from varied genetic and environmental exposure. The methods applied must be highly precise to avoid confounding the metabolic patterns associated with these phenotypes even further due to analytical variance, and to allow direct comparison of data obtained from hundreds or thousands of individuals. To achieve this, the challenges associated with sensitive and specific detection of the components of such a complex mixture, spanning multiple orders of magnitude in concentration and potentially laced with unknown substances, must be addressed. Furthermore, successful molecular profiling of human urine therefore requires multiple coordinated analytical methods which are high performing in their own right, as the chemical diversity of urine prohibits any one method from successfully achieving comprehensive metabolite coverage. The collation of such methods, together with fit-for-purpose data processing tools, is pursued here for the efficient analysis of urine to facilitate deeper understanding of human phenotypes.

Chapter 2: Analytical strategies

2.1 Introduction

Any instrument capable of measuring fundamental physical and/or chemical properties of small molecules is eligible for use in urinalysis. As chemical mixtures grow in complexity, the specificity required of the instrument increases commensurately. Sensitivity is a fundamental criterion as many important small molecules such as hormones are present in biofluids in very small quantities. Furthermore, a large dynamic range is desired for accurate quantification of metabolites across a wide physiological range (Lentner, 1981). Few analytical detection platforms perform with sufficient specificity, sensitivity, and range to be useful in metabonomic studies. Of these qualifying platforms, nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) have emerged as the top performing candidates (Lindon and Nicholson, 2008) as evidenced by their ubiquitous application in the field (Coen et al., 2008, Nevedomskaya et al., 2011, Lenz and Wilson, 2007, Skogerson et al., 2009, Athersuch et al., 2010). Although no single technique is without limitation, the molecular coverage and specificity, as well as excellent dynamic range in measurement, propels these tools forward in metabonomic application on the basis that they are unmatched in their ability to deconvolve hundreds or thousands of signals from complex biological matrices (Fernie et al., 2004, Lu et al., 2010, Buscher et al., 2009). Mass spectrometry, when coupled to a liquid chromatographic system, is particularly apt in the deconvolution and measurement of complex molecular mixtures (Buscher et al., 2009), and is increasingly relied upon in metabonomic research and population screening (Swann et al., 2013, Maitre et al., 2014, Wang et al., 2011). Development and application of the combined (or "hyphenated") LC-MS system is therefore the central focus of the work presented herein. However, in order to provide an appropriate foundation and context for the development of the system as a whole as well as the methods for extracting the data produced, an introduction of each component technique is warranted.

2.2 Fundamental principles of column-based liquid chromatography

Liquid chromatography, while unable to directly detect the presence and quantity of metabolites, is an important tool for the separation of complex molecular mixtures over time. This distribution of metabolite content can make downstream detection and measurement more sensitive, more specific, and more accurate, thereby contributing to the fundamental tenets of metabonomic measurement. The chromatographic environment generally consists of two distinct but contacting phases; one of which is immobile (the stationary phase) and the other being mobile (mobile phase). Chemical solutes introduced to this environment will have a differential affinity for each phase depending on the chemical composition of the system. Some analytes may preferentially interact with the stationary phase and therefore be immobilised or *retained*, while others may preferentially interact with the mobile phase and be carried through the system, or *eluted*. For any given analyte species, the ratio of its distribution between stationary and mobile phases is defined as its retention factor.

Column chromatography is a common format for the application of this principle, whereby an empty cylinder is filled with particles (typically composed of silica or polymer) which act as the stationary phase, and a liquid solvent acting as the mobile phase is passed through it. A liquid sample may be directly injected to the pre-column flow of mobile phase creating a *band* of solutes that is carried through the column for separation (Figure 2-1). Thus, the system is particularly well suited for the separation of the non-volatile molecular species that compose the majority of human biofluids. The chemical selectivity of each phase determines the extent and selectivity of the retention and separation of molecular species as the individual component species travel through and eventually emerge from the column as *peaks*. The distribution, width, and shape of these chromatographic peaks are related to the performance of the chromatographic separation.

2.2 Fundamental principles of column-based liquid chromatography

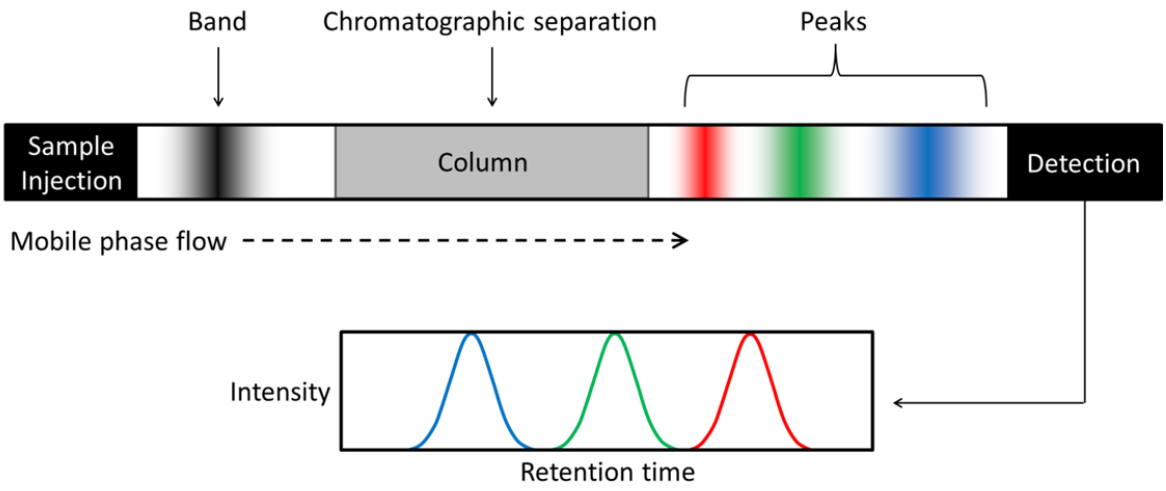


Figure 2-1. A cartoon depiction of chromatography. Sample is injected into a flow path of mobile phase creating a band which is separated into its components according to their differential interactions with the column. The components are eluted as peaks and passed onward for detection.

Chromatographic separation performance is largely dependent on the efficiency achieved by the chromatographic column used. Column efficiency is characterised by the presence of theoretical plates which are a concept originally used to describe stages of contact between the vapour and liquid phases in distillation columns. Theoretical plates were first adapted for use in describing the theory of chromatographic separations by Martin and Synge (Martin and Synge, 1941), and characterise a column's efficiency in terms of its plate count (N) whereby greater efficiency separations result from higher values of N. Plate count itself is a function of the plate height (H, also defined at the height equivalent to one theoretical plate or *HETP*) and the column length (L), meaning greater separation efficiencies are achieved by columns that are longer (for a given value of H) and/or more plate-dense (for a given length). Defined mathematically, we arrive at equation 2.1.

$$N = \frac{L}{H} \tag{2.1}$$

Plate height is antagonised by the diffusion of solutes in a liquid, and thus broadening of a sample band's constituent analyte peaks reduces the effective separation efficiency. This blurring of the separation otherwise produced limits the quality of results obtained from column chromatography. The extent of the overall diffusion is partly dependent on the time allowed for the process to occur, whether it precedes the column (extra-column band broadening), comes after the column (extra-column peak broadening), or occurs while in the column by a process known as *longitudinal diffusion*. In each of these cases, performing faster separations reduces the amount of time available for diffusion, producing narrower chromatographic peaks by minimising the processes of longitudinal diffusion and peak/band broadening. One common approach to producing faster separations is increasing the linear velocity of the mobile phase (*i.e.* the speed at which the solvent front travels the length of the packed column) by increasing the flow rate. However, the gains in peak sharpness produced by increasing the linear velocity of a separation (thereby limiting longitudinal diffusion) are antagonised by a second type of diffusion caused by resistance to *mass transfer* in both the stationary and mobile phases. As the stationary phase particles used in column chromatography are generally semi- or entirely porous (increasing the available surface area for mobile phase and sample contact), analytes may migrate through them at different rates. Molecules of a given species that migrate further into the pores of a particle have a slower overall forward progress through the column when compared with molecules that do not penetrate the particle pores as deeply, resulting in peak broadening. Furthermore, analytes in the centre of mobile phase streams in and around particles will be carried with higher velocity than those closer to the surrounding surfaces. At faster flowrates, the effects of resistance to mass transfer are exacerbated. Because of this interplay between longitudinal diffusion and mass transfer, a tuneable flow rate is required to achieve an optimum of narrow peak shape (and therefore high chromatographic resolution) in minimal time.

Not all diffusion is so largely influenced by mobile phase flow, however. *Eddy diffusion* occurs when the molecular contents of the sample travel different paths through the particle bed of the stationary phase, experiencing varying constriction and thus achieving different relative migration speeds. This type of diffusion can occur with approximate independence from the absolute rate of mobile phase

flow, and can therefore be thought of as a characteristic of a given column. Columns that are well packed, utilising smaller and more uniformly sized particles, will suffer less eddy diffusion according to the relationship shown in Equation 2.2, where λ is the packing factor (a measure of the flow inequality in a packed column, and therefore reflective of the quality of packing) and d_p is the particle size diameter.

$$H = 2\lambda d_p \tag{2.2}$$

Unless otherwise specified, all chromatographic separations conducted within this thesis utilise columns packed with the smallest commercially available class of stationary phase particles (less than 2 micrometres in diameter), thereby limiting the contribution of eddy diffusion to band broadening and increased values of H. Additionally, columns with a narrow inner diameter yield less independent paths of travel for eluting molecules, further reducing the contribution of eddy diffusion to increased H. Therefore, unless otherwise specified, all chromatographic separations conducted within this thesis utilise columns of relatively small (2.1 mm) inner diameter.

Together, these three means for chromatographic peak dispersion (eddy diffusion = A term; longitudinal diffusion = B term; mass transfer = C term) are captured in the well-known equation derived from the work of van Deemter (van Deemter et al., 1956) which relates the velocity of mobile phase (v) to the achieved chromatographic efficiency (for a given column length) in terms of the empirical quantity H where lower values indicate higher resolution separations (Carr and Sun, 1998).

$$H = A + \frac{B}{v} + Cv \tag{2.3}$$

Although subsequent derivative works (most notably that of Knox (Bristow and Knox, 1977) which makes the A term weakly dependent on mobile phase velocity) further refine the relationships between peak dispersion and mobile phase linear velocity, the more simple expression above is sufficiently

descriptive for the intended purpose of illustrating the basic relationships among terms. This information is represented visually by the composite curve shown in Figure 2-2.

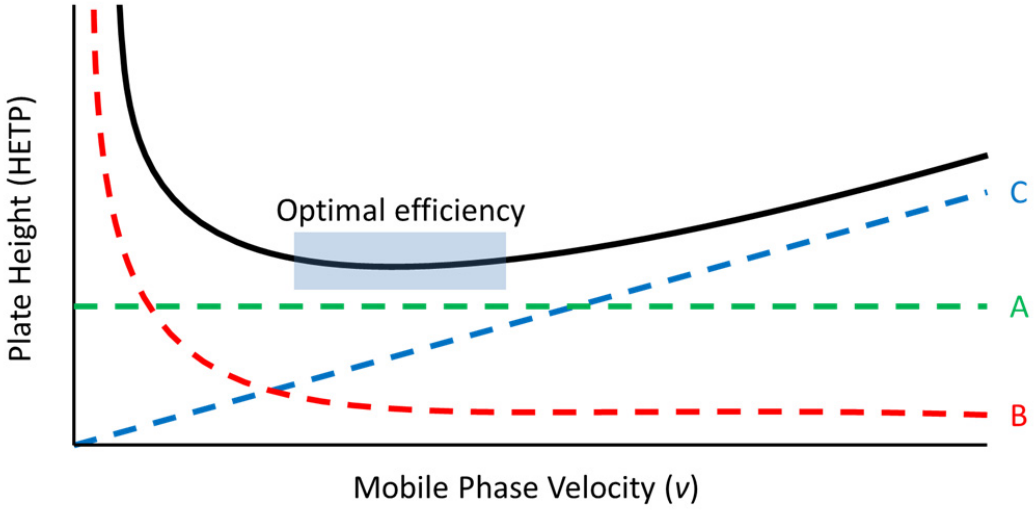


Figure 2-2. Chromatographic dispersion illustrated in a van Deemter plot. A curve of theoretical plate height (HETP) values achieved for a continuum of mobile phase velocities (v), constrained by the three types of diffusion captured in the van Deemter equation: eddy diffusion (A), longitudinal diffusion (B) and resistance to mass transfer (C).

A range of linear velocities therefore exist whereby the collective effects of peak broadening are minimised, achieving optimal efficiency for a given chromatographic system configuration. This range may be extended in the beneficial direction of faster mobile phase velocity (thereby enabling faster analyses) by reducing the contribution of the mass transfer term (C), which in turn is related to the size of the stationary phase particles. The use of small particles reduces both the mass transfer term and, as previously mentioned, the eddy diffusion term (A) allowing for both wider ranges of optimal mobile phase velocity and lower overall values of H. These combined effects allow high performance separations to be conducted in a reduced amount of time thanks to the use of faster mobile phase flow rates without sacrificing chromatographic efficiency (Mazzeo et al., 2005). However, as the particle size decreases, and/or as the mobile phase linear velocity increases, the force required to move the mobile phase increases. Together, these phenomena impose a practical limit on the performance achievable by a chromatographic system which is dependent on its ability to apply force.

2.2 Fundamental principles of column-based liquid chromatography

In cases where the particle size is relatively large and the rate of solvent flow is not required to be high, the force of gravity is sufficient to move the mobile phase vertically through a column (gravity chromatography). However, as the particle size shrinks, and/or as higher linear velocities are desired, liquid pumping systems are commonly applied to generate the requisite force. Ultimately the amount of force required to move the mobile phase is dependent on parameters including the desired speed of the separation, the particle size of the stationary phase material contained within the column, the dimensions of the column, and the viscosity of the mobile phase. As higher linear velocities and therefore greater force (perceived by the pump system as *back pressure*) are able to produce higher performance results characterised by better resolved peaks in shorter periods of time, this approach to column chromatography is known as high performance liquid chromatography (HPLC) or less frequently as high *pressure* liquid chromatography, underscoring the interdependency of pressure and performance. A model system is illustrated in Figure 2-3. The pump systems used in HPLC are capable of sustaining constant liquid flow at pressures approaching 6000 psi (approximately 400 times the atmospheric pressure at sea level). Advanced pump systems capable of delivering constant liquid flow at pressures 2-3 times greater than the traditional limit of HPLC are generally used in conjunction with very small particle sizes (less than 2µm in diameter), generating very high separation efficiencies. This approach is therefore known as ultra-performance liquid chromatography (UPLC) (Plumb et al., 2004). It is used exclusively as the platform for LC separations throughout this thesis.

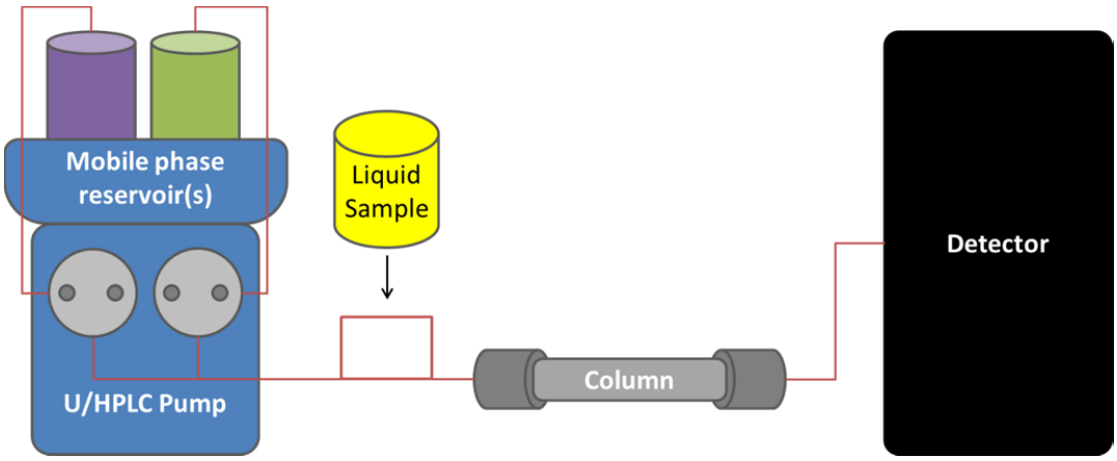


Figure 2-3. A U/HPLC pump system for column chromatography.

U/HPLC systems are capable of pumping either a single mobile phase solvent or mixing multiple solvents, giving rise to two types of chromatographic elution. The first, called *isocratic elution*, is performed when the mobile phase composition is homogenous for the duration of the separation, and is best applied to the separation of a few selected targets for which a mobile/stationary phase pair can be specifically formulated for optimal results. The second, called *gradient elution*, is performed by changing the mobile phase composition by mixing solvents in various proportions over the course of the separation in order to affect the retention factor of solutes. Modulation of this additional parameter contributes versatility to the system, increasing the range of chemical diversity over which a single stationary phase can be utilised for chromatographic separation. This principle is particularly favoured in molecular profiling of biological fluids where the solute diversity (*e.g.* range of polarity) is great. Finally, the composition of mobile phase solvents, as well as the composition of the stationary phase, affects the chemical selectivity of the separation. However these parameters can be constrained by the choice of detector, and therefore will be discussed in a subsequent section.

2.3 Methods for assessing chromatographic performance

Assessing the performance of a chromatographic separation is important for both the development and application of chromatographic methods. By reviewing the data generated as eluted peaks observed by the detector, an analyst may compare the relative capabilities of multiple methods, or evaluate the effects of a host of possible changes to the system such as mobile phase composition and velocity, stationary phase composition, and column dimensions. The following sections provide a foundation for the types of performance assessment used later in this thesis.

2.3.1 Defining chromatographic peak width

Baseline peak width (W_b) is measured as the distance between the baseline intersection points of tangents drawn through the peak inflection points as illustrated in Figure 2-4, method A. In practice, measurement of the full peak width at half of the maximum peak height (W_h) is more convenient and more precise as it does not require linear extrapolation from the points of inflection and mitigates the effects of peak distortions near the peak base (*e.g.* tailing). Measurement of W_h is illustrated in Figure 2-4, method B.

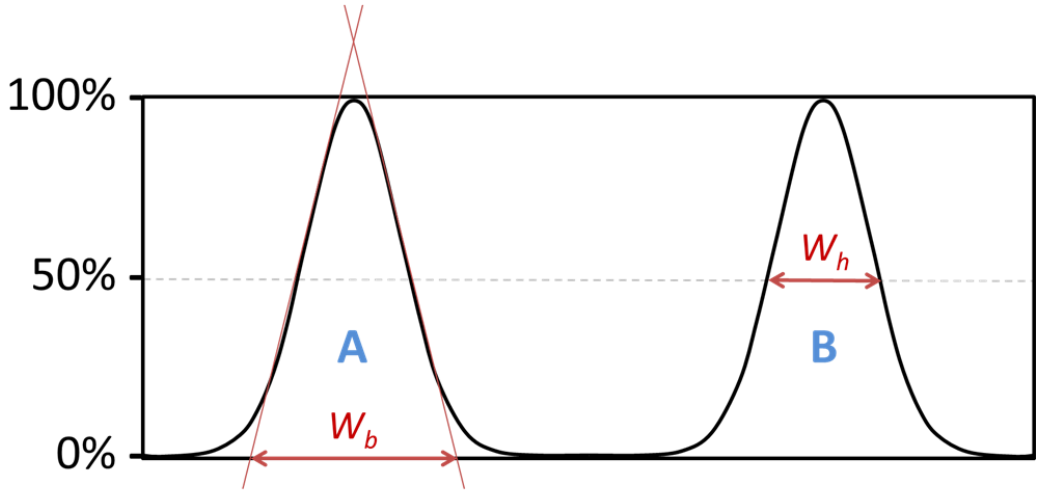


Figure 2-4. Measurement of peak width at the peak base (A) and at half of the peak height (B).

As a rule of thumb, values of peak width (W) may be estimated from measured W_h values using the conversion shown below (Snyder et al., 2010). The values of W reported herein are calculated in this manner.

$$W \equiv \frac{W_h}{0.588} \approx W_b \tag{2.4}$$

2.3.2 Measurement of chromatographic efficiency

Under isocratic separation conditions, the efficiency of a column is reported in terms of its number of theoretical plates (N). As illustrated in Equation 2.1, this value is equal to the quotient of column length (L) divided by H (the empirical value determined by the mobile phase velocity and the effects of dispersion previously discussed). H is therefore a measure of efficiency per unit length, while N describes the efficiency achievable with a given column. N may be determined empirically for a target analyte using the following equation, where t_R is the retention time of an example analyte and W is its peak width.

$$N = 16\left(\frac{t_R}{W}\right)^2 \quad (2.5)$$

Accounting for the equality between estimates of W from observed W_h values ($W_h \equiv 0.588W$), W_h can be directly utilised with the modified equation below. Chromatographic efficiency is calculated in this manner within this thesis.

$$N = 5.54\left(\frac{t_R}{W_h}\right)^2 \quad (2.6)$$

2.3.3 Measurement of chromatographic resolution and peak capacity

Chromatographic resolution, or the ability of a separation to distinguish independent solutes, is calculated as the difference in retention time values between two peaks (t_1 and t_2) divided by the average peak width, as illustrated in Figure 2-5 and the following equation.

$$R_S = \frac{t_2 - t_1}{(W_2 + W_1)/2} \quad (2.7)$$

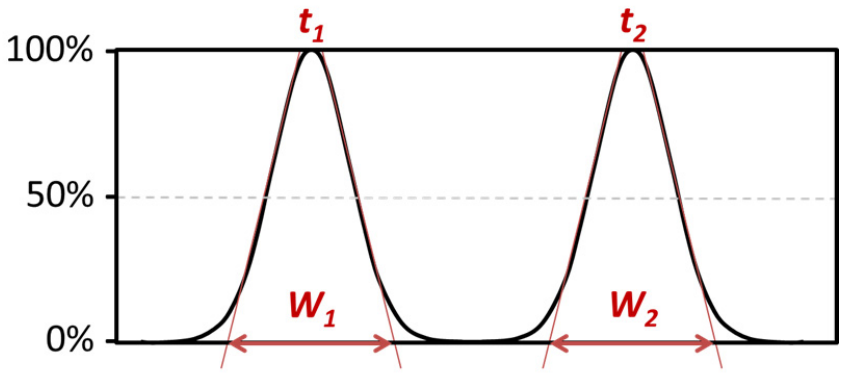


Figure 2-5. Measurements required for the calculation of chromatographic resolution.

The theoretical maximum number of peaks which may be resolved (with R_S=1) in a given separation is known as the peak capacity (P_C). Therefore peak capacity is a measure of the potential ability of a chromatographic separation to resolve solutes, assuming that they are evenly distributed with elution

time. It is most often applied for the assessment of gradient separations. This ideal spacing of chromatographic peaks is illustrated in Figure 2-6.

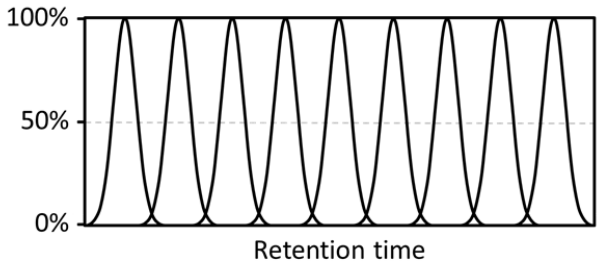


Figure 2-6. Ideal distribution of chromatographic peaks illustrating the maximum number of peaks that can be fitted into a chromatogram with a resolution of one.

A separation's P_C may be measured by addition of all resolution values between peak pairs in a sequential series spanning the duration of the gradient, starting with the injection peak (0) and ending with the last peak (l) as follows (Neue, 2008):

$$P_C = 1 + R_S(0,1) + R_S(1,2) + R_S(2,3) + \cdots + R_S(l-1,l) \quad (2.8)$$

The peak capacity of any gradient segment may be calculated by addition of all resolution values between the first (f) and last peaks that flank that segment, yielding a value known as the sample peak capacity (P_C^{**}) (Neue, 2008, Dolan et al., 1999):

$$P_C^{**} = R_S(f,f+1) + R_S(f+1,f+2) + R_S(f+2,f+3) + \cdots + R_S(l-1,l) \quad (2.9)$$

By carefully selecting the area of a gradient for inclusion, a more representative value of capacity may be obtained where the separation span of the matrix of interest is well defined.

Where peak width is approximately constant across the gradient, the measurement of sample peak capacity may be simplified by dividing the difference in retention time between the first and last peaks by the average peak width (W_{ave}) of those peaks and all between them.

$$P_C^{**} = \left(\frac{t_l - t_f}{W_{ave}}\right) \tag{2.10}$$

Greater peak capacities are characteristic of methods that have the potential to resolve more solutes, and are preferred in human biofluid molecular profiling applications where high sample complexity ensures a high density of analytes.

2.4 LC detection by mass spectrometry

Analytes eluting from a column require measurement by a detector. Spectroscopic detectors are well suited to this purpose as they are able to measure the content of a liquid flow, most commonly by measurement of the associated absorption or emission spectra (*e.g.* ultraviolet-visible spectroscopy or fluorescence spectroscopy, respectively). Such spectral data serve as both quantitative signals proportional to the concentration of the solute present and characteristic indicators of chemical composition. However these techniques tend to be selective in the chemicals they are able to detect (*e.g.* only those chemicals with chromophores or flurophores) and are therefore inherently limited in their capability to broadly capture data from chemically diverse mixtures. Furthermore, the chemical specificity afforded by spectroscopic analysis of complex mixtures can be lacking with many techniques as the presence of shared functional groups may confound the interpretation of spectral data. Therefore a more characteristic and descriptive property of molecular identity is desired for application to human biofluid screening.

Mass spectrometry (MS), the gas phase separation and detection of molecular ions (carrying an electric charge) by their mass and charge, is therefore an attractive alternate means for the specific detection of diverse chemicals. The wide natural distribution of mass among physiological metabolites (illustrated in Figure 2-7) inherently benefits the application of MS to complex biofluids such as human urine. Furthermore, pico- and femtogram sensitivity for many ionisable molecular species and an increasingly wide dynamic range make MS well paired to the analytical challenges faced in metabonomic applications. Indeed, chromatographic separation coupled to mass spectrometric detection is one of the staple platforms of metabonomic research (Want et al., 2010, Lenz and Wilson, 2007, Bajad et al., 2006, Rainville et al., 2007).

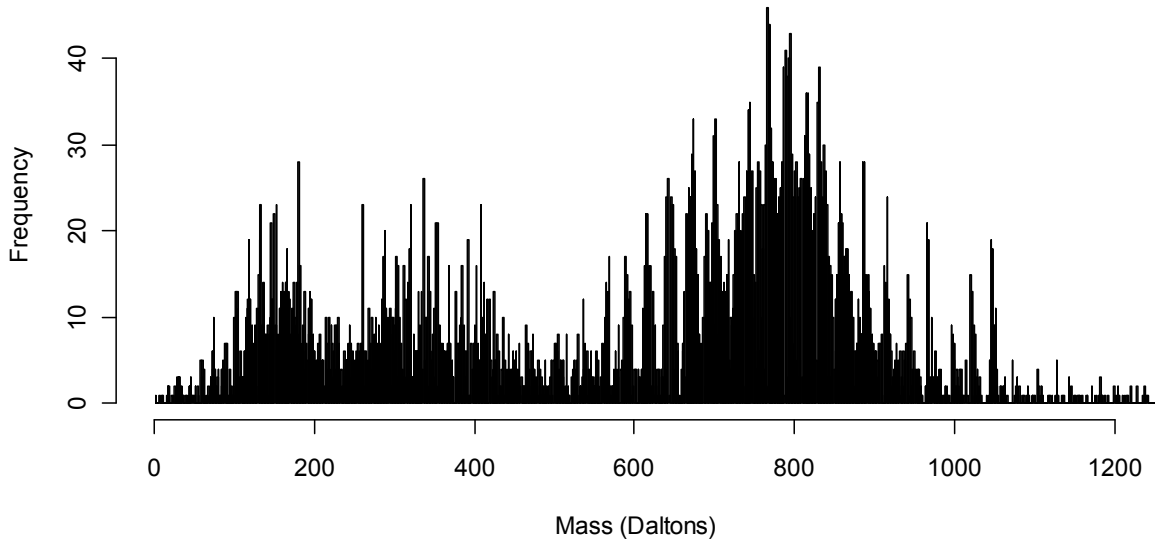


Figure 2-7. Histogram (1 Da. bins) of compounds in the Human Metabolome Database v2.5 (Wishart et al., 2009) with the mass range 1-1200 Da.

2.4.1 Ionisation

Application of MS to chromatographic eluate detection is only limited by the requirement to ionise the eluting chemicals of interest and ensure their gas phase availability for mass analysis. MS has therefore proven a popular detection system for biofluid separations by gas chromatography systems where separations also occur in the gas phase (Kuhara, 2005, Dettmer et al., 2007). However mating MS to LC requires both ionisation and desolvation (removal of the eluent from the effluent), producing gas phase ions for mass analysis. From the wide range of ionisation techniques that exist, only a select few are applicable for mating LC to MS (*e.g.* electrospray ionisation, atmospheric pressure chemical ionisation, and atmospheric pressure photoionisation). The most popular of these is electrospray ionisation (ESI) (Want et al., 2005), whereby the dispersion of liquid is accomplished by applying a voltage to a conductive tube channelling the chromatographic effluent creating a Taylor cone from which a fine jet of liquid is emitted. As the droplets produced desolvate (commonly assisted by application of heated gas such as diatomic nitrogen), the charge density on the surface of the droplets approaches a theoretical limit of space-charge density known as the Raleigh limit, and the droplet dissociates in a process called Coloumbic fission. This process creates smaller droplets which continue the cycle of desolvation and coloumbic fission until only ions in the gas phase remain. While the exact nature of

the final ionisation of analytes is still a topic of discussion and research (Wilm, 2011), electrospray has become a *de facto* standard in untargeted molecular profiling due to its ability to ubiquitously ionise biochemicals. The utility of ESI is also a result of its "soft" nature, producing molecular ions without introducing a high degree of molecular fragmentation which is characteristic of other common ionisation techniques such as electron ionisation.

However, ESI is a competitive event, meaning that analytes and additives present in the mobile phase can potentially compete for available charge. In many cases, strongly ionising species can reduce the apparent intensity of species that are ionised with less efficiency, thus confounding interpretation of otherwise quantitative detection. In other cases, and to a similar effect, the presence of some chemicals may enhance the ionisation and therefore artificially increase the observed intensity of analytes. These effects are mitigated by the introduction of fewer chemical species at any given time, leading to additional gains in apparent sensitivity (Buhrman et al., 1996). High resolution chromatographic separation is therefore beneficial in LC-MS applications utilising competitive ionisation processes such as ESI, providing as close to a sequential stream of isolated molecular species as practically achievable for a given sample matrix.

2.4.2 Mass spectrometry for LC-based molecular profiling

Once molecular ions are formed, they are guided by ion optics into a mass analyser where they may be differentiated by their mass-to-charge number ratios (m/z). While the numerator is a physical property of the molecule, the denominator reflects the potentially variable number of charges adopted by the ion. Molecular species with greater relative mass (*e.g.* those originating from peptides and proteins) commonly take multiple charges during ionisation while most small molecule metabolites (less than approximately 1500 Da.) will typically adopt a single charge (Hoffmann and Stroobant, 2007). This simplifies the interpreteation of metabolite mass spectra, relating m/z most often to mass alone.

Separation of these ionic species is achieved when they are introduced into the mass analysers electric and magnetic fields. Multiple mass analyser configurations exist, however all are underpinned by the physical laws defined by Lorentz (Lorentz force) and Newton (Newton's second law of motion).

2.4 LC detection by mass spectrometry

Together they equate the force of a particle (F) to its mass (m) and acceleration (a) as well as to its charge (q, distinct from the number of charges z) and velocity (v) when in an applied electric (E) and/or magnetic (B) field, forming the fundamental relationship governing motion of charged particles.

$$F = m\mathrm{a} \tag{2.11}$$

$$F = q(E + vB) \tag{2.12}$$

$$(m/q)\mathrm{a} = E + vB \tag{2.13}$$

Mass analyzers therefore leverage this relationship to separate charged particles of differing mass-to-charge number ratios by separately modulating, holding constant, or measuring the acceleration, velocity, and trajectory of a particle via modulation of applied electric and magnetic fields. Although different mass analysers utilise different approaches to effectively solve this differential equation, all are fundamentally rooted in the same principal.

Of the many mass analysers available, untargeted molecular profiling applications generally require the use of mass spectrometers that are capable of rapid acquisition of a broad m/z range of metabolites at high resolution and high accuracy. The first of these four criteria is critical for any LC detector (mass spectrometer or other), as accurate quantitation of a chromatographic peak requires at least 10 sampling points as a rule of thumb (Fillatre et al., 2010), and UPLC separations can produce peaks on the order of W = 1 second (Li et al., 2008). The second criterion is intrinsically satisfied by mass analysers that evaluate large spectral ranges simultaneously. This precludes the use of popular mass analysers such as the quadrupole mass filter which analyse m/z values nominally by individual selection over a small window (typically not smaller than 0.5 Da.). In order to measure m/z values across a wide range with a Q mass filter, this window must be swept or "scanned" across the range in order to assemble a complete spectrum from each low resolution measurement, requiring time and therefore antagonising the first criterion, which is not ideal for detection.

2.4 LC detection by mass spectrometry

The third criterion of high mass resolution is beneficial to the deconvolution of complex mixtures such as human biofluids which contain hundreds or thousands of distinct small molecule species of similar mass. Mass spectrometric resolution (R), commonly expressed as the full width of a spectral peak at half of its maximum intensity (FWHM), is mathematically defined as the quotient of a given mass (m_x) and the smallest mass difference resolvable (resolving power) at the given mass.

$$R = \frac{m_x}{|m_x - m_y|} \tag{2.14}$$

High resolution separations confer a high degree of specificity to the analysis, and when combined with the ability to accurately measure m/z (the fourth criterion), the number of molecular formula potentially responsible for a spectral signal is greatly limited, aiding efforts at mass-based metabolite identification. The accuracy of a mass measurement is expressed as percentage error of the known theoretical mass. This error is calculated by dividing the difference between the measured (observed) and theoretical (calculated) mass values by the theoretical value. Given the high relative accuracy of some mass spectrometric instrumentation, a constant multiplier of 1,000,000 is applied to the percentage error, providing a more convenient integer or single decimal value to report. Mass accuracy is therefore reported in units of parts per million (ppm).

$$ppm = \left(\frac{measured - theoretical}{theoretical}\right) 10^6 \tag{2.15}$$

Two types of mass analyser are used almost exclusively for routine nontargeted metabolite profiling, satisfying the four aforementioned criteria and also demonstrating other characteristics of an able detector such as measurement sensitivity and precision. They are the time-of-flight (ToF) and relatively newer Orbitrap mass analysers. The latter is primarily commended for its ability to achieve very high m/z resolution. However as a trapping instrument (which collects and holds ions for analyses over a period of time), this requires taking the instrument momentarily "offline" to perform the measurement, as the resolution generated and data acquisition speed are inversely proportional. While commercial ToF instruments also achieve high resolution measurements, they currently fail to surpass

the resolution produced by Orbitrap instruments. However, ToF resolution and data acquisition rate are independent, and therefore the rate of data acquisition (scan speed) required to adequately detect UPLC peaks must be considered when comparing the effective resolution of ToF and Orbitrap instruments. Subject to the data acquisition parameters used and exact type of analysis required, ToF instruments have been shown to outperform their Orbitrap counterparts based on achieving superior data acquisition rate and higher sensitivity (Rousu et al., 2010). Furthermore, ToF instruments have been noted to have superior ability to accurately establish isotopic abundance patterns (Prof. Zoltan Takats, personal communication, October 2014) which can greatly contribute to molecular annotation efforts (Kind and Fiehn, 2006). However, as both technologies remain in development, the analyst may expect a tendency toward homogenisation of their future capabilities in profiling applications. Therefore, for the studies presented herein, all work was performed using ToF mass analysis for no better reason than the availability of the instrumentation.

ToF mass analysers operate by separating ions following acceleration in an electric field of known strength towards a detector (Stephens, 1946). The resulting differential velocity is related to the m/z value of the ion, and therefore the time between acceleration and detection can be converted into an m/z value, where heaver mass ions require more time than lighter mass ions. Depending on the mass range the analyst wishes to observe, this process can be repeated tens-of-thousands of times per second. A unit of acquired data is therefore a timed accumulation of acceleration and detection events. This near-constant and simultaneous observation of an m/z range sufficiently wide to capture small molecules makes the ToF mass spectrometer a highly applicable detector for UPLC-MS profiling.

However, mass analysers are not necessarily mutually exclusive within a single instrument where the path of ions is not terminated by detection or other means. A versatile configuration of both quadrupole (Q) and ToF mass analysers (called a Q-ToF) is commonly used in place of ToF analysis, conferring the benefits of both analyser types to one instrument. In this configuration, the quadrupole precedes the ToF, separated by a collision cell capable of performing molecular fragmentation. A common method for fragmenting molecules involves the use of collision induced dissociation (CID) whereby a molecule's kinetic energy is converted to internal energy when collided with neutral atoms

such as argon, helium, or diatomic nitrogen in the collision cell. The internal energy can in turn result in the cleavage of covalent bonds and the resulting production of characteristic molecular fragments in a reproducible manner. The quadrupole (optionally) provides pre-selection of molecular targets of interest to pass on for fragmentation analysis, and the full fragmentation spectrum is then captured at high resolution with high sensitivity by the ToF mass analyser. The process of using two mass analysers in concert is known as tandem mass spectrometry, or MS/MS, and can be useful in structure-based elucidation efforts at metabolite identification as evidenced by the increasing amount of MS/MS data available in LC-MS databases (Smith et al., 2005, Horai et al., 2010, Wishart et al., 2007). Alternatively, the quadrupole and collision cell can also be selectively disabled, simply passing ion content through to the ToF mass analyser. The latter approach was used herein for routine profiling applications; however the quadrupole was used where necessary to generate CID fragmentation spectra for reference chemicals and metabolite targets of interest.

2.4.3 Orthogonal separations in LC-MS applications

While the combination of LC and MS creates a powerful system for the multidimensional separation and detection of analytes from complex biofluids, the use of MS as a detection system for LC separations also imposes constraints on the types of chromatography that are appropriate for use. Specifically, the composition of the mobile phase (and therefore the effluent) must be compatible with the ionisation source interface. For this reason, volatile additives such as formic acid, acetic acid, ammonium formate, ammonium acetate, and ammonium bicarbonate are heavily favoured in LC-MS applications for the control of pH and ionic strength over other classical chromatographic mobile phase additives, making some chromatographic methods such as ion exchange limited in applicability. Furthermore, the electrospray ionisation from LC effluent greatly benefits from the (minor) presence of water, again restricting the use of wholly-organic solvent containing methods such as normal phase chromatography. Because of these restrictions, only two types of chromatography have been widely adopted in LC-MS applications to small molecule research.

The first of these is reversed-phase chromatography (RPC), whereby a sample in predominantly aqueous solution is typically applied to a hydrophobic column in predominantly aqueous conditions.

2.4 LC detection by mass spectrometry

The retained solutes are then eluted with a gradient of increasing organic solvent, eventually removing the increasingly non-polar and hydrophobic content from the stationary phase. In this manner, RPC is the "reverse" of normal phase chromatography whereby the stationary phase is hydrophilic and the mobile phase hydrophobic. The stationary phase particles used in RPC are typically bound with alkyl hydrocarbons, with the most common being octadecyl (C18). This stationary phase is adequate for the retention of nonpolar and mildly polar analytes, but fails to retain many of the small polar analytes that constitute aqueous human biofluids such as urine. Specialised RPC C18 stationary phases such as the high strength silica (HSS) T3 column (Waters Corp., Milford MA, USA) have been shown to enable the improved retention of small polar analytes (New and Chan, 2008), and are therefore a preferred candidate for application to the study of human urine (Want et al., 2010). RPC mobile phase solvents are commonly modified by the addition of MS compatible volatile acids such as formic or acetic in relatively low (0.1%) concentration. The resulting separations are characterised by uniform peak shape and robust reproducible performance, making their use predominant in LC-MS applications.

The second common LC-MS chemistry is hydrophilic interaction liquid chromatography (HILIC). HILIC may be considered a variant of normal-phase chromatography in that it utilises a hydrophilic stationary phase (usually unbonded silica or particles bonded with charged functional groups) and elution of increasingly hydrophilic solutes is performed by increasing the polarity of the eluent (Alpert, 1990). However, unlike normal-phase chromatography, RPC solvents are used but in the reverse order. Typically a sample in mixed aqueous/organic solvent is introduced to the column which has been equilibrated with a high concentration of an aprotic organic solvent (often acetonitrile) and a small amount of dissolved water. The water is thought to form a semi-immobile pseudo-layer around the hydrophilic stationary phase, and therefore the retention of polar analytes is achieved by a combination of partitioning into this layer (and further into the stationary phase) as well as a number of intermolecular forces including weak-electrostatic mechanisms and hydrogen bonding (Alpert, 1990). While the prevailing mechanisms for analyte retention may not yet be fully elucidated (Buszewski and Noga, 2012), the retention of polar analytes makes HILIC an important approach to biofluid profiling as its selectivity is largely complimentary to that obtained by RPC. The elution of the majority of

analytes at higher organic concentrations also benefits the sensitivity of HILIC-MS assays, enhancing the efficiency of ESI. However, the separations can be more troublesome than those obtained by RPC, with poor retention reproducibility, poor peak shape, and long equilibration times reported in the literature (Tang et al., 2014, Gray et al., 2013).

2.6 Data review and pre-processing

2.6.1 Manual data review and illustrations

When combined, LC and MS instrumentation produce three dimensional datasets that describe three fundamental measurements about the analytes separated and detected by the hyphenated system. The first is the measured m/z value for an ion, or rather the m/z distribution for many ions of the same species, creating a spectral peak. The second is the intensity of that spectral peak as observed by the detector, which is (under normal circumstances within the linear range of the detector) proportional to the number of ions present in the group detected. Finally, as mass spectra are recorded over chromatographic retention time, each analyte species has a measured retention time, or (again) rather a retention time distribution for many ions of the same species, creating a chromatographic peak. This three-dimensional data is presented in a number of different forms throughout this thesis, and therefore a brief description of these illustrations is presented here using the data obtained from a single UPLC-MS analysis of a human urine sample.

The most fully descriptive method for visualising data is in an interactive two-dimensional projection of all three dimensions. One such projection, generated with LCMS3D v0.14 (Waters Corp., Milford MA, USA) is shown in Figure 2-8.

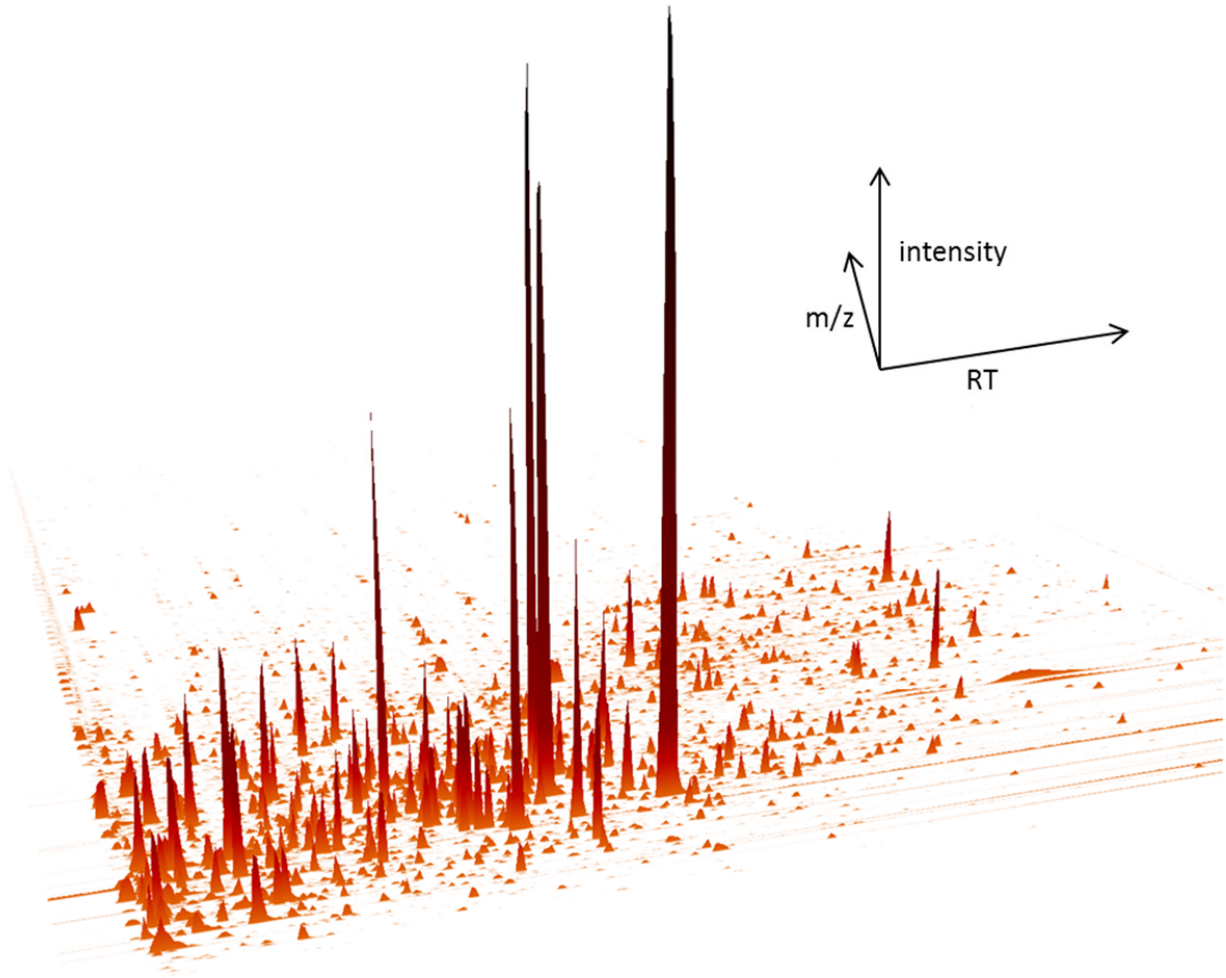


Figure 2-8. A two-dimensional screen capture of interactive three-dimensional data obtained by LC-MS analysis of a single urine sample. Peaks occupy a footprint in both the m/z and retention time (RT) dimensions, and extend upward with increasing intensity.

However, lacking the ability to move and interact with the plot, a two-dimensional map, looking down from the top of the intensity axis, may be more informative for visualisation on paper. An example of such a map illustrating LC-MS peak intensity (color) in the retention time and m/z dimensions (x and y axes, respectively) is shown in Figure 2-9. The map was produced using MassLynx v4.1 software (Waters Corp., Milford MA, USA),

2.6 Data review and pre-processing

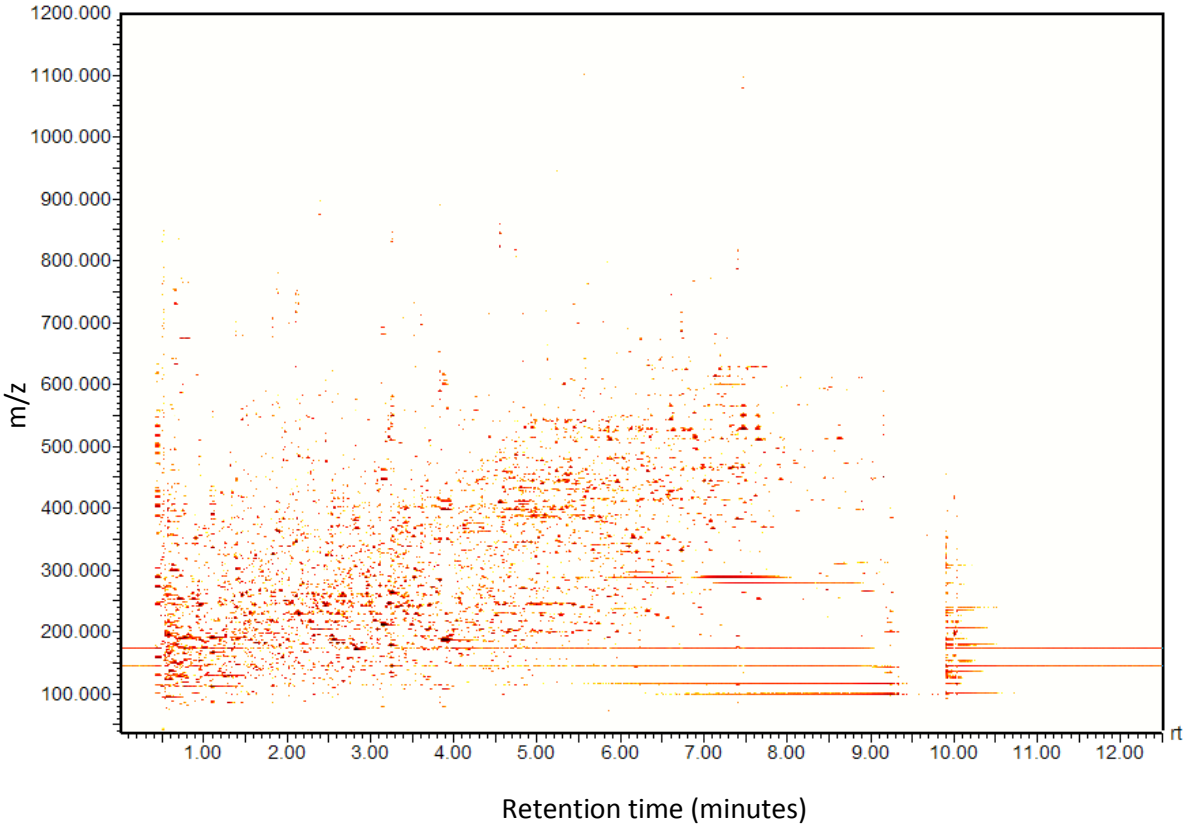


Figure 2-9. A two-dimensional projection of three-dimensional data obtained by UPLC-MS analysis of a single urine sample. Chromatographic retention time and m/z are set as the *x* and *y* axes, respectivly. UPLC-MS peak intensity is shown using a color gradient, however much of this information is lost when peaks are viewed in an overview of the entire high resolution analysis, as it is presented here.

These types of visualisations can be complex to interpret, however, and therefore more tailored visualisations are often used by analysts for specific purposes. In order to evaluate the chromatographic aspects of the data, the m/z domain is often collapsed by summation of the intensities of all spectral peaks present in a given MS scan. The resulting total ion chromatogram (TIC, more formally called total ion current) illustrates the sum of all detected signals (plotted on the y axis) for each moment in time (plotted on the x axis). A TIC plot for the urine sample is shown in Figure 2-10. This figure and all similar data visualisations introduced herein were produced using MassLynx v4.1 software.

2.6 Data review and pre-processing

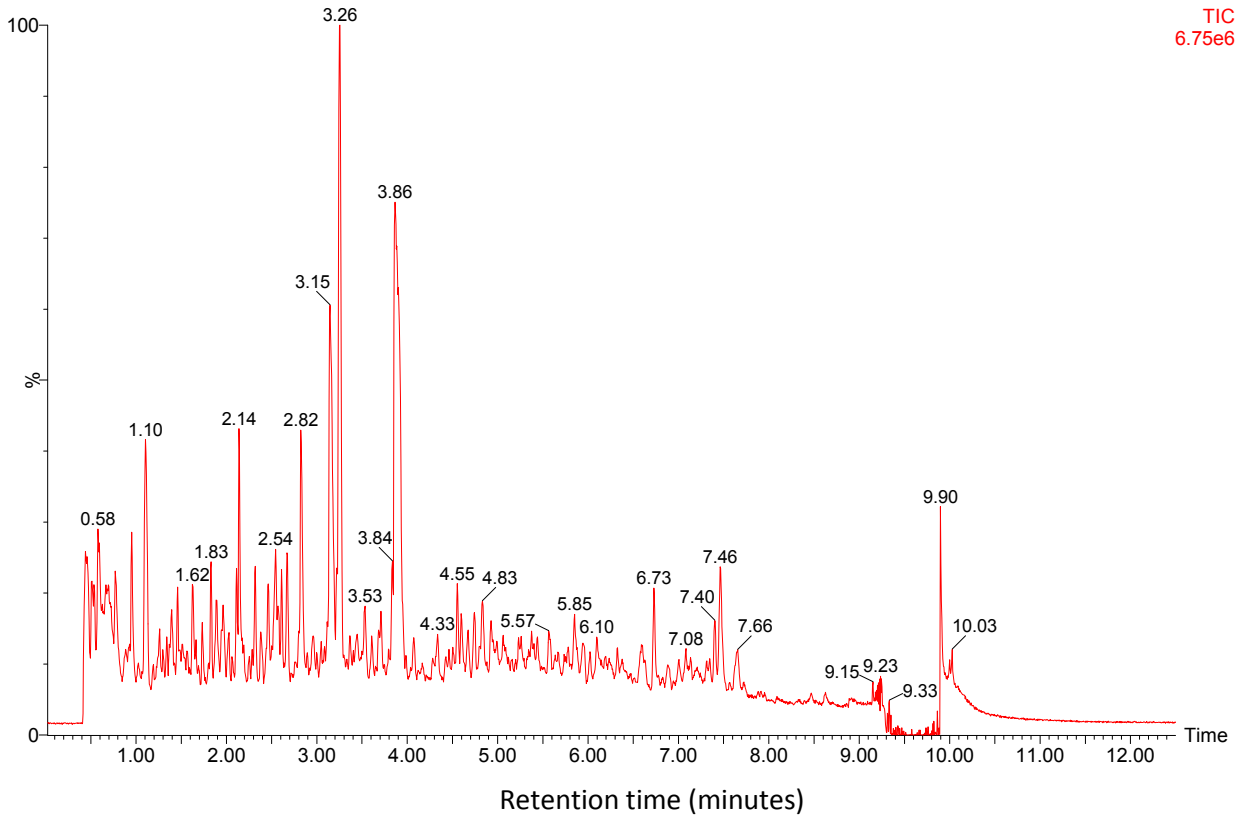


Figure 2-10. TIC visualisation of a human urine sample analysed by UPLC-MS. Chromatographic retention time (minutes) and total ion current (intensity normalised to the most intense peak) are plotted on the *x* and *y* axes respectively.

Spectral noise can potentially obscure peaks in TIC representations where many small noise-derived signals can overwhelm a single large analyte-derived signal. For this reason, the chromatogram is sometimes constructed using the intensity of the most intense peak from every MS measurement in the analysis. This kind of representation is called a base peak intensity chromatogram (BPI). A comparison between TIC (red) and BPI (green) plots for the urine sample is shown in Figure 2-11.

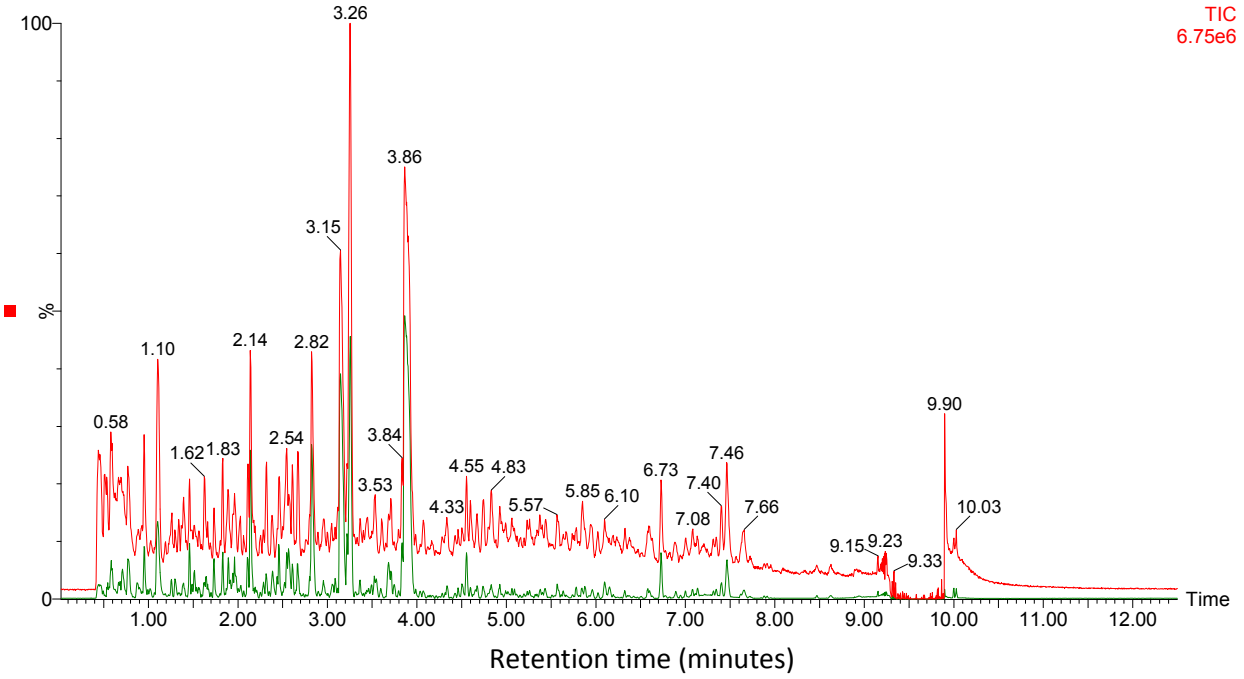


Figure 2-11. Comparison between TIC and BPI visualisations for the same human urine sample analysed by UPLC-MS. The TIC trace is shown in red, while the BPI trace is shown in green. Chromatographic retention time and intensity are plotted on the *x* and *y* axes respectively.

Finally, if the chromatogram of a single analyte of interest is desired, that chromatogram can be generated by extracting the intensity signal derived only for the m/z value corresponding to the analyte of interest. This type of chromatogram is called an extracted ion chromatogram (EIC). An example EIC is illustrated in Figure 2-12 for the expected negative ion mass of hippuric acid (a common urinary component) with an m/z value of 178.0504 +/- 0.1 Da. The inverse transformation of collapsing the retention time domain into a single mass spectrum of summed intensity is rarely performed. However, the mass spectrum obtained at any given point in time can be selected from the series and evaluated for relationships among observed spectral peaks such as the presence of isotope peaks (from ions containing heavy isotopes such as carbon-13), non-proton adducts (ions produced by addition of alkali metal ions), multimers (clusters of ions), or fragments (*e.g.* by CID or in-source fragmentation). A mass spectrum from the apex of each of the two clearly visible peaks in the EIC is illustrated in Figure 2-12. From these spectra, it can be observed that a mass (m/z) of 178.0542 is responsible for the peak at 2.8 minutes, while a mass of 178.0506 is responsible for the peak at 3.24. Utilising the calculation of

mass accuracy presented in Section 2.4.2, it can be determined that the second peak (RT = 3.24 min) is most likely hippuric acid, as the recorded mass values is within 1.2 ppm of the expected value for hippuric acid as opposed to the first peak with a measured mass that is within 21.4 ppm of the expected value.

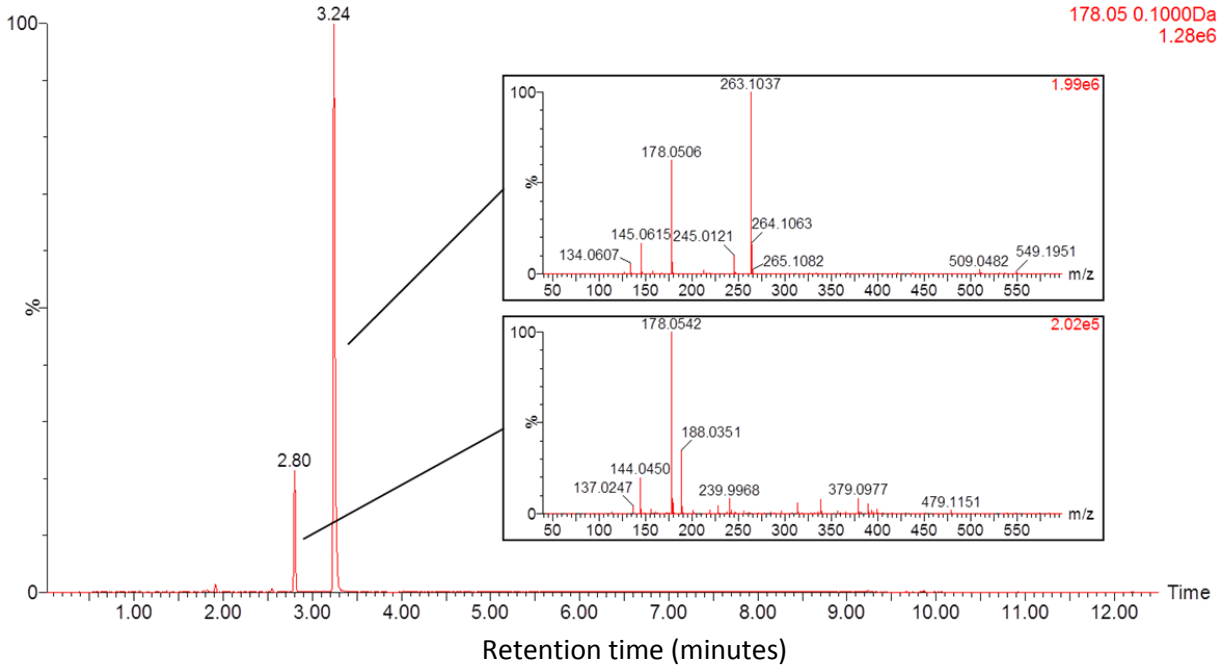


Figure 2-12. EIC of a selected m/z, with inset mass spectra. The calculated deprotonated mass of hippuric acid (178.0504 +/- 0.1 Da) was extracted from the dataset produced by a LC-MS analysis of human urine. A single mass spectrum from the apex of each chromatographic peak is shown.

2.6.2 Automated pre-processing of LC-MS data

In order to compare many LC-MS datasets collected in the course of a multi-sample experiment, the distinct signals captured (herein referred to as *features*, each with a descriptive m/z, retention time, and intensity) must be detected, extracted, and collated. The goal of these procedures is to accurately represent the analyte-derived information captured by LC-MS analysis in a single matrix, free of LC-MS system noise and processing artefacts, in order to facilitate further analysis (discussed subsequently) and interpretation of the results. The size and complexity of molecular profiling datasets precludes a manual approach to pre-processing, and automated tools that leverage computational resources are instead required. A number of software packages have been developed for the pre-

processing of LC-MS datasets, both commercially and freely available (Katajamaa and Oresic, 2007). To the extent that LC-MS raw data in vendor-specific formats can be converted into an open-source format, the open source software packages have been ubiquitously utilised in the field due to their flexibility, modular design, community-driven evolution, and open-source mechanism of data treatment (Coble and Fraga, 2014).

XCMS (Smith et al., 2006) was selected for use in the pre-processing of LC-MS data presented throughout this thesis because of its modular design containing methods that have been developed for application to high resolution profiling. Within XCMS (operated in the R software environment (R Core Team, 2014)), the centWave method of peak detection was used for the detection, extraction, and integration (intensity measurement based on the calculated area of EIC peaks) of features from each dataset (Tautenhahn et al., 2008). The centWave method assesses the LC-MS data for regions of interest (mass traces with sufficient m/z precision to suggest that a true signal has emerged from random noise) and applies a continuous wavelet transformation in the chromatographic domain to establish detected peaks. The centWave method requires each MS spectral peak to be collapsed into a single line (or *centroid* peak) representing the mass and intensity of the full spectral (*continuum*) peak. LC-MS data are commonly collected in a so-called "centroid mode" in which continuum spectra are converted to centroid spectra on a scan-by-scan basis during acquisition. In instances where data has been collected in continuum mode with full spectral peak shape detail, the datafiles can be converted to centroid-spectrum files post-acquisition using conversion tools. High resolution continuum data acquired on Waters brand ToF MS systems is conveniently converted to centroid format appropriate for centWave using the AutoAFAMM function of MassLynx 4.1. Provision is also made for the requisite conversion of centroid data files into open source formats appropriate for input to XCMS. The DataBridge executable function (Waters Corp., Milford MA, USA) was used to convert all proprietary format (Waters) data files to the open format named NetCDF prior to XCMS import and analysis.

Use of centWave on centroid mode data requires the user to define a small number of parameters to guide the criteria by which features are discerned from noise. Estimates for these parameters are

generally obtained by evaluation of the raw data in relation to the width of chromatographic peaks, and the mass error observed among sequential MS scans across a chromatographic peak. Guidelines for what intensity of signal a user considers to be noise may be specified, as well as the number of sequential scans required above a given signal intensity threshold to consider signal a true feature. The relevant parameters that require consideration and adjustment are listed in Table 2-1 with brief descriptions. Specific values used vary throughout the work contained herein and are therefore reported on a case-by-case basis.

Parameter	Description	Example values
ppm	Maximum allowed m/z deviation in consecutive scans	30
peakwidth	Range of expected chromatographic peak widths (in seconds)	1 to 8
snthresh	Signal to noise threshold, calculated as the maximum peak intensity minus the baseline value divided by the standard deviation of the local chromatographic noise	10
noise	An absolute filter that removes all spectral features of intensity less than the specified value	500
prefilter	Peaks are only considered for extraction if they contain at least *x* scans of intensity *y* or greater	*x* = 6 *y* = 5000

Table 2-1. User defined parameters for LC-MS feature extraction using centWave in XCMS.

Once features have been detected among all samples in an experiment, features derived from the same chemical species must be grouped across samples into a single table before broad comparative analyses can be performed. Unless specified otherwise, this was accomplished using the *group* function within XCMS, relying on the density method of grouping which utilises a user-defined window of mass error

and a chromatographic density profile (effectively a smoothed histogram of peak retention time values) to match features. Grouping is optionally augmented by rounds of alignment to correct for global shifts in chromatographic retention of analytes among samples. However, unless otherwise specified, retention time alignment was not utilised in the subsequent studies due to the close matching retention time across samples produced by UPLC analysis, which is believed (herein) to be compensated for by grouping alone.

The product of the feature extraction and grouping procedures is a matrix of integrated peak areas (intensity values) for every sample (one row per sample) and every feature group (one column per group). Feature groups are described by the median m/z and retention time values of each individual feature within that group. An example matrix of feature intensities is shown in Table 2-2. Missing values, arising where a feature was detected within the experimental dataset but not in every sample (resulting in an integrated area of zero for those samples) are replaced with non-zero values by integrating the EIC area (representing noise) where a peak would be expected in those samples using the *fillPeaks* function within XCMS.

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
median m/z	68.9943	86.0759	88.0919	88.092	88.0921	90.5414
median RT	25.19	89.732	513.348	486.752	521.6	535.134
Sample 01	8358.68	52405.2	10797.7	6530.27	35580.6	3519.25
Sample 02	7245.5	16150.2	10882	16409.7	13401.1	3394.77
Sample 03	7127.96	51931.5	8206.4	9243.43	18160	3432.88
Sample 04	8210.97	51027	12222.3	7661.1	5998.86	3814.35
Sample 05	7391.57	49323.9	7582.45	8561.09	12670.1	3276.64
Sample 06	7564.92	52503.2	7338.85	7748.79	7171.07	3394.79
Sample 07	8760.62	52018.3	11696.8	7298.02	12893.5	3458.68
Sample 08	8645.15	51778.2	11008.9	7922.45	13339.7	3664.05
Sample 09	7179.3	50977.4	8376.49	7165.67	13037.2	3568.88

Table 2-2. The data matrix product of feature extraction and grouping, showing the feature intensity for six distinct features across nine hypothetical samples.

2.7 Multivariate data analysis

The overall purpose of molecular profiling is to evaluate the metabolic signatures related to known phenotypes (*e.g.* disease states), clinical measures (*e.g.* blood pressure or BMI) or other known metadata (*e.g.* age or gender). In order to do so, the data obtained by LC-MS measurements are evaluated and interpreted in the context of the study design. However, the data matrices produced by profiling of complex biofluids such as human urine can be extremely large, containing thousands of features which represent a consensus average metabolome subset (as captured by the LC-MS system). When hundreds or thousands of samples are analysed in the context of large population screening, the matrices produced are enormous, potentially containing millions of individual feature intensities. Comparison of the intensities of all feature groups (called *variables* in subsequent analysis) among all samples (called *observations*) with respect to the study design using classical univariate statistical approaches (*e.g.* Student's t-test and typical 95% confidence threshold (Student, 1908)) is likely to produce many false positives, given the large number of comparisons. While multiple testing correction approaches have been utilised to surmount this issue (Broadhurst and Kell, 2006), multivariate approaches to data analysis and interpretation are generally favoured within the metabonomics field (Liland, 2011, Want and Masson, 2011). Such approaches reduce the risk of producing false positive associations with respect to the study design, and furthermore are excellent at detecting patterns of correlation among metabolites (potentially indicating the presence of biochemical pathway associations or co-regulation in biochemical networks).

A number of methods exist for the multivariate analysis of profiling data matrices. Of these, principal components analysis (PCA: (Pearson, 1901, Hotelling, 1933)), partial least squares (PLS: (Wold et al., 2001)) analysis and orthogonal projection to latent structures (OPLS: (Trygg and Wold, 2002)) are staple methods applied to metabolite profiling experiments. Use of these methods allows the investigator to explore the sources of variance within complex multi-dimensional datasets in either a top-down approach (unsupervised analysis using PCA) or in a targeted manner with respect to known variables (supervised analysis using PLS or OPLS). In both cases, the complex dataset is simplified,

highlighting the major sources of variation or sources of variation with known relevance to the study design.

2.7.1 Principal components analysis (PCA)

As an unsupervised technique, PCA does not require any input related to the study design (known metadata or phenotypic information). Rather, the technique is used to project the dataset in a manner that highlights the maximum sources of variance present using a minimal number of orthogonal planes called *principal components* (PCs). Each individual PC is a linear combination of variables describing a latent (*i.e.* hidden or inferred) variable which in turn describes variability within the dataset and is independent from all other PCs. PCs are generated in order of the amount of variance they explain, aiding in data-reduction. In this manner, PCA produces a summary of the data which is useful for detecting trends, correlated variables, and outlying observations (samples). The results of PCA may be visualised by either a scores plot or a loadings plot, illustrating the relation of observations (samples) or variables (feature groups) respectively. Scores plots are used exclusively within this thesis as illustrations for PCA. All scores plots contained herein were generated using SIMCA-P+ software (Umetrics, Umeå Sweden). Mean-centering is applied by default, whereby the mean of each variable is subtracted from the data, producing a plot oriented around the intersection of the x and y axes. Furthermore, SIMCA utilises the non-linear iterative partial least squares (NIPALS) algorithm for computing principal components which does not normalise the scales (coordinates in scores space) on the x and y axes. Therefore, although the scales are directly related to the variance explained by each component, it is convenient (for intuitive interpretation of the data) to additionally report the percent of total variance explained by each component. An example is shown in Figure 2-13 for a set of 23 replicate UPLC-MS analyses of a single urine sample, followed by a single analysis of a serum sample (not shown) and 23 subsequent analyses of the original urine sample. In this example, the first and second PCs are shown on the x and y axes, representing 34 percent and 19 percent of the dataset variance, respectively. The PCA scores plot clearly illustrates the perturbation to the precision otherwise observed in replicate urine analyses, as well as a tendency of the metabolic profiles in the post-serum analyses toward returning to the pre-serum injection state.

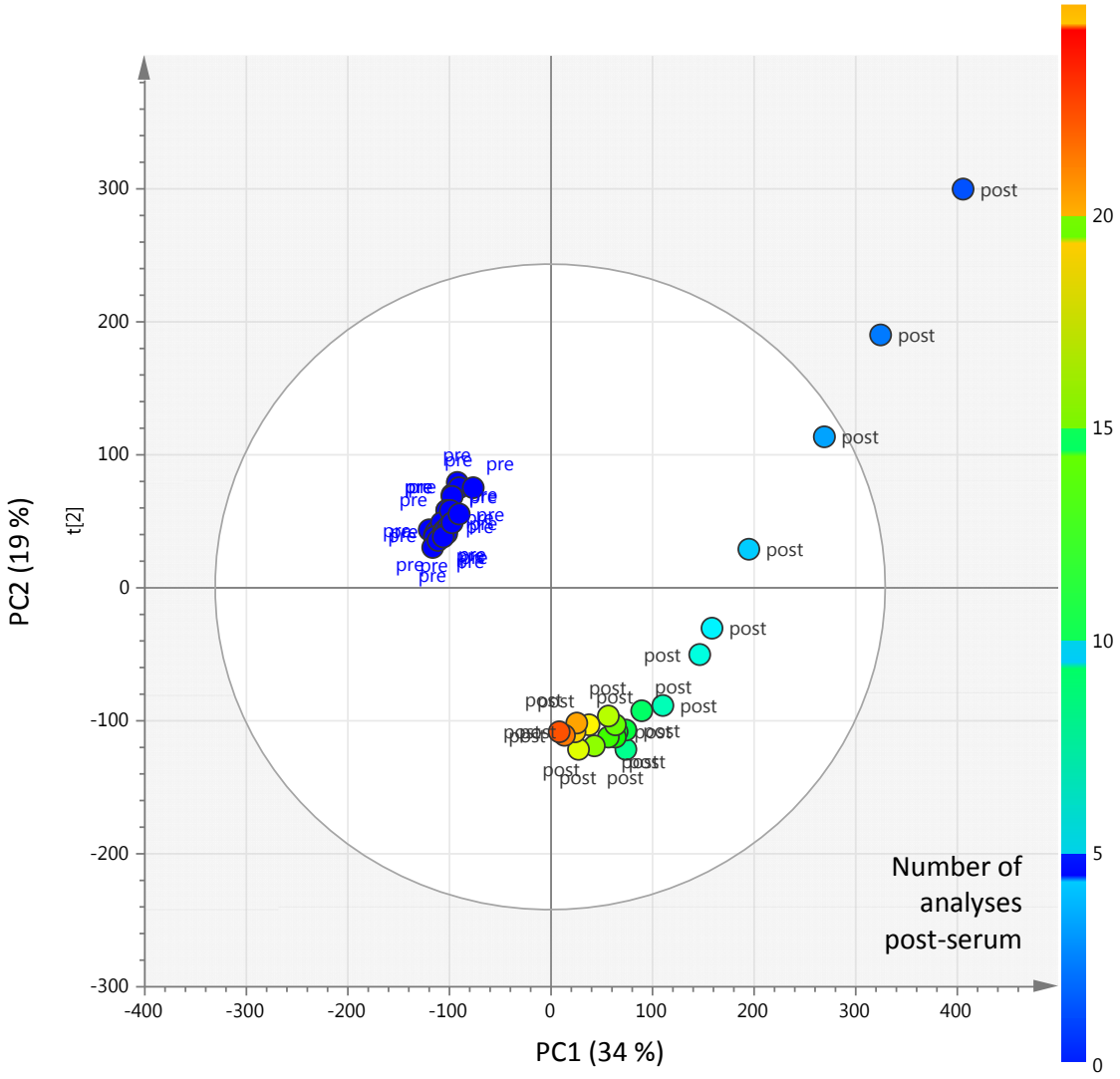


Figure 2-13. PCA scores plot of an analytical reproducibility intervention study. The precision of the UPLC-MS measurement is tested on an unchanging urine sample before (pre) and after (post) the injection of a single serum sample. PCA illustrates that the urine analyses immediately following the serum analysis yield profiles that are different from the otherwise reproducible urine analyses, but tend toward homogenisation and a return to initial conditions as the urine sample is repeatedly analysed.

2.7.2 Partial least squares (PLS) and orthogonal projection to latent structures (OPLS) analyses

Often epidemiological studies are searching for subtle phenotypic effects among large populations (*e.g.* metabolic correlation with increased risk of heart disease or type 2 diabetes), and therefore PCA is rarely sufficient for the elucidation of features correlated to the study design due to the presence of greater confounding variation. For these purposes, supervised methods of multivariate analysis exist whereby a model of the data is built using additional information beyond the previously discussed matrix of observations and measured variables. The first of these, PLS, models the relationship between the profiling data matrix and additional data (*e.g.* retrospective patient outcome with respect to developing a disease) to discover the maximum covariance between them, described by PLS components. In this manner, PLS may be regarded as a specialised extension of PCA which is targeted to the study design or specific relationship of interest.

In PLS, as with PCA, the maximum (co)variance between the profiling data matrix and a single *Y* variable of additional data should be represented by the first component. However, systematic variation in the profiling data matrix that is independent of *Y* can confound the interpretation of the resulting model, as PLS must model this variation together with the variation of interest. As a result, the variation of interest can become distributed across multiple PLS components. In order to remove the confounding variation in these instances, OPLS has been developed as an extension of PLS. Using OPLS, confounding variation in the profiling dataset is modelled and removed, allowing PLS to generate the maximum covariance between the corrected dataset and *Y* with the fewest (usually one) components. This in turn allows for a more convenient interpretation of results and elucidation of the relevant covarying LC-MS features.

In PLS and OPLS, the scores and especially the cross-validated scores are used to estimate the impact of the experimental design on the data and to identify potential confounding factors and outliers. The relevant statistical values R^2_Y and Q^2_Y are also important to consider when evaluating a PLS or OPLS model. The first (R^2_Y) represents the proportion of *Y* explained by the profiling data matrix. This parameter is considered in the context of the second value (Q^2_Y) which is the proportion of *Y* predicted

by the profiling data matrix through a cross-validation process. A higher Q^2_Y value is indicative of a more reliable PLS model. However, the difference between the R^2_Y and Q^2_Y values is also important, as a greater difference may indicate the influence of over-fitting on the model produced by noise in the data. Over-fitting must be minimized as much as possible to allow reliable interpretation of the PLS and OPLS models. Once the PLS and OPLS models are validated, loadings plots are often used to highlight the underlying features responsible for the desired covariance. In a biochemical context, these patterns of metabolites may represent biomarkers which indicate biological processes, pathogenic processes, or pharmacologic responses to therapeutic interventions, depending on the design of the study (Biomarkers Definitions Working Group, 2001). However O/PLS loadings plots can equally be used to highlight the LC-MS features driving analytical variance. An example of a specialised loadings plot called an "S-plot" (Wiklund et al., 2008) is shown in Figure 2-14A for the OPLS discriminant analysis (OPLS-DA, meaning comparison between distinct classes) of the urine sample analyses described above and illustrated in Figure 2-13. The S-plot juxtaposes covariance with its reliability (correlation), sending the most discriminant features to the extremes of each axis. The intensities of the red and blue highlighted features are shown for each pre and post-serum injection in Figure 2-14B, illustrating the opposing patterns with respect to the binary study design.

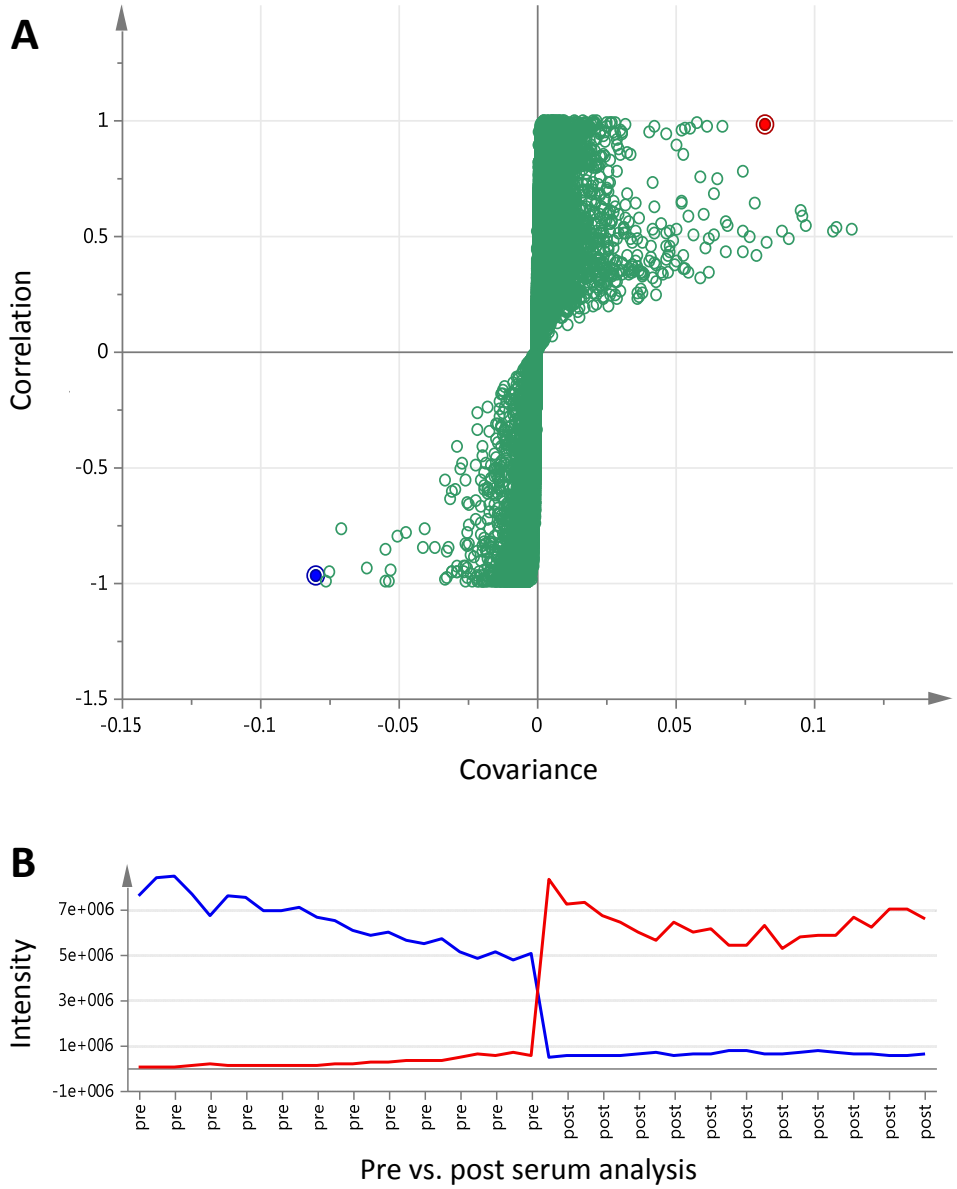


Figure 2-14. OPLS-DA loadings "S-plot" (A) and variable line plots of selected features (B). Features observed in the repeated UPLC-MS analysis of a human urine sample before and after the single injection of a serum sample are plotted above with respect to their covariance and correlation, while the intensities of selected red and blue highlighted features are shown for each pre and post-serum injection illustrating the opposing patterns.

2.8 Metabolite identification

Once features of interest have been selected using multivariate analysis, their interpretation in a biological context is limited only by the extent to which their chemical identities are known. A single analyte can produce a number of features, any of which may not be intuitive to relate back to the expected monoisotopic mass of the molecule. For this reason, metabolite assignment and identification of selected features among the thousands monitored by LC-MS analysis remains a challenge within the field (Wishart, 2011). It is important to draw a distinction between assignment and identification, as the former is an annotation of a feature based on measured properties and the latter is claimed after verification by comparison to an authentic standard (Salek et al., 2013, Sumner et al., 2007).

In order to properly assign features with a chemical name to benefit biological interpretation, the analyst must first interpret his or her chromatographic and spectral data. Determination of the true monoisotopic parent in a cluster of masses detected at a given retention time and all potentially derived from the same molecular species is not always trivial or even possible. Many molecular species simply are not observed as simple [M+H]+ or [M-H]- protonated or deprotonated ions, but rather as adducts, multimers, or fragments, even when fragmentation is not purposefully induced. Therefore, even a spectrum of a known chemical can require significant interpretation. The inverse approach of constructing formulae and molecular structures from spectral information, especially that confounded by the presence of multiple molecular species, can be prohibitively challenging. Yet, some understanding of a mass spectrum and its relation to the suspected molecule is generally required for subsequent steps in the identification process.

2.8.1 The elemental composition approach to elucidating a molecular formula

The most rudimentary approach to molecular assignment is the mathematical combination and permutation of atoms (selected by the analyst based on their assumptions of the molecular composition) until molecular solutions are found with the same mass as the observed feature (+/- some set window of error). More liberal estimation of the component atoms can result in greatly amplified numbers of potential matching formula. One effective filter applied to reduce erroneous matches is derived from the similarity of the observed isotopic pattern to each theoretical pattern, a property that

is accurately calculable if the molecular species is natural in origin (and therefore the population of product ions conforms to the expected natural distribution of isotopes). While popular for its ease of automation across many masses of interest, the results of the "elemental composition" approach are often not definitive at routinely achievable mass accuracy (approximately 1ppm). In the best case, a molecular formula can be determined; however the molecular structure remains unsolved.

2.8.2 Use of mass and spectral databases for metabolite/biomarker assignment

Where a monoisotopic value can be derived from the spectral data, that mass may be searched against a number of freely available databases (Williams, 2008, Smith et al., 2005, Horai et al., 2010, Wang et al., 2012, Wishart et al., 2009). Some databases such as HMDB are restricted to known chemicals relevant to the human metabolome. Other databases such as ChemSpider or PubChem contain a far greater wealth of chemical structures, but may return assignments that are biologically redundant or irrelevant to the system being studied. The number of potential assignments returned from a database search of a mass therefore depends on the size and structure of that database, as well as the mass error used in the search (and therefore the mass error inherent to the MS instrumentation). In recent years, efforts have been increased to include additional spectral information beyond that of parent mass and molecular structure. The inclusion of CID MS/MS reference spectra and spectral searching in freely available databases such as METLIN and MassBank has simultaneously reduced the need for spectral pre-interpretation and increased the confidence in proper formula and structure matching.

2.8.3 Validation to authentic standards for metabolite identification

Despite the growing utility of database searching and the popularity of automated elemental composition calculations, the assignments gathered from these approaches must be considered tentative. Validation of retention time, mass, isotope distribution, and (if available) fragmentation pattern between the candidate compound and an authentic standard remains the gold standard required for absolute assignment in the LC-MS-based metabonomic field (Kind and Fiehn, 2010, Sumner et al., 2007). Standards are therefore run by the original analytical method, often at a later date following the original analysis, along with one or more representative samples from the original experiment in order to assess the chromatographic and spectral similarity. On well characterized and

reproducible instrumentation, LC retention time results similar to those obtained in the original experiment can be expected. However, the variance between separate experiments is generally greater than that experienced within any one experiment, where the solvents are of an identical batch and preparation and the stationary phase is at a certain point of use and conditioning. At times, returning to the process of molecular identification after the data from the original experiment has been collected and analysed can be confounded by slight differences between experimental conditions. To avoid this, if the user has an idea of which molecules they expect to be of interest at the conclusion of an experiment, they can prepare and analyse those individual standards within the original sample set, therefore obtaining maximally comparable retention time data. Where this approach is not possible, an idealised approach might be to run a full complement of thousands of individual standards to serve as authentic references within each profiling experiment in the event that some may be discriminating features of interest requiring subsequent identification. This approach is not practical on a per-experiment basis, given that the time to collect the standards data could equal or outweigh the time required to analyse the biological samples in many large profiling experiments.

2.8.4 Method-specific LC-MS retention time databases

Where individual chemical standards have been obtained and analysed in the course of newly sought absolute identifications, their retention times and spectra can be recorded. Indeed, many metabonomic research laboratories maintain in-house LC-MS libraries with the retention time of authentic standards run by their established chromatographic methods. These data can be referenced in later experiments to generate tentative molecular assignments (absolute assignments are still made by direct comparison to the authentic standard). This approach requires that the reproducibility of the method is robust to changes in the condition of stationary phase, variance in mobile phase batch quality and preparation, and the physical configuration of the instrument. Furthermore, such a database requires that the method not be tailored, improved, or otherwise changed. The approach is therefore valuable but strict, and has the potential to limit the speed and quality of future analyses. The methods may not be accurately implemented in other labs with slightly differing instrument configurations, and therefore

are of limited use across the greater field. A potential alternative to method-specific databases is the conceptual focus of Chapter 5.

Chapter 3: Development of coupled liquid chromatography and mass spectrometric methods for the profiling of human urine from large patient cohorts.

3.1 Introduction

The goal of metabolic profiling is to maximise the range of molecular species measured in complex sample matrices while achieving a high degree of specificity and sensitivity. These traits are inherent strengths of the multidimensional separation and sensitive detection provided by high resolution hyphenated LC and MS technologies, and as a consequence, UPLC-MS has found widespread application in the profiling field. This chapter explores the development and application of complementary chromatographic methods with special emphasis on addressing the challenges of large-scale deployment in the context of molecular epidemiology studies.

Unlike clinical chemistry approaches, whereby a small number of molecular targets are analysed (by LC-MS or other means) in a quantitative manner that allows direct comparison to established reference values, the broad scope of profiling and the lack of commensurate quantitative benchmarking limit LC-MS profiling studies to relative comparisons among samples or sample sets. This requires each study to contain multiple samples or sample subsets which span a physiological range of interest or adequately represent phenotypes (often simply "control" and "diseased") for direct comparison. The data generated among samples within a study must also be captured with sufficient precision to facilitate cross-sample comparison. Unfortunately, both LC and MS face inherent challenges to precision due to the direct interaction of the sample and requisite carriers (e.g. LC mobile phase components) with the analytical system, resulting in gradual changes in the measurements produced.

When analysing many samples in sequence, the dynamics of the UPLC-MS system manifest collectively as the so-called "run-order effect". This phenomenon reflects the combined changes in UPLC, ion source, and MS hardware performance that ultimately cause time-dependent deviation in the analytical measurements taken and reduce the overall performance of the analysis. Perhaps the most striking component of the run order effect is the commonly observed decrease in sensitivity across continuous

sample measurement, originating from contamination of the ion source and initial ion optics (resulting in reduced formation and transmission of ions) and/or longitudinal fatigue of the ion detection system. Additional run order effect components are common such as the chromatographic migration of molecular species and changes in peak shape. Both the UPLC and MS systems, as well as the ionisation interface between them, are therefore dynamic, and should each be the subject of scrutiny when assessing the precision of UPLC-MS analyses.

It is commonly understood that UPLC-MS instrumentation is most dynamic when all components are clean and new, as made evident by numerous publications citing specific efforts at conditioning a "fresh" LC-MS system (clean ion source and new chromatographic column) through repeated exposure to representative sample material (Want et al., 2010, Spagou et al., 2011, Gika et al., 2007, Sangster et al., 2006). Such conditioning efforts are aimed at equilibration of the analytical system, thereby increasing the precision of subsequent analyses. This process is considered to be largely finite and compensated for by a short if-not-arbitrary number of conditioning injections (*e.g.* five to ten (Zelena et al., 2009, Want et al., 2010)).

Somewhat contrary to the practice of conditioning is the practice of performing LC-MS analyses in distinct sets of continuous sample analyses (commonly, "batches") interrupted for cleaning and maintenance of the LC-MS system components intended to restore the initial state of performance (herein referred to as an "analytical batch"). This interruption of continuity is often considered necessary to prevent excessive decline in LC-MS platform performance which would severely compromise the molecular coverage and measurement precision if left unattended (Zelena et al., 2009). This is most commonly performed to restore the MS instrument's original sensitivity, and involves disassembly and cleaning of the ion source components and adjacent ion optics. Batches may also be defined by the replacement of chromatographic stationary or mobile phases as well as ancillary hardware such as components of the sample injection system.

While the run order effect is gradual, batch effects are sudden deviations in measurement whereby the chromatographic retention, mass measurement, and intensity are subject to change in a global or

metabolite-specific manner. These deviations complicate downstream data processing, requiring sophisticated alignment tools, normalisation strategies, and more recently dedicated batch correction efforts (Vaughan et al., 2012, Dunn et al., 2011, Draisma et al., 2010, Wagner et al., 2007) all of which risk introducing error and artefacts (e.g. over-correction) to the dataset. Furthermore, additional conditioning of the system may be necessary at the commencement of each batch to ensure that the system is again equilibrated, reducing the efficiency of the overall workflow. Minimising the number of analytical batches required for sample set analysis and ensuring continuous operation of the LC-MS platform is therefore of specific interest, especially as sample sets are driven to be larger in scale to accommodate the desire for enhanced statistical power for identification of candidate biomarkers of disease.

Continuous operation of the system requires an uninterrupted supply of samples for analysis, which creates the potential for a second type of batch based on the schedule of sample preparation and the stability of each sample's molecular content. Modern commercial LC-MS instrument packages include a dedicated sample handling device capable of introducing limited numbers of samples to the instrument in an autonomous manner. However, as the capacity of such systems is indeed limited, samples are generally prepared in discrete batches (herein referred to as "preparation batches") by laboratory personnel or centralised robotics instrumentation and then transferred to the LC-MS sample handling device for storage prior to analysis. The preparation schedule must therefore be achievable within a typical laboratory working environment, periodic to ensure ease of management, and frequent to limit the amount of time a prepared sample ages between preparation and analysis. While the first two requirements are functional and related to efficiency, the last requirement relates to data quality, reducing the time allowed for metabolite reaction or degradation, and potentially improving the total number of metabolites accurately measured. The need to establish a schedule and number of sample preparation batches is magnified when analysing vast numbers of samples per study, such as in typical epidemiology studies of biobank-derived sample cohorts.

The size of the preparation batch is dependent on the capacity of the sample preparation laboratory as well as the quality requirements and duration of the analysis. For example, in targeted applications, the

stability of target molecules ageing while queued for analysis is commonly evaluated in method validation, and an appropriate batch size (or compound-stabilising preparation procedure) may be tailored accordingly to uphold a predetermined standard of measurement precision. However, in profiling studies of complex human biofluids, the wide range of detectable molecular species virtually assures that some of the observed content will be highly unstable, to an extent that may not be practical to address through an intensive sample preparation schedule. The stability of molecules therefore ultimately impacts the molecular coverage observed in profiling studies. As the wide breadth of molecular coverage is a fundamental tenet of molecular profiling, this too deserves special emphasis in workflow development.

A template for large-scale analysis is therefore proposed in Figure 3-1, wherein a single profiling experiment is composed of one or more analytical batches (*x*), and each analytical batch is composed of one or more sample preparation batches (*y*). In turn, each sample preparation batch is comprised of one or more groups of assay samples (*z*). Although it will be discussed subsequently, it is worth noting here that the 96-well plate is an established standard utilised in many high throughput/high volume processes including cell screening, PCR amplification, and immunoassays (*i.e.* ELISA) and is therefore an appropriate building block for preparation batches in high throughput metabolic profiling applications.

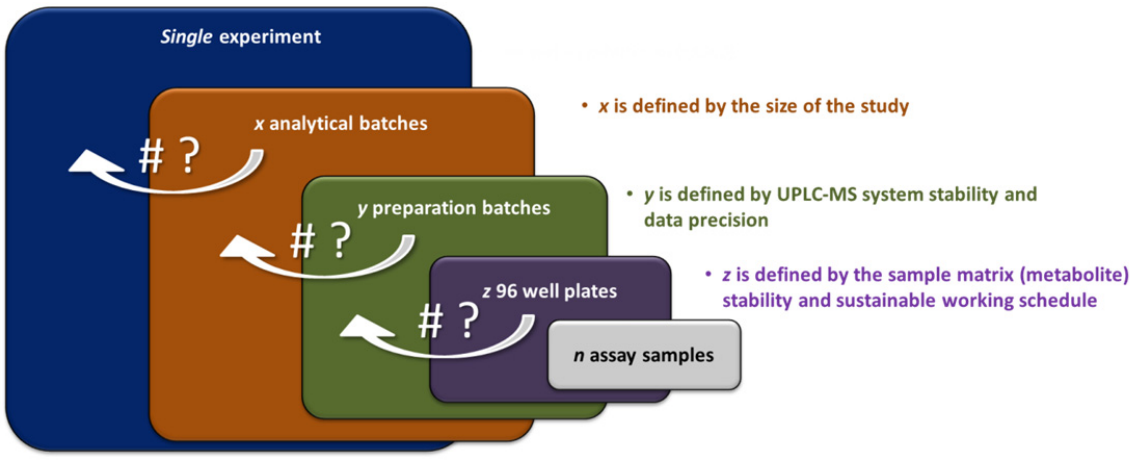


Figure 3-1. A schematic of the basic structure of large-scale UPLC-MS profiling analysis

3.1 Introduction

In order to successfully conduct large-scale experiments and generate high quality profiling data for comparative analysis, the values for x and y and z must be carefully considered and optimised in a holistic manner that is reasonable within the constraints of the laboratory environment. The analytical methods and LC-MS system configuration should be harmonised to sustain uninterrupted analysis for as long as possible to minimise or preclude batching effects, facilitating downstream data handling and conserving the statistical power which is expected of large studies. To allow for potential data fusion and meta-analysis among distinct large populations, an additional emphasis is placed on minimising inter-experimental variance and establishing quality criteria. Working these improvements into LC-MS methodology will positively impact the applicability of LC-MS in mating broad screening to large populations for unparalleled insight to human metabolism.

To realize the benefits of LC-MS analysis in this context, the development of high-metabolite coverage methods must be undertaken with emphasis on reproducibility, efficiency, and throughput. These aspects should be developed without conceding the principles of high molecular coverage, sensitivity, and specificity which propel LC-MS forward as a leading profiling technology (Gika et al., 2014). As application of profiling to epidemiological sample cohorts means studying an array of phenotypes, sample preparation techniques should be minimally selective. The chromatographic methods must therefore be robust and capable of handling crude biological samples with minimal manipulation, additionally benefitting the working laboratory by being more convenient, more rapid, and less error prone than more complex processing strategies. To avoid undermining the ability to complete large experiments, these aims need to be prioritised without significantly impacting the overall speed of analysis achieved using modern rapid methods.

This chapter is therefore organised in three main components of development:

1. Adaptation of two complementary chromatographic assays (RPC and HILIC) to improve information density and maximise molecular coverage with high precision.
2. LC-MS system optimisation and characterisation of the variance underlying analytical batch size.

3. Defining optimal preparation batch size for limiting sample age within a practical working environment.

Here it is hypothesised that the conventional method of acquiring LC-MS data in small batches is detrimental to the overall precision of a large-scale assay. Moreover, run order effects are easier to correct for than batch effects because of their systematic nature, which is subject to modelling (a concept which is further developed in the subsequent chapter). The overarching goal of this chapter is therefore to maximise the coverage of the metabolome by developing complementary RPC and HILIC assays while ensuring adequate precision, both in the chromatographic technique and system performance to preclude the need for acquiring large-scale projects in multiple small batches.

3.2 Specific Objectives

- Generate standard materials for use in both method development and routine application to large cohort analyses.
- Develop high capacity chromatographic methods with complementary metabolite coverage that are fit-for-purpose in human urine LC-MS profiling studies.
- Define a system configuration using cutting edge analytical instrumentation that maximises the analytical batch size, optimising the balance between system sensitivity and longitudinal precision of measurements.
- Establish an optimum sample preparation and analysis schedule that is fit for the purpose of molecular profiling, demonstrating adequate throughput, minimal sample age, and maximal laboratory efficiency.

3.3 Reagents and biofluids for method development and system optimisation.

Throughout this chapter, two main sources of metabolite content are used for the development of chromatographic methods and assessment of their performance. The first is a urine sample generated from a pool of 76 freshly voided samples, donated by willing volunteers with approved informed consent, for use as a long term reference (LTR) material across multiple urine profiling studies. This

pool is used herein as a representative matrix for both RPC and HILIC method development. The second source of metabolites is a pair of synthetic mixtures of chemical reference standards, each formulated for a single chromatographic assay (RPC and HILIC). Given the ubiquitous use of these materials throughout this chapter, a brief description of their development is warranted.

3.3.1 Development of a pooled urine sample for use as a representative matrix

In order to create a single representative urine matrix for the standardisation of ongoing research and analysis within the MRC-NIHR National Phenome Centre, a large pool of urine was created and aliquoted. Approximately 20 litres of urine were pooled from fresh voids of 78 individual participants in a single day of collection according to the standard operating procedure included in Appendix 1 (PCSOP.036 revision 7: *Generation of Urine Long Term Reference (LTR)*, Matthew R. Lewis, 2013). Briefly, urine was voided directly into 500 mL Corning centrifuge tubes which were stored at 4° C overnight. A single mL of each sample was reserved for NMR analysis to ensure that polyethylene glycol (PEG), a common contaminant observed in collected human urine specimens, was not present in appreciable quantity. Two samples were removed as a result of this analysis. The day after collection, the remaining 76 samples were centrifuged at 4° C and the supernatants were combined in a single Nalgene 20L polypropylene carboy and homogenised at 4° C by stirring using a Teflon-coated stir-bar and magnetic stir-plate. The homogenised urine was aliquoted into 15 mL Corning centrifuge tubes and stored at -80° C. This urine LTR was used regularly for the development and illustration of methods within this chapter.

3.3.2 Development of chemical reference mixtures

To augment the use of urine LTR in LC-MS method development and assessment, synthetic mixtures of chemical reference materials were created, with a subsequent intended use of application in pre-experiment hardware suitability testing (evaluation of the UPLC-MS system for performance within predetermined boundaries), within-run targeted quality control (QC) evaluation, and retention locking for LC-MS data fusion (e.g. of new data to a database of annotated molecules). Purposeful metabolite selection is required to ensure mixtures are economical and fit-for-purpose, providing specific targets that sparsely represent the molecular diversity observed in human biofluids. The Division of

3.3 Reagents and biofluids for method development and system optimisation.

Computational Systems Medicine at Imperial College London houses a large library of metabolically relevant small molecule reference standards which served as the pool for further selection and testing via the algorithm illustrated in Figure 3-2.

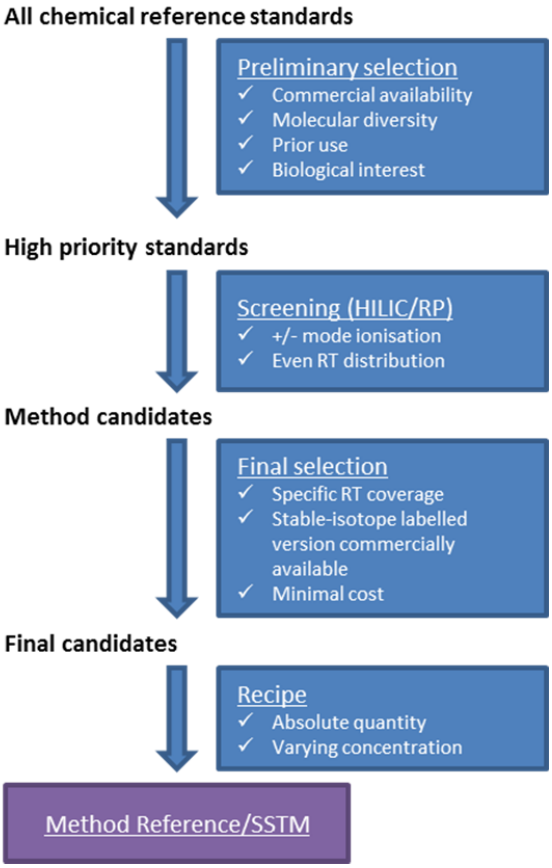


Figure 3-2: Strategy for selection of chemical standards for reference testing

By considering the empirically observed water solubility, commercial availability, and potential biological significance of each metabolite, the total library of over 1800 available reference materials was reduced to the 57 high priority standards listed in Table 3-1. Special care was taken to ensure the collection was appropriately diverse in relation to the observed contents of human urine and not entirely composed of a single molecular class (*e.g.* amino acids) that may fail to broadly represent the analytical behaviour of urinary metabolites. Specific inclusion was granted to chemical standards with a record of prior use in a standards mixture from published literature (Evans et al., 2009, Zelena et al., 2009) or publically available standard operating procedure (HUSERMET, 2008).

3.3 Reagents and biofluids for method development and system optimisation.

Ref.	Nominal Name
	2-Deoxyguanosine
	2'-Deoxyyadenosine
	2-Fluoro-DL-alpha-phenylglycine
	3-Hydroxybutyric acid
	3-Hydroxytyramine
	4-guanidinobutyric acid
	5-Hydroxytryptamine
	Adenine
	Adenosine
H, H2	Alanine
	Arginine
H	Benzoic acid
	Benzyltrimethylammonium bromide
	Betaine
	Carnitine
H, H2	Citric Acid
	Citrulline
	Creatine
	Creatinine
	Cyclohexylacetic Acid
	Cytidine
	Cytidine 5'-monophosphate
M	Dioctyl phthalate
	Folic acid
H	Fructose
M	Glucose
H2	Glutamic acid
	Glutamine
H2	Uridine

Ref.	Nominal Name (continued)
	Glutathione (reduced)
H, H2	Glycine
M, H2	Hippuric Acid
	Inosine
	Isoleucine
M, H2	Leucine
H	Lysine
H	Malonic Acid
M	Methionine
	N-acetyl-glycine
	N-Acetyl-L-aspartic acid
	N-acetyl-L-cysteine
	N-acetyl-L-glutamine
	Nicotinamide
M, H	Octanoic acid
M, H2	Phenylalanine
M	Progesterone
	Proline
	Riboflavin
H	Stearic acid
H	Succinic acid
	Taurine
	Threonine
	Trigonelline hydrochloride
	Trimethylamine-*N*-oxide
M, H, H2	Tryptophan
M	Tyrosine
	Uracil
H2	Uridine

Table 3-1. List of 57 high priority standards assessed for inclusion in the development of RPC and HILIC test mixtures. Molecules appearing in previously published literature are indicated in blue (HUSERMET), orange (Metabolon) and green (both). For specific reference, H = HUSERMET (HUSERMET, 2008), H2 = HUSERMET (Zelena et al., 2009), and M = Metabolon (Evans et al., 2009).

Chemical reference solutions were made in a qualitative manner by aliquoting a small scoop (for solids) or drop (for liquids) of the chemical to a clean storage tube and diluting with 5 mL of ultrapure water. The approach resulted in concentrated solutions for water soluble chemicals. Incomplete solubility was

observed for some chemical preparations despite vortexing and sonication at room temperature (for a maximum of 30 minutes), however in most cases a sufficient amount of material had dissolved to produce a signal by UPLC-MS analysis. Therefore, regardless of the visible outcome, these stock materials were used as the basis for subsequent analysis, as this development was not dependent on specific concentrations of metabolites. Each solution or suspension was pipetted into an individual well of a 96-well deep well plate. The plate was centrifuged to separate soluble and insoluble materials. The supernatant was decanted to a 96-well microplate, and carried through three rounds of 1:10 (volume in total volume) serial dilution, with the aim of at least one concentration being appropriate for LC-MS analysis (neither too dilute to be detected nor too concentrated to saturate the chromatographic loading). Plates of 1:1 (stock), 1:10, 1:100, and 1:1000 diluted reference materials were frozen at -80° C until required for analysis.

Each 1:100 dilution reference material plate was thawed and prepared for UPLC-MS analysis by further dilution with either water or acetonitrile (3 volumes to 1 of sample) for RPC and HILIC analyses respectively. RPC and HILIC reference methods (Want et al., 2010, Spagou et al., 2011) were used to determine the nominal retention and MS signal response in both positive and negative ion modes by electrospray ionisation. A Xevo TQ-S (Waters Corp., Manchester UK) tandem quadrupole mass spectrometer was selected for use in rapid screening because of its fast polarity switching and high dynamic range, the former allowing for interleaved near-simultaneous detection in both ion modes and the latter creating a broad target for reference material concentration to fall within. Full scan mass analysis and detection was utilised across the range of 50 to 1200 Daltons. Where chromatographic overloading was observed, the 1:1000 dilution plate was thawed and used for reanalysis to obtain a more accurate chromatographic retention time. In the event that a chemical was not observed in either positive or negative ionisation mode, 1:10 or 1:1 dilution plates were used for reanalysis. If a signal was not observed at stock concentration (1:1), the compound was considered not suitable for MS detection.

Metabolite species spanning the gradient elution portions relevant to the separation of urine by HILIC and RPC methods (and therefore spanning the relevant ranges of polarity) were determined to be method candidates, provided they were also detected in both ionisation modes. The method candidate

list underwent further refinement to ensure adequate retention coverage of standards that were preferentially inexpensive, and for which stable labelled isotope versions were commercially available at a reasonable cost (stable isotope label variants of the mixture were intended for application in QC and retention alignment of study data). Molecular diversity of the final candidates was considered, as was the short term stability of the mixture as assessed by both UPLC-MS and NMR spectroscopy. With the knowledge that chemical reaction did not pose a short term risk to the stability of either mixture, a final recipe was crafted for each method.

Two versions of each recipe were made. The first is an un-labelled mixture of the compounds listed in Table 3-2, and is referred to as a system suitability test mixture (SSTM). The second contains stable isotope labelled versions of all chemicals, made to the same recipe without two chemicals (hippuric acid and phenylalanine) which are instead added to all samples as internal standards. This latter material is referred to as the method reference (MR). The SSTM is used routinely in the NPC to qualify instruments as performing adequately for use prior to initiating an experiment, while the latter is used in NPC research as a sample dopant for real-time and post-acquisition QC assessment and retention time marking. Within this chapter, the RPC and HILIC SSTMs are exclusively used for chromatographic method development.

RPC SSTM	HILIC SSTM
L-Glutamine	L-Phenylalanine
L-Glutamic Acid	Hippuric Acid
Creatinine	Adenosine
Cytidine	Adenine
Citric Acid	Taurine
L-Isoleucine	Creatine
L-Leucine	L-Arginine
L-Phenylalanine	L-Tryptophan
L-Tryptophan	Uracil
Hippuric Acid	
Benzoic Acid	
Octanoic Acid	

Table 3-2. Chemical reference standards composition of the RPC SSTM and HILIC SSTM.

The overlap of these reference mixtures with the published reference mixture content used by Metabolon (Evans et al., 2009) and the HUSERMET consortium (Zelena et al., 2009) (including unpublished SOP HU0005) is illustrated below in Figure 3-3.

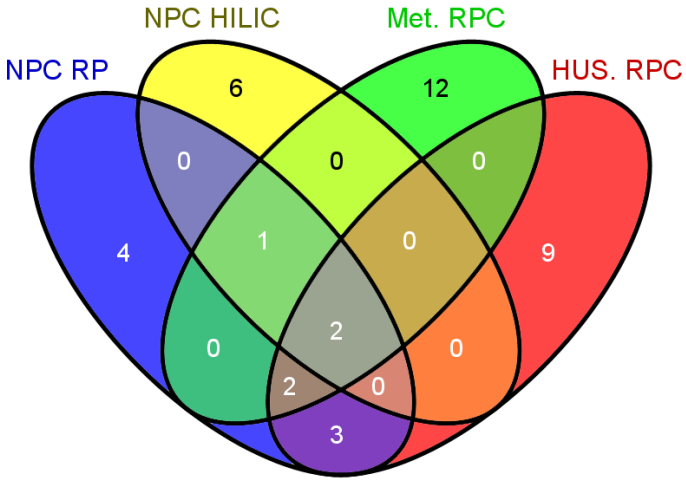


Figure 3-3. Venn diagram illustrating overlap of chemicals used in standardising LC-MS instrument and chromatographic performance across three research groups. RPC and HILIC mixtures from NPC are shown separately. Graphic generated using Venny (Oliveros, 2007).

3.4 Adaptation of chromatographic separations

A molecular profiling system should be well suited to the partitioning of the chemical composition of the biological matrix, yielding specificity and sensitivity by minimising chemical interference. However, all applicable technologies impose a degree of selectivity that is counterproductive to achieving global coverage. Perhaps the greatest source of unwanted selectivity in LC-MS analyses is imposed by the finite ability of any single chromatographic system to retain and separate the vast diversity of chemical species present in human biofluids. Given the wide range of hydrophobicity and chemical structure (e.g. from urea to bile salts) potentially present in the small molecule content of human urine, no single chromatographic approach is suitable for achieving comprehensive coverage. Therefore, two or more assays with orthogonal selectivity may be used in sequence or in parallel, depending on the configuration of the instrumentation platform, to maximise the molecular class

coverage. It is the aim of this section to develop these two chromatographic systems for continuous analysis, maximising the coverage of each individual assay.

3.4.1 Urine analyte hydrophobicity and retention in a reversed-phase system

Reversed-phase chromatography is highly regarded for its broad range of retention, high quality and uniform peak shape, fast equilibration, and excellent precision. Urine nevertheless poses a challenge in RPC retention, as the hydrophilicity of many analytes exceeds that which is retainable by a reversed-phase system. Stationary phases with low ligand density have been engineered to be compatible with completely aqueous sample loading environments (a low ligand density allows for greater interaction of the stationary phase particle material and prevents phase collapse), enhancing the retention of small polar analytes to obtain greater coverage. The method of Want and colleagues (Want et al., 2010) utilises such a stationary phase, and was therefore selected for application to large-scale profiling. This method is referred to hereafter as the "reference method".

An initial goal was to investigate the retention and separation of early eluting molecular species. Analyses were conducted on a representative pooled urine sample using the reference method. The chromatographic details have been explained in the publication by Want et. al., but to summarise, a 2.1 x 100mm Acquity HSS T3 (trifunctional C18 alkyl phase bonded to high strength silica with proprietary endcapping in a 1.8 µm particle size) column (Waters Corp., Milford MA, USA) was held at 40° C and used together with a combination of mobile phases (A = 0.1% formic acid in water; B = 0.1% formic acid in acetonitrile). The separation was performed using the programmed mobile phase compositions and gradients listed below in Table 3-3. The flow rate was held at 0.5 ml/min for the duration of the analysis, and the chromatographic eluate was directed to a mass spectrometer operated arbitrarily in the negative mode (unless indicated otherwise) by electrospray ionisation. Throughout this chapter, all MS detection is made using a Xevo G2-S Q-ToF mass spectrometer (Waters Corp., Manchester UK). Furthermore, due to the improved sensitivity of this instrument model over older generations of ToF mass spectrometers, 2 µl injections of prepared sample were made to the system rather than the 5 µl injections specified by the reference method. This was accomplished in all cases by loading 10 µl of sample onto a 2 µl fixed-volume loop, with the excess draining off as waste.

3.4 Adaptation of chromatographic separations

Time (minutes)	A (%)	B (%)
0	99	1
1	99	1
3	85	15
6	50	50
9	5	95
10	5	95
10.1	99	1
12	99	1

Table 3-3. Chromatographic gradient of the reference method showing the duration and mobile phase composition of each step. A = H2O + 0.1% formic acid; B = acetonitrile + 0.1% formic acid.

An analysis by the above method is illustrated in Figure 3-4 (purple trace), overlaid with an isocratic separation of the same sample at the initial conditions (green trace). The metabolite-dense region of the chromatogram was estimated to be between 0.4 and 8 minutes. Comparison of their TIC traces demonstrates that the first 20% of the elution profile (from 0.4 to 2 minutes) from both separations is virtually identical, indicating the large contribution of isocratic separation to the overall method.

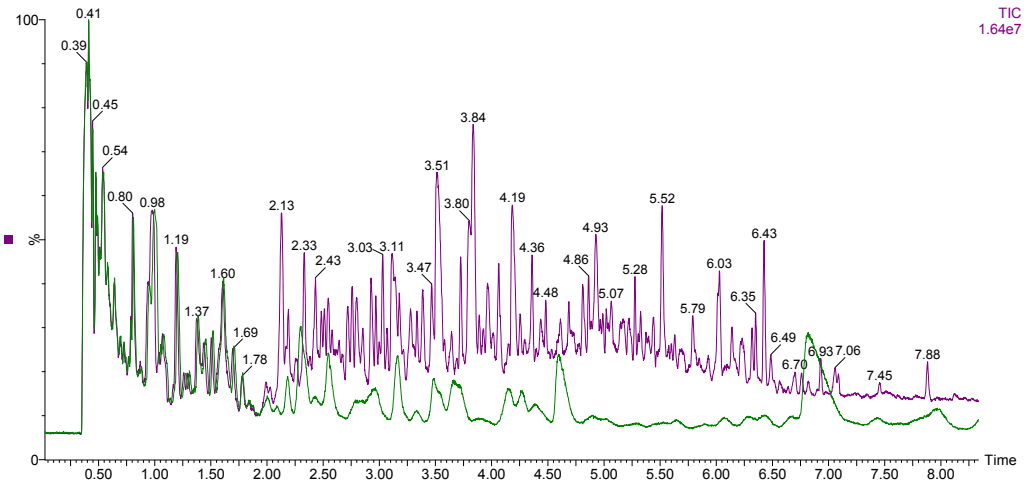


Figure 3-4. TIC traces of a representative urine separation by the method of Want and colleagues (purple trace) and by isocratic elution at initial conditions (green trace).

3.4 Adaptation of chromatographic separations

This initial portion of the chromatogram was observed to be highly metabolite dense. To further elucidate the feature density of the reversed-phase urine separation, feature extraction was performed on the gradient separation data file shown in Figure 3-4 using the centWave algorithm of the XCMS package (described in Section 2.6.2). While conventional feature extraction relies on the response of features across multiple samples within an experiment to differentiate true signals from noise, such a method was not applicable for use on a single sample. Therefore, high noise thresholds were utilised to preclude the incorporation of noise signals as features. Those thresholds along with other relevant centWave parameters are listed in Table 3-4.

The density of the number of detected features (regardless of their intensity) versus their chromatographic retention time were plotted in R, and illustrated in Figure 3-5. While this simple analysis does not preclude the possibility that noise distribution is also higher in the earlier part of the chromatogram, it is a fair indication together with the TIC trace that the feature density is highest in the initial part of the chromatogram. This observation warrants exploration of increasing the initial separation efficiency and distributing the metabolic content more evenly across the chromatogram.

parameter	value
ppm	30
peakwidth	1 to 8
snthresh	50
noise	1000
prefilter	$x = 6$
	$y = 5000$

Table 3-4. CentWave parameters used for feature extraction within XCMS.

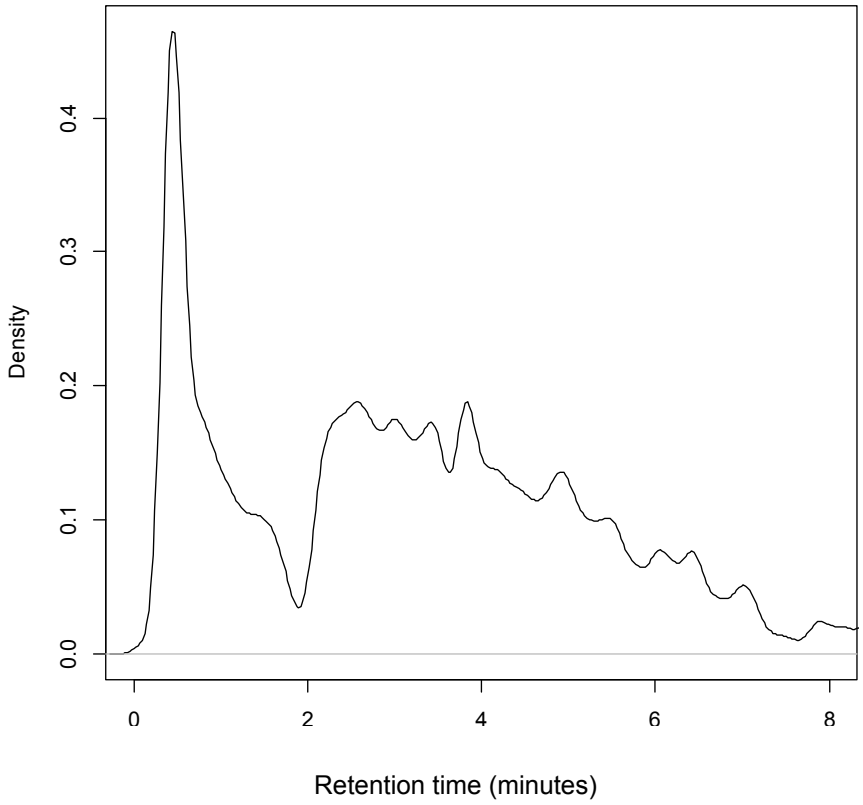


Figure 3-5. Distribution (smoothed count) of chromatographic peaks detected (N = 5103) using the centWave method in XCMS.

The goal of method enhancement is to maximise the density of feature information while improving the overall distribution, as well as to improve chromatographic resolution where possible given the constraints on handling liquid flow imposed by the ESI interface and MS system. As the most information-dense portion of the separation was shown to be obtained by isocratic elution, initial efforts were focused specifically on this region. As separation efficiency under isocratic conditions is directly proportional to column length (as described in Section 2.3.2), a longer column of the same stationary phase (and therefore same plate height) should increase the chromatographic efficiency (and therefore resolution) in the early region of the chromatogram.

To test this, the reversed-phase standards mixture was injected 3 times (sequentially) on three 150mm length columns and three 100mm length columns. The columns were of the same stationary phase type (Acquity HSS T3), and the same inner diameter (2.1mm), but from three different manufacturer

batches to incorporate the real-world effects of column batch variation into the measurement. All injections were performed on the same UPLC-MS system, and columns were tested in order of alternating length. Isocratic separations were performed at the initial conditions of the reference method (99% water, 1% acetonitrile plus 0.1% formic acid) at a flow rate of 0.5 mL/min. The scan rate of mass spectrometric detection was increased to 20 scans per second ensuring accurate definition of peak shape. Three iterations spanning six scans of Savitzky Golay smoothing (Savitzky and Golay, 1964) were applied to all chromatographic peaks prior to measurement of peak width in order to minimise the contribution of detection noise. All peak width measurements were performed by manual evaluation in MassLynx software.

Chromatographic efficiency (N) was calculated on two pairs of early eluting species from the RPC standards mixture; glutamine/glutamic acid and isoleucine/leucine. The glutamine/glutamic acid pair represents very early elution near the injection peak (0.46 and 0.48 minutes, respectively, in a representative analysis using the reference method, compared to an injection peak at 0.37 minutes), while the isoleucine/leucine peak pair elute toward the end of the isocratic elution period (1.46 and 1.58 minutes, respectively, in a representative analysis using the reference method). The resolution of each pair, based on the average peak widths and retention times across triplicate injections for each of three unique columns, was also calculated to provide a representative estimate of the ability of each method to resolve early eluting species. The mass of the [M-H]- ion was used to extract the chromatographic trace of each standard (glutamine = 145.0613; glutamic acid = 146.0453; isoleucine & leucine = 130.0680). Feature extraction and peak width measurements were performed in MassLynx 4.1 software. The retention time of the leading and tailing peak slopes were recorded at half peak height. The difference between the two measurements was recorded as the peak width (W_h). The leading value plus half of the width was calculated as the peak retention time.

As expected, use of the 150mm column (50% longer than the 100mm column) yielded proportional increases in observed column efficiency for the isoleucine (51%) and leucine (47%) chemical standards in the reference mixture as illustrated in Figure 3-6. Greater increases in column efficiency were observed for for the glutamine (98%) and glutamic acid (88%) chemical standards, seemingly due to

their closer proximity to the injection peak. The resolution of the glutamine/glutamic acid separation achieved by the 150mm column was increased by 40% over that value achieved by the 100mm column. The resolution of the isoleucine/leucine acid separation achieved by the 150mm column was increased by 25% over that value achieved by the 100mm column. Greater variation was observed in the peak widths of the later eluting isoleucine and leucine peaks, yielding a greater variance in the calculated efficiency values. This is likely the result of substantial peak broadening and consequential distortion observed in the later part of the isocratic elution.

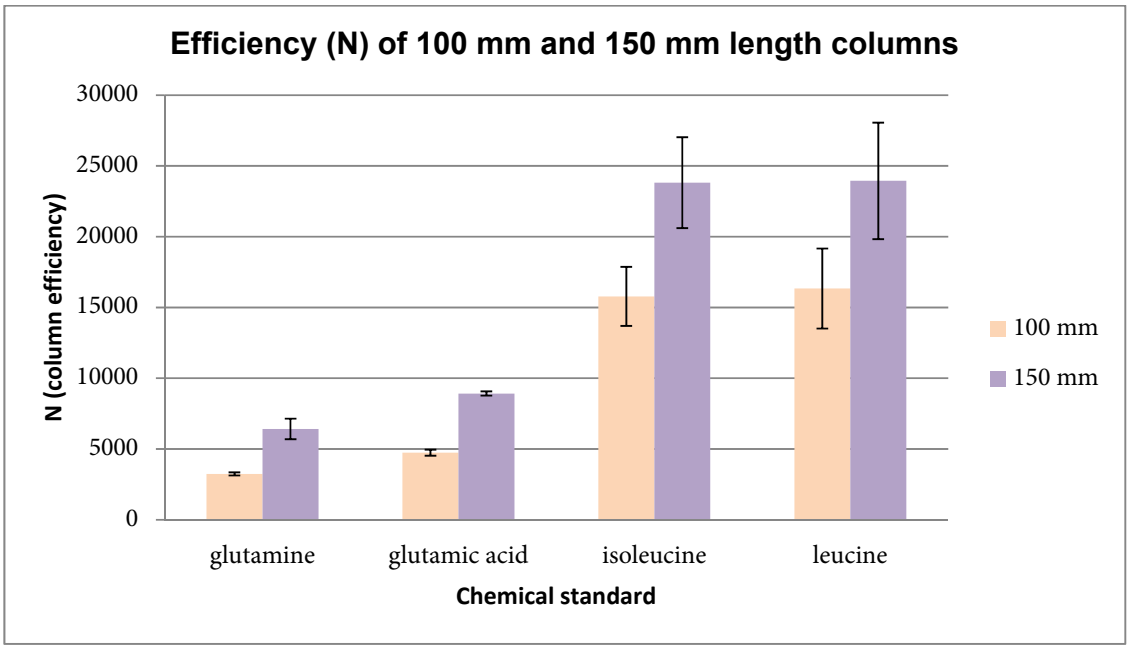


Figure 3-6. Column efficiency calculated using the average values of three replicate injections of the standards mixture, run on three independent columns of each length (100mm and 150mm).

The observed peak broadening of isoleucine and leucine indicated that the isocratic "hold" at initial conditions was excessive in length, contributing to distorted peak shape and low feature density (observable near the two minute retention time mark in Figure 3-5) within the affected chromatographic area. In order to more evenly distribute the metabolic content as well as limit the effects of band broadening observed late in the isocratic portion of the elution, the initial hold was shortened from one minute to 0.1 minute. This change reduced the chromatographic area of purely

isocratic behaviour to approximately the first minute of elution following the injection peak. Subsequent peak elution was uniformly sharp, and the metabolic content was more evenly distributed. These effects are illustrated in the overlaid chromatogram TIC traces from a purely isocratic separation at initial conditions (black), the gradient of Want and colleagues (green) and the same separation with a shortened initial hold (red) shown in Figure 3-7. Collapsing the latter isocratic region also served to shorten the method by nearly one minute, allowing for either a shorter method or a longer gradient separation within the original method duration.

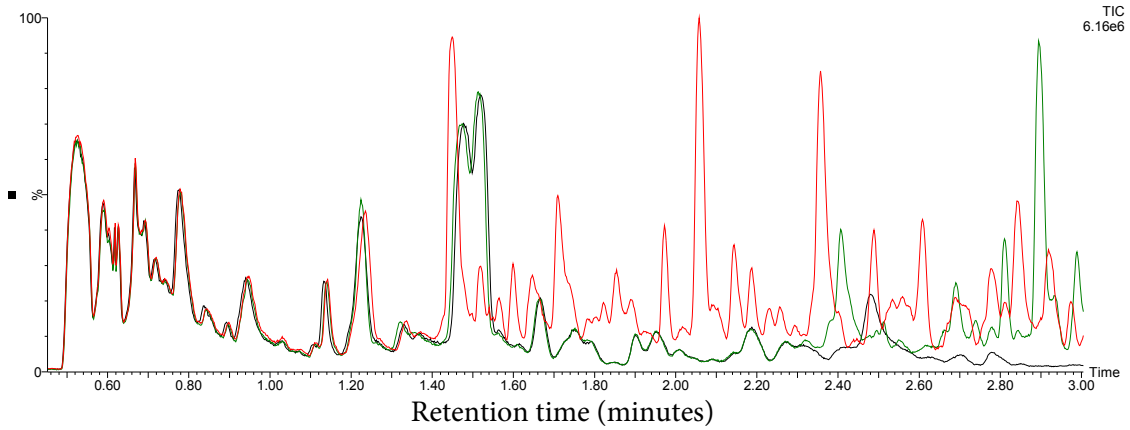


Figure 3-7. Chromatographic separation of the urine LTR using a 2.1 x 150mm Acquity HSS T3 reversed-phase column and three distinct methods. The separation using the reference method is shown in green, with an isocratic separation at initial conditions shown in black. Finally, an adaptation of the reference method with a shortened initial hold is shown in red, contributing to more uniform peak shape and feature density across the first three minutes of elution.

3.4.2 Optimisation of reversed-phase gradient elution conditions

With the initial retention and requisite isocratic separation optimised, the focus of method development was turned to the subsequent gradient elution and the remainder of the chromatogram. Where the sample matrix is complex and the molecular content is well distributed across the chromatogram, a linear gradient is a prudent choice resulting in even and predictable performance across the analysis. Whereas the reference method's gradient separation was segmented into three independently linear stages, a single linear gradient covering the majority of molecular content was desired for the adapted version on the grounds of simplicity and robustness. It was hoped that this decision would translate to increased assay precision, as well as potentially simplifying future retention

time correction, retention time prediction, and potential method transfer efforts (*e.g.* to smaller inner diameter columns, requiring re-mapping of annotated molecules).

The first linear gradient segment from the reference method (ΔB = 7% per minute) was extended to completion (1% to 100% B in 14.14 minutes) in order to elucidate the entire RPC urine chromatogram in linear form (Figure 3-8). Visual evaluation of feature density with respect to chromatographic retention time was used to determine a practical gradient endpoint for the analysis, therefore setting the gradient duration. Both positive and negative mode MS detection were utilised to ensure the total detectable feature density was represented as accurately as possible. In both modes of detection, virtually all of the observable molecular content was eluted before the eight minute mark, corresponding to a solvent composition of 56.3% B (acetonitrile + 0.1% formic acid). This composition was therefore chosen as the end of the analytical gradient segment.

3.4 Adaptation of chromatographic separations

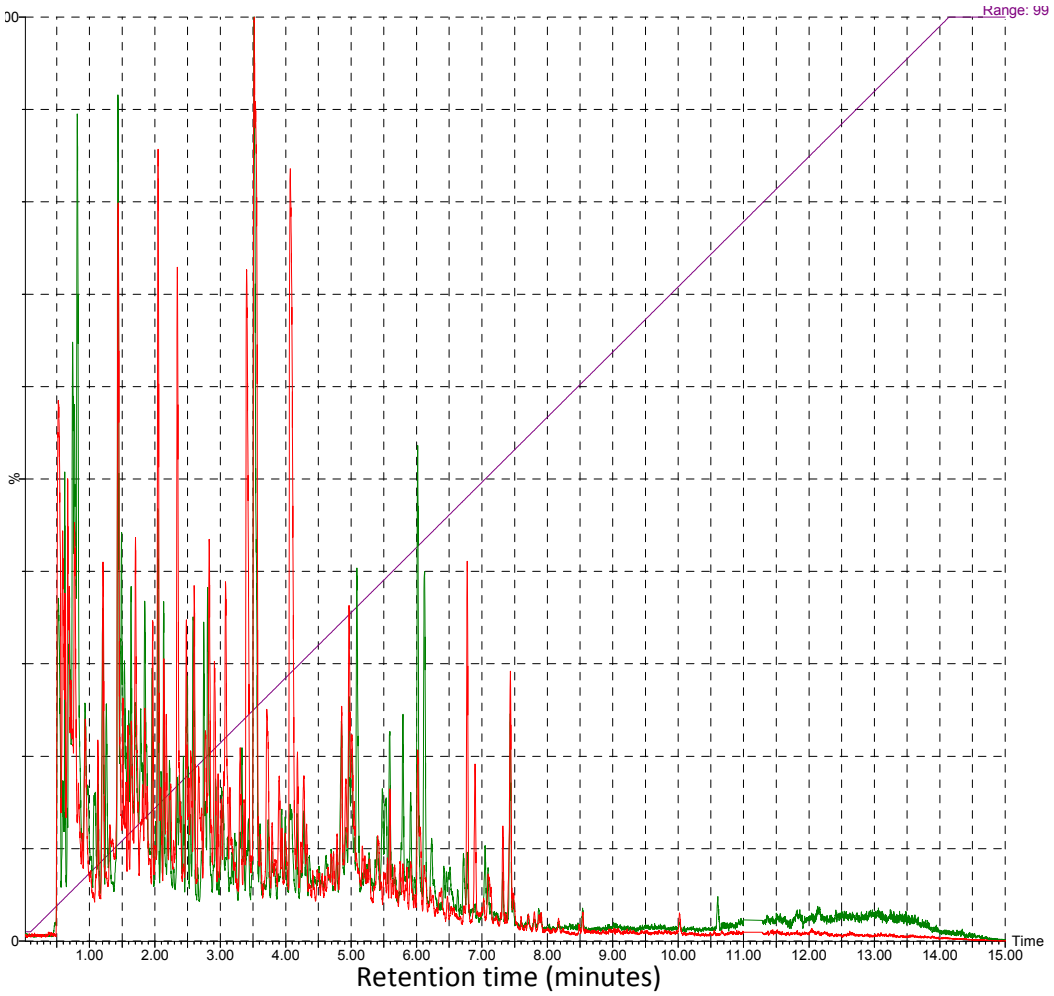


Figure 3-8. Assessment of feature density within a linear separation of the urine LTR. TIC traces are shown for positive (green) and negative (red) MS detection which closely overlap in the data rich region. Finally, the chromatographic linear gradient expressed in %B (acetonitrile + 0.1% formic acid) from 1% to 100% is overlaid in purple.

It is important to note that the contents of urine can vary widely among samples, and therefore development of analytical conditions conducted on a single sample may not be sufficiently representative of the matrix for all steps of development. In the course of evaluating the analytical gradient endpoint, many observations were made of urine specimens donated from individual volunteers, both male and female and of various ages. However, the urine LTR was found to be sufficiently representative of normal healthy urine (perhaps unsurprisingly as it is a combination of 76 unique samples) in this instance and is therefore used herein for illustration. It is also noteworthy that the gradient endpoint chosen based on healthy urine may not be appropriate for all urine samples, as

some pathological conditions are known to result in the "spilling" of more hydrophobic molecular species into the urine. For example, patients with cholestasis will have an abnormally large quantity of bile acids in their urine, many of which elute beyond the bounds of the linear separation presented here (Bove et al., 2004). For this reason, a second "gradient wash" step was introduced to rapidly raise the concentration of organic solvent to 100%, preventing the accumulation of potential hydrophobic species. This gradient was set to approximately 10x the slope of the analytical gradient (rising to 100% B in 0.7 minutes), resulting in rapid elution. Consequentially, chromatographic peaks observed in this area will be composed of less scans than those eluted within the analytical gradient, and some quantitative accuracy may be compromised. However, where the data are sufficiently indicative of the presence of more hydrophobic species, additional investigation by specialised analyses (ie. bile acid profiling) may be warranted for those samples (Muto et al., 2012).

Within the boundaries of the separation defined, the raw performance of the gradient separation may be addressed. The ultimate consideration in the development of a gradient separation of a complex biofluid is the maximisation of the theoretical number of peaks that could be fitted into a given chromatographic space with a resolution of one, also known as the peak capacity. While the Knox equation and van Deemter plots have traditionally been used to identify the optimum flow rate for maximal peak capacity (simultaneously limiting the peak broadening effects of Eddy-diffusion, longitudinal diffusion, and mass transfer) in liquid chromatography, a recent investigation has demonstrated that flow rates far in excess of the theoretical optimum value continue to provide performance gains with small particle size columns (Petersson et al., 2008). Therefore, efforts were undertaken to increase the mobile phase flow rate, requiring specific consideration of the potential limitations of system pressure tolerance and the desolvation capability of the downstream LC-MS interface.

System pressure tolerance is ultimately limited by the specification of the LC pumps at a given flow rate. Fortunately, UPLC pump hardware has been specifically designed to withstand high system pressures. The Acquity UPLC (Waters Corp., Milford MA, USA) model utilised for all work within this thesis has a maximum recommended operating pressure of 15,000 psi at the flow rates considered

herein. The pressure experienced by the pumps is related to the viscosity of the solvent(s) used, which is in turn a function of their composition and their temperature. To minimise system pressure, low viscosity solvents that form low viscosity mixtures are preferentially used. Water and acetonitrile are well suited to this task, yielding a maximum system pressure at approximately 75/25 water/acetonitrile which is only slightly higher than the pressure produced by water alone. This is in contrast to other common solvent combinations for reversed-phase LC-MS such as water and methanol that reach a higher relative maximum pressure when mixed, which limits their use in high flow rate separations.

The viscosity of all LC-MS solvents is reduced at the point of greatest restriction by applying heat to the chromatographic column and column inlet, allowing for increased flow rates to be achieved at a given system pressure. However, increased column temperature can have negative effects on the stability of the stationary phase and therefore the longevity of the column and precision of retention data across an experiment. The retention of early eluting species will also be improved at lower temperatures, benefitting the coverage of the assay. Finally, temperature sensitivity of the analytes in a complex matrix is a concern, as some labile molecular species have been observed to yield lower signals of detection at higher column temperatures (e.g. trichloroacetic acid). However, the relationship between the use of higher flow rates at increased temperatures and the consequential faster elution means that the analytes are exposed to higher temperature for less time, confounding the outcome. While optimisation for a set of target compounds is possible based on empirical observation, application in profiling studies where discovery is often the goal generally adopt a more conservative nature. Therefore, the column temperature was increased 12.5% from the reference method's specification of 40° C to 45° C, which is the manufacturer's maximum suggested operating limit for the stationary phase material.

At this temperature, the maximum pressure endured by the system when performing the water to acetonitrile gradient described above (ΔB = 7% per minute) was recorded for 10 replicate analyses at four distinct flowrates (0.4, 0.5, 0.6, and 0.7 mL/minute). The observed data were extrapolated to determine the flow rate achievable at the maximum system pressure (15,000 psi) recommended by the instrument manufacturer, circumventing the need to physically assess this upper limit and avoiding

risking damage to the pump components (eg. seals), as illustrated in Figure 3-9. Using this approach with lightly used columns (<200 injections), flow rates of up to 0.77 mL/min are estimated to be possible. However, it is prudent to allow some working room to ensure the assay will be robust to variations in system pressure within and among experiments. Doing so helps to buffer against variation in system pressure arising from either upward drift in pressure caused by blockage *via* particulate matter and small amounts of protein precipitated on-column by the strong eluent, or as a consequence of variation in stationary phase manufacturing batches, which produce particles of variable size that contribute to variance in system pressure exerted by each unique column. For these reasons, it is a common rule of thumb to routinely operate at 80% of the maximum tolerable system pressure (12,000 psi) (Petersson et al., 2008). The maximum pressure produced with a flowrate of 0.6 mL/minute fits well with this guidance.

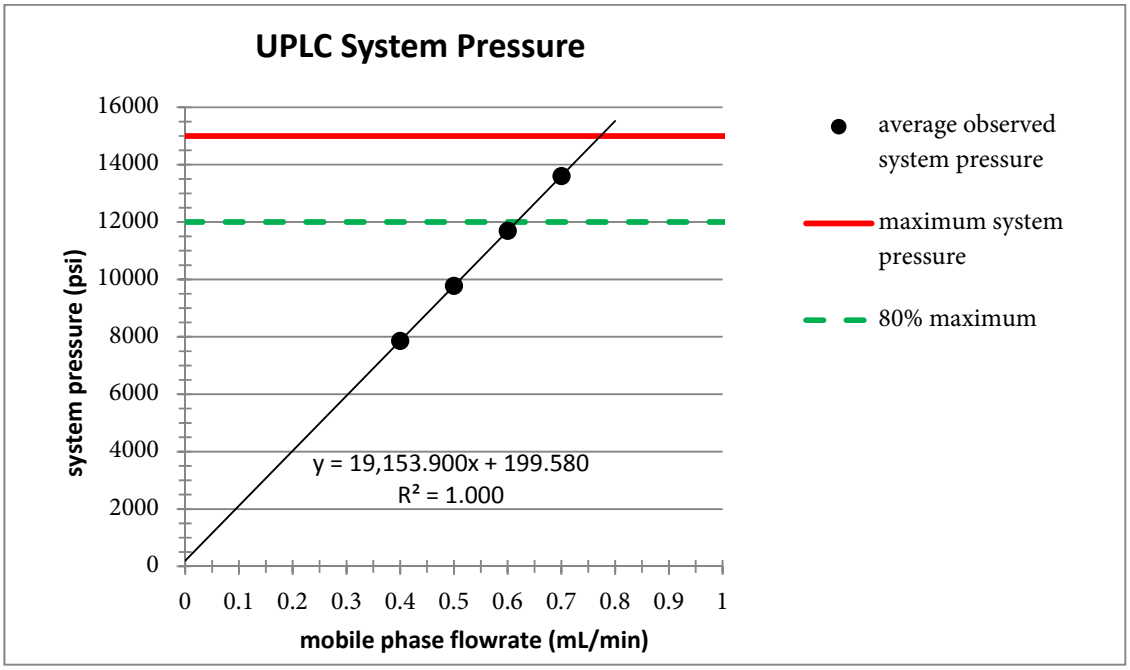


Figure 3-9. The relationship between UPLC system pressure and mobile phase flow rate. Each black dot represents the average of 10 measurements of the maximum system pressure observed during a linear A (water + 0.1% formic acid) to B (acetonitrile + 0.1% formic acid) gradient at 7% per minute for flow rates of 0.4, 0.5, 0.6, and 0.7 mL/minute. Extrapolation of the system pressure as a function of the mobile phase flowrate (black line) shows the maximum tolerated system

pressure (15,000 psi, red line) achieved at approximately 0.77 mL/minute. A flowrate of 0.6 mL/minute is very near 80% of the maximum tolerated system pressure.

However, before making a final determination regarding LC flowrate in a hyphenated system, the capability of the downstream components must also be considered. In the case of LC-MS, the (electrospray) ionisation interface has a finite ability to ionise and desolvate LC effluent. Greater flow can reduce the ionisation and desolvation efficiency resulting in an apparent decrease in detected signal for a given chemical quantity. This effect was illustrated by repeated analysis of the RPC SSTM at the flowrates of 0.4 to 0.7 mL/minute as tested above, and the integrated areas of selected reference standards eluting at various mobile phase compositions are illustrated in Figure 3-10. Selection of flowrate in such a hyphenated system is therefore a compromise between separation performance and sensitivity (and therefore molecular coverage). For all LC-MS development presented herein, a maximum flow rate of 0.6 mL/min was chosen as desolvation appeared complete in the source (with aggressive desolvation settings of 600° C and 1000 L/hr nitrogen flow), leaving no residual liquid at mobile phase conditions from 100% aqueous to 100% organic. Flow rates greater than 0.6 ml/min of mostly aqueous LC effluent were found to be incompatible with lockspray hardware (specific to Waters MS instruments), resulting in the rapid accumulation of liquid on the lockspray baffle. During longer periods of blocking the mobile phase spray, such as during the automatic tuning of the detector gain between analyses, this accumulation was occasionally accompanied by suction of droplets into the source cone, disrupting the data from subsequent analyses. The resulting 20% increase in mobile phase flow rate (and 12.5% higher column temperature) resulted in earlier solute elution across the gradient. Therefore the gradient slope was made slightly shallower to compensate (6.667% B per minute instead of 7%, equalling 1% per 9 seconds) and the analytical gradient endpoint was rounded down to 55%.

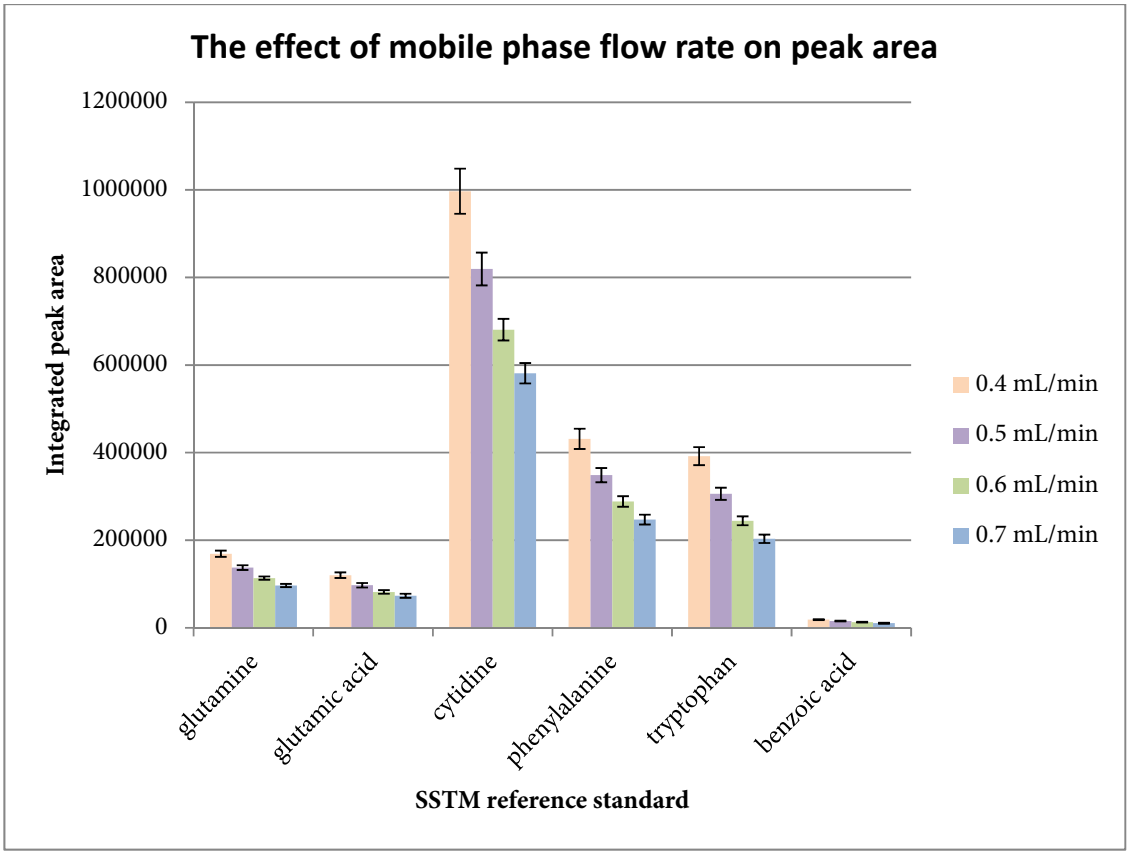


Figure 3-10. Chromatographic peak area of selected SSTM reference standards decreases with respect to increased mobile phase flow rate.

Finally, prior to the calculation of peak capacity achieved by the modified method, all washing and equilibration steps were carefully optimised to maximise the volume of solvent delivered (column volumes) in a given amount of time without exceeding the system pressure limits imposed by the UPLC pumps. Chromatographic column volume (V_M) was estimated by calculating the volume of the column as an empty cylinder and assuming that 40% of that space is occupied by the stationary phase material. The remaining 60% around and within the porous particles may therefore be occupied by mobile phase. The equation used herein for V_M is illustrated below, where r is the cylinder's inner radius and h is its height (*i.e.* column length).

$$V_M = 0.6(\pi r^2 h) \tag{3.1}$$

3.4 Adaptation of chromatographic separations

By this estimate, the 100mm and 150mm columns have column volumes of 0.21 and 0.31 mL, respectively. The gradient flow rates were tuned to achieve a high and nearly constant pressure as the solvent composition (and therefore viscosity) changed from completing the gradient, through the high organic wash, return to initial conditions, and equilibration at initial conditions. A comparison of the system pressure traces generated by the optimised and reference methods is illustrated in Figure 3-11.

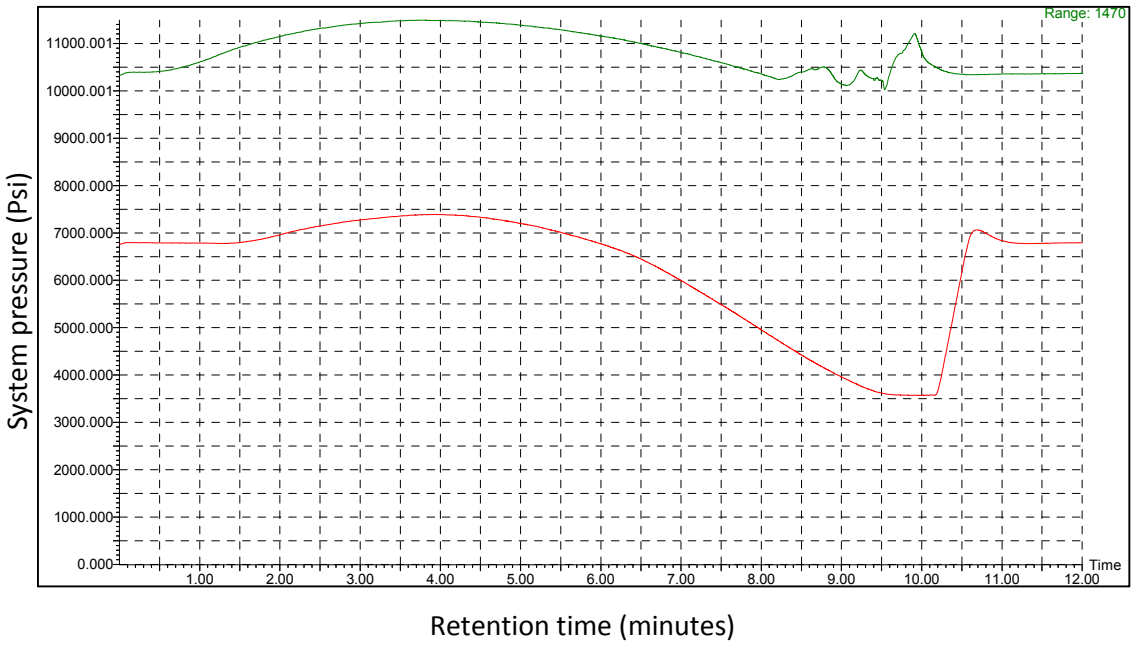


Figure 3-11. System pressure traces of the optimised (green) and reference (red) methods. The trace of the optimised method illustrates the efficient utilisation of available system pressure to enable faster column washing and equilibration to compensate for its greater length and therefore increased volume.

In this manner, the number of column volumes applied to the column for washing and equilibration steps were held similar to those used in the reference method (Table 3-5). This was achieved within the same overall gradient time despite the use of a 50% longer column.

Chromatographic method	Reference method	New method	
Column length	**100 mm**	**150mm**	
initial hold	2.41	0.19	Column volumes
gradient	19.25	15.59	
gradient wash	n/a	1.24	
high organic wash	2.41	1.76	
return to initial	0.24	0.32	
equilibration	4.57	5.03	

Table 3-5. Comparison of the chromatographic column volumes used in the reference and optimised RPC methods.

Following careful real-time evaluation of the change in system pressure (Δ = psi/minute) during the column equilibration step, it was concluded that, as an independent measure, an additional half minute of equilibration at a flowrate of 0.6 mL/min (0.96 column volumes) should be added to the method. Doing so consistently produced a lower Δ value, indicating a more complete equilibration of the column in 99% aqueous conditions. Functionally this would be expected to manifest in greater and more precise retention of solutes, and therefore an assessment was attempted using the SSTM, randomly varying column equilibration times among values of 10, 5, 2.5, 1.25, 0.625, 0.3125 and 0.15626 minutes. A tendency for all analytes to elute slightly earlier with each subsequent injection (regardless of the equilibration time applied) was observed, requiring a linear correction to compensate for this run order effect. Once the correction was applied, the chromatographic retention of all species were observed to rapidly decline when equilibration was conducted for less than two minutes. The data suggest a trajectory of equilibration whereby increasing time yields improved analyte retention, but with diminishing returns. This result is illustrated in Figure 3-12 using cytidine (RT ≈ 0.88 min) as a representative example. The total column volumes for equilibration were therefore set at 3.6, and the overall method length set at 12.5 minutes. The final method is described in Table 3-6.

3.4 Adaptation of chromatographic separations

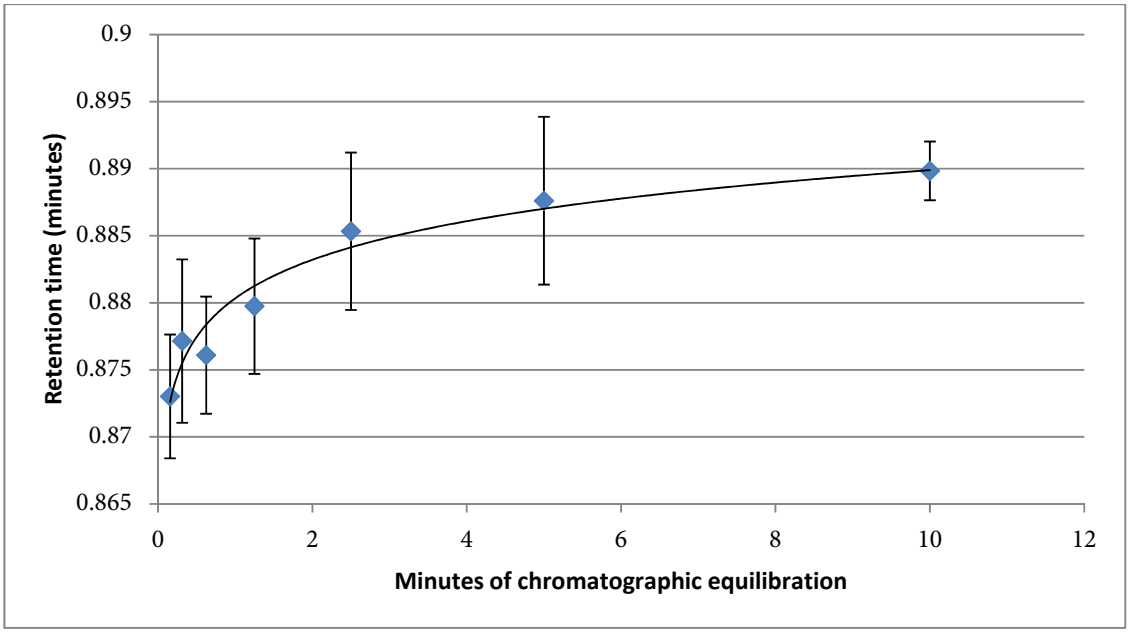


Figure 3-12. Cytidine retention as a function of column equilibration time at initial chromatographic conditions (99:1 water-to- acetonitrile with 0.1% formic acid added).

Time (minutes)	Flow Rate	A (%)	B (%)	purpose
0.00	0.60	99	1	
0.10	0.60	99	1	initial hold
8.20	0.60	45	55	gradient
8.35	0.61	35	65	gradient wash
8.50	0.63	25	75	gradient wash
8.65	0.67	15	85	gradient wash
8.80	0.75	5	95	gradient wash
8.90	0.80	0	100	high organic wash
9.20	1.00	0	100	high organic wash
9.40	1.00	0	100	high organic wash
9.50	1.00	99	1	return to initial
9.55	0.90	99	1	equilibration
9.65	0.80	99	1	equilibration
9.75	0.70	99	1	equilibration
9.85	0.65	99	1	equilibration
9.95	0.61	99	1	equilibration
10.00	0.60	99	1	equilibration
12.50	0.60	99	1	equilibration

Table 3-6. Chromatographic gradient of the optimised RPC method, showing the programmed gradient times and mobile phase composition (A = H2O + 0.1% formic acid; B = acetonitrile + 0.1% formic acid).

3.4.3 Assessment of reversed-phase peak capacity

The peak capacity of the optimised method was calculated and compared to that of the reference method. For additional comparison, the optimised chromatographic separation was also applied to the 100mm column. Capacity was calculated using the reversed-phase reference mixture and the same columns and approach implemented in the measurement of chromatographic efficiency (Section 2.3.3), with two notable exceptions. First, the reversed-phase standards mixture was injected ten times (instead of three) on each column in order to more precisely assess retention and peak width. The first three (of ten) injections were used as conditioning injections, and the last seven used in peak capacity calculations. Second, while three iterations of Savitzky Golay smoothing were again used, the span for each round of smoothing was reduced from six to two scans, as the peak shapes were sharper than those observed in the isocratic separations. All peak width measurements were again performed by manual evaluation in MassLynx software. Glutamine, isoleucine, tryptophan, hippuric acid, benzoic

acid and octanoic acid were selected from the reversed-phase standards mixture for the calculation of peak capacity due to their even spacing across the chromatogram, defining five discrete chromatographic segments. Peak capacity was calculated and reported using the sum of each segment resolution as described in Section 2.3.3. The overlaid EIC's for each marker in both the reference method (top) and optimised method (bottom) along with the gradient of each expressed in %B are illustrated in Figure 3-13.

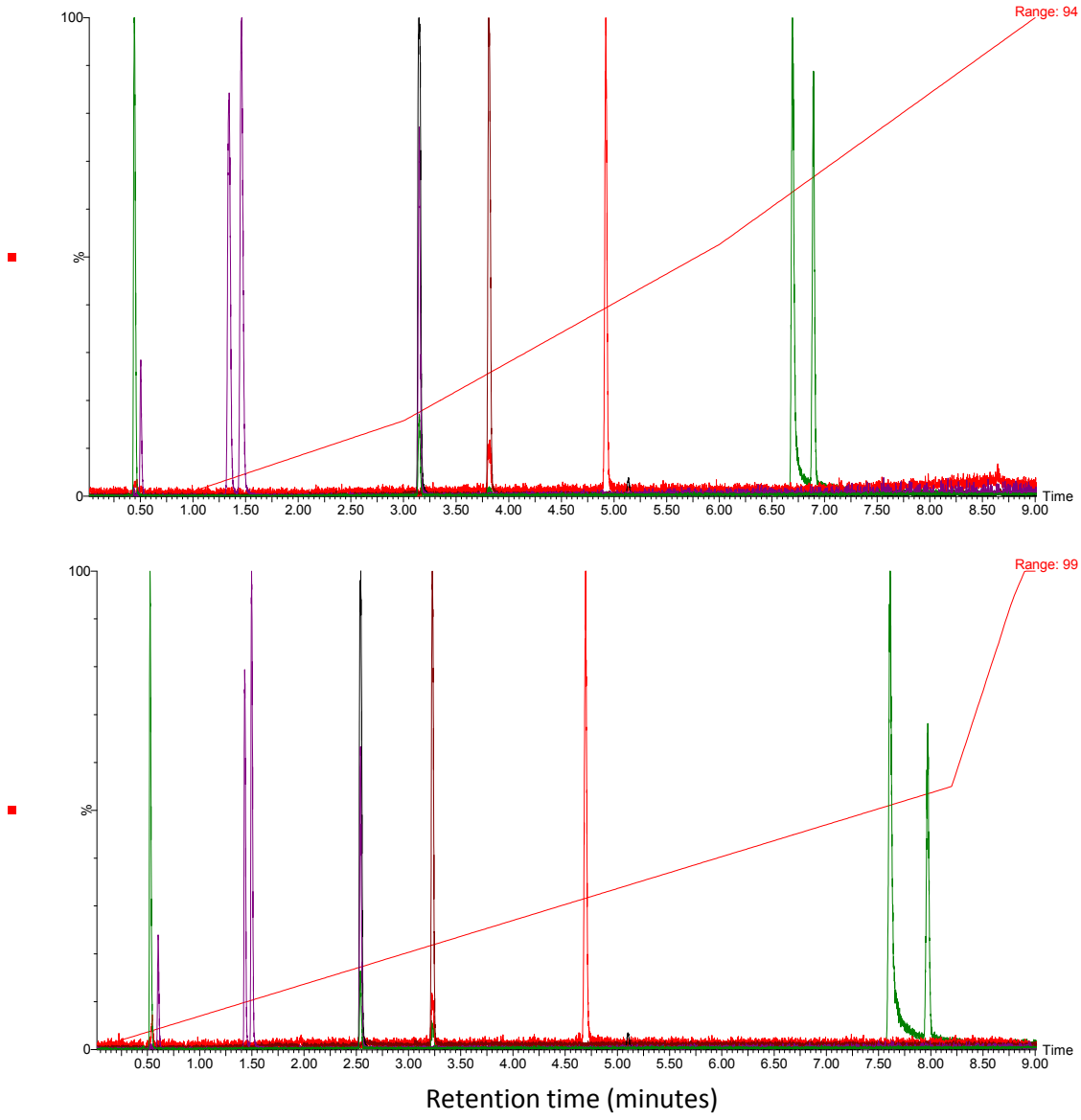


Figure 3-13. Selected SSTM reference standards were extracted from a single analysis by the reference method (top) and optimised method (bottom). In order of elution, the chemicals are: glutamine (green), isoleucine (purple, first of two closely eluting peaks), tryptophan (black), hippuric acid (brown), benzoic acid (red) and octanoic acid (green, first of two closely eluting peaks – the second is an unknown contaminant). The gradient for each method (shown as %B) is overlaid in red.

Ten injections of urine LTR were also made on each column in order to assess the feature distribution of urinary metabolites using each method. Feature extraction was performed on injections 4-10 for each batch using XCMS and the parameters listed in Table 3-7.

parameter	RPC-MS
ppm	20
peakwidth	1 to 8
snthresh	50
noise	1000
prefilter	x = 6
	y = 5000

Table 3-7. XCMS parameters for the extraction of features from urine LTR analyses using both the reference and optimised chromatographic RPC methods.

Taken as a whole, the optimised method produces 21% more peak capacity (using Equation 2.9, or 24% using the average method of calculating peak capacity shown in Equation 2.10) over the course of the analysis than the reference method between the first and last retention time markers glutamine and octanoic acid. Interestingly, those gains were not evenly distributed across the analysis. The optimised method was found to have increased peak capacity in the first, third, fourth, and fifth segments by an average of 29%, 24%, 27%, and 37% respectively. A 13% decrease in peak capacity was observed in segment two was a consequence of closing the gap in feature density at two minutes which represents an area of unusable capacity. These values are illustrated for each chromatographic segment in Figure 3-14 as projected onto a feature density map from one representative set of urine analyses using the reference method.

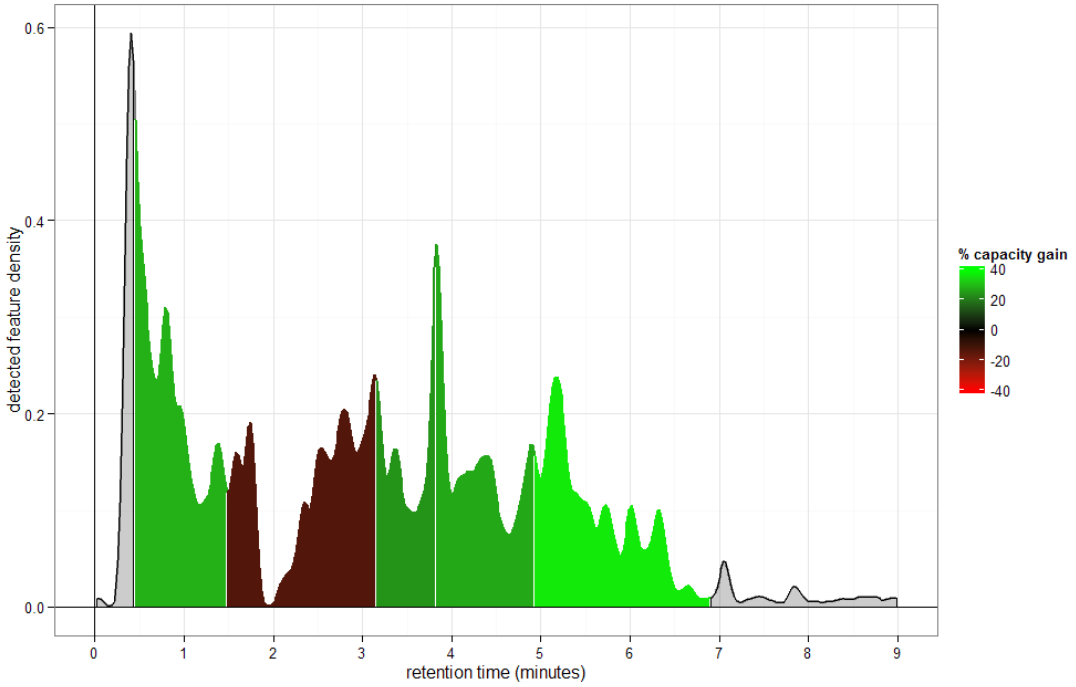


Figure 3-14. Density of detected features in relation to chromatographic retention time using the reference method. The colour represents the percent gain (or loss) in peak capacity achieved by the optimised method versus the reference method as measured between retention time markers from the RPC SSTM.

The gap in feature density was closed by the shortening of the one minute isocratic hold at initial conditions. As stated previously, this precludes peak distortion from excess isocratic elution, allowing for more uniform peak shape and distribution in the early-to-mid portion of the chromatogram. This is clearly illustrated by extraction of a selected metabolite cluster spanning the relevant area of chromatographic retention such as the EIC (m/z=153.058 +/- 0.1 Da) shown in Figure 3-15, generated from representative analyses of the urine LTR using the reference (top, red) and optimised (bottom, black) methods. The observed decrease in peak capacity is therefore largely an artefact based on the insertion of an area of distorted peak elution, violating the assumption in calculating capacity that peak shape is approximately uniform throughout the area of measurement.

3.4 Adaptation of chromatographic separations

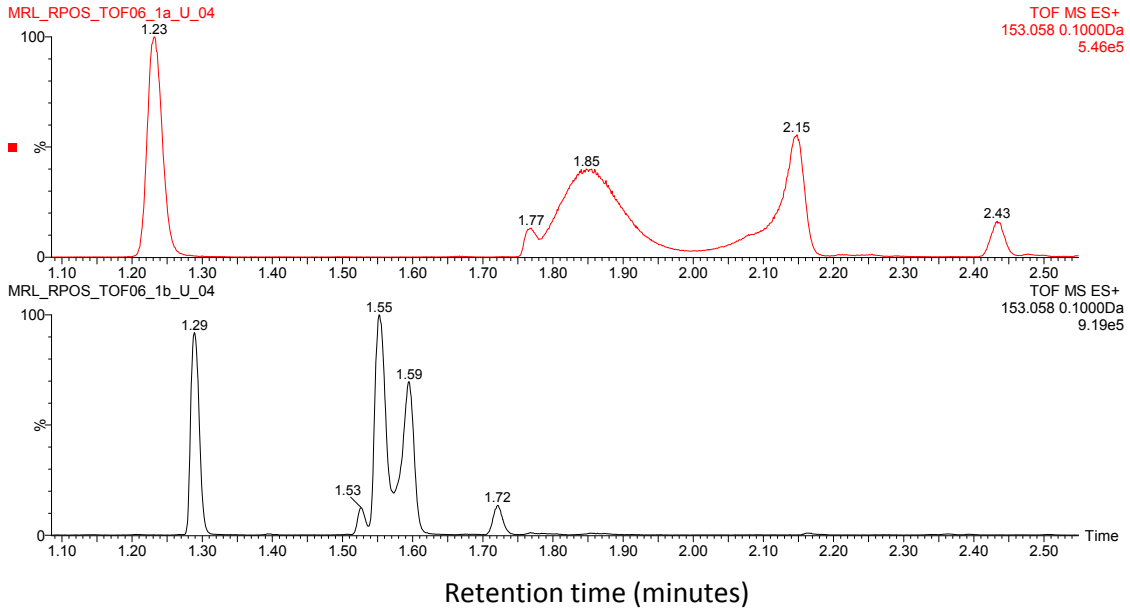


Figure 3-15. EIC of a metabolite cluster (m/z=153.058 +/- 0.1 Da) spanning the area of chromatographic distortion caused by one minute of isocratic elution at initial conditions. Distorted peaks produced by the reference method (top) are shown against the same peaks, eluting earlier and with more uniform shape, produced by the optimised method (bottom).

The start and end retention time bounds of the analyte content distributed across the analysis remained largely static despite the change in column length and gradient, but with a different distribution of features throughout each elution. These gains in capacity were therefore made within the same approximate chromatographic "footprint", and not due to simple elongation of method. To illustrate this, the feature density from two representative urine analysis sets (reference method = grey; optimised method = blue) are shown below in Figure 3-16.

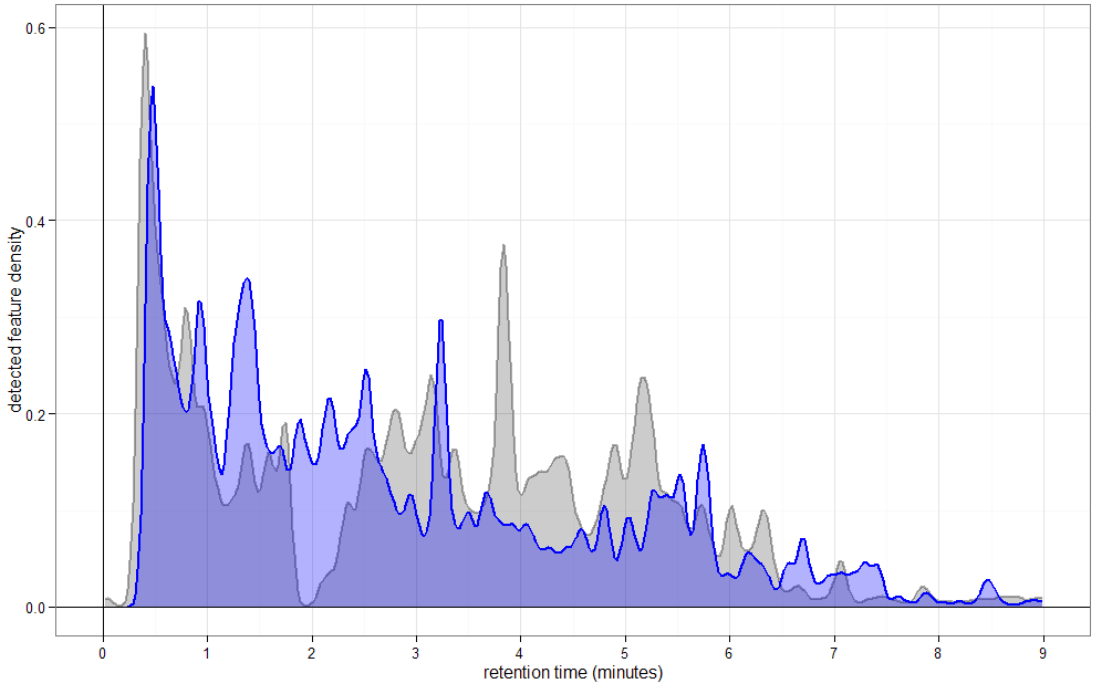


Figure 3-16. Overlay of the feature density distribution between the reference method (grey) and optimised method (blue), illustrating the approximately equivalent footprint of chromatographic elution between the reference and optimised methods.

In order to assess more directly the effect of the 150mm column length versus the changes made to the elution method, the capacity achieved by the optimised method was also assessed using the 100mm column. The overall gain in peak capacity across the entire analysis (from glutamine to octanoic acid) was more marginal at 11%. However, the majority of that improvement was observed within the first segment (21%) as expected from the outcome of the efficiency and resolution testing under isocratic conditions.

3.4.4 Adaptation of HILIC for complementary retention and separation of small polar analytes in urine

Despite the emphasis on polar metabolite retention in the selection and refinement of the reversed-phase chromatographic separation, the technique is inherently limited in what it can achieve as it is ultimately dependent on solute interaction with a hydrophobic medium. As a consequence, many molecular species still escape its grasp, warranting the development and implementation of a

complementary approach for metabolite retention and detection. HILIC is a strong candidate for this, as its mechanisms of metabolite retention allow for the elution of solutes in order of increasing polarity (Tang et al., 2014). Using a polar stationary phase, the sample matrix is loaded onto the column in the presence of an organic solvent (typically aprotic, such as acetonitrile) with a small amount of water. It is commonly believed that the water forms a pseudo-layer on the polar stationary phase, and therefore analyte retention is largely based on liquid-liquid partitioning between the organic and aqueous components of the mobile phase. Use of reversed-phase solvents makes the application of HILIC to LC-MS an attractive alternative to other separations aimed at polar analyte separation, and similarly make it an attractive solution for application here, as much of the development in Sections 3.4.1 and 3.4.2 can be applied in a similar manner to yield increased performance. The HILIC method of Want *et. al.* (Want et al., 2010) again serves as an appropriate reference method for adaptation, as it has since been successfully applied in the analysis of urine (Spagou et al., 2011). The gradient program is shown in Table 3-8.

Time (minutes)	A (%)	B (%)
0	99	1
1	99	1
12	0	100
12.1	99	1
15	99	1

Table 3-8. Chromatographic gradient of the HILIC reference method showing the duration and mobile phase composition of each step. A = 95% acetonitrile and 5% ammonium acetate; B = 50% acetonitrile and 50% ammonium acetate. The final concentration of ammonium acetate in both mobile phases is 10mM.

Experience among colleagues within CSM indicates that the successful preparation of HILIC solvents as described is not straightforward, owing to the low solubility of the volatile buffer salts (eg. ammonium acetate or ammonium formate, required for adequate peak shape) in a 95:5 mixture of acetonitrile and water. Their solubility may be encouraged by sonication or direct heating of the solvent, yet it is often observed that the aqueous buffer salt separates after preparation as the solvent

cools, appearing first as a visible haze (emulsion) and finally coalescing as droplets of immiscible liquid at the bottom of the solvent. Furthermore, a volumetric preparation of 95:5 organic solvent-to-water ratio in bulk quantity (*e.g.* 1L) is not convenient to produce with commonly sized volumetric flasks. As any small error in the percentage of water present in the initial separation can have dramatic effects on analyte retention (Gray et al., 2013), the precision of solvent preparation is of the utmost importance in HILIC separations. Therefore, it was rationalised that greater consistency in solvent composition could be achieved by simplifying the solvent B to a completely aqueous salt buffer, and solvent A to a completely organic (acetonitrile) solvent. The UPLC instrument is then used to mix A and B in a precise and reproducible manner, achieving 95:5 initial conditions.

As a consequence of this design, the HILIC separation became both a gradient separation in terms of aqueous content as well as buffer concentration, as only one mobile phase component contained the salt buffer. For this reason, the buffer type and concentration was briefly assessed. Whereas the reference method as implemented by Spagou *et. al.* contains both ammonium acetate and formic acid, potentially facilitating the creation of both formate and acetate adducts in negative mode ionisation, ammonium formate and formic acid were used, simplifying the anions present in solution. It was reasoned that a 20mM preparation in water (mobile phase B) would provide an amount of ammonium formate commonly used in LC-MS separations across the range of HILIC elution, from 1mM at initial conditions (5% A) to 10mm at final conditions (50% A). In this manner, the amount of buffer salt required to be solubilised by the 95:5 acetonitrile and water mixture was reduced 10-fold to 1mM at initial conditions. Finally, both A and B solvents were doped with 0.1% formic acid as used in the reference method. The final solvent compositions were therefore: A = acetonitrile + 0.1% formic acid; B = 20mM ammonium formate + 0.1% formic acid.

With the chemistry adapted for increased reproducibility of solvent preparation and mixing over the reference method, the focus was turned to adaptation of the gradient for improved performance. As with the RPC separation, the amount of available system pressure represents an opportunity for performance enhancement. As the elution gradient in HILIC is essentially the reverse of a reversed-phase system, with high organic initial conditions and increasing the aqueous component as the strong

eluent, it can be expected that the pattern of system pressure is reversed as well, but otherwise similar. However, when utilising acetonitrile and water as the mobile phase components, the maximum pressure produced in HILIC applications is less than that produced in RPC applications as the amount of water required to cleanly elute all solutes generally does not surpass 50%, and therefore the maximum pressure-producing ratio of approximately 75:25 water:acetonitrile is not reached. For this reason, the potential exists to use mobile phase flowrates in HILIC separations that are higher than those used under similar conditions for RPC. Furthermore, the higher proportion of organic solvent makes the eluent easier to desolvate and yields excellent signal intensity.

Given the availability of system pressure, and following the principles demonstrated in the RPC development, the 100mm length column was exchanged for one of 150mm length. Furthermore, the flowrate was increased by 50%, from 0.4mL/minute in the reference method to 0.6 mL/minute. The availability of system pressure at the start of the method allows for flowrates in excess of 1mL/minute, which were utilised in the equilibration phase which is reported to be critical for general chromatographic reproducibility in HILIC (Gray et al., 2013). The one minute isocratic hold was again shortened to 0.1 minute in order to achieve more uniform peak shape and even the solute distribution in the eluate. The overall method duration was standardised to 12.5 minutes in order to match the cycle time of the RPC method, allowing a pair of instruments to analyse samples simultaneously using both methods, and on the same sample preparation schedule. These changes culminated in the chromatographic method listed in Table 3-9, applying the column volumes listed in Table 3-10 in a manner that is 16.6% shorter than the reference method.

Time (minutes)	Flow Rate	A (%)	B (%)	purpose
0.00	0.60	5	95	
0.10	0.60	5	95	initial hold
6.85	0.60	50	50	gradient
8.00	0.60	50	50	high aqueous wash
8.10	0.605	5	95	return to initial
8.20	0.61	5	95	equilibration
8.30	0.62	5	95	equilibration
8.40	0.65	5	95	equilibration
8.50	0.70	5	95	equilibration
8.60	0.80	5	95	equilibration
8.70	0.90	5	95	equilibration
8.80	0.90	5	95	equilibration
10.80	1.00	5	95	equilibration
11.00	0.60	5	95	equilibration
12.50	0.60	5	95	equilibration

Table 3-9: Chromatographic gradient of the optimised HILIC method, showing the programmed gradient times and mobile phase composition. A = 20mM ammonium formate + 0.1% formic acid; B = acetonitrile + 0.1% formic acid.

Chromatographic method	Reference method	New method	
Column length	**100 mm**	**150mm**	
initial hold	1.92	0.19	Column volumes
gradient	21.17	12.99	
high aqueous wash	n/a	2.21	
return to initial	0.19	0.19	
equilibration	5.58	11.11	

Table 3-10: Comparison of chromatographic column volumes used in the reference and optimised HILIC methods.

With respect to sample preparation, the reference method uses entirely aqueous sample (dilute urine). However, it is well known that approximating the sample solvent to the initial mobile phase conditions of any chromatographic separation will lead to improved analyte retention and chromatographic performance. Therefore, the addition of three volumes of acetonitrile to the sample was tested, and

observed to dramatically improve chromatographic peak shape. In order to accurately target the available dynamic range of the Xevo G2-S Q-ToF, human urine samples required dilution with an equal volume of water prior to preparation with acetonitrile. In the final method, a 2 µl injection of the 8x diluted urine is used for each analysis.

HILIC SSTM was analysed by the optimised method utilising the sample preparation procedure described above. The plot of combined EIC's in Figure 3-17 illustrates the distribution of the diverse molecular content, further highlighting the method's applicability to urinary solutes. The peak capacity achieved by both methods was compared among five replicate analyses of the SSTM using the manual assessment approach described for RPC peak capacity analysis in Section 3.4.3, with the exception that the average peak width was used to calculate the average peak capacity (Equation 2.10) rather than summation of the segments between reference standards. This was necessary due to an observed change in the elution order of some analytes. Using this approach, the peak capacity of the optimised method was calculated to be 90% greater than that of the reference method (44.4 for the optimised method *vs.* 23.4 for the reference method). This improvement was achieved in 16.6% shorter analysis time, compared to the reference method (12.5 minutes *vs.* 15 minutes).

Finally, it was observed during initial tests with urine samples that negative mode ionisation and detection yielded very few features when paired with the separation conditions described. An analogous separation was attempted using ammonium bicarbonate and an amide-bound stationary phase (Acquity BEH Amide). However when applied to the separation of human urine the results were not encouraging in relation to observed feature distribution across the gradient elution, and the development was suspended. Therefore, the optimised method is only used with ESI+ detection.

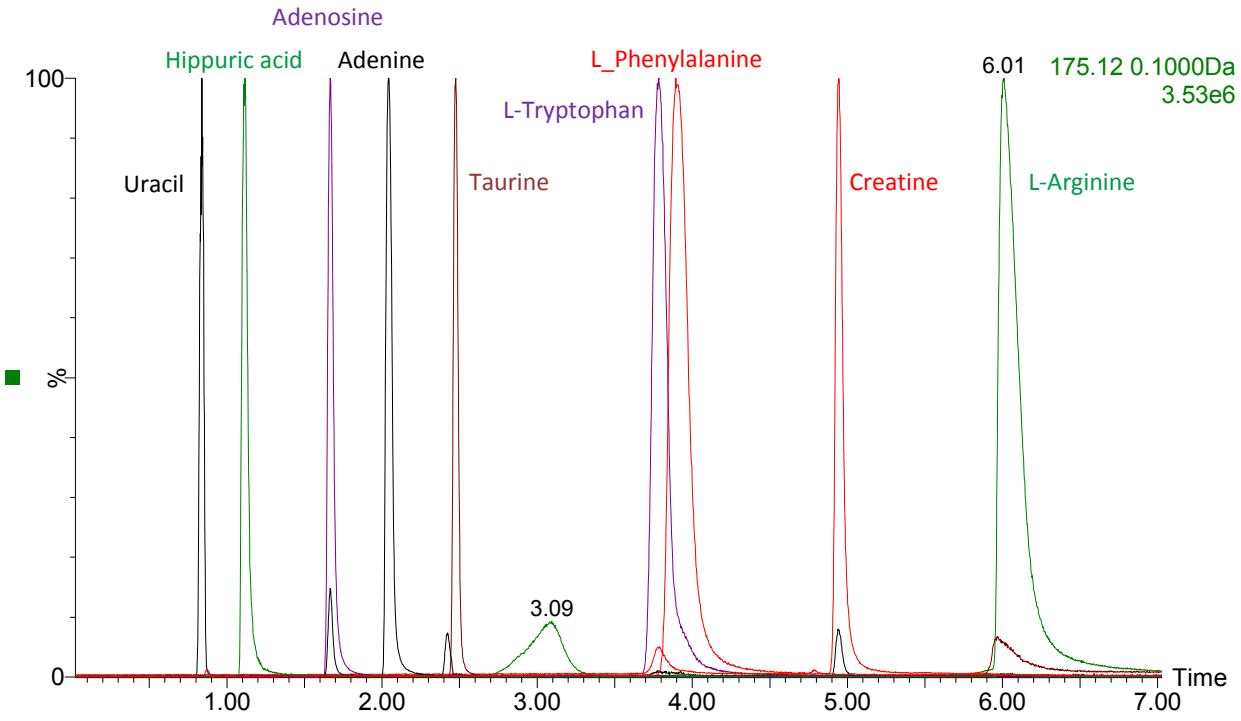


Figure 3-17. HILIC separation of the HILIC SSTM. The SSTM is observed to cover the chromatographic space of the developed assay, providing points of retention reference throughout.

3.4.5 Assessment of separation complementarity

In order to maximise the amount of urinary metabolite species measured in a set of assays, the individual methods must be shown to be complementary in nature, with each elucidating unique matrix content. Specifically, between the RPC and HILIC assays developed here, each should be able to demonstrate complementary retention for the content of a representative urine sample. However, a purely global comparison is not practical given the complexity of the matrix as demonstrated by visual comparison of the LTR urine analysed by both RPC-MS and HILIC-MS (in positive ionisation mode) shown in Figure 3-18.

3.4 Adaptation of chromatographic separations

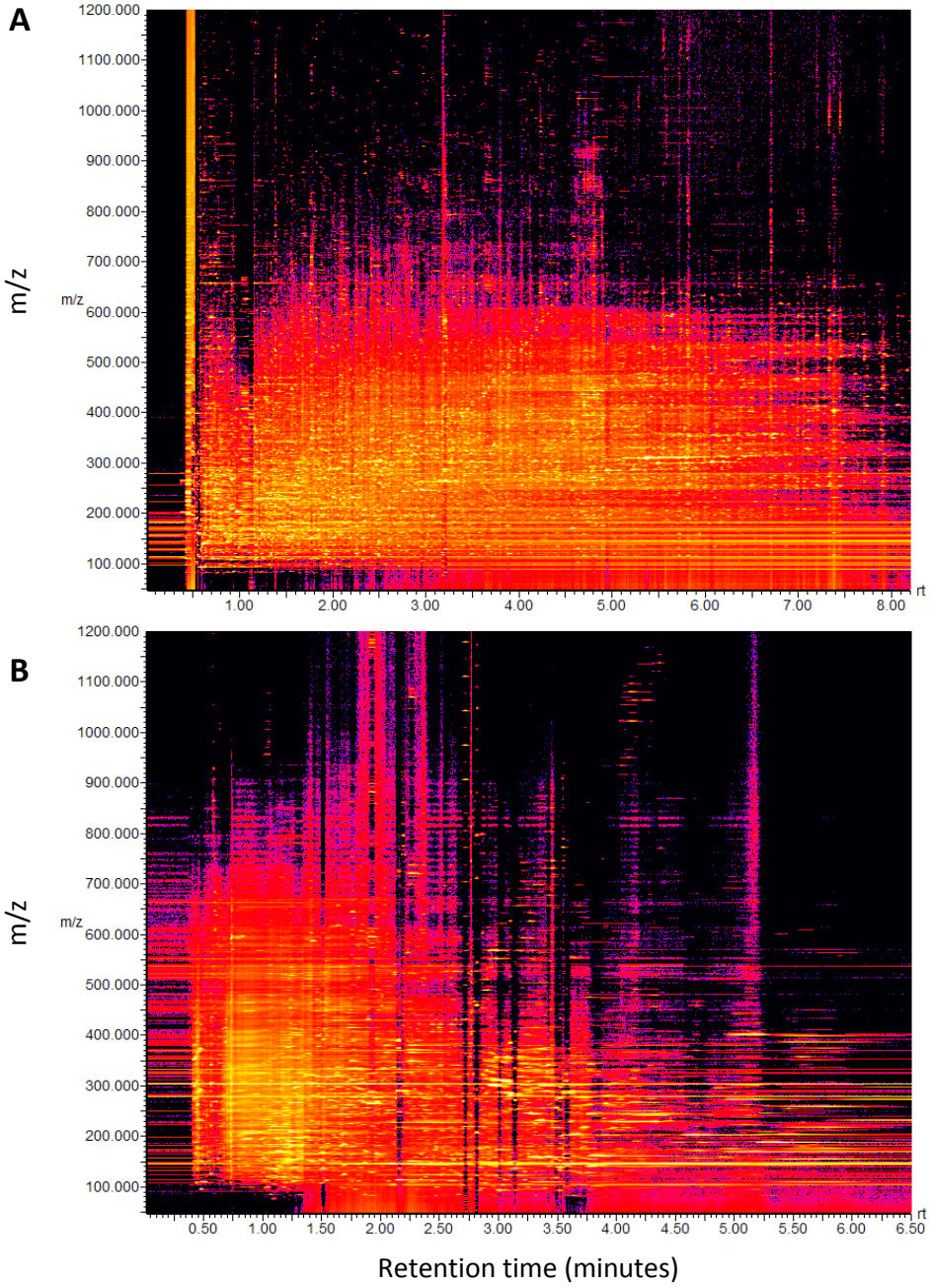


Figure 3-18: Two-dimensional plots of orthogonal LC-MS urine separations. Both RPC (A) and HILIC (B) separations are complex and feature-rich, making direct comparison of the complementarity of retention and chromatographic elution an impractical task that requires simplification.

3.4 Adaptation of chromatographic separations

While an evaluation of the standards mixture developed in Section 3.3.2 may provide insight to the specific molecules contained within, its complexity does not adequately represent that of a urine matrix to a sufficient degree required for assessing overall complementarity. Therefore, a compromise was sought in order to evaluate the complementarity of the RPC and HILIC methods whereby the urine sample, representing all species likely to be present, was fractionated into a series of more simple mixtures of endogenous content for use in direct comparison of the methods.

Briefly, LTR urine was concentrated by overnight freeze drying of 20 mL in a 50 mL Florence flask. The dried urine was solubilised by addition of 2 mL LTR urine followed by brief vortexing, creating an 11x concentrated solution. LTR urine was used for solute dissolution rather than water in an effort to replace any metabolite species lost in the process of freeze drying (albeit at a 1x concentration). Fractionation by reversed-phase chromatography was performed using a column similar to that used in the analytical separation described above, except that the particle size was larger (by approximately 67%) and the column inner diameter larger (by approximately 120%) to accommodate greater sample loading and therefore the separation of more biomass per injection. For this, a 4.6 x 150 mm Atlantis T3 column with 3µm particle size (Waters Corp., Milford MA, USA) was held at 25° C during the separation, and a gradient elution performed with water+ 0.1% formic acid (A) and methanol + 0/.1% formic acid (B) at a constant flowrate of 1mL/minute and the program shown in Table 3-11. Methanol was selected as the strong eluent to give a degree of elution complementarity to the analytical method described above which instead utilises acetonitrile. Twenty microliter (full loop) injections of the 11x urine preparation were made for each separation.

3.4 Adaptation of chromatographic separations

Time (minutes)	A (%)	B (%)
0	100	0
1	100	0
16	5	95
20	5	95
21	100	0
25	100	0

Table 3-11: Chromatographic method used for fractionation of the concentrated urine LTR sample. Programmed gradient times and mobile phase composition (A = H2O + 0.1% formic acid; B = methanol + 0.1% formic acid) are shown.

Initial injections were observed by MS detection using a Xevo TQ-S tandem quadrupole mass spectrometer operating in scanning mode, and the method was found to produce baseline peakwidths of approximately 6 seconds. Fractions were therefore collected at a closely matching rate of 9 seconds per fraction, allowing the collection of 120 fractions from the start of analysis to halfway through the high organic washing step (0 to 18 minutes). In order to accumulate sufficient biomass for further analysis, each injection and separations cycle was repeated for 44 consecutive rounds, yielding a total of 880ul fractionated 11x concentrated urine. Fractions from all separations were collected into a single set of 120 10mL polystyrene culture tubes using a Waters Fraction Collector III (Waters Corp., Milford MA, USA). The eluate content of each tube was evaporated under nitrogen (10psi) in a 37° C water bath using a TurboVap LV nitrogen dryer. Drying time was variable, usually between 5 and 16 hours, depending on the eluent composition (more aqueous or more methanol), and also on the specific desolved solutes. Once dry, all fractions were solubilised by vortexing and sonication in a volume of ultrapure water sufficient for multiple subsequent analyses (1.76mL). The final concentration of an analyte in a fraction was estimated to be between 2.75 and 5.5 times the native concentration (in the original urine sample), depending on the distribution of an analyte peak across adjacent fractions (best case in a single fraction, worst case split evenly between two). Each solution was transferred to a well of a 96-well plate for convenient aliquoting for further analyses and stored at -80° C until required.

3.4 Adaptation of chromatographic separations

All fractions were analysed by the RPC and HILIC UPLC-MS analyses described above, with interleaving analysis of the original LTR pooled urine sample. The overlaid chromatograms of fraction 14 (green trace) and the adjacently analysed LTR (red trace) are shown in Figure 3-19, with RPC analysis shown at the top, and HILIC analysis shown at the bottom. This selected example of a single fraction illustrates the complementarity of the two chromatographic approaches, showing molecular content eluting near the injection peak in the RPC system well distributed across the HILIC separation.

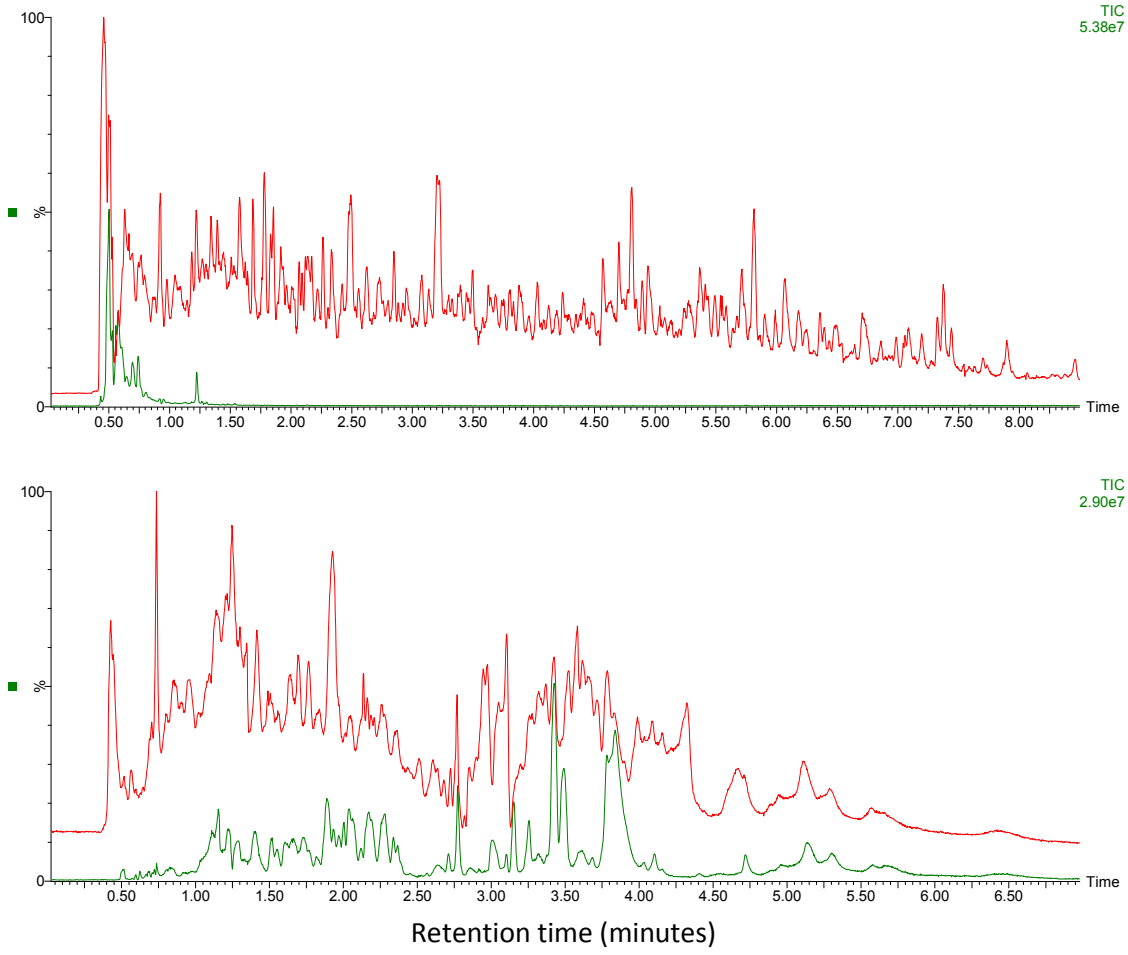


Figure 3-19: Illustration of a selected fraction (fraction 14, shown in green) analysed by RPC (top) and HILIC (bottom) chromatography. Adjacent urine LTR analysis is shown overlaid (in red) to provide context for the urine separation for each method. Fraction 14 is representative of other early fractions which are poorly retained by RPC but well retained and distributed by HILIC.

3.4 Adaptation of chromatographic separations

In order to more globally assess the complementarity of the two separation techniques, the density of feature distribution for each fraction analysed by each chromatographic method was assessed. Feature extraction using the centWave algorithm of XCMS was performed on each data file individually as previously described. The settings for peak detection varied slightly between methods, as the peak shape in HILIC is more variable and potentially wider than that observed in RPC due to the more diverse mechanisms present in HILIC for solute retention and separation. The centWave parameters used for feature extraction of each dataset are summarised in Table 3-12.

parameter	RPC-MS	HILIC-MS
ppm	20	20
peakwidth	1 to 8	2 to 30
snthresh	50	50
noise	1000	1000
prefilter	$x = 6$	$x = 6$
	$y = 5000$	$y = 5000$

Table 3-12: XCMS parameters for the extraction of features from urine fractions analysed by both RPC and HILIC methods.

Features detected in each data file (from each fraction for each chromatographic method) were collected into retention time groups of 15 seconds. Feature density was then calculated as the number of detected features per retention time group. The collated results are presented in the heat maps in Figure 3-20. A high and constant background of detected features was observed in the second retention time bin of all fractions in the HILIC dataset including the earliest "blank" fractions. This was determined to correspond to chemical noise at the injection peak, and therefore the values in that single retention bin were substituted with zero values in order to avoid artificially biasing the plot scale and obscuring the density of observed features in the remainder of the plot.

3.4 Adaptation of chromatographic separations

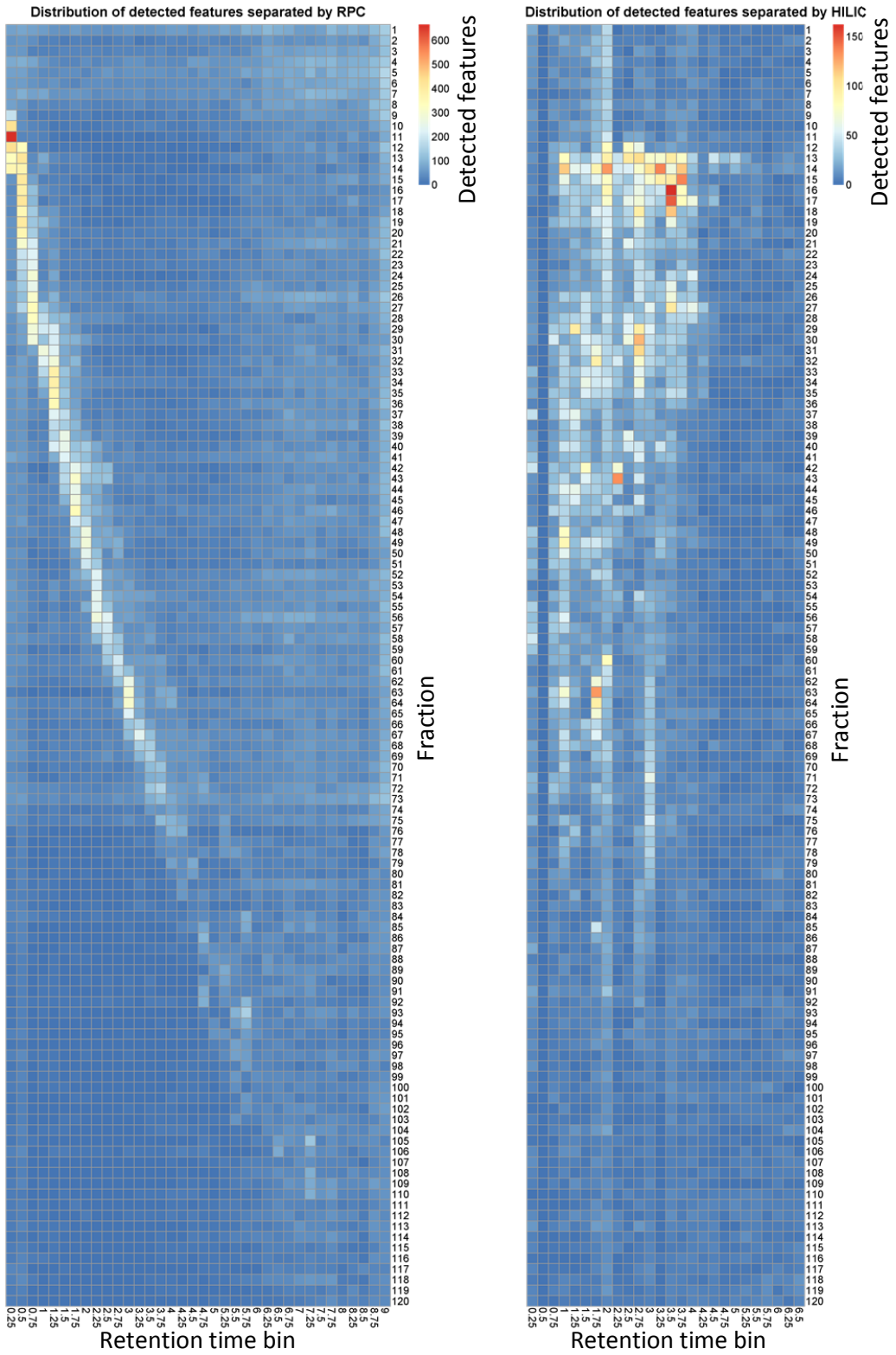


Figure 3-20. Heat map representation of the density of detected features in 0.25 minute retention time bins (x-axis) for each fraction (y-axis) by both RPC and HILIC optimised methods.

3.4 Adaptation of chromatographic separations

As expected, the distribution of detected features in the RPC analysis across increasing retention time is observed to correlate with increasing fraction number (Figure 3-20, left panel), as the fractionation itself was performed by RPC on a larger scale. However, the correlation between retention time and fraction number is less than perfect due to the use of methanol instead of acetonitrile as the strong eluent in the separation. The difference between the selectivity of the preparative separation (using methanol) and analytical separation (using acetonitrile) can be observed in many fractions where individual molecular species clearly differentiate in retention from the bulk of the eluted material. An example of this is illustrated in fraction 17 (Figure 3-21, top), where the majority of the molecular content elutes in the narrow band between 0.4 and 0.9 minutes, except for two clearly deviating peaks at 1.94 and 2.47 minutes. The latter was tentatively identified as 5′-Deoxy-5′-(methylthio)adenosine by targeted MS/MS-derived fragmentation pattern match to reference spectra. The contents of some later fractions show a wide retention time distribution in the analytical method, as illustrated by the chromatogram of fraction 49 (Figure 3-21, bottom).

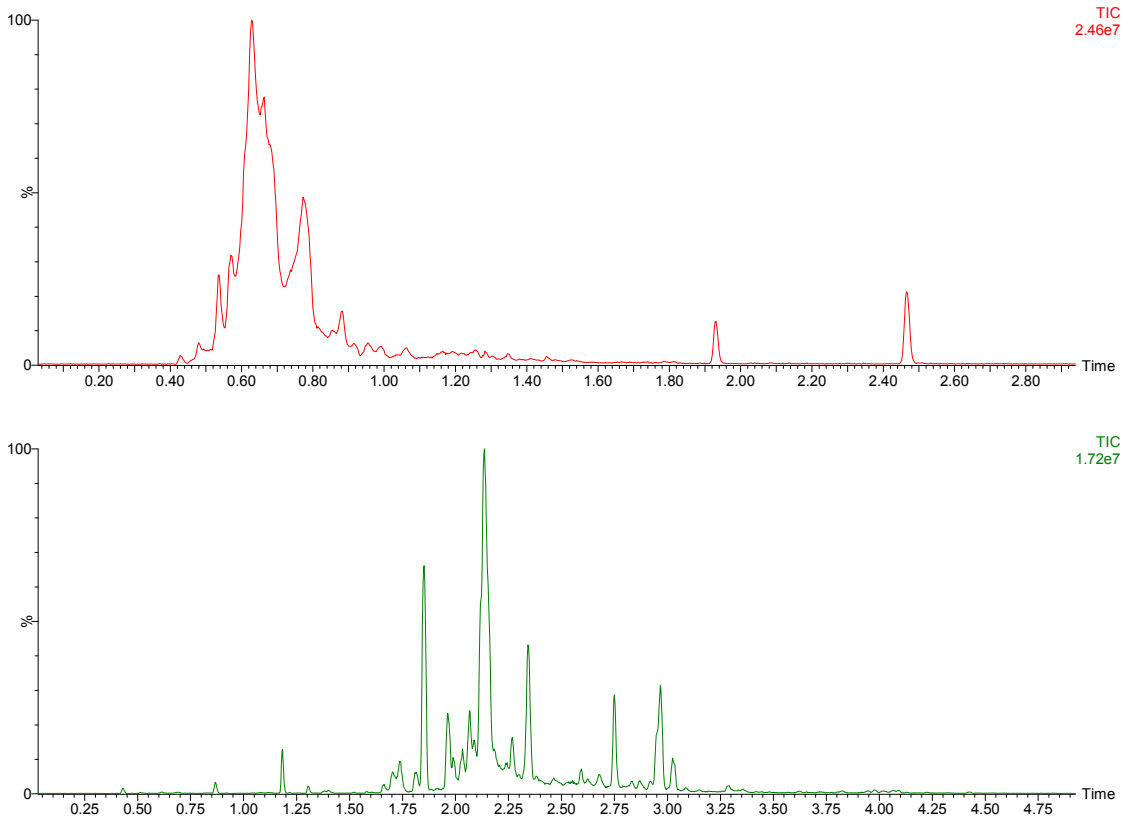


Figure 3-21: The molecular contents of representative fractions 17 (top) and 49 (bottom) derived from HPLC RPC separation of LTR urine using methanol as the strong eluent are mildly dispersed by the analytical RPC method utilising acetonitrile. This is due to the difference in selectivity between the two non-polar solvents when used in combination with the same stationary phase chemistry.

The HILIC separation, on the other hand, demonstrates no visible correlation with the fraction order, indicating orthogonally and good complementarity to RPC. Specifically, early-fraction molecular content not well retained by RPC is observed to be well retained and chromatographically distributed by HILIC. This analysis therefore explicitly demonstrates, in the context of a complex biofluid made interpretable by fractionation, that the methods adapted herein contribute a more complete coverage of the urine metabolome than either method used alone.

3.5 Maximisation of analytical batch size through optimisation of the LC-MS platform.

The adaptations of existing chromatographic approaches made in section 3.4 were intended to render the chromatographic methods fit for long term use with excellent long term precision and, for HILIC separations, ease of solvent batch preparation. With these refined and complementary chromatographic methodologies in place to provide broad urinary metabolome coverage, the focus of enhancing the precision of measurements in large studies turned to the downstream components of the analytical system. In order to ensure that the LC-MS platform yields maximally reproducible results over a long duration, the impact of the analysis on the instrumentation must be minimised, allowing the extension of analytical batch size and the reduction of distinct batches that need to be corrected for post-acquisition.

As the sample matrix itself interacts with all components of the analytical system, influencing their performance over time, it is logical to attempt to reduce the amount of sample used for analysis. In this manner, the major sources of drift in chromatographic retention, ionisation, and detection efficiency caused by repeated analysis may be minimised. Although migration to micro or nano-scale LC is an increasingly attractive means for scaling down, it often comes at the expense of assay robustness (Noga et al., 2007), making such approaches less attractive for deployment on large-scale. As the aim of this thesis is to optimise traditional UPLC, those options will not be explored. Rather, a reduction in the requisite materials will be developed within the constraints of the UPLC methods developed in the previous section.

3.5.1 MS sensitivity as a currency for longitudinal precision

The selection and use of highly sensitive instrumentation is key in achieving large-scale analysis, as it provides both improved metabolic coverage and acts as a currency to be traded for gains in longitudinal precision. The relatively wide pore in the sampling cone of the Xevo G2-S Q-ToF combined with the high ion transmission efficiency of the stepwave ion guide make the instrument highly sensitive, allowing for smaller volumes of more dilute sample to be assayed, in turn minimising the impact of the sample on the chromatographic and ion source components. Ideally the ToF mass

spectrometer should be designed such that the whole population of ions available are utilised, further maximising sensitivity. However, sensitivity and mass resolution are inherently at odds with each other in a ToF mass analyser, as a narrower ion beam with controlled energy spread will produce a higher resolution m/z measurement. Commercial constraints placed upon instrument manufacturers to meet or exceed a specified resolution value result in common tuning configurations that which narrow the ion beam through use of ion optical elements and slits (e.g. an aperture plate). By shaping the ion beam, very high mass resolution values can be achieved at great expense to sensitivity and dynamic range, as less of the full beam is passed on to the detector.

The philosophy adopted herein is that the importance of instrument sensitivity and dynamic range are prioritised above that of mass resolution in UPLC-MS profiling applications where the chromatographic separation greatly reduces the incidence of mass interference from co-eluting species. Custom tuning of the mass spectrometer's optics is therefore warranted to maximise sensitivity at the reasonable expense of resolution. To achieve this routinely, voltages on the ion optics affecting the focus of the ion beam onto the entrance slit prior to the pusher region of the ToF assembly (acceleration lens and aperture) were tuned for maximum signal of the monoisotopic peak of cytidine (m/z = 244.0933) from infusion of the RPC SSTM. The voltage across the top and bottom half plates of the steering lens (steering) were adjusted for symmetrical peak shape, while the pusher offset (voltage applied to the pusher with respect to the entrance voltage) and reflectron grid voltage were tuned for best peak shape and resolution without sacrificing the intensity gains. The resolution values produced by the Xevo G2-S Q-ToF mass spectrometer tuned as specified here are generally between 14,000 and 17,000 FWHM.

3.5.2 Source optimisation for sensitivity and minimal impact of sample on MS inlet

When the HILIC and RPC methods described are mated to the Q-ToF instruments configured as outlined above, the amount of sample required for analysis may be reduced to better fit the instruments sensitivity and linear range of detection. The consequence of this is less biomass accumulation in the source area, reducing the need to stop analysis and clean the instrument, thus introducing a new batch. Both methods described here utilise a 2 µl injection of dilute urine. Urine for RPC is diluted with an

equal volume of water, and therefore each analysis only introduces 1 µl of urine to the instrument. Urine for HILIC is diluted even further, owing to the increased sensitivity achieved under HILIC conditions, with only 12.5% of the 2 µl (0.25 µl per injection) being urine.

Furthermore, because of the efficiency of ion intake and transmission to the mass analyser, the electrospray position may be taken further away from the source inlet than experience with older generation Q-ToFs would allow, again keeping the source cone from accumulating biomass for longer periods of analysis (see Figure 3-22 for a visualisation of the source cone and probe angle). Greater liquid flow rates utilised in both the HILIC and RPC methods (as compared to the flow rates in the reference methods) inherently necessitate a greater distance between the electrospray capillary and MS inlet orifice, as the electrospray droplets produced are larger and require more time for gas phase ion generation from a greater number of coulombic fission events. The effect of probe position (in terms of adjustable probe angle relative to the cone) on observed signal intensity was therefore assessed using the RPC SSTM and RPC separation method previously developed. Analysis was repeated across the range of probe positioning (positions 4 to 10 in intervals of 1). Maximum values for signal intensity were obtained at intermediate settings for most chemical reference standards, being both far enough from the cone to facilitate more complete desolvation of the LC eluent, and close enough for efficient ion intake (for example, tryptophan and hippuric acid, bottom panel of Figure 3-23). However, other molecular species showed clear preferences for more direct spray into the cone (creatinine) while others showed the opposite preference, gaining signal intensity as the distance between the spray and cone was increased (phenylalanine). The observed variation likely reflects complex arrangements among molecular species within charged droplets formed by electrospray. A probe position of seven was chosen as an intermediate value providing excellent signal intensity across the board, but distancing the spray sufficiently from the cone to allow for minimal accumulation of biomass on the cone surface. In this manner, the sensitivity gained by the use of sensitive instrumentation and fit-for-purpose ToF tuning allows for more conservative approach to sample volumes and probe positioning, minimising the impact of repeated sample analysis on the UPLC and ionisation source.

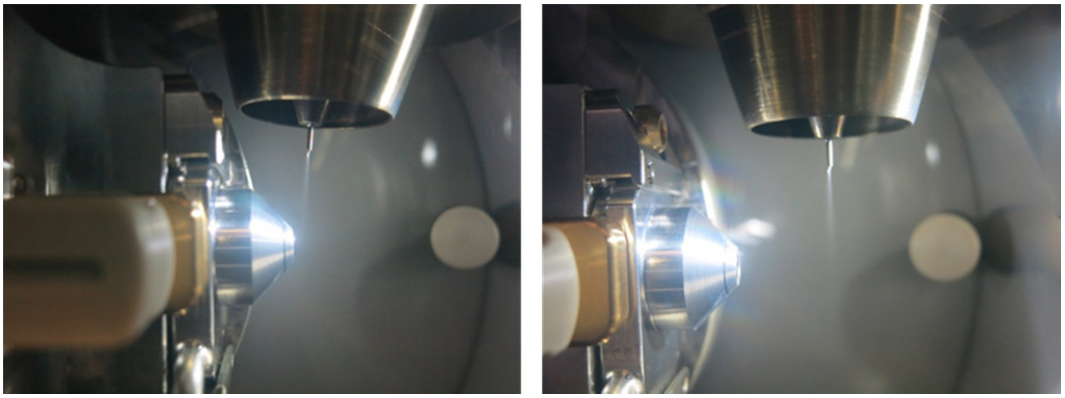


Figure 3-22: Illustration of an electrospray probe angle in relation to the MS inlet cone. The photograph on the left shows an aggressive probe position (lower values, closer to 4) while the photograph on the right shows a more conservative angle (higher values, closer to 10). Note that these photos were obtained on a Xevo TQ-S for clarity, as the lockspray assembly present on a Xevo G2-S Q-ToF obscures the view of the spray and cone during operation. The source designs are otherwise the same for the purposes of this illustration.

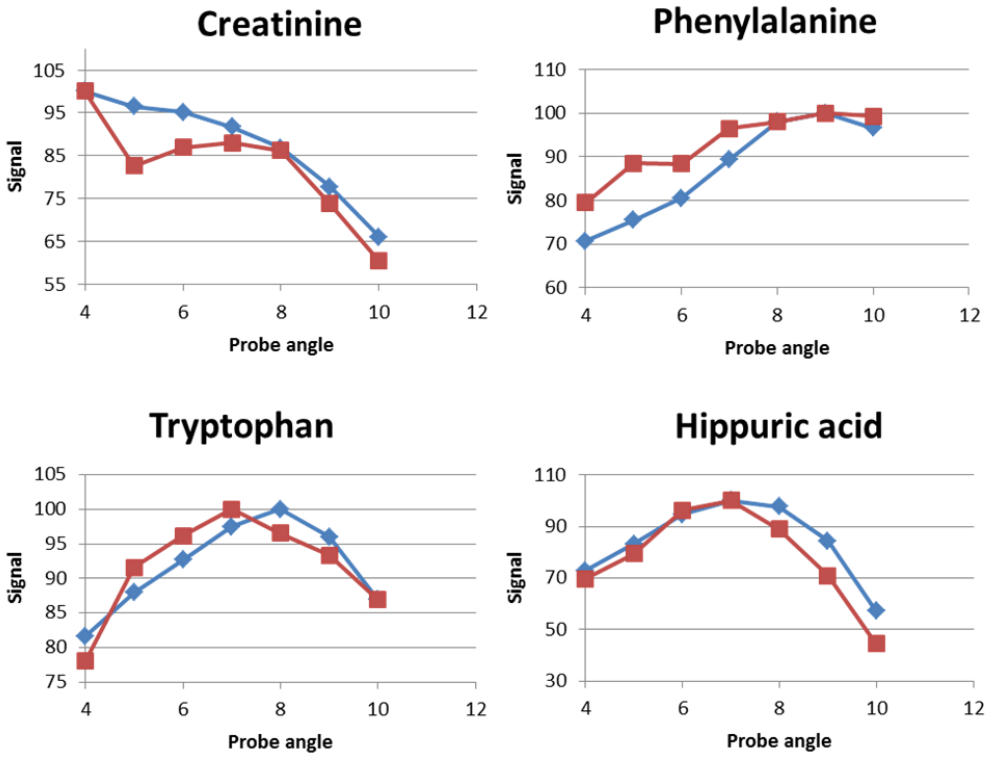


Figure 3-23: Normalised observed signal (integrated peak area) of selected reference chemicals from the RPC SSTM analysed by RPC using electrospray ionisation with varying probe angle relative to the MS source cone. The experiment was repeated on two different Xevo G2-S Q-

ToF instruments (red and blue lines) to ensure the results observed were not unique to a single system.

3.5.3 Testing the limits of UPLC-MS system stability for an optimised configuration

The RPC chromatographic method was chosen to test the performance boundaries of the overall system including the ionisation interface and MS, and a test set of human urine samples was designed specifically for this purpose. The sample set was engineered from a set of individual urine voids acquired from four volunteers. Portions of each sample were used to generate six 1:1 (v/v) pairwise mixtures. A bulk pooled sample (equal parts of all 4 original urine samples) was generated to serve as a study reference (SR). Each sample type was diluted 1:1 with ultrapure water (Fisher Optima LC-MS grade) and homogenized prior to aliquoting to a 96 well plate (2 mL deep-well) for analysis, minimizing any potential difference among technical replicates of the same sample. The four original and six pair-mixed assay samples (10 in total) were respectively aliquoted to cells within columns 1-10 of the plate such that each column contained replicates of a biofluid of distinct but interrelated composition. The pooled sample was aliquoted to cells within columns 11 and 12, such that each plate row contained samples 1-10 and two pooled study reference samples. The sample combination scheme and plate layout are summarized in Table 3-13. The contents of the 2 mL deep well plate were then further sub-aliquoted to 9 analytical plates (350 µl per well), with 200 µl in each well.

Sample ID	96 well plate column #	Composition (in equal parts)
1	1	1
2	2	2
3	3	3
4	4	4
5	5	1+2
6	6	1+3
7	7	1+4
8	8	2+3
9	9	2+4
10	10	3+4
SR	11, 12	1+2+3+4

Table 3-13: Sample composition and distribution within the 96-well format.

3.5 Maximisation of analytical batch size through optimisation of the LC-MS platform.

UPLC-MS profiling was carried out using the RPC analytical method described above. The column (Waters Acquity HSS T3 2.1 x 150mm) was new and not conditioned as described previously (Want et al., 2010). The source parts of the Xevo G2-S Q-ToFMS including the electrospray probe, MS inlet cone (and cone guard) and source enclosure were newly cleaned. These measures were taken to allow observation of the total precision of the system from the first introduction of mobile phase and sample to the conclusion of the experiment. The 10 assay samples from a given plate row were injected in a randomized order, with the study reference samples in wells 11 and 12 injected after the first and second set of five assay samples, respectively. A study reference sample was thus assayed every 6th sample for the duration of the experiment. In the following analysis of the sample set, only the study reference samples are considered, providing a frequent and consistent reference for assessing method and system performance.

MS data were collected in continuum mode. Data conversion to continuum spectra and reformatting to NetCDF of all study reference samples were accomplished using the AutoAFAMM function of MassLynx 4.1 and Databridge executable function, respectively. All definable parameters for feature extraction using the centWave method within XCMS were set to values determined to be appropriate based on manual review of the raw data files, listed in Table 3-14. Peak integration was performed on the raw data rather than the fitted peaks (integrate=2) to reflect the original measurements as closely as possible.

parameter	value
ppm	30
peakwidth	1 to 8
snthresh	50
noise	300
prefilter	$x = 6$
	$y = 5000$

Table 3-14: XCMS parameters for the extraction of features from urine analysed by LC-MS using the optimised RPC method.

Density-based feature grouping was performed in XCMS using a bandwidth (smoothing kernel adjustment) of 1 and an m/z grouping window (mzwid) of 0.01. A lower threshold of absolute noise was used in peak detection for more thorough peak picking, balanced by the use of a simple noise filtration scheme by which groups containing features detected in less than 50% of the SR samples in a single plate were discarded. Feature intensities were not normalized, as doing so could potentially obscure the true variation yielded by the analytical system and method.

Analysis of the feature detection results was performed using R software. Figure 3-24 illustrates the comparison of the total number of features detected in each plate, revealing a surprising level of consistency among all but the first plate. The lack of a severe decline in the number of detected features over the duration of the analysis indicates that the UPLC-MS system has been optimized for maximum longitudinal performance.

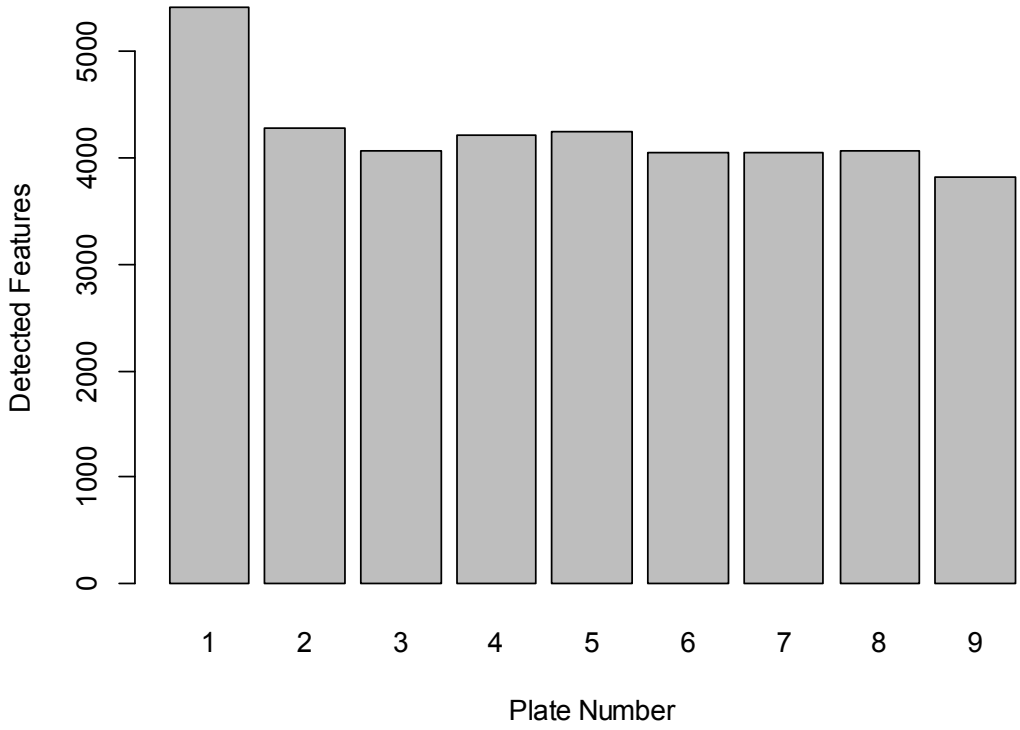


Figure 3-24: The total number of features detected in the 16 SR samples of each plate of 96 sample analyses, across 9 sequential plates in total.

Similarly, a surprising amount of precision was observed across all plates, with only the first plate having a median coefficient of variation (CV) across SR-derived features in excess of 15%. The median values from plates 3-9 were between 5 and 7% demonstrating the excellent overall precision achieved by the chromatographic method. The distribution of SR-derived feature intensity CV from each plate is collated in a box-and-whisker plot shown in Figure 3-25. The decline and stabilization of feature CV distribution across the first three plates reflects the conditioning of the UPLC-MS system.

3.5 Maximisation of analytical batch size through optimisation of the LC-MS platform.

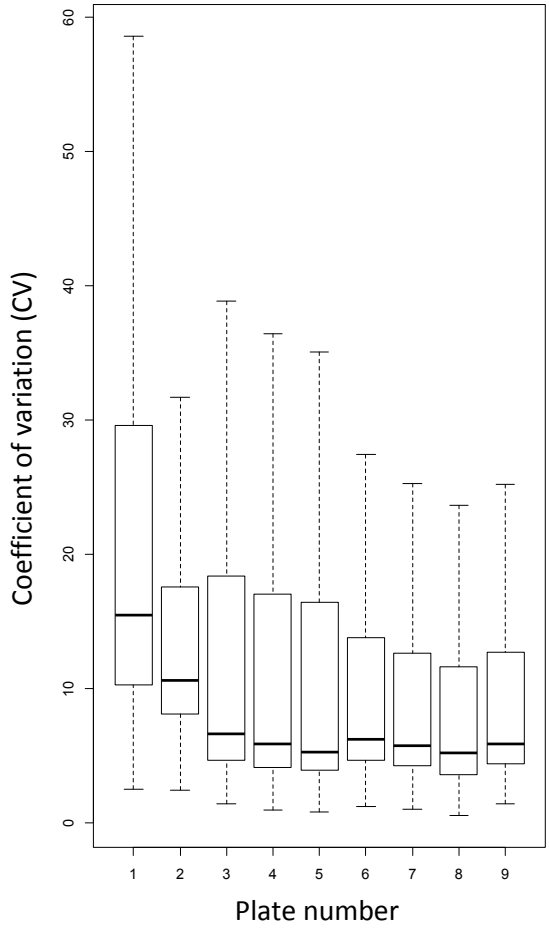


Figure 3-25: The plate-wise distribution of CV observed for featured extracted from each plate's 16 SR samples.

To investigate the data as a unified set, individual feature sets extracted from each plate were grouped together using the same density-based method and parameters as used for each individual plate. The minimum fraction filter was again used, this time excluding feature groups that contained features not found in at least 50% of the SR samples present in any one plate.

Additional noise filtration was implemented by utilizing the SR dilution series appended to the end of the sample analysis. The Spearman's rank correlation coefficient was calculated for each feature in the dilution series, correlating the linear dilution to the intensity response of the average value of three replicate measurements for each of 5 dilutions. Features with a correlation of 0.8 or more were retained for further analysis, yielding a final of 4391 features from 6560 originally detected.

3.5 Maximisation of analytical batch size through optimisation of the LC-MS platform.

Principal Components Analysis (PCA) of the un-normalized SR sample dataset was conducted using the SIMCA-P+ v. 13.0.2 software (Umetrics, Umeå Sweden). Unit variance scaling was applied to the dataset, and a scores plot was generated to illustrate the majority of variance within the dataset as illustrated in Figure 3-26.

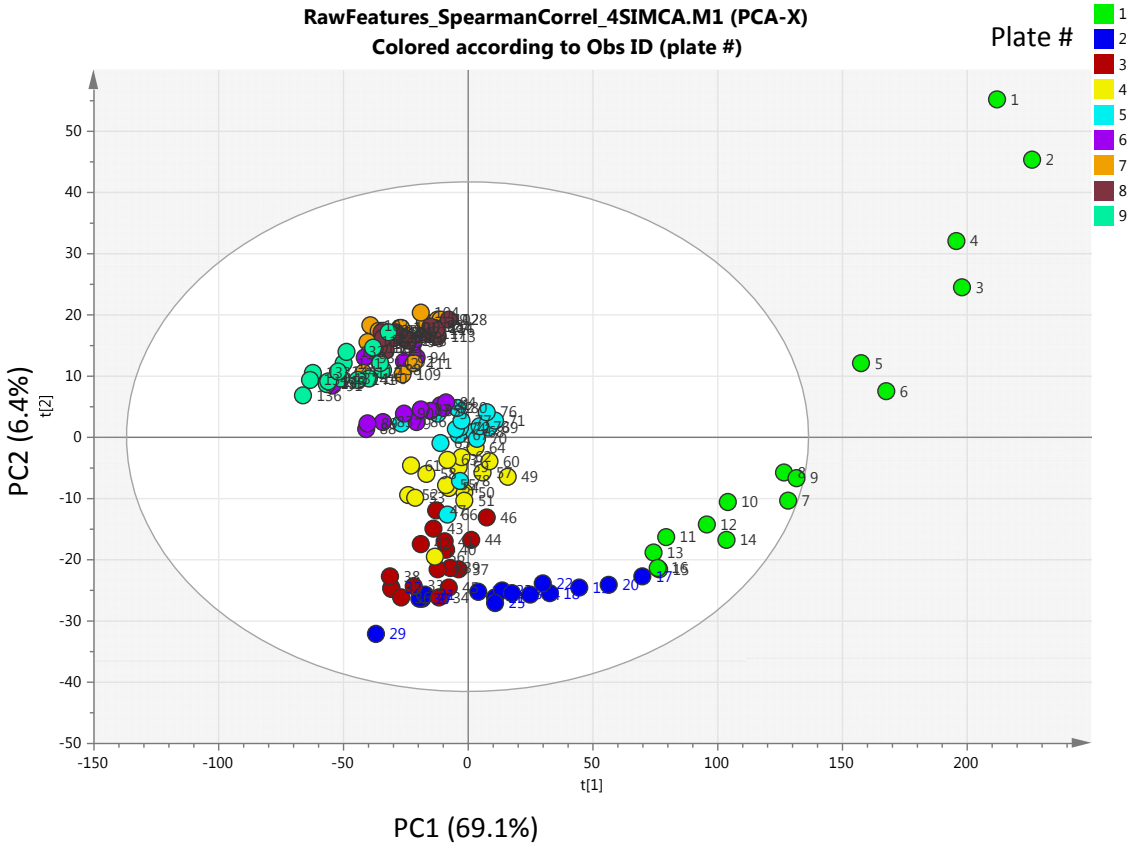


Figure 3-26: PCA scores plot showing the distribution of SR samples (coloured by plate number) across principal components 1 and 2, accounting for 69.1% and 6.4% of the total dataset variance, respectively.

The scatter of SR samples from plates one and two across PC1 (accounting for 69.1% of the total dataset variance) reflects the conditioning required by the LC-MS system to achieve a state of equilibrium. The 16 SR samples from the first plate represent 96 total sample injections, suggesting a conditioning period that is far greater than previously reported. However it must be noted that the analysis was started without any preliminary washing of the column or equilibration at initial conditions, potentially

contributing to this difference. The low amount of variance explained by PC2 (6.4%) and all subsequent PC's indicates a high degree of homogeneity within the dataset once equilibrium is achieved.

Detailed examination of the features most responsible for the distribution across PC1 reveal an underlying combination of conditioning effects. First, many chemical species demonstrate a higher loss of sensitivity across the first 2 plates than is observed in the remaining seven plates. The intensity of an exemplary feature (m/z = 460.285, RT = 3.8 min) as detected across all SR samples is illustrated in Figure 3-27A (top). This effect is attributed to loss of instrument sensitivity due to initial soiling of the source with sample residue, and/or initial conditioning of the MS detector. The second (and less prevalent) intensity behaviour responsible for sample scatter in PC1 is a fast rise in signal intensity from near zero (baseline noise) in plate one to a steady value in plates three to nine. The intensity of a representative feature (m/z = 358.259, RT = 6.4 min) as detected across all SR samples is illustrated in Figure 3-27B (bottom). This effect is attributed to the detection of peaks found within plate three to nine analyses that are not present at the expected retention time in earlier plates due to retention time migration. As the chromatographic peaks migrate into the area of integration defined by the consistent position in SR analyses from later plates, the integrated intensity increases to eventually reach the maximum value representing the entire peak.

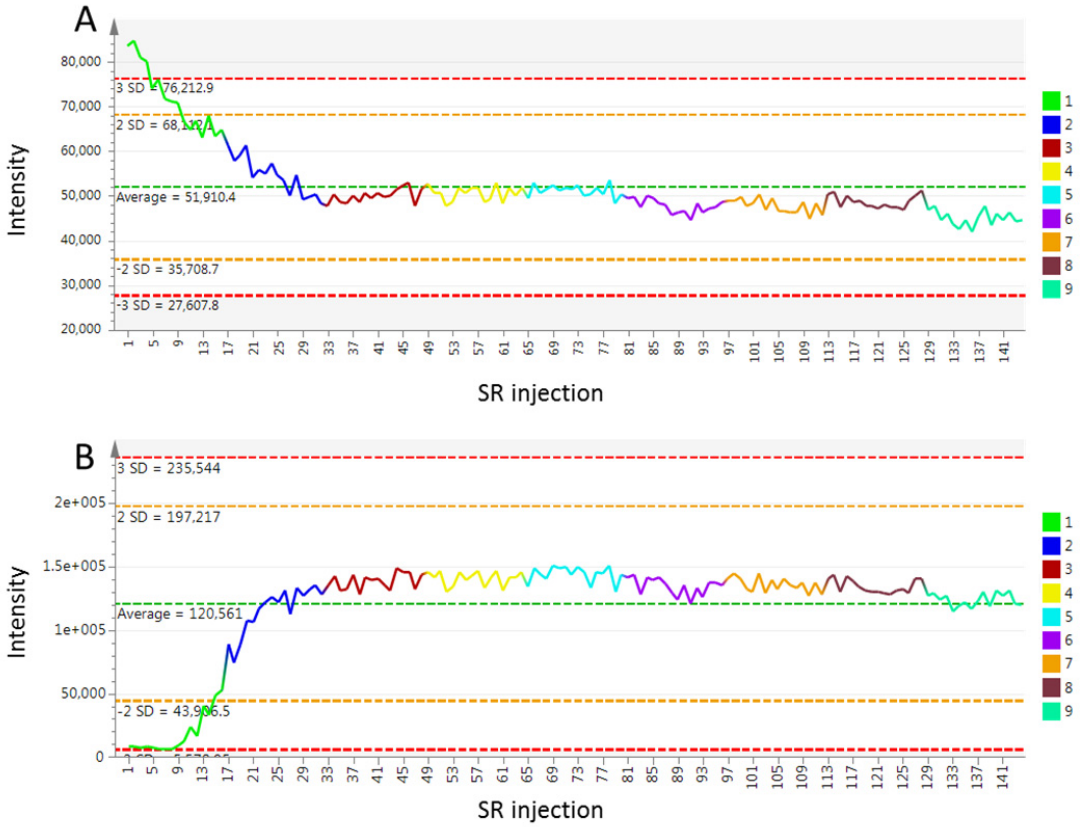


Figure 3-27. Variable line plots illustrating the intensity of two selected features as measured across all SR samples in each of 9 plates, shown in ascending order (left to right). These features were selected because they are exemplary of the two patterns of feature intensity observed to be responsible for scatter of samples across PC1, and each represent a unique manner of instrument conditioning as described in the text. The intensity pattern observed in panel A (top) is a consequence of MS conditioning, whereas the pattern observed in panel B (bottom) is a consequence of chromatographic conditioning.

To explicitly illustrate that this effect is due to feature migration, chromatograms of the feature plotted in Figure 3-27B were extracted from the first SR samples analysed in plates one through four. The results, shown in Figure 3-28, indicate that the peak is present in the earlier analyses (bottom), but eluting later than the 6.4 minute retention time that characterizes the peak group. Therefore, the baseline area is integrated, giving rise to the pattern observed in Figure 3-27B. Taken together, these results indicate that achieving system equilibrium requires independent conditioning of the LC and MS analytical subunits.

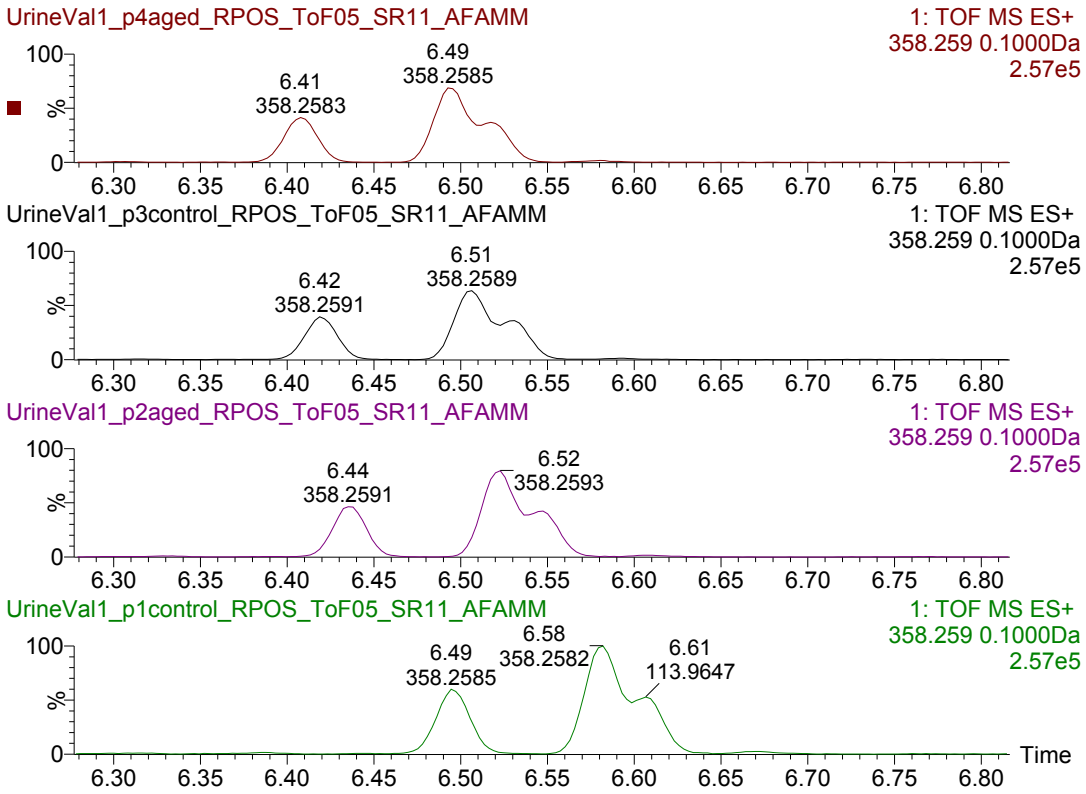


Figure 3-28. EICs of feature m/z = 358.259 @ 6.4 min from the first SR samples analysed in plates 1 through 4 (bottom to top). The peak group is defined at 6.4 minutes, representing the equilibrium position of the peak (observed here in plate 4, top). However, as the peak's initial retention time was 6.49 minutes in the first SR analysis in plate 1, its intensity was not recorded accurately as belonging to the 6.4 minute peak group.

3.6 Optimisation of sample preparation batch size.

Although the methodology and instrument platform have been shown to be stable for nearly 1000 sequential analyses following a period of initial conditioning, it is doubtful that a prepared complex biofluid sample would be stable for an equivalent amount of time. While it is true that samples can be processed to be compatible with analytical technologies and methods (eg. via derivatisation to enhance stability over time or amenability to a particular type of chromatographic separation), the molecular profiling approach tends towards the reverse implementation, instead utilising methods that are robust to minimally prepared samples. LC-MS is inherently well suited to the analysis of aqueous biofluids such as urine which, under normal healthy conditions, is sterile and free from excessive protein which is not compatible with LC-MS solvents. Small debris including any cellular material from the urinary

tract as well as cryoprecipitate and urinary sediment is easily removed from the sample by centrifugation or filtration, rendering the sample nominally ready for LC-MS analysis.

The methods utilised above require only dilution and centrifugation of a human urine sample, with water for RPC and acetonitrile for HILIC, making for a convenient and efficient workflow. The minimal preparation approach reduces or precludes the introduction of error and unwanted selectivity, facilitating robustness and benefitting both the research laboratory (in terms of time and cost) and the quality of data produced. The final consideration for achieving large-scale analysis is therefore the impact of sample age and molecular stability on the coverage achieved, warranting detailed consideration of the sample stability, preparation batch size, and frequency of sample preparation.

3.6.1 Assessment of urine stability

In the absence of automated online sample preparation, samples are thawed and prepared in batches which are stored in an autosampler (usually operated at reduced temperature) for an interim period during the batch analysis. This elapsed time is defined herein as the sample age, as opposed to the elapsed time between sample collection and analysis which is rarely within the analyst's control for human-derived specimen. As the prepared sample ages, components of its molecular content may undergo chemical cross reaction, reaction upon prolonged exposure to air, precipitation, or selective sequestering by adsorption to the sample container. These changes, broadly characterised as molecular instability, may modulate observed signals and potentially confound both qualitative detection (number of molecular species detected) and quantification.

As a sub-aim of the previously described experiment, the effect of sample ageing on the stability of the molecular content was investigated. All nine plates were frozen at -80° C and thawed once between preparation and analysis, but five of the nine were thawed immediately prior to analysis, while the remaining four were thawed at the start of the analysis despite not being needed immediately. The latter group were aged at 4° C until they were analysed in an alternating order with control (freshly thawed) plates. Interleaving of plates was necessary to deconvolute the effects of sample age and run

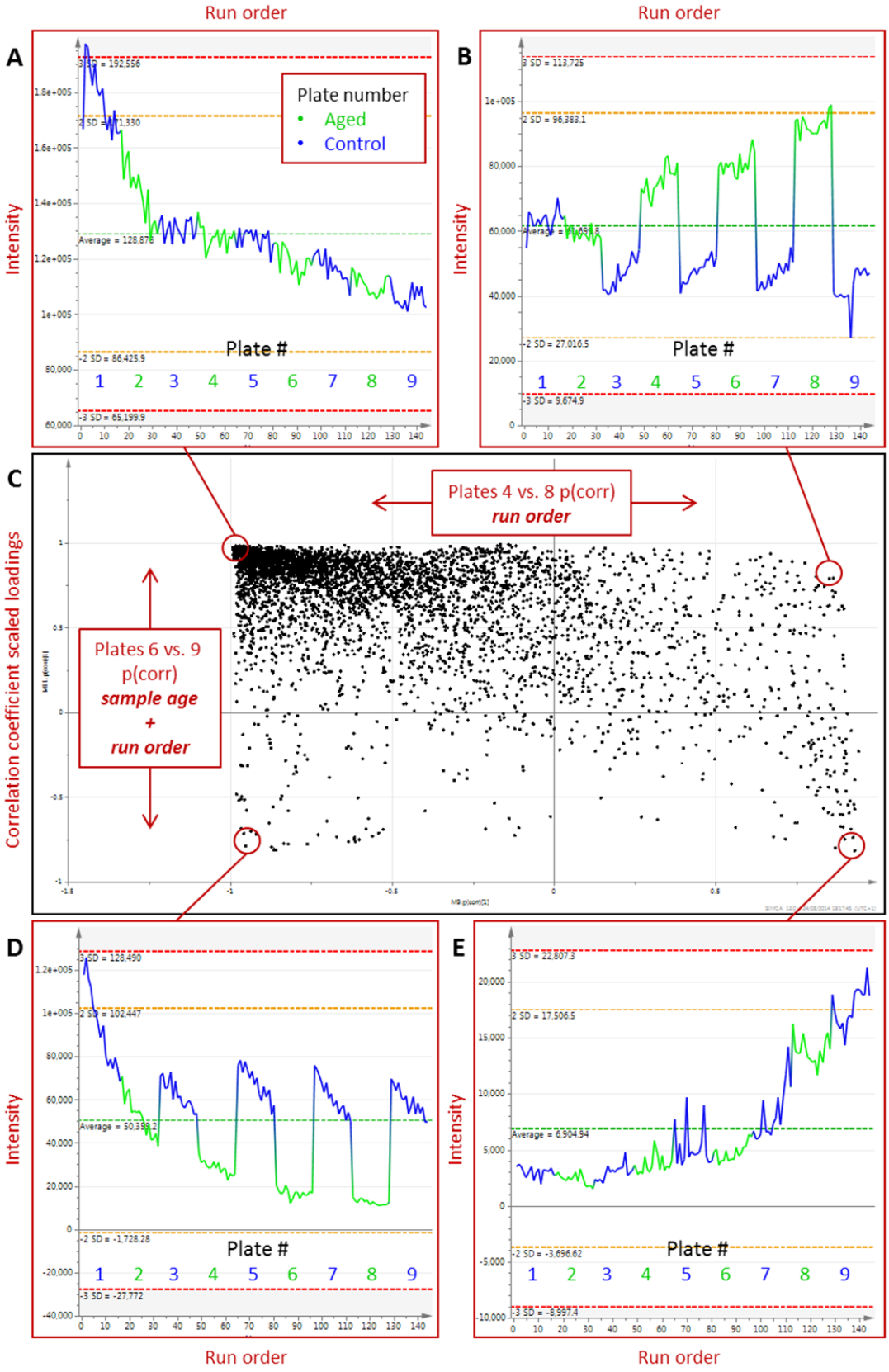
order, each of which was expected to potentially cause a decrease in the observed intensity of a given molecular species.

No obvious differentiation between control and aged plates was observed in PCA of the dataset, or in the number and precision of metabolites detected, indicating that the instability of the matrix overall is minor in comparison to the analytical variance. In order to elucidate the differences in metabolite composition between aged and control plates, OPLS-DA analysis was conducted on plates six (aged) and nine (control). To minimise the confounding effect of run order on the gain or loss of molecular species, the correlation coefficient scaled loadings (p(corr)) loadings from the OPLS-DA model for plates six vs. nine were plotted against the same loadings from an OPLS-DA of plates four vs. eight (both aged). The R2Y and Q2 values obtained with a single calculated component were 0.72 and 0.70 for the plate six vs. plate nine OPLS model, and 0.84 and 0.80 for the four vs. eight OPLS model, indicating that the discriminant analyses are valid. Features oriented in the resulting plot (illustrated in Figure 3-29, panel C) correlate either negatively or positively with run order (panels A and E, respectively) or negatively or positively with sample age (panels D and B, respectively). In this illustration it is easily discerned that the majority of features detected associate with a slight decrease in intensity with relation to analysis order, while very few associate with differences due to sample storage and age alone.

(See figure on next page)

Figure 3-29: Illustration of feature behaviour related to run order and sample age. Features that decrease and increase with respect to run order are shown in the upper left and lower right areas of the loadings scatter plot (C). Features that decrease and increase with respect to sample age are shown in the lower left and upper right areas of the loadings scatter plot. For extreme features of each behaviour, the intensity across all SR samples of is plotted in panels A, B, D, and E.

3.6 Optimisation of sample preparation batch size.



3.6 Optimisation of sample preparation batch size.

Given that the vast majority of analytes were robust to ageing effects, it is not necessary to make great compromises to the ease and efficiency of sample preparation to achieve greater molecular stability. However, as profiling studies are often utilised for biomarker discovery and thus complete knowledge of the content of a given sample is never assumed, limiting the sample age to the greatest extent practical is a prudent measure which should be considered in the development of a workflow for large-scale analysis.

3.6.2 Simulation of analytical cycles in a large profiling experiment conducted within a model working environment

In the course of this analysis, it was noted that in order to ensure control plates were thawed immediately prior to their scheduled analysis, manual removal from -80° C storage and loading into the sample manager were necessary at odd and inconvenient working hours. While this was conducted for the experiment described above, it was found to be unsustainable for routine application. To address this, plates of samples could have been analysed discontinuously, loading each subsequent plate as early as possible the next day after a completed analysis, however this approach would have introduced batch effects and reduced the number of analyses achievable, eroding platform precision and efficiency and countering the aims of the development in this chapter. Alternatively, sample plates could have been thawed and loaded in reasonable advance of when they were required (eg. at the end of the working day), but doing so would have compromised the accuracy of the aging results. Therefore, maintenance of a practical working environment necessary for industrialised analysis of large sample cohorts and standardisation of sample age appear to be at odds with one another. To address this, an optimal solution was sought to allow for a standard maximum sample age, limited to the greatest extent possible within a practical working environment. Doing so is the last step to achieving large-scale analysis, as well as ensuring quality results in data generation.

In order to determine practical bounds for the preparation and loading of samples, the constraints of the working day must be defined. The calculations in this section assume that the sample preparation and analytical systems are not fully integrated and automated, requiring intervention by an operator during or between preparation and analysis (ie. racking prepared plates in their pre-analysis storage

compartment). Defining an example laboratory work schedule requires adherence to generally adopted practices if the model is to be widely applicable. The following assumptions regarding sample preparation and the working day are estimated to be commensurate with the work required and in line with common practices.

- The working day starts at 9:00.
- Sample preparation requires three hours, making 12:00 the earliest time at which freshly prepared samples may be added to the queue for analysis.
- The working day ends at 18:00, after which no more samples may be added to the queue until the following day.

This definition of the working day is illustrated in Figure 3-30.

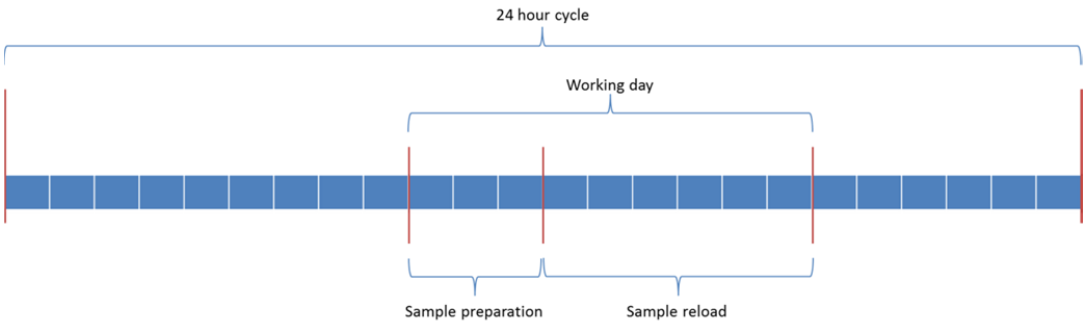


Figure 3-30: A graphical illustration of a 24 hour period (00:00 to 24:00) annotated with the assumed components of the working day. Each hour is represented by a single blue square. The working day is defined as 9:00 to 18:00, with three hours set aside for sample preparation allowing the remaining 6 for loading samples to the analytical instrument.

In order to calculate optimal schedules of sample preparation, both the demand of the system and the preparation batch size must also be known. To this end, the size of a batch has been standardised via the use of 96-well plates as reasoned previously, leaving only the rate of sample analysis (the amount of time elapsed between injections during continuous analysis, herein defined as the "cycle time") as a variable in order to seek periodicity while limiting sample age.

3.6 Optimisation of sample preparation batch size.

A script was written and implemented in the R software environment to establish, for a range of cycle times, the minimal batch duration that allows for continuous sample analysis in compliance with the laboratory working environment previously defined. The script was extended to simulate, for all analytical periods tested, the continuous analysis of twenty 96-well plates (1920 individual samples in total) and to calculate the maximum requisite sample age. Twenty plates was chosen as an optimistic upper limit for the number of plates comprising an analytical batch, limited only by the finite lifetime of a chromatographic column (commonly assumed to be approximately 2000 injections). The script may be found in Appendix 2.

Utilising this script, a simulation was performed for analytical method durations between 2 and 32 minutes (in steps of 0.05 minutes) to determine the number of plates per batch required to produce a batch analysis duration spanning from the start of analysis (set to 15:00 which represents the middle of the sample reload time) to or beyond the following day's minimum reload time of 12:00. The resulting plate-per-batch values and total batch durations are illustrated in Figure 3-31. Where the analytical method duration is equal to or greater than 13.15 minutes, analysis of the batch completes after the minimum reload time (12:00) at which additional samples can be loaded to the instrument, ensuring continuity. Where the analytical method duration is less than 13.15 minutes, analysis of a single plate would end before additional samples can be prepared and appended, breaking the continuity, and therefore at least one additional plate must be included in the sample preparation batch to extend the batch duration and regain continuity.

3.6 Optimisation of sample preparation batch size.

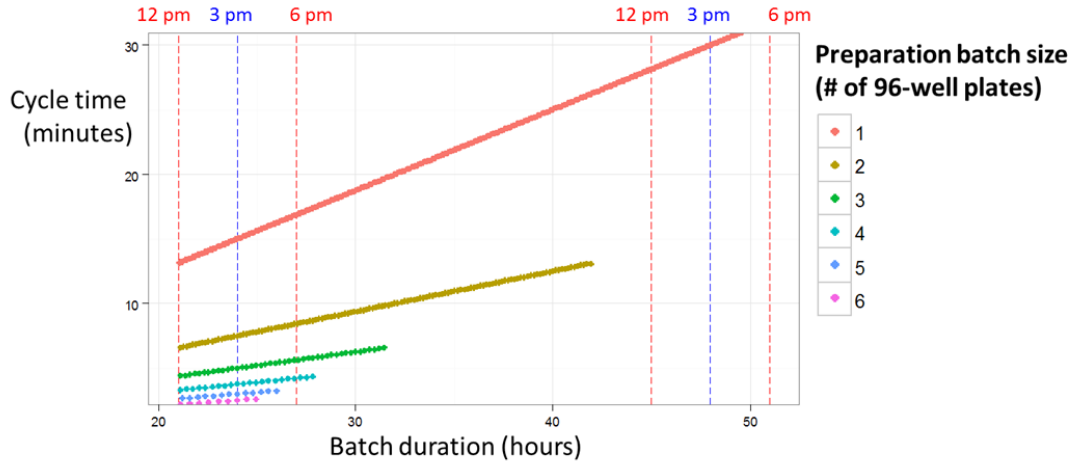


Figure 3-31: Illustration of sample batch reload times, assuming analysis is initiated at 3pm for analytical cycle times between 2 and 32 minutes. Red and blue dashed lines indicate important time intervals of the next two following days. Where a reload is required before another batch can be prepared and submitted (12pm the next day, noted as the first dashed red line), an additional plate is prepared within the batch to extend the duration so that another plate may be appended the following day. The number of plates per batch is denoted by the dot colour.

Figure 3-31 illustrates the substantial impact of additional plates per batch on the batch duration. As the batch duration increases, so does the maximum sample age, and in this manner, small changes in analytical method duration that push reload times past working day thresholds can potentially cause large changes in the maximum age of a sample between preparation and analysis.

Utilising the same script, a simulation of continuous analysis of twenty 96-well sample plates was performed to determine the maximum age experienced by samples for each method duration tested (again, 2 to 32 minutes in steps of 0.05 minutes). The simulated experiment was set to initiate analysis in the middle of the sample reload day (15:00). On each day, the smallest number of plates were prepared that allowed the analysis duration to reach the minimum reload time (12:00) of the subsequent day. Where a new batch was scheduled to start after 12:00 but before the end of the working day (18:00), the new batch preparation was assumed to conclude exactly when the samples were needed for appending to the sequence, minimising sample age. Where a new batch was required to start between the end of the working day (18:00) and the earliest reload time of the following day

(12:00), the new batch preparation was assumed to conclude at the latest possible time (18:00), again minimising sample age.

The results of the simulation are illustrated in Figure 3-32, where the maximum sample age has been plotted for each method duration tested. Close inspection of these data shows that maximum sample age is minimised when batch duration cycle times are regular (or approximately regular). In such a scenario where plate batches are prepared once per day, an analytical method duration of approximately 15 minutes (1 plate of 96 samples per day) produces the lowest maximum sample age of any method duration above approximately 7.5 minutes (2 plates of 96 samples per day). The inability of other cycle times to establish a regular period eventually requires that one or more additional plates be appended to a batch in order to bridging what otherwise would be a gap in analysis, increasing the overall maximum sample age by over 24 hours in most cases.

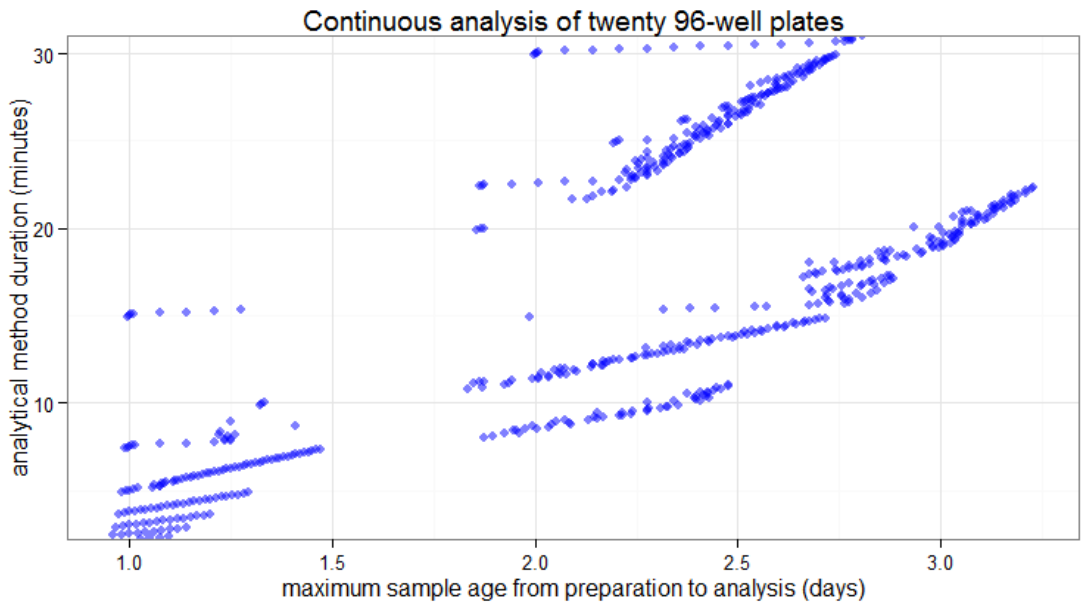


Figure 3-32: Simulation revealing the maximum age accumulated by a sample during the continuous analysis of 20 96-well plates for analytical cycle times between 2 and 32 minutes.

Additional regular cycles are possible as indicated in Figure 31. For example, a 30 minute analysis time yields a repeating 48 hour cycle per plate (96 analyses * 0.5 hours per analysis = 48 hours). A 10 minute

analysis also fits in a 48 hour cycle, with 3 plates being analysable in 48 hours (3*96*(1/6) hours = 48 hours). Very fast methods of 7.5 and 5 minutes achieve 2 and 3 plates (respectively) with regularity every 24 hour period, offering substantial throughput. However, for the subsequent development in this chapter the method duration of 15 minutes was selected as an optimal value, allowing the maximal method duration to develop high performance results while minimising the greatest sample age per plate to 24 hours. The resulting 24-hour per-plate cycle provides the opportunity for daily intervention at regular intervals, and analysis time is easy to quantise and manage. Finally, the ease of instrument scheduling facilitates maximal utilisation, promoting whole laboratory efficiency. The theoretical maximum number of analyses possible in one year by a single LC-MS system is therefore approximately 35,000.

It is acknowledged that the calculation may be reversed in laboratory environments where stringently defined requirements on sample throughput require a set method duration. In those instances the batch size may be modulated to achieve the same outcome. However, doing so may require specialised preparation instrumentation, or may result in reduced laboratory efficiency.

3.7 Method finalisation

Determination of a 15 minute cycle time as optimal for large-scale continuous analysis required the lengthening of both developed chromatographic methods. The time required between analyses for the loading of each subsequent sample and initiation of data acquisition was calculated to be 0.35 minutes (21 seconds). The duration of the programmable chromatographic method was therefore standardised to 14.65 minutes, together creating a 15 minute analysis cycle. The additional time was allocated to key areas throughout each method. For the RPC method, the majority of the available time was allocated to lengthening the linear gradient duration, extending it by 22.2%. Lengthening the gradient separation while holding all other parameters constant (*e.g.* column length, flowrate, etc.) theoretically produces improved peak resolution and therefore peak capacity, although the gains are not proportional to the contribution of increased time-dependent band broadening from longitudinal diffusion. The wash at high organic concentration wash was also extended by 0.35 minutes (175%) to ensure complete

removal of hydrophobic any species encountered, pre-empting their accumulation on the column. The finalised method is outlined in detail in Table 14 (left).

The HILIC method was modified more substantially, primarily due to the extra analytical time available from the standardisation to 15 minutes, but secondarily due to the feasibility of doubling sample throughput by utilising a column switching approach with a single detector. Given that HILIC chromatography requires such extensive equilibration, it was recognised that the original method was nearly half cleaning and equilibration. If modified such that the gradient area rich in metabolic content was half (or slightly less) of the overall cycle time, two chromatographic systems could be mated to a single MS using a valve to switch between eluates, economically doubling throughput. Although such a system was not used in application, the method was consequently modified to render it eligible for column switching applications, perhaps as such systems become more commonplace commercially.

To accomplish this, the latest eluting metabolites intended for measurement should be cleared from the column in advance of approximately 7 minutes to allow a small amount of time for the hypothetical valve changeover and initiation of a new MS acquisition. The latest eluting peak shapes suffered from substantial peak tailing, however, making clean elution challenging. A feature of m/z = 170.093 was identified as the last eluting molecular species present in the urine LTR sample as illustrated in Figure 3-33. Investigation of the peak(s) from data-dependent targeted MS/MS generated on the LTR urine fractions produced in section 3.4.3 revealed fragmentation patterns consistent with 3-methylhistidine (the main peak apex near 5.25 minutes) and 1-methylhistidine (the shoulder peak at 6.5 minutes) when compared to those on file with the Human Metabolome Database (HMDB00479 and HMDB00001, respectively).

3.7 Method finalisation

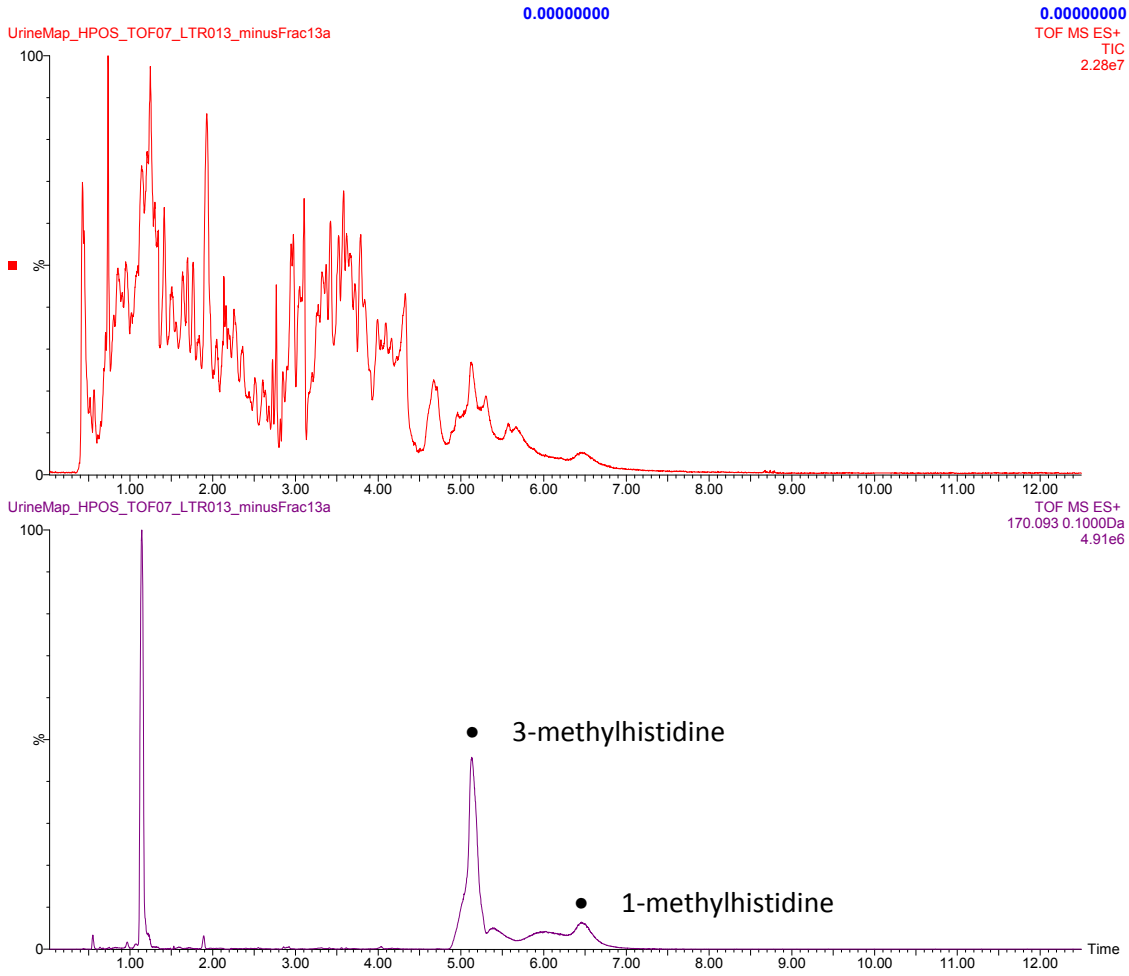


Figure 3-33: Within the HILIC analysis of LTR urine (TIC shown in red, top), m/z=170.093 (EIC shown in purple, bottom) is the latest eluting species observed.

The elution of broad chromatographic peaks after approximately 4.5 minutes was compressed by breaking the linear gradient into two independently linear portions; a shallower early portion and a steeper latter portion, simultaneously improving the separation of early eluting species while ensuring the clean elution of 1-methylhistidine in advance of 7 minutes. This had the desirable side effect of making the peak shapes more uniform across the entire elution. The remaining time was logically allocated to the equilibration step, completing the method modification. The finalised method is outlined in detail in Table 3-15 (right).

RPC			
Time (minutes)	Flow Rate	A (%)	B (%)
0.00	0.60	99	1
0.10	0.60	99	1
10.00	0.60	45	55
10.15	0.61	35	65
10.30	0.63	25	75
10.45	0.67	15	85
10.60	0.75	5	95
10.70	0.80	0	100
11.00	1.00	0	100
11.55	1.00	0	100
11.65	1.00	99	1
11.70	0.90	99	1
11.80	0.80	99	1
11.90	0.70	99	1
12.00	0.65	99	1
12.10	0.61	99	1
12.15	0.60	99	1
14.65	0.60	99	1

HILIC			
Time (minutes)	Flow Rate	A (%)	B (%)
0.00	0.60	5	95
0.10	0.60	5	95
4.60	0.60	20	80
5.50	0.60	50	50
7.00	0.60	50	50
7.10	0.605	95	5
7.20	0.61	5	95
7.30	0.62	5	95
7.40	0.65	5	95
7.50	0.70	5	95
7.60	0.80	5	95
7.70	0.90	5	95
7.80	1.00	5	95
12.50	1.00	5	95
13.50	0.60	5	95
14.65	0.60	5	95

Table 3-15: Finalised RPC (left) and HILIC (right) chromatographic methods after extending to 14.65 minutes (for a 15 minute injection-to-injection cycle with 0.35 minute inter-analysis delay) and distribution of excess time to key method steps.

3.8 Application to large-scale molecular profiling

The finalised RPC and HILIC methods were applied to a set of 2035 unique urine specimens in the course of ongoing work at the MRC-NIHR National Phenome Centre. All samples were prepared as previously described, by dilution with an equal volume of water followed by either centrifugation and analysis of the supernatant (RPC) or by further dilution with 3 volumes of acetonitrile to 1 volume of diluted urine, centrifugation, and analysis of the supernatant (HILIC). Batches of 80 urine samples were prepared together with 16 QC samples, together comprising a single 96 well plate of urine samples. Eight of the QC samples were aliquoted from a study reference (SR) urine pool generated by combining a small volume of all study samples in equal parts. The SR is therefore representative of the total study matrix. The other eight QC samples were aliquoted from freshly thawed LTR urine,

representing an external reference for the urine matrix. Sample plates were prepared daily except for those analysed on weekends which were, by necessity, prepared on the Friday prior to analysis.

Reversed-phase analysis was conducted on two Xevo G2-S Q-ToF instruments, one running in positive ion detection mode and the other in negative ion detection mode (herein referred to as RPC+ and RPC-). HILIC analysis was conducted on a single Xevo G2-S Q-ToF instrument running in positive ion detection mode (HILIC+). Alternating SR and LTR urine samples were analysed every 5 study samples. Prior to starting the experiment, an automated detector gain test was performed on each Q-ToF, whereby the signal obtained from a constant infusion of leucine enkephalin was assessed to determine the optimal voltage to be applied to the detector in order to generate approximately 90% of the possible signal intensity. This ensures that the signal obtained is nearly maximised, but does not risk applying too great a detector voltage whereby the detector would age more rapidly with no benefit to observed signal. The ToF mass analyser was calibrated with reference peaks generated by infusion of sodium formate solution. Finally, the system, complete with a new column (2.1 x 150mm HSS T3 and 2.1 x 150 mm BEH HILIC columns for RPC and HILIC chromatography, respectively) was conditioned with approximately one plate's worth of SR sample injections based on the data obtained in Section 3.5.3

The experiment of 2035 study samples (approximately 2400 injections, including the QC samples) was deliberately broken into two analytical batches near the midpoint of the analysis representing a slight extension in batch size based on the previous successful analysis of nine sample plates. The ionisation source was evaluated between analytical batches with the intention of cleaning all relevant components (e.g. the inlet cone and cone guard as well as the capillary and probe assembly). However, all components were observed to be in good condition, relatively unsoiled, and were therefore not serviced. Therefore, only the detector gain test was performed between batches. The batch 1 and batch 2 gain settings are listed in Table 3-16, indicating a substantial decrease in gain between batches in each analysis. No other intervention was performed. A single column per method was used for the duration of the experiment.

3.8 Application to large-scale molecular profiling

Instrument	Batch 1	Batch 2
ToF 1 (RPC+)	2602	2702
ToF 2 (RPC-)	2813	2988
ToF 3 (HILIC+)	2600	2750

Table 3-16: Detector gain voltages applied during the first and second batches of data acquisition for each instrument and method type.

3.8.1 Chromatographic precision

The chromatographic precision produced by all methods was observed to be excellent when used across thousands of sample injections. For the HILIC analyses, the variance observed in the retention times of peaks was minimal despite multiple preparations of mobile phase made and used throughout the analysis. The TIC traces from the first and last LTR urine analysis are illustrated in Figure 3-34 (top and bottom, respectively). This result is indicative of the success of both the method development and the efforts toward making solvent preparation more reproducible. In addition, no significant degradation in peak shape was observed, despite the use of only a single column for all 2400 analyses. The TIC traces from the first available and last LTR urine analysis are illustrated in Figures 3-35 and 3-36 (top and bottom, respectively) for RPC+ and RPC- analyses.

Unfortunately, both the RPC positive and negative mode acquisitions were plagued by corruption of data files that were later traced to a hardware incompatibility within the data acquisition computer. The result is an incomplete dataset for both of these analyses, necessitating a second round of analysis at a later date to recapture the corrupted files. For this reason, the largest continuous subset of data acquisition was selected for illustration of system and method performance.

3.8 Application to large-scale molecular profiling

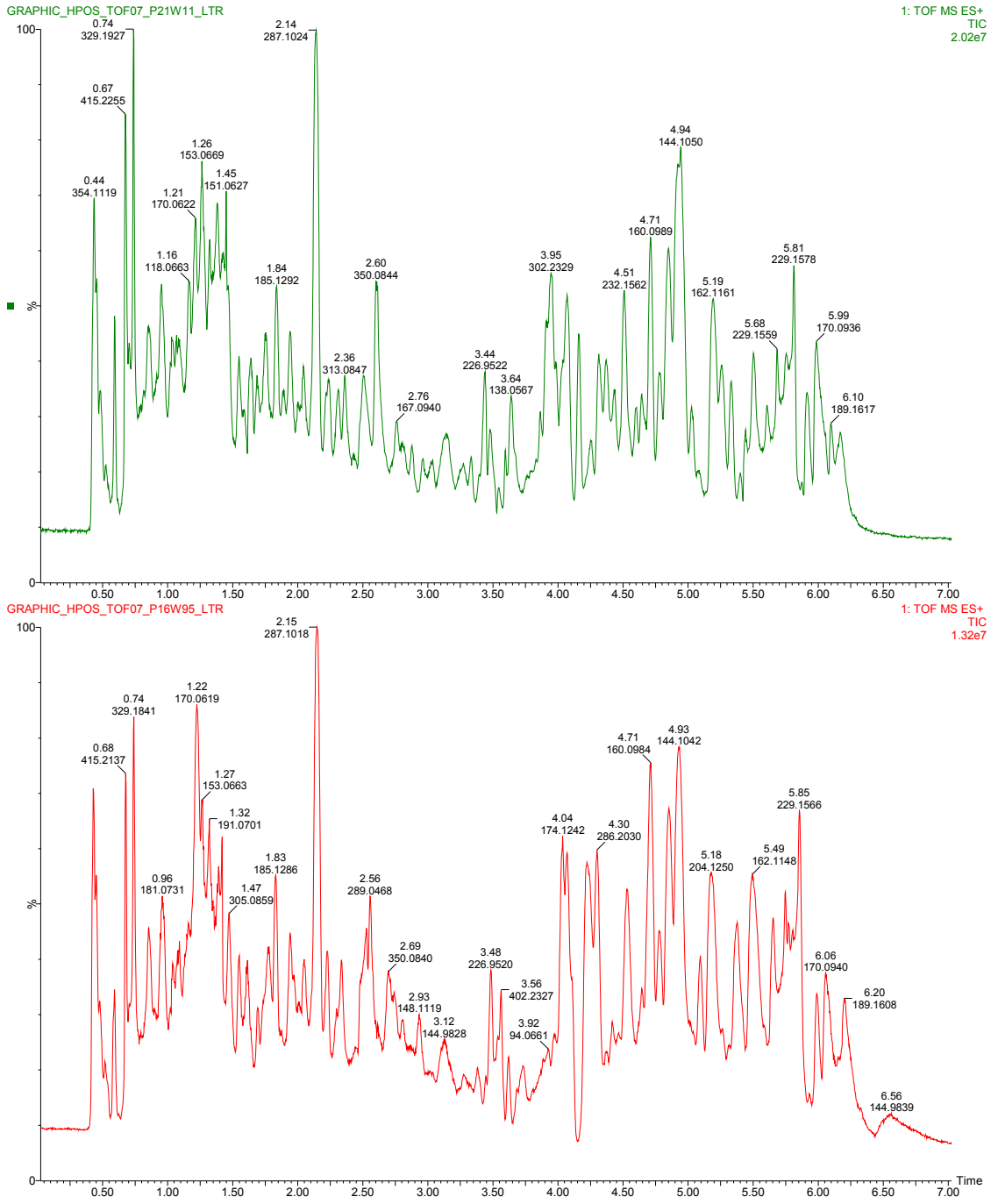


Figure 3-34: TIC traces from the first (top) and last (200th, bottom) HILIC+ urine LTR analyses in the set of 25 randomised 96-well plates, approximately 2400 injections apart, showing similarity.

3.8 Application to large-scale molecular profiling

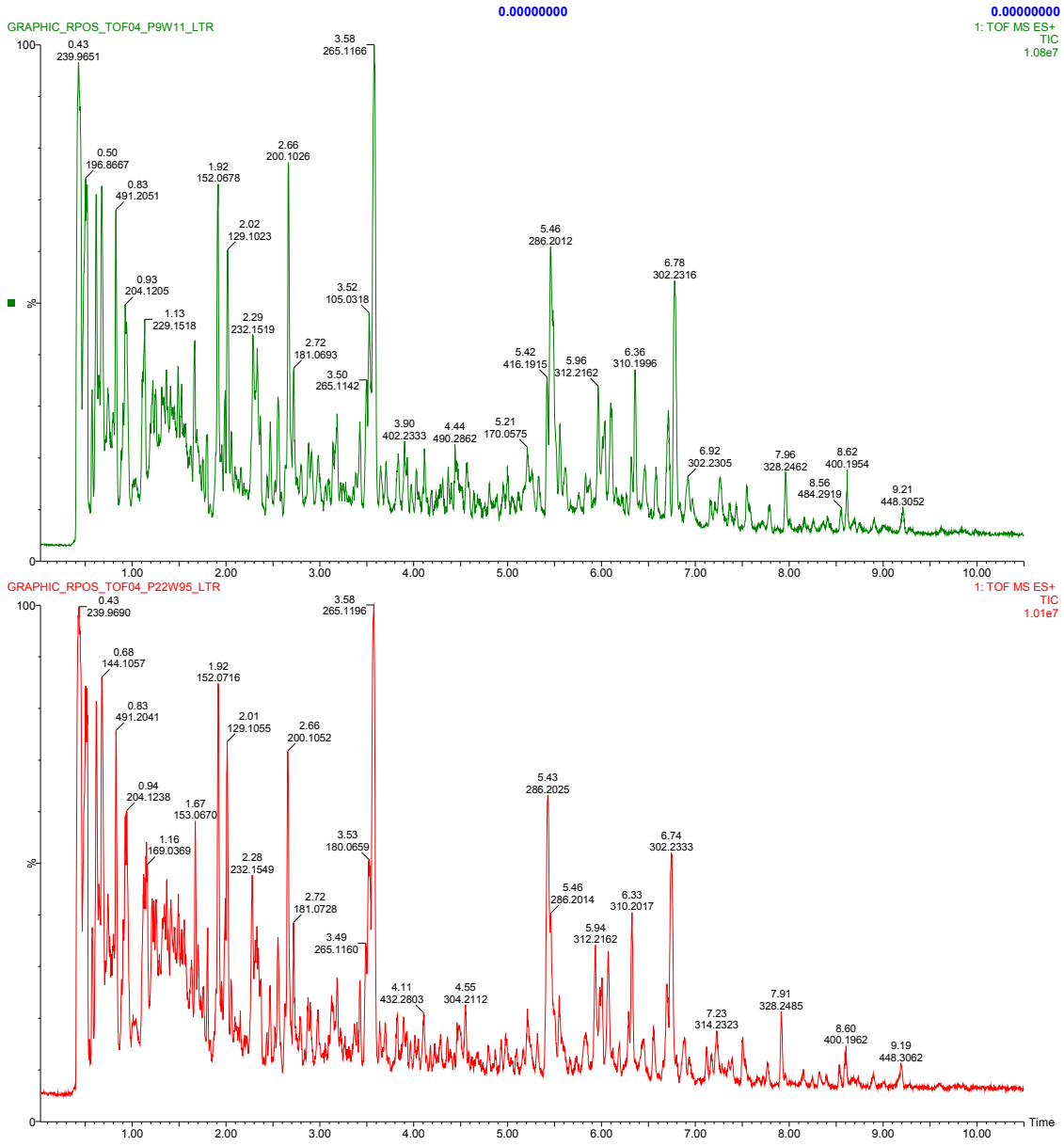


Figure 3-35: TIC traces from the 33rd (top) and last (200th, bottom) RPC+ urine LTR analyses in the set of 25 randomised 96-well plates, approximately 2016 injections apart, showing remarkable similarity. Analyses on plates 1-4 suffered from a small leak at the UPLC solvent mixer, and required reanalysis at a later date.

3.8 Application to large-scale molecular profiling

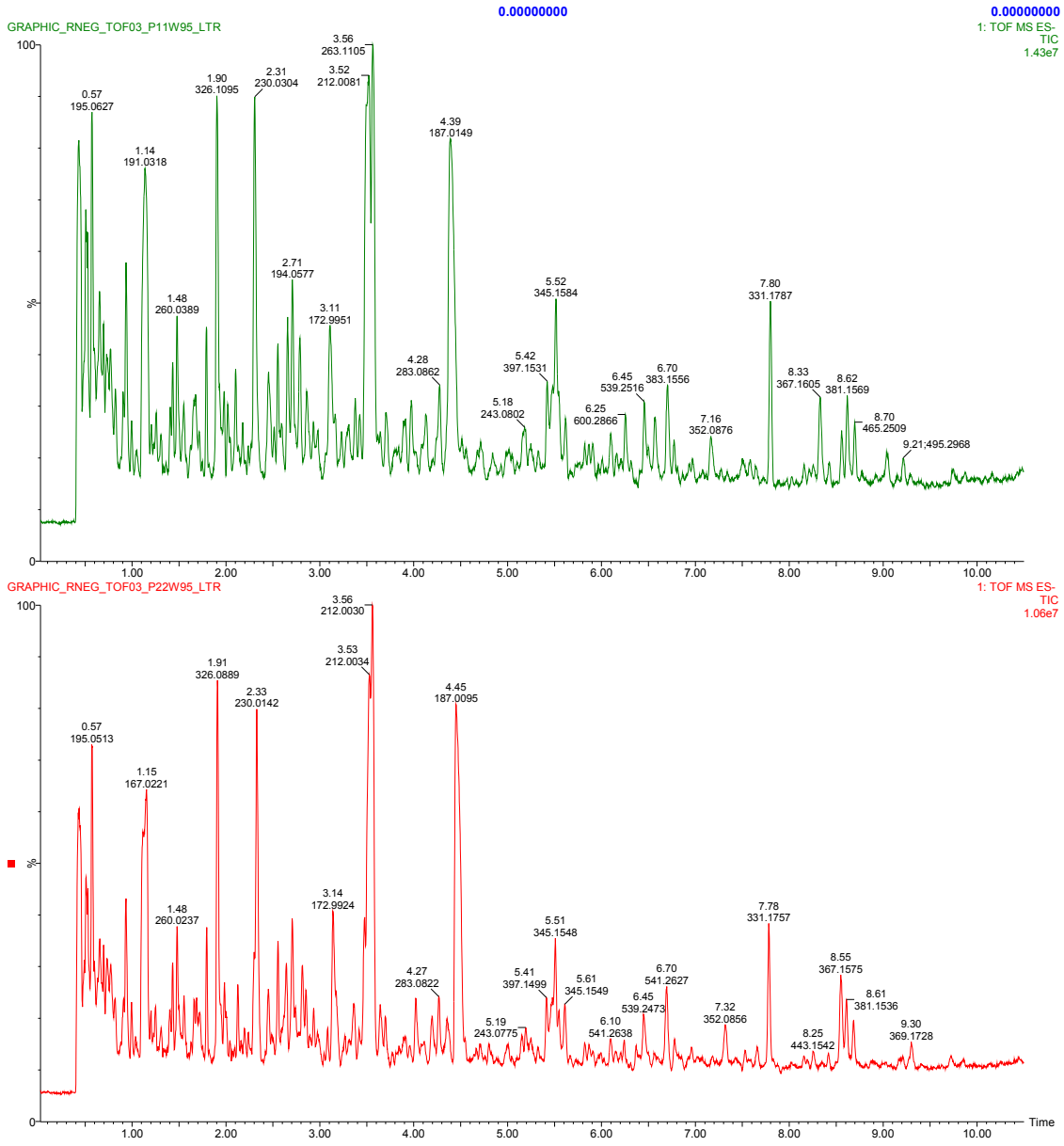


Figure 3-36: TIC traces from the 8th (top) and last (200th, bottom) RPC- urine LTR analyses in the set of 25 randomised 96-well plates, approximately 2300 injections apart, showing remarkable similarity. LTR analyses 1-7 on the first plate required reanalysis due to corrupted data files.

3.8.2 Intensity precision

The decline in observed signal intensity with respect to run order was greater in the application described here than observed in the nine-plate analysis described in Section 3.5. However, optimisation of the detector voltage at the mid-point of the analysis was observed to fully restore the signal, indicating the losses were rooted in loss of gain at the detector, and not due to contamination of the source. The exact reasons for the observed decline in detector performance within this experiment are not known, however it should be noted that the instrument manufacturer has introduced a function to assess and fine tune the voltage on the detector prior to each analysis in an effort to compensate for sensitivity lost throughout continuous analysis. Yet, initial testing of this automatic tuning function caused undesirable side effects such as the accumulation of LC eluent on the lockspray baffle which is used to block the analyte probe while gain analysis is performed using the lockspray probe. This accumulation occasionally led to the suction of liquid into the MS source when the automatic tuning step lasted for more than 30 seconds. The tuning step was also observed to last for a variable duration depending on the results and therefore could potentially cause variance in the amount of column equilibration experienced prior to each analysis. For these reasons, the function was determined to be unsuitable for use in the application described here.

As noted previously, a number of modern informatics solutions exist for the adjustment of signal loss observed across the analysis. Both the internal standards in the method reference (added to the LTR) as well as the full chromatographic data obtained from the LTR itself may be used as points of reference for such data correction. However, their application is beyond the scope of this chapter, as the aim is to generate high data precision from the analytical system itself, minimising the need for such informatics solutions. Therefore discussions are ongoing with the instrument manufacturer to design an improved method of detector voltage tuning that is more suitable for routine LC-MS analysis applications. Further work will be focused on this theme, creating software that is able to measure background chemical noise produced during chromatographic equilibration and utilise that to ensure the detector gain is adjusted prior to each injection, producing a consistent signal output and further contributing to the overall stability of the LC-MS platform.

3.9 Conclusions

The efforts presented within this chapter are not intended to produce chromatographic methods that are incrementally better performing than their predecessors, but rather to make methods that are fit for a purpose. Here, that purpose is the industrialised application of UPLC-MS profiling to large sample cohorts, which heavily weights the practical and operational aspects of analysis as well as the quality of the data produced. It is clear in retrospect, having learned from the process of development described in the preceding sections, that the most appropriate model for developing profiling methods for industrialised application suitable for population phenotyping is one where the practical constraints of the working laboratory are considered from the outset. These constraints then define the boundaries in which the analyst may subsequently work to optimise analytical performance.

In this case, the establishment of a regular 24h cycle for each batch of prepared samples allows for continuous analysis on a predictable, easy to manage, and efficient schedule. The cycle simultaneously limits the maximum age of a sample to an extent that is otherwise unachievable for sustained analyses within a common laboratory working environment as previously described (Figure 3-30). Limiting sample age increases the molecular coverage achievable by the assay as feature degradation is minimised during the analysis of each batch of prepared samples.

Of the method durations that achieve a 24 hour cycle (where 96-well plates are used as the fundamental unit of sample preparation) the longest (15 minutes) was chosen for the molecular profiling applications developed herein. This choice allows for a moderately high throughput (96 samples per day, per instrument) while leaving a suitable amount of time for the development of highly retentive, high capacity separations with adequate column cleaning and equilibration which are well matched to the performance capabilities of modern separations hardware. The cycle time calculations also indicate logical targets for future method development, with 7.5 and 5 minute analysis durations allowing for 2 and 3 plates to be analysed within the same 24h period.

Within the 15 minute time allotment, methods were created which generate high performance results with respect to overall peak capacity, peak distribution, and resolution of early eluting peaks, thereby

maximising the molecular coverage of each method. Longer column lengths were used to dramatically improve the resolution of early eluting species, particularly in RPC analysis where a large amount of urine metabolite content is early eluting. Variable mobile phase flow rates were used in order to increase the number of column volumes available for tasks of specific importance to each method. By selectively using 1mL/minute flow rates at high organic mobile phase concentrations, column cleaning was improved for the RPC separation, while column equilibration was improved for the HILIC separation. Furthermore, in an effort to combat the reputation of HILIC analysis for being unstable, the mobile phase preparation steps required by the analyst were simplified for the HILIC separation, relying more heavily instead on the UPLC device for accurate proportioning and mixing of solvents at very skewed distributions (e.g. 95:5 A:B).

These enhancements to the chromatographic methodology were supported by MS source and ToF tuning procedures designed to generate and trade sensitivity in exchange for precision during continuous analysis. By simply injecting less sample material (2µl of prepared dilute sample) and using source geometry that favoured the maintenance of clean source components over ultimate sensitivity, a high degree of chromatographic and intensity measurement precision was achieved in controlled testing. While that chromatographic precision was maintained in the large-scale application to population phenotyping (Section 3.8), the precision in measured feature intensity suffered specifically because of loss of gain at the ToF detector over the course of the analysis. This phenomenon appears to be under reported in the literature, perhaps because the effect is confused with a loss of sensitivity originating from the soiling of the ionisation source, or because analysis batch sizes are typically smaller than those achieved here (e.g (Zelena et al., 2009)). As a consequence, this observation lays the foundation for further work in the stabilisation of the ion detector system which is currently ongoing.

Notwithstanding this, sufficient precision has been demonstrated to allow for accurate comparative analysis across thousands of samples. While it is impractical to directly compare the long-term precision of the newly adapted methods vs. the reference methods at large-scale, their current implementation has been more successful at achieving measurement precision during large-scale continuous analysis than previous attempts with identical UPLC instrumentation utilising the RPC

reference method and similar bench top ToF instrumentation (Swann et al., 2013). In addition, the quality of LC-MS results can now be better tracked thanks to the development of QC reference materials such as the LTR urine and mixtures of chemical reference standards discussed previously. Both materials may be used for retrospective data correction (e.g. retention time alignment). While the LTR represents the complete matrix and may be used to bridge the complex profiles arising across many independent studies, the synthetic mixtures provide specific targets more suitable for prospective hardware suitability testing and real time assessment of method performance and quality control.

Finally, the complementarity of HILIC and RPC methods as configured herein has been successfully demonstrated in direct application to human urine, thereby ensuring the use of two chromatographic approaches for molecular profiling is warranted and efficient. While all technologies impose a degree of selectivity, the combination of these approaches provides broad coverage of diverse metabolite classes as illustrated by the excellent orthogonal peak distribution of the RPC and HILIC SSTM contents. The urine fractions generated are of potential future benefit, as the coordination of results from advanced analyses such as MS/MS, ion mobility, or NMR across the simplified urine fractions may assist in the rapid annotation of unknown compounds from subsequent studies.

Chapter 4: LC-MS feature grouping suitable for real-time application in large-scale profiling

4.1 Introduction

Achieving broad coverage in the measurement of chemical species is the crux of the metabolic profiling workflow. While the previous chapter focuses on expanding metabolite coverage in LC-MS data acquisition and improving the precision of measurement for application over large studies, this chapter's focus is on maintaining those qualities in the data processing steps which prepare a dataset for analysis. Before the acquired data can be collectively modelled or interrogated, the measurements from each individual sample must be collated to a single unified dataset. Such a dataset generally takes the form of a matrix of measured intensities for all observed chemical species (rows) across all samples (columns). To produce this matrix, each detected chemical species must first be differentiated from all other species present in the sample, as well as chemical and electronic noise, and then measured for intensity. This process is repeated for each sample in the study generating a series of independent datasets, each containing thousands of extracted chemical species representing the metabolic content of the sample. Identical species must then be grouped across all samples within the study to allow comparison of their measured intensities among individual samples or subset groups of samples as desired.

Individual chemical species are defined by their measured properties. The most fundamental of these in LC-MS datasets are the mass-to-charge (m/z) ratio and the chromatographic retention time. The use of high resolution mass spectrometric and chromatographic technologies (e.g. time of flight mass analysis and ultra-performance liquid chromatography) as well as high performance separations such as those described in the previous chapter greatly benefit the differentiation of features observed in complex biofluids. Yet, high resolution alone does not necessarily ensure accurate grouping of features across samples, as the independent quality of precision is also required to accurately group features across samples in an experiment. Feature grouping is therefore confounded by variance in the measurements used to assign identity to features.

4.1 Introduction

Both chromatographic and mass spectrometric instrumentation measurements are susceptible to systematic and unsystematic variation (Lange et al., 2008) referred to herein as measurement drift and error respectively. Error is the product of random deviation in measurement, whereas drift is a consequence of environmental and system changes over the course of the analysis, and is therefore time-dependent. Drift in mass measurement, for example, may occur when changes in air temperature surrounding the time of flight tube of a ToF instrument affect its length, causing a change in ion flight time and therefore the measured m/z ratio (Chernushevich et al., 2001). Fortunately, strategies that mitigate this effect have been routinely implemented in commercial mass spectrometric instrumentation for many years, such as hardware to allow selected sampling of a known reference mass for linear and global measurement adjustment (Wolff et al., 2001). However, no such strategies exist for the on-instrument correction of chromatographic retention drift which can be more dynamic, more selective, and non-linear in nature. As a consequence, the problem of drifting retention times is typically overcome post-acquisition using feature alignment and grouping software solutions.

The most relevant and widely used of the freely available software packages for feature grouping are discussed in detail in a subsequent section. However, all methods share two possible modes of failure when grouping features from multiple samples (Lange et al., 2008, Pluskal et al., 2010). The first mode of potential failure exists when features from identical chemicals across samples are not grouped together, likely resulting from analytical variance in the observed m/z and retention time measurements. The second potential failure is the erroneous grouping of features from distinct chemicals resulting from the inability of the grouping method to distinguish features with similar m/z and retention time characteristics. In both of these cases, feature grouping errors negatively impact upon the quality of the unified dataset. Furthermore, the potential occurrence of these grouping errors is exacerbated in high density datasets such as those produced by molecular profiling of human biofluids (Johnson et al., 2003, Lommen, 2009).

Fortunately, the resolution, scan to scan precision and overall accuracy of m/z measurements achieved by modern mass spectrometric instrumentation continues to improve as advances are made in commercial hardware. These benefits enable more stringent parameters to be set in the m/z dimension

when defining groups of features across samples, simplifying the exercise and improving the quality of the outcome. The performance of LC too has improved, specifically with the advent of commercial UPLC in 2004 making higher resolution separations possible in shorter times (Plumb et al., 2004). Furthermore, throughout Chapter 3, efforts were made to enhance the resolution (peak capacity) and precision of fast LC separations for application to large studies. The results of these efforts should likewise benefit feature grouping, producing more accurate experiment datasets. However the common tools in use for the extraction and grouping of LC-MS features were neither developed (Smith et al., 2006) nor benchmarked (Lange et al., 2008, Pluskal et al., 2010) using UPLC-MS data, and therefore may not be fit for purpose when applied to such datasets.

However, it is true that, when applied to large numbers of samples requiring days or weeks of acquisition time and potentially spanning multiple batches of chromatographic reagents, the cumulative analytical drift can become substantial across the duration of the experiment. Large profiling datasets are therefore especially at risk for loss of valuable molecular coverage and accurate feature intensity information. For these reasons, feature grouping remains a critical and delicate intermediate step between data collection from individual samples and analysis of the unified dataset. Fortunately, the chromatographic drift processes observed in modern UPLC applications are generally slow and non-Markovian, and are therefore embedded in the order of analysis (Eilers, 2004). Yet, no mainstream software package explicitly leverages the analysis order (also referred to as the "multisample advantage" by Tengstrand *et. al.* in their recent publication (Tengstrand et al., 2014)) when performing feature alignment and grouping. It is therefore a hypothesis within this chapter that the consideration of analysis run order in data processing will yield greater accuracy in feature grouping and produce datasets which better represent the breadth and quality of chemical measurement originally present in the raw data.

Finally, to support the holistic goal of high throughput analysis, the speed of feature extraction and grouping must be considered as well as the speed at which the data are generated. Under ideal circumstances a dataset for further analysis would be produced as soon as possible following data acquisition, allowing further dedicated investigation (e.g. targeted MS/MS or ion mobility

measurements) of features of interest elucidated by statistical modelling (e.g. regression of known metadata to the profiling dataset). However, recent experience with the feature extraction and grouping of profiling datasets in excess of 2000 samples has proven arduous, requiring an amount of time approximately equal to that required for data acquisition (with a high rate of complete failure as some software simply cannot achieve such large volume analysis). A conceptual fix would therefore be to allow the data analysis to run in parallel with data acquisition. Yet, while the advent of high throughput metabonomics has helped prioritise the efficiency and processing speed of feature alignment and grouping algorithms (Lange et al., 2008), none are so far implemented as the data are generated (herein nominally referred to as "real-time").

The opportunity therefore exists to develop and implement a fit-for-purpose method of feature matching that is appropriate for applications in large-scale profiling. Such a method should be adapted for the types of retention time variance observed in state of the art UPLC-MS analyses, including invulnerability to the time dependent drift produced by run order effect across large sample set. Furthermore the method should be suitable for implementation during data acquisition to support holistic high-throughput analysis. Such advancements are of fundamental importance and immense practical benefit to large profiling studies. The aim of this chapter is therefore to investigate the types of variation observed in large UPLC-MS datasets, and to leverage this knowledge in the creation and implementation of a novel iterative pairwise feature matching mechanism capable of realtime deployment and suitable for application to large-scale study.

4.2 Specific Objectives of Method Development

- Characterise the types of systematic retention migration observed in continuous profiling analyses.
- Develop and implement an algorithm that links features detected in subsequent samples which is appropriate for use in real time as data are generated.
- Demonstrate feasibility of the method in comparison to popular open source grouping methods using a synthetic dataset for performance assessment.

4.3 Existing tools and strategies for feature alignment and grouping

Prior to the development of advanced automated feature extraction methods, chromatographic data were preferentially aligned for peak grouping by warping of raw data rather than by detection and integration of peaks (Nielsen et al., 1998). While this technique, when applied to the TIC trace itself, was suitable for matching well resolved chromatographic features among samples, all techniques operating in a single dimension naturally struggle to cope with complex chromatograms and high peak density observed in human biofluid analysis (Christin et al., 2010). The development of Component Detection Algorithm (CODA) later allowed the extraction and selection of high quality single mass chromatograms, and COW was applied to each, leveraging the MS separation provided by LC-MS measurements (Christin et al., 2008). However, these and similar techniques perform alignment and grouping by warping the raw data (either the TIC or selected ion chromatograms), and are therefore capable of distorting chromatographic peak shape and introducing associated artefacts (Chae et al., 2008).

Advances in feature detection and extraction offer an orthogonal approach whereby features are distinguished within complex LC-MS datasets and integrated prior to alignment and grouping efforts. Grouping is then a matter of collation of a reduced dataset (de-noised, to the extent that feature extraction is selective for features of given intensity, signal to noise threshold, etc.) rather than manipulation of raw data (Robinson et al., 2007). This approach is commonly applied throughout modern LC-MS processing software packages which in turn provide a complete solution for feature extraction, integration, and multi-sample alignment and grouping. Of those that are freely available for use (not commercially produced and restricted to proprietary data formats and means of operation), the most widely used of these software packages determined by number of citations at the time this chapter was prepared (Coble and Fraga, 2014) include XCMS (Smith et al., 2006) and MZmine (Katajamaa et al., 2006), although new software packages continue to emerge (Tengstrand et al., 2014). The former two have also been shown to be the highest performing amongst their peers (Lange et al., 2008). Each of these established packages has been developed as a modular collection of individual procedures which perform distinct processes including feature detection and grouping. The feature

grouping approaches of these most pervasive approaches are reviewed here to provide a context for the subsequent developments within this chapter.

XCMS, the most highly cited of all related open source LC-MS data pre-processing software (Coble and Fraga, 2014), was first introduced in 2006 and has undergone various updates in implementation since that time (Tautenhahn et al., 2012, Gowda et al., 2014). One such update was the addition of the popular centWave peak detection algorithm, appropriate for high resolution MS data (Tautenhahn et al., 2008). The original grouping algorithm, still commonly used within Computational Systems Medicine because of its convenience and speed, considers the spatial density of extracted features of similar mass throughout the entire dataset (consisting of all samples), evaluating their distribution using a Gaussian-smoothed histogram assessment tool (kernel density estimator) to define boundaries of peak groups. While the method is indeed fast, the application of a feature-density based grouping scheme applied to high feature-density datasets (such as the fast profiling of human urine) creates a logical challenge. As the method considers all of the data at once, variance in peak position across the sample analyses contributes to a widening of the calculated density for a given feature which can sometimes lead to the erroneous grouping of two or more distinct features of similar mass and retention time within the same sample, where the analytical variance is on the same order as the distance between the peaks. This problem is compounded by difficulty in estimating the appropriate smoothing parameter (bandwidth) to apply within a given dataset, in turn determining the effective resolution at which feature density across the dataset is calculated. Despite the existence of published guidelines for general application for density grouping to UPLC-MS data (Patti et al., 2012a), this parameter remains unintuitive to adjust for a given separation, as it bears no direct relevance to the observable retention time variance (*i.e.* it is not a window of estimated retention time variance). Notwithstanding these difficulties, the grouping method is optionally augmented by iterative rounds of non-linear chromatographic alignment and re-grouping, however such alignment is limited to correcting for global shifts in chromatographic retention and is not feature specific.

The original implementation of MZmine (published a year earlier in 2005) utilises a feature alignment and grouping method referred to as the "join aligner" whereby detected features from a given sample

are matched to those in a master feature list by scoring their closeness in observed m/z and retention time (Katajamaa et al., 2006). With each new round of matching, the m/z and retention time values that represent each group in the master list are updated to reflect the average values from of all features within that group. User-set windows of m/z and retention time are centred on those master list average values, defining the new bounds for subsequent feature matching.

Despite this approach being released in advance of XCMS, it proved so popular that it served as the inspiration for an updated XCMS grouping method called "nearest" which works in the same manner but is augmented by XCMS's inbuilt optional non-linear chromatographic alignment option. MZmine, lacking this, recognised the join aligner method as not suitable for grouping peaks with non-linear deviation in chromatographic retention and released MZmine 2 in 2010 (Pluskal et al., 2010) adding the capability for non-linear correction of the chromatographic space between two samples. In this manner, the alignment and grouping methods of the two approaches have largely homogenised. This is convenient for the work presented herein, as we are able to use XCMS to test both the "density" and "nearest" methods (the latter representing the join-aligner method of MZmine) on a single featureset without breaking the embedded multi-step processing pipeline to feed the output of one method (*e.g.* feature detection) into another (*e.g.* grouping) (Robinson et al., 2007).

4.4 Definitions

In order to clearly present the intended workflow and pre-requisite procedures, a number of terms used throughout this chapter must first be defined. Those terms are listed here in an order that allows more complex terms to build on those previously defined:

- **Sample set:** A set of samples (either an entire experiment, or a batch) analysed in a continuous sequential manner.
- **Run order:** The order of samples sequentially analysed.
- **Feature:** A distinct detected signal with measured mass (m/z), retention time, and intensity (also characteristic peak shape parameters, if measured). Note that a single molecular species

will frequently be detected as multiple features due to the presence of isotopes, fragments, and adducts which are all separable by mass.

- **Feature set:** The complete set of all extracted features from a single raw data file (from a single sample).
- **Cluster:** A subset of features, extracted from a single raw data file (single sample), that are within close m/z and retention proximity. Clustered features are identified by matching the feature set against itself using 2x the window of retention time error used for matching features between samples.
- **Independent feature:** a feature extracted from a single raw data file (single sample) that is not in close m/z and retention proximity of other features in the set, and therefore not a member of a cluster.
- **Independent feature match:** A match between two independent features.
- **Community:** one or more cluster where at least one feature is matched to an independent feature or cluster of features in the paired feature set.
- **Link:** A pair of matched features from two samples in sequential run order.
- **Chain:** a continuous series of links across multiple samples of sequential run order
- **Ledger:** A matrix containing all features detected across all samples in the sample set. Each column represents a single sample, and contains all features extracted from that sample. Each row represents a unique feature within the full dataset, containing the individual entries, links, or chains of matched features.

4.5 Experimental LC-MS dataset

4.5.1 LC-MS data acquisition

In order to assess the variation in real LC-MS measurements, as well as provide an example dataset for the testing of multiple grouping strategies, a test set of human urine samples was designed and analysed as described previously in Section 3.5.3 (Testing the limits of UPLC-MS system stability for an optimised configuration). Feature extraction was performed on subsets of this dataset (specifically defined in the following sections) using the centWave method within XCMS and the controlling

parameters described in Table 4-1. The signal-to-noise threshold and prefilter intensity were set to lower values than those used throughout Chapter 3 in an effort to extract the maximum available feature information.

parameter	value
ppm	30
peakwidth	1 to 8
snthresh	10
noise	1000
prefilter	x = 8
	y = 2000

Table 4-1. XCMS parameters for the extraction of features from urine analyses by LC-MS using the optimised RPC method developed in Chapter 3.

4.5.2 Expected and observed patterns chromatographic retention deviation

The chromatographic retention of any given molecular species is subject to variance over the course of sequential analyses. While measurement error contributes to this variance, systematic drift within or among sample analyses is more generally problematic to downstream feature grouping. As much of the development in the previously discussed feature alignment and grouping methods has focused on the correction of nonlinear retention time warping in chromatographic space among samples, the model dataset was first explored for signs of such perturbations.

To accomplish this, all sample injections from the first plate of the test analysis were assessed for perceived nonlinear chromatographic drift using the retention time correction provided by XCMS. Using all of the samples rather than just the SR urine replicate injections ensured that any non-linearity due to differences in sample composition would be represented in the data subset. Features were grouped using the "nearest" grouping method with retention time and mass error windows of 6

seconds and 0.1 m/z respectively. Of those groups, 633 were defined as "well behaved" (each containing exactly 1 feature per sample) and therefore suitable for use in determining the global chromatographic retention time correction required for each sample to better align the feature content among samples and enhance subsequent grouping. These "de-warping" curves were generated for each sample by applying locally weighted polynomial regression (LOESS) across the retention time deviations observed in the features of the 633 correction groups. A span of 0.2 was used to control the degree of LOESS smoothing. The results of this non-linear alignment are shown in the output produced by XCMS, illustrated in Figure 4-1. The correction curve for each sample is shown as a single line, coloured (using a red-to-violet "rainbow" gradient) corresponding to the order in which they were analysed.

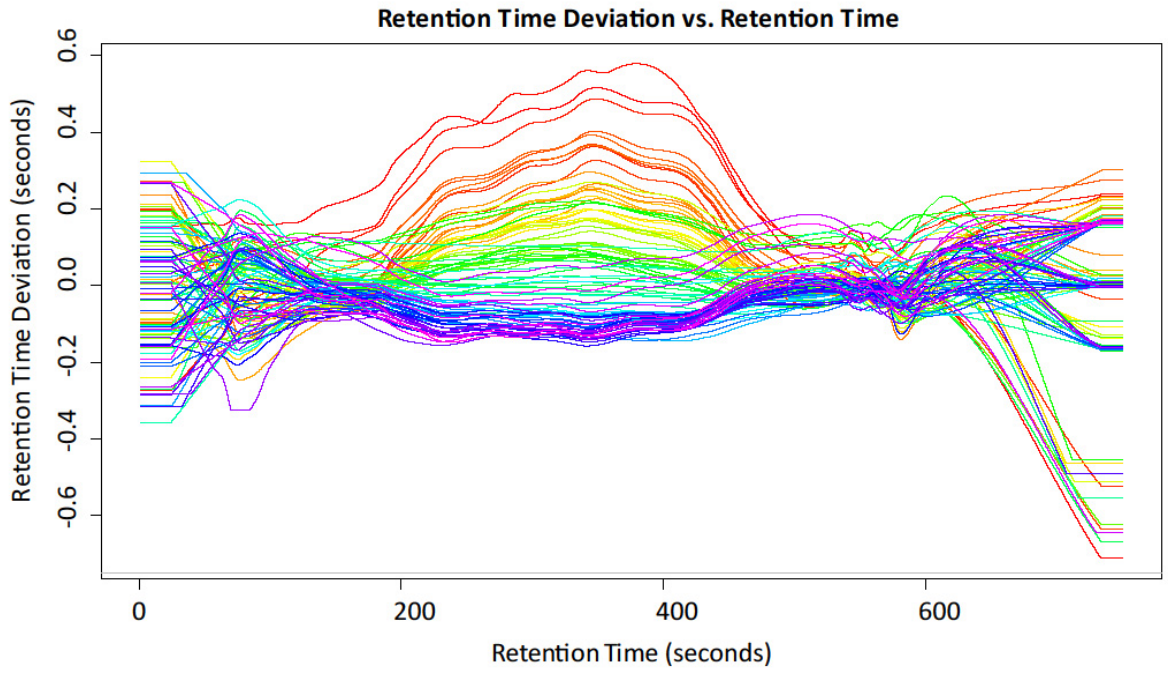


Figure 4-1. LOESS retention time deviation curves for 96 sequential urine analyses from plate 1 of the test sample set. The curve for each sample is coloured according to the order in which the sample was analysed, revealing a run order associated progression of features between 200 and 500 seconds in an otherwise stable chromatographic environment. The sample analysis order is coded in the rainbow-styled color of the correction curves, with red curves indicating earlier samples in run order and violet indicate later samples in run order.

4.5 Experimental LC-MS dataset

The LOESS smoothed retention time deviation curves show a slight tendency for features from well behaved peak groups between 200 and 500 seconds to migrate in a nonlinear manner with respect to the remainder of the gradient portion of the chromatogram (to approximately 600 seconds) which is otherwise very stable. The observed overall deviation is not substantial, however, and would be expected to be easily subsumed by a small retention time error window in feature grouping. For any individual features strongly contributing to this trend (perhaps with deviations in excess of that shown illustrated by the LOESS smoothed retention time correction curve), it is furthermore important to note that the deviation itself appears to be dynamic and correlated with analysis order. This indicates that even where features deviate in retention time among chromatograms, they do so in a gradual time-dependent manner.

With this in mind, the dataset was further explored for evidence of retention time drift across multiple samples. Such changes are expected to be gradual in nature (barring acute system failure) and, if tracked for a given analyte species, are expected to follow one of a few indicative patterns recognisable to an experienced chromatographer. Some retention deviations are the consequence of a developing hardware malfunction such as the development of a leak in the solvent delivery system. In these scenarios, as the problem worsens, the change in retention time increases with each subsequent analysis. An illustration of such retention drift patterns is shown in Figure 4-2A. Such drift tends not to be selective, instead affecting entire portions of the chromatographic space. Conversely, where the deviation is the result of system exposure to mobile phase or sample matrix, the changes tend to become less severe with each subsequent analysis as the system reaches equilibrium. An illustration of equilibrium-derived retention drift patterns is shown in Figure 4-2B. Such drift tends to be chemically selective, with different species exhibiting different behaviour over time, resulting in the mixing of molecular content in the chromatographic space.

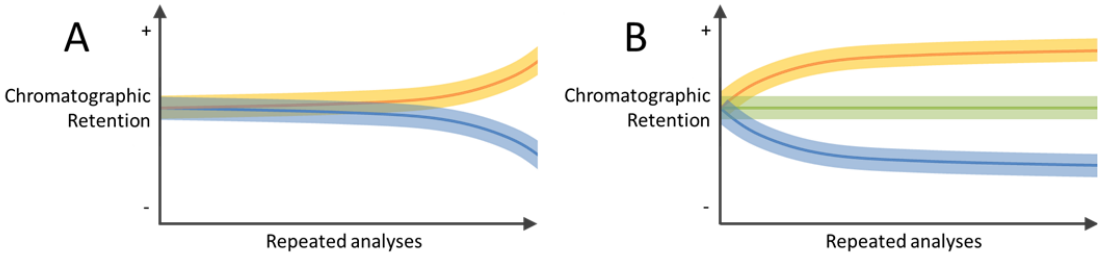


Figure 4-2. Panel A illustrates two possible patterns of analyte retention drift due to system instability. Developing system problems such as leaks in the pumps or fluid delivery path may cause increased (yellow) or decreased (blue) retention of an analyte species across repeated analyses. Panel B illustrates three possible patterns of analyte retention drift due to system equilibration, whereby analyte retention may increase (yellow) or decrease (blue) before stabilising, or remain constant (green) across all repeated analyses.

As the LC-MS analysis of the test set was completed without hardware errors, the drift illustrated in Figure 4-2A was not observed in the dataset. However, evaluation of the raw data produced by the profiling experiment reveals the presence of complex and selective equilibration drift throughout the dataset. To best illustrate this, an example was chosen whereby features with all three retention migration behaviours (outlined in Figure 4-2B) are observed in a single small chromatographic space. Figure 4-3 illustrates three sets of EICs generated from even numbered SR analyses 2-18 (stacked top to bottom according to run order). Panel A shows a cluster of features (m/z = 302.196) which elute earlier as the analysis order increases, while C (m/z = 295.008) shows a distinct feature that elutes later with increasing run order. A feature of intermediate mass (m/z = 297.156) is shown in panel B at a static retention time (red dotted line) as it does not migrate. Note that the feature shown in panel C changes elution order with the feature in panel B, violating the assumption of traditional feature grouping methods that elution order remains constant.

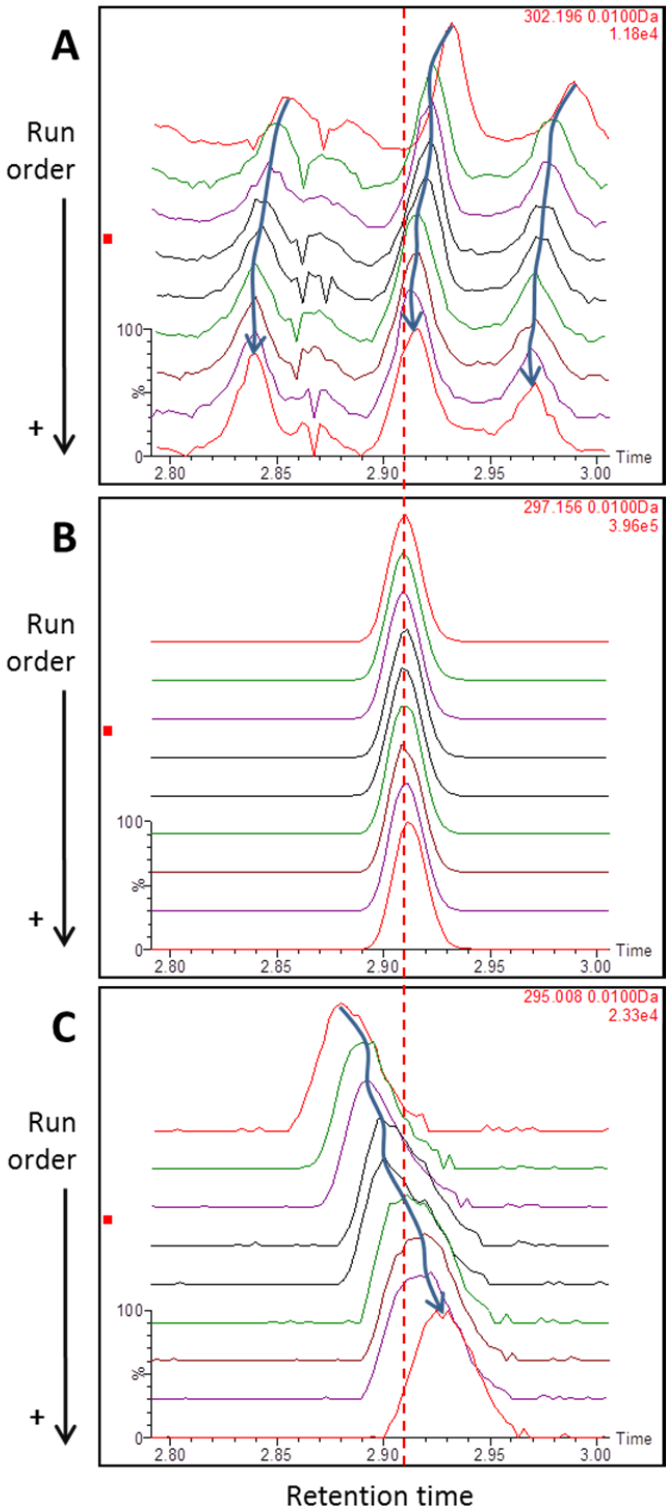


Figure 4-3. Heterogeneity in chromatographic feature drift. Extracted ion chromatograms of three m/z values (A = 302.196; B = 297.156; C = 295.008) eluting between 2.8 and 3 minutes in a series

of LC-MS analyses of a SR urine sample. Nine analyses are shown representing every other SR sample from the first 18 SR analyses of the experiment. Chromatographic drift to both earlier (A) and later (C) retention times are observed with increasing run order in comparison to a feature that is not observed to drift (B).

4.5.3 Feature clusters and cluster migration

Despite the efforts presented in Chapter 3 at increasing chromatographic peak capacity, clusters of poorly resolved features may readily be observed in complex biofluids such as human urine. The components of these clusters are often related in chemical composition and structure making them difficult to separate by LC, and sometimes impossible to differentiate by accurate mass measurement (e.g. isobars such as leucine and isoleucine). A representative feature cluster (m/z = 310.2015 +/- 0.0025) from the test set SR urine analyses is shown in Figure 4-4. Fortunately, the prerequisite peak detection algorithms such as XCMS's centWave are well suited to differentiating closely eluting features, even where baseline chromatographic separation is not achieved. A high degree of accuracy can be seen in the 2D plot of features detected in centWave when compared to the complex EIC immediately above. This complex cluster of features will be used later in this chapter as an example to illustrate the performance of feature alignment and grouping between sample pairs.

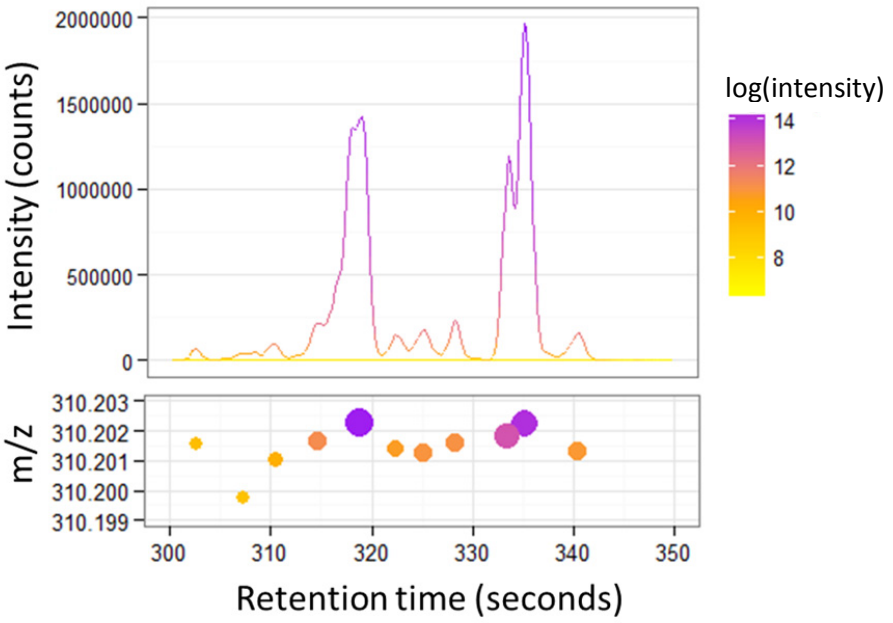


Figure 4-4. A representative feature cluster (m/z = 310.2015 +/- 0.0025) from the test set SR urine analyses shown as an EIC and as featured detected by XCMS centWave.

Clustered features are subject to the same measurement error and systematic variance that potentially affect all features, but their close proximity compounds the challenge of accurate feature grouping, warranting special consideration in the design and performance testing of a grouping algorithm (Tengstrand et al., 2014). Furthermore, as noted in the literature (Chae et al., 2008), it is known to chromatographers that features within such clusters often display similar chromatographic behavior. An example of this from the test set SR analyses is illustrated in Figure 4-5 (panel A), with a distinct yet simultaneously eluting feature shown for reference (panel B) demonstrating that the migration observed is specific to the features within the cluster and not due to generic chromatogram warping (as the reported retention time deviation curves reported in Figure 4-1 would otherwise suggest for features eluting near 380 seconds). While this "whole cluster migration" may be visually intuitive to deconvolve when run order is considered (*i.e.* the plotting order in Figure 4-5), feature grouping methods that consider the entire dataset and discard analysis order are easily confounded by cluster migration, as peaks of near identical mass migrate into retention spaces previously occupied by

adjacent peaks within the same cluster. It is therefore difficult, with no knowledge of the sample analysis order, to determine how to set and adjust bounds for accurate feature grouping.

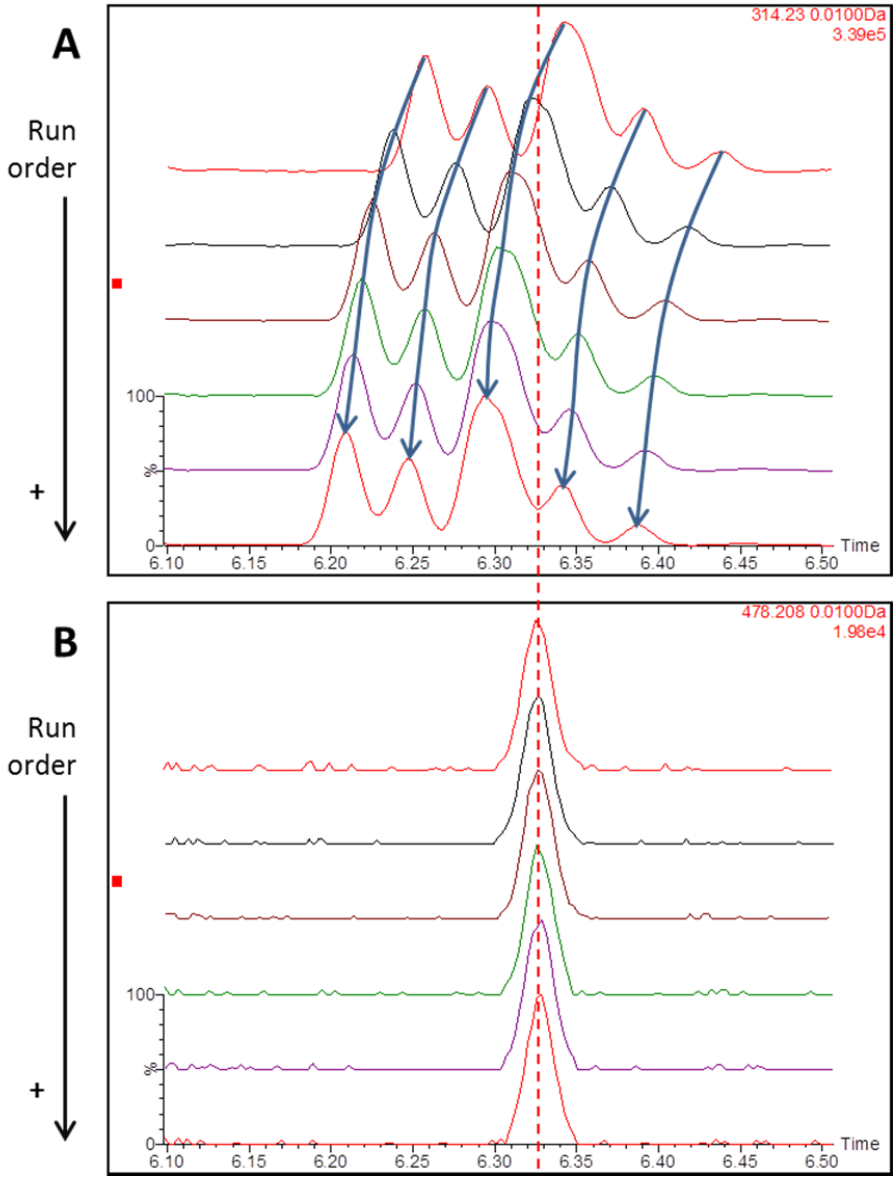


Figure 4-5. A cluster of features (m/z = 314.23) exhibiting the same retention migration behavior (A) with respect to increasing run order. A static feature (m/z = 478.208) within the same chromatographic elution window is shown to provide a reference ensuring the drift observed is not due to generic warping.

4.5 Experimental LC-MS dataset

Simple and fast grouping methods that consider the entire dataset such as simple binning (Gurdeniz et al., 2012) (illustrated in Figure 4-6, panel B) and feature density calculation therefore risk the erroneous inclusion of nearby peaks of similar mass and retention time in the formation of groups. A solution is needed which is sensitive and adaptive to the migration of features from sample to sample (Figure 4-6, panel C), and is able to leverage this knowledge to perform more accurate and specific feature grouping, yielding higher quality datasets which better represent the underlying raw data.

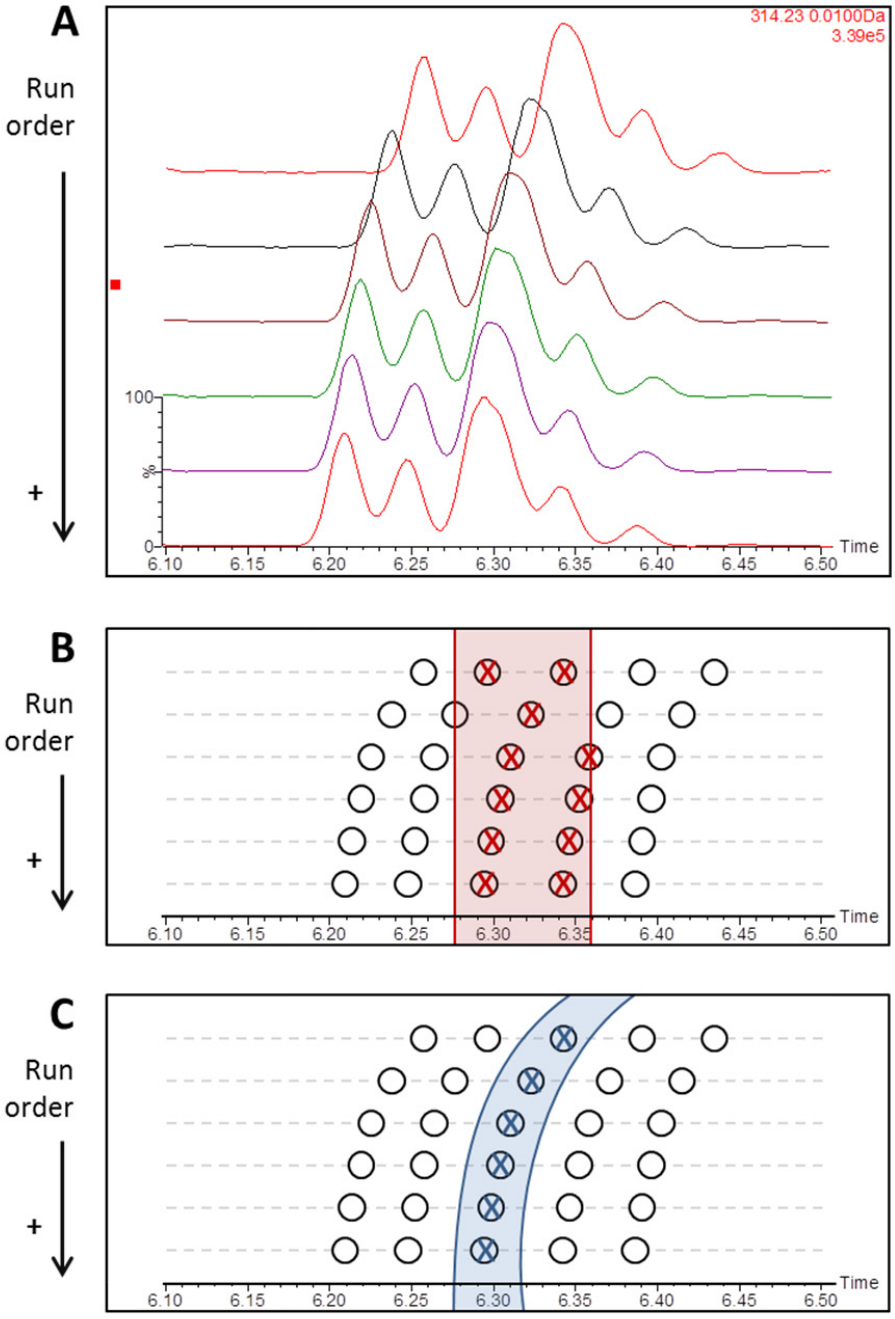


Figure 4-6. A cluster (m/z = 314.23) exhibiting retention migration behavior with respect to increasing run order is illustrated (A) along with cartoon representations of feature density and binning approaches to feature grouping (B) and the proposed pairwise approach to grouping in run order (C).

4.5.4 Pairwise comparison

A key observation in the test dataset is that the chromatographic variance of peaks between sequential analyses tends to be small in comparison to the variance observed over an entire analysis, especially on a population screening scale. This phenomenon is illustrated in Figure 4-7, which shows the exemplar m/z = 310.2015 cluster in sequential pairs of SR analyses (each 5 sample injections apart) from the start, middle, and end of the 864 sample test analysis. Pairwise comparison may therefore be made between features detected in sequential analyses using minute windows of retention error. It is this observation, applicable to both feature clusters and independent features, that guided the conceptual foundation for the grouping algorithm presented here. By connecting features between samples through a process of iterative pairwise matching using small windows of retention error, the reproducibility of LC-MS across short analytical durations can be exploited for improved grouping across the entire dataset, regardless of the size or duration of collection.

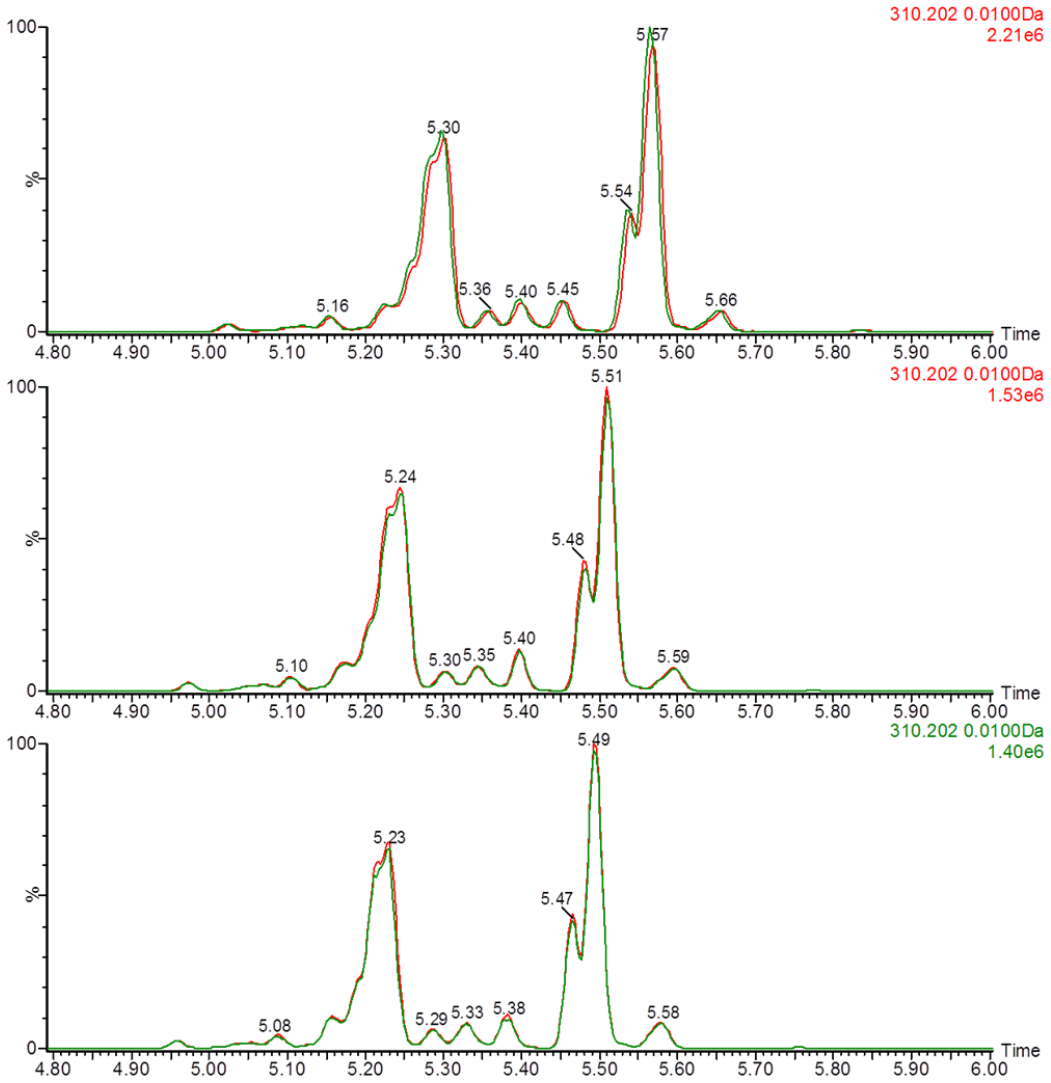


Figure 4-7. Pairwise agreement in feature retention between sequential quality control samples from the early, mid, and late portions of the 864-sample analysis. EIC's are shown for the m/z = 310.2015 (+/- 0.01 Da.) feature cluster. While there is a substantial difference in the retention times among each set, the pairwise retention times are virtually identical for each pair.

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

Code for iterative pairwise feature matching was developed and implemented within the R software environment, and may be found in Appendix 3. The script, referred to as the "ROgroup" method, is engineered to be compliant with a modular framework whereby any data preprocessing and feature

detection algorithm can be utilized, provided that it is capable of generating a table of m/z, retention time, and intensity values for features on a sample-by-sample basis. The first and second samples in the analysis order are imported to XCMS for independent feature extraction using the centWave algorithm to generate feature sets. The feature set generated from sample n is designated as the template, while the feature set from sample n+1 is designated as the candidate. The feature matching procedure is then implemented between the template and candidate feature sets as described in the following sections, using the algorithm illustrated in Figure 4-8. The matching process is iterated for each sample pair within the sample set, with the candidate set from the previous round of pairwise matching becoming the template set in the next round. All detected features within a sample set, as well as any links established among them through matching, are collated in a link ledger.

Of all measurements potentially associated with a feature (*e.g.* chromatographic peak shape, signal to noise, etc.), only the observed m/z ratio and chromatographic retention time are used to determine feature matches (except in special cases of complete ambiguity, detailed below). Furthermore, in all cases, matching is performed by setting a window of retention time and m/z error, centered on the target value, to create "intelligent bins". This simple approach was chosen to highlight the efficacy of the concept of pairwise matching in run order, and the overall algorithm, rather than rely on the complexity of the matching method itself.

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

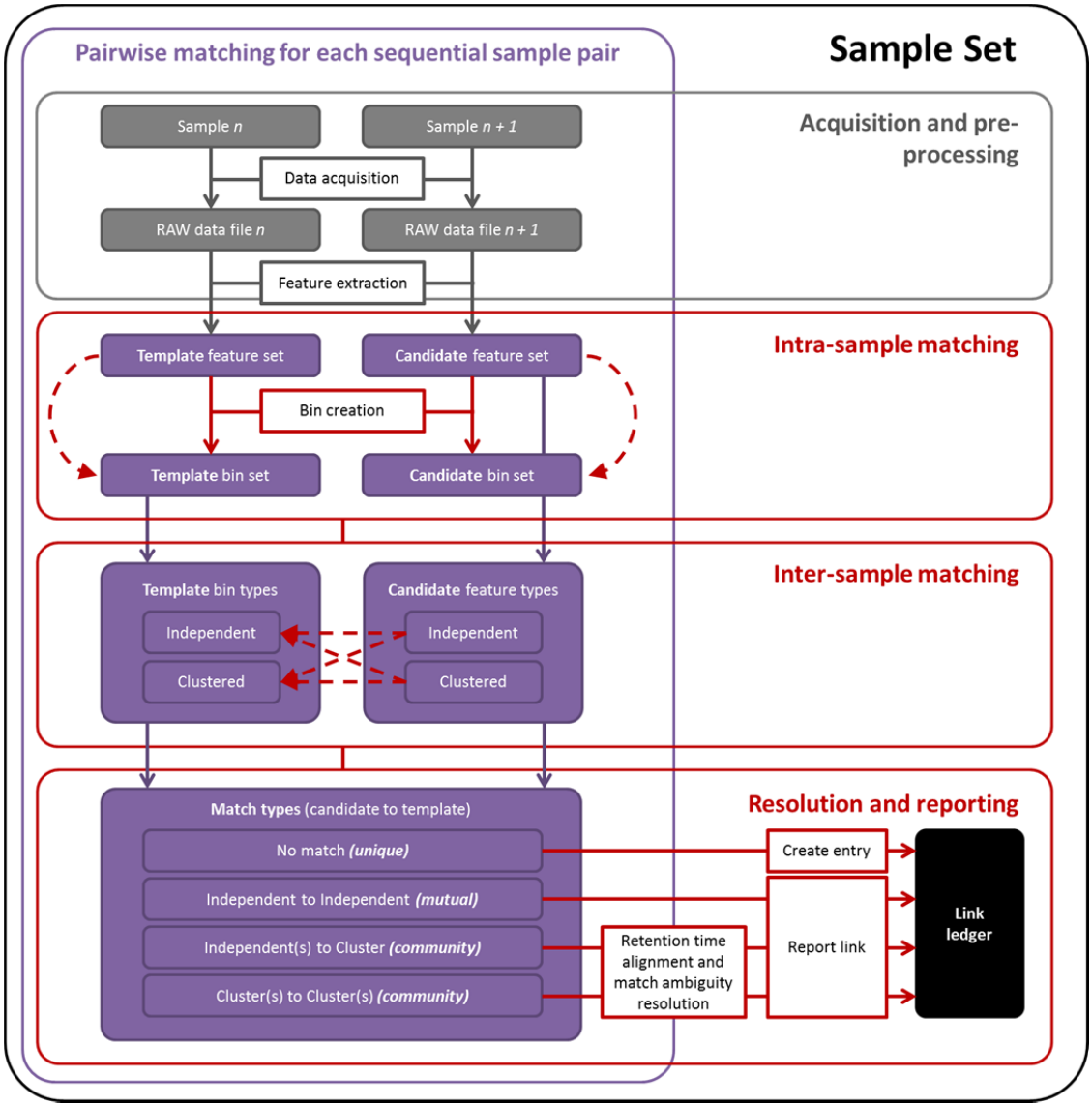


Figure 4-8. For a set of samples, iterative rounds of pairwise matching (purple) are conducted between feature sets from sequential sample analyses. Data acquisition and feature extraction (grey) are prerequisites for pairwise matching between template and candidate sets. The matching and reporting workflow steps and procedures (red) are detailed in the following sections.

4.6.1 Intra-sample matching to define feature clusters

As discussed in Section 4.5.3, clusters of features are likely to produce ambiguous pairwise matches where more than one feature matches a single feature in the opposing dataset. Therefore, early detection of feature clusters can help to shunt those features into a specialised matching scheme,

avoiding the occurrence of ambiguous matching in the rest of the dataset. Features that are very close in both m/z and retention are therefore identified and marked as clusters in a process of intra-sample matching whereby the feature set is matched against itself. To achieve this, intelligent bins are built for every feature in the set. The feature set is then matched against its own feature-derived bins, with matches indicating the presence of closely related features in the set. Together, these matched features are determined to be clusters, and are designated with a unique cluster identification number. Features with no intra-sample matches are considered to be "independent". A single feature cluster as detected by the software is illustrated in Figure 4-9. It is important to note that individual features in a cluster may be relatively distant from one another in either m/z and/or retention time, as long as they are indirectly connected through other features present in the cluster. This process is conducted for both the template and candidate feature sets, defining both clusters and independent features in each.

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

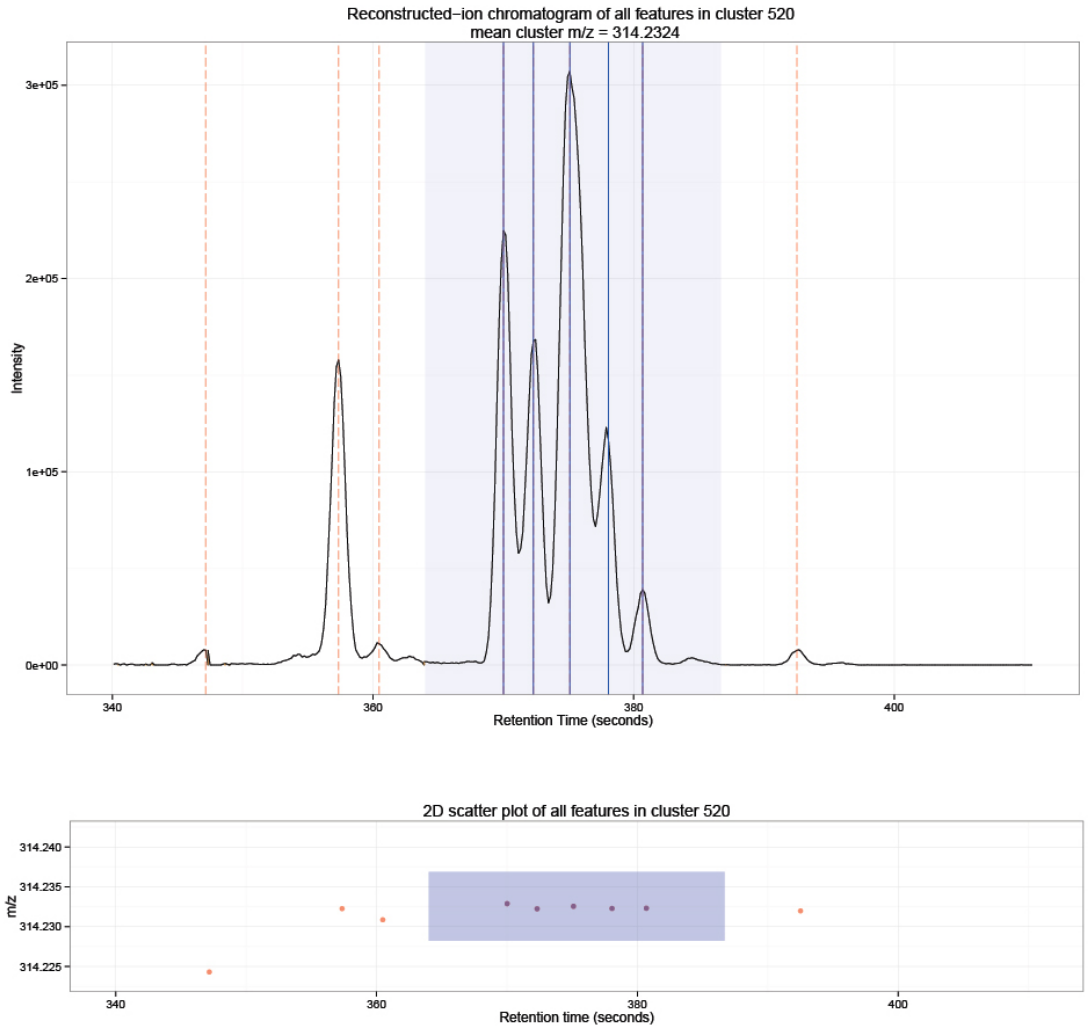


Figure 4-9. A single feature cluster detected in a representative sample of human urine. The top panel shows the cluster EIC (mean m/z = 314.2324 +/- 30ppm). All detected features are indicated with vertical lines, with solid blue lines indicating inclusion in the cluster, and dotted-red lines indicating exclusion from the cluster based on retention time, mass, or both. The bottom panel shows the same detected features in both m/z and retention time dimensions. In both panels, the blue shaded area indicates the boundaries of the cluster for retention time (top) and both retention time and m/z (bottom).

To illustrate the procedure's outcome on a representative example, feature clustering was performed on human urine sample p5-SR8, analyzed at the mid-point of the 9-plate reversed-phase LC-MS test

analysis. Of the 10,280 features extracted from the template data file, 6451 (63%) were found to be independent, while the remaining 3829 (37%) were associated in clusters. Figure 4-10 illustrates all 10,280 features extracted from the template data file, each colored by the size of its associated cluster (independent features are shown in black with 90% transparency). Colocation of the largest feature clusters is observed in areas that are inherently noise-prone, such as the signals from solvent contaminants at the end of the chromatogram, and during the post-gradient equilibration period (after 600 seconds).

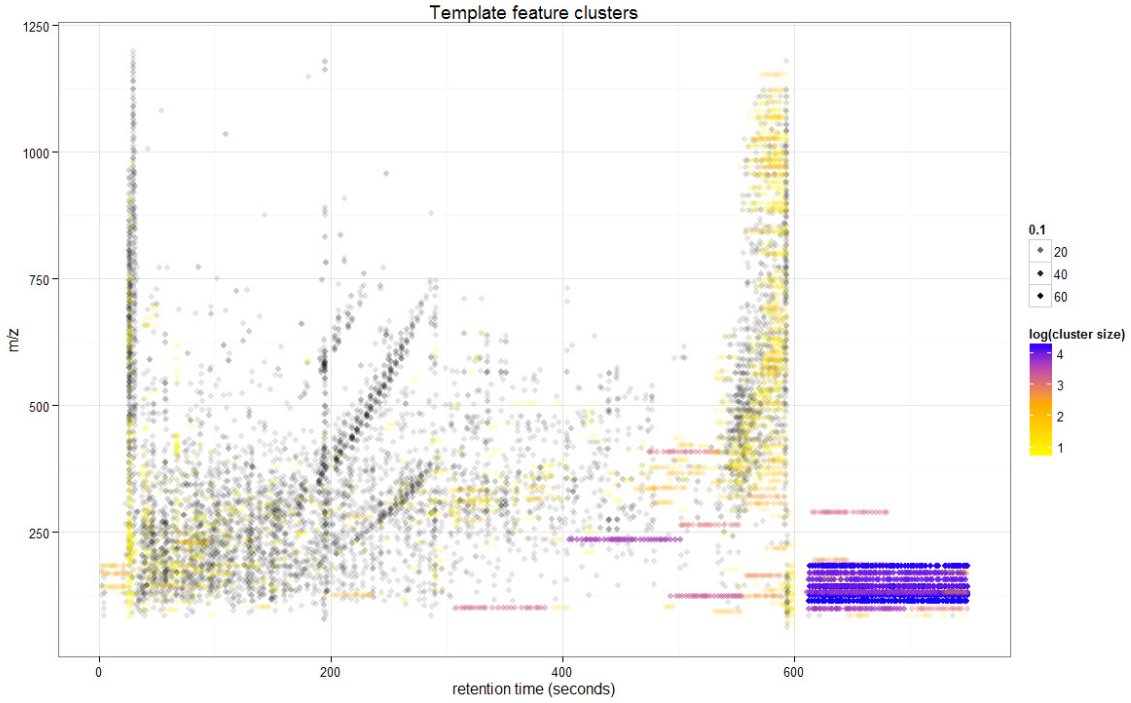


Figure 4-10. Spatial orientation of large clusters in a representative reversed-phase LC-MS dataset. Independent features are shown in black, with a transparency of 90% in an effort to minimize the over plotting effect observed with the given information density. Clustered features are shown in color, with the color and transparency set relative to the log of the cluster size in order to highlight the location of clusters of greater size.

The extent to which noise data were detected as features by the detection software was an unexpected result. In this example, 10.9% of all features in the dataset belong to clusters with sizes of 10 or greater (38 clusters in total). EIC plots of those clusters were automatically generated for manual evaluation,

and three such clusters are illustrated in Figure 4-11. It is clear that erroneous feature detection within constant signals such as those originating from mobile phase contaminants produces features of near-identical mass with high retention time density, and that these are in turn likely to generate large feature clusters. Evaluation of features on a per-cluster basis is an efficient form of data review, as 10.9% of the dataset (1,120 features) were verified as fit for omission in less than one minute of manual review. While within this example it must be acknowledged that the most severe source of solvent-derived noise in the dataset could be avoided by truncating the feature detection scan range to a more limited subset within the raw data file, more severely contaminated datasets may still benefit from such an approach to de-noising.

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

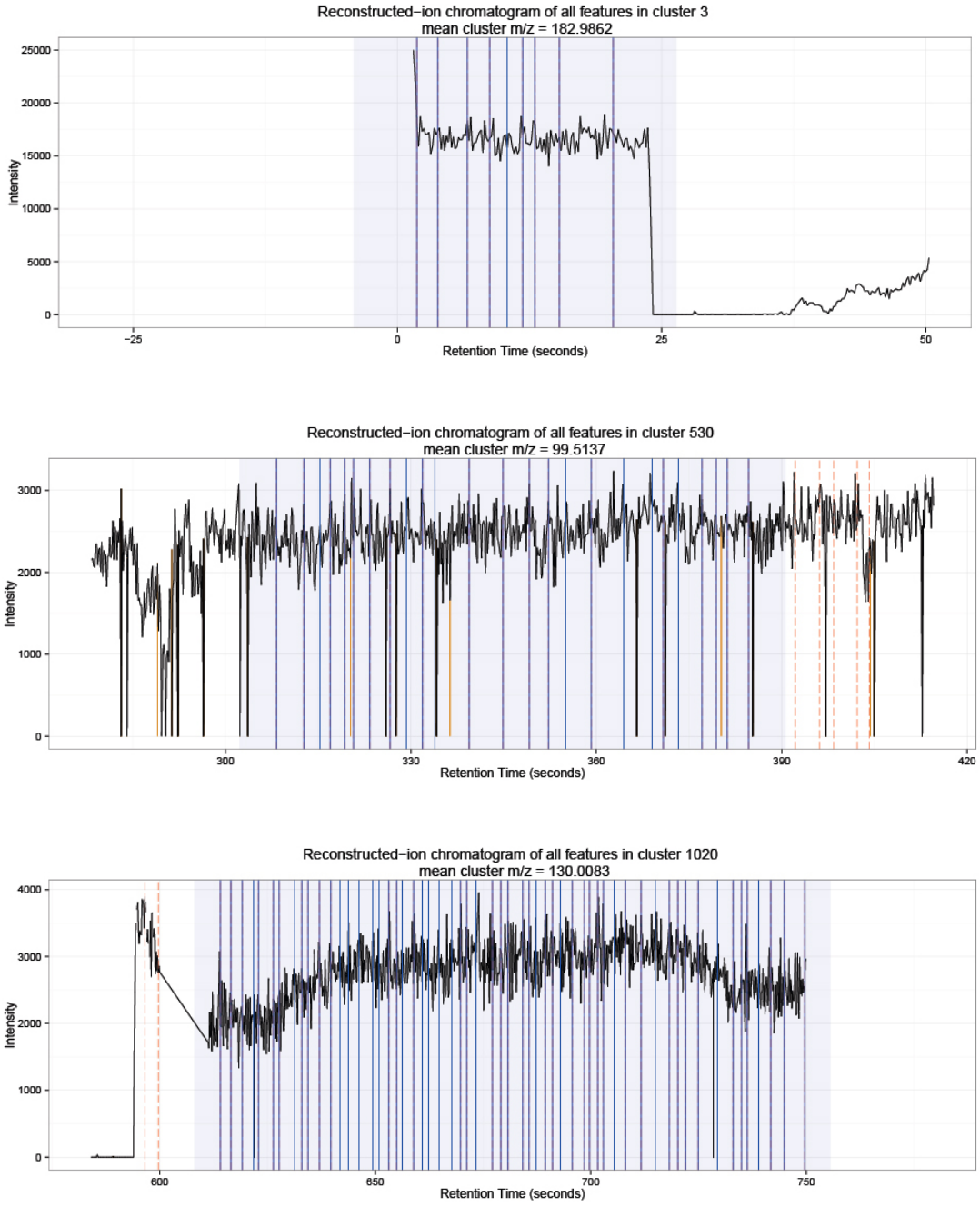


Figure 4-11. Representative large clusters of 10 or more features from early, mid, and late retention times (top to bottom, respectively.)

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

4.6.2 Inter-sample matching

Having identified feature clusters in the template and candidate datasets, the next step is to identify matching features between them. Inter-sample feature matches are established in the same manner used for intra-sample matching whereby intelligent bins are constructed in one feature set (in this case, the template) and features from the opposing set (candidate set) are matched against them. Bins are again built by applying a tolerance window to both the m/z and retention time values of each template feature. However, the tolerance windows used are not equal to those used in intra-sample matching. Use of equal tolerances would allow for fringe cases where two features in the candidate set do not match each other, but both match a feature oriented perfectly between them (in retention time) from the template set, and thus an ambiguous match is made from three independent features (as illustrated in Figure 4-12, left). In order to avoid creating a separate resolution pipeline for these special cases, the values used for inter-sample matching are set to 50% of those used in intra-sample matching, thereby ensuring that ambiguous matches only occur where feature clusters are involved (Figure 4-12, right).

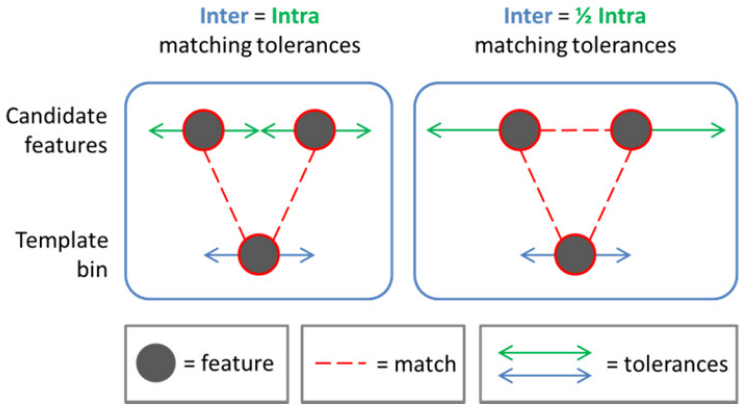


Figure 4-12. Tolerance windows for inter (blue) and intra (green) sample feature matching. Red dotted lines denote matches. To avoid the scenario whereby a single template bin may match two independent candidate features (left), the intra-sample matching tolerance is set to double the value of the inter-sample matching tolerance. Thus, ambiguous matches always involve clusters, allowing for a single scheme of ambiguity resolution.

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

Matching produces three outcomes:

1. Candidate features with no matching template feature are designated as ***unique***, and are appended as new entries to the link ledger.
2. Independent (unclustered) candidate features matching a single independent template feature are designated as ***mutual***.
3. Any match involving a feature cluster, whether it be in the candidate or template feature sets, is designated as a ***community***, containing an ambiguous match requiring conflict resolution (discussed in the subsequent section).

To demonstrate intra-sample matching in the context of the test dataset, detected features from samples p5-SR8 (as template) and p5-SR9 (as candidate) were matched. The majority of candidate features (58%) had only a single match in the template feature set. Of the remaining candidate features, 15% matched multiple template features, and 27% were unique, having no template matches. The majority of single matches (92%) were mutual, being reciprocated as single matches in the reverse direction (template to candidate matching). The remainders were ambiguous matches in the reverse direction. At this stage, mutual matches may be reported to the link ledger as links, and unique features may be appended as new entries.

4.6.3 Feature communities and resolution of ambiguous matching

Match ambiguity can give rise to networks whereby independent or clustered candidate features match independent or clustered template features, which may in turn match other candidate features, and so on. Where such networks exist, the resolution of a single ambiguous match may have knock-on effects that influence the resolution of other matches. For example, choosing which match to resolve first may influence the outcome for the entire network where a cascade of incorrect matching is initiated by the consumption of a single key feature in an erroneous initial match. Resolution of ambiguous matching is therefore a process which benefits from complete prior knowledge of the network (or "community" as defined herein), and a method of resolution that provides the best matching scheme for the community as a whole, rather than for any one pair of features.

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

Feature communities are built by an iterative process of collating all connecting clustered and matched features within and between the candidate and template feature sets. A cartoon illustration of this procedure is shown in Figure 4-13. Step 1 starts with the selection of a single arbitrary feature (blue), and seeks to include any clustered features (solid red lines) in the growing community. In step 2, inter-sample matches for all clustered features (dotted red line) are included in the community. All features which cluster with those newly found are included (step 3), and matches back to features in the original dataset are sought (step 4). The processes repeat until the community no longer grows (as shown in steps 5 and 6) indicating the successful definition of the community boundary where neither additional pairwise matching nor additional intra-sample interference is observed in either sample. Each community receives a unique ID number which is shared for all community features in both template and candidate datasets.

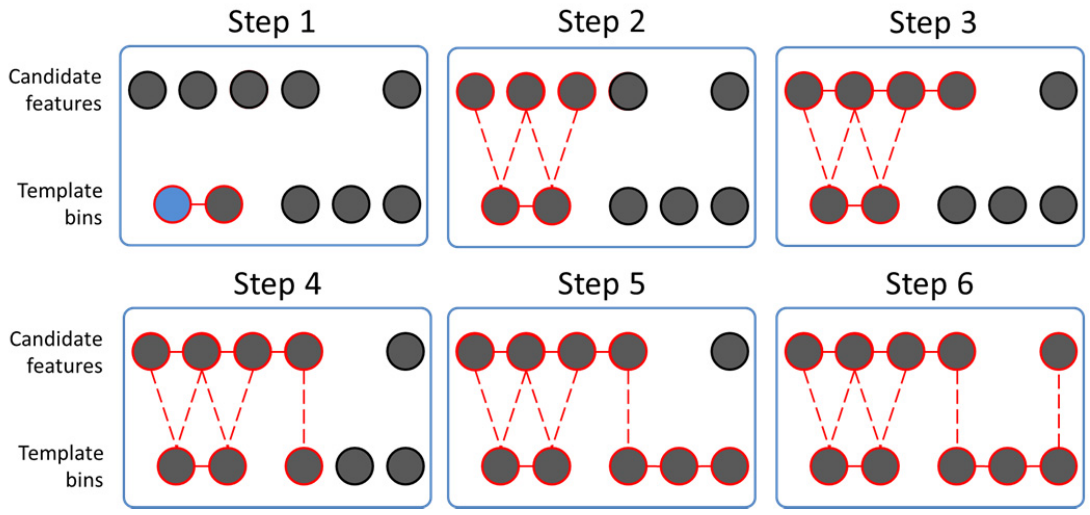


Figure 4-13. Community boundary definition by iterative collation of clustered and matching features. In this simplified illustration, the linear spatial closeness of features (dots) is a simplified surrogate for combined m/z and RT similarity, whereby features close to each other match (cluster), and features distant from each other do not.

Once a distinct feature community has been defined, the ambiguous matches contained within are resolved in a manner determined by the best matching scheme for the community as a whole. To illustrate the community resolution workflow, a challenging example was sought whereby a complex

cluster was found to migrate by a relatively large degree between template and candidate samples. The first and last SR analyses from the 1st plate in the test set were chosen as the template and candidate datasets for this purpose, and the m/z = 310.2015 cluster illustrated previously (Figure 4-4) was specifically observed. EICs (m/z = 310.2015 +/- 0.01Da) from both analyses are shown in Figure 4-14 (top) illustrating the shift in cluster retention time and resulting overlap of disparate features. The 11 features of the m/z = 310.2015 cluster were detected in both samples and collated as community number 627 as illustrated in Figure 4-14 (bottom).

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

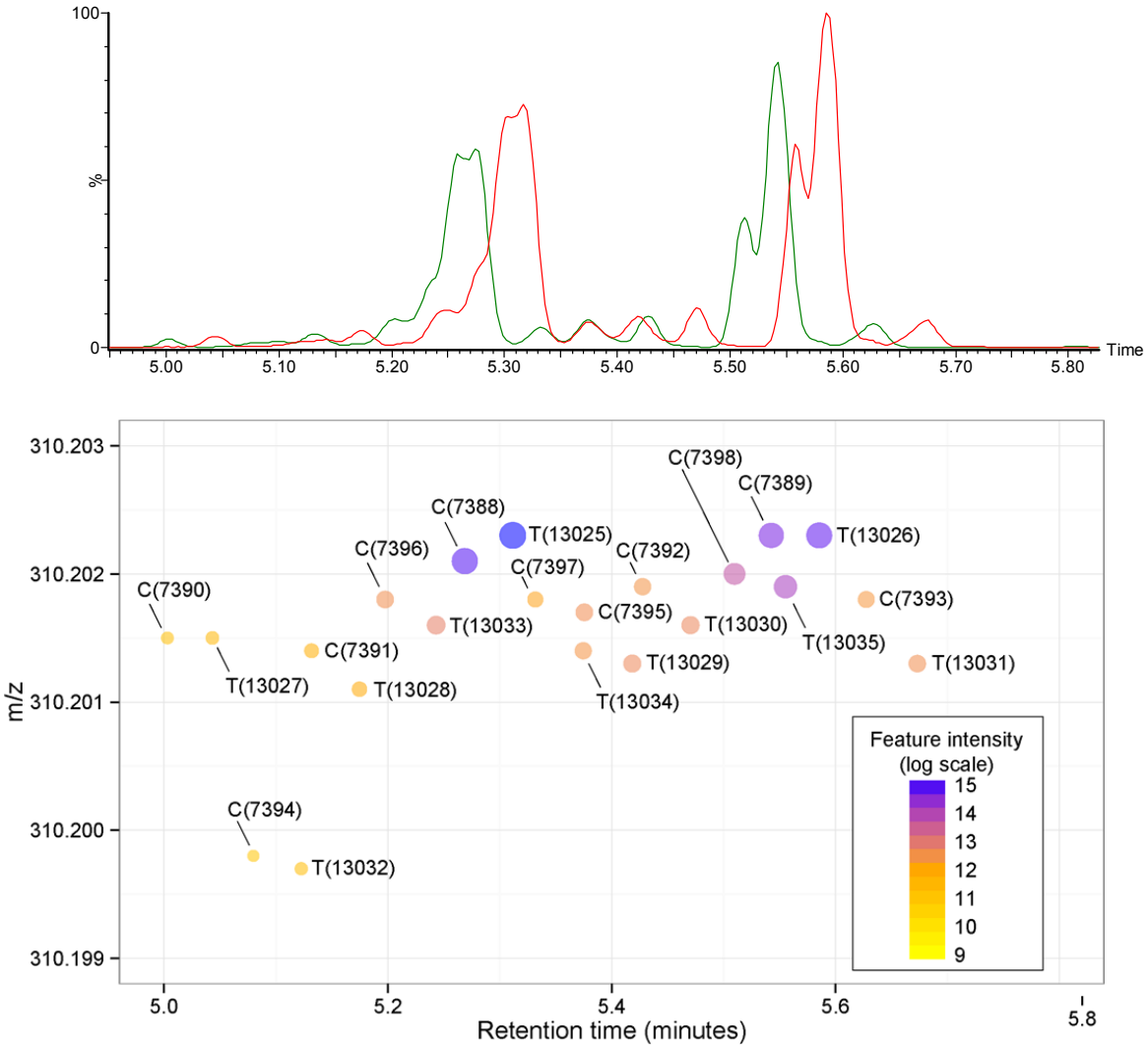


Figure 4-14. Feature community #627, illustrated as EIC peaks and as detected features. EICs of m/z = 310.2015 +/- 0.01Da in the first (red) and last (green) SR analyses from plate 1 of the test set, representing the template and candidate for matching, respectively (top). The centWave-extracted features from these two clusters were collected by the ROgroup script into a single community, and are shown in the two-dimensional plot below (bottom). Extracted features are annotated according to their origin from either the template (T) or candidate (C) feature set, as well as their individual feature ID number. For visual clarity, features are both sized and colored according to the log value of their intensity.

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

Following on from the assumption that clustered features of similar mass and retention time will experience similar drift in retention time across repeated analyses, pattern matching between the candidate and template community features is attempted. This is accomplished by applying a systematic linear shift in retention time to all candidate features while leaving the retention times of the template features unchanged. The shift that produces the best overall match for the community is determined by summation of the difference between the retention time of each candidate and its closest match among the template features. For example, the residual values between all candidate and template features are calculated and illustrated in the heatmap matrix shown in Figure 4-15. The community match score is the sum of the smallest residual in each (candidate) row.

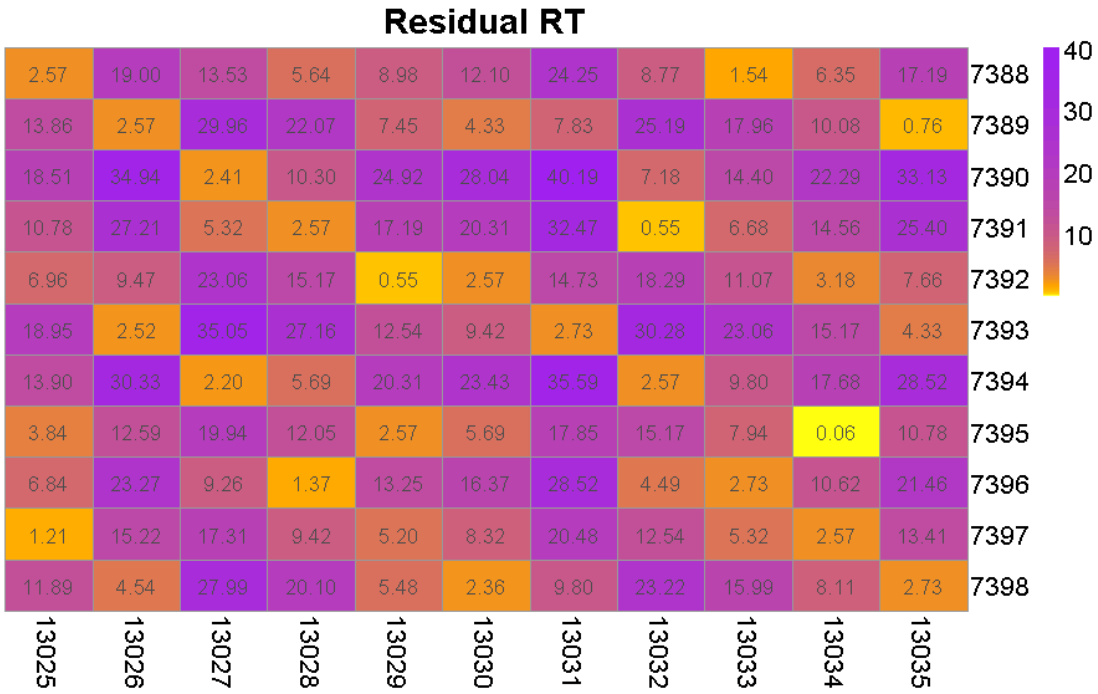


Figure 4-15. A heatmap of the retention time differences (in seconds) among all candidate and template features in community #627. Candidate features are listed along the y axis, while template features are listed along the x axis. Closest matches are shown in a yellow.

The retention times of all candidate features are then changed in increments of 0.1 seconds, and the community match score is recalculated for each iteration. This is repeated across a fixed window which is based on the intra-sample clustering retention time window (in this case, +/- 6 seconds). The community match scores are collated to find the minimum overall value (+2.6 seconds, in this example) as illustrated in Figure 4-16.

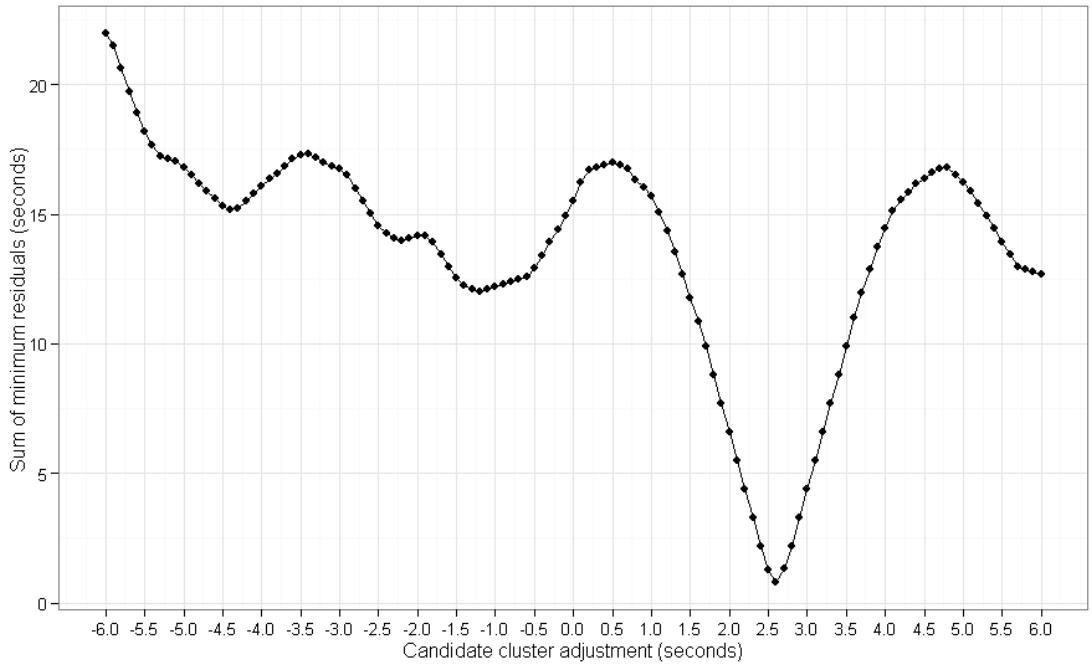


Figure 4-16. Pattern matching of feature clusters within community #627. The optimal set of matches is determined by shifting one cluster relative to the other in a linear manner, and calculating the sum of lowest retention time differences across all candidate features.

The matrix of retention time differences with the optimal correction (the retention time shift value that produces the lowest minimum residual sum for the whole feature community) applied to the candidate features is shown in Figure 4-17, clearly a change of individual matches driven by a global metric of best fit for the community.

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

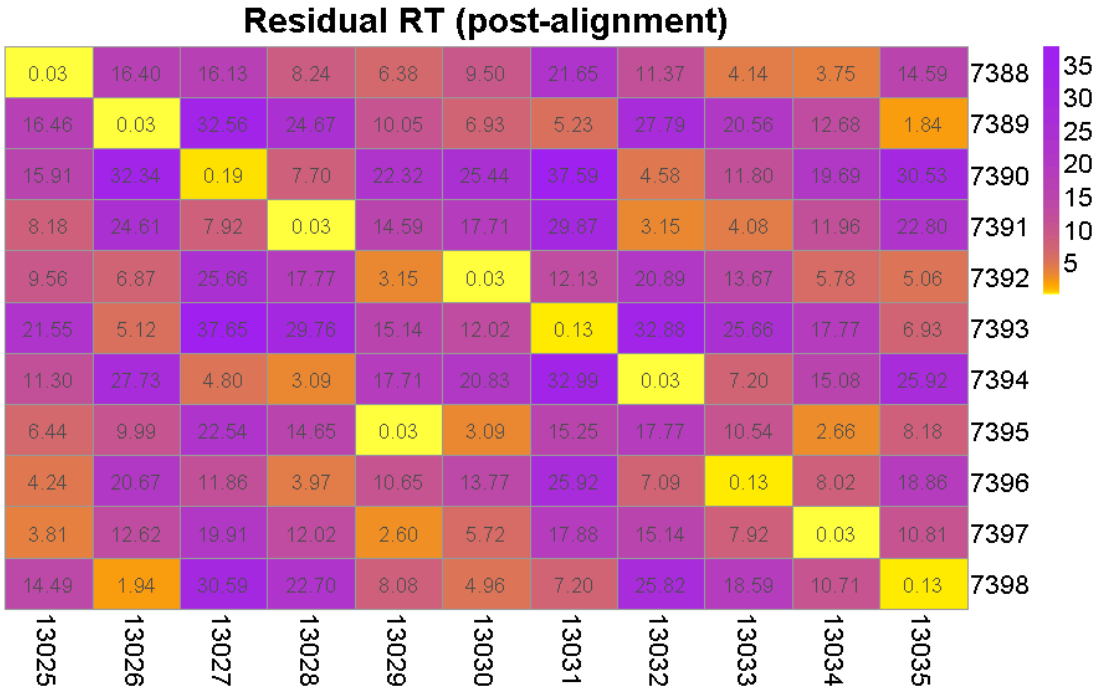


Figure 4-17. A heatmap of the community-optimised retention time differences (in seconds) among all candidate and template features in community #627. Candidate features are listed along the y axis, while template features are listed along the x axis. Closest matches are shown in a yellow.

Prior to reporting the optimized matches as linked features, a check is performed which ensures that each individual match refined in this process was originally defined as a potential match by inter-sample matching, This is visually represented in Figure 4-18 as a superimposed mask on the heatmap from Figure 4-17 where original inter-sample non-matches are represented by white cells (NA = not originally a match). This action is performed to prohibit the creation of new matches simply because they are the closest (but still distant) option presented. This is consistent with the goal of resolving ambiguity rather than creating entirely new matching schemes.

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

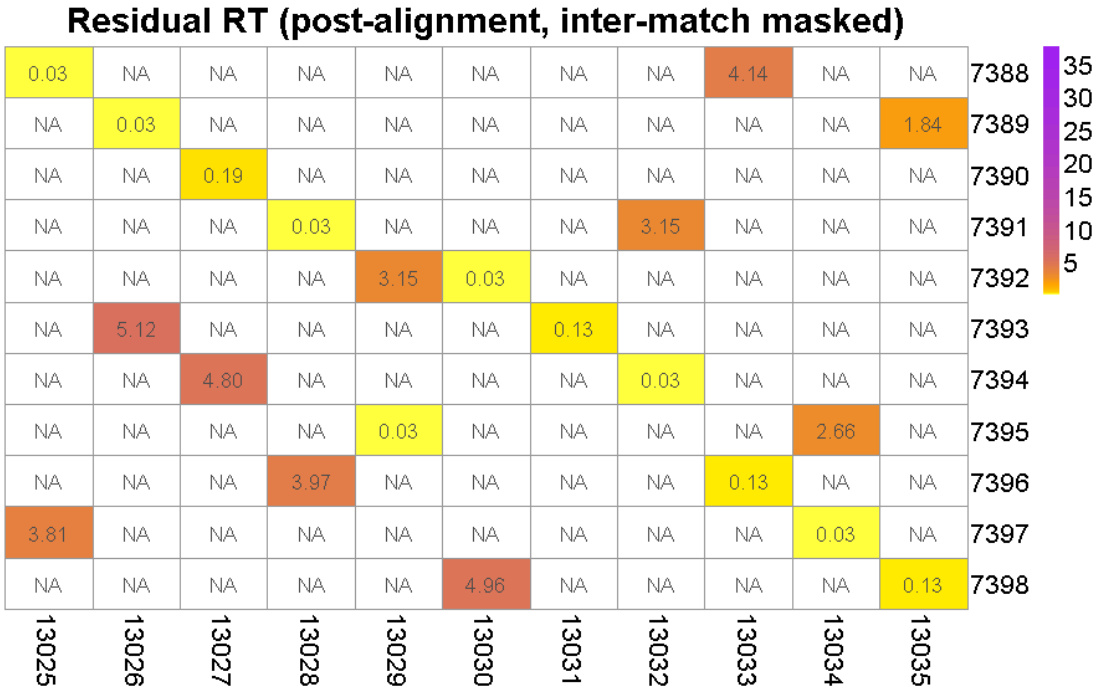


Figure 4-18. A masked heatmap of the community-optimised retention time differences (in seconds) among all candidate and template features in community #627. Masking (white cell with value = NA) indicates that the candidate/template feature pair was not initially matched prior to alignment.

Finally, individual matches are reported as links between features. This process starts by finding the lowest residual value in the masked matrix (*e.g.* 0.03, in Figure 4-18, rounded for visual calrity), and reporting the corresponding row and column as linked candidate and template features. The row and column are then removed from the matrix, and the process is repeated. This continues until all rows and columns with potential matches are depleted. In rare cases, retention time alignment alone is unable to unambiguously resolve all potential matches within a community. In these cases, the mass difference is used to resolve the ambiguity. However, it should be noted that while comparison of observed m/z values between features may produce accurate matches where all other match possibilities have a relatively large mass difference (*e.g.* in the matching of candidate feature 7394 to template feature 13032 illustrated in Figure 4-19), once the competing mass differences are within the analytical error of the MS instrument, m/z comparison alone loses virtually all of its ability to produce accurate matches. This is illustrated by the presence of many mediocre matches between features in

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

Figure 19, colored in orange and purple, with no clear pattern of optimal matches (yellow cells) unlike the pattern clearly obtained by retention time matching (Figure 17).

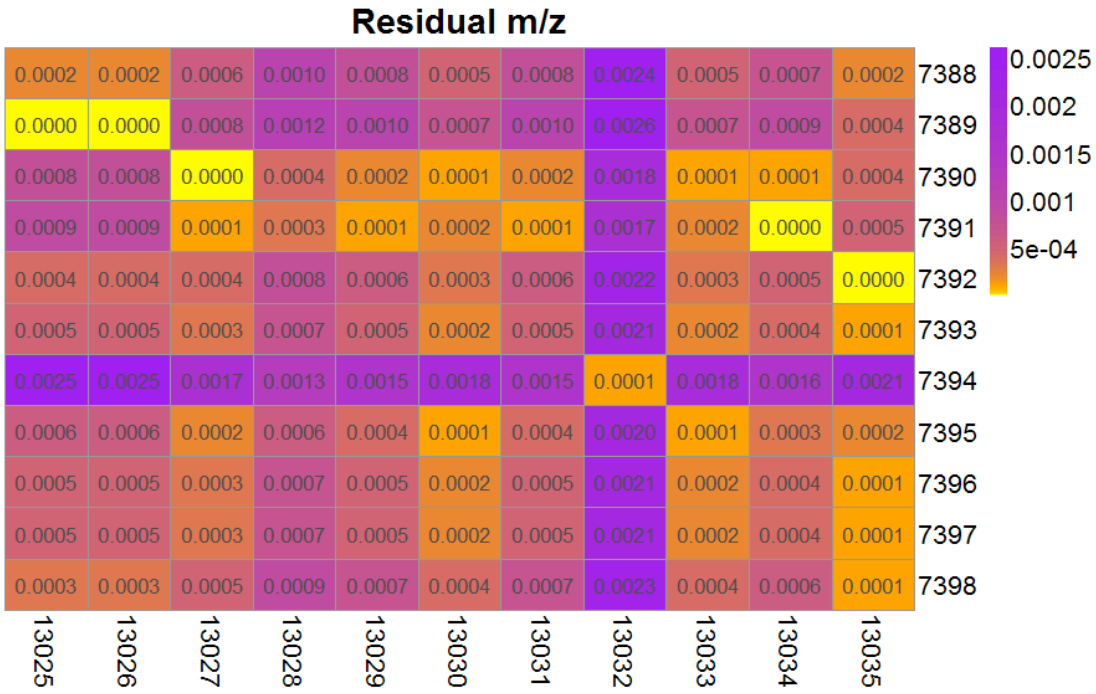


Figure 4-19. A heatmap of the m/z value differences among all candidate and template features in community #627. Candidate features are listed along the y axis, while template features are listed along the x axis. Closest matches are shown in a yellow.

4.6.4 Reporting links between features

The process of matching features to discover links is repeated for each sequential pair of samples in the dataset, with each candidate sample becoming the new template in the next round of matching. As the process continues, links between features among all samples are accumulated, requiring a set of repositories for the identifying data (eg. unique feature numbers) as well as the measured data associated with each feature (m/z, retention time, and intensity). Each of these sets of data are stored and organized in a link ledger which resembles a traditional matrix of extracted LC-MS data whereby each sample is represented by a single column, and each grouped set of features across the total dataset is represented by an individual row. As the process of iterative pairwise matching proceeds, the link ledger is grown in the following manner, as illustrated in Table 1.

4.6 Implementation of pairwise feature set matching across replicate quality control samples (ROgroup)

Experiment Feature ID	Sample 1	Sample 2	Sample 3
1	**1**	**1563**	**256**
2	**2**	**57**	NA
3	**3**	**425**	**4861**
4	**4**	NA	**24**
5	**5**	**426**	**1159**
6	**6**	**2128**	NA
7	**7**	**3581**	**658**
…	…	…	…
4527	**4527**	NA	**2897**
4528	NA	**3578**	NA
4529	NA	**3579**	**545**
4530	NA	**3582**	NA
…	…	…	…
7894	NA	**6144**	**3487**
7895	NA	NA	**4456**
7896	NA	NA	**4457**
7897	NA	NA	**4458**
…	…	…	…
9214	NA	NA	**6247**

Table 4-2: The feature matching (link) ledger, abbreviated, for hypothetical feature sets from three samples.

First, the identities of all features extracted from the first sample of the experiment are stored as a column of sequential unique numbers (orange cells). In this example, 4527 features were extracted from the first sample. These feature numbers are then used, in the broader context of the experiment, to describe any subsequently associated (linked) feature, and therefore are set as "Experiment Feature ID's" (black cells). Features from the second sample are then linked to those from the first sample by the matching algorithm described above. These features are reported in an adjacent column (green) using their numerical identifying number from the second sample feature set. In the example shown,

feature 1563 from sample two was positively matched to feature one of sample one, and therefore is captured in the same row, denoted overall as experiment feature 1. Where no feature from sample two could be matched into an existing row, an NA is entered (eg. in row 4). Conversely, where features from sample two have no match in the sample one featureset, they are appended to the bottom of the link ledger as new rows. In the example shown, the ledger is extended by 3367 new rows to create 7894 in total. Values of "NA" are retroactively assigned to those rows in columns from all earlier samples, explicitly stating the absence of a matching feature (grey cells). The process is repeated for all new samples (sample three shown in blue), reporting feature links, NA values, and appending new rows to the ledger. Separate matrices are generated in an identical manner for the storage of feature m/z, retention time, and intensity information. By referring to an Experiment Feature ID across these coordinated matrices, one can evaluate the measured properties of all linked features across the experiment.

A visual example of such evaluation is plotted in Figure 4-20, whereby the average m/z, average RT, and average intensity was plotted for each group made using the ROgroup approach on plate 1 SR urine sample feature sets (16 in total). Each feature is colored by the number of linked features in the group (i.e. the chain length), with blue dots indicating complete chains and yellow dots indicating chains composed of only a few linked features. In this example it can be observed that low chain length co-localizes with areas of noise within the chromatogram (e.g. m/z streaks from background contamination and the swell of background noise between 550 and 600s corresponding to the high percentage organic washing phase). Medium chain lengths (orange) are observed within the first approximate minute of the chromatogram, potentially indicating a higher variance in retention time than observed throughout the remainder of the chromatogram.

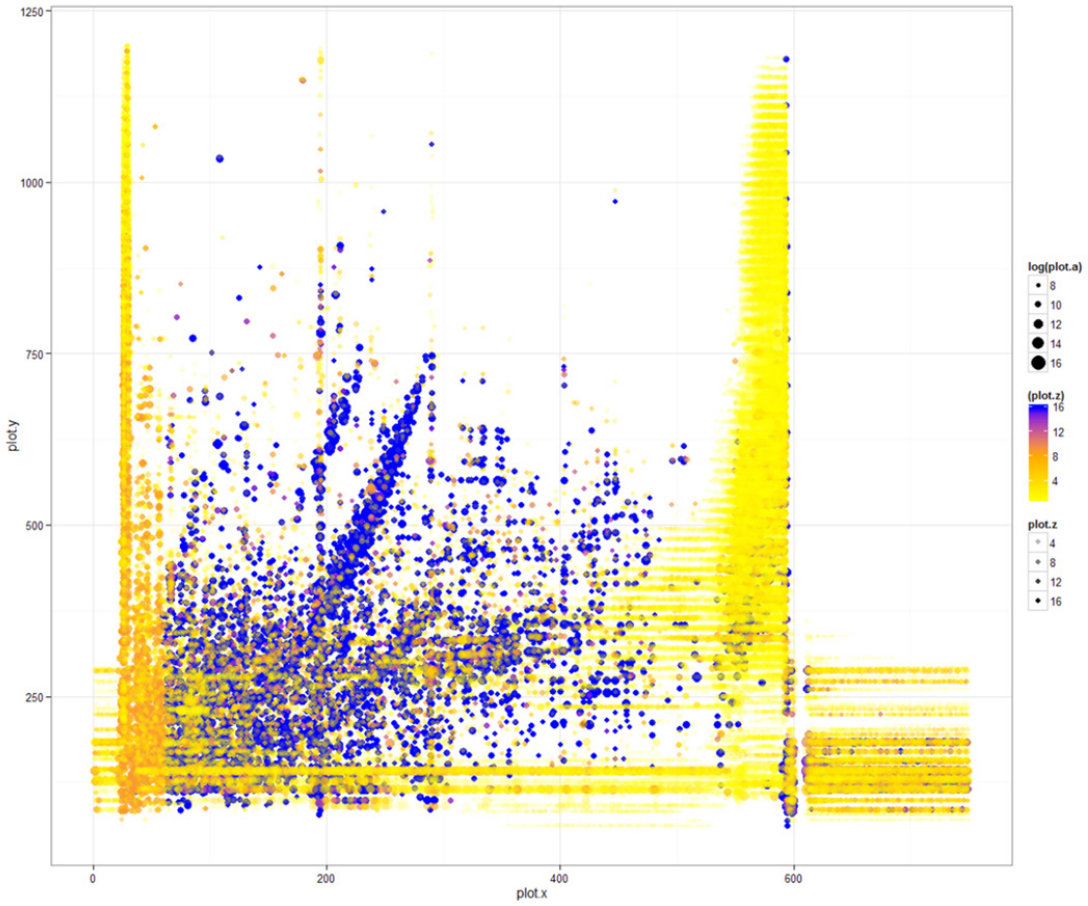


Figure 4-20. Two dimensional plot of feature groups created using the ROgroup method on the first 16 SR urine analyses from the test dataset. The mean m/z and RT values for each group are plotted as a single dot per group, sized by the mean of the group intensity, and colored by the number of features that compose the group (chain). Groups of low chain length (yellow) are observed to co-localise with areas of expected chemical noise, including the organic solvent wash and equilibration steps in the latter half of the analysis.

4.7 Comparison to existing techniques

4.7.1 Creation of a synthetic dataset

In order to benchmark the performance of the grouping approach detailed above (hereafter referred to as the ROgroup method) in the context of other methods, the results of matching must be compared to a known true answer (or "ground truth") so that performance can be quantified (Lange et al., 2008). However, it is not straightforward to deduce ground truth for a real profiling dataset with thousands of features per sample. Some public proteomic and metabolomic datasets do exist where supplemental

analysis and feature identification have provided some knowledge of which peaks across samples share the same identity (Lange et al., 2008). However, the small molecule datasets are based on analysis of plant extracts using larger particle size columns and longer analyses (HPLC separation), and therefore are not representative of the data produced by high throughput UPLC-MS profiling of biofluids from molecular epidemiology sample sets. Furthermore, the feature selection process used in the creation of the datasets was specifically inclusive to features that showed limited deviation. Taken together, it is clear that these datasets are neither representative of the challenge at hand with the separation of human biofluids nor do they reflect the state-of-the art with respect to UPLC-MS profiling. A more appropriate benchmarking solution is therefore needed.

Computational synthesis of a dataset has a number of advantages over the use of organically created data. First, establishing ground truth is applicable to all features, precluding selection bias, and is guaranteed to be accurate. Second, the types of retention time drift observed and reported in Sections 4.5.2 and 4.5.3 can be separately implemented in multiple iterations of the synthetic dataset in order to independently test the effects of random and systematic retention time noise on the accuracy of grouping methods. Finally, the confounding variable of feature extraction ("peak picking") performance across samples may be eliminated. Finally, when seeded with an extracted feature table from a real representative sample, the synthetic dataset remains representative of the biochemical diversity found in human urine.

To accomplish this, a representative urine sample (the 16th SR urine sample from plate 1 of the test set) was extracted using centWave as described above to produce a feature list. The resulting feature list was replicated 99 times to create a dataset of 100 identical feature sets. Random noise (normal distribution with a standard deviation of 0.001 m/z) was applied to all m/z values in order to mimic the variance observed in ToF measurements. From this dataset, four derivative datasets were constructed to represent four types of retention time variance among features:

4.7 Comparison to existing techniques

A. No noise applied (no RT variance)
B. Random noise applied to 100% of features (normal distribution, SD = 0.1 second)
C. Run-order correlated logarithmic retention time drift applied to approximately 20% of features
    a. Approximately 10% elute increasingly earlier with run order
    b. Approximately 10% elute increasingly later with run order
D. Both random noise (100% of features) AND run-order correlated RT drift (20% of features) applied

To mimic the retention time drift observed in the test dataset, the natural log was taken of the series of numbers from 1 to 100, producing a maximum drift of 4.61 seconds. The pattern was then added to or subtracted from the retention times of all features in a selected feature group to produce features that migrate (earlier or later in retention time) with increasing run order. The retention times of a single selected feature with both run-order correlated drift and random noise applied across all 100 samples are shown in Figure 4-21. Twenty percent of independent feature groups were selected for the application of run order noise. To mimic the observed migration of feature clusters, group clusters were defined and twenty percent were selected for the application of run order noise. These clusters were defined using the "kde2d" function of the MASS package in R (Venables et al., 2002) in an effort to orthogonalise the clustering process to the approach used in the ROgroup algorithm. Two-dimensional kernel density estimation (RT bandwidth = 10, m/z bandwidth = 0.5) was applied to locate areas of high feature density within the dataset, and a threshold of 1e-5 was used to define cluster-containing regions. Assignment of drift direction (earlier or later elution with run order) was random among the independent features and clusters selected for application of drift noise.

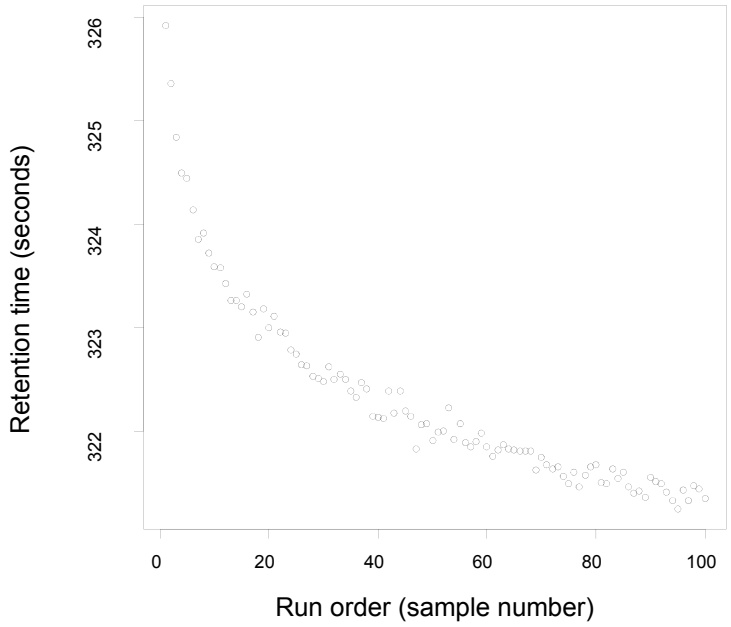


Figure 4-21. The programmed logarithmic drift in retention time for a given feature as a function of run order was applied to a feature from the dataset, along with random noise, to produce a realistic retention profile across the sample set.

No global shifts in feature retention time, linear or otherwise, were added to the data, precluding the need for non-linear dewarping of each chromatogram and allowing a direct comparison of grouping methods only. All other descriptive data produced for features detected by centwave (e.g. peak shape information) were nullified by setting all values in the dataset equal to the total median value observed. In this manner, the ability of each method to group features was based solely on performance using the basic parameters. Furthermore, the highly controlled nature of the experiment precludes intensity variance (biologically or analytically derived) that can manifest as a difference in the presence or absence of a given feature, and also precludes variance or artefacts introduced by the peakpicker itself (erroneously failing to detect a feature that is present in the raw data). As a result, the grouping methods only have to overcome the various types of retention time variance introduced.

4.7 Comparison to existing techniques

4.7.2 Means of assessing grouping method performance.

The grouping of features from the synthetic dataset described above was evaluated using the "density" and "nearest" methods within XCMS and the ROgroup method developed herein. In each case, the user-specified m/z error window was set to the commonly used value of 0.01 m/z (Liu et al., 2012, Vaughan et al., 2012), which is only slightly smaller than the value suggested by Patti *et. al.* for UPLC-ToF data (Patti et al., 2012a) and more appropriate for the amount of noise introduced here. Given that the maximum drift in retention time for any given feature within the dataset was made to be 4.61 seconds +/- the random noise, a threshold of 6 seconds (0.1 minutes) was chosen for the retention time grouping in the nearest and ROgroup approaches. As the density method requires a bandwidth setting in place of a window, the recommended value for UPLC-ToF data was used (bw = 2) (Patti et al., 2012a).

The resulting feature groups (observed groups) were classified using two distinct criteria. The first criterion is the completeness of a group, referring to the number of features grouped across the dataset. If a group contains one feature for every sample, that group is said to be complete. If less than 1 feature per sample are grouped, the observed group is incomplete. Within either of those group types, if the identities are all grouped features are matching, the observed group is said to be homogeneous, and therefore accurate grouping has occurred. However, if the identities are mixed, indicating the erroneous matching of features, they are said to be heterogeneous. These observed group types may be summarised as follows:

1. Homogeneous complete
2. Homogeneous incomplete
3. Heterogeneous complete
4. Heterogeneous incomplete

One approach to summarising these data, and therefore the performance of each grouping method tested, is by calculation of precision and recall as adapted for the assessment of feature grouping outcome by Lange and colleagues (Lange et al., 2008). Precision, when used in this context, is an

assessment of the quality of the composition of observed groups, and is penalised by heterogeneity within groups across the dataset. Recall, when used in this context, is an assessment of the erroneous splitting of features belonging to a single true group across multiple observed groups, and is penalised where more than one observed group contains a given feature. Perfect precision and recall (a value equal to one for each) is achieved when all groups are homogenious complete, and the number of observed groups equals the expected number of true groups. The calculation of the precision and recall metrics for the assessment of grouping method performance as applied to the grouping outputs of the derivative datasets (introduced in Section 4.7.1) is described in the subsequent pseudo-code. As each dataset is composed of 100 samples, and each sample composed of the same compliment of 12756 features, a total of 12756 homogeneous complete true groups is the expected result of perfect grouping. Because of this design, a given feature and the true group to which that feature belongs are both referred to as $\boldsymbol{x}$ (*i.e.* feature 18 in each sample belongs to true group 18).

- For every true group (for each $\boldsymbol{x}$ in 12756), determine which observed groups contain at least one instance of feature $\boldsymbol{x}$. Store the identity of these groups as $\boldsymbol{A}$.
    - Determine the total number of features (number of non-NA values) in all observed groups in $\boldsymbol{A}$. Store this value as $\boldsymbol{B}$.
        - For each $\boldsymbol{x}$, append values of $\boldsymbol{B}$ to create vector $\boldsymbol{C}$.
    - Determine the ratio of true groups containing at least once entry of feature $\boldsymbol{x}$ (1) to the number of observed groups containing at least one entry of feature $\boldsymbol{x}$. Store this value as $\boldsymbol{D}$.
        - For each $\boldsymbol{x}$, append values of $\boldsymbol{D}$ to create vector $\boldsymbol{E}$.
- Calculate precision and recall as follows:
    - Precision = sum(100/$\boldsymbol{C}$)/12756
    - Recall = sum($\boldsymbol{E}$)/12756

Precision is therefore the mean of each ratio between the expected number of features per group (100, from 100 samples) and the observed number of features that constitute all observed groups containing

feature ***x***, for each true group ***x***. Recall is the mean of the ratios of the number of true groups (1) to the number of observed groups (1 or more) for each feature ***x***.

The density grouping method has the unique ability (of the three methods tested: density, nearest, and ROgroup, and representative of any global binning-based approach) to further complicate this scheme, as it can erroneously combine multiple features from the same sample within a feature group (herein referred to as overgrouping). As a result, it is possible for this method to produce a lower number of observed groups than true groups. Because of this substantial difference in output, precision and recall was not calculated for the output of density method, but the number of groups containing more features than there are samples is reported as an independent value in addition to the reporting of complete/incomplete homogeneous/heterogeneous groups.

4.7.3 Results

The results of applying three grouping methods (density, nearest, and ROgroup) to the four datasets previously generated (A-D, described in Section 4.7.1) have been tabulated in the series of tables below. The first of these, shown as Table 4-3, reports the total number of feature groups produced by each method relative to the expected "true" number of groups (12,756). For example, grouping of dataset A (no retention time noise or drift applied) yields a dataset that is nominally 100% the expected number of groups when using the nearest and ROgroup methods. However, the density method produces a smaller number of groups than expected (84%), regardless of the dataset used, indicating overgrouping. The nearest method, when applied to dataset D (retention time noise and run order drift applied) returns datasets with 125% and 131% the number of expected groups, depending on the order of analysis. The ROgroup method under the same circumstances returns a number of groups which is closer to the true value, at 101% for run order analysis and 106% for random order analysis.

dataset	density	nearest (run)	nearest (random)	ROgroup (run)	ROgroup (random)
A	84	101	101	100	100
B	84	103	103	100	100
C	84	118	125	100	105
D	84	125	131	101	106

Table 4-3. Volume of feature groups produced by each grouping method on each dataset (A-D) expressed in terms of percentage of the true number of groups (12,756).

The composition of these groups is presented in Table 4-4, stratified into the four classes of group type introduced in Section 4.7.2 and a separate overgrouped classification unique to the density method. It is immediately clear that the density method fails in its ability to form homogenous complete groups in comparison to the other methods. This is in part due to the presence of overgrouping creating groups with more features than samples. The focus of group quality is therefore split between the nearest and ROgroup methods. Both achieve similar results with the simplest dataset (A). However, as the datasets become more complicated with the introduction of random retention noise (B) and run order drift (C), or both (D), characteristic behaviour is observed. The nearest method tends to produce a larger number of incomplete groups (both homogenous and heterogenous) than the ROgroup method, while the ROgroup method tends to produce more groups that are complete but heterogenous. The ROgroup method outperforms both the nearest and density methods in terms of its ability to produce homogenous complete groups.

4.7 Comparison to existing techniques

Group Type	Dataset	Density	Nearest (run)	Nearest (random)	ROgroup (run)	ROgroup (random)
Homogeneous (complete)	A	7391	12738	12738	12738	12738
	B	7381	12680	12681	12647	12650
	C	7297	11834	11267	12621	12210
	D	7297	11429	10998	11907	11565
Homogeneous (incomplete)	A	1007	17	15	0	0
	B	1004	100	98	0	0
	C	1095	2233	2946	3	443
	D	1093	1909	2231	51	444
Heterogenous (complete)	A	0	0	0	18	18
	B	0	0	0	109	106
	C	1	60	27	134	335
	D	3	24	17	809	904
Heterogenous (incomplete)	A	399	117	120	0	0
	B	405	318	319	0	0
	C	399	945	1754	0	382
	D	402	2640	3411	62	581
Overgrouped	A	1896				
	B	1896				
	C	1961				
	D	1964				

Table 4-4. Tabulation of observed feature groups (of 12756 true groups) created by applying three feature grouping methods to four synthetic datasets. Each dataset is characterised by distinct retention time variance (A = no noise or drift applied; B = normally distributed noise; C = run order drift; D = both normally distributed noise and run order drift). Nearest and ROgroup methods were run in either the same order as the data files were acquired (run) or in random order (random). As the density method can only analyse datasets as a whole, the order of analysis is not relevant. Overgrouping occurs where there are more features per group than samples, and is only possible with the density method.

To assist in the interpretation of grouping quality and therefore method performance, the accepted metrics of performance and recall were calculated for each method/dataset pair. The results are shown in Table 4-5. While both methods perform with high scores (likely a symptom of the highly controlled conditions present within the datasets), the ROgroup method outperforms the nearest method when analysed in run order on the most complex and therefore representative dataset. For both methods, analysis in run order improves the precision and recall scores obtained on datasets C and D but not A

and B, indicating the generality of performance enhancement when considering run order where retention drift is present.

Group Type	dataset	nearest (run)	nearest (random)	ROgroup (run)	ROgroup (random)
precision	A	**0.999**	**0.999**	**0.999**	**0.999**
	B	**0.998**	**0.998**	0.996	0.996
	C	0.980	0.963	**0.995**	0.978
	D	0.964	0.950	**0.965**	0.947
recall	A	**0.999**	**0.999**	**0.999**	**0.999**
	B	0.995	0.995	**0.996**	**0.996**
	C	0.949	0.919	**0.995**	0.973
	D	0.923	0.898	**0.964**	0.941

Table 4-5. Precision and recall values for the nearest and ROgroup methods of feature grouping across datasets A-D. The maximum values per dataset are highlighted in green.

4.8 Discussion and conclusions

4.8.1 Chromatographic retention time variance and grouping method performance

The overall goal of alignment and feature grouping is to overcome variations within a set of raw data to produce an accurate collated dataset for further analysis. The challenge in doing so is in maintaining the excellent chemical specificity produced by high resolution UPLC-ToF MS despite confounding imprecision in measurement. While the methods developed in Chapter 3 aim to maximise chromatographic precision among sequential analyses, a degree of both random and systematic variance is still observed. Careful evaluation of the variance in retention time of features across the test dataset has revealed the feature-specific nature of retention time deviation and furthermore demonstrated that such migration (where present) is encoded in the run order of the analysis. The global deviations that developers of alignment and grouping algorithms sought to correct with whole-chromatogram de-warping methods are simply not a substantial source of variance in high precision UPLC-MS analysis. In addition to the overall improvements in chromatographic reproducibility, this is perhaps due to the pace at which UPLC analyses are able to be conducted. Whereas fluxuations in the chromatographic system and laboratory environment over the course of an hour may have yielded

nonlinear and nonspecific perturbations in the chromatographic retention of features across a single 1 hour HPLC chromatogram, the same deviation would be spread across four 15 minute UPLC chromatograms and therefore encoded in the analysis order.

These results warranted the construction of a synthetic dataset whereby the feature content of a single representative urine sample was replicated to create an experiment of 100 synthetic samples by introducing mass error, retention time error, and run order retention drift among those features. Of the datasets created, dataset D is the most representative of the observed data and therefore the most useful in assessing the potential performance of grouping methods on a real dataset obtained during large-scale phenotyping analysis. Leveraging this knowledge and performing grouping analysis in the same order as the synthesized run order improves both the precision and the recall of the nearest and ROgroup methods in the datasets with retention drifting features (datasets C and D). However it is the ROgroup method which yields the most accurate dataset in terms of overall size (101% for dataset D by ROgroup in run order) precision (0.965), and recall (0.964). These results suggest that the matching approach utilising feature clustering and ambiguous match resolution by whole-community alignment is a viable if not superior grouping strategy. Meanwhile the density method, as a representative of all grouping methods which consider the entire dataset at once and perform grouping by cross-sectioning the dataset, fails to accurately group the complex profiling data even in the most simple case where no RT noise or drift is applied (dataset A).

4.8.2 Potential for application in real-time

Application in real time is important for a true high throughput system, as existing methods of data pre-processing can in some cases rival the amount of time required to acquire the raw data. The ROgroup script is intended to be run iteratively following the completion of each new sample analysis. In the first instance, the time required for this procedure is slightly longer than in subsequent iterations as both the first and the second sample datafiles must be peakpicked prior to feature matching. However in subsequent rounds, the candidate feature set is repurposed as the new template feature set and only the newly acquired sample is peakpicked for pairwise matching. It should be noted that any method which is capable of analysing samples in the order in which they were analysed may be applied

in a real time manner. While this excludes approaches such as the density method or methods that suggest retrospective selection of a reference chromatogram (Chae et al., 2008), it does not necessarily exclude the join aligner/nearedt method tested here. This approach could be modified for application in real time where the first sample analysed seeds the master template, and all subsequent samples are matched in to the master averaged feature list. However it is important that the speed at which the method performs grouping across the entire analysis is less than the data acquisition time.

While both the nearest grouping method and the ROgroup method are capable of matching features between two samples in fewer than 15 minutes, the manner in which the time required for matching increases with repeated analyses must be considered. With each new set of features grouped using the join-aligner/nearest method, unmatched features are appended to the bottom of the master feature list. As this list grows longer, more time is required for each subsequent match. The rate of appending unmatched features is related to the prevalence of noise in the extracted feature set, as random noise would not be expected to match into existing groups. Therefore, with no inbuilt noise detection capabilities at the grouping stage, these noise-derived features accumulate in the master list. When applied across thousands of samples, the time required for data acquisition may rapidly outpace the time required to match in new features into an unwieldy master list. The use of true pairwise matching (between two samples, rather than one sample and a master list) ensures a more constant processing time, as approximately the same number of features are matched (assuming reasonable biofluid homogeneity) in each iteration. The master list in the ROgroup method is simply a repository for matched data, and does not participate in matching itself. This allows the process to scale linearly, and therefore does not impose a practical limit on the number of sequential analyses that can be performed and processed.

It is worth noting that the length of the total repository may also be kept in check by the ROgroup method on a sample-by-sample basis when recognizing and eliminating feature clusters of excessive size, indicating the presence of features derived from constant background signals. In this manner, the grouping process becomes an active noise removal step rather than a passive collation or process by

which noise is specifically introduced (*e.g* the 25% erroneous excess of feature groups produced by the nearest method operating in run order on dataset D).

4.8.3 Application to datasets with increased biological variance

The approach outlined herein was tested on replicate injections of a pooled reference (SR) sample and on synthetic data created by modulating the measurements of features from a single sample. However, to be of use in the large-scale screening of biofluid samples from human populations, the performance of such an approach must be considered in the context of a variable sample matrix. A weakness of the "true" pairwise approach is evident when considering that a complete feature group across 1000 samples is only achieved when the matching process completes successfully 999 times without error. Should a feature fail to be detected in the peak picking process, either erroneously or because it is below the limit of detection in the raw data of a single sample, its absence will inevitably break the chain, disjoining the growing feature group. This is a consequence of having no means by which features in the following samples can be connected to past chains, as would be possible using an iteratively updated master template. In this manner, a single dilute sample of urine could potentially break all feature chains, disrupting the coherency of the dataset on a broad scale.

Future work therefore includes the creation of a post-acquisition phase of feature grouping which connects broken feature chains. This task will greatly benefit from the longitudinal observations in m/z and retention time variation captured in chain fragments, as these may provide guidance for seeking matching features or other chain fragments that are from samples farther downstream in the analysis run order. A proposed mechanism is one where the longest incomplete chain in the dataset is analysed to determine m/z and retention time error boundaries (e.g. based on standard deviations of variance within the observed data) as well as model any systematic retention time trajectory into the preceding or subsequent samples. Armed with this knowledge, an area of high probability for finding a match can be extrapolated into neighbouring samples subsuming features with matching values. The calculation could be iterative to reduce the "distance" of extrapolation required. The entire process could be repeated, targeting the next longest incomplete feature chain and so on until the dataset has been depleted of chainable features. Such a mechanism would maximally leverage the value captured

in the known trends and distributions of formed chains to efficiently deplete the residual dataset. The remainder could be either discarded, or further grouped by a method suitable for application in a sparse dataset (*e.g.* the density grouping method). For very rare signals (e.g. some xenobiotics), a targeted approach to exploration of profiling datasets for these chemicals may be the best course of action. Finally, while this step is intended to occur after acquisition, it is hoped that such a solution could be coded efficiently to "clean" the data without adding substantial processing time and compromising the desired benefits of real time analysis.

Chapter 5: In-solution databases to facilitate rapid metabolite identification in metabolic profiling studies

Objectives

1. To construct a large scale mixture of chemical reference compounds to use as a potential surrogate for pure reference materials in the identification of metabolites.
2. To develop and implement a scripted method for the automatic deconvolution of a known complex mixture of chemical standards using tandem mass spectrometry.
3. To demonstrate the feasibility of prospective metabolite identification in an example metabolite profiling application.

5.1 Introduction

Chapter 3 describes the development and application of LC-MS methods that have been optimised for the purpose of large-scale profiling. It is hoped that the quality produced by these methods and the efficiency their use confers to the laboratory facilitate their establishment as standard approaches, warranting intensive characterisation to support rapid molecular assignment of the data produced. Broad annotation of these methods with individual chemical reference standards would further increase their value, but would require a substantial investment in both cost and time. If such an investment were to be made, it would create an energy barrier to the pursuit of incremental methodological performance gains achieved by introducing subtle changes in the chromatographic separation requiring the repetition of these laborious processes. To overcome this barrier, advancements to methodology should yield substantial savings in time and/or cost that are greater than the value of the investment in method characterisation. This scenario is inevitable, as technologies are constantly being developed to facilitate chromatographic separations with higher performance and higher efficiency. The advents of sub-2 micron stationary phase particles and ultra-high performance chromatographic systems, core shell columns, nano-scale chromatography, and novel mobile phase dopants (*e.g.* (Yanes et al., 2011)) have all challenged previously established methods. Looking forward, it is therefore prudent to address the inevitability of change by proposing a strategy for quickly

annotating the signals produced by UPLC-MS methods, enabling the further development of such methodology.

Early in the development of the chemical reference mixtures described in Chapter 3, it became apparent that unambiguous deconvolution of a mixture of known composition is possible using the mass spectrometric data of accurate mass, provided that the mixtures did not contain multiple isobaric species. It is expected that the use of molecular fragmentation by CID would further extend the ability to deconvolve a known mixture. The approach of spectral matching is certainly not novel, as one need not look further than the popular AMDIS software (Stein, 1999) from the National Institute of Standards and Technology (NIST, Gaithersburg, Maryland, USA) for GC-MS analysis. However, its application has been largely limited to spectra obtained by fragmentation-intensive ionisation (*e.g.* electron ionisation) and in practice appears to be limited to the annotation of unknown features in a sample matrix. As the content of complex biofluid matrices is variable, annotations by this method are always tentative and require comparison to a chemical reference standard for authentication (Sumner et al., 2007).

However, it is hypothesised that if such an approach were applied to the unambiguous deconvolution of a finite mixture of known composition, those synthetic mixtures could be then be efficiently used in place of individual reference standards in facilitating metabolite identification. Using this approach, it should be possible to build chemical libraries in solution and rapidly convert them to *de novo* databases with inherently accurate retention time. By broadly capturing a large number of reference materials at the time of the original profiling analysis, the data required for metabolite identification could be captured *prospectively*, precluding the need to return to the instrument at a later date and attempt to match the original system conditions to perform metabolite identification work. It is therefore of interest to determine the extent to which chemical reference standards can be combined, unambiguously deconvolved, and applied to metabolite identification in profiling studies. In addition to application in routine analysis, the envisioned approach should serve to lower the barrier to method characterisation, facilitating and potentially directly contributing to method development (the latter being true if the level of success in generating a *de novo* database from an in-solution library was used

as a metric to assess changes in chromatographic conditions). The envisioned workflow for both of these potential applications is outlined in Figure 5-1.

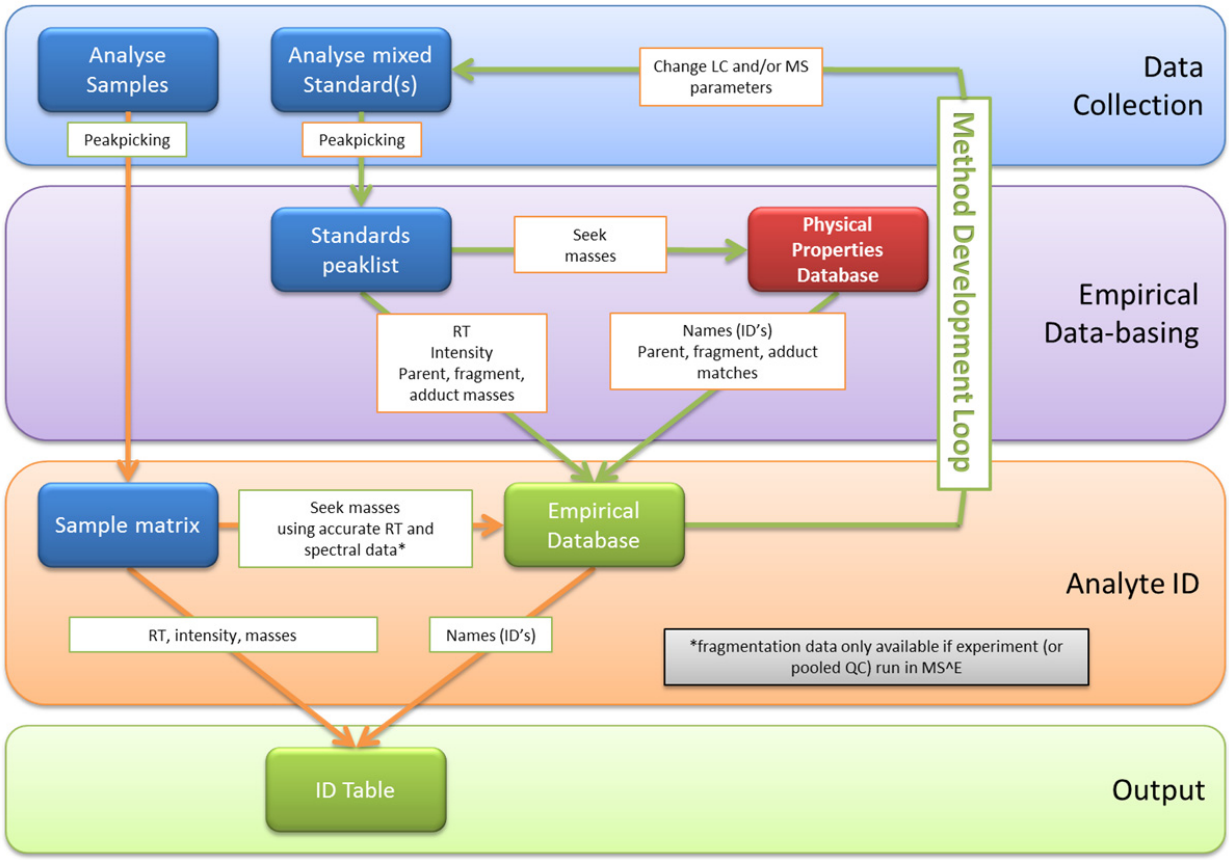


Figure 5-1. Envisioned workflow for the use of an in-solution database (mixed standards) and known physical properties (e.g mass spectra) of the mixture components (physical properties database) for the *de novo* generation of an empirical database. The empirical database can then be used as a profile of known reference material to facilitate metabolite identification of unknown features in a sample matrix derived from a biological study, generating an ID table. Alternatively, the extent of success in building the empirical database can be used as a metric for UPLC-MS method development, effecting changes in LC or MS conditions that produce better profiling of the standards contained in the mixture.

5.2 Methods

All work contained herein was performed concurrently with the method development documented in Chapter 3, and therefore all UPLC-MS analyses have been performd using the established method of Want *et. al* (Want et al., 2010) rather than by the method optimised for large scale analysis. However, as the approach aims to create *de novo* databases for any given chromatographic method, the results obtained are representative and easily extendable.

5.2.1 Preparation of individual chemical standards and standards mixtures

A "shotgun" approach for the creation of standards mixtures was adopted by selecting approximately 400 reference chemicals at random from the library of chemical reference materials maintained by the Division of Computational Systems Medicine. These materials were prepared as described in Chapter 3 (Section 3.3.2). Briefly, individual chemical reference solutions were made in a qualitative manner by aliquoting a small spatula scoop (for solids) or drop (for liquids) of the chemical (of reasonable purity, generally exceeding 95%) to a clean storage tube and diluting with 5mL of ultrapure water. The approach resulted in concentrated solutions for water soluble chemicals. Incomplete solubility was observed for some chemical preparations despite vortexing and sonication at room temperature (for a maximum of 30 minutes). Regardless, each solution or suspension was pipetted into an individual well of a 96-well deep well plate. This process was repeated until four plates were produced. Plate-wise mixtures were then prepared by combining all individual solutions from a single 96 well plate in equal parts to yield four distinct mixtures (each containing a maximum of 96 standards). Those mixtures were then combined in equal parts to make a master mixture. All individual standards and standard mixtures were frozen at -80° C.

5.2.2 Acquisition of reference spectra

Frozen plates of solubilised reference standards were transported on dry ice to Waters Corporation small molecule profiling laboratory (Milford, MA, USA) for characterisation using their in-house instrumentation. Prior to the analysis of each individual plate, the standards were thawed, transferred to new 96 well plates and diluted 1:100 (v/v) with water. The diluted standards plate was held at 4° C for the duration of the analysis. Using an Acquity UPLC system (Waters Corp., Milford MA, USA), a

flow injection approach was initially attempted in order to present each standard for multiple mass spectrometric measurements over a prolonged period of time (20 to 30 seconds). However it was observed that sample desalting by slight retention in a reversed-phase system was beneficial for the separation of sodium and other adduct-forming cations found in many standards from the analyte of interest. Therefore, brief separations were performed using an Acquity 2.1 x 50 mm HSS T3 column held at 30° C. Five microliter injections were made for the analysis of each standard. Solvents used for the separation were water + 0.05% formic acid (A) and acetonitrile + 0.05% formic acid (B). A rapid loading and elution program (Table 5-1) was designed to retain analytes briefly and elute them in broad peaks to facilitate the collection of many targeted mass spectrometric measurements.

Time (minutes)	Flow Rate (ul/min)	% Solvent B
Initial	200	0
.5	200	0
1.5	500	100
2.5	500	100
2.6	500	0
3.0	200	0
3.5	200	0

Table 5-1. A rapid elution program for the high throughput generation of reference spectra from chemical standards.

Eluate was directed to a Synapt G2 Q-ToF HDMS mass spectrometer by means of an electrospray ionization interface operating in either the positive or negative ion mode. Full scan MS data were acquired for the mass range between 30 and 1200 m/z. Four additional interleaved channels of data were simultaneously acquired using the instrument's data dependent analysis (DDA) function whereby the most intense feature(s) in an MS scan are automatically targeted for MS/MS analysis. In this manner, analyte-derived features could be conveniently targeted for MS/MS analysis with no prior knowledge or assumptions of the spectra generated by each reference material. Feature targeting in DDA was triggered by the intensity of a given spectral signal exceeding 1000 counts and continued for five scans before a new target mass was selected. During CID fragmentation of selected targets, a ramp of collision energy from 10 V to 50 V was applied per scan ensuring a broad representation of

fragmentation spectra from both labile and collision resistant molecular species. Data were acquired in continuum mode at approximately 18,000 resolution (FWHM), and with approximately 1 ppm mass accuracy following calibration with a sodium formate solution. A solution of leucine enkephalin was infused via a secondary orthogonal electrospray probe to provide a known mass value for adjustment of the calibration over time (known as the *lock mass*).

5.2.3 Generation and interpretation of a spectral library

Spectral data from a subset of the total number of reference standards, drawn from a single plate, were manually interpreted and peaks related to each known standard were recorded in database format (Appendix 4). Briefly, a monoisotopic mass [M] was calculated from each molecular formula, ignoring the contribution of salts and water molecules in hydrates, using ChemFolder v 12.0 software (ACDlabs). This mass was then adjusted to reflect the theoretical gain or loss of a proton resulting in positive [M+H]+ and negative [M-H]- ionization, respectively. The expected ion mass was then extracted from the corresponding standard's total ion intensity chromatogram (TIC). Where the expected peak was observed, the intensity-DDA functions were opened for inspection. In most cases, the ion mass was present with sufficient intensity to trigger quadrupole selection and CID fragmentation yielding a fragmentation spectrum. These spectral data were combined for the duration of the targeted aquisition, mass corrected to the LockMass signal using the "Automatic Peak Detection" setting in the MassLynx software, and deisotoped using the "ToF Transform" function with the mass range set to 30-1200 and a maximum charge state equal to one. Both absolute and relative intensity thresholds were applied to the resulting spectrum such that the lowest included intensities were greater than 200 counts (for the combined scans) and greater than 5% of the most intense peak present. The resulting spectral masses were recorded in the database as CID fragments of the targeted molecular parent.

Data obtained from selected CID fragmentation of non-parent ion masses were extracted in the same manner and subjected to interpretation before inclusion in the database. MS features that match the expected ionised mass of the reference chemical were classified as *parent* ions. Fragments specifically originating from MS/MS of parent ions were classified as *fragment* ions. Where an ion species with

mass greater than that of the standard was found to fragment to the original standard ion mass, the targeted mass was recorded as an *adduct*. Fragments of adduct masses that were greater than the original parent mass, or less than the parent mass but not observed to be fragments of the original parent mass, were recorded separately as *adduct fragments*. In the event that a distinct chromatographic peak was detected but the masses not immediately relatable to the expected parent mass or a common adduct (ie. [M+Na]+), the data were not included for further consideration. The spectral library produced is hereafter referred to as the *physical properties database*.

5.2.4 Development and implementation of a deconvolution script for the assignment of mixed standards

A simple script for the deconvolution of standard mixtures and *de novo* generation of mass and retention time-containing databases was coded in the R programming language for the freely available R software environment. The script is intended to match data from a paired physical properties database and standards mixture, generating a chromatographic method-specific empirical database, complete with theoretical mass spectra and empirical retention times. A visual algorithm summarizing the method is represented in Figure 5-2, and the full script is available in Appendix 5. Briefly, the script seeks matches between the recorded m/z values in the physical properties database and those observed in the XCMS-generated peak list from the UPLC-MS analysis of the reference mixture using a user-selectable window of m/z error (set generously here as 100 ppm). Despite the potential for multiple features in the empirical dataset to match any one reference standard m/z value in the database, a consensus (or best) match is established when all (or the majority) of the matched features share the same retention time. Where a consensus or best match can be determined, the standard is reported to an *empirical database*, combining the recorded spectrum from the physical properties database with the observed retention time from the UPLC-MS peak list.

In order to test the automated deconvolution script, UPLC-MS was used to profilg the master mixture, the subset mixture, and all individual components of the subset mixture using the chromatographic method described by Want *et. al.* (Want et al., 2010) on an Acquity UPLC mated to a Q-ToF Premier (Micromass/Waters Corp., Manchester UK) and electrospray interface operating in the negative

ionization mode, which was shown to yield more peaks than initial screening with positive mode detection. The standards mixtures were profiled using a low-to-high CID collision energy ramp (20 to 40 eV / scan) to obtain both parent and fragment information in each MS scan. MS data was collected across the range of 40 to 1000 m/z.

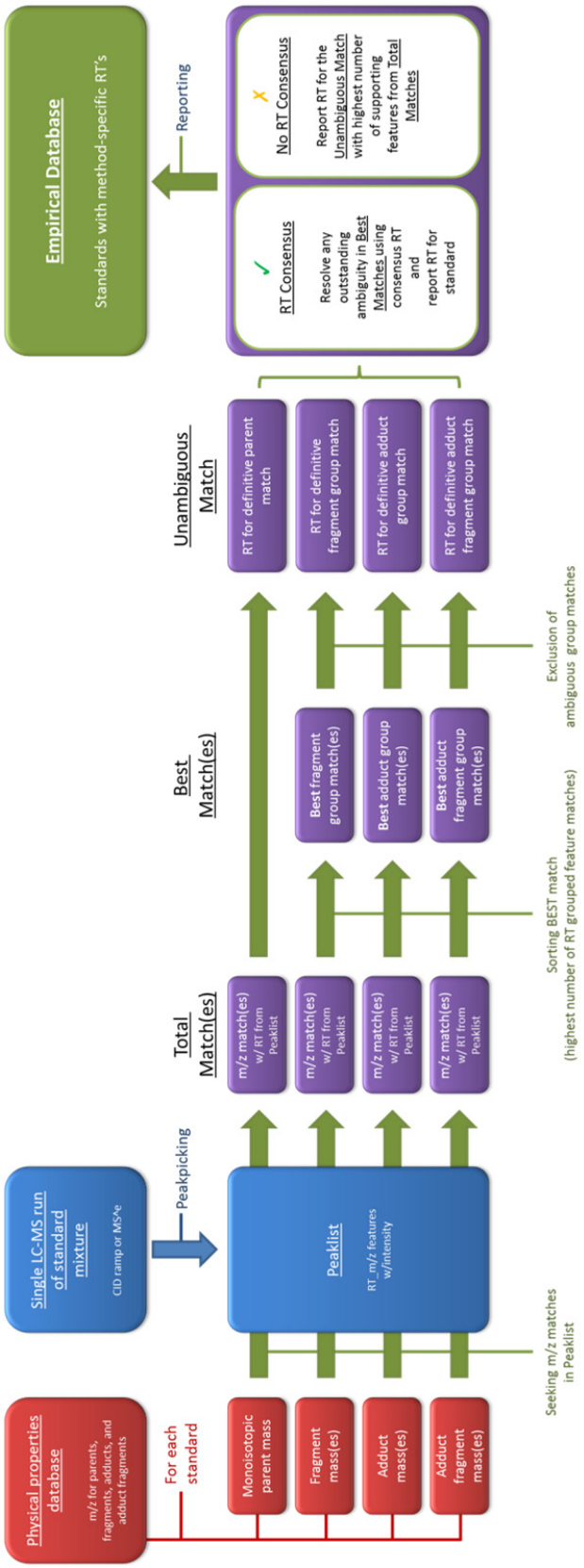


Figure 5-2. Visual algorithm of the deconvolution workflow.

5.2 Methods

5.2.5 UPLC-MS profiling of human urine from a bariatric surgery subject cohort and standards mixtures

Paired urine samples were collected from 57 patients immediately before and approximately 2 months after receiving bariatric surgery (114 samples total). A composite urine sample (QC) was prepared for column conditioning and quality control as described by Want *et. al.* (Want et al., 2010). In addition, a QC dilution series (dQC) was created by serial dilution of the QC sample (in steps of 1:1 volume-to-volume dilution with water) for the purpose of data filtration as suggested by Cloarec et. al. (Croixmarie et al., 2009). Extrapolating from this concept, group-specific composite samples (gQCs) were prepared for the separate pre- and post-surgery groups, allowing cross dilution of potential biomarkers and facilitating molecular assignment by signal intensity-directed targeted MS/MS (data dependent analysis, or DDA), ensuring a higher concentration of group-specific biomarkers (i.e. metabolites that are significantly up or down-regulated are easier to identify from the pooled sample where all samples are exhibiting higher concentration). Cross dilution of the gQCs was performed by mixing them in the following ratios: 0:100, 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, 9:1, 100:0. These are collectively referred to as xQC samples. All samples (including the various QC samples except for the highest concentration dQC) were diluted with water 1:1 (volume-to-volume) and centrifuged to remove particulate matter prior to analysis. Profiling was conducted using the RPC UPLC conditions described previously (Want et al., 2010), with an Acquity UPLC mated to a Q-ToF Premier (Micromass/Waters Corp., Manchester UK) and electrospray interface operating in either the positive or negative ionization mode. Dynamic range extension was utilised, expanding the linear range of the detector by collecting and combining both full strength and attenuated strength signals (a correction factor was automatically applied by the instrument software to the latter) to allow for accurate intensity measurements of otherwise saturated peaks. Data was collected across the 40 to 1000 m/z range with a scan rate of 0.2s/scan and a minimum (0.02s) interscan delay. The ESI source capillary and sampling cone voltages were 3000 and 30 volts, respectivly for positive mode inoisation and 2500 and 25 volts respectivly for negative mode inoisation. A source temperature of 120 °C was used, along with a desolvation flow of nitrogen gas at 800L/h and 400 °C.

5.2 Methods

For each analysis, four blank samples were analysed to establish the background signal of the UPLC-MS system, followed by 10 injection of system conditioning using the QC sample (these were excluded from subsequent data processing). This was followed by two QC injections marking the start of the analysis (and therefore included in data prosessing), followed by the randomised analysis of all study samples with a single QC injection interleaved every 11th injection. The final QC injection was followed by the analysis of all 11 xQC samples and DDA of the gQCs (for the non-specific capture of MS/MS information to facilitate metabolite ID). Next, the standards mixtures were profiled using a low-to-high collision energy ramp (20 to 40 eV CID) for each scan to obtain both parent and fragment information. Finally, the 12 dQC samples were analysed, followed by two QC injections and a single blank sample.

5.2.6 Pre-processing, multivariate statistical analysis, and biomarker discovery

Data extraction via feature detection, alignment, grouping, and integration was performed on each dataset (positive and negative inoisation mode) using XCMS as described in Chapter 2. The processing variables used for centWave feature detection are reported in Table 5-2. The grouping function was performed using the nearest method with retention time and mass axis grouping boundaries of 12 seconds and 0.07 Da., respectively. Only feature groups demonstrating a correlated response to the dilution series (dQCs) of 0.8 or greater were passed on to the final data matrix. Finally, to account for the observed variable dilution of urinary contents, median fold change mornalisation was applied to the filtered dataset using the approach of Veselkov *et. al.* (Veselkov et al., 2011). The resulting matrix of feature intensities across all samples and feature groups was imported to SIMCA-P+ v. 12.0.1 (Umetrics, Umeå Sweden) for principal component and orthogonal projection to latent structure discriminant analyses (PCA and OPLS-DA). Pareto scaling was applied to the dataset within Simca.

parameter	RPC-MS
ppm	30
peakwidth	2 to 10
snthresh	10
noise	(default)
prefilter	(default)

Table 5-2. XCMS centWave parameters for feature detection.

5.3 Results

5.3.1 Collection of reference spectra for a physical properties database

The use of a short chromatographic separation, rather than the originally planned flow injection approach, aided in the DDA acquisition of intensity-triggered targeted MS/MS spectra by separating high intensity chemical noise from the reference-chemical derived features. This was most often the case for reference standards obtained as sodium salts, where the sodium cation was observed to form complexes with the formic acid solvent additive, yielding a strong sodium formate signal eluting immediately in the injection peak (Figure 5-3). Brief retention and prolonged elution of analyte species (*e.g.* cytidine 5'-monophosphate shown below) were sufficient to remove most analytes from the injection peak, allowing DDA and acquisition of spectral information related to the analyte including lower intensity multimers and adducts. The information from these analyses were manually interpreted as described in the methods (Section 5.2.3). An example of this process is illustrated in Figure 5-4, using the reference standard cytidine 5'-monophosphate.

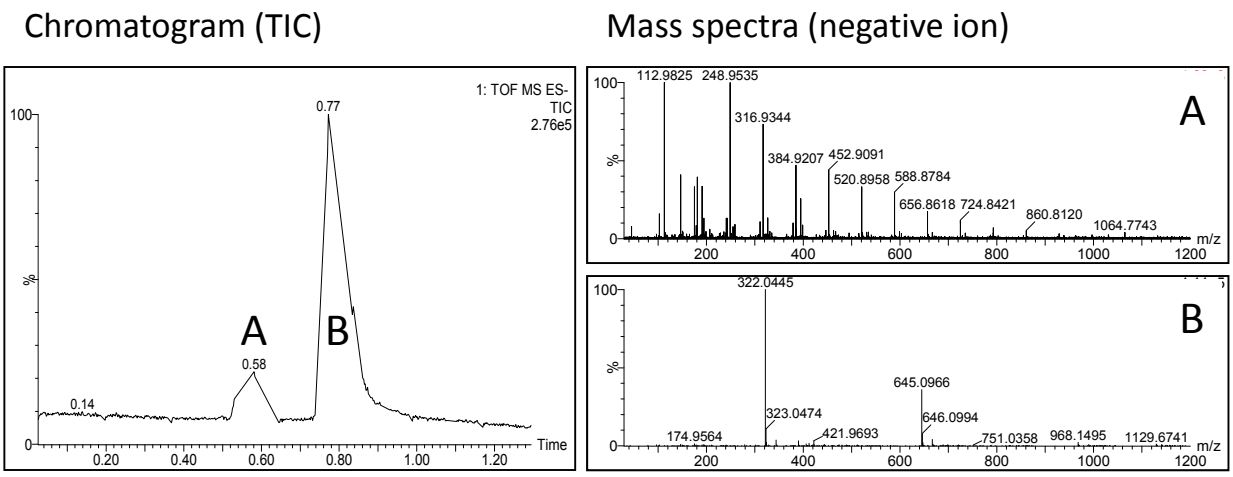


Figure 5-3. Chromatographic separation of a representative standard from potentially obscuring salt. Sodium in the sample (left panel, chromatographic peak A) presents in the mass spectrometer as a sodium formate mass envelope (upper right panel, spectrum A) via in-source combination with the formic acid solvent modifier. It is chromatographically separated from standard analyte cytidine 5'-monophosphate (left panel, chromatographic peak B), resulting in clean MS and MS/MS reference spectra (lower right panel, spectrum B).

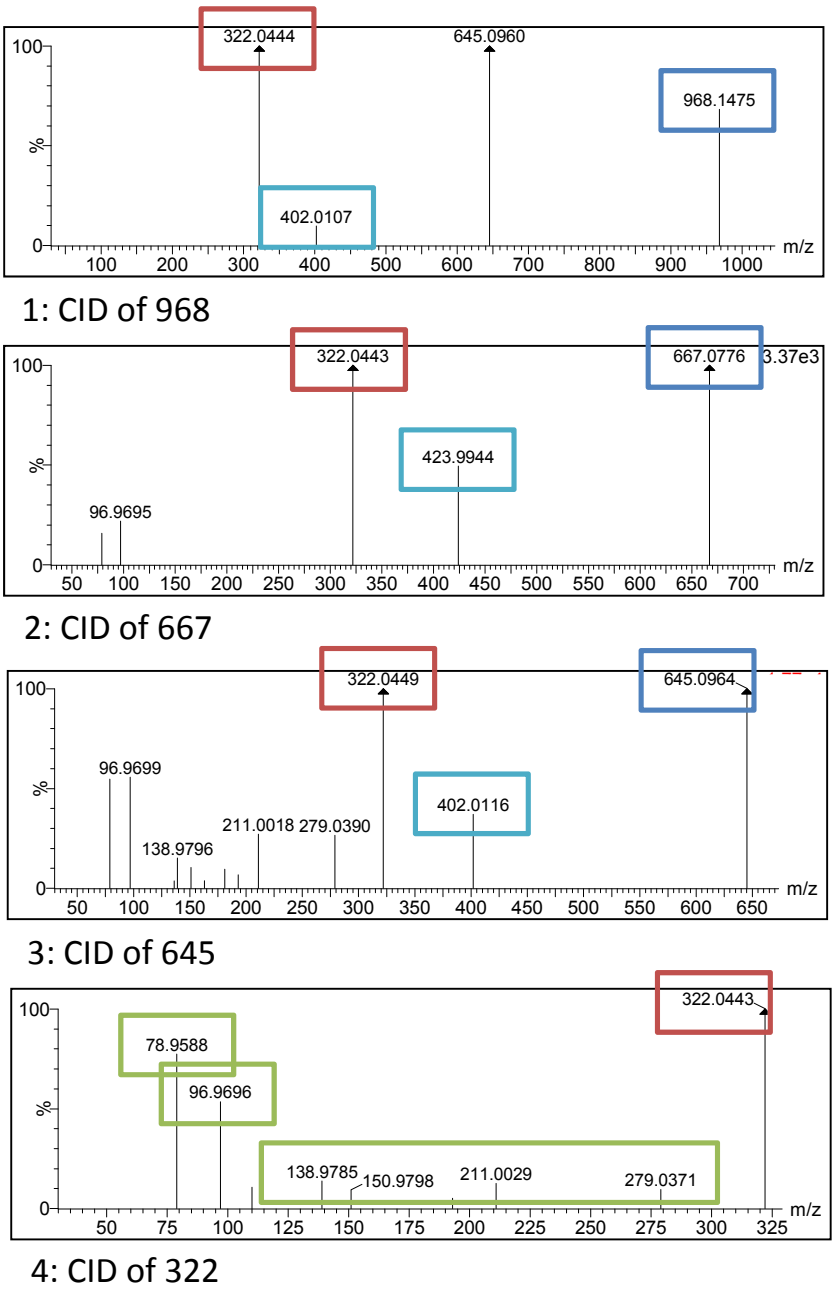


Figure 5-4. Interpretation of spectral data from four interleaved DDA MS/MS acquisitions for the UPLC-MS analysis of cytidine 5'-monophosphate. Species matching the mass of the expected monoisotopic ion for deprotonated cytidine 5'-monophosphate ([M-H]- = 322.0440 m/z) are highlighted in red. The products of targeted fragmentation of the [M-H]- species are classified as fragments and highlighted in green. Species greater in mass than [M-H]- which fragment to [M-H]- were classified as adducts, highlighted in blue. The products of adduct fragmentation which have not been independently validated as adducts themselves, as well as those smaller than [M-H]- which have not been independently found to be fragments of the parent mass, were classified

as adduct fragments (highlighted in light blue). Mass values not indicated by color are redundant with those defined elsewhere in the figure.

Due to the labour intensive nature of this manual characterisation, the DDA MS/MS data were manually interpreted for a subset of 77 standards analysed in negative ionisation mode from a single plate as a test application. Nineteen standards failed to ionize in negative mode when introduced individually via electrospray at high concentration. Those 19, as well as three additional standards with spectra not clearly related to their expected masses, were withheld from the physical properties database and excluded from subsequent analysis. A summary of the spectral information obtained and interpreted for the remaining 55 standards is summarized in Table 5-3. The resulting subset database is included in Appendix 4.

	Monoisotopic parent	One or more fragments	One or more adducts	One or more adduct fragments
# of standards	54	33	44	27
% total (n=55)	98%	60%	80%	49%

Table 5-3. Number of standards (from 55 total) with observed monoisotopic parent mass, fragment mass(es), adduct mass(es), and adduct fragment mass(es).

5.3.2 Implementation and testing of automated mixture deconvolution and annotation

The script developed for the deconvolution of known chemical mixtures and generation of an empirical database was implemented using the UPLC-MS analysis of the 77 standards subset mixture. Selected results from the matching procedure are shown, illustrating the possible outcomes of empirically assigning retention times from the UPLC-MS data to the reference materials with their known spectra in the physical properties database (Figure 5-5). In the case of danylsarcosine (A), a consensus retention time of approximately 6.2 minutes was achieved by the unambiguous matching of the expected parent (red) as well as expected adducts (blue) and fragments (green). In the case of 3-hydroxybutyric acid (B), multiple chromatographic peaks were obtained from the standard with similar mass profiles (perhaps indicating an origin of the material from poly[(R)-3-hydroxybutyric acid]). This was therefore an unusual challenge for the matching script, as each chemical was expected to yield a single chromatographic peak. However, as the greatest number of matching features shared the retention time of approximately 4.5 minutes, that group of features was chosen as the best match. Finally, in a minority of cases, a determination of a consensus or best match was not possible due to complete ambiguity such as that seen in the matching of (S)-2-hydroxybutyric acid, where multiple matches are found for the parent mass and an adduct, producing three equally plausible retention times that match both the parent and adduct m/z values.

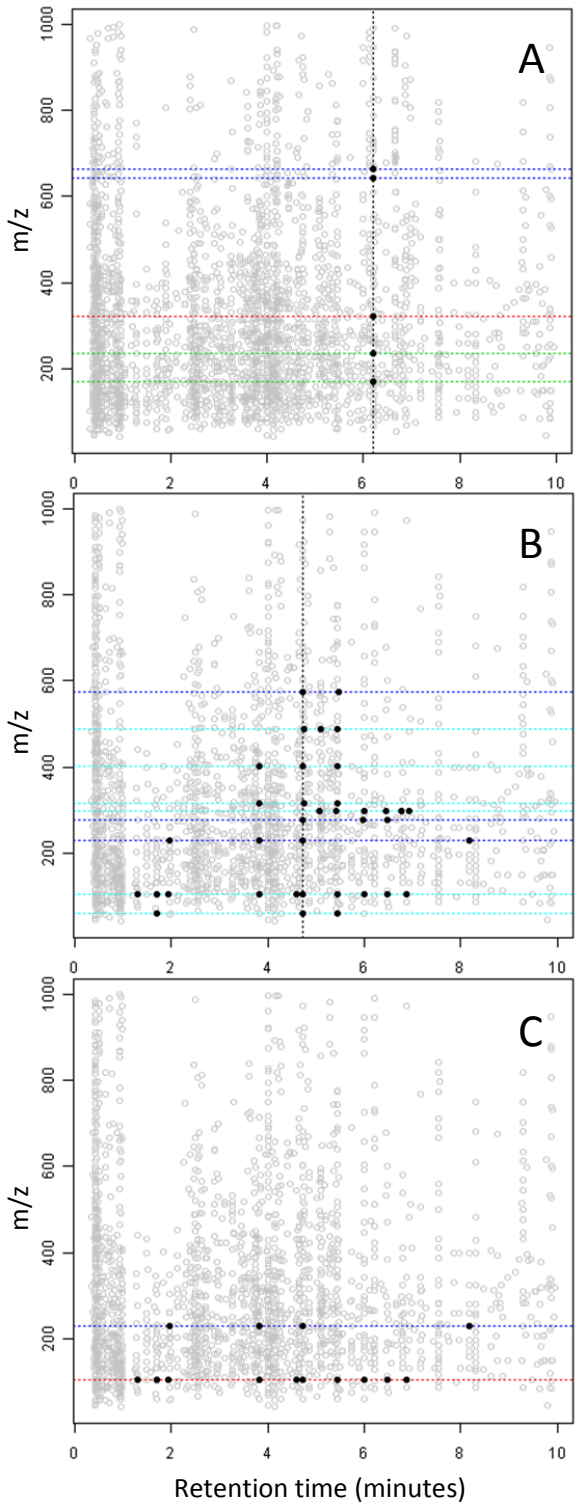


Figure 5-5. Selected results from the feature matching procedure between expected spectral features from the physical properties database and the UPLC-MS data acquired by profiling of the subset standards mixture. The matching of danylsarcosine (A), 3-hydroxybutyric acid (B), and (S)-2-hydroxybutyric acid (C) are shown. Dashed lines are shown for the m/z values representing the spectra of each reference chemical, colored according to their classification in the database (red = parent, green = fragment, blue = adduct, and light blue = adduct fragment). All features detected in the analysis of the reference mixture are shown as grey circles except for features that match database spectral values shown in black. The consensus or best match retention time is shown as a black vertical dashed line in A and B. The absence of a line in C indicates that no concensus or best match was chosen.

5.3 Results

To validate the automated annotations, reference retention times were obtained for all but four of the 55 ionisable and interpretable standards by UPLC-MS analysis of the individual standards. Interestingly, multiple chromatographic peaks were observed for nine standards, indicating chemical impurity, the presence of polymerisation (as previously discussed) or an inherent tendency towards the formation heterogeneous structures prior to chromatographic separation. An example of the latter is shown in Figure 5-6 for the reference chemical N-acetyl-cysteine.

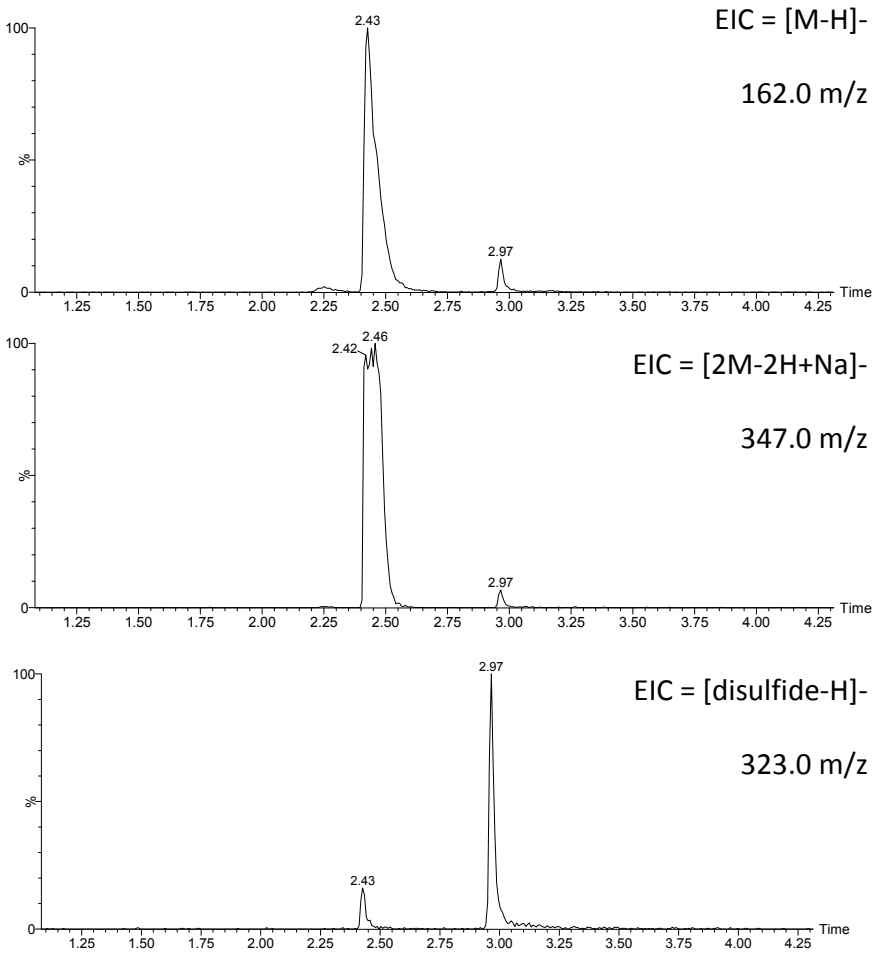


Figure 5-6. Retention reference of N-acetyl-cysteine. N-acetyl-cysteine (top) elutes in two distinct chromatographic peaks, likely dependent on oxidation state. The early eluting peak at 2.3 min is interpreted as the reduced monomer, as monomer adduct ions such as [2M-2H+Na]- (middle) are favoured at that retention time. The expected disulphide mass is dominant in the later eluting peak (bottom). The two distinct chromatographic peaks indicate that the two species exist prior to chromatographic separation. The presence of each mass signal at both retention

times indicates some interconversion after chromatographic separation, likely occurring in the electrospray process.

Of the 51 molecular standards represented in the physical properties database with reference empirically derived retention times, 38 were annotated with a new empirical retention time from the subset mixture. The given retention time was found to match the reference value in 35 of the 38 annotations giving 6% false annotations. Twenty-five percent of the standards remained undetected or not annotated. Annotation of the same 51 standards was attempted in the peaklist generated from the profiling of the master mixture, and retention times were again accurately assigned to 35 of the standards. One additional false annotation was made from a previously un-annotated standard. Together, these results indicate that the additional feature complexity did not confound the annotation.

5.3.3 Evaluation of the master standards mixture in comparison to a biological sample

The base peak intensity (BPI) chromatogram (with detection in the negative ionization mode) for the master mixture of chemicals is shown in Figure 5-7 (top). The BPI chromatogram of the composite urine sample is shown in Figure 5-7 (bottom) for visual reference to demonstrate the general similarities in complexity and intensity, as well as the superficial similarity in composition.

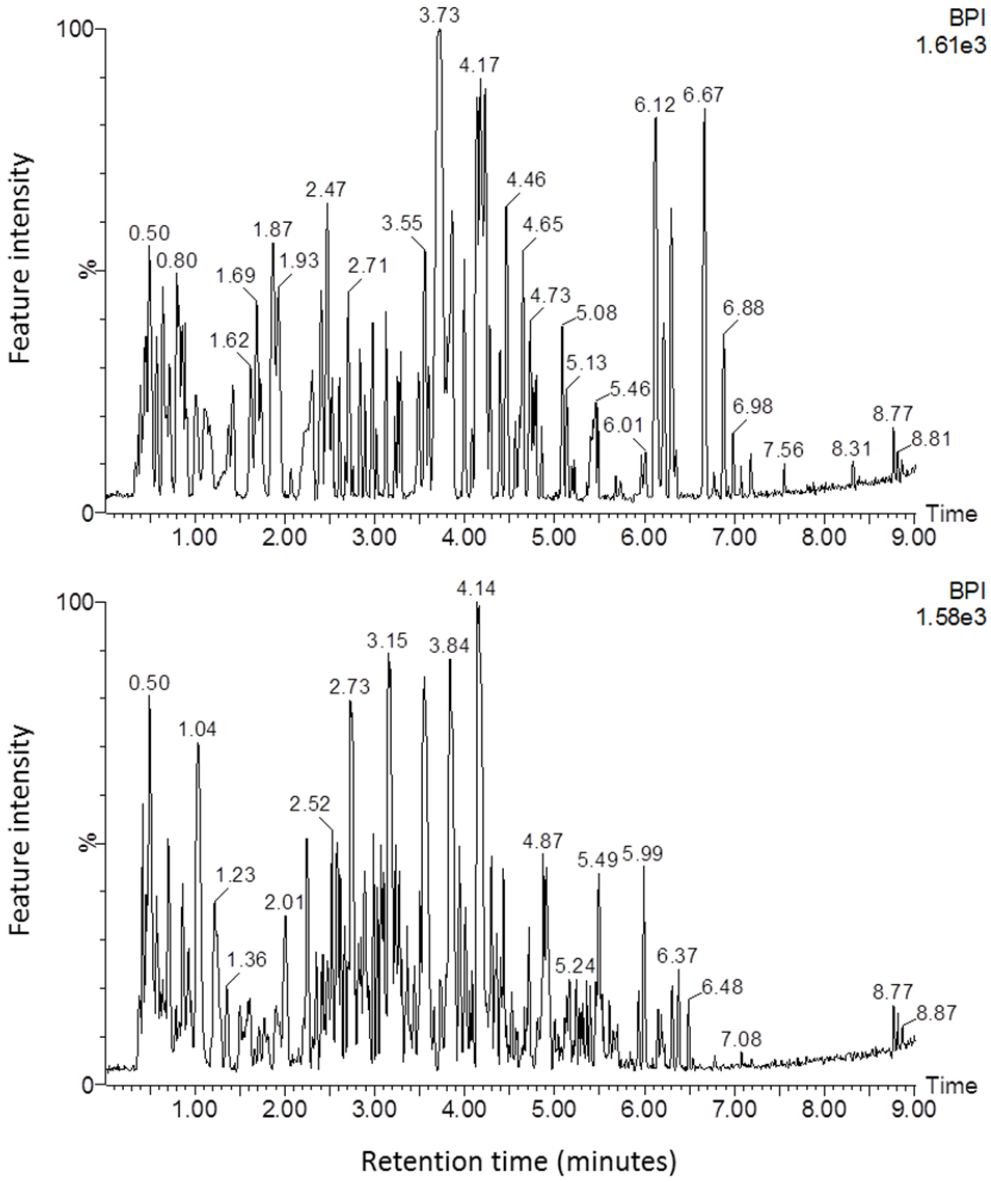


Figure 5-7. Base-peak intensity (BPI) chromatograms of a synthetic standards mixture (top) and a composite sample of human urine (bottom) from a pre vs. post bariatric surgery study. Each sample is shown separated by reversed-phase chromatography with detection (arbitrarily) in negative ionization mode.

5.3.4 Use of a standards mixture for molecular identification in UPLC-MS profiling

Variance in the positive mode profiling dataset is illustrated in a PCA scores plot (Figure 5-8). A slight discrimination between urine samples from patients before and after surgery is observed in both PC1 (representing 8.3% of the total dataset variance) and PC2 (representing 5.9% of the total dataset

variance), highlighted by the diagonal orientation of the cross-diluted group QC samples (xQCs, shown in green) which sits atop the complete pooled QC sample cluster (orange). These groups were further resolved by orthogonal projection to latent structure discriminant analysis (OPLS-DA). The R2Y and Q2 values obtained with a single calculated component were 0.59 and 0.44, indicating that the discriminant analyses are valid. The features most responsible for the discrimination of pre- and post-surgery urine samples were observed using a loadings S-plot of correlation vs. covariance (Figure 5-9A). Discriminant features with differing mass but nearly identical retention time were suspected to be related to a common molecular species (or "component"), and therefore manually collated using common mathematical relationships in mass. Among the most discriminant features, eight (highlighted with red boxes in Figure 5-9A) were found to share a common retention time, and were therefore selected as collectively representing a potential candidate biomarker which is significantly elevated in post-surgery patient samples. These were interpreted as the monoisotopic ion, multimers of that ion, and a fragment of that ion representing the neutral loss of water (Figure 5-9B).

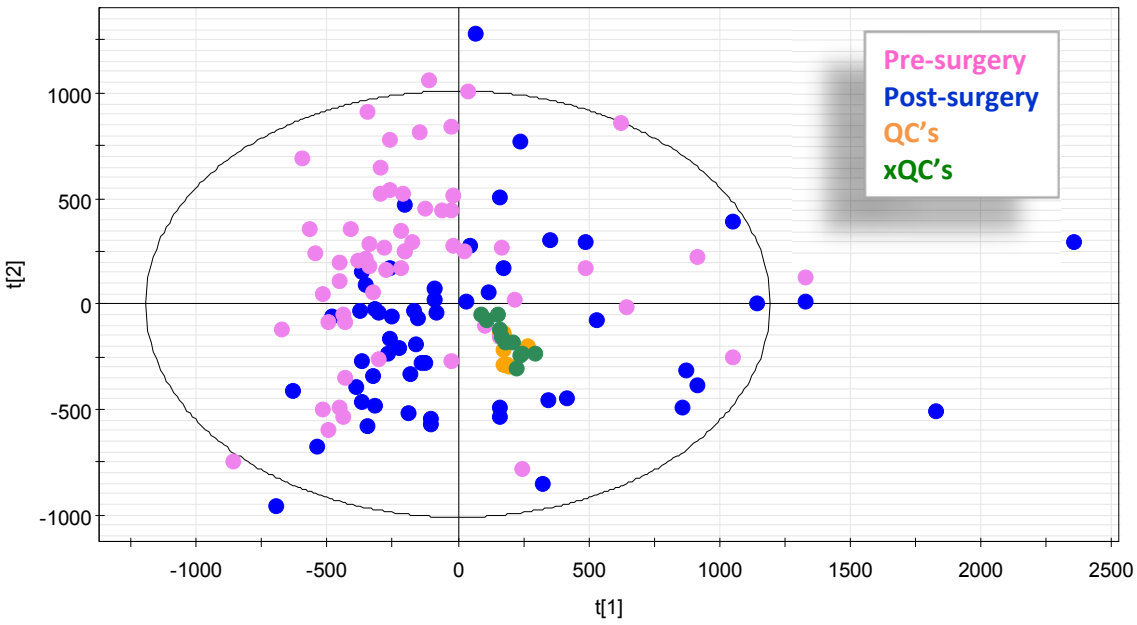
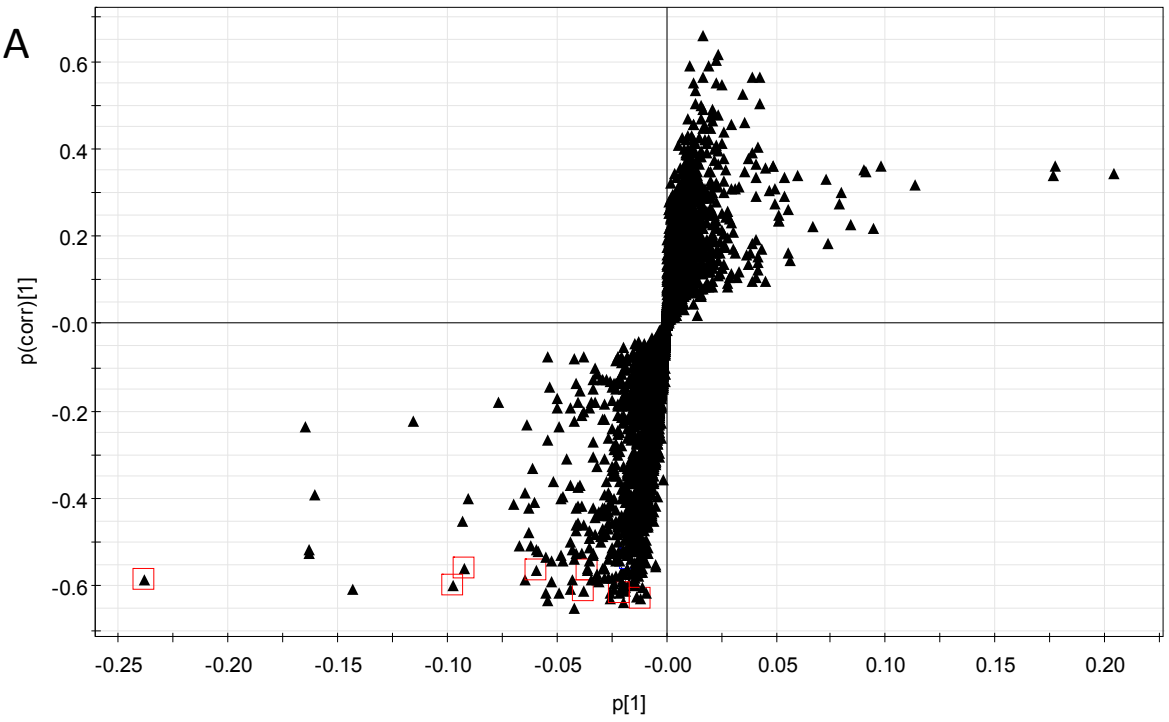


Figure 5-8. A PCA scores plot of principal components 1 vs. 2 of urine samples from a cohort of subjects before and after surgical intervention. Data from the positive ionisation mode analysis is shown. A general trend towards separation of urine samples collected before surgery (pink) and after surgery (blue) is observed across both visualized components. The xQC samples are shown in green. Replicate injections of the QC sample throughout the analysis are shown in orange, demonstrating a high degree of analytical reproducibility with respect to the observed biological variance. PC1 (x axis) is responsible for 8.3% of the total dataset variance, while PC2 (y axis) is responsible for 5.9% of the total dataset variance.

A



B

Retention time (min)	m/z	Component interpretation
3.859	**247.1068**	**[M-H2O+H]+**
3.859	248.092	Isotope of 247
3.859	249.0974	Isotope of 247
3.859	**265.118**	**[M+H]+**
3.859	266.1228	Isotope of 266
3.859	**529.2295**	**[2M+H]+**
3.859	530.2344	Isotope of 529
3.859	**793.3427**	**[3M+H]+**

Figure 5-9. OPLS-DA loadings plot of the feature-set from pre- and post-bariatric surgery subjects, and manual interpretation of selected features elevated in post-surgery patient samples. The individual features responsible for this discrimination are illustrated in an OPLS "S" plot where features responsible for the suggested class differentiation are displayed at the upper and lower bounds of the Y axis (A). Features of greater average magnitude are found towards the extremes of the X axis. The discriminative feature cluster indicated by red squares is manually interpreted by common adduct and neutral loss calculations (B).

5.3 Results

In an effort to identify the discriminant analyte of interest using the *in solutio* database, EICs of the [M+H]+ ion mass (m/z=265.118 +/- 20ppm) were generated from both the composite QC sample and standards mixture analyses, revealing the presence of a peak of the same mass and retention time (Figure 5-10). The spectral patterns of both the full scan and DDA MS/MS analyses of the candidate biomarker match the spectrum obtained by ramped collision energy CID MS/MS in the standards mixture (Figure 5-11).

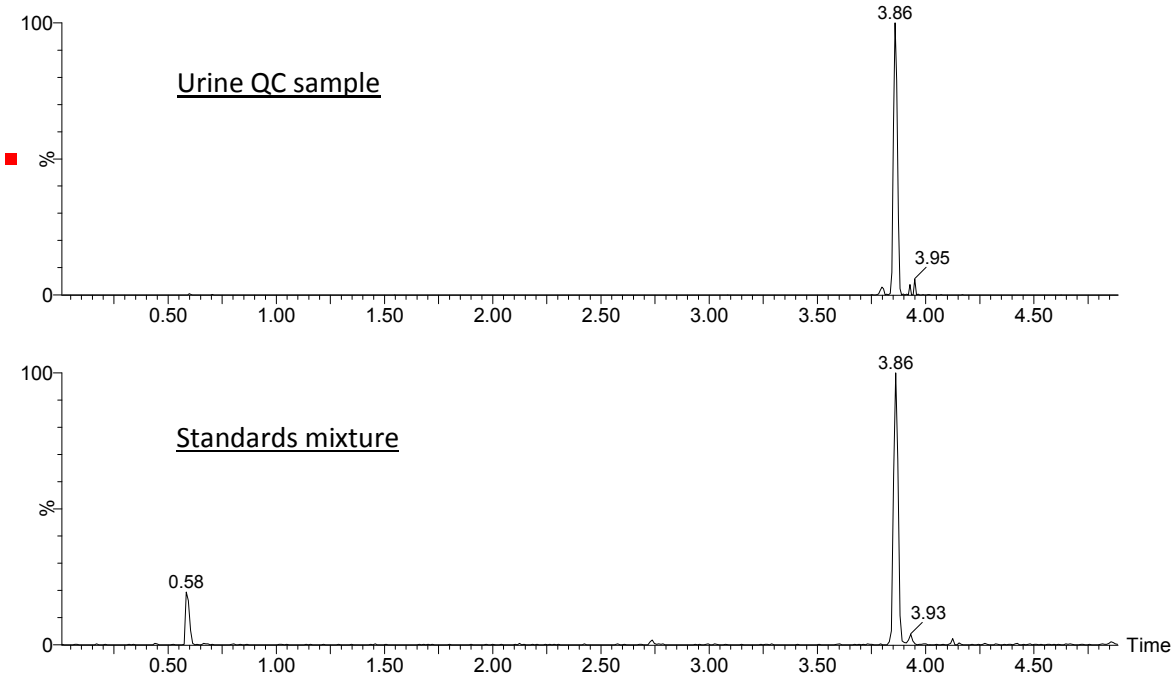


Figure 5-10. Comparison of extracted ion chromatograms (EIC m/z = 265.118 at 20ppm) between a urine QC sample analysis (top) and the standards mixture analysis (bottom). A chromatographic peak of m/z = 265.118 and matching retention is observed in both the urine QC sample and standards mixture, indicating the presence of the unknown discriminant metabolite in the standards library. An additional chromatographic peak of matching mass is observed at 0.58 min in the standards mixture, making assignment of either peak ambiguous if based on mass alone.

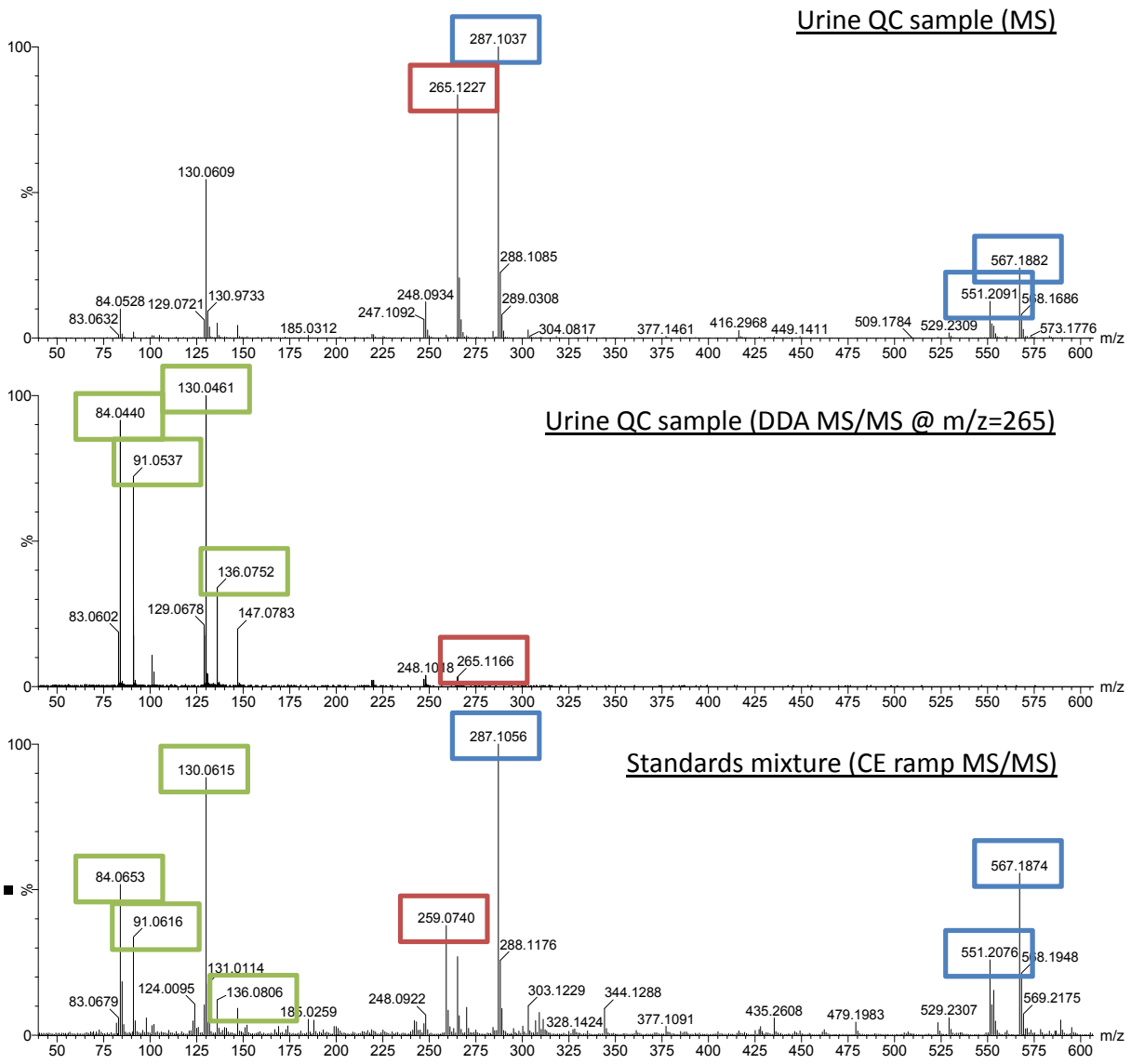


Figure 5-11: Comparison of summed MS and MS/MS spectra from the chromatographic peak at 3.86 minutes in the QC urine sample (top and middle, respectively) and the standards mixture (bottom). Matching spectral adducts, fragments, and the parent mass are indicated with blue, green, and red respectively.

5.3 Results

In order to claim an accurate identification of the unknown analyte in urine, the matching standard reference chemical present in the mixture must be unambiguously annotated. Alpha-N-Phenylacetyl-L-glutamine ($C_{13}H_{16}N_2O_4$) is the only standard present in the mixture matching the expected [M+H]+ ion mass (m/z=265.118) at 20ppm mass accuracy. However, the presence of a second chromatographic peak with the same mass (at 0.58 minutes, visible in Figure 5-10) indicates the presence of ambiguity and precludes the assignment of alpha-N-phenylacetyl-L-glutamine based on mass alone. However, by comparing the fragmentation patterns of the peaks from the standards mixture to those contained within the physical properties database, it became clear by full mass-spectral match that the peak at 3.86 minutes was accurately assigned as alpha-N-phenylacetyl-L-glutamine. With the standard unambiguously located within the mixture, the mixture could therefore be used to accurately identify the unknown biomarker as alpha-N-phenylacetyl-L-glutamine with high confidence.

It should be noted that of the other discriminant features highlighted by the OPLS-DA analysis of the positive mode data, only alpha-N-phenylacetyl-L-glutamine was matched to the empirical database. The process was repeated for the negative mode analysis yielding additional biomarkers of interest, including a number of features determined by analysis of the DDA-derived MS/MS spectra to be both free and glucuronic acid conjugated hydroxylated saturated fatty acids. However, no molecules of this type were included in the standards mixture. Together, these results suggest a need for more comprehensive mixtures.

5.4 Discussion

The principal goal of this work is to demonstrate the potential for a workflow whereby confident feature identification may be made in profiling studies using only prospective data. By combining reference chemicals into mixtures and applying a basic spectral deconvolution script, annotation of those mixtures was demonstrated in a manner that suggests their utility in place of individual reference chemicals. The time savings produced by theis method of multiplexing have the potential to make the large scale screening of chemical standards in routine analysis practical. However, for such mixtures to be truly useful in profiling studies, they should be as large as possible, allowing the construction of large empirical databases with minimal number of mixture injections. It is envisioned that a reasonable number of mixtures of reasonable complexity (*i.e.* perhaps 10 mixtures of 100 chemical reference standards each to avoid severe ionisation suppression effects) could be analysed at the start or end of a profiling analysis of biological samples, providing the data to allow an empirical database of 1000 metabolites to be built with minimal consumption of analytical resource. This database would be complete with method-specific accurate retention time and empirical spectral intensity measurements (including isotopic ratios) that reflect the exact state of the system used for the profiling analysis. When combined with prospective modes of MS/MS data capture on QC samples such as DDA or MS^e (Bateman et al., 2002), it is conceivable that confident metabolite identification of unknown features could be performed by retrospective review of data captured during the initial profiling experiment, precluding the need to return to the instrument at a later date to attempt MS/MS work. By potentially eliminating the need for post-analysis assessment and confirmation of molecular species identity, the approach is targeted to impact a major bottleneck of mass spectrometry-based metabonomic research. Such an approach is equally targeted to the rapid and comprehensive development of chromatographic methods and mass spectrometric detection parameters. Annotated analysis of hundreds of molecular species in a single injection has the potential to catalyse the testing of novel stationary phase chemistries and mobile phase modifiers as well as the more subtle effects of MS source and ion optic voltages on the selectivity and sensitivity of the profiles produced.

5.4 Discussion

The efficacy of such an approach is limited only by (a) the ability to combine many chemical reference standards and (b) the ability to unambiguously deconvolve them. The former would likely be limited by the chemical stability of the mixture, which was not tested in the study presented here, but remains a serious concern for further investigation. The latter is more conceptually open ended, as increasingly complex mixtures can be unambiguously assigned as long as all of the components are unique in at least one manner that is a measureable physical property. This could be achieved both by the conscious design of mixtures only containing metabolites with disparate physical properties (rather than the random combinations made here in the interest of time) as well as the collection of additional data which is descriptive of the metabolites. For example, a mixture of two chemicals with known disparate accurate mass values, no potential for shared molecular sub-structure (and therefore shared fragments), and no potential for adduction to make one chemical the exact same mass as the other, can be unambiguously deconvolved simply by knowing the accurate minoisotopic mass of each. However, pushing the limits of mixture complexity will require expansion of the physical properties database, potentially including the use of ion mobility measurements of molecular collisional cross section (Wickramasekara et al., 2013) to bolster the specificity and accuracy of annotations. Ion mobility data was collected during the analysis of the standards, but has yet to be extracted and integrated into a database format and remains planned further work.

The successful creation of a standards mixture which resembles the complexity of a human biofluid such as urine when profiled by UPLC-MS is therefore a step in what is planned to be an iterative process of refinement of the databases, deconvolution algorithm, and mixtures themselves. The successful application of such a standards mixture to the prospective metabolite identification of alpha-N-phenylacetyl-L-glutamine, a urinary biomarker that discriminates pre- and post-bariatric surgery patient cohorts, serves to demonstrate the potential applicability of the approach to routine analysis and is consistent with published data showing increased excretion of alpha-N-phenylacetyl-L-glutamine post-bariatric surgery in rat models associated with a shift in the gut bacteria responsible for the conversion of phenylalanine to phenylacetate in the colon (Li et al., 2011).

Chapter 6: General discussion and future work

Within this thesis, an advanced molecular profiling pipeline for human population screening has been developed and tested. This pipeline is composed of:

1. Fit-for-purpose and complimentary chromatographic methods facilitating broad molecular coverage with a high degree of analytical precision and laboratory efficiency.
2. A robust UPLC-MS system configuration that permits sustained analysis of large sample sets, reducing the need for analytical sample batching and associated batch correction.
3. A method of UPLC-MS feature extraction and grouping suitable for real time application to sustained analysis.
4. A concept for prospective biomarker identification using in-solution databases to generate empirical UPLC-MS method-specific and experiment-specific databases.

Parts of the pipeline deviate from conventional wisdom surrounding molecular profiling (*e.g.* the generation of high quality datasets by dramatically extending analytical batch sizes) while others are incremental but logical and impactful changes to established practices (*e.g.* considering the analysis order of samples in their pre-processing). Still others remain largely conceptual (e.g. in-solution databases), requiring further development in application to realise their full potential. Taken together, these approaches enable efficient and quality data capture in large scale application. However, they have also uncovered areas where further work is now needed to materialise greater gains.

Achieving large scale continuous analysis has allowed the observation of strong decay in the signal obtained from ToF detectors, which now appears to be the limiting factor in continuous analysis of large sample sets. This phenomenon has not yet received mainstream attention in the literature, and consequentially may not be a priority target of ongoing hardware improvement (e.g. increasing detector speed, dynamic range, and total lifetime). To combat this with the existing hardware, we (in conjunction with the instrument manufacturer Waters Corp.) are co-developing a rapid pre-analysis measurement scheme whereby the detector gain (and therefore signal output) is stabilised across each

individual analysis. Knock-on effects to detector linearity require testing, but early prototype implementations of the stabilising method are yielding promising results.

Additionally, the data presented herein demonstrate that chromatographic conditioning to a state of equilibrium is not general and finite, but rather analyte specific and potentially persisting for the duration of the analysis. On one hand, lengthy conditioning may be performed (e.g. a single plate of 96 QC samples) in an attempt to absorb the most severe effects of UPLC-MS system conditioning. Such an approach is not necessarily impractical, as the relative duration of pre-experiment procedures lessens at larger scale. However, performing feature extraction and pairwise matching in analysis order confers the ability to track molecules as they move, lessening the overall requirement for chromatographic equilibrium. By optomising the profiling methods and UPLC-MS configuration to sustain large batch analysis with high precision, analytical variance is coded in the run order rather than in a series of disparate batches. Complimenting this data with a run-order based grouping mechanism ensures that analytes can be tracked from the start of the analysis to the end with high precision and recall without a need for complex batch correction. The results of testing show that the algorithm produces a dataset with more true-feature groups and resists the bloating of the overall size of the dataset. However, further work in the implementation of this approach to real datasets is required to realise its full potential, including real-time application.

One specific area of interest is in the replacement of user-defined grouping thresholds with metric-driven automatoically determined thresholds for m/z and retention time error. Early efforts show that sensible values can be achieved through the automated testing of matching tolerance windows, expanding in both m/z and retention time dimensions with each subsequent round, until the number of unambiguous matches across all features between two datasets has been maximised (i.e. windows that are too small will match fe features, but windows that are too large will produce match ambiguity, creating a metric-based optimum). This approach will require additional computational time and resource, however if performed in the context of real-time feature extraction and grouping, the rate of data acquisition creates a buffer thought to be ample for these calculations.

5.4 Discussion

The current and planed future approaches to real-time feature grouping and run-order analysis pave the way for the development of real-time QA monitoring procedures for high throughput metabolic profiling. For example, the m/z and retention windows automatically optimised in the proposed future approach may be monitored for significant deviation between sample pairs indicating an acute change in UPLC-MS conditions warranting investigation by an analyst. More simply, the overall chromatographic and MS signal intensity among replicate QC samples may be tracked by looking for a zero-difference in feature intensity between adjacent QC samples. Real time tracking of the number of accurately matched feature pairs between QCs can also help identify when chromatographic drift has accelerated outside of what the grouping algorithm is able to correct. This information may be fed back to the analyst in real time, sparing both precious biofluid sample and acquisition time. All of these systems are currently being developed in an effort to further develop the pipeline for efficient large scale analysis.

Significant challenges in feature annotation and identification are also addressed, using mixtures of known reference chemicals in place of pure reference standards to assist in the prospective identification of unknown biomarkers. One practical implementation of this concept would be the creation of bio-panels of analytes thought to be relevant to the biological system and study design, or suggested as potentially relevant by earlier analyses (e.g. NMR). In this manner, the chances of capturing the data required for confident metabolite identification would be improved within the original profiling experiment, aiding in overall efficiency of the high throughput large scale phenotyping laboratory.

Finally, it is noteworthy that while large-scale analysis provides unique challenges to the analyst, other practical aspects become easier as experiments grow larger. All operations performed prior to each experiment (e.g. cleaning, calibration, and conditioning) require less time relative to the analysis as the experiments grow larger. Furthermore, labs which focus on very large-scale work may be more able to dedicate instruments to specific methodology and biofluid types, reducing the incidence of system contamination (eg. lipid contamination from the analysis of human blood products prior to analysis of human urine) and hardware fatigue introduced by constant instrument reconfiguration. When

individual experiments are large enough, the complete replacement of consumables (eg. the column) and non-consumable parts alike (eg. sample injection syringe, loop, and peak tubing) becomes increasingly economical, further reducing the incidence of contamination and fatigue per project. In this manner, natural benefits exist to compliment the challenges of population phenotyping addressed within this thesis, further enabling the continued growth of UPLC-MS application to large scale analysis.

The ultimate goal of these efforts is to be able to generate metabolic data related to human phenotypes with sufficient efficiency and precision to allow for multi-centre participation in human phenotyping and cross comparison of acquired data. In this manner, the scope of phenotyping can be greater than the capabilities of any one laboratory, amassing unprecedented statistical power for meta-analyses of profiling data. It is hoped that this pipeline is a foundation on which future developments facilitate harmonisation of platforms and databases across multiple laboratories. Extension of this would include translation to other commonly collected biofluids and sample types such as blood product samples (serum and plasma) in order to capture unparalleled insight into human phenotypes.

References

AHMED, A. M. 2002. History of diabetes mellitus. *Saudi Med J,* 23**,** 373-8.

ALPERT, A. J. 1990. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *J Chromatogr,* 499**,** 177-96.

ASSFALG, M., BERTINI, I., COLANGIULI, D., LUCHINAT, C., SCHAFER, H., SCHUTZ, B. & SPRAUL, M. 2008. Evidence of different metabolic phenotypes in humans. *Proc Natl Acad Sci U S A,* 105**,** 1420-4.

ATHERSUCH, T. J., CASTRO-PEREZ, J., RODGERS, C., NICHOLSON, J. K. & WILSON, I. D. 2010. UPLC-MS, HPLC-radiometric, and NMR-spectroscopic studies on the metabolic fate of 3-fluoro-[U-14C]-aniline in the bile-cannulated rat. *Xenobiotica,* 40**,** 510-23.

AUSTDAL, M., SKRASTAD, R. B., GUNDERSEN, A. S., AUSTGULEN, R., IVERSEN, A. C. & BATHEN, T. F. 2014. Metabolomic biomarkers in serum and urine in women with preeclampsia. *PLoS One,* 9**,** e91923.

BAJAD, S. U., LU, W., KIMBALL, E. H., YUAN, J., PETERSON, C. & RABINOWITZ, J. D. 2006. Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *J Chromatogr A,* 1125**,** 76-88.

BATEMAN, R. H., CARRUTHERS, R., HOYES, J. B., JONES, C., LANGRIDGE, J. I., MILLAR, A. & VISSERS, J. P. 2002. A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation. *J Am Soc Mass Spectrom,* 13**,** 792-803.

BECKONERT, O., COEN, M., KEUN, H. C., WANG, Y., EBBELS, T. M., HOLMES, E., LINDON, J. C. & NICHOLSON, J. K. 2010. High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues. *Nat Protoc,* 5**,** 1019-32.

BERNINI, P., BERTINI, I., LUCHINAT, C., NEPI, S., SACCENTI, E., SCHAFER, H., SCHUTZ, B., SPRAUL, M. & TENORI, L. 2009. Individual human phenotypes in metabolic space and time. *J Proteome Res,* 8**,** 4264-71.

BICTASH, M., EBBELS, T. M., CHAN, Q., LOO, R. L., YAP, I. K., BROWN, I. J., DE IORIO, M., DAVIGLUS, M. L., HOLMES, E., STAMLER, J., NICHOLSON, J. K. & ELLIOTT, P. 2010. Opening up the "Black Box": metabolic phenotyping and metabolome-wide association studies in epidemiology. *J Clin Epidemiol,* 63**,** 970-9.

BIOMARKERS DEFINITIONS WORKING GROUP 2001. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther,* 69**,** 89-95.

BOLODEOKU, J. & DONALDSON, D. 1996. Urinalysis in clinical diagnosis. *J Clin Pathol,* 49**,** 623-6.

BOUATRA, S., AZIAT, F., MANDAL, R., GUO, A. C., WILSON, M. R., KNOX, C., BJORNDAHL, T. C., KRISHNAMURTHY, R., SALEEM, F., LIU, P., DAME, Z. T., POELZER, J., HUYNH, J., YALLOU, F. S., PSYCHOGIOS, N., DONG, E., BOGUMIL, R., ROEHRING, C. & WISHART, D. S. 2013. The human urine metabolome. *PLoS One,* 8**,** e73076.

BOVE, K. E., HEUBI, J. E., BALISTRERI, W. F. & SETCHELL, K. D. 2004. Bile acid synthetic defects and liver disease: a comprehensive review. *Pediatr Dev Pathol,* 7**,** 315-34.

BRISTOW, P. A. & KNOX, J. H. 1977. Standardization of Test Conditions for High-Performance Liquid-Chromatography Columns. *Chromatographia,* 10**,** 279-289.

BROADHURST, D. I. & KELL, D. B. 2006. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics,* 2**,** 171-196.

BROECKLING, C. D., HEUBERGER, A. L. & PRENNI, J. E. 2013. Large scale non-targeted metabolomic profiling of serum by ultra performance liquid chromatography-mass spectrometry (UPLC-MS). *J Vis Exp***,** e50242.

References

BUHRMAN, D. L., PRICE, P. I. & RUDEWICZCOR, P. J. 1996. Quantitation of SR 27417 in human plasma using electrospray liquid chromatography-tandem mass spectrometry: A study of ion suppression. *J Am Soc Mass Spectrom,* 7**,** 1099-105.

BUSCHER, J. M., CZERNIK, D., EWALD, J. C., SAUER, U. & ZAMBONI, N. 2009. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Anal Chem,* 81**,** 2135-43.

BUSZEWSKI, B. & NOGA, S. 2012. Hydrophilic interaction liquid chromatography (HILIC)--a powerful separation technique. *Anal Bioanal Chem,* 402**,** 231-47.

CARR, P. W. & SUN, L. F. 1998. An approximate expression for the minimum plate height produced by Knox's equation. *Journal of Microcolumn Separations,* 10**,** 149-152.

CHAE, M., SHMOOKLER REIS, R. J. & THADEN, J. J. 2008. An iterative block-shifting approach to retention time alignment that preserves the shape and area of gas chromatography-mass spectrometry peaks. *BMC Bioinformatics,* 9 Suppl 9**,** S15.

CHERNUSHEVICH, I. V., LOBODA, A. V. & THOMSON, B. A. 2001. An introduction to quadrupole-time-of-flight mass spectrometry. *J Mass Spectrom,* 36**,** 849-65.

CHRISTIN, C., HOEFSLOOT, H. C., SMILDE, A. K., SUITS, F., BISCHOFF, R. & HORVATOVICH, P. L. 2010. Time alignment algorithms based on selected mass traces for complex LC-MS data. *J Proteome Res,* 9**,** 1483-95.

CHRISTIN, C., SMILDE, A. K., HOEFSLOOT, H. C., SUITS, F., BISCHOFF, R. & HORVATOVICH, P. L. 2008. Optimized time alignment algorithm for LC-MS data: correlation optimized warping using component detection algorithm-selected mass chromatograms. *Anal Chem,* 80**,** 7012-21.

COBLE, J. B. & FRAGA, C. G. 2014. Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery. *J Chromatogr A,* 1358**,** 155-64.

COEN, M., HOLMES, E., LINDON, J. C. & NICHOLSON, J. K. 2008. NMR-based metabolic profiling and metabonomic approaches to problems in molecular toxicology. *Chem Res Toxicol,* 21**,** 9-27.

CROIXMARIE, V., UMBDENSTOCK, T., CLOAREC, O., MOREAU, A., PASCUSSI, J. M., BOURSIER-NEYRET, C. & WALTHER, B. 2009. Integrated comparison of drug-related and drug-induced ultra performance liquid chromatography/mass spectrometry metabonomic profiles using human hepatocyte cultures. *Anal Chem,* 81**,** 6061-9.

DALLMANN, R., VIOLA, A. U., TAROKH, L., CAJOCHEN, C. & BROWN, S. A. 2012. The human circadian metabolome. *Proc Natl Acad Sci U S A,* 109**,** 2625-9.

DETTMER, K., ARONOV, P. A. & HAMMOCK, B. D. 2007. Mass spectrometry-based metabolomics. *Mass Spectrom Rev,* 26**,** 51-78.

DOLAN, J. W., SNYDER, L. R., DJORDJEVIC, N. M., HILL, D. W. & WAEGHE, T. J. 1999. Reversed-phase liquid chromatographic separation of complex samples by optimizing temperature and gradient time I. Peak capacity limitations. *J Chromatogr A,* 857**,** 1-20.

DONA, A. C., JIMENEZ, B., SCHAFER, H., HUMPFER, E., SPRAUL, M., LEWIS, M. R., PEARCE, J. T., HOLMES, E., LINDON, J. C. & NICHOLSON, J. K. 2014. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal Chem,* 86**,** 9887-94.

DRAISMA, H. H., REIJMERS, T. H., VAN DER KLOET, F., BOBELDIJK-PASTOROVA, I., SPIES-FABER, E., VOGELS, J. T., MEULMAN, J. J., BOOMSMA, D. I., VAN DER GREEF, J. & HANKEMEIER, T. 2010. Equating, or correction for between-block effects with application to body fluid LC-MS and NMR metabolomics data sets. *Anal Chem,* 82**,** 1039-46.

DUNN, W. B., BROADHURST, D., BEGLEY, P., ZELENA, E., FRANCIS-MCINTYRE, S., ANDERSON, N., BROWN, M., KNOWLES, J. D., HALSALL, A., HASELDEN, J. N., NICHOLLS, A. W., WILSON, I. D., KELL, D. B., GOODACRE, R. & HUMAN SERUM METABOLOME, C. 2011. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc,* 6**,** 1060-83.

EILERS, P. H. 2004. Parametric time warping. *Anal Chem,* 76**,** 404-11.

References

EVANS, A. M., DEHAVEN, C. D., BARRETT, T., MITCHELL, M. & MILGRAM, E. 2009. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal Chem,* 81**,** 6656-67.

FERNIE, A. R., TRETHEWEY, R. N., KROTZKY, A. J. & WILLMITZER, L. 2004. Metabolite profiling: from diagnostics to systems biology. *Nat Rev Mol Cell Biol,* 5**,** 763-9.

FILLATRE, Y., RONDEAU, D., JADAS-HECART, A. & COMMUNAL, P. Y. 2010. Advantages of the scheduled selected reaction monitoring algorithm in liquid chromatography/electrospray ionization tandem mass spectrometry multi-residue analysis of 242 pesticides: a comparative approach with classical selected reaction monitoring mode. *Rapid Commun Mass Spectrom,* 24**,** 2453-61.

GARROD, A. E. 1902. About Alkaptonuria. *Med Chir Trans,* 85**,** 69-78.

GARROD, A. E. 1909. *Inborn errors of metabolism : the Croonian Lectures delivered before the Royal College of Physicians of London, in June 1908,* London, Frowde and Hodder & Stoughton.

GARROD, A. E. 1923. *Inborn errors of metabolism,* London,, H. Frowde and Hodder & Stoughton.

GIKA, H. G., THEODORIDIS, G. A., WINGATE, J. E. & WILSON, I. D. 2007. Within-day reproducibility of an HPLC-MS-based method for metabonomic analysis: application to human urine. *J Proteome Res,* 6**,** 3291-303.

GIKA, H. G., WILSON, I. D. & THEODORIDIS, G. A. 2014. LC-MS-based holistic metabolic profiling. Problems, limitations, advantages, and future perspectives. *J Chromatogr B Analyt Technol Biomed Life Sci,* 966**,** 1-6.

GOWDA, H., IVANISEVIC, J., JOHNSON, C. H., KURCZY, M. E., BENTON, H. P., RINEHART, D., NGUYEN, T., RAY, J., KUEHL, J., AREVALO, B., WESTENSKOW, P. D., WANG, J., ARKIN, A. P., DEUTSCHBAUER, A. M., PATTI, G. J. & SIUZDAK, G. 2014. Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. *Anal Chem,* 86**,** 6931-9.

GRAY, N., HEATON, J., MUSENGA, A., COWAN, D. A., PLUMB, R. S. & SMITH, N. W. 2013. Comparison of reversed-phase and hydrophilic interaction liquid chromatography for the quantification of ephedrines using medium-resolution accurate mass spectrometry. *J Chromatogr A,* 1289**,** 37-46.

GURDENIZ, G., KRISTENSEN, M., SKOV, T. & DRAGSTED, L. O. 2012. The Effect of LC-MS Data Preprocessing Methods on the Selection of Plasma Biomarkers in Fed vs. Fasted Rats. *Metabolites,* 2**,** 77-99.

HABER, M. H. 1988. Pisse prophecy: a brief history of urinalysis. *Clin Lab Med,* 8**,** 415-30.

HOFFMANN, E. D. & STROOBANT, V. 2007. *Mass spectrometry : principles and applications,* Chichester, England ; Hoboken, NJ, J. Wiley.

HOLMES, E., LOO, R. L., STAMLER, J., BICTASH, M., YAP, I. K., CHAN, Q., EBBELS, T., DE IORIO, M., BROWN, I. J., VESELKOV, K. A., DAVIGLUS, M. L., KESTELOOT, H., UESHIMA, H., ZHAO, L., NICHOLSON, J. K. & ELLIOTT, P. 2008. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature,* 453**,** 396-400.

HORAI, H., ARITA, M., KANAYA, S., NIHEI, Y., IKEDA, T., SUWA, K., OJIMA, Y., TANAKA, K., TANAKA, S., AOSHIMA, K., ODA, Y., KAKAZU, Y., KUSANO, M., TOHGE, T., MATSUDA, F., SAWADA, Y., HIRAI, M. Y., NAKANISHI, H., IKEDA, K., AKIMOTO, N., MAOKA, T., TAKAHASHI, H., ARA, T., SAKURAI, N., SUZUKI, H., SHIBATA, D., NEUMANN, S., IIDA, T., TANAKA, K., FUNATSU, K., MATSUURA, F., SOGA, T., TAGUCHI, R., SAITO, K. & NISHIOKA, T. 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom,* 45**,** 703-14.

HOTELLING, H. 1933. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.,* 24**,** 22.

HUSERMET. 2008. *SOP HU005: Sample Scheduling, Picking and Preparation* [Online]. Available: http://www.husermet.org/files/SOP_HU005_SamplePicking_v10_pub.pdf.

References

JOHNSON, K. J., WRIGHT, B. W., JARMAN, K. H. & SYNOVEC, R. E. 2003. High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. *J Chromatogr A,* 996**,** 141-55.

KADDURAH-DAOUK, R., KRISTAL, B. S. & WEINSHILBOUM, R. M. 2008. Metabolomics: a global biochemical approach to drug response and disease. *Annu Rev Pharmacol Toxicol,* 48**,** 653-83.

KATAJAMAA, M., MIETTINEN, J. & ORESIC, M. 2006. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics,* 22**,** 634-6.

KATAJAMAA, M. & ORESIC, M. 2007. Data processing for mass spectrometry-based metabolomics. *J Chromatogr A,* 1158**,** 318-28.

KEUN, H. C. & ATHERSUCH, T. J. 2011. Nuclear magnetic resonance (NMR)-based metabolomics. *Methods Mol Biol,* 708**,** 321-34.

KIND, T. & FIEHN, O. 2006. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics,* 7**,** 234.

KIND, T. & FIEHN, O. 2010. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal Rev,* 2**,** 23-60.

KRUG, S., KASTENMULLER, G., STUCKLER, F., RIST, M. J., SKURK, T., SAILER, M., RAFFLER, J., ROMISCH-MARGL, W., ADAMSKI, J., PREHN, C., FRANK, T., ENGEL, K. H., HOFMANN, T., LUY, B., ZIMMERMANN, R., MORITZ, F., SCHMITT-KOPPLIN, P., KRUMSIEK, J., KREMER, W., HUBER, F., OEH, U., THEIS, F. J., SZYMCZAK, W., HAUNER, H., SUHRE, K. & DANIEL, H. 2012. The dynamic range of the human metabolome revealed by challenges. *FASEB J,* 26**,** 2607-19.

KUHARA, T. 2005. Gas chromatographic-mass spectrometric urinary metabolome analysis to study mutations of inborn errors of metabolism. *Mass Spectrom Rev,* 24**,** 814-27.

LANGE, E., TAUTENHAHN, R., NEUMANN, S. & GROPL, C. 2008. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics,* 9**,** 375.

LARIVE, C. K., BARDING, G. A. & DINGES, M. M. 2014. NMR Spectroscopy for Metabolomics and Metabolic Profiling. *Anal Chem*.

LAW, C. Y., LAM, C. W., CHING, C. K., YAU, K. C., HO, T. W., LAI, C. K. & MAK, C. M. 2014. NMR-based urinalysis for beta-ketothiolase deficiency. *Clin Chim Acta,* 438C**,** 222-225.

LENTNER, C. 1981. *Geigy scientific tables,* West Caldwell, N.J., Medical Education Division, Ciba-Geigy Corp.

LENZ, E. M. & WILSON, I. D. 2007. Analytical strategies in metabonomics. *J Proteome Res,* 6**,** 443-58.

LI, F., MAGUIGAD, J., PELZER, M., JIANG, X. & JI, Q. C. 2008. A novel 'peak parking' strategy for ultra-performance liquid chromatography/tandem mass spectrometric detection for enhanced performance of bioanalytical assays. *Rapid Commun Mass Spectrom,* 22**,** 486-94.

LI, J. V., ASHRAFIAN, H., BUETER, M., KINROSS, J., SANDS, C., LE ROUX, C. W., BLOOM, S. R., DARZI, A., ATHANASIOU, T., MARCHESI, J. R., NICHOLSON, J. K. & HOLMES, E. 2011. Metabolic surgery profoundly influences gut microbial-host metabolic cross-talk. *Gut,* 60**,** 1214-23.

LILAND, K. H. 2011. Multivariate methods in metabolomics - from pre-processing to dimension reduction and statistical analysis. *Trac-Trends in Analytical Chemistry,* 30**,** 827-841.

LINDON, J. C., HOLMES, E., BOLLARD, M. E., STANLEY, E. G. & NICHOLSON, J. K. 2004. Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. *Biomarkers,* 9**,** 1-31.

LINDON, J. C. & NICHOLSON, J. K. 2008. Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics. *Annu Rev Anal Chem (Palo Alto Calif),* 1**,** 45-69.

LIU, Q., SHI, Y., GUO, T., WANG, Y., CONG, W. & ZHU, J. 2012. Metabolite discovery of helicidum in rat urine with XCMS based on the data of ultra performance liquid chromatography coupled to time-of-flight mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci,* 907**,** 146-53.

LOMMEN, A. 2009. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal Chem,* 81**,** 3079-86.

References

LU, W., CLASQUIN, M. F., MELAMUD, E., AMADOR-NOGUEZ, D., CAUDY, A. A. & RABINOWITZ, J. D. 2010. Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. *Anal Chem,* 82**,** 3212-21.

LUAN, H., LIU, L. F., MENG, N., TANG, Z., CHUA, K. K., CHEN, L. L., SONG, J. X., MOK, V. C., XIE, L. X., LI, M. & CAI, Z. 2014. LC-MS-Based Urinary Metabolite Signatures in Idiopathic Parkinson's Disease. *J Proteome Res*.

MAITRE, L., FTHENOU, E., ATHERSUCH, T., COEN, M., TOLEDANO, M. B., HOLMES, E., KOGEVINAS, M., CHATZI, L. & KEUN, H. C. 2014. Urinary metabolic profiles in early pregnancy are associated with preterm birth and fetal growth restriction in the Rhea mother-child cohort study. *BMC Med,* 12**,** 110.

MARTIN, A. J. & SYNGE, R. L. 1941. A new form of chromatogram employing two liquid phases: A theory of chromatography. 2. Application to the micro-determination of the higher monoamino-acids in proteins. *Biochem J,* 35**,** 1358-68.

MAZZEO, J., DNEUE, U., KELE, M. & PLUMB, R. 2005. Advancing LC Performance with Smaller Particles and Higher Pressure. *Analytical Chemistry,* 77**,** 460 A-467 A.

MENNI, C., KASTENMULLER, G., PETERSEN, A. K., BELL, J. T., PSATHA, M., TSAI, P. C., GIEGER, C., SCHULZ, H., ERTE, I., JOHN, S., BROSNAN, M. J., WILSON, S. G., TSAPROUNI, L., LIM, E. M., STUCKEY, B., DELOUKAS, P., MOHNEY, R., SUHRE, K., SPECTOR, T. D. & VALDES, A. M. 2013. Metabolomic markers reveal novel pathways of ageing and early development in human populations. *Int J Epidemiol,* 42**,** 1111-9.

MONLEON, D., MORALES, J. M., BARRASA, A., LOPEZ, J. A., VAZQUEZ, C. & CELDA, B. 2009. Metabolite profiling of fecal water extracts from human colorectal cancer. *NMR Biomed,* 22**,** 342-8.

MORELL-GARCIA, D., BARCELO, B., RODRIGUEZ, A., LINEIRO, V., ROBLES, R., VIDAL-PUIGSERVER, J., COSTA-BAUZA, A. & GRASES, F. 2014. Application of nuclear magnetic resonance spectroscopy for identification of ciprofloxacin crystalluria. *Clin Chim Acta,* 438C**,** 43-45.

MUTO, A., TAKEI, H., UNNO, A., MURAI, T., KUROSAWA, T., OGAWA, S., IIDA, T., IKEGAWA, S., MORI, J., OHTAKE, A., HOSHINA, T., MIZUOCHI, T., KIMURA, A., HOFMANN, A. F., HAGEY, L. R. & NITTONO, H. 2012. Detection of Delta4-3-oxo-steroid 5beta-reductase deficiency by LC-ESI-MS/MS measurement of urinary bile acids. *J Chromatogr B Analyt Technol Biomed Life Sci,* 900**,** 24-31.

NEUE, U. D. 2008. Peak capacity in unidimensional chromatography. *J Chromatogr A,* 1184**,** 107-30.

NEVEDOMSKAYA, E., MAYBORODA, O. A. & DEELDER, A. M. 2011. Cross-platform analysis of longitudinal data in metabolomics. *Mol Biosyst,* 7**,** 3214-22.

NEW, L. S. & CHAN, E. C. 2008. Evaluation of BEH C18, BEH HILIC, and HSS T3 (C18) column chemistries for the UPLC-MS-MS analysis of glutathione, glutathione disulfide, and ophthalmic acid in mouse liver and human plasma. *J Chromatogr Sci,* 46**,** 209-14.

NEZAMI RANJBAR, M. R., ZHAO, Y., TADESSE, M. G., WANG, Y. & RESSOM, H. W. 2013. Gaussian process regression model for normalization of LC-MS data using scan-level information. *Proteome Sci,* 11**,** S13.

NG, D. P., SALIM, A., LIU, Y., ZOU, L., XU, F. G., HUANG, S., LEONG, H. & ONG, C. N. 2012. A metabolomic study of low estimated GFR in non-proteinuric type 2 diabetes mellitus. *Diabetologia,* 55**,** 499-508.

NICHOLSON, G., RANTALAINEN, M., MAHER, A. D., LI, J. V., MALMODIN, D., AHMADI, K. R., FABER, J. H., HALLGRIMSDOTTIR, I. B., BARRETT, A., TOFT, H., KRESTYANINOVA, M., VIKSNA, J., NEOGI, S. G., DUMAS, M. E., SARKANS, U., THE MOLPAGE, C., SILVERMAN, B. W., DONNELLY, P., NICHOLSON, J. K., ALLEN, M., ZONDERVAN, K. T., LINDON, J. C., SPECTOR, T. D., MCCARTHY, M. I., HOLMES, E., BAUNSGAARD, D. & HOLMES, C. C. 2011. Human metabolic profiles are stably controlled by genetic and environmental variation. *Mol Syst Biol,* 7**,** 525.

References

NICHOLSON, J. K., LINDON, J. C. & HOLMES, E. 1999. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica,* 29**,** 1181-9.

NICHOLSON, J. K., O'FLYNN, M. P., SADLER, P. J., MACLEOD, A. F., JUUL, S. M. & SONKSEN, P. H. 1984. Proton-nuclear-magnetic-resonance studies of serum, plasma and urine from fasting normal and diabetic subjects. *Biochem J,* 217**,** 365-75.

NIELSEN, N. P. V., CARSTENSEN, J. M. & SMEDSGAARD, J. 1998. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A,* 805**,** 17-35.

NOGA, M., SUCHARSKI, F., SUDER, P. & SILBERRING, J. 2007. A practical guide to nano-LC troubleshooting. *J Sep Sci,* 30**,** 2179-89.

OLIVEROS, J. C. 2007. VENNY. An interactive tool for comparing lists with Venn Diagrams. .

PATTI, G. J., TAUTENHAHN, R. & SIUZDAK, G. 2012a. Meta-analysis of untargeted metabolomic data from multiple profiling experiments. *Nat Protoc,* 7**,** 508-16.

PATTI, G. J., YANES, O. & SIUZDAK, G. 2012b. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol,* 13**,** 263-9.

PEARSON, K. 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philos. Mag.,* 2.

PETERSSON, P., FRANK, A., HEATON, J. & EUERBY, M. R. 2008. Maximizing peak capacity and separation speed in liquid chromatography. *J Sep Sci,* 31**,** 2346-57.

PLUMB, R., CASTRO-PEREZ, J., GRANGER, J., BEATTIE, I., JONCOUR, K. & WRIGHT, A. 2004. Ultra-performance liquid chromatography coupled to quadrupole-orthogonal time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom,* 18**,** 2331-7.

PLUSKAL, T., CASTILLO, S., VILLAR-BRIONES, A. & ORESIC, M. 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics,* 11**,** 395.

R CORE TEAM 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

RAINVILLE, P. D., STUMPF, C. L., SHOCKCOR, J. P., PLUMB, R. S. & NICHOLSON, J. K. 2007. Novel application of reversed-phase UPLC-oaTOF-MS for lipid analysis in complex biological mixtures: a new tool for lipidomics. *J Proteome Res,* 6**,** 552-8.

ROBINSON, M. D., DE SOUZA, D. P., KEEN, W. W., SAUNDERS, E. C., MCCONVILLE, M. J., SPEED, T. P. & LIKIC, V. A. 2007. A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments. *BMC Bioinformatics,* 8**,** 419.

ROUSU, T., HERTTUAINEN, J. & TOLONEN, A. 2010. Comparison of triple quadrupole, hybrid linear ion trap triple quadrupole, time-of-flight and LTQ-Orbitrap mass spectrometers in drug discovery phase metabolite screening and identification in vitro--amitriptyline and verapamil as model compounds. *Rapid Commun Mass Spectrom,* 24**,** 939-57.

SALEK, R. M., STEINBECK, C., VIANT, M. R., GOODACRE, R. & DUNN, W. B. 2013. The role of reporting standards for metabolite annotation and identification in metabolomic studies. *Gigascience,* 2**,** 13.

SANGSTER, T., MAJOR, H., PLUMB, R., WILSON, A. J. & WILSON, I. D. 2006. A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabonomic analysis. *Analyst,* 131**,** 1075-8.

SAVITZKY, A. & GOLAY, M. J. E. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry,* 36**,** 1627-1639.

SKOGERSON, K., RUNNEBAUM, R., WOHLGEMUTH, G., DE ROPP, J., HEYMANN, H. & FIEHN, O. 2009. Comparison of gas chromatography-coupled time-of-flight mass spectrometry and 1H nuclear magnetic resonance spectroscopy metabolite identification in white wines from a sensory study investigating wine body. *J Agric Food Chem,* 57**,** 6899-907.

References

SMITH, C. A., O'MAILLE, G., WANT, E. J., QIN, C., TRAUGER, S. A., BRANDON, T. R., CUSTODIO, D. E., ABAGYAN, R. & SIUZDAK, G. 2005. METLIN: a metabolite mass spectral database. *Ther Drug Monit,* 27**,** 747-51.

SMITH, C. A., WANT, E. J., O'MAILLE, G., ABAGYAN, R. & SIUZDAK, G. 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem,* 78**,** 779-87.

SNYDER, L. R., KIRKLAND, J. J. & DOLAN, J. W. 2010. *Introduction to modern liquid chromatography,* Hoboken, N.J., Wiley.

SPAGOU, K., WILSON, I. D., MASSON, P., THEODORIDIS, G., RAIKOS, N., COEN, M., HOLMES, E., LINDON, J. C., PLUMB, R. S., NICHOLSON, J. K. & WANT, E. J. 2011. HILIC-UPLC-MS for exploratory urinary metabolic profiling in toxicological studies. *Anal Chem,* 83**,** 382-90.

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

STEPHENS, W. 1946. Pulsed Mass Spectrometer with Time Dispersion. *Bull. Am. Phys. Soc.,* 21**,** 22.

STUDENT 1908. The probable error of a mean. *Biometrika,* 6**,** 1-25.

SUMNER, L. W., AMBERG, A., BARRETT, D., BEALE, M. H., BEGER, R., DAYKIN, C. A., FAN, T. W., FIEHN, O., GOODACRE, R., GRIFFIN, J. L., HANKEMEIER, T., HARDY, N., HARNLY, J., HIGASHI, R., KOPKA, J., LANE, A. N., LINDON, J. C., MARRIOTT, P., NICHOLLS, A. W., REILY, M. D., THADEN, J. J. & VIANT, M. R. 2007. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics,* 3**,** 211-221.

SWANN, J. R., SPAGOU, K., LEWIS, M., NICHOLSON, J. K., GLEI, D. A., SEEMAN, T. E., COE, C. L., GOLDMAN, N., RYFF, C. D., WEINSTEIN, M. & HOLMES, E. 2013. Microbial-mammalian cometabolites dominate the age-associated urinary metabolic phenotype in Taiwanese and American populations. *J Proteome Res,* 12**,** 3166-80.

TANG, D. Q., ZOU, L., YIN, X. X. & ONG, C. N. 2014. HILIC-MS for metabolomics: An attractive and complementary approach to RPLC-MS. *Mass Spectrom Rev*.

TAUTENHAHN, R., BOTTCHER, C. & NEUMANN, S. 2008. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics,* 9**,** 504.

TAUTENHAHN, R., PATTI, G. J., RINEHART, D. & SIUZDAK, G. 2012. XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal Chem,* 84**,** 5035-9.

TENGSTRAND, E., LINDBERG, J. & ABERG, K. M. 2014. TracMass 2--a modular suite of tools for processing chromatography-full scan mass spectrometry data. *Anal Chem,* 86**,** 3435-42.

TRYGG, J. & WOLD, S. 2002. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics,* 16**,** 119-128.

TZOULAKI, I., EBBELS, T. M., VALDES, A., ELLIOTT, P. & IOANNIDIS, J. P. 2014. Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. *Am J Epidemiol,* 180**,** 129-39.

VAN DEEMTER, J. J., ZUIDERWEG, F. J. & KLINKENBERG, A. 1956. Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography. *Chemical Engineering Science,* 5**,** 271-289.

VAUGHAN, A. A., DUNN, W. B., ALLWOOD, J. W., WEDGE, D. C., BLACKHALL, F. H., WHETTON, A. D., DIVE, C. & GOODACRE, R. 2012. Liquid chromatography-mass spectrometry calibration transfer and metabolomics data fusion. *Anal Chem,* 84**,** 9848-57.

VENABLES, W. N., RIPLEY, B. D. & VENABLES, W. N. 2002. *Modern applied statistics with S,* New York, Springer.

VESELKOV, K. A., VINGARA, L. K., MASSON, P., ROBINETTE, S. L., WANT, E., LI, J. V., BARTON, R. H., BOURSIER-NEYRET, C., WALTHER, B., EBBELS, T. M., PELCZER, I., HOLMES, E., LINDON, J. C. &

References

NICHOLSON, J. K. 2011. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Anal Chem,* 83**,** 5864-72.

WAGNER, S., SCHOLZ, K., SIEBER, M., KELLERT, M. & VOELKEL, W. 2007. Tools in metabonomics: an integrated validation approach for LC-MS metabolic profiling of mercapturic acids in human urine. *Anal Chem,* 79**,** 2918-26.

WALDRAM, A., HOLMES, E., WANG, Y., RANTALAINEN, M., WILSON, I. D., TUOHY, K. M., MCCARTNEY, A. L., GIBSON, G. R. & NICHOLSON, J. K. 2009. Top-down systems biology modeling of host metabotype-microbiome associations in obese rodents. *J Proteome Res,* 8**,** 2361-75.

WANG, S. Y., KUO, C. H. & TSENG, Y. J. 2013. Batch Normalizer: a fast total abundance regression calibration method to simultaneously adjust batch and injection order effects in liquid chromatography/time-of-flight mass spectrometry-based metabolomics data and comparison with current calibration methods. *Anal Chem,* 85**,** 1037-46.

WANG, Y., XIAO, J., SUZEK, T. O., ZHANG, J., WANG, J., ZHOU, Z., HAN, L., KARAPETYAN, K., DRACHEVA, S., SHOEMAKER, B. A., BOLTON, E., GINDULYTE, A. & BRYANT, S. H. 2012. PubChem's BioAssay Database. *Nucleic Acids Res,* 40**,** D400-12.

WANG, Z., KLIPFELL, E., BENNETT, B. J., KOETH, R., LEVISON, B. S., DUGAR, B., FELDSTEIN, A. E., BRITT, E. B., FU, X., CHUNG, Y. M., WU, Y., SCHAUER, P., SMITH, J. D., ALLAYEE, H., TANG, W. H., DIDONATO, J. A., LUSIS, A. J. & HAZEN, S. L. 2011. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature,* 472**,** 57-63.

WANT, E. & MASSON, P. 2011. Processing and analysis of GC/LC-MS-based metabolomics data. *Methods Mol Biol,* 708**,** 277-98.

WANT, E. J., CRAVATT, B. F. & SIUZDAK, G. 2005. The expanding role of mass spectrometry in metabolite profiling and characterization. *Chembiochem,* 6**,** 1941-51.

WANT, E. J., WILSON, I. D., GIKA, H., THEODORIDIS, G., PLUMB, R. S., SHOCKCOR, J., HOLMES, E. & NICHOLSON, J. K. 2010. Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc,* 5**,** 1005-18.

WICKRAMASEKARA, S. I., ZANDKARIMI, F., MORRE, J., KIRKWOOD, J., LEGETTE, L., JIANG, Y., GOMBART, A. F., STEVENS, J. F. & MAIER, C. S. 2013. Electrospray Quadrupole Travelling Wave Ion Mobility Time-of-Flight Mass Spectrometry for the Detection of Plasma Metabolome Changes Caused by Xanthohumol in Obese Zucker (fa/fa) Rats. *Metabolites,* 3**,** 701-17.

WIKLUND, S., JOHANSSON, E., SJOSTROM, L., MELLEROWICZ, E. J., EDLUND, U., SHOCKCOR, J. P., GOTTFRIES, J., MORITZ, T. & TRYGG, J. 2008. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry,* 80**,** 115-122.

WILLIAMS, A. J. 2008. Public chemical compound databases. *Curr Opin Drug Discov Devel,* 11**,** 393-404.

WILM, M. 2011. Principles of electrospray ionization. *Mol Cell Proteomics*.

WISHART, D. S. 2011. Advances in metabolite identification. *Bioanalysis,* 3**,** 1769-82.

WISHART, D. S., KNOX, C., GUO, A. C., EISNER, R., YOUNG, N., GAUTAM, B., HAU, D. D., PSYCHOGIOS, N., DONG, E., BOUATRA, S., MANDAL, R., SINELNIKOV, I., XIA, J., JIA, L., CRUZ, J. A., LIM, E., SOBSEY, C. A., SHRIVASTAVA, S., HUANG, P., LIU, P., FANG, L., PENG, J., FRADETTE, R., CHENG, D., TZUR, D., CLEMENTS, M., LEWIS, A., DE SOUZA, A., ZUNIGA, A., DAWE, M., XIONG, Y., CLIVE, D., GREINER, R., NAZYROVA, A., SHAYKHUTDINOV, R., LI, L., VOGEL, H. J. & FORSYTHE, I. 2009. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res,* 37**,** D603-10.

WOLD, S., SJOSTROM, M. & ERIKSSON, L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* 58**,** 109-130.

WOLFF, J. C., ECKERS, C., SAGE, A. B., GILES, K. & BATEMAN, R. 2001. Accurate mass liquid chromatography/mass spectrometry on quadrupole orthogonal acceleration time-of-flight mass

analyzers using switching between separate sample and reference sprays. 2. Applications using the dual-electrospray ion source. *Anal Chem,* 73**,** 2605-12.

YANES, O., TAUTENHAHN, R., PATTI, G. J. & SIUZDAK, G. 2011. Expanding coverage of the metabolome for global metabolite profiling. *Anal Chem,* 83**,** 2152-61.

YOUSRI, N. A., KASTENMULLER, G., GIEGER, C., SHIN, S. Y., ERTE, I., MENNI, C., PETERS, A., MEISINGER, C., MOHNEY, R. P., ILLIG, T., ADAMSKI, J., SORANZO, N., SPECTOR, T. D. & SUHRE, K. 2014. Long term conservation of human metabolic phenotypes and link to heritability. *Metabolomics,* 10**,** 1005-1017.

ZELENA, E., DUNN, W. B., BROADHURST, D., FRANCIS-MCINTYRE, S., CARROLL, K. M., BEGLEY, P., O'HAGAN, S., KNOWLES, J. D., HALSALL, A., CONSORTIUM, H., WILSON, I. D. & KELL, D. B. 2009. Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. *Anal Chem,* 81**,** 1357-64.

ZHANG, T., WATSON, D. G., WANG, L., ABBAS, M., MURDOCH, L., BASHFORD, L., AHMAD, I., LAM, N. Y., NG, A. C. & LEUNG, H. Y. 2013. Application of Holistic Liquid Chromatography-High Resolution Mass Spectrometry Based Urinary Metabolomics for Prostate Cancer Detection and Biomarker Discovery. *PLoS One,* 8**,** e65880.

Appendices

Appendix 1

PCSOP.036

MRC-NIHR Phenome Centre

Document: **PCSOP.036**
Revision: 7

Author: Matthew R. Lewis
Effective: Sep 17, 2013 13:38

Generation of Urine Long Term Reference (LTR)

1. **Purpose**
   The purpose of this standard operating procedure (SOP) is to provide step-by-step instructions for the creation and aliquoting of a Long Term Reference (LTR) sample of human urine for use in UPLC-MS and NMR assays.
2. **Scope**
   Human urine is a potentially infectious (Class 2) biofluid commonly handled for study within the Phenome Centre. Standardized procedure for handling of unscreened human urine and its use in the creation of a homogenous LTR is required for the safety of laboratory personnel, assurance of data quality, and interpretability.
3. **Definition**
   The Long Term Reference (LTR) is a study-independent biofluid pool which is analyzed regularly across all experiments performed within the Phenome Centre to provide a biologically relevant quality control reference.
4. **Required Materials**
    1. Collection and pooled sample materials
        1. Ultrapure water (Milli-Q or equivalent) for rinsing collection tubes
        2. Nalgene 20 L polypropylene carboy with spigot
        3. Large stir plate with large stir bar
        4. 3x cases of 500mL PP Centrifuge Tubes with Plug Seal Cap, Sterile, 6/Pack, 36/Case (Corning product **#431123**)
    2. Centrifugation materials
        1. Eppendorf 5810R centrifuge and S-4-104 rotor
        2. 4x swing-bucket inserts (Eppendorf cat **#5825 745.000**)
    3. Dispensing materials
        1. 2x Eppendorf 5-25 mL Varispenser Plus (Eppendorf cat **#4961000047**)
        2. 2x 1 L bottles to fit Varispenser
        3. 2x small stir plates (for 500mL bottle)
        4. 2x small stir bars (for 500mL bottle)
        5. ~2200x 15 mL Corning CentriStar polypropylene centrifuge tubes (Corning product **#430791**)
        6. Sequential unique barcodes (-80C appropriate) for 15 mL Corning centrifuge tubes
5. **Procedure**
    1. Specimen collection

1. 500mL Corning centrifuge tubes are pre rinsed with ultrapure water (Milli-Q or equivalent).
2. Individual urine specimens are obtained in pre-rinsed 500mL Corning centrifuge tube from volunteers who have completed the donor consent form (PCDOC.009).
    1. Specimen volume in excess of 500 mL is decanted to the toilet by the specimen provider.
3. The sample is labeled with sequential number
4. 1mL of sample is removed for UPLC-MS to a 1.7mL eppendorf tube and transported to IRDB for analysis (and labeled with appropriate number).
5. 1mL of sample is removed for NMR to a 1.7mL eppendorf tube (and labeled with appropriate number).
6. The bulk 500mL specimen is transported to IRDB building and stored at 4C.

2. Specimen testing
    1. Individual specimen are analyzed overnight by NMR and UPLC-MS to ensure no PEG contamination or other cause for exclusion from the pooled reference.
3. Pooled sample homogenization
    1. The following morning, post analysis, outlying samples are disposed of according to (SOP.XXX)
    2. The remaining specimen are balanced according to volume and centrifuged at 4C for 15 min at max speed (3,171 x g as limited by Eppendorf swing bucket/insert type 5825745.000).
    3. Supernatant is carefully decanted by pouring into a pre-rinsed Nalgene 20 L polypropylene carboy with spigot, held in a refrigerator at 4C. This is the ***pooled sample.***
        1. The total number of combined specimens is recorded: _____.
    4. When 20 L pooled sample has been accumulated, a large stirbar is added to the carboy, and the carboy is loosely capped (not air-tight).
    5. The carboy is placed on a large stir plate and supported such that it is stable while stirring.
    6. The pooled sample is stirred with sufficient speed for sample homogenization for 5 minutes.
4. Pooled sample dispensing (NMR)
    1. While stirring, 1 L of pooled sample is dispensed to a clean 1L pyrex bottle with stir bar.
    2. The bottle is capped with an Eppendorf Varispenser.
    3. With gentle stirring, the pooled sample is dispensed from the container to sequentially barcoded Corning 15mL centrifuge tubes using a 5-25mL Eppendorf varispenser plus in set volumes of **7.8 mL**
        1. NMR requirements have been calculated as follows:
            1. 1 LTR per plate
            2. 13 plates / 1000 assay samples
            3. 13 LTR samples total * 600ul = **7.8 mL (per study)**
    4. Any aliquots less than 7.8 mL (ie those dispensed at the end of the batch) are disposed of according to the Biological Waste Disposal SOP (SOP_TBD)
    5. All tubes are tightly capped, racked in barcoded containers, and stored at -80C in the space designated for LTR storage.
5. Pooled sample dispensing (MS)
    1. While stirring, 1 L of pooled sample is dispensed to a clean 1L pyrex bottle with stir bar.
    2. The bottle is capped with an Eppendorf Varispenser.
    3. With gentle stirring, the pooled sample is dispensed from the container to sequentially barcoded Corning 15mL centrifuge tubes using a 5-25mL Eppendorf varispenser plus in set volumes of **11.5 mL**
        1. MS requirements have been calculated as follows:
            1. 8 LTR per plate
            2. 13 plates / 1000 assay samples

3. 104 LTR samples total * 110ul = **11.5 mL (per assay, per study)**
4. Any aliquots less than 11.5 mL (ie those dispensed at the end of the batch) are disposed of according to the Biological Waste Disposal SOP (SOP_TBD)
5. All tubes are tightly capped, racked in barcoded containers, and stored at -80C in the space designated for LTR storage.

6. Repeated centrifugation and dispensing for NMR and MS*
    1. Step 5.e (MS) is repeated 3x
    2. Step 5.d (NMR) is repeated 1x, followed by 4x repeats of step 5.d (MS) until the pooled sample has been entirely aliquoted.
    3. *Note: Because 4x MS assays are expected per every 1x NMR assay (per sample, per experiment), the steps above are repeated to generate 4x the MS aliquots but not to bias either LTR to the first or last batches.

6. **Related Documents**

    Document Number Title

7.

    #Revision History

Version	Date	Author	Comment
7	Sep 17, 2013 13:38	Lewis, Matthew R	Migrated to Confluence 5.3
6	Sep 17, 2013 13:38	Lewis, Matthew R	Migrated to Confluence 4.0
5	Sep 17, 2013 13:38	Lewis, Matthew R	
4	Sep 17, 2013 13:36	Lewis, Matthew R	minor edits
3	Aug 30, 2013 13:27	Lewis, Matthew R	
2	Jul 31, 2013 20:28	Lewis, Matthew R	minor edit
1	Jul 31, 2013 20:12	Lewis, Matthew R	

Appendix 2

R script for determining the maximum age of a sample during a continuous 20 plate analysis.

```
library(reshape2)
library(ggplot2)

  day.minutes<-60*24                    # definition of a day
  work.start<-9*60                      # ASSUMES WORK STARTS AT 9AM
  reload.start<-work.start+(3*60)       # ASSUMES 3H PREP TIME REQUIRED FOR PLATE BATCH, regardless
of size
  reload.end<-18*60                     # ASSUMES DAY ENDS AT 6PM
  reload.mid<-(reload.start+((reload.end-reload.start)/2))
  plates<-20                            # total number of plates to be analysed continuously
  batch<-96                             # sample batch size
  run.time<-seq(2.2, 31, by = .05)      # run times to assess
    run.time<-round(run.time, 2)        # fixes floating point errors
  duration<-(run.time*batch)            # calculates the duration of analysis, given a batch size.
  tally.duration<-0                     # dummy variable for later use
  tally.plates<-0                       # dummy variable for later use

# sets starting time for analysis of the first plate
# ASSUMES START TIME IS IN THE MIDDLE OF THE RELOAD LENGTH - IE 3:00.  Rationale is that it gives
the most room for drift in method time, either up or down in time
  time.started<-rep.int(reload.mid, length(run.time))


# NOTE: to obtain "minimisation of batch duration plot", run above code and code within this repeat
ONCE (do not allow repeat)
repeat{

#  if the run duration doesn't last until reload.start time next day, more plates are needed in the
batch
  more.plates<-as.vector(rep(1,length(run.time)))
  batch.duration<-duration

# add more plates to a batch to get short analyses over one night.
  repeat{

  # tests if the run goes into the next reload day (T) or falls short (F)
    day.lapse<-(batch.duration+time.started)>1440+reload.start

  # stops the loop when all runs at least go into the next reload day
    if (length(day.lapse)==sum(day.lapse)) break

  # adds a plate to each method where the batch does not lapse into the next day's reload time.
    more.plates[which(day.lapse==F)]<-((more.plates[which(day.lapse==F)])+1)

  # updates the time finished
    batch.duration<-more.plates*duration

  }

  #limits number of plates to the number specified in the cycles.
    more.plates[which((tally.plates+more.plates)>plates)]<-plates-
(tally.plates[which((tally.plates+more.plates)>plates)])
  # updates batch duration post limiting
    batch.duration<-more.plates*duration

  # plates that can be reloaded during the working day are done so immediatly before analysis.
  # batches that start after the working day are prepared at end of working day and must age extra,
beyond their duration.
  # however, batches must only wait since reload.end of the previous day, regardless of how many
days lapsed in batch
    # dummy string
    wait.time<-rep.int(0, length(run.time))
    # adds rest of night plus morning hours to those reloading before 720 min
    wait.time[which(time.started<reload.start)]<-
time.started[which(time.started<reload.start)]+(1440-reload.end)
```

Appendix 2

```
    wait.time[which(time.started>reload.end)]<-time.started[which(time.started>reload.end)]-
reload.end

    # calculates the max sample age as the duration of the batch analysis + the wait time between
prep and start
    plate.age<-batch.duration+wait.time


    #tally.duration<-cbind(tally.duration, batch.duration)
    tally.duration<-cbind(tally.duration, plate.age)
    tally.plates<-(tally.plates+more.plates)

    hours.plot<-batch.duration/60
    schedule<-as.data.frame(cbind(hours.plot, run.time, more.plates))
    ggplot(schedule, aes(hours.plot, run.time)) +
    geom_vline(xintercept = c(24, 48), colour = "blue", size = .5, linetype = "longdash") +
    geom_vline(xintercept = c(21, 45), colour = "red", size = .5, linetype = "longdash") +
    geom_vline(xintercept = c(27, 51), colour = "red", size = .5, linetype = "longdash") +
    geom_point(aes(colour = factor(more.plates))) +
    scale_colour_discrete(name = "Plates per batch") +
    scale_y_continuous(expand=c(0,0)) +  # eliminates margin on y axis
    xlab("batch duration (hours)") +
    ylab("analytical method duration (minutes)") +
    labs(title="Minimisation of batch duration") +
    theme_bw()

  # calculates the time of day the analysis will finish
    time.finished<-(batch.duration+time.started)-((floor((batch.duration+time.started)/1440)*1440))

  # adjust the new start time.
    time.started<-time.finished


 if (sum(tally.plates<plates)==0) break

}
# NOTE: to obtain "continuous analysis of twenty 96-well plates" plot, run all code above,
including the repeat, and code below

# removes seed column
  tally.duration<-tally.duration[,-1]
  max.age<-(apply(tally.duration, 1, max))
  #max.age<-rowMeans(tally.duration)
  max.age<-max.age/1440
  schedule<-as.data.frame(cbind(max.age, run.time))


  ggplot(schedule, aes(max.age, run.time)) +
  #geom_vline(xintercept = 1, colour = "darkgrey", size = .5) +
  #geom_vline(xintercept = 2, colour = "darkgrey", size = .5) +
  #geom_vline(xintercept = 3, colour = "darkgrey", size = .5) +
  geom_point(colour = "blue", alpha = 0.5, shape = 16) +
  #geom_hline(yintercept = 12, colour = "red", size = .5, linetype = "longdash") +
  #scale_x_continuous(expand=c(0,0)) +  # eliminates margin on x axis
  scale_y_continuous(expand=c(0,0)) +  # eliminates margin on y axis
  xlab("maximum sample age from preparation to analysis (days)") +
  ylab("analytical method duration (minutes)") +
  labs(title="Continuous analysis of twenty 96-well plates") +
  theme_bw()
```

Appendix 3

R script for pairwise feature matching (ROgroup).

```
# ROgroup
# Version 30
# Matthew R. Lewis

##########################################################################################
##########################################################################################
#################################----1ST TIME LOADING----#################################

# sets working directory
  #myDir = "C:/Users/Matthew R. Lewis/Dropbox/1_Phenome Center/Research and
Development/RunOrderExtraction/NetCDF"
  myDir = "D:/NPC Research/Urine Phase 1 Validation/ToF05Centroids.PRO/NetCDF"
  setwd(myDir)
  (WD <- getwd())

# loads required packages
  library(xcms)
  library(ggplot2)
  library(reshape2)
  library(pheatmap)
  library(gridExtra)

# load run order
  run.order <- read.csv("AnalysisOrder.csv", header = F)

# load initial template sample for feature detection and grouping
  first.sample<-xcmsRaw(as.character(run.order[1,1]))     # use this to start at the beginning
  #first.sample<-xcmsRaw(as.character(run.order[72,1]))   # project mid point is plate 5, samples
48 and 59.  select this to start at 48.

# sets the initial sample
  sample.count<-1                                          # use this when starting from the
beginning
  #sample.count <- 72                                      # use this when starting from the mid
point

 detailed.plotting <- F

##########################################################################################
##########################################################################################
####################----detect features, build template----####################

# perform centWave-based feature detection
# note: keep noise thresholds low - as the entire signal response drops, so will the noise level,
and real signals.
  peakpick <- function(sample.x){
    peaks<-findPeaks.centWave(sample.x,
      ppm=30,                              # maxmial tolerated m/z deviation in consecutive scans,
in ppm (parts per million)
      peakwidth=c(1,8),                    # Chromatographic peak width, given as range (min,max) in
seconds
      snthresh = 10,                       # Signal/Noise ratio: ([maximum peak intensity] -
[estimated baseline value]) / standard deviation of local chromatographic noise
      noise=1000,                          # centroids with intensity < noise are omitted from ROI
detection
      prefilter=c(8, 2000),                # Mass traces are only retained if they contain at least
x scans with intensity y
      mzCenterFun="wMean",                 # m/z centre of feature (wMean=intensity weighted mean of
the feature m/z values)
      integrate=2,                         # integration type (1=on bounds decided by waves, 2=on
raw data)
      verbose.columns=T)                   # provides additional peak metadata
      #sleep=.01,
      #scanrange=c(314,315)
      #nSlaves=6,                          # number of core processors
```

```
    return(peaks)
  }

  template<-peakpick(first.sample)

  # centWave very annoyingly returns duplicate features.  Sometime ALL measurements are duplicated,
sometimes just the m/z, RT, and into.
  # template<-unique(template) on the whole dataset works only where all measurements are
duplicated.
  # therefore, the dataset is stripped back to only what is used in matching, and unique entries
are retained.
  # also, presumably because of floating point errors... numbers require rounding too, before they
can be ID'd as uique.

  template <- cbind(round(template[,"mz"], digits=4), round(template[,"rt"], digits=3),
round(template[,"into"], digits=3))
  colnames(template) <- c("mz","rt","into")
  template<-unique(template)              # duplicate features (identical rows) are removed

##############################################################################################
##############################################################################################
####################----MATCHING PARAMETERS----####################

  half.mz.win<-0.002  # 1/2 error window (Daltons).  This should be the same value as that used in
peakpicking (ie. 30ppm)
  half.rt.win<-3      # 1/2 error window (seconds)
  #half.it.win<-30    # 1/2 error window (percent for log transformed it)

##############################################################################################
##############################################################################################
####################----RESERVE master matrices----####################

 master.ID<-(1:(nrow(template)))          # creates a matrix for storing matched feature IDs
 master.MZ<-(template[,"mz"])             # creates a matrix for storing matched feature m/z values
 master.RT<-(template[,"rt"])             # creates a matrix for storing matched feature RTs
 master.IT<-(template[,"into"])           # creates a matrix for storing matched feature ITs

##############################################################################################
##############################################################################################
####################----REPEAT FROM HERE----####################

repeat {

# load next candidate sample file.
  new.sample<-xcmsRaw(as.character(run.order[(sample.count+1),1]))    # chooses next sample in run
order
  #new.sample<-xcmsRaw("UrineVal1_p1control_RPOS_ToF05_SR96_AFAMM01.CDF")      # only used for
illustration where the candidate sample is specifically chosen.

# peakpick new sample file
  candidate<-peakpick(new.sample)

  candidate <- cbind(round(candidate[,"mz"], digits=4), round(candidate[,"rt"], digits=3),
round(candidate[,"into"], digits=3))
  colnames(candidate) <- c("mz","rt","into")

  candidate<-unique(candidate)            # centWave seems to produce a small number of duplicate
features (identical rows).  This removes them.


##############################################################################################
##############################################################################################
####################----set boundaries and create double-width bins for self-matching----
####################

# calculates upper and lower limits for self-matching (uses 2x windows to ensure all ambiguous
matches are clusters).

  upper.mz.t.dbl<-(template[,"mz"])+(half.mz.win*2)
  lower.mz.t.dbl<-(template[,"mz"])-(half.mz.win*2)
```

```
  upper.rt.t.dbl<-(template[,"rt"])+(half.rt.win*2)
  lower.rt.t.dbl<-(template[,"rt"])-(half.rt.win*2)

  upper.mz.c.dbl<-(candidate[,"mz"])+(half.mz.win*2)
  lower.mz.c.dbl<-(candidate[,"mz"])-(half.mz.win*2)

  upper.rt.c.dbl<-(candidate[,"rt"])+(half.rt.win*2)
  lower.rt.c.dbl<-(candidate[,"rt"])-(half.rt.win*2)

################################################################################
################################################################################
####################----INTRA-sample matching (clustering)----####################

  # herds are groups of features that match other features within the same dataset
  # note that mass error values should be in absolute terms, otherwise features may fall into more
than 1 herd as the absolute inclusion parameters shift.  ????

  herder <- function(dataset, lowerbounds.mz, upperbounds.mz, lowerbounds.rt, upperbounds.rt){

    herd.ID<-rep(NA, nrow(dataset))                                    # creates a reporting vector
where the position relates to the dataset ID and the value is unique to the herd
    herd.val<-1                                                        # makes designation for first
herd value

    for (i in 1:length(herd.ID)){
          if (is.na(herd.ID[i])==F) next                               # only runs calculation on
values that have no herd designation from inclusion in earlier loops

          matches<-which(                                              # performs mz and rt matching.
            dataset[i,"mz"]>lowerbounds.mz & dataset[i,"mz"]<upperbounds.mz &
            dataset[i,"rt"]>lowerbounds.rt & dataset[i,"rt"]<upperbounds.rt
          )

          if (length(matches)>1) {                                     # seeks boundaries of a herd
only if a match is not unique
            old.matches<-0                                             # primes the subsequent while
condition to run (via TRUE), as matches can not = 0
            loop.start<-1                                              # creates a val that can be
updated, avoiding the need to re-match values as matches grows
            while (length(matches)!=length(old.matches)){              # loops as long as new matches
are being added on.
              old.matches<-matches                                     # after above test, matches
becomes the old.matches, as matches is updated subsequently

              for (j in loop.start:length(matches)){                   # start with 1, unless updated
(see below) - avoids re-match values as matches grows
                matches<-append(matches,(which(                        # performs self-matching using
mz and rt on newly associated features to grow herd
                  dataset[matches[j],"mz"]>lowerbounds.mz & dataset[matches[j],"mz"]<upperbounds.mz
&
                  dataset[matches[j],"rt"]>lowerbounds.rt & dataset[matches[j],"rt"]<upperbounds.rt
                )))
                loop.start<-loop.start+1                               # updates starting value in for
loop with each value matched
              }
              matches<-matches[!duplicated(matches)]                   # updates "matches" by removing
duplicate values
            }
          } else herd.ID[i]<-0                                         # enters a 0 value as herd.ID
for things that are unique within the dataset

          if (length(matches)==1) next                                 # skips rest of loop if dataset
feature was self-unique.
          herd.ID[matches]<-herd.val                                   # enters a unique herd.ID
number for all associated features
          herd.val<-herd.val+1                                         # goes to the next herd.val
    }

    return(herd.ID)
  }
```

```
  candidate.herd.ID<-herder(candidate, lower.mz.c.dbl, upper.mz.c.dbl, lower.rt.c.dbl,
upper.rt.c.dbl)
  template.herd.ID<-herder(template, lower.mz.t.dbl, upper.mz.t.dbl, lower.rt.t.dbl,
upper.rt.t.dbl)

### Plot template features after self-matching

    cluster.count <- table(template.herd.ID)[-1]                    # Stores the number of features
in each cluster (except for cluster = 0) as a table.

  ### only TEMPLATE features which are in clusters.  colored (and alpha) by size of cluster (number
of participants).

    plot.x <- template[(which(template.herd.ID!=0)),"rt"]
    plot.y <-template[(which(template.herd.ID!=0)),"mz"]
    plot.z <- as.numeric(cluster.count[template.herd.ID])           # reports the cluster size for
each clustered feature in the template dataset.
    plot.a <-template[(which(template.herd.ID!=0)),"into"]

    plot.df<-as.data.frame(cbind(plot.x, plot.y, plot.z, plot.a))

  ### only TEMPLATE features which are NOT in clusters.

    plot2.x <- template[(which(template.herd.ID==0)),"rt"]
    plot2.y <-template[(which(template.herd.ID==0)),"mz"]
    plot2.a <-template[(which(template.herd.ID==0)),"into"]

    plot2.df<-as.data.frame(cbind(plot2.x, plot2.y, plot2.a))

  ### create scatter plot

     p <- ggplot(plot.df, aes(plot.x, plot.y, label=NULL)) +
              geom_point(data=plot2.df, aes(plot2.x, plot2.y, alpha=.5)) +     #  plots template
features NOT in clusters (on bottom layer)
              #geom_point(aes(alpha=plot.z, colour = log(plot.z))) +             # alpha related to
cluster size, highlighting larger clusters more
              geom_point(aes(alpha=.5, colour = log(plot.z))) +             # plots template
features in clusters (on top layer).  Note the log of the
              scale_colour_gradientn(colours = c("yellow","orange","blue"), name = "log(cluster
size)") +
              #scale_x_continuous(limits=c(5, 5.8), breaks=seq(5, 5.8, .2)) +
              #scale_y_continuous(limits=c(310.199, 310.203), breaks=seq(310.199, 310.203, 0.001))
+
              labs(y = "m/z", x = "retention time (seconds)", title = "Template feature clusters")
+
              theme_bw()

     ggsave(p,filename=as.character(paste(as.character(run.order[sample.count,]),
"TemplateFeatClusters.pdf", sep = "_")), width = 14, height = 9, units = "in") # ID will be the
unique identifier. and change the extension from .png to whatever you like (eps, pdf etc).




if (detailed.plotting==T) {


   clusters.per.cluster.size <- NULL
   features.per.cluster.size <- NULL

   for (i in 1:(max(cluster.count))){
    features.per.cluster.size <-
append(features.per.cluster.size,(sum(as.numeric(cluster.count[template.herd.ID])==i)))
    clusters.per.cluster.size <- append(clusters.per.cluster.size, (sum(cluster.count==i)))

   }

   cluster.tally <- as.data.frame(cbind((1:(max(cluster.count))), (clusters.per.cluster.size),
(features.per.cluster.size)))
   colnames(cluster.tally) <- c("size","clusters","features")
```

Appendix 3

```
   ggplot(cluster.tally) +
    #geom_bar(stat="identity", aes(x=size, y=clusters)) +
    geom_bar(stat="identity", aes(x=size, y=features)) +
    theme_bw()

  # percentage of dataset features that are in clusters of 10 or more.
  (sum(cluster.tally[(10:(nrow(cluster.tally))),"features"]) )/(nrow(template))*100
  # cluster numbers of clusters with 10 or more features
  as.numeric(which(cluster.count>=10))


### EIC and scatter plotting for all template clusters

  # controls the EIC window of m/z variance for plotting only
    ppm <- 60                                                    # m/z window oriented around
each feature in the cluster.  this should reflect the value used in centWave peak picking
      ppm.error <- ((ppm/1000000)*(template[,"mz"]))
      upper.mz.t.ppm<-(template[,"mz"])+(ppm.error/2)
      lower.mz.t.ppm<-(template[,"mz"])-(ppm.error/2)

  # controls the window of RT variance for plotting only
    RT.window <- 60                                              # RT window oriented around
each feature in the cluster
      upper.rt.display <- (template[,"rt"])+(RT.window/2)
      lower.rt.display <- (template[,"rt"])-(RT.window/2)


  for (i in 1:length(cluster.count)){                            # for every cluster
  #for (i in (as.numeric(which(cluster.count>=10)))){            # for clusters of size >= 10


    # actual feature values in the cluster
    mz.cluster.val <- template[(which(template.herd.ID==i)),"mz"]   # m/z values from all features
in cluster i
    rt.cluster.val <- template[(which(template.herd.ID==i)),"rt"]   # RT values from all features
in cluster i
    it.cluster.val <- template[(which(template.herd.ID==i)),"into"] # IT values from all features
in cluster i


    # cluster bounds (derived from the original matching error window)
    mz.cluster.min <- lower.mz.t.dbl[which(template.herd.ID==i)]    # lower m/z bounds, defined in
self-matching, of all features in cluster i
    mz.cluster.max <- upper.mz.t.dbl[which(template.herd.ID==i)]    # upper m/z bounds, defined in
self-matching, of all features in cluster i
    rt.cluster.min <- lower.rt.t.dbl[which(template.herd.ID==i)]    # lower m/z bounds, defined in
self-matching, of all features in cluster i
    rt.cluster.max <- upper.rt.t.dbl[which(template.herd.ID==i)]    # upper m/z bounds, defined in
self-matching, of all features in cluster i

    # display bounds (derived from the stated ppm/RT error window for visualisation)
    mz.display.min <- lower.mz.t.ppm[which(template.herd.ID==i)]    # lower m/z bounds with centre-
oriented 30ppm window of all features in cluster i
    mz.display.max <- upper.mz.t.ppm[which(template.herd.ID==i)]    # upper m/z bounds with centre-
oriented 30ppm window of all features in cluster i
    rt.display.min <- lower.rt.display[which(template.herd.ID==i)]  # lower m/z bounds with centre-
oriented 30ppm window of all features in cluster i
    rt.display.max <- upper.rt.display[which(template.herd.ID==i)]  # upper m/z bounds with centre-
oriented 30ppm window of all features in cluster i

    # cluster limits based on feature bounds
    cluster.bounds.mz.high <- max(mz.cluster.max)
    cluster.bounds.mz.low <- min(mz.cluster.min)
    cluster.bounds.rt.high <- max(rt.cluster.max)
    cluster.bounds.rt.low <- min(rt.cluster.min)

    # display limits based on feature bounds
    display.bounds.mz.high <- max(mz.display.max)
    display.bounds.mz.low <- min(mz.display.min)
```

Appendix 3

```
    display.bounds.rt.high <- max(rt.display.max)
    display.bounds.rt.low <- min(rt.display.min)

    # create EICs for each feature in cluster i
      eics <- NULL                                                  # place holder for reporting
eic intensity values (1 row = 1 feature's eic intensity values)
      for (j in 1:cluster.count[i]){                                # for every feature in cluster
i...
        eic <- rawEIC(first.sample,mz=c(mz.display.min[j], mz.display.max[j]))    # extracts eic
for feature j in cluster i using the display window for EIC
        eics<-rbind(eics, eic$intensity)                            # appends intensity values as
new row to eics.
      }
      eics.m <- melt(eics)                                          # reshapes the matrix to be
ggplot2 friendly
      colnames(eics.m) <- c("feature", "scan", "intensity")         # renames columns to
appropriate meaning
      r.time <-first.sample@scantime[eics.m[,"scan"]]               # creates scan-to-time
conversion
      eics.m <- cbind(eics.m, r.time)                               # appends scan-to-time
conversion to eics data-frame

    # finds the maximum EIC intensity value in the display area
      max.win.it <- max(eics.m[(which(((eics.m[,"r.time"])>cluster.bounds.rt.low) &
((eics.m[,"r.time"])<(cluster.bounds.rt.high)))),"intensity"])

    # extracts the RT of all features that are within the display RT window and within the selected
m/z range
      foo <- which((template[,"rt"]>(display.bounds.rt.low)) &
(template[,"rt"]<(display.bounds.rt.high)) & ((template[,"mz"]>(display.bounds.mz.low))) &
((template[,"mz"]<(display.bounds.mz.high))))
      others.rts <- template[(foo),"rt"]
      others.mzs <- template[(foo),"mz"]
      others.its <- template[(foo),"into"]


    # EIC (template)

     cluster.eic <- ggplot(eics.m, aes(r.time, intensity, color=feature)) +
                       annotate("rect", xmin = cluster.bounds.rt.low, xmax = cluster.bounds.rt.high,
ymin=-Inf, ymax=Inf, alpha = 0.05, fill = "blue") +     # blocks out cluster RT bounds
                       geom_line() +
                       scale_colour_gradient(low="orange", high="black") +
                       geom_vline(xintercept = c(others.rts), alpha=.5, colour="red", linetype =
"longdash") + # displays all features detected in plotting view
                       geom_vline(xintercept = c(rt.cluster.val), alpha=1, colour="blue", linetype =
"solid") + # places a sold line over dashed ones for clustered features

scale_x_continuous(limits=c((display.bounds.rt.low),(display.bounds.rt.high))) +
                       scale_y_continuous(limits=c(0, max.win.it)) +
                       labs(y = "Intensity", x = "Retention Time (seconds)", title =
as.character(paste("Reconstructed-ion chromatogram of all features in cluster",i, "\n mean cluster
m/z =", (round(mean(mz.cluster.val),4)), sep=" "))) +
                       theme_bw()
     cluster.eic

     ggsave(cluster.eic,filename=as.character(paste("Cluster", i, "EIC.pdf", sep = "_")), width =
14, height = 9, units = "in") # ID will be the unique identifier. and change the extension from
.png to whatever you like (eps, pdf etc).


  ### create scatter plot
     cluster.plot <- as.data.frame(cbind(mz.cluster.val, rt.cluster.val, it.cluster.val))
     display.plot <- as.data.frame(cbind(others.mzs, others.rts, others.its))

     cluster.scatter <- ggplot(cluster.plot, aes(rt.cluster.val, mz.cluster.val, label=NULL)) +
                          geom_rect(xmin = cluster.bounds.rt.low, xmax = cluster.bounds.rt.high,
ymin = cluster.bounds.mz.low, ymax = cluster.bounds.mz.high, alpha = 0.05, fill = "blue") +     #
blocks out cluster RT bounds
                          #annotate("rect", xmin = cluster.bounds.rt.low, xmax =
cluster.bounds.rt.high, ymin = 229.1540, ymax= 229.1604, alpha = 0.05, fill = "blue") +
```

```
                          geom_point(data = display.plot, aes(others.rts, others.mzs, alpha = 1
),colour = "red") +
                          geom_point(aes(alpha = 1),colour = "blue") +              # plots template
features in clusters (on top layer).  Note the log of the

scale_x_continuous(limits=c((display.bounds.rt.low),(display.bounds.rt.high))) +
                          scale_y_continuous(limits=c(display.bounds.mz.low,
display.bounds.mz.high)) +
                          labs(y = "m/z", x = "Retention time (seconds)", title =
as.character(paste("2D scatter plot of all features in cluster",i, sep=" "))) +
                          theme_bw()
     cluster.scatter

     ggsave(cluster.scatter,filename=as.character(paste("Cluster", i, "Scatter.pdf", sep = "_")),
width = 14, height = 3, units = "in") # ID will be the unique identifier. and change the extension
from .png to whatever you like (eps, pdf etc).
  }

  } else NULL

  #first.sample@env$profile

##############################################################################################
##############################################################################################
#######----set boundaries and create template bins for inter-sample matching----########

# calculates upper and lower limits for cross-matching

  upper.mz<-(template[,"mz"])+half.mz.win
  lower.mz<-(template[,"mz"])-half.mz.win

  upper.rt<-(template[,"rt"])+half.rt.win
  lower.rt<-(template[,"rt"])-half.rt.win

##############################################################################################
##############################################################################################
####################----INTER-sample matching----####################

# Match candidates to bins, creating pairwise connections.

# input = template, candidate, lower.mz, upper.mz, lower.rt, upper.rt

  matchcount<-rep(NA, nrow(candidate))                                  # seeds a vector containing the
# of template bins matched by the candidate
  bins.matched<-vector("list", nrow(candidate))                         # list of lists, each
containing the ID's of template bins matched by the candidate

  bincount<-rep(NA, nrow(template))
  bins<-vector("list", nrow(template))                                  # creates bins (lists) into
which candidate matches may be collected

  for (i in 1:nrow(candidate)){                                         # for every candidate
feature...

      #i=1570 # no matches
      #i=255  # 1 match only
      #i=1901 # multiple matches

        # matches candidates, one by one, into all bins.
          matches<-which(                                                       # "matches" are
template bin #s.  (i is the candidate feature #)
            candidate[i,"mz"]>lower.mz & candidate[i,"mz"]<upper.mz &       # candidate feature
must fall within mz bin
            candidate[i,"rt"]>lower.rt & candidate[i,"rt"]<upper.rt         # candidate feature
must ALSO fall within RT bin
          )

        # record info on # and ID of template bins matched by the candidate
          # matchcount: 0  = no matches in the template; 1 = unique match to template bin; >1=
multiple bins matched
```

```
          matchcount[i]<-length(matches)              # # of template bins that do contain the
candidate feature
          bins.matched[[i]]<-matches                  # stores ID's of all template bins that the
candidate feature fall into

        if (length(matches)==0) next else NULL        # if there are no matches, then skip to next
candidate.  Nothing to place in bins.

        # record info on how many/which candidates are falling into each bins
          for (m in 1:length(matches)){               # for every match (template bin) listed...
            bins[[matches[m]]]<-append(bins[[matches[m]]], i)        # append (to not overwrite)
candidate IDs (i) into all matched bins.
          }
  }

  # if candidate falls into  0 template bins, it is unmatched and will need to be appended to the
master.ID list as a new entry.
    #orphans<-which(matchcount==0)       # candidate ID's that do not match any template bins

  # counts number of candidate features in each bin
    for (i in 1:length(bins)){
        bincount[i]<-length(bins[[i]])
    }

### PLOTTING

#Candidate matches to template bins

matchcount.type <- matchcount
matchcount.type[which(matchcount==0)] <- "No matches"
matchcount.type[which(matchcount==1)] <- "Single match"
matchcount.type[which(matchcount>1)] <- "Multiple matches"

matchcount.df <-as.data.frame(cbind(matchcount, matchcount.type))

p <- ggplot(data = matchcount.df, aes(factor(matchcount.type))) +
      geom_bar(width=.8, fill="#707070") +
      #coord_polar(theta="y") +
      coord_flip() +
      #ggtitle("Template bins matched by candidate features") +
      theme(panel.grid.major.y = element_blank(),
          panel.grid.minor.y = element_blank(),
          panel.grid.major.x = element_line(colour="#909090"),
          panel.grid.minor.x = element_line(colour="#909090", linetype = "dotted"),
          panel.background = element_rect(fill = "white"),
          axis.ticks = element_blank(),
          axis.title.x = element_blank(),
          axis.title.y = element_blank(),
          axis.text.y = element_text(size=14))

ggsave(p,filename=as.character(paste(as.character(run.order[sample.count,]),
"CandidateFeatureMatches.pdf", sep = "_")), width = 7, height = 2, units = "in")

#Template matches to candidate features

bincount.type <- bincount
bincount.type[which(bincount==0)] <- "No matches"
bincount.type[which(bincount==1)] <- "Single match"
bincount.type[which(bincount>1)] <- "Multiple matches"

bincount.df <-as.data.frame(cbind(bincount, bincount.type))

p <- ggplot(data = bincount.df, aes(factor(bincount.type))) +
      geom_bar(width=.8, fill="#707070") +
      #coord_polar(theta="y") +
      coord_flip() +
      #ggtitle("Template bins matched by candidate features") +
      theme(panel.grid.major.y = element_blank(),
          panel.grid.minor.y = element_blank(),
          panel.grid.major.x = element_line(colour="#909090"),
          panel.grid.minor.x = element_line(colour="#909090", linetype = "dotted"),
```

```
          panel.background = element_rect(fill = "white"),
          axis.ticks = element_blank(),
          axis.title.x = element_blank(),
          axis.title.y = element_blank(),
          axis.text.y = element_text(size=14))

ggsave(p,filename=as.character(paste(as.character(run.order[sample.count,]),
"TemplateFeatureMatches.pdf", sep = "_")), width = 7, height = 2, units = "in")



matchcount.return <- NULL

for (i in 1:length(matchcount)){
  if (matchcount[i]==1) {
    matchcount.return <- append(matchcount.return, (bincount[bins.matched[[i]]]))
  } else NULL
}

(sum(matchcount.return==1)/length(matchcount.return))*100 # number of single matches that are
returned as signle matches (mutual)


################################################################################
################################################################################
##############################----DEFINE COMMUNITIES...----#####################################

### COMMUNITY FUNCTIONS
#1
# takes a set of features (ie. candidate), pulls their herd designations, and appends other
(candidate) features from those herds.
extend.features <- function(herd.ID, features){
  #herd.ID = candidate.herd.ID
  #features = compile.candidates

  bar<-herd.ID[features]                        # returns the (ie. CANDIDATE) herd id's of matched
(ie. candidate) features
  bar<-bar[!duplicated(bar)]                    # removes redundancy from the candidate herd list
  bar<-bar[which(bar>0)]                        # removes herd=0 herds

  if (length(bar)==0) (extended.features<-features) else NULL   # if bar is zero, the candidates
must be of herd=0.  Therefore, there is no extension by herd, and extended.candidates is just the
compiled candidates list as before

  if (length(bar)==1) (extended.features<-which(herd.ID==bar)) else NULL # if all candidates are of
a single herd, then find all candidates from that herd only.

  if (length(bar)>1) extended.features<-NULL else NULL
  if (length(bar)>1) (for (r in 1:length(bar)){                          # if candidates are of
multiple herds, compile each herd's features iterativly.  excludes herd=0
      (extended.features<-append(extended.features, which(herd.ID==bar[r])))      # redundancy is
solved in a following step below
  }) else NULL

  # in any case where bar!=0, the original compile.candidates list is appended to the extended list
in case herd=0 candidates were removed.
  if (length(bar)!=0) (extended.features<-append(extended.features, features)) else NULL
  if (length(bar)!=0) (extended.features<-extended.features[!duplicated(extended.features)]) else
NULL        #  List redundancy is again reduced (also important for step above).

  return(extended.features)
}

#2
# gathers all features matched to the feature subset (ie. returns candidate features matching
template feature input)
compile.features <- function(feature.subset, match.list){

  #foo =  feature.subset
  #bins = matchlist
```

```
  reporter<-0                                          # seeds reporting vector for appending
results
  for (f in 1:length(feature.subset)){                 # repeat for every template ID belonging
to the selected TEMPLATE herd.
      reporter<-append(reporter, match.list[[feature.subset[f]]])   # compile all candidate matches
belonging to all template bins in the selected TEMPLATE herd
  }
  reporter<-reporter[-1]                               # unseeds vector
  reporter<-reporter[!duplicated(reporter)]            #remove redundant candidates from compiled
list

  #compile.candidates = reporter
  return(reporter)
}

###/ COMMUNITY FUNCTIONS

# assign a community ID to all associated features in both template and candidate lists.  community
ID is shared between datasets.
  template.community.ID<-rep(NA, nrow(template))       # creates reporting vector (position =
template feature) for communiy groups
  candidate.community.ID<-rep(NA, nrow(candidate))     # creates reporting vector (position =
candidate feature) for communiy groups

  community.val<-1                                     # sets the first community value to 1.
updated at end of the following for loop.

  for (i in 1:max(template.herd.ID)){                  # loop sequentially for each herd,
starting at 1 so no need to worry about 0 herd val.
      #i=1 # all bins contain 0 matches
      #i=10 # bins all contain unique matches.
      #i=1940 #  herd has 44 template members
      #i= XXXX mixed herd (including 0) example needed

    # skip herd if already assigned to a community.
    if ((length(na.omit(template.community.ID[which(template.herd.ID==i)]))!=0) &
((sum(is.na(template.community.ID[which(template.herd.ID==i)]))>0))) print("WARNING") else NULL
    if (length(na.omit(template.community.ID[which(template.herd.ID==i)]))!=0) next else NULL

    feature.subset<-which(template.herd.ID==i)         # gives template ID's that belong to the
selected TEMPLATE herd

    # 1) finds all candidates which match all feature.subset template bins.
    # 2) extends candidates to all those in herds of bin-matched candidates from #1.
    # 3) finds all template bins which match all candidates from #2
    # 4) extends template bins to all those in herds of candidate-matched bins from #3 (necessary
to extend here because each loop starts with a single template herd)
    # 5) tests to see if # of template bins has grown in the last 4 steps.  If not, breaks loop.
If so, repeats from 1 after updating feature.subset.
    repeat{
      compiled.candidates<-compile.features(feature.subset, bins)         # returns all candidate
features matching all template bins in template herd i
      if (length(compiled.candidates)==0) break else NULL                 # if all template bins in
the template herd are matchless, break the repeat

      # extends the candidate feature subset to all those associated with a herd of the
compile.candidates set.
      extended.candidates<-extend.features(candidate.herd.ID, compiled.candidates)

      #candidate[extended.candidates,]

      compiled.bins<-compile.features(extended.candidates, bins.matched)  # returns all template
bins matching all candidate features across involved candidate herds (above)

      compiled.bins<-(append(feature.subset, compiled.bins))              # in case any original
template features were lost from the new list.  not sure this is necessary.
      compiled.bins<-compiled.bins[!duplicated(compiled.bins)]            #remove redundancy

      extended.bins<-extend.features(template.herd.ID, compiled.bins)     # extends the template
bins to all those associated with the herd(s) of the compile.bins set.
```

```
      # update plot as features are added???

      # if extended bin list has the same # of bins as the original feature subset, it is not
growing.  Repeat can be broken.  If it grows, update feature.subset and repeat.
      if (length(feature.subset)==length(extended.bins)) break else (feature.subset<-extended.bins)
    }

    if (length(compiled.candidates)==0) next else NULL                    # if all template bins in
the template herd are matchless, skip to next herd (continued from above)

    # checks that all features values are the same.  This should always be true - havn't found a
counter example yet.
    if (sum((sort(feature.subset, decreasing = FALSE))==(sort(extended.bins, decreasing =
FALSE)))!=length(feature.subset)) print("mismatch in features....") else NULL


    if ((sum(na.omit(template.community.ID[extended.bins])))!=0) print("overwriting") else NULL  #
checks the candidate.community.ID slots to ensure they're empty...
    template.community.ID[extended.bins]<-community.val                   # stores community ID in
template community list
    if ((sum(na.omit(candidate.community.ID[extended.candidates])))!=0) print("overwriting") else
NULL  # checks the template.community.ID slots to ensure they're empty...
    candidate.community.ID[extended.candidates]<-community.val            # stores community ID in
candidate community list
    community.val<-community.val+1                                        # update community.val

  }

  # basic check to make sure that max number of community values reported to candidate and template
match
    if (max(na.omit(template.community.ID))==max(na.omit(candidate.community.ID))) NULL else
print("WARNING: max community values mismatch")

  # output = template.community.ID and candidate.community.ID

################################################################################################
################################################################################################
###############----resolving free herds from the candidate list----#######################

# herd=0 template bins may still match candidate features with herd designations.  Those candidates
may in turn match template features, but those must (?) also be herd=0 or the community would have
been picked up in the first place.  Such "closed systems" go undetected by the method above.  Needs
to be fixed.

# all herds (template or candidate) with at least one cross match (to template/candidate) need to
be counted as a community.
# thus a community is defined as a series of features with at least one self match (in EITHER
template/candidate datasets) and at least one cross match.
# all template herds have been taken care of above, in order.
# here, the remaining candidate herds not assigned to communities are inspected for potential
community building in a process that is the mirror image to above.

# to inspect candidate herds NOT a part of any community:

# candidates with a herd designation ==1
#as.numeric(candidate.herd.ID!=0)
# candidates withOUT a community designation = 1
#as.numeric(is.na(candidate.community.ID))

#((as.numeric(candidate.herd.ID!=0))+(as.numeric(is.na(candidate.community.ID))))

#0 = herd=0, but with community.  Must have been a dead-end match to a herded template feature.
#1= herd=0, without community OR with herd, with community.  Independent or community-based.
#2 = with herd, without community. This is what we want to investigate.
   #Only OK if all features in the herd have 0 matches to template features.
   wHwoC.can<-
which((((as.numeric(candidate.herd.ID!=0))+(as.numeric(is.na(candidate.community.ID)))))==2)  #
candidate features that have a herd but no community.

   free.herds.can<-candidate.herd.ID[wHwoC.can]                           # candidate herds which
have NO community associated.
```

Appendix 3

```
   # candidate.community.ID[lost.can]                                    # just checking that the
above line is true...
   free.herds.can<-free.herds.can[!duplicated(free.herds.can)]           # removes duplicates
called by multiple candidate features in the same herd.

   # note that community.val is already primed from the previous step.

   for (i in 1:length(free.herds.can)){            # for each herd in the free herds list.  Note
that from the implementation above, "i" becomes "free.herds.can[i]"


    # skip herd if already assigned to a community.
    if ((length(na.omit(candidate.community.ID[which(candidate.herd.ID==free.herds.can[i])]))!=0) &
((sum(is.na(candidate.community.ID[which(candidate.herd.ID==free.herds.can[i])]))>0)))
print("WARNING") else NULL
    if (length(na.omit(candidate.community.ID[which(candidate.herd.ID==free.herds.can[i])]))!=0)
next else NULL

    feature.subset<-which(candidate.herd.ID==free.herds.can[i])         # gives candidate ID's that
belong to the selected CANDIDATE herd

    # 1) finds all template features which match all feature.subset candidate features
    # 2) extends template features to all those in herds of candidate-matched template features
from #1.
    # 3) finds all candidate features which match all template features from #2
    # 4) extends candidate features to all those in herds of template-matched features from #3
(necessary to extend here because each loop starts with a single candidate herd)
    # 5) tests to see if # of candidate features has grown in the last 4 steps.  If not, breaks
loop.  If so, repeats from 1 after updating feature.subset.
    repeat{
      compiled.templates<-compile.features(feature.subset, bins.matched)         # returns all
template features matching all candidate features in candidate herd free.herds.can[i]
      if (length(compiled.templates)==0) break else NULL                # if all template bins in
the template herd are matchless, break the repeat

      # extends the template feature subset to all those associated with a herd of the
compile.templates set.
      extended.templates<-extend.features(template.herd.ID, compiled.templates)

      #candidate[extended.candidates,]

      compiled.cans<-compile.features(extended.templates, bins)  # returns all template bins
matching all candidate features across involved candidate herds (above)

      compiled.cans<-(append(feature.subset, compiled.cans))
      compiled.cans<-compiled.cans[!duplicated(compiled.cans)]

      extended.cans<-extend.features(candidate.herd.ID, compiled.cans)     # extends the candidate
features to all those associated with the herd(s) of the compile.cans set.

      # update plot as features are added???

      # if extended candidate list has the same # of candidates as the original feature subset, it
is not growing.  Repeat can be broken.  If it grows, update feature.subset and repeat.
      if (length(feature.subset)==length(extended.cans)) break else (feature.subset<-extended.cans)
    }

    if (length(compiled.templates)==0) next else NULL                  # if all candidate features
in the candidate herd are matchless, skip to next herd (continued from above)

    # checks that all features values are the same.  This should always be true - havn't found a
counter example yet.
    if (sum((sort(feature.subset, decreasing = FALSE))==(sort(extended.cans, decreasing =
FALSE)))!=length(feature.subset)) print("mismatch in features....") else NULL

    if ((sum(na.omit(candidate.community.ID[extended.cans])))!=0) print("overwriting") else NULL  #
checks the candidate.community.ID slots to ensure they're empty...
    candidate.community.ID[extended.cans]<-community.val                  # stores community ID in
candidate community list
    if ((sum(na.omit(template.community.ID[extended.templates])))!=0) print("overwriting") else
NULL  # checks the template.community.ID slots to ensure they're empty...
```

Appendix 3

```
    template.community.ID[extended.templates]<-community.val            # stores community ID in
template community list
    community.val<-community.val+1                                      # update community.val

  }

  # basic check to make sure that max number of community values reported to candidate and template
match
    if (max(na.omit(template.community.ID))==max(na.omit(candidate.community.ID))) NULL else
print("WARNING: max community values mismatch")

##############################################################################################
##############################################################################################
#################################----PLOTTING BREAK!----#######################################

# for plotting, find feature cluster of interest (m/z = 310.202)

foo<-as.data.frame(cbind((candidate[(which(candidate[,"mz"]>310.19 &
candidate[,"mz"]<310.21)),"mz"]),
      (candidate[(which(candidate[,"mz"]>310.19 & candidate[,"mz"]<310.21)),"rt"]/60),
      ((candidate[(which(candidate[,"mz"]>310.19 & candidate[,"mz"]<310.21)),"into"])),
      (candidate.community.ID[which(candidate[,"mz"]>310.19 & candidate[,"mz"]<310.21)]),
      (which(candidate[,"mz"]>310.19 & candidate[,"mz"]<310.21))))

foo[(which(is.na(foo[,4]))),4]<-0 # sets NAs as 0's in community column to avoid problems with
plotting

foo<-(cbind(foo,(rep("C",nrow(foo)))))
colnames(foo)<-c("mz","rt","intensity","community","featureID", "origin")

bar<-as.data.frame(cbind((template[(which(template[,"mz"]>310.19 & template[,"mz"]<310.21)),"mz"]),
      (template[(which(template[,"mz"]>310.19 & template[,"mz"]<310.21)),"rt"]/60),
      ((template[(which(template[,"mz"]>310.19 & template[,"mz"]<310.21)),"into"])),
      (template.community.ID[which(template[,"mz"]>310.19 & template[,"mz"]<310.21)]),
      (which(template[,"mz"]>310.19 & template[,"mz"]<310.21))))

bar[(which(is.na(bar[,4]))),4]<-0 # sets NAs as 0's in community column to avoid problems with
plotting

bar<-(cbind(bar,(rep("T",nrow(bar)))))
colnames(bar)<-c("mz","rt","intensity","community","featureID", "origin")

pair<-rbind(foo, bar)
plot.annotation<-paste(pair[,"origin"],"(",pair[,"featureID"],")", sep="")
pair<-cbind(pair,plot.annotation)

  p1 <- ggplot(pair, aes(rt, mz, label=plot.annotation))
  #p1 <- ggplot(pair, aes(rt, mz, label=community))

  p1<- p1 +
    geom_point(aes(size = log(intensity), alpha=.1, colour = log(intensity)), xlim = c(5, 5.8),
ylim = c(310, 311)) +
    scale_colour_gradientn(colours = c("yellow","orange","blue")) +
    geom_text(size=3.5, hjust=.5, vjust=1.7) +
    #geom_text(size=3.5, hjust=-0.2, vjust=.3) +
    scale_x_continuous(limits=c(5, 5.8), breaks=seq(5, 5.8, .2)) +
    scale_y_continuous(limits=c(310.199, 310.203), breaks=seq(310.199, 310.203, 0.001)) +
    labs(x = "Retention time (minutes)", y = "m/z") +
    theme_bw()

  ggsave(p1,filename="selectCommunity.pdf", width = 9, height = 4.5, units = "in") # ID will be the
unique identifier. and change the extension from .png to whatever you like (eps, pdf etc).

    # overlay EIC????



##############################################################################################
##############################################################################################
#######----community match scoring and linear micro-alignment----############
```

```
template.matches<-rep(NA, nrow(template))       # creates a reporting vector for matches emerging
from this loop.  Gets filled with candidate ID values.
candidate.matches<-rep(NA, nrow(candidate))     # creates a reporting vector for matches emerging
from this loop.  Gets filled with template ID values.

match.tally.t<-NULL        # starts a tally of template features which have been definitivly
matched (successfully or otherwise)
match.tally.c<-NULL        # starts a tally of candidate features which have been definitivly
matched (successfully or otherwise)

for (g in 1:max(na.omit(c(candidate.community.ID, template.community.ID)))){   # loop for every
community ID

  # g=800     (SR 1 and 2)
  # g=627  (SR 1 and 16... first and last from plate 1)

  candidate.community<-which(candidate.community.ID==g)      # returns all candidates feature ID's
that match the given community number
  template.community<-which(template.community.ID==g)        # returns all template feature ID's
that match the given community number

  match.tally.t<-append(match.tally.t, template.community)   # notes that the template features in
this community underwent a round of matching and are "spent"
  match.tally.c<-append(match.tally.c, candidate.community)  # notes that the candidate features in
this community underwent a round of matching and are "spent"

  # returns the matching status of all features in the community, in matrix form (1=match, NA= no
match)
    match.matrix<-matrix(nrow=length(candidate.community), ncol=length(template.community))     #
creates a matrix of NAs for palceholders.
    for (i in 1:length(candidate.community)){
        match.matrix[i,(match(bins.matched[[candidate.community[i]]], template.community))]<-1  #
places a 1 where a candidate feature matches a bin.  0 where it does not.
    }
    match.matrix


  # function to calculate residual values between all candidate and template features in the
community
    get.residuals<- function(candidate.community, template.community, candidate.vals, data.type){
      scores<-matrix(nrow=length(candidate.community), ncol=length(template.community))
      for (i in 1:length(template.community)){                                  # for each bin...
          scores[,i]<-(template[template.community[i],data.type])-candidate.vals      # one
template bin minus all candidates
      }
      scores<-abs(scores)
      scores
      return(scores)
    }

if (detailed.plotting==T){
  # PLOTTING ONLY
  # plotting of RT differences w/o alignment

      foo.rt.scores<-get.residuals(candidate.community,                     # computes RT residuals
for modulated candidate RT vals (against template vals)
      template.community,
      candidate.vals=(candidate[candidate.community,"rt"]),                 # original vals are
explicitly called here
      data.type="rt")

      ### non-corrected RT scores
      rt.scores.plot <- foo.rt.scores+0.0000001 # adding a small value avoids the occurance of -Inf
return from log transformation of a zero residual
      rownames(rt.scores.plot) <-candidate.community
      colnames(rt.scores.plot) <-template.community

      pheat.col <- colorRampPalette(c("white","yellow", "orange", "purple"))(100)
      log.seq<-rev(1 * 1.1^(0:100))
      pheat.breaks<-(max(rt.scores.plot)/log.seq)
```

```
      pheatmap((rt.scores.plot),
        color=pheat.col,
        breaks=pheat.breaks,
        cluster_rows=F,
        cluster_cols=F,
        scale="none",
        show_rownames=T,
        show_colnames=T,
        main="Residual RT",
        display_numbers=T,
        fontsize=16,
        number_format="%.2f")
} else NULL



  # COMMUNITY-BASED MICRO-ALIGNMENT
  # This approach evaluates the total* RT residual for all candidates in the community (* total =
sum of minimum values for each candidate)
  # Move all candidate RT's by a set fraction of the original RT window
  # Find new minima (minima may change as new features match better)
  # Sum all new minima
  # Iterate this process for each modulation of the RT values, and look for a minimum value in the
sum of each, indicating best overall fit for the community
  if (length(candidate.community)>1 & length(template.community)>1) {       # miro-align only if
there are >1 template and candidate features in the community.

    #modulation<-seq(from = -(half.rt.win), to = half.rt.win, by =.1)        # calls on original rt
error wondow to define bounds, and sets steps for modulation of RTs   ORIGINAL VALUE
    modulation<-seq(from = -(2*half.rt.win), to = 2*half.rt.win, by =.1)        # calls on original
rt error wondow to define bounds, and sets steps for modulation of RTs   THESIS VISUALISATION
    min.val.sum<-NULL                                                          # creates reporting
vector
    min.bin.matrix<-NULL                                                       # creates reporting
vector

    for (i in 1:length(modulation)){                                           # for every set step in
modulating candidate RT's...

      mod.rt.vals<-(candidate[candidate.community,"rt"])+modulation[i]      # ... modulates
candidate RT values

      mod.rt.scores<-get.residuals(candidate.community,                        # computes RT residuals
for modulated candidate RT vals (against template vals)
      template.community,
      candidate.vals=mod.rt.vals,                                              # modulated vals are
explicitly called here
      data.type="rt")

    # computes the minimum residual value for each candidate (each matrix row)
      min.val<-NULL                                                            # creates reporting
vector for all candidate minimum values
      min.bin<-NULL                                                            # creates reporting
vector for template bins corresponding to minimum values
      for (s in 1:nrow(mod.rt.scores)){                                        # For every candidate
(matrix row)...
        min.val<-append(min.val, min(mod.rt.scores[s,]))                       # ...report the minimum
residual value...
        min.bin<-append(min.bin, which.min(mod.rt.scores[s,]) )                # ... and the template
bin corresponding to that minimimum value.
      }

      min.val.sum<-append(min.val.sum, sum(min.val))                           # sums the minimum
values of all candidates, and appends it to this vector.
      min.bin.matrix<-rbind(min.bin.matrix, min.bin)                           # corresponding
candidate to template matching patterns that produced the above min.val.sum minima.
    }

    rownames(min.bin.matrix) <- modulation                                     # renames the
min.bin.matrix rows with the values used in RT modulation
```

Appendix 3

```
    if (detailed.plotting==T){
      # plots the modulation value (added to candidate RT) vs. the sum of all minimum residuals
across all candidates

        modulation.plot.vals <- as.data.frame(cbind(modulation, min.val.sum))
        modulation.plot <- ggplot(modulation.plot.vals, aes(x=modulation, y=min.val.sum)) +
                           geom_point() +
                           geom_line() +
                           scale_x_continuous(breaks=seq((min(modulation)), (max(modulation)),
0.5)) +
                           labs(x = "Candidate cluster adjustment (seconds)", y = "Sum of minimum
residuals (seconds) ") +
                           theme_bw()

        ggsave(modulation.plot,filename=as.character(paste("Community", g, "alignment.pdf", sep =
"_")), width = 14, height = 9, units = "in")

        #  plot template community eic in blue, corrected eic in black, and uncorrected eic in
transparent grey.

    } else NULL

    # if there are two or more minima, choose the one originating from the least modulation to the
candidate dataset.
    if (length(modulation[which.min(min.val.sum)])>1) print("two minimum values in candidate RT
modulation.") else NULL
    # min(abs(modulation[which.min(min.val.sum)])) #this picks the right one, but looses the sign.
have to fix this....

    rt.correction<-modulation[which.min(min.val.sum)]                    # reports the
correction value that was added to all community candidate rt's to achieve optimal RT match.

    best.community.match<-min.bin.matrix[(which.min(min.val.sum)),]      # reports the matching
structure (position=candate, value = template) that produced the optimal RT match.

  } else NULL

  # if either candidate or template have only 1 feature...
  if (length(candidate.community)==1 | length(template.community)==1) {

    # runs function to return residual values for the candidate/template feature matrix.
    rt.scores<-get.residuals(candidate.community,
      template.community,
      candidate.vals=(candidate[candidate.community,"rt"]),              # specifies the
candidate values to be matched (explicit so they can be later modulated)
      data.type="rt")                                                    # sets the data type in
the template that candidate.vals are up against

  } else NULL

  # if both candidate and template have more than 1 feature...
  if (length(candidate.community)>1 & length(template.community)>1) {
    # FROM ABOVE: RE-runs RT scoring function with aligned retention time..
    rt.scores<-get.residuals(candidate.community,
      template.community,
      candidate.vals=((candidate[candidate.community,"rt"])+rt.correction), # note that these are
modified as determined to be optimal above in alignment.
      #candidate.vals=(candidate[candidate.community,"rt"]),              # use this instead to
see unaligned results. ie. for dev comparison, plotting, etc.
      data.type="rt")

  } else NULL

  # mz scores must be kept because in rare cases, it is used to resolve best matches where both rt
and it are identical
  # runs function to return residual values for the candidate/template feature matrix.
  mz.scores<-get.residuals(candidate.community,
    template.community,
    candidate.vals=(candidate[candidate.community,"mz"]),                # specifies the
candidate values to be matched (explicit so they can be later modulated)
```

```
    data.type="mz")                                                    # sets the data type in
the template that candidate.vals are up against

  # IT (log) residuals
  it.scores<-matrix(nrow=length(candidate.community), ncol=length(template.community))
  for (i in 1:length(template.community)){                # for each bin...
      it.scores[,i]<-(log(template[template.community[i],"into"]))-
(log(candidate[candidate.community,"into"]))      # one template bin minus all candidates
  }
  it.scores<-abs(it.scores)

if (detailed.plotting==T){
  ### PLOTTING (requires "pheatmap" package)
    # original matches

      match.matrix.plot<-match.matrix
      match.matrix.plot[is.na(match.matrix.plot)]<-0
        rownames(match.matrix.plot) <-candidate.community
        colnames(match.matrix.plot) <-template.community

       pheatmap(match.matrix.plot,
        cluster_rows=F,
        cluster_cols=F,
        scale="none",
        show_rownames=T,
        show_colnames=T,
        main="Inter-sample matches",
        display_numbers=F,
        fontsize=16,
        color = c("grey","white"))
        #filename = "Matches.matrix.pdf")

    # corrected RT scores
      rt.scores.plot <- rt.scores+0.0000001 # adding a small value avoids the occurance of -Inf
return from log transformation of a zero residual
        rownames(rt.scores.plot) <-candidate.community
        colnames(rt.scores.plot) <-template.community

        pheat.col <- colorRampPalette(c("white","yellow", "orange", "purple"))(100)
        log.seq<-rev(1 * 1.1^(0:100))
        pheat.breaks<-(max(rt.scores.plot)/log.seq)

        pheatmap((rt.scores.plot),
          color=pheat.col,
          breaks=pheat.breaks,
          cluster_rows=F,
          cluster_cols=F,
          scale="none",
          show_rownames=T,
          show_colnames=T,
          main="Residual RT (post-alignment)",
          display_numbers=T,
          fontsize=16,
          number_format="%.2f")

    # corrected masked RT scores
      rt.scores.plot <- rt.scores+0.0000001 # adding a small value avoids the occurance of -Inf
return from log transformation of a zero residual
        rownames(rt.scores.plot) <-candidate.community
        colnames(rt.scores.plot) <-template.community

        pheat.col <- colorRampPalette(c("white","yellow", "orange", "purple"))(100)
        log.seq<-rev(1 * 1.1^(0:100))
        pheat.breaks<-(max(rt.scores.plot)/log.seq)

        rt.scores.plot[is.na(match.matrix)]<-NA

        pheatmap((rt.scores.plot),
          color=pheat.col,
          breaks=pheat.breaks,
          cluster_rows=F,
```

```
          cluster_cols=F,
          scale="none",
          show_rownames=T,
          show_colnames=T,
          main="Residual RT (post-alignment, inter-match masked)",
          display_numbers=T,
          fontsize=16,
          number_format="%.2f")


    # corrected MZ scores
      mz.scores.plot <- mz.scores+0.000005 # adding a small value avoids the occurance of -Inf
return from log transformation of a zero residual
        rownames(mz.scores.plot) <-candidate.community
        colnames(mz.scores.plot) <-template.community

        pheat.col <- colorRampPalette(c("white","yellow", "orange", "purple"))(100)
        log.seq<-rev(1 * 1.1^(0:100))
        pheat.breaks<-(max(mz.scores.plot)/log.seq)

        pheatmap((mz.scores.plot),
          color=pheat.col,
          breaks=pheat.breaks,
          cluster_rows=F,
          cluster_cols=F,
          scale="none",
          show_rownames=T,
          show_colnames=T,
          main="Residual m/z",
          display_numbers=T,
          fontsize=16,
          number_format="%.4f")

    # corrected IT scores
      it.scores.plot <- it.scores+0.0000001
        rownames(it.scores.plot) <-candidate.community
        colnames(it.scores.plot) <-template.community

        pheat.col <- colorRampPalette(c("white","yellow", "orange", "purple"))(100)
        log.seq<-rev(1 * 1.1^(0:100))
        pheat.breaks<-(max(it.scores.plot)/log.seq)

        pheatmap((it.scores.plot),
          color=pheat.col,
          breaks=pheat.breaks,
          cluster_rows=F,
          cluster_cols=F,
          scale="none",
          show_rownames=T,
          show_colnames=T,
          main="Residual intensity",
          fontsize=16,
          display_numbers=T)

} else NULL

  ### Combining RT/mz/IT differences for further resolution of ambiguity unresolvable by RT alone

    # # combination by multiplication - advised by Olivier
    #
    #   combined.scores.1 <- (rt.scores+0.0001)*(mz.scores+0.0001)
    #   combined.scores.2 <- (rt.scores+0.0001)*(mz.scores+0.0001)*(it.scores+0.0001)

      # scaling of rt and mz matching scores: makes their maximumm contribution equal so they can
be combined with equal weight.
        # Note - if all scores are 0, don't scale (dividing by zero can introduce errors)
        if (sum(rt.scores)>0) (rt.scores.scaled<-rt.scores/sum(rt.scores)) else (rt.scores.scaled
<- rt.scores)
        if (sum(mz.scores)>0) (mz.scores.scaled<-mz.scores/sum(mz.scores)) else (mz.scores.scaled
<- mz.scores)
```

```
        if (sum(it.scores)>0) (it.scores.scaled<-it.scores/sum(it.scores)) else (it.scores.scaled
<- it.scores)

      # generates combined scores options, in the event that rt.scores (and combined.scores.1)
doesn't resolve ambiguity
        combined.scores.1 <- rt.scores.scaled+mz.scores.scaled
        combined.scores.2 <- combined.scores.1+it.scores.scaled

      if (detailed.plotting==T){
      ### PLOTTING (requires "pheatmap" package)
      ### combined scaled scores

      ### combined1
      combined.scores1.plot <- combined.scores.1+0.00001
        rownames(combined.scores1.plot) <-candidate.community
        colnames(combined.scores1.plot) <-template.community

        pheat.col <- colorRampPalette(c("white","yellow", "orange", "purple"))(100)
        log.seq<-rev(1 * 1.1^(0:100))
        pheat.breaks<-(max(combined.scores1.plot)/log.seq)

        pheatmap((combined.scores1.plot),
          color=pheat.col,
          breaks=pheat.breaks,
          cluster_rows=F,
          cluster_cols=F,
          scale="none",
          show_rownames=T,
          show_colnames=T,
          main="Residual RT (post-alignment)",
          display_numbers=T,
          fontsize=16,
          number_format="%.4f")

      ### combined2
      combined.scores2.plot <- combined.scores.2
        rownames(combined.scores2.plot) <-candidate.community
        colnames(combined.scores2.plot) <-template.community

        pheat.col <- colorRampPalette(c("white","yellow", "orange", "purple"))(100)
        log.seq<-rev(1 * 1.1^(0:100))
        pheat.breaks<-(max(combined.scores2.plot)/log.seq)

        pheatmap((combined.scores2.plot),
          color=pheat.col,
          breaks=pheat.breaks,
          cluster_rows=F,
          cluster_cols=F,
          scale="none",
          show_rownames=T,
          show_colnames=T,
          main="Residual RT (post-alignment)",
          display_numbers=T,
          fontsize=16,
          number_format="%.4f")

      } else NULL

  ### MASKING SCORES by what was originally a match in inter-sample matching
    # differences between features not originally matched are removed across all matrices and
replaced with "NA".
    # this stops the forced linking of poor matches, just because they're the least bad.  This is a
decision making step, not match-determining.

      rt.scores.scaled <- rt.scores.scaled*match.matrix
      mz.scores.scaled <- mz.scores.scaled*match.matrix
      it.scores.scaled <- it.scores.scaled*match.matrix
      combined.scores.1 <- combined.scores.1*match.matrix
      combined.scores.2 <- combined.scores.2*match.matrix

  ### FINAL RESOLUTION OF MATRIX MATCHES
```

Appendix 3

```
    # where the matrix is 1 row or column, choosing the best match is easy - it's just the lowest
number present
      # 1 candidate feature
      if (length(candidate.community)==1) {
        best.bin<-(which(rt.scores.scaled==min(na.omit(rt.scores.scaled[1,]))))                #
pulls minimum value as best bin for the candidate

        if (length(best.bin)>1) (best.bin<-
(which(combined.scores.1==min(na.omit(combined.scores.1[1,])))) ) else NULL
        if (length(best.bin)>1) (best.bin<-
(which(combined.scores.2==min(na.omit(combined.scores.2[1,])))) ) else NULL
        if (length(best.bin)>1) (print(paste("Error 01: sample", sample.count, "community", g,
sep=" "))) else NULL

        template.matches[template.community[best.bin]]<-candidate.community                    #
connects the candidate ID with the template feature in the reporting vector
        candidate.matches[candidate.community]<-template.community[best.bin]
      } else NULL

      # 1 template bin
      if (length(template.community)==1) {
        best.can<-(which(rt.scores.scaled==min(na.omit(rt.scores.scaled[,1]))))                #
pulls minimum value as best bin for the candidate

        if (length(best.can)>1) (best.can<-
(which(combined.scores.1==min(na.omit(combined.scores.1[,1])))) ) else NULL  # uses combined.scores
1 (rt/mz) if rt alone can't resolve
        if (length(best.can)>1) (best.can<-
(which(combined.scores.2==min(na.omit(combined.scores.2[,1])))) ) else NULL  # uses combined.scores
2 (rt/mz/it) if rt alone can't resolve
        if (length(best.can)>1) (print(paste("Error 02: sample", sample.count, "community", g,
sep=" "))) else NULL

        template.matches[template.community]<-candidate.community[best.can]                    #
connects the candidate ID with the template feature in the reporting vector
        candidate.matches[candidate.community[best.can]]<-template.community
      } else NULL

### where matrix is more than 2x2, a slightly more complex approach is needed:
  # micro-alignment maximises the effect of rt.scores

  # rely on the best.community.match obtained in the earlier micro-alignment.
  if (length(candidate.community)>1 & length(template.community)>1) {


    repeat {                                                                    # this bit picks
the lowest residual, calls that a definitive match, and removes its column and row from the matrix.
then repeats
      # test if matrix is empty (T = break cycle)
        if (sum(is.na(rt.scores.scaled))==(length(template.community)*length(candidate.community)))
break else NULL

      # find the lowest residual
        lowest<-which(rt.scores.scaled == min(rt.scores.scaled, na.rm = TRUE), arr.ind = TRUE)

        # if there are multiple lowest scores...
        if (length(lowest)>2){

           # first, see if the multiple lowest scores are in conflict.
           # if not.... pass all onward...
            if ((length(lowest[,"row"])==length(unique(lowest[,"row"]))) &
(length(lowest[,"col"])==length(unique(lowest[,"col"])))) {
                      # report it as a definitive match
                        template.matches[template.community[lowest[,"col"]]]<-
candidate.community[lowest[,"row"]]
                        candidate.matches[candidate.community[lowest[,"row"]]]<-
template.community[lowest[,"col"]]
                      # remove it and others from both the row and column
                        rt.scores.scaled[,lowest[,"col"]]<-NA
                        rt.scores.scaled[lowest[,"row"],]<-NA
                      # do the same to combined.scores.1/2 in the event that they are required
```

```
                        combined.scores.1[,lowest[,"col"]]<-NA
                        combined.scores.1[lowest[,"row"],]<-NA
                        combined.scores.2[,lowest[,"col"]]<-NA
                        combined.scores.2[lowest[,"row"],]<-NA
            } else NULL

          # ... and move on to the next
            if ((length(lowest[,"row"])==length(unique(lowest[,"row"]))) &
(length(lowest[,"col"])==length(unique(lowest[,"col"])))) next else NULL

          # if so (ELSE...), reserve those which require additional parameters to resolve, and
report the rest (if any).

            lowest.reserve <- lowest[which(duplicated(lowest[,"row"], fromLast=T) |
duplicated(lowest[,"row"], fromLast=F) | duplicated(lowest[,"col"], fromLast=T) |
duplicated(lowest[,"col"], fromLast=F)),]  # reserves all duplicated values
            # lowest.reserve will have multiple entries by nature, which are naturally stored as a
matrix. - no need to fix formatting
            lowest <- lowest[which(!duplicated(lowest[,"row"], fromLast=T) &
!duplicated(lowest[,"row"], fromLast=F) & !duplicated(lowest[,"col"], fromLast=T) &
!duplicated(lowest[,"col"], fromLast=F)),]      # only non-duplicated values
            if (length(lowest)==2) (lowest <- t(lowest)) else NULL              # if there is only
one lowest entry, it's returned as an interger that needs to be made into a matrix for the
following step

                      # report it as a definitive match
                        template.matches[template.community[lowest[,"col"]]]<-
candidate.community[lowest[,"row"]]
                        candidate.matches[candidate.community[lowest[,"row"]]]<-
template.community[lowest[,"col"]]
                      # remove it and others from both the row and column
                        rt.scores.scaled[,lowest[,"col"]]<-NA
                        rt.scores.scaled[lowest[,"row"],]<-NA
                      # do the same to combined.scores.1/2 in the event that they are required
                        combined.scores.1[,lowest[,"col"]]<-NA
                        combined.scores.1[lowest[,"row"],]<-NA
                        combined.scores.2[,lowest[,"col"]]<-NA
                        combined.scores.2[lowest[,"row"],]<-NA

          # then go on to resolution.
             lowest.again <- NULL
             for (i in 1:nrow(lowest.reserve)) {
                 lowest.again<-append(lowest.again, combined.scores.1[lowest.reserve[i,1],
lowest.reserve[i,2]])
             }

             lowest.sorted <- sort(lowest.again, decreasing=FALSE, na.last=NA)

             if (lowest.sorted[1]==lowest.sorted[2]) {

               lowest.again <- NULL
               for (i in 1:nrow(lowest.reserve)) {
                   lowest.again<-append(lowest.again, combined.scores.2[lowest.reserve[i,1],
lowest.reserve[i,2]])
               }
             }

             # this should really be done per conflicting subset, as lowest.reserve may actually
be 2 sets of independently conflicting vals.  However, this will get us there, just less
efficiently.
               lowest <-
t(as.matrix(lowest.reserve[(which(lowest.again==min(lowest.again))),1:2]))     # the good lowest is
the row that gives the lowest val.

             # again, see if the multiple lowest scores are in conflict.
             # if not.... pass all onward...
               if ((length(lowest[,"row"])==length(unique(lowest[,"row"]))) &
(length(lowest[,"col"])==length(unique(lowest[,"col"])))) {
                      # report it as a definitive match
                        template.matches[template.community[lowest[,"col"]]]<-
candidate.community[lowest[,"row"]]
```

```
                        candidate.matches[candidate.community[lowest[,"row"]]]<-
template.community[lowest[,"col"]]
                      # remove it and others from both the row and column
                        rt.scores.scaled[,lowest[,"col"]]<-NA
                        rt.scores.scaled[lowest[,"row"],]<-NA
                      # do the same to combined.scores.1/2 in the event that they are required
                        combined.scores.1[,lowest[,"col"]]<-NA
                        combined.scores.1[lowest[,"row"],]<-NA
                        combined.scores.2[,lowest[,"col"]]<-NA
                        combined.scores.2[lowest[,"row"],]<-NA
                } else print("Error!!!!")

        }


        if (length(lowest)==2){

                      # report it as a definitive match
                        template.matches[template.community[lowest[,"col"]]]<-
candidate.community[lowest[,"row"]]
                        candidate.matches[candidate.community[lowest[,"row"]]]<-
template.community[lowest[,"col"]]
                      # remove it and others from both the row and column
                        rt.scores.scaled[,lowest[,"col"]]<-NA
                        rt.scores.scaled[lowest[,"row"],]<-NA
                      # do the same to combined.scores.1/2 in the event that they are required
                        combined.scores.1[,lowest[,"col"]]<-NA
                        combined.scores.1[lowest[,"row"],]<-NA
                        combined.scores.2[,lowest[,"col"]]<-NA
                        combined.scores.2[lowest[,"row"],]<-NA
        }
    }
  }
} # ALIGNMENT LOOP ENDS HERE

# plots the intensity correlation between matched template and candidate features
plot(log(template[(which(is.na(template.matches)==F)),"into"]),log(candidate[(na.omit(template.matc
hes)),"into"]))
cor(log(template[(which(is.na(template.matches)==F)),"into"]),log(candidate[(na.omit(template.match
es)),"into"]))

# plots the rt vs. mz of matched features (from the candidate data set's perspective)
plot(candidate[na.omit(template.matches),"rt"], candidate[na.omit(template.matches),"mz"])

##############################################################################################
##############################################################################################
#########----report failed attempts at community matching as no matches----#############

  match.cleanup <- function(match.list, tally){
    # list of unmatched template bins
    unmatched<-which(is.na(match.list))  # ID= template bins without matches

    # removes failed matches
    failed<-NULL
    for (i in 1:length(unmatched)){
      # if unmatched bin has attempted matches in community matching and remains unmatched, it is a
failure and needs to be marked as such.
      if (sum(unmatched[i]==tally)>0) (failed<-append(failed,unmatched[i])) else NULL
    }

    match.list[failed]<-0

    return(match.list)
  }

# insert 0's for failed community match participants
  template.matches<-match.cleanup(template.matches, match.tally.t)
  candidate.matches<-match.cleanup(candidate.matches, match.tally.c)

##############################################################################################
```

Appendix 3

```
#############################################################################################
#######----remove orphans----###########

  # outside of communities, independent features with no matches are useless.  either empty bins,
or require appending to the master.ID.
  orphans.can<-which(matchcount==0)     # candidate ID's that do not match any template bins
  orphans.bin<-which(bincount==0)       # template ID's that do not match any candidate bins

  # replaces NAs with 0's indicating matching has occured and has failed.
  # why do some of these have 0's already???...
  template.matches[orphans.bin]<-0
  # ..and none of these do??
  candidate.matches[orphans.can]<-0

#############################################################################################
#############################################################################################
#######----non-community matching----###########

# 1.  remove mutually unique matches to reduce the size of residual matrix matching.
  # leave templates with 2 matches and 1-match templates with multiple return matches for matrix
matching.

# goal is to see if two things point at each other (and only each other).  doesn't matter which
side starts.
unmatched.bins<-which(is.na(template.matches))  # ID= template bins without matches
#unmatched.candidates<-which(is.na(candidate.matches))  # ID= template bins without matches

# do sequentially for every unmatched bin...
for (i in 1:length(unmatched.bins)){

# template matches one candidate only
if (bincount[unmatched.bins[i]]==1) {
   return.matches<-matchcount[bins[[unmatched.bins[i]]]]     # gets the number of return (template)
matches for the single candidate ID matched by the template
   # single return match
   if (return.matches==1) {
      # for fun, check that the ID's are the same.  They should be always... just checking for dev.
      if (bins.matched[[bins[[unmatched.bins[i]]]]]==unmatched.bins[i]) {
         template.matches[unmatched.bins[i]]<-bins[[unmatched.bins[i]]]  # stores candidate ID as
unique match to template ID
         candidate.matches[bins[[unmatched.bins[i]]]]<-unmatched.bins[i] # stores template ID as
unique match to candidate ID
      } else print("WARNING: matches are exclusivly mutual, but ID's do not match")

   } else print("returns multiple matches - pass to matrix matching")
   } else print("template matches more than 1 independent candidate - pass to matrix matching")
}

#############################################################################################
#############################################################################################
############################## compile ledger and write out data ############################

### recording matched data
# replaces 0's with NAs, now that functionality of NA's and 0's from matching is no longer needed.
  template.matches[which(template.matches==0)]<-NA



# orders the template matches in the same order as the master.ID matrix (in "experimental order")
ordered.template.matches<-NULL
  for (o in 1:(length(as.matrix(master.ID)[,sample.count]))){
    # if the master.ID row has no template entry, then the match must be NA.  Otherwise, the match
is a value as determined by the template.matches
    if (is.na(as.matrix(master.ID)[o,sample.count])) ordered.template.matches[o]<-NA else
ordered.template.matches[o]<-template.matches[(as.matrix(master.ID)[o,sample.count])]
  }

# adds candidate matches to new column in master.ID reporting matrix
  master.ID<-cbind(master.ID, ordered.template.matches)
```

Appendix 3

```
### appending orphans to master.ID as new rows
# duplicates all successfully matched featuresfeatures that have been matched
  candidates.plus.matches<-na.omit(c((1:(nrow(candidate))),template.matches))  #all candidate
features + those that have been matched

# finds only features that have a single instance (are not duplicated, and tehrefore are have not
been successfully matched)
  candidate.orphans<-candidates.plus.matches[(!duplicated(candidates.plus.matches, fromLast=T) &
!duplicated(candidates.plus.matches, fromLast=F))] # removing all instances of duplicates
(including the first of each)

# assures that all candidate features are in order (not critical).
  candidate.orphans <- sort(candidate.orphans, decreasing = FALSE)

# append candidate.orphans, in order, to the bottom of the master.ID matrix as new rows.
  new.addition <- matrix(data = NA, nrow = (length(candidate.orphans)), ncol = (ncol(master.ID)-1),
byrow = FALSE, dimnames = NULL)

# grafts on to master.ID
  master.ID <- rbind(master.ID, (cbind(new.addition, candidate.orphans)))


### use master.ID to update master.mz, rt, and it

master.MZ <- cbind((rbind(as.matrix(master.MZ),new.addition)),
(candidate[(master.ID[,(sample.count+1)]),"mz"]))
master.RT <- cbind((rbind(as.matrix(master.RT),new.addition)),
(candidate[(master.ID[,(sample.count+1)]),"rt"]))
master.IT <- cbind((rbind(as.matrix(master.IT),new.addition)),
(candidate[(master.ID[,(sample.count+1)]),"into"]))


### looping controls and procedures

# stop matching analysis when all samples in the run order have been matched
  if (sample.count==nrow(run.order)) break else NULL

# prints the name of the file that was just matched and added to the ledger
  print (paste((as.character(run.order[(sample.count+1),1])), "matched in, reported to ledger", sep
= " "))

# updates the sample count for the next round of matching
  sample.count <- sample.count+1

# sets the candidate from round n to be the template from round n+1
  template<-candidate

} # end of program repeat



# NOTE:  this part of the code will allow the script to monitor the creation of new files and run
in real-time.  However, it was not used in the Thesis.
# looks for next file in specified order.  If present, it loads and repeats the matching.  If not,
it waits.
file.ready=0
repeat{
print("repeat!")
if (file.exists((as.character(run.order[(sample.count+1),1])))==F) (file.ready=0) else
(file.ready=1)   # sets switch to either sleep or repeat
if (file.ready==1) (print("next file found") & break) else (Sys.sleep(10))
}
```

Appendix 4

Database of empirically observed spectral mass values for 55 standard compounds in negative ionization mode

ID	formula/name (M)	parent	frag	adducts	adductfrags
1	C4H6O3	101.0239		225.0377	192.9933
1	a-ketobutyric acid			264.9778	
3	C19H19N7O6	440.1397	132.0452	462.1131	
3	Folic acid		175.051		
3			311.089		
3			378.1313		
3			396.1426		
3			147.0284		
3			119.0353		
5	C10H17N3O6S	306.076	87.057	482.0998	482.1096
5	Glutathione reduced		99.05593	328.0579	338.0491
5			99.0569	635.1392	310.0461
5			128.0353	613.1608	304.0611
5			135.0556	611.1435	294.0707
5			141.0624	328.0584	282.0727
5			143.0452		264.0565
5			146.0477		253.0263
5			160.0074		250.0805
5			166.0929		242.0784
5			177.0381		238.0788
5			179.0471		221.0554
5			197.0542		219.0384
5			210.0871		207.0376
5			254.0779		199.0162
5			272.0878		184.1058
5			288.0656		181.0595
5					175.0174
5					167.0443
5					150.017
6	C8H9NO2	150.0555	107.0374	230.0124	
6	Paracetamol sulfate				
7	C4H8O3	103.0395		275.1138	59.0137
7	3-hydroxybutyric acid			229.0698	103.0402
7				573.2156	297.0963
7					315.104

7					401.1432
7					487.1779
9	C5H6O5	145.0137	101.0245	334.9995	166.9961
9	a-ketoglutaric acid			313.0164	
10	C12H14N2O2	217.0977	131.0375		
10	N-acetyl-5-hydroxy-tryptamine		144.0457		
10			157.0534		
10			158.0612		
12	C8H15NO6	220.0821			
12	N-Acetyl-D-galactosamine				
13	C5H9NO3S	162.0225	84.0456	347.0354	217.972
13	N-acetyl-L-cysteine			323.0376	193.9957
13				184.0044	162.023
13					128.0353
13					116.0178
13					74.00675
14	C7H11NO5	188.0559	100.0773	399.1016	
14	N-acetyl-DL-glutamic acid		102.056	210.0378	
14			126.0557		
14			128.0354		
14			144.0667		
14			170.0461		
15	C5H12N2O2S	163.0541	76.02253	545.0976	381.0356
15	S-(2-aminoethyl)-L-cysteine hydrochloride			349.098	364.0092
15					294.0026
15					250.9592
15					216.9753
15					206.9691
15					199.9469
16	C5H11N3O2	144.0773	102.0554	289.1631	
16	4-guanidinobutyric acid			212.0653	
16				311.1451	
17	C9H10O5	197.045	137.0245	417.0799	
17	DL-4-hydroxy-3-methoxymandelic acid			265.0331	
18	C6H10O3	129.0552		281.1005	169.0489
18	DL-a-keto-ß-methyl-n-valeric acid			443.1	151.0379
18				433.1454	313.0397
19	C10H18N2O5	245.1137	116.0712		
19	?-L-glutamyl-L-valine		128.0354		
19			165.1037		
19			183.1141		

19			209.0933		
19			227.1033		
20	C11H13NO2	190.0868	144.0456		
20	5-methoxytryptophol		145.051		
20			175.0643		
22	C9H9O3	165.0552	103.0544	353.1015	59.0138
22	(S)-3-hydroxy-3-phenylpropionic acid				83.0119
22					119.0514
22					143.0459
22					165.0553
22					187.0391
22					205.0467
22					229.0485
22					247.06
25	C4H7NO3	116.0348	74.0245	255.0585	
25	N-acetyl-glycine				
26	C9H13NO3	182.0817	122.0376		
26	(+/-)-Epinephrine		148.0406		
26			149.0484		
26			164.0718		
28	C6H10O3	129.0552		201.1126	99.0811
28	4-Methyl-2-oxovaleric acid			229.1075	123.1188
28				281.1006	139.1113
28				425.214	141.1291
28				353.1568	157.1231
28				505.2042	185.1178
28					151.0371
28					169.0476
28					223.0964
28					295.152
28					393.1479
29	C3H4O4	103.0031			
29	ß-Hydroxypyruvic acid				
30	C4H8O3	103.0395		229.0689	
30	(S)-2-hydroxybutyric acid				
31	C9H14N4O3	225.0988	81.045	473.1863	
31	L-carnosine		93.0456	451.2039	
31			110.0722		
31			137.0354		
31			154.062		
31			163.0986		

31			181.1095		
32	C9H14N3O8P	322.044	78.9589	645.0968	181.0627
32	Cytidine 5'-monophosphate		96.9697	667.0787	402.0117
32			110.0363	968.1483	423.9945
32			138.9786		312.9507
32			150.9799		
32			192.9881		
32			211.0029		
32			279.0371		
33	C6H10O3	129.0552			
33	(+/-)-3-methyl-2-oxovaleric acid				
34	C4H10O2S2	153.0044			
34	Dithiothreitol				
35	C7H12N2O4	187.0719	125.0722	397.1327	143.0818
35	N-acetyl-L-glutamine		127.0504		109.0398
35			145.062		
35			169.0619		
37	C15H18N2O4S	321.0909	234.059	643.1888	
37	Dansylsarcosine piperidinium salt		170.0975	665.1707	
39	C9H8O3	163.0395	91.0554	349.0699	185.0222
39	phenyl-pyruvate				
40	C3H4O3	87.0082		197.0067	
40	Pyruvic acid				
41	C4H8O3	103.0395		229.0683	
41	?-Hydroxybutyric acid			171.0271	
42	C6H10O3	129.0552		281.1	151.0384
42	a-Ketoisocaproic acid			443.099	169.0485
42				433.1454	209.1164
42					184.988
42					241.0532
42					281.1019
42					313.0356
43	C6H9NO5	174.0402	88.0405	371.063	70.028
43	N-acetyl-DL-aspartic acid		114.0203		112.0378
43			130.0511		156.0271
43			112.0429		196.0137
45	C5H11NO2S	148.0432		319.076	170.0255
45	L(+)-penicillamine			295.0775	180.0143
45					114.0564
47	C6H12O2	115.0759		183.0637	
47	n-Caproic acid			253.1415	

50	C5H10O5	149.045			
50	D(-)Ribose				
51	C19H17N3O4S2	414.0582			
51	Cephaloridine				
52	C6H12O3	131.0708	85.0658	285.131	
52	L-a-hydroxyisocaproic acid				
54	C5H8O3	115.0395		253.0693	155.0337
54	a-Ketoisovaleric acid			401.054	143.9887
54					156.9612
54					170.9733
54					213.0239
54					214.0299
54					215.0357
54					285.0063
54					329.0708
54					101.9393
55	L-homocarnosine	239.1144	80.0386	501.2193	102.0568
55	C10H16N4O3		81.0458	479.2381	
55			84.0458		
55			93.0455		
55			101.0709		
55			108.056		
55			110.0722		
55			136.0513		
55			137.0357		
55			141.0669		
55			150.1005		
55			154.0625		
55			177.1149		
55			193.1131		
55			195.125		
55			221.1006		
57	DL-a-Hydroxybutyric acid	103.0395	57.0348	229.0687	145.0866
57	C4H8O3			189.077	
58	Homogentisic acid	167.0344	122.0367	357.0572	189.0164
58	C8H8O4		108.0211		123.0447
58			93.035		
58			121.0288		
59	a-Keto-?-methiolbutyric acid sodium salt	147.0116		317.0137	99.0084
59	C5H8O3S				
60	Sulfamethazine	277.0759	236.0486	309.0649	121.0404

60	C12H14N4O2S		213.1143	345.064	80.9662
60			196.0172	887.1623	79.9552
60			195.0237		
60			155.0057		
60			132.0053		
60			122.0719		
60			106.0404		
61	C6H8O7	191.0192	85.029	441.0487	423.0334
61	D/a-saccharic acid 1,4-lactone		147.0289	405.0272	231.0118
61			129.0194	383.045	141.0166
61				209.0296	147.0298
61					129.0186
61					111.009
61					133.0136
61					89.0232
61					71.01305
61					173.0101
62	C11H11NO3	204.0661	186.0555	431.1206	226.0491
62	DL-indole-3-lactic acid		158.0607	409.1389	
62			142.0656		
62			130.0658		
62			128.05		
62			116.0498		
63	C8H15NO5	204.0872	72.0087	431.1642	283.017
63	Boc-Ser-OH		100.004		357.0916
63			130.0141		
64	C5H6O4	129.0188		281.0273	85.0278
64	Mesaconic acid				
65	C6H8O7	191.0192	147.03	423.0377	387.0206
65	D-Saccharic acid 3,6-Lactone		85.0291	405.0276	361.0359
65			57.0338	209.0299	343.0264
65					317.0434
65					231.0103
65					141.0181
65					133.0131
65					129.0209
65					115.0028
65					111.0073
65					89.0234
65					71.0131
69	C5H6N2O2	125.0351	81.0452	147.0176	

69	Imidazole-4-acetic acid			273.0606	
73	C33H34N4O6	581.24	537.2485		
73	Biliverdin		285.1234		
73			239.1173		
73			213.1015		
74	C4H8O3	103.0395	59.0129	533.2224	171.0649
74	DL-?-hydroxybutyric acid			573.2144	189.0761
74				447.1869	85.0291
74				167.0265	229.0687
74					315.1046
74					297.0972
74					315.1054
74					401.1421
74					487.1768
75	C3H6O3	89.0239		201.0375	
75	L-lactate				
77	C4H8O3	103.0395		229.0684	
77	ß-hydroxybutyric acid				

Appendix 5

Deconvolution and *de novo* database generating script

```
### IMPERICAL - de novo database synthesizer
### Version 12dev - October 3, 2011
        # Stable parent matching, parent validation via targeted fragment matching, independent
fragment matching, and independent adduct matching
### Matthew R. Lewis
### Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College
London
### Much thanks to Florian Geier and Paul Benton for their assistance and contributions

### ------------------------------------------------------------------------------------------
----------------- ###

### Conditions ###
  # Data must be given in minutes (RT).  Can be changed easily.... #
### /Conditions ###

### data import and setup ###
  # load working directory and xcms package #
    myDir = "C:/Users/mrlewis1/Dropbox/EMPERIAL/chrom development"
    setwd(myDir)
    WD <- getwd()
  # clean up workspace - start from scratch to ensure no variable carryover #
    rm(list = ls(all = TRUE))
  # database style may change when adducts are incorporated #
  # imports data from .csv files - physical property database (tDB) and peakpicked mixed-standard
run (STD) #
    tDB<-read.csv("XferDB.csv", header=TRUE)
    STD<-read.csv("MasterMix.csv", header=TRUE)
### /data import and setup ###

### variables for consideration ###
  # use intensity to break ambiguity between potential parent matches? (1 = yes, 0 = no)#
    pTie = 1
  # multiplication factor to differentiate acceptable match from next most intense possible match #
    factor = 10
  # mass error (ppm) #
    xfragppm = 100
    xparentppm = 100
    xadductppm = 100
  # retention time error window (seconds, +/- 1/2 window) #
    xdeltaRT=2
  # arbitrary x variable necessary to make later x[[i]] stuff work #
    x<-numeric()
    y<-numeric()
    z<-numeric()

### /variables for consideration ###


### ------------------------------------------------------------------------------------------
----------------- ###


### STEP 1: Creation of initial visualization (base scatter plot) ###

par(mfrow=c(2,1))

  # sets overall plot X and Y axis limits from min/max values in the peaklist (STD) #
    xlim.min<-min(STD[,"rt"])
    xlim.max<-max(STD[,"rt"])
    ylim.min<-min(STD[,"mz"])
    ylim.max<-max(STD[,"mz"])

  # creates blank plot for assigned features #
```

```
    plot(-1,-1, col="1", xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max),xlab="retention
time", ylab="m/z", main="Assignments")

  # creates m/z vs rt plot of the peaklist #
    plot(STD[,"rt"],STD[,"mz"], col="1", xlim=c(xlim.min,xlim.max),
ylim=c(ylim.min,ylim.max),xlab="retention time", ylab="m/z", main="STD mix")

### /STEP 1 ###


### ------------------------------------------------------------------------------------------
----------------- ###


### STEP 2: Conversion of data from input form to accessible form ###

  # convert the theoretical database (tDB) to a matrix of unique standard compounds (uniqueID),
each with one ID, mass, and number of databased fragments.#
  # In the process, individual fragment lists are generated for each unique standard compound #
  # once later references are generalized to row names rather than numbers, i'd like to add more
information here, such as compound NAME.#

  # create matrix of unique standard compounds (across) by ID and mass (down) from the tDB #
  # rbind won't include text - therefore names can not be included here #
  # original uniqueID line would not extract DB entries from a table of only unique entries - it
seemed to depended on the occurance of duplication.  It has therefore been rewritten from scratch #

  # Binds id and parent from tDB, maintaining redundancy that may or may not exist in the tDB #
    tDB.ID<-rbind(tDB[,"id"],tDB[,"parent"])
    rownames(tDB.ID)<-c("id","tMZ")
  # entries that are NOT duplicates are stored in uniqueID.  Only the first instance of a redundant
"id" entry is kept #
    uniqueID<-tDB.ID[,(which(!duplicated(tDB.ID["id",])))]

  # saves the number of unique standard compounds present in the tDB as a variable for later use #
    nSTDs<-length(uniqueID["id",])

  # creates and fills a row of # fragments (tFrags) per ID.  Simultaneously creates individual
fragment lists for each standard compound,
    fragsperID<- matrix(nrow=1,ncol=nSTDs)
      rownames(fragsperID)<-"tFrags"
    # loops for all unique ids #
      for(i in 1:nSTDs){
           # generates a T/F vector for each standard compound referring to the tDB #
            temp<-tDB[(tDB$id==uniqueID["id",i]),]
           # saves each fragment list by its uniqueID column number, and removes NA entries from
fragment count and list #
             x[[i]]<-na.omit(temp["frag"])
           # fills in [# fragments per ID = tFrags] to table #
             fragsperID[1,i]<-length(x[[i]]$frag)
    }

  # creates and fills a row of # adducts (tAdducts) per ID.  Simultaneously creates individual
adduct lists for each standard compound,
    adductsperID<- matrix(nrow=1,ncol=nSTDs)
      rownames(adductsperID)<-"tAdducts"
    # loops for all unique ids #
      for(i in 1:nSTDs){
           # generates a T/F vector for each standard compound referring to the tDB #
            temp<-tDB[(tDB$id==uniqueID["id",i]),]
           # saves each adduct list by its uniqueID column number, and removes NA entries from
adduct count and list #
             y[[i]]<-na.omit(temp["adducts"])
           # fills in [# adducts per ID = tAdducts] to table #
             adductsperID[1,i]<-length(y[[i]]$adducts)
    }

  # creates and fills a row of # adduct fragments (tAdductFrags) per ID.  Simultaneously creates
individual adductfrags lists for each standard compound,
    adductfragsperID<- matrix(nrow=1,ncol=nSTDs)
      rownames(adductfragsperID)<-"tAdductfrags"
```

Appendix 5

```
    # loops for all unique ids #
      for(i in 1:nSTDs){
           # generates a T/F vector for each standard compound referring to the tDB #
            temp<-tDB[(tDB$id==uniqueID["id",i]),]
           # saves each adduct list by its uniqueID column number, and removes NA entries from
adduct count and list #
             z[[i]]<-na.omit(temp["adductfrags"])
           # fills in [# adducts per ID = tAdductfrags] to table #
             adductfragsperID[1,i]<-length(z[[i]]$adductfrags)
    }

  # adds fragments/ID and adducts/ID to the unique ID matrix #
    uniqueID<-rbind(uniqueID, fragsperID, adductsperID, adductfragsperID)

### /STEP 2 ###


### ----------------------------------------------------------------------------------------------
---------------- ###


### STEP 3: PARENTS ###

### STEP 3a:  match all STD feature masses (m/z's) to a list of unique parents, and return a matrix
of possible hits #
  # matchmaker and MatchedList functions rewritten from code originally provided by Florian Geier #

  # eliminates tDB standard entries with missing parent values from parent matching #
    tDB.rents<-uniqueID[,which(uniqueID["tMZ",]!="NA")]

  # matchmaker finds matching values in matchR for each value in matchE, with a specified window of
error
    matchmaker<-function(matchE,matchR,window,idList.colnames){

#####matchE<-STD[,"mz"]
#####matchR<-uniqueID["tMZ",]
#####window<-xparentppm
#####idList.colnames<-uniqueID["id",]

    # this section covers the matching of STD features to the parents detailed in tDB #

    # primary is a matrix derived from the mz column of the STD matrix.  Column 1 is minus error,
and column2 is plus error #
      primary<- matrix(nrow=length(matchE),ncol=2,dimnames=list(NULL,c("min","max")))
    # this loop calculates the error to be applied to each mz entry in STD and fills the empty
primary matrix explained above #
      for(i in 1:length(matchE)){
          # calculates ppm error #
      # there may be a way to generalize this so that it can be used for nonppm calcs - split error
out as a variable function?
            error<-((window/10^6)*matchE[i])
            less<-(matchE[i]-error)
        more<-(matchE[i]+error)
        primary[i,1]<-less
        primary[i,2]<-more
      }

    # this section matches and reports hits only (y/n = 1/0) #
    # creates the empty 1 column matrix that will be filled with parent match hits between the tDB
and STD mzs given the supplied error #
      MatchedList<- matrix(nrow=length(matchE),ncol=1,dimnames=list(NULL,c("hits")))

    # fills the above table, replacing NA with digits where matching values were found (1) or not
(0) #
      for(i in 1:length(matchE)){
        temp<-which(matchR >= primary[i,1] & matchR <= primary[i,2])
        MatchedList[i,1]<-length(temp)
      }

    # this section matches, as above, but reports sorted results #
```

```
    # creates a matrix where nSTDs = columns and length(STD) = rows.  Matches will be ID'd and
assigned to one of the database standards in this list # #
    # old version = idList<- matrix(nrow=length(matchE), ncol=nSTDs)
      idList<- matrix(nrow=length(matchE), ncol=length(matchR))
      colnames(idList)<-idList.colnames

    # rematches and sorts hits (displayed as TRUE or FALSE into proper tDB standard column #
      for(i in 1:length(matchE)){
    # this matching really runs the wrong way around - but it works.
      idList[i,]<-matrix(matchR >= primary[i,1] & matchR <= primary[i,2])
      }

    # combines output from both rounds of matching (hits and sorted values) #
    # this step makes Ts and Fs into 1s and 0s.  Intended for later addition #
      FullList<-cbind(MatchedList,idList)
      return(FullList)
    }

  # runs the matching function with respect to parents (STD --> uniqueID tMZ (from tDB)) #
    possibleParents<-matchmaker(STD[,"mz"], uniqueID["tMZ",], xparentppm, uniqueID["id",])


### ----PLOT-BREAK---- ###

  # fades all datapoints on STD peak list plot to grey
  # for(i in seq(153:230){
    for(i in seq(153, 230, by = 10)){
      par(new=T)
      pPlot<-plot(STD[,"rt"],STD[,"mz"], col=colours()[i], xlab="retention time", ylab="m/z",
main="STD mix")
    }

  # plotting - this could be done at once, but it looks more interesting if done one standard at a
time #
    for(i in 1:nSTDs){
#i=20
    # plots a horizontal line for "i" standard tMZ representing the value to which emperical
measurements from the peak list (STD) will be matched #
      par (new=T)
      abline(h=(uniqueID["tMZ",i]),col=1,lty=3, xlim=c(xlim.min,xlim.max),
ylim=c(ylim.min,ylim.max))
      par (new=T)
    # replots all points on each iteration, but colors black only those with a positive match to a
tDB parent mass (sequentially, for each of nSTDs)
      plot(STD[,"rt"],STD[,"mz"], col=possibleParents[,(1+i)], xlab="", ylab="", pch=16,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))
      Sys.sleep(0.1)
    }

  # to color ambiguous hits (where more than 1 feature in the peaklist matches any one tDB standard
parent mass) yellow:
    pBlotter.ambiguous<-STD[(which(possibleParents[,"hits"]>1)),]
    par (new=T)
    plot(pBlotter.ambiguous[,"rt"],pBlotter.ambiguous[,"mz"], col=7, xlab="", ylab="", pch=20,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))

### ----/PLOT-BREAK---- ###


### STEP 3b: finding and reporting unambiguous hits based on parent matching alone #

# tiebreaking function for when the number of matches reported (column sum) for any tDB standard is
greater than 1.  Must be turned on in settings to be implemented #
  tiebreaker<-function(i, temp, factor, matchedParents){
  # creates a list of all into (intensity) values #
    intensities<-temp[,"into"]
  # unlists so that sorting can occur #
    intensities<-unlist(intensities)
  # sorts largest to smallest #
    intensities<-sort(intensities, decreasing = TRUE , na.last = NA)
```

Appendix 5

```
  # fills intensity from the above #
    if (intensities[1]>=(intensities[2]*factor)) matchedParents["pIntensity",i]<-intensities[1]
else NULL
  #fills retention time value #
    if (intensities[1]>=(intensities[2]*factor)) matchedParents["pRetention",i]<-
(temp[which(temp[,"into"]==intensities[1]),])[,"rt"] else NULL
    return(matchedParents)
  }

  tiebreaker.plot<-function(i, temp, factor, pBlotter, solved){
  # creates a list of all into (intensity) values #
    intensities<-temp[,"into"]
  # unlists so that sorting can occur #
    intensities<-unlist(intensities)
  # sorts largest to smallest #
    intensities<-sort(intensities, decreasing = TRUE , na.last = NA)
  # pulls out best match STD id from above processing
    best<-temp[(which(temp[,"into"]==intensities[1])),"id"]
  # returns best STD feature from peaklist
    if (intensities[1]>=(intensities[2]*factor)) return(STD[best,]) else NULL
  }

# reduces parent hitlist matrix to only those entries with a single match across the row (single
match to a single tDB standard) #
# this eliminates STD matrix features that match multiple tDB standard parent masses.  This would
be the case with two (or more) isobaric species in the same standard mixture #
# the match is probably correct for ONE of the tDB standards, but can't be correct for all three,
therefore it is considered ambiguous #
  unambiguousRows<-possibleParents[which(possibleParents[,"hits"]==1),]

# removes hits column so the values can be input to the matchedParents matrix, created below #
  unambiguousRows<-unambiguousRows[,-1]

# creates a matrix resembling uniqueID for reporting matches of tDB standards to the STD matrix.
Where results are not ambiguous (no more than 1 match) RT and intensity of the matched STD parent
feature is reported #
  matchedParents<-matrix(nrow=3,ncol=length(uniqueID["id",]))
  colnames(matchedParents)<-uniqueID["id",]
  rownames(matchedParents)<-c("pMatches","pRetention","pIntensity")

# sums each column in the unambiguousRows matrix #
# sums of 1 have passed both tests (row, above, and column, here) for ambiguity.
# sums >1 are ambiguous, with multiple STD features are matching that tDB standard compound parent
mass #
# stores results in matchedParents as the pMatches row #
  matchedParents["pMatches",]<-colSums (unambiguousRows, na.rm = FALSE, dims = 1)

# binds STD matrix to possibleParents hitlist (moves "hits" column to end of list), and preforms
the same-as-above reduction to eliminate nonmatches and ambiguous matches #
# possibleParents2 only exists in this block, and should not be called outside of this block.  #
  possibleParents2<-possibleParents[,-1]
  hits<-possibleParents[,"hits"]
  parentsSTD<-cbind(possibleParents2,STD,hits)
# maintains full list w/o removing ambiguous matches and nonmatches - this is used in rematching
section
  full.parentsSTD<-parentsSTD[which(parentsSTD[,"hits"]!=0),]
# removes ambiguous and nonmatches
  parentsSTD<-parentsSTD[which(parentsSTD[,"hits"]==1),]
  parentsSTD


### ----PLOT-BREAK---- ###


# create empty matrix for plotting unambiguous matches of STD features to tDB standards ##
  pBlotter<-matrix(nrow=nSTDs,ncol=2)
  colnames(pBlotter)<-c("mz","rt")

  solved<-STD[0,0]

### ----/PLOT-BREAK---- ###
```

Appendix 5

```
  # fills matchedParents with data from unambiguous matches
  # for each uniqueID, where there is a single unambiguous match to the STD matrix, fill
retention/intensity data from that matched feature #
  # i was having trouble with this, because i'd like to specifically target the id number of each
column in parentsStd ((parentsSTD[,"i"]==1)).  However, "i" is not valid.  So i'm relying on
ordering....
    for(i in 1:nSTDs){
      temp<-parentsSTD[which(parentsSTD[,i]==1),]
    # skips i for which pMatches = NA (ie missing parent values in uniqueID #
      if (is.na(matchedParents["pMatches",i]==1)==TRUE) next
    # if only 1 single match exists for a given tDB standard parent, report that retention time
(from STD peak list) as the RT of the tDB parent #
      if (matchedParents["pMatches",i]==1) matchedParents["pRetention",i]<-temp[1,"rt"] else
matchedParents["pRetention",i]<-NA
    # if only 1 single match exists for a given tDB standard parent, report that intensity (from
STD peak list) as the RT of the tDB parent #
    # see if this replacement can be combined with the above, possibly with the & operator #
      if (matchedParents["pMatches",i]==1) matchedParents["pIntensity",i]<-temp[1,"into"] else
matchedParents["pIntensity",i]<-NA

### ----PLOT-BREAK---- ###

    # if only 1 single match exists for a given tDB standard parent, report that mass and RT (from
STD peak list) as uniquely ID'd in the blotter vector (for plotting) #
      if (matchedParents["pMatches",i]==1) (pBlotter[i,"mz"]<-temp[1,"mz"])&(pBlotter[i,"rt"]<-
temp[1,"rt"])  else NULL

### ----/PLOT-BREAK---- ###

    # insert holding matrix for ambiguous hits here?????  As above pBlotter?????
    # if the number of matches reported (column sum) for any tDB standard is greater than 1, and if
the tiebreak function has been implemented by the user, run the tiebreak function #
      if (matchedParents["pMatches",i]>1 & pTie==1) matchedParents<-tiebreaker(i, temp, factor,
matchedParents) else NULL

### ----PLOT-BREAK---- ###

    # if the number of matches reported (column sum) for any tDB standard is greater than 1, and if
the tiebreak function has been implemented by the user, run the tiebreak function #
      if (matchedParents["pMatches",i]>1 & pTie==1) solved<-rbind(solved,(tiebreaker.plot(i, temp,
factor, pBlotter, solved))) else NULL

### ----/PLOT-BREAK---- ###

       }

  # binds parent matches and related retention and intensity information to the uniqueID matrix #
    report<-rbind(uniqueID,matchedParents)
  # in pMatches row, 1 means an unambiguous match has been made from the STD dataset.  >1 means
isobaric species (within the given ppm range) have been detected in the STDS matrix.  0 means no
match was found for the mass #
    report


### /STEP 3 ###


### ----------------------------------------------------------------------------------------------
----------------- ###


### STEP 4: PLOTTING OUT AND RESET ###

  ## update plot with uniquely ID'd parent masses - puts a smaller dot over ID'ing as unique the
existing one ID'ing a match.
  par (new=T)
  plot(pBlotter[,"rt"],pBlotter[,"mz"], col=3, xlab="", ylab="", pch=20, xlim=c(xlim.min,xlim.max),
ylim=c(ylim.min,ylim.max))
  ## update plot with uniquely ID'd parent masses that were resolved in the tiebreaker function #
  if (pTie==1) par (new=T) else NULL
```

Appendix 5

```
  if (pTie==1) (plot(solved[,"rt"],solved[,"mz"], col=3, xlab="", ylab="", pch=20,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))) else NULL

    par(mfrow=c(2,1))

  ## update ASSIGNMENT plot with uniquely ID'd parent masses - puts a smaller dot over ID'ing as
unique the existing one ID'ing a match.
  plot(pBlotter[,"rt"],pBlotter[,"mz"], col=3, pch=20, xlim=c(xlim.min,xlim.max),
ylim=c(ylim.min,ylim.max), xlab="retention time", ylab="m/z", main="Assignments")
  ## update plot with uniquely ID'd parent masses that were resolved in the tiebreaker function #
  if (pTie==1) par (new=T) else NULL
  if (pTie==1) (plot(solved[,"rt"],solved[,"mz"], col=3, xlab="", ylab="", pch=20,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))) else NULL


  # creates m/z vs rt plot of the peaklist #
    plot(STD[,"rt"],STD[,"mz"], col=colours()[230], xlim=c(xlim.min,xlim.max),
ylim=c(ylim.min,ylim.max),xlab="retention time", ylab="m/z", main="STD mix")

# old code
#  # fades non-ID'd features #
#  # pBlotter.unmatched<-STD[(which(possibleParents[,"hits"]==1)),]
#  # par (new=T)
#  # plot(pBlotter.ambiguous[,"rt"],pBlotter.ambiguous[,"mz"], col="red", xlab="", ylab="", pch=16,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))
#  # par (new=T)
#
#  # clean up plot
#
#    plot(STD[,"rt"],STD[,"mz"], col=colors()[230], xlim=c(xlim.min,xlim.max),
ylim=c(ylim.min,ylim.max),xlab="retention time", ylab="m/z", main="STD mix")
#    par (new=T)
#    plot(pBlotter[,"rt"],pBlotter[,"mz"], col=1, xlab="", ylab="", pch=16,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))
#    par (new=T)
#    plot(pBlotter[,"rt"],pBlotter[,"mz"], col=3, xlab="", ylab="", pch=20,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))
#    if (pTie==1) par (new=T) else NULL
#    if (pTie==1) (plot(solved[,"rt"],solved[,"mz"], col=1, xlab="", ylab="", pch=16,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))) else NULL
#    if (pTie==1) par (new=T) else NULL
#    if (pTie==1) (plot(solved[,"rt"],solved[,"mz"], col=3, xlab="", ylab="", pch=20,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))) else NULL


### /STEP 4 ###


### ------------------------------------------------------------------------------------------
----------------- ###


### STEP 5: SUBLIST MATCHING ###
### match all STD feature masses (m/z's) to a list of fragments, adducts, or adduct fragments and
return a matrix of possible hits #

# matchmaker2 finds matching values in matchR for each value in matchE, with a specified window of
error, and adds a nested loop for searching fragment groups
    matchmaker2<-function(matchE,matchR,window,idList.colnames, nEntries, w){

#####matchE<-STD[,"mz"]
#####matchR<-tDB[,"frag"]  #####matchR<-tDB[,"adducts"]
#####window<-xfragppm  #####window<-xadductppm
#####idList.colnames<-uniqueID["id",]
#####nEntries<-"tFrags" #####nEntries<-"tAdducts"
#####w<-x #####w<-y #####w<-z

  # primary is a matrix derived from the mz column of the STD matrix.  Column 1 is minus error, and
column2 is plus error #
    primary<- matrix(nrow=length(matchE),ncol=2,dimnames=list(NULL,c("min","max")))
```

```
  # this loop calculates the error to be applied to each mz entry in STD and fills the empty
primary matrix explained above #
    for(i in 1:length(matchE)){
          # calculates ppm error #
    # there may be a way to generalize this so that it can be used for nonppm calcs - split error
out as a variable function?
            error<-((window/10^6)*matchE[i])
            less<-(matchE[i]-error)
      more<-(matchE[i]+error)
      primary[i,1]<-less
      primary[i,2]<-more
  }

  # this section matches and reports hits only (y/n = 1/0) #
  # creates the empty 1 column matrix that will be filled with parent match hits between the tDB
and STD mzs given the supplied error #
  MatchedList<- matrix(nrow=length(matchE),ncol=1,dimnames=list(NULL,c("hits")))

  # fills the above table, replacing NA with digits where matching values were found (1) or not (0)
#
  for(i in 1:length(matchE)){
    temp<-which(matchR >= primary[i,1] & matchR <= primary[i,2])
    MatchedList[i,1]<-length(temp)
  }

  # this section matches, as above, but reports sorted results #
  # creates a matrix where nSTDs = columns and length(STD) = rows.  Matches will be ID'd and
assigned to one of the database standards in this list # #
  # old version = idList<- matrix(nrow=length(matchE), ncol=nSTDs)
    idList<- matrix(nrow=length(matchE), ncol=length(idList.colnames))
    colnames(idList)<-idList.colnames

  # for every feature in the STD matrix #
    for(i in 1:length(matchE)){

  # for every standard in the tDB (uniqueID)
    for(z in 1:nSTDs){
      # checks that the STD in question has any associated fragments.  If not, this inner loop
continues for the NEXT value of z #
        if (uniqueID[nEntries,z]==0) next else newVec<-(w[[z]])
      # stores frag list, which appears to be itself a vector, for tDB standard "z" in newVec #
      #newVec<-(w[[z]])
        temp<-which(newVec >= primary[i,1] & newVec <= primary[i,2])
              idList[i,z]<-length(temp)
    }
  }


  # combines output from both rounds of matching (hits and sorted values) #
  # this step makes Ts and Fs into 1s and 0s.  Intended for later addition #
    FullList<-cbind(MatchedList,idList)

  return(FullList)
}


### /STEP 5 ###


### ----------------------------------------------------------------------------------------------
----------------- ###


### STEP 6: PREFORM MATCHING FOR FRAGMENTS, ADDUCTS, AND ADDUCT FRAGMENTS ###


# runs the matching function with respect to fragments #
  possibleFragments<-matchmaker2(STD[,"mz"], tDB[,"frag"], xfragppm, uniqueID["id",], "tFrags", x)
# runs the matching function with respect to adducts #
  possibleAdducts<-matchmaker2(STD[,"mz"], tDB[,"adducts"], xadductppm, uniqueID["id",],
"tAdducts", y)
```

Appendix 5

```
# runs the matching function with respect to adduct fragments #
  possibleAdductfragments<-matchmaker2(STD[,"mz"], tDB[,"adductfrags"], xadductppm,
uniqueID["id",], "tAdductfrags", z)


### /STEP 6 ###


### ------------------------------------------------------------------------------------------
----------------- ###


### STEP 7: REDUCE MATCHED SUBLIST DATASETS TO INCLUDE MATCHES ONLY ###

# reduces each possibleXXXX matrix to only those entries with a hit to the STD matrix #
  #hitlist<-possibleFragments[which(possibleFragments[,"hits"]>=1),]
  hitlist.x<-possibleFragments[which(possibleFragments[,"hits"]>=1),]
  hitlist.y<-possibleAdducts[which(possibleAdducts[,"hits"]>=1),]
  hitlist.z<-possibleAdductfragments[which(possibleAdductfragments[,"hits"]>=1),]

# removes hits column so the values can be input to the matchedXXXX matrix, created below #
  #hitlist2<-hitlist[,-1]
  hitlist.x2<-hitlist.x[,-1]
  hitlist.y2<-hitlist.y[,-1]
  hitlist.z2<-hitlist.z[,-1]

# binds STD matrix to possibleXXXX hitlist (moves "hits" column to end of list), and eliminates
nonmatches #

  possibleFragments2<-possibleFragments[,-1]
  hits.x<-possibleFragments[,"hits"]
  fragmentsSTD<-cbind(possibleFragments2,STD,hits.x)

  possibleAdducts2<-possibleAdducts[,-1]
  hits.y<-possibleAdducts[,"hits"]
  adductsSTD<-cbind(possibleAdducts2,STD,hits.y)

  possibleAdductfragments2<-possibleAdductfragments[,-1]
  hits.z<-possibleAdductfragments[,"hits"]
  adductfragmentsSTD<-cbind(possibleAdductfragments2,STD,hits.z)

# removes nonmatched rows (features) - used later for rematching

  full.fragmentsSTD<-fragmentsSTD[which(fragmentsSTD[,"hits.x"]!=0),]
  full.adductsSTD<-adductsSTD[which(adductsSTD[,"hits.y"]!=0),]
  full.adductfragmentsSTD<-adductfragmentsSTD[which(adductfragmentsSTD[,"hits.z"]!=0),]



### /STEP 7 ###


### ------------------------------------------------------------------------------------------
----------------- ###


### STEP 8: REPORTING NUMBER OF MATCHES FROM SUBLIST MATCHING ###

# creates a matrix resembling uniqueID for reporting matches of tDB standards to the STD matrix.
Where results are not ambiguous (no more than 1 match) RT and intensity of the matched STD parent
feature is reported #

  matchedFragments<-matrix(nrow=3,ncol=length(uniqueID["id",]))
  colnames(matchedFragments)<-uniqueID["id",]
  rownames(matchedFragments)<-c("xMatches","xRetention","xIntensity")

  matchedAdducts<-matrix(nrow=3,ncol=length(uniqueID["id",]))
  colnames(matchedAdducts)<-uniqueID["id",]
  rownames(matchedAdducts)<-c("yMatches","yRetention","yIntensity")

  matchedAdductfragments<-matrix(nrow=3,ncol=length(uniqueID["id",]))
```

```
  colnames(matchedAdductfragments)<-uniqueID["id",]
  rownames(matchedAdductfragments)<-c("zMatches","zRetention","zIntensity")

# sums each column of the fragment hitlist2 matrix, stores results in matchedXXXX as the nMatches
row #

  matchedFragments["xMatches",]<-colSums (hitlist.x2, na.rm = FALSE, dims = 1)
  matchedAdducts["yMatches",]<-colSums (hitlist.y2, na.rm = FALSE, dims = 1)
  matchedAdductfragments["zMatches",]<-colSums (hitlist.z2, na.rm = FALSE, dims = 1)

### /STEP 8 ###


### ----------------------------------------------------------------------------------------------
----------------- ###


### STEP 9: GROUPING AND DETERMINATION OF BEST ASSIGNMENT FOR SUBLIST MATCHING ###

# this step ignores previous parent assignments and attempts to independently match m/z groups in
the peak list to fragment vectors (or v/v??) #
# completely possible that m/z's assigned as parents (or fragments, adducts, etc.) in above steps
are reassigned as fragments (adducts, etc.) here - truely independent look at the data #

#things called on:

grouper<-function(uniqueID, generalSTD, xdeltaRT, uniqueIDrow){

# creates the empty matrix for reporting best frag group matches #
  xMatches<-matrix(nrow=4,ncol=length(uniqueID["id",]))
  colnames(xMatches)<-uniqueID["id",]
  rownames(xMatches)<-cbind("xMatches", "xRetention", "xIntensity", "xMatched%")

# for each standard in the tDB... #
  for(i in 1:nSTDs){
#i=4
  # retrieves the number of databased fragments for the standard compound targeted #
    max.x<-uniqueID[uniqueIDrow,i]
  # checks that fragments exist for matching.  If so, reduces generalSTD table to relevant hits
only.  If not, goes to NEXT i (so NA's don't break the loop). #
    if (uniqueID[uniqueIDrow,i]==0) next else generalSTDi<-generalSTD[(which(generalSTD[,i]!=0)),]

  # if no matches are found, report "0" as number of matches and continue to next i
  # this line is difficult to test - should validate once loop is finished #
    if (length(generalSTDi[,i])==0) xMatches["xMatches",i]<-0 & next else NULL
  # if only 1 match is found, report it immediatly.  No need to do further sorting to establish
best match. #
    if (length(generalSTDi[,i])==1) (xMatches["xMatches",i]<-1)&(xMatches["xRetention",i]<-
generalSTDi[,"rt"])&(xMatches["xIntensity",i]<-generalSTDi[,"into"]) else NULL

  # loop only progresses to here if more than 1 match exists, requiring sorting #

### histogram approach ###
# creates a histogram distribution of number of matched fragments per bin of retention time #
# rt bin size reflects the original window of error specified up front.  the absolute bin size is
calculated by applying this window to the entire rt spread #
# this method is convenient, but has the disadvantage that by imposing hard limits (bins) -
splitting of rt groups may occur #
# foo<-hist(generalSTDi[,"rt"],((max(generalSTDi[,"rt"])-min(generalSTDi[,"rt"]))/(xdeltaRT/60)))
# max(foo$counts)
###

### moving match approach ###

  # this approach matches each rt in fragmentSTDi to itself and all others, effectivly moving the
window along each matched fragment.  This way the match that includes the most additional matches
within its rt window can be selected, and peak splitting is negated #

  # parameters for matching
    matchE<-generalSTDi[,"rt"]
    matchR<-generalSTDi[,"rt"]
```

```
    window<-xdeltaRT
    idList.colnames<-uniqueID["id",]

  # primary is a matrix of max and min rt values (from the generalSTDi matrix).  Column 1 is minus
error, and column2 is plus error #
    primary<- matrix(nrow=length(matchE),ncol=2,dimnames=list(NULL,c("min","max")))
  # this loop calculates the error to be applied to each rt entry in generalSTDi and fills the
empty primary matrix explained above #
    for(j in 1:length(matchE)){
    # this needs to be conditional if rt is to be in seconds OR in minutes.... #
      error<-((window/60)/2)
        less<-(matchE[j]-error)
      more<-(matchE[j]+error)
      primary[j,1]<-less
      primary[j,2]<-more
    }

  # this section matches and reports hits only (y/n = 1/0) #
  # creates the empty 1 column matrix that will be filled with match hits between the tDB and STD
rts given the supplied error window #
    MatchedList<- matrix(nrow=length(matchE),ncol=1,dimnames=list(NULL,c("rtMatches")))

  # fills the above table, replacing NA with digits where matching values were found (1) or not (0)
#
    for(b in 1:length(matchE)){
      temp<-which(matchR >= primary[b,1] & matchR <= primary[b,2])
      MatchedList[b,1]<-length(temp)
    }

    generalSTDi<-cbind(generalSTDi,MatchedList)

# test ambiguous match seeking code by changing rt of one match
# generalSTDi[4,"rt"]<-5.00

  # ambiguity must be resolved here
  # if all matches are found to be without a group, no one can be selected as the best without
additional criteria #
  # if each match is its own group (1, reported in rtMatches), the sum of rtMatches will == the
length of the column
    if (sum(generalSTDi[,"rtMatches"])==length(generalSTDi[,"rtMatches"])) next else NULL

# reports the number of matched features in the largest group(s)
# does not tell user if there are multiple groups with that number of matched features - that is
resolved below
  matches<-(max(generalSTDi[,"rtMatches"]))

# filling table of results #
  xMatches["xMatches",i]<-matches

# what if two groups are tied for highest number of matches?
# validation that all features marked with the max # of matches match in RT - OR, validate that
there are a max of X features in a group with max X matches
# if the median rt value of all matches in the highest-match# group matches all values
individually, +/- the given RT error, the group is homogenious.  Otherwise, it is ambiguous, and no
best assignment can be made #
# stores median rt of highest-match# group as "med"
  med<-median(generalSTDi[(which(generalSTDi[,"rtMatches"]==matches)),"rt"])
# checks that all matches belong to the same rt group
  rt.upper<-sum((med+(xdeltaRT/60))<generalSTDi[(which(generalSTDi[,"rtMatches"]==matches)),"rt"])
  rt.lower<-sum((med-(xdeltaRT/60))>generalSTDi[(which(generalSTDi[,"rtMatches"]==matches)),"rt"])
# if sum of rows for both tests = 0, all matches belong to same rt group - proceed with assignment.
If not, multiple "best" groups exist (ambiguity).  Do not make assignment.
  if ((rt.upper+rt.lower)!=0) next else NULL

# filling table of results with assignments #
# fills median retention time for all fragments in best group
  xMatches["xRetention",i]<-median(generalSTDi[(which(generalSTDi[,"rtMatches"]==matches)),"rt"])
# fills summed intensity for all fragments in best group
  xMatches["xIntensity",i]<-sum(generalSTDi[(which(generalSTDi[,"rtMatches"]==matches)),"into"])
  xMatches
  }
```

Appendix 5

```
# calculates % fragments matched
# this should be modified to calculate % only where assignments have been made.  Currently not the
case!
  xMatches["xMatched%",]<-((xMatches["xMatches",])/(uniqueID[uniqueIDrow,]))*100
# write NA over % matched where no assignment has been made
  for(i in 1:nSTDs){
    if (is.na(xMatches["xRetention",i])) (xMatches["xMatched%",i]<-NA) else NULL
  }

  return(xMatches)
}


 xMatches<-grouper(uniqueID, fragmentsSTD, xdeltaRT, "tFrags")
 yMatches<-grouper(uniqueID, adductsSTD, xdeltaRT, "tAdducts")
  # applies a bandaid to change row names - this could be delt with properly in the above function,
but would require cautious changes
    row.names(yMatches)<-c("yMatches","yRetention","yIntensity","yMatched%")
 zMatches<-grouper(uniqueID, adductfragmentsSTD, xdeltaRT, "tAdductfrags")
  # applies a bandaid to change row names - this could be delt with properly in the above function,
but would require cautious changes
    row.names(zMatches)<-c("zMatches","zRetention","zIntensity","zMatched%")


# binds new data independent fragment matching to report
  report<-rbind(uniqueID,matchedParents, xMatches, yMatches, zMatches)
  report

### /STEP 9 ###


### ----------------------------------------------------------------------------------------------
----------------- ###


### STEP 10: FUNCTION FOR DECIDING WHICH ASSIGNMENT IS BEST BY TALLYING TOTAL HITS

decisionsdecisions<-function(i,matchedParents,xMatches,yMatches,zMatches){

#i=15
  consensus.rt<-
rbind(matchedParents["pRetention",],xMatches["xRetention",],yMatches["yRetention",],zMatches["zRete
ntion",])
  # omit results where no assignments were made at all
  if (sum(consensus.rt[,i], na.rm=T)==0) next else NULL

  assignments<-na.omit(consensus.rt[,i])

# creates a reporting matrix for rematching tally
    thedecider<- matrix(nrow=9,ncol=length(assignments))
      rownames(thedecider)<-
c("assignment","pRematch","xRematch","yRematch","zRematch","pRematchInto","xRematchInto","yRematchI
nto","zRematchInto")
      thedecider["assignment",]<-assignments

# rematches parents, going back to the full parentsSTD table (w/o ambiguous matches removed)
  for (k in 1:length(assignments)){
  #k=2
    assignments[k]
    # returns all parent hits for tDB standard i
      pHits<-full.parentsSTD[which(full.parentsSTD[,i]==1),]
    # sums number of matches in total parent match list (with ambiguous matches retained)
      rematches<-sum((pHits[,"rt"]<(assignments[k]+(xdeltaRT/60)))&(pHits[,"rt"]>(assignments[k]-
(xdeltaRT/60))))
    # reports # of matches to the tested assignment in the parents row
      thedecider["pRematch",k]<-rematches
    # something to account for new intensity values must go here - i think this code works, but
have not validated it.
```

Appendix 5

```
      newInto<-
pHits[which(((pHits[,"rt"]<(assignments[k]+(xdeltaRT/60)))&(pHits[,"rt"]>(assignments[k]-
(xdeltaRT/60)))))),]
      newInto<-newInto[,"into"]
      thedecider["pRematchInto",k]<-sum(newInto)
  }
# rematches fragments, going back to the full fragmentsSTD table (w/o ambiguous matches removed)
  for (k in 1:length(assignments)){
  #k=2
    assignments[k]
    # returns all fragment hits for tDB standard i
      xHits<-full.fragmentsSTD[which(full.fragmentsSTD[,i]==1),]
    # sums number of matches in total parent match list (with ambiguous matches retained)
      rematches<-sum((xHits[,"rt"]<(assignments[k]+(xdeltaRT/60)))&(xHits[,"rt"]>(assignments[k]-
(xdeltaRT/60))))
    # reports # of matches to the tested assignment in the parents row
      thedecider["xRematch",k]<-rematches
    # something to account for new intensity values must go here - i think this code works, but
have not validated it.
      newInto<-
xHits[which(((xHits[,"rt"]<(assignments[k]+(xdeltaRT/60)))&(xHits[,"rt"]>(assignments[k]-
(xdeltaRT/60)))))),]
      newInto<-newInto[,"into"]
      thedecider["xRematchInto",k]<-sum(newInto)
  }
# rematches adducts, going back to the full adductsSTD table (w/o ambiguous matches removed)
  for (k in 1:length(assignments)){
  #k=2
    assignments[k]
    # returns all adduct hits for tDB standard i
      yHits<-full.adductsSTD[which(full.adductsSTD[,i]==1),]
    # sums number of matches in total parent match list (with ambiguous matches retained)
      rematches<-sum((yHits[,"rt"]<(assignments[k]+(xdeltaRT/60)))&(yHits[,"rt"]>(assignments[k]-
(xdeltaRT/60))))
    # reports # of matches to the tested assignment in the parents row
      thedecider["yRematch",k]<-rematches
    # something to account for new intensity values must go here - i think this code works, but
have not validated it.
      newInto<-
yHits[which(((yHits[,"rt"]<(assignments[k]+(xdeltaRT/60)))&(yHits[,"rt"]>(assignments[k]-
(xdeltaRT/60)))))),]
      newInto<-newInto[,"into"]
      thedecider["yRematchInto",k]<-sum(newInto)
  }

# rematches adductfrags, going back to the full adductsSTD table (w/o ambiguous matches removed)
  for (k in 1:length(assignments)){
  #k=2
    assignments[k]
    # returns all adduct hits for tDB standard i
      zHits<-full.adductfragmentsSTD[which(full.adductfragmentsSTD[,i]==1),]
    # sums number of matches in total parent match list (with ambiguous matches retained)
      rematches<-sum((zHits[,"rt"]<(assignments[k]+(xdeltaRT/60)))&(zHits[,"rt"]>(assignments[k]-
(xdeltaRT/60))))
    # reports # of matches to the tested assignment in the parents row
      thedecider["zRematch",k]<-rematches
    # something to account for new intensity values must go here - i think this code works, but
have not validated it.
      newInto<-
zHits[which(((zHits[,"rt"]<(assignments[k]+(xdeltaRT/60)))&(zHits[,"rt"]>(assignments[k]-
(xdeltaRT/60)))))),]
      newInto<-newInto[,"into"]
      thedecider["zRematchInto",k]<-sum(newInto)
  }

# sum of matches
thedecider<-rbind(thedecider, (thedecider[2,]+thedecider[3,]+thedecider[4,]+thedecider[5,]))
# calculates highest number of matches

# calculates number of max matches - checks for ambiguity
replicate.best<-sum(thedecider[10,]==max(thedecider[10,]))
```

Appendix 5

```
# tiebreaker function could be used here to break ties

#sets stillambig to false by default
stillambig=F

# check to see if ties are from same-group RT's in the original report.  If so, they form a partial
consensus and would be expected to give similar results, minimicing ambiguity
if (replicate.best>1) multiple<-assignments[which(thedecider[10,]==max(thedecider[10,]))] else NULL
if (replicate.best>1) stillambig<-(max(abs(multiple-mean(multiple))))>(xdeltaRT/60) else NULL

# if replicate.best>1, ambiguous result (multiple possibilities with same # total matches - could
be from two near identical RT matches!)
if (replicate.best>1 & stillambig==T) print(c(j,"still ambiguous after total retally")) else NULL
if (replicate.best>1 & stillambig==T) decision<-NA else NULL

if (replicate.best>1 & stillambig==F) (best<-which(thedecider[1,]==multiple[1])) else NULL
# if replicate best = 1, report best result
if (replicate.best==1) best<-which(thedecider[10,]==max(thedecider[10,])) else NULL
if (replicate.best<1) print(c(j,"um... how did this happen?")) else NULL

if (stillambig==F) (decision<-as.numeric(thedecider["assignment",best])) else NULL
return(decision)
}

### /STEP 10 ###


### ----------------------------------------------------------------------------------------------
----------------- ###


### STEP 11: RESOLUTION OF AMBIGUOUS MATCHES BASED ON CONSENSUS DATA ###

# replicates report so that new RT and intensity values can be added w/o changing the original
report
pre.report<-report
write.csv(pre.report, file="prereport.csv")


# reduces report matrix to only the RT results from p, x,y, and z
  report.rt<-
rbind(report["pRetention",],report["xRetention",],report["yRetention",],report["zRetention",])
    rownames(report.rt)<-c("pRT","xRT","yRT","zRT")
# if true, all RT values in that column agree (all matches agree, or no matches!).
  consensus.rt<-((apply(report.rt,2,max,na.rm=T))-(apply(report.rt,2,min,na.rm=T)))<(xdeltaRT/60)


# clean up step for consensus matches only - go back and see if any matches were not resolved to
assignments.  Search for the concensus rt match in the ambiguous results and report.

# reduces report matrix to only the #matches results from p, x,y, and z
  report.matches<-
rbind(report["pMatches",],report["xMatches",],report["yMatches",],report["zMatches",])
    rownames(report.matches)<-c("pMatches","xMatches","yMatches","zMatches")

# determines if # of assignments in a column equals the number of matched params
noambiguity<-(colSums(report.matches>0, na.rm=T))==(colSums(report.rt!="NA", na.rm=T))

# replicates list with same col numbering as consensus.rt (and report), replaces all values with
NA.
retention<-consensus.rt
retention[which(retention==T)]<-NA
retention[which(retention==F)]<-NA

for (i in 1:length(uniqueID["id",])){
# i=1
j<-as.numeric(uniqueID["id",i])

# pre scenario

# if there are no RT's reported at all, skip the standard
```

```
    if ((sum(report.rt[,i], na.rm=T))==0) next else NULL

# scenario 1
  # if there is RT concensus and no ambiguous matches, simply report the mean RT of the column
    if (noambiguity[i]==T & consensus.rt[i]==T) retention[i]<-mean(report.rt[,i],na.rm=T) else NULL
# scenario 2
  # if there is RT concensus but ambiguous matches, first see if ambiguity can be resolved, then
report mean RT and into values
  # establish target consensus RT
    if (noambiguity[i]==F & consensus.rt[i]==T) target<-mean(report.rt[,i],na.rm=T) else NULL

  # create a blank tally so that script doesn't crash where no tally exists
    tally<-((report.matches[,i])=="z")
    tally["pMatches"]<-0
    tally["xMatches"]<-0
    tally["yMatches"]<-0
    tally["zMatches"]<-0
  # find whether the parent/frag/adduct/addfrag (or multiple) has the ambiguous matches
    if (noambiguity[i]==F & consensus.rt[i]==T) (tally.match<-
((report.matches[,i])>0))&(tally.match[is.na(tally.match)]<-0) else NULL
    if (noambiguity[i]==F & consensus.rt[i]==T) (tally.rt<-
((report.rt[,i])>0))&(tally.rt[is.na(tally.rt)]<-0) else NULL
    if (noambiguity[i]==F & consensus.rt[i]==T) (tally<-tally.match-tally.rt) else NULL
    if (noambiguity[i]==F & consensus.rt[i]==T & tally["pMatches"]==1) print(c(j,"requires parent
rematch")) else NULL

# scenario 2a - if the parent was ambiguous

  # create a default rescheck
    rescheck<-F
  # if there is ambiguity in the parent, but no matches to help resolve in frag, adduct, adduct
frag, then skip to next std.
    if (noambiguity[i]==F & consensus.rt[i]==T & tally["pMatches"]==1) (rescheck<-
((tally.match["xMatches"]+tally.match["yMatches"]+tally.match["zMatches"])==0)) else NULL
    if (noambiguity[i]==F & consensus.rt[i]==T & tally["pMatches"]==1 & rescheck==T)
print(c(j,"however no definitive frag adduct or adduct frag matches to resolve parent ambiguity"))
& next else NULL

  # pulls all features with a match to the std(i) parent mass from the full list of matches
    if (noambiguity[i]==F & consensus.rt[i]==T & tally["pMatches"]==1) (ambiguous<-
full.parentsSTD[which(full.parentsSTD[,i]==1),]) else NULL

  # if one of the ambiguous feature matches is within the RT window, report the RT and intensity of
that feature to the report (for record keeping) and report.rt (for inclusion in the RT averaging)
    # creates a blank test
      test<-0
    if (noambiguity[i]==F & consensus.rt[i]==T & tally["pMatches"]==1) (test<-
min(abs((ambiguous[,"rt"])-(mean(report.rt[,i],na.rm=T))))) else NULL
    if (noambiguity[i]==F & consensus.rt[i]==T & tally["pMatches"]==1 & test>(xdeltaRT/60))
print(c(j,"no additional parent matches")) else NULL
    if (noambiguity[i]==F & consensus.rt[i]==T & tally["pMatches"]==1 & test<(xdeltaRT/60))
(report["pRetention",i]<-ambiguous[which((abs((ambiguous[,"rt"])-
(mean(report.rt[,i],na.rm=T))))<(xdeltaRT/60)),"rt"]) else NULL
    if (noambiguity[i]==F & consensus.rt[i]==T & tally["pMatches"]==1 & test<(xdeltaRT/60))
(report.rt["pRT",i]<-ambiguous[which((abs((ambiguous[,"rt"])-
(mean(report.rt[,i],na.rm=T))))<(xdeltaRT/60)),"rt"]) else NULL
    if (noambiguity[i]==F & consensus.rt[i]==T & tally["pMatches"]==1 & test<(xdeltaRT/60))
(report["pIntensity",i]<-ambiguous[which((abs((ambiguous[,"rt"])-
(mean(report.rt[,i],na.rm=T))))<(xdeltaRT/60)),"into"]) else NULL
  # report mean RT (including newly picked ambiguous match) to the retention vector as a final
answer.
    if (noambiguity[i]==F & consensus.rt[i]==T & tally["pMatches"]==1) (retention[i]<-
mean(report.rt[,i],na.rm=T)) else NULL
# scenario 2b - if the frag, adduct, or adduct frag was ambiguous (currently not delt with)
    if (noambiguity[i]==F & consensus.rt[i]==T & tally["pMatches"]==0) print(c(j,"Ambiguous frag
adduct or adduct frag list")) else NULL
  # this fix copied from scenarios 3 and 4.  may not be appropriate here.
    if (noambiguity[i]==F & consensus.rt[i]==T) retention[i]<-
decisionsdecisions(i,matchedParents,xMatches,yMatches,zMatches) else NULL
```

Appendix 5

```
# scenario 3 - no absolute consensus, but no ambiguous sections either.  Action = rematch each best
RT across all params to see which has most support.
  if (noambiguity[i]==T & consensus.rt[i]==F) retention[i]<-
decisionsdecisions(i,matchedParents,xMatches,yMatches,zMatches) else NULL
  if (noambiguity[i]==T & consensus.rt[i]==F) print(c(j,"indecision in RTs returned by parent,
frag, adduct, and adductfrag matching")) else NULL

# scenario 4 - no absolute consensus, some sections ambiguous.  Action is same as in #3.  However,
this scenario may include resolution of ambiguity and filling to the report in the future.
  if (noambiguity[i]==F & consensus.rt[i]==F) retention[i]<-
decisionsdecisions(i,matchedParents,xMatches,yMatches,zMatches) else NULL
  if (noambiguity[i]==F & consensus.rt[i]==F) print(c(j,"indecision and ambiguity in RTs returned
by parent, frag, adduct, and adductfrag matching")) else NULL



}

report<-rbind(report,retention)

write.csv(report, file="report.csv")

summary.report<-rbind(report["id",],report["tMZ",],report["retention",])
colnames(summary.report)<-c("id","tMZ","eRT")

summary.report<-t(summary.report)
colnames(summary.report)<-c("id","tMZ","eRT")
write.csv(summary.report, file="summary_report.csv")




### /STEP 11 ###


### ----------------------------------------------------------------------------------------------
----------------- ###


### STEP 12: PLOT PARENT AND SUBLIST MATCHING DATA ###

# red (or non-black or grey) dots indicate duplicated entries in the tDB.
par(mfrow=c(1,1))

## plotting - this could be done at once, but it looks more interesting if done one standard at a
time #
for(i in 1:nSTDs){
plot.new()
plot(STD[,"rt"],STD[,"mz"], col=colors()[230], xlim=c(xlim.min,xlim.max),
ylim=c(ylim.min,ylim.max),xlab="retention time", ylab="m/z", main="STD mix")

#i=15
# plots a horizontal line for "i" standard parents (red), frags (blue), adducts (red), and adduct
fragments (orange)  representing the values to which emperical measurements from the peak list
(STD) will be matched #
  xBlotter<-x[[i]]
  yBlotter<-y[[i]]
  zBlotter<-z[[i]]

# plots parent (red)
  par (new=T)
  # plots the individual tDB parent
    abline(h=(uniqueID["tMZ",i]),col=2,lty=3, xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))
  par (new=T)
  # replots all points on each iteration, but colors black only those with a positive match to a
tDB parent mass (sequentially, for each of nSTDs)
    plot(STD[,"rt"],STD[,"mz"], col=possibleParents[,(1+i)], xlab="", ylab="", pch=16,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))
```

Appendix 5

```
# plots fragments (green)
  # plots all fragments for tDB standard i
    abline(h=(xBlotter[,"frag"]),col=3,lty=3, xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))
  par (new=T)
  # plots fragment matches
    plot(STD[,"rt"],STD[,"mz"], col=fragmentsSTD[,(i)], xlab="", ylab="", pch=16,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))

#plots adducts (blue)
  # plots all adducts for tDB standard i
    abline(h=(yBlotter[,"adducts"]),col=4,lty=3, xlim=c(xlim.min,xlim.max),
ylim=c(ylim.min,ylim.max))
  par (new=T)
  # plots fragment matches
    plot(STD[,"rt"],STD[,"mz"], col=adductsSTD[,(i)], xlab="", ylab="", pch=16,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))

#plots adductfragments
  # plots all fragments for tDB standard i
    abline(h=(zBlotter[,"adductfrags"]),col=5,lty=3, xlim=c(xlim.min,xlim.max),
ylim=c(ylim.min,ylim.max))
  par (new=T)
  # plots fragment matches
    plot(STD[,"rt"],STD[,"mz"], col=adductfragmentsSTD[,(i)], xlab="", ylab="", pch=16,
xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))

#plots consensus RT
  # plots vertical line at the decided consensus RT
    abline(v=retention[i],col=1,lty=3, xlim=c(xlim.min,xlim.max), ylim=c(ylim.min,ylim.max))

var_filename <- uniqueID["id",i]
dev.copy(png, paste(var_filename, '.png', sep='' ))
dev.off()

  Sys.sleep(.1)
# sets the background to be white for plotting - this helps when saving the plots to png
  par (new=F)
#erase all plotting, replace with grey base full plot for next tDB standard
  plot(STD[,"rt"],STD[,"mz"], col=colors()[230], xlim=c(xlim.min,xlim.max),
ylim=c(ylim.min,ylim.max),xlab="retention time", ylab="m/z", main="STD mix")

}


### /STEP 12 ###
```

Appendix 6: Publications (2012-2014)

1. Sahota A, Parihar JS, Capaccione KM, Yang M, Noll K, Gordon D, Reimer D, Yang I, Buckley BT, Polunas M, Reuhl KR, **Lewis MR**, Ward MD, Goldfarb DS, Tischfield JA. Novel Cystine Ester Mimics for the Treatment of Cystinuria-induced Urolithiasis in a Knockout Mouse Model. *Urology*. 2014, 84(5), 1249.e9-1249.e15.
2. Dona AC, Jimenez B, Schaefer H, Humpfer E, Spraul M, **Lewis MR**, Pearce JTM, Holmes E, Lindon JC, Nicholson JK. Precision High-Throughput Proton NMR Spectroscopy of Human Urine, Serum, and Plasma for Large-Scale Metabolic Phenotyping. *Analytical Chemistry*. 2014, 86(19), 9887-9894.
3. Ladep NG, Dona AC, **Lewis MR**, Crossey MME, Lemoine M, Okeke E, Shimakawa Y, Duguru M, Njai HF, Fye HKS, Taal M, Chetwood J, Kasstan B, Khan SA, Garside DA, Wijeyesekera A, Thillainayagam AV, Banwat E, Thursz MR, Nicholson JK, Njie R, Holmes E, Taylor-Robinson SD. Discovery and Validation of Urinary Metabotypes for the Diagnosis of Hepatocellular Carcinoma in West Africans. *Hepatology*. 2014, 60(4), 1291-1301
4. Sarafian MH, Gaudin M, **Lewis MR**, Martin F-P, Holmes E, Nicholson JK, Dumas M-E. Objective Set of Criteria for Optimization of Sample Preparation Procedures for Ultra-High Throughput Untargeted Blood Plasma Lipid Profiling by Ultra Performance Liquid Chromatography-Mass Spectrometry. *Analytical Chemsitry*. 2014, 86(12), 5766-5774.
5. Mirnezami R, Spagou K, Vorkas PA, **Lewis MR**, Kinross J, Want E, Shion H, Goldin RD, Darzi A, Takats Z, Holmes E, Cloarec O, Nicholson JK. Chemical mapping of the colorectal cancer microenvironment via MALDI imaging mass spectrometry (MALDI-MSI) reveals novel cancer-associated field effects. 2014. *Molecular Oncology*, 8(1), 39-49.
6. Schumacher J, Behrends V, Pan Z, Brown DR, Heydenreich F, **Lewis MR**, Bennett MH, Razzaghi B, Komorowski M, Barahona M, Stumpf MPH, Wigneshweraraj S, Bundy JG, Buck M. Nitrogen and Carbon Status Are Integrated at the Transcriptional Level by the Nitrogen Regulator NtrC In Vivo. *MBio*. 2013, 4(6), e00881-13.

Appendix 6: Publications (2012-2014)

7. Balog J, Sasi-Szabo L, Kinross J, **Lewis MR**, Muirhead LJ, Veselkov K, Mirnezami R, Dezso B, Damjanovich L, Darzi A, Nicholson JK, Takats Z. Intraoperative Tissue Identification Using Rapid Evaporative Ionization Mass Spectrometry. 2013, *Science Translational Medicine*, 5(194).
8. Swann JR, Spagou K, **Lewis M**, Nicholson JK, Glei DA, Seeman TE, Coe CL, Goldman N, Ryff CD, Weinstein M, Holmes E. Microbial-Mammalian Cometabolites Dominate the Age-associated Urinary Metabolic Phenotype in Taiwanese and American Populations. *Journal of Proteome Research*. 2013, 12(7), 3166-3180. ***(Selected for reproduction within this thesis)***
9. Hong M-G, Karlsson R, Magnusson PKE, **Lewis MR**, Isaacs W, Zheng LS, Xu J, Gronberg H, Ingelsson E, Pawitan Y, Broeckling C, Prenni JE, Wiklund F, Prince JA. A Genome-Wide Assessment of Variability in Human Serum Metabolism. *Human Mutation*. 2013, 34(3), 515-524.
10. Heuberger AL, Broeckling CD, **Lewis MR**, Salazar L, Bouckaert P, Prenni JE. Metabolomic profiling of beer reveals effect of temperature on non-volatile small molecules during short-term storage. *Food Chemistry*. 2012, 135(3), 1284-1289.
11. Saric J, Want EJ, Duthaler U, **Lewis M**, Keiser J, Shockcor JP, Ross GA, Nicholson JK, Holmes E, Tavares MFM. Systematic Evaluation of Extraction Methods for Multiplatform-Based Metabotyping: Application to the Fasciola hepatica Metabolome. *Analytical Chemistry*. 2012, 84(16), 6963-6972.
12. Thompson MD, Mensack MM, Jiang W, Zhu Z, **Lewis MR**, McGinley JN, Brick MA, Thompson HJ. Cell signaling pathways associated with a reduction in mammary cancer burden by dietary common bean (Phaseolus vulgaris L.). *Carcinogenesis*. 2012, 33(1), 226-232.

Microbial-Mammalian Cometabolites Dominate the Age-associated Urinary Metabolic Phenotype in Taiwanese and American Populations

Swann JR, Spagou K, Lewis M, Nicholson JK, Glei DA, Seeman TE, Coe CL, Goldman N, Ryff CD, Weinstein M, Holmes E.

Journal of Proteome Research 2013, 12(7), 3166-3180

Journal of proteome research

Article

pubs.acs.org/jpr

Microbial−Mammalian Cometabolites Dominate the Age-associated Urinary Metabolic Phenotype in Taiwanese and American Populations

Jonathan R. Swann,[†] Konstantina Spagou,[‡] Matthew Lewis,[‡] Jeremy K. Nicholson,[‡] Dana A. Glei,[§] Teresa E. Seeman,[∥] Christopher L. Coe,[⊥] Noreen Goldman,[#] Carol D. Ryff,[¶] Maxine Weinstein,[§] and Elaine Holmes*[,‡]

[†]Department of Food and Nutritional Sciences, School of Chemistry, Food and Pharmacy, University of Reading, Whiteknights, Reading, RG6 6AP, United Kingdom

[‡]Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, South Kensington, London SW7 2AZ, United Kingdom

[§]Center for Population and Health, Georgetown University, Washington, D.C., United States

[∥]Division of Geriatrics, UCLA David Geffen School of Medicine, Los Angeles, California 90095, United States

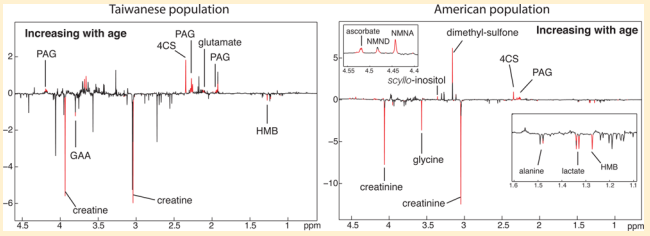
[⊥]Harlow Center for Biological Psychology, University of Wisconsin, Madison, Wisconsin, United States

[#]Office of Population Research, Princeton University, 243 Wallace Hall, Princeton, New Jersey 08544-2091, United States

[¶]Institute of Aging, Department of Psychology, Medical Science Center, University of Wisconsin, Madison, Wisconsin 53706, United States

Supporting Information

ABSTRACT: Understanding the metabolic processes associated with aging is key to developing effective management and treatment strategies for age-related diseases. We investigated the metabolic profiles associated with age in a Taiwanese and an American population. ^{1}H NMR spectral profiles were generated for urine specimens collected from the Taiwanese Social Environment and Biomarkers of Aging Study (SEBAS; n = 857; age 54−91 years) and the Mid-Life in the USA study (MIDUS II; n = 1148; age 35−86 years). Multivariate and univariate linear projection methods revealed some common age-related characteristics in urinary metabolite profiles in the American and Taiwanese populations, as well as some distinctive features. In both cases, two metabolites—4-cresyl sulfate (4CS) and phenylacetylglutamine (PAG)—were positively associated with age. In addition, creatine and β-hydroxy-β-methylbutyrate (HMB) were negatively correlated with age in both populations ($p < 4 \times 10^{-6}$). These age-associated gradients in creatine and HMB reflect decreasing muscle mass with age. The systematic increase in PAG and 4CS was confirmed using ultraperformance liquid chromatography−mass spectrometry (UPLC−MS). Both are products of concerted microbial−mammalian host cometabolism and indicate an age-related association with the balance of host−microbiome metabolism.



KEYWORDS: *age, sex, metabolic profiling, NMR spectroscopy, 4-cresyl sulfate, phenylacetylglutamine*

INTRODUCTION

The chronic nature of most diseases associated with aging, coupled with the increased probability of elderly individuals presenting with multiple pathologies requiring complex therapeutic management strategies, makes analysis of age-related conditions challenging. Aging is associated with a general decline in physiological function, particularly in the intestine, where a decrease in intestinal motility, a reduction in the capacity of the immune system and changes in the beneficial and hostile gut microbiota contribute to the general decline in health. Many elegant studies in short-lived model organisms such as the nematode worm *Caenorhabditis elegans* and the mouse have contributed to our current understanding of the aging process.[1,2] However, the true complexity of aging in human populations cannot be fully characterized in these animal models, given the diverse exposure of humans to a myriad of physical, environmental and social stressors.[3,4] Thus, in parallel to exploring experimental models of aging, there is a need for research into the mechanisms and consequences of aging in human populations. Epidemiological studies investigating population differences in the prevalence of diseases across countries[5−7] and between men and women[8] offer a particularly useful resource for studying aging.

Received: January 5, 2013
Published: May 23, 2013

Metabolic phenotyping and metabolome-wide association studies (MWAS) offer a powerful new means for discovering molecular biomarkers and metabolic pathways that underlie disease risk.[9,10] This approach uses high-resolution spectroscopic techniques and mathematical modeling to generate a molecular fingerprint of a biological specimen[11] and can provide a novel framework for identifying appropriate therapeutic intervention strategies at the individual and population level. A particular strength of metabolic phenotyping lies in its ability to reveal a representative overview of host, extra-genomic and environmental contributions to metabolism.

Metabolic profiling approaches have been applied to studies on age-associated diseases in both nonhuman[2,12] and human populations, with a focus on identifying age-related changes in the biochemical composition of serum or plasma. Several groups have reported decreased serum carnitines, acylcarnitines and amino acids with age and increased free fatty acid levels in aging rodents.[13,14] In contrast, other studies have found an increase in free serum carnitine with age in humans.[15] While plasma provides a useful system-level readout of the physiological status of an organism at a given point in time, urine provides time-averaged information on the metabolic events that have occurred throughout the whole animal. The metabolic signature of urine is influenced by the host's genome and physiology but also provides a window on extrinsic input from dietary factors and the gut microbiome.

Here we apply a spectroscopic profiling approach to define the metabolic signature of aging in two distinct human populations—the Taiwanese Social Environment and Biomarkers of Aging Study (SEBAS)[16] and the Mid-Life in the USA (MIDUS II)[17] cohorts—using ^{1}H nuclear magnetic resonance (NMR) spectroscopy and ultraperformance liquid chromatography–mass spectrometry (UPLC–MS) of urine specimens. Through this approach we identify the global sources of metabolic variation and sex-specific elements within the metabolic signatures of these geographically and culturally distinct populations. In addition, we identify clear metabolic correlates of biological aging in relation to declining muscle metabolism and also age-related variation in the functionality of several pathways involved in gut microbial–host metabolic regulation.

METHODS AND MATERIALS

Description of Populations and Specimen Collections

SEBAS Study. A total of 857 urine specimens from the 2000 SEBAS study (age range 54–91; mean 68 years) were shipped from the Lombardi Comprehensive Cancer Center, Georgetown University to Imperial College London. This specimen set comprised urine from 368 females and 489 males. Specimens were stored at Imperial College at −80 °C prior to analysis.

MIDUS Study. A total of 1148 urine specimens from the MIDUS II study (age range 35–86; mean 57 years) were shipped from the Harlow laboratory, University of Wisconsin and stored at −80 °C at Imperial College prior to analysis. Participants included 651 females and 497 males. Both sample sets were 12-h overnight urine collections.

The demographic characteristics of the SEBAS and MIDUS participants are summarized in Table 1.

^{1}H NMR Spectroscopic Analysis

Quality control (QC) aliquots for NMR analysis were prepared by combining aliquots of urine from randomly selected subgroups of individuals. For each cohort, SEBAS and MIDUS, specimens were randomized and interspersed with QC aliquots (using a total of 129 QC aliquots) in order to assess data quality and variation over the analytical measurement period. Specimens were prepared and spectra acquired using in-house protocols[18] adopting a standard one-dimensional pulse sequence with suppression of the water resonance. Briefly, urine specimens were prepared by the addition of phosphate buffer made up in deuterium oxide containing 1 mM 3-(trimethylsilyl)-[2,2,3,3-$^2H_4$]-propionic acid sodium salt (TSP) as an external reference and 2 mM sodium azide as a bacteriocide. For each specimen, a standard one-dimensional NMR spectrum was acquired with water peak suppression using a standard pulse sequence (recycle delay (RD)-90°-t_1-90°-t_m-90°-acquire free induction decay (FID)). A mixing time (t_m) of 100 ms was used and the RD was set at 2 s. The 90° pulse length was approximately 12 μs and t_1 was set to 3 μs. An acquisition time per scan was 2.73 s and, for each specimen, 8 dummy scans were followed by 128 scans. The spectra were collected into 64K data points using a spectral width of 20 ppm.

Table 1. Study Participant Information for SEBAS and MIDUS

	SEBAS	MIDUS
Total specimens NMR[a]	857	1148
Total specimens MS	725	1196
Age range	54–91	35–86
Sex (female/male)	368/489	651/497

[a]The number of urine specimens for NMR and MS differ due to the number of specimens excluded based on the differing analytical constraints of the two techniques. For NMR analysis, specimens were excluded if the glucose levels or ethanol concentrations were too high, which caused bias in the models. For MS specimens were excluded where there was insufficient specimen volume or where specimens contained a polyethylene glycol contaminant, possibly leached from the storage vials. Outliers in the PCA scores plots of the NMR data were evaluated using the Hotellings T ellipse and discarded where appropriate in order to remove undue influence of artifacts on the models.

Preprocessing and Modeling of the NMR Spectral Data

Spectra were phased, corrected for baseline distortions and referenced to the TSP signal at δ 0.00. The region between δ 4.70 and 6.20 containing the residual water resonance and the urea peak was removed for all spectra. For the MIDUS spectral data, the region containing the methyl resonance of acetate (δ 1.92) was removed owing to pretreatment of these aliquots with acetate. The remaining spectral variables between δ 0.70–4.70 and δ 6.20–10.00 were normalized to the sum of the spectral integral prior to analysis using principal components analysis (PCA). Data were analyzed with and without peak alignment using the algorithm defined by Veselkov et al.[19] The main sources of variation in the data were identified and further explored. Partial least-squares discriminant analysis (PLS-DA) was applied to the data with and without the application of an orthogonal filter to remove extraneous variation and to establish metabolic patterns relating to a variety of participant variables including age and sex. The predictive performance of the models was assessed using a 7-fold cross-validation approach and the Q^2Y (goodness of prediction) values are provided. Permutation testing (1000 permutations) has been performed to ensure the validity of the PLS models. Linear regression was used to measure the statistical significance of the metabolic variations. A cutoff of $p < 4 \times 10^{-6}$ was used based

dx.doi.org/10.1021/pr4000152 | J. Proteome Res. 2013, 12, 3166–3180

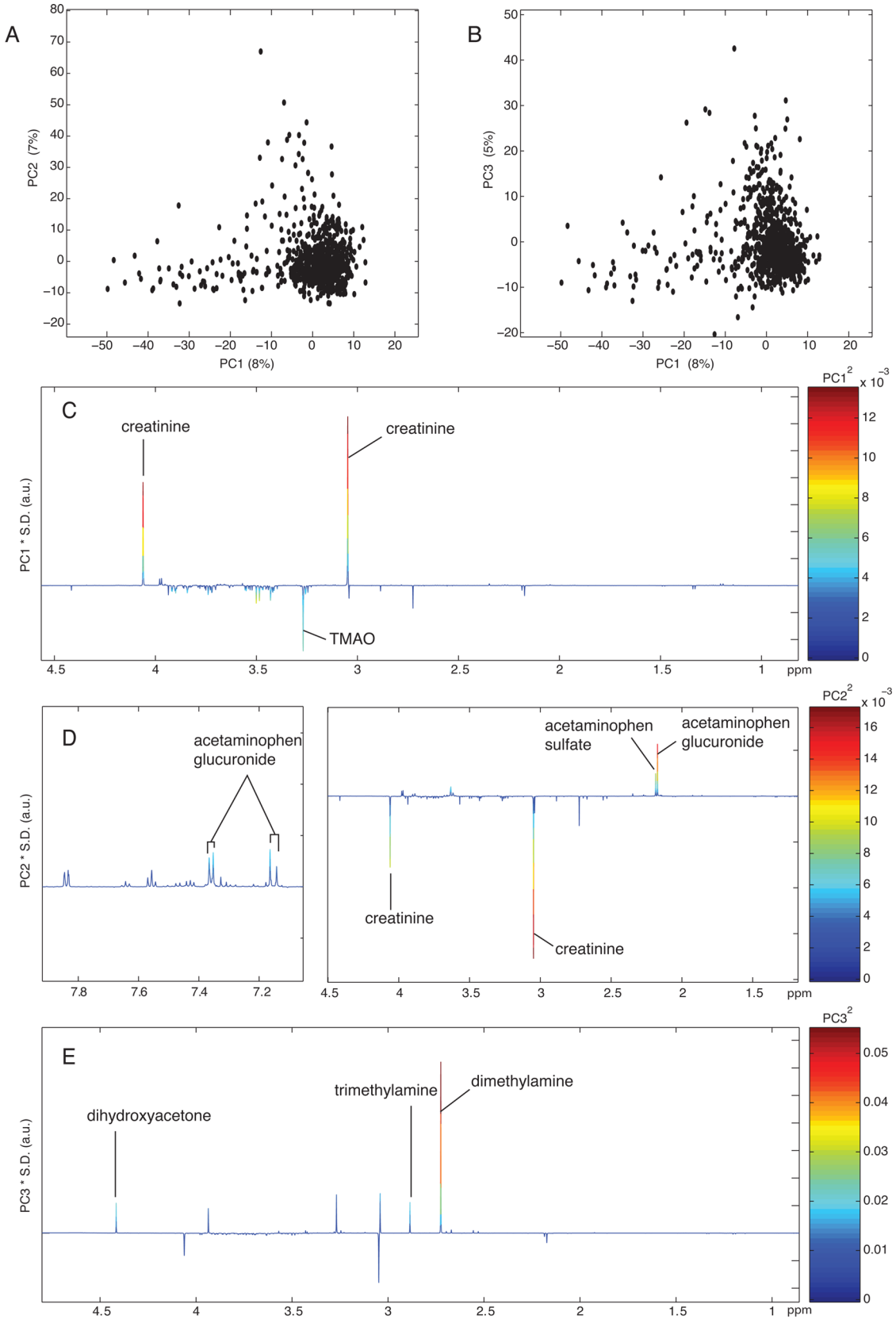


Figure 1. PCA model of the urinary profiles of all SEBAS participants. Scores plots for (A) PC1 vs PC2 and (B) PC1 vs PC3 (% variance explained in parentheses). Product of PC loadings with standard deviation of the entire data set, colored by the square of the PC shown for (C) PC1, (D) PC2 and (E) PC3.

on the method described by Chadeau-Hyam et al.[20] for selecting a suitable level of significance in metabolome wide association studies (MWAS) with an expected family wise error rate of 5% for 13 000 variables.

dx.doi.org/10.1021/pr4000152 | J. Proteome Res. 2013, 12, 3166−3180

UPLC–MS Spectral Analysis

UPLC–MS analysis was performed to validate the NMR-detected correlation of PAG and 4-cresyl sulfate with age and to explore other possible age related variation in the urinary metabolome using optimized protocols for urine metabolite profiling.[21] Briefly, urine specimens were prepared by dilution (1:1) with water (Sigma, LC–MS grade), vortexed for ten seconds, and centrifuged at 16 000× *g* for 10 min. Two hundred microliters were aliquoted into 96-well 350 μL plates (Waters Corporation, Milford, MA) with cap mats (VWR, U.K.). A composite quality control (QC) aliquot was prepared by combining 50 μL from 775 randomly selected SEBAS and MIDUS specimens. The QC aliquot was subaliquoted to minimize freeze–thaw cycle effects and stored frozen until required for the analysis. Ten analyses of the QC aliquot were performed at the beginning of the analytical run for system conditioning. A single QC aliquot injection was performed at 10-aliquot intervals throughout the subsequent data acquisition to provide data for the assessment of analytical reproducibility including peak retention times and detector response. Additionally, five blanks were injected prior to the injection of QC-conditioning aliquots in order to ensure that there was no contamination from the UPLC system, and again at the end of the experiment to ensure that specimen carryover was not observed.

Metabolic profiling was performed on an Acquity UPLC system (Waters Corp., Milford, MA) coupled to an LCT Premier time-of-flight mass spectrometer (Waters Corp., Manchester, U.K.). UPLC–MS conditions were optimized in terms of peak shape, reproducibility and retention times of analytes. Chromatography was performed using an Acquity HSS T3 column, 2.1 × 100 mm column (Waters Corp., Milford, MA) held at 40 °C. Separation was performed using gradient elution with 0.1% (v/v) formic acid in H_2O (A) and 0.1% (v/v) formic acid in ACN (B) at a flow rate of 0.5 mL/min. Starting conditions were 99.9% A and 0.1% B for 1.0 min, changing linearly to 15% B over the next 2 min, and then to 50% B over the next 3 min, and finally to 95% B in the next 3 min and kept for 1 min. Afterward the solvent composition returned to starting conditions over 0.1 min, followed by re-equilibration for 2 min prior to the next injection.

Mass spectrometry was performed using electrospray in both positive and negative ionization modes (ESI+ and ESI−). The capillary voltage was 3.2 kV (ESI+) and 2.4 kV (ESI−), cone voltage was 35 V, desolvation temperature was 350 °C, and source temperature was 120 °C. The cone gas flow rate was 25 L/h, and desolvation gas flow rate was 900 L/h. The LCT Premier was operated in V optics mode with a scan time of 0.2 s and interscan delay of 0.01 s. For mass accuracy, a LockSpray interface was used with a 20 μg/L leucine enkephalin (555.2645 amu) solution (50/50 ACN/H_2O with 0.1% v/v formic acid) at 70 μL/min as the lock mass. Data were collected in centroid mode with a scan range of 50–1000 *m/z*, with lockmass scans collected every 15 s and averaged over 3 scans to perform mass correction.

Preprocessing and Modeling of the UPLC–MS Data

Since the system is not generally stable during the first injections, the first 10 QC samples were used to ensure that stability had been attained, after which the QC-conditioning aliquots were excluded from further data processing. The rest of the raw data (i.e., the target specimens plus the remaining QC aliquots) within the run were converted to netCDF format using the DataBridge tool implemented in MassLynx software (Waters Corporation, Milford, MA).

The data were preprocessed using the freely available XCMS software. The Centwave algorithm was used for peak picking with a peak width window of 3–15 s, the *m/z* width for the grouping was changed to 0.1 Da, the bandwidth parameter was kept to default (30 s) for the first grouping and was subsequently determined from the time deviation profile plot after retention time correction. An output table was obtained at the end comprising *m/z*, RT and intensity values of the detected metabolite features in each specimen.

The data were then normalized in R with an in-house script.[22] The coefficient of variation (CV = standard deviation/mean) values were calculated for all the intensities of metabolite features (mz_Rt) in the QC samples analyzed within the run (see Supporting Information for details). In the generated data sets features with a CV higher than 30% in replicated injections of the QC aliquots interspersed within the run were removed. The output table was exported into SIMCA-P+ 12.0.1 software (Umetrics, Umeå, Sweden) for multivariate analysis. Principal component analysis (PCA), partial least-squares-discriminant analysis (PLS-DA) and orthogonal projection on latent structures-discriminant analysis (OPLS-DA) were performed on all data.

Adjustment of Data Sets for Differential Age Ranges between the SEBAS and MIDUS Studies

Owing to different age ranges between the two study populations (SEBAS 54–91 years, mean 68 years; MIDUS 35–86 years, mean 57 years), auxiliary models were constructed using a restricted age range that comprised the overlap between the two studies (ages 54–86 years); the results are reported in Supporting Information (Figures S3–S5).

RESULTS

The analytical platforms and methods were robust and reliable, as indicated by the coefficients of variation for the quality control specimens. Moreover, the analytical quality of the data was good across both the NMR spectroscopy and the UPLC–MS data, obtained for both the SEBAS and the MIDUS data sets, with the one exception of ESI negative mode data for the MIDUS cohort. No adjustment of the MS data for run order was necessary. For the UPLC–MS in ESI+ ion mode, the coefficients of variation for the QC samples were 25.2 ± 19.1 and 23 ± 17.7 for SEBAS and MIDUS, respectively. ESI− ion mode gave similar results with CV values 31.8 ± 19.3 for the SEBAS study. For the MIDUS study, the CV ESI− ion values were high (50 ± 53.3); therefore, we refrained from further analysis of the negative ionization mode data set.

Global Analysis of the ^{1}H NMR Urine Data

The scores and loadings plots from the global PCA model for the SEBAS data set (Figure 1) show that the first component was dominated by creatinine and trimethylamine-*N*-oxide (TMAO), which represented the greatest sources of variation across the specimen set. Creatinine is a crude indicator of muscle mass and can vary with sex and age. TMAO is associated with consumption of certain fish and shellfish, where it functions as an antifreeze agent and an osmolyte and has been shown to be elevated in urine after consumption of diets rich in phytoestrogens, for example, soy or miso. The variance in the second component was dominated by metabolites related to acetaminophen, namely acetaminophen glucuronide and acetaminophen sulfate. Methylamines and a singlet (δ 4.41)

dx.doi.org/10.1021/pr4000152 | J. Proteome Res. 2013, 12, 3166–3180

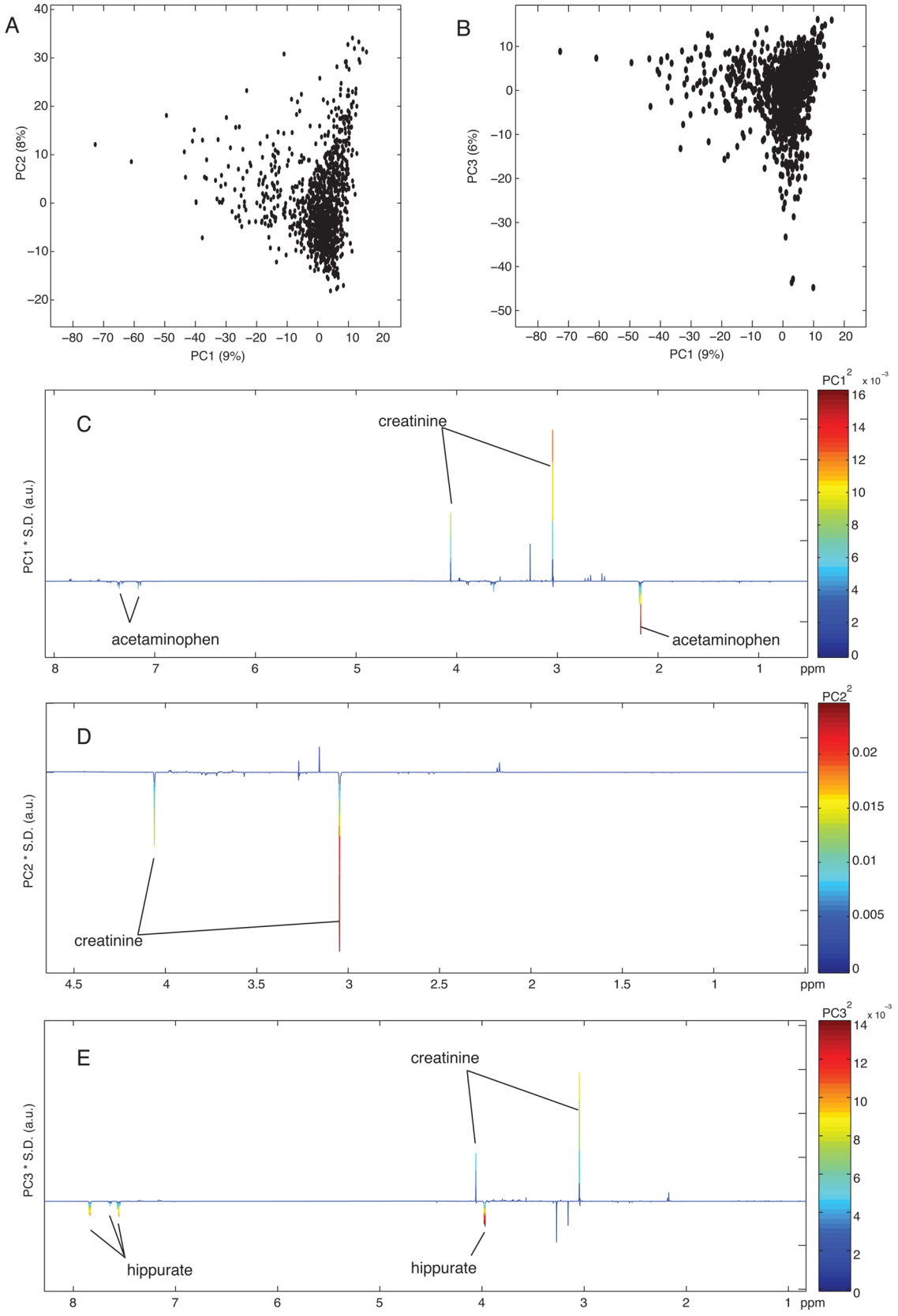


Figure 2. PCA model of the urinary profiles of all MIDUS participants. Scores plots for (A) PC1 vs PC2 and (B) PC1 vs PC3 (% variance explained in parentheses). Product of PC loadings with standard deviation of the entire data set, colored by the square of the PC shown for (C) PC1, (D) PC2 and (E) PC3.

tentatively assigned as dihydroxyacetone exerted the greatest influence on the third principal component.

Similarly to the SEBAS data set, the first component of the PCA model calculated for the MIDUS data set was strongly

dx.doi.org/10.1021/pr4000152 | J. Proteome Res. 2013, 12, 3166−3180

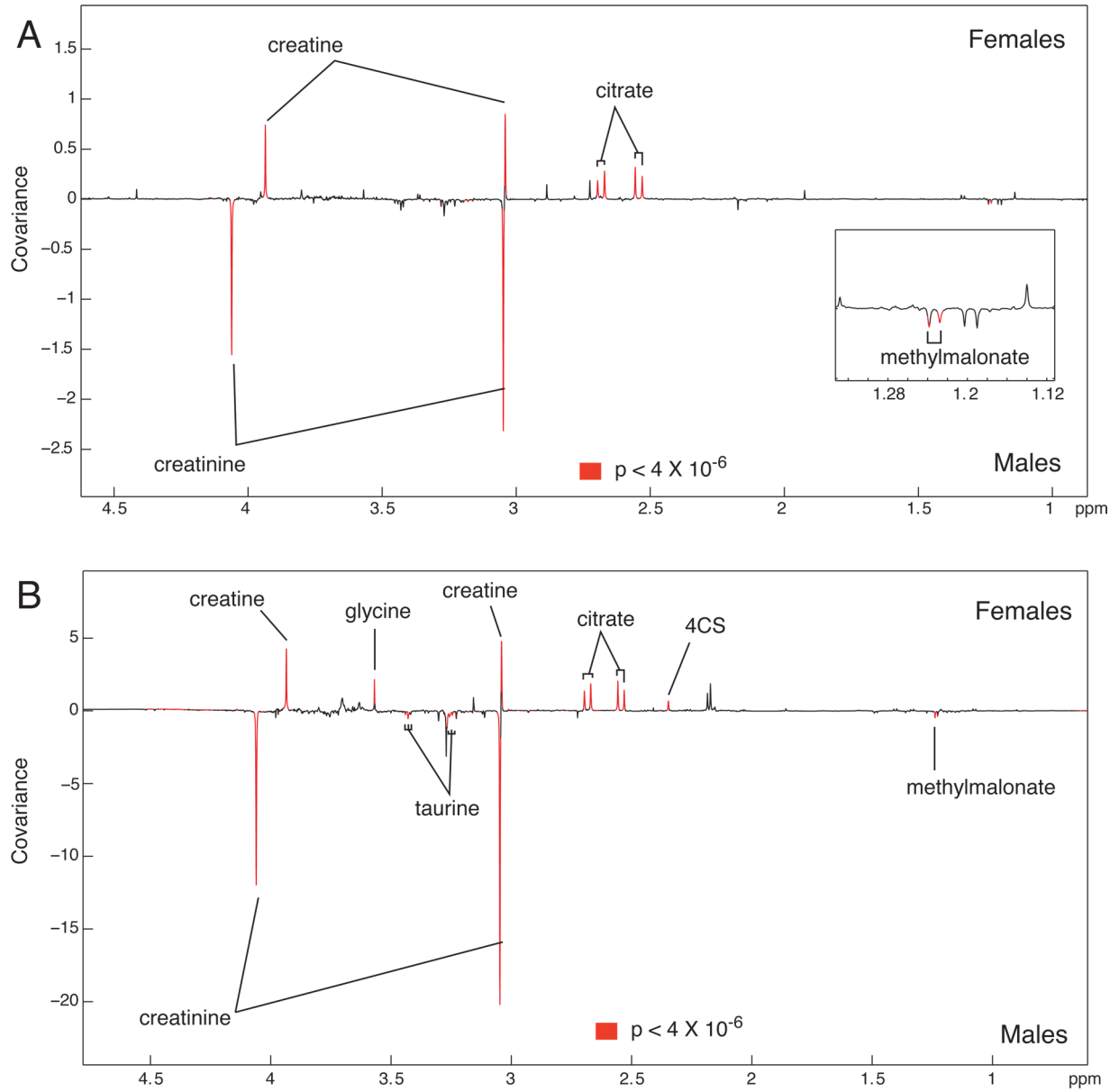


Figure 3. Linear regression analysis correlating ^{1}H NMR spectral profiles of urine with sex. Covariance plots derived from linear regression analysis for (A) SEBAS and (B) MIDUS, color-coded by significance. Significance determined by $p < 4 \times 10^{-6}$, the metabolome-wide significance level (MWSL).

influenced by creatinine (Figure 2). In addition, acetaminophen metabolites also made a substantial contribution to the first component. Although the principal components are linear and orthogonal, creatinine also dominated the second component. When a metabolite is influential in the loadings explaining more than one component, it is generally because the variance of that metabolite is determined by more than one major source of variation in the data set. The mammalian−microbial cometabolite hippurate accounted for the majority of the variance in the third component of the MIDUS II model.

Since methylamines contributed strongly to the variation in the SEBAS but not the MIDUS II data set, the urinary concentrations of trimethylamine (TMA) and dimethylamine (DMA) were calculated from the integrals at δ 2.88 and δ 2.72 respectively and found to be significantly different for the Taiwanese (mean concentration TMA = 0.11 ± 0.11 mM and DMA = 0.44 ± 0.46 mM) and American populations (mean concentration TMA = 0.02 ± 0.01 mM and DMA = 0.15 ± 0.1 mM). Because of overlap with taurine and other metabolites, the integral values for the TMAO signal were not calculated but visual inspection of the data suggested that TMAO was found in higher concentrations in the urine of Taiwanese participants.

Sex-related Differences in Urinary Metabolic Phenotypes

Because creatinine was one of the major sources of variation found in both the SEBAS and MIDUS cohorts, and is known to differ with both age and sex, the influence of sex on the NMR derived metabolic profiles was characterized prior to focusing on age-related metabolic differences. Using an unsupervised PCA approach, no clear discrimination of specimens according to sex could be seen for either the SEBAS or the MIDUS cohorts (Supporting Information Figure S1) indicating that the major sources of variation in urine composition across the populations were not sex-related.

dx.doi.org/10.1021/pr4000152 | J. Proteome Res. 2013, 12, 3166−3180

Table 2. Age-related Variation in SEBAS and MIDUS Urinary Metabolic Profiles using Linear Regression[a]

	SEBAS						MIDUS					
	all		females		males		all		females		males	
metabolite	*R*	*P*-value	*R*	*P*-value	*R*	*P*-value	*R*	*P*-value	*R*	*P*-value	*R*	*P*-value
4CS	+0.32	1.53×10^{-21}	+0.34	2.66×10^{-11}	+0.30	1.12×10^{-11}	+0.23	9.83×10^{-16}	+0.20	3.21×10^{-7}	+0.20	3.21×10^{-7}
PAG	+0.32	1.20×10^{-21}	+0.34	1.53×10^{-11}	+0.31	4.08×10^{-12}	+0.29	6.55×10^{-23}	+0.29	4.57×10^{-14}	+0.29	4.57×10^{-14}
glutamate	+0.23	1.32×10^{-11}	–	–	–	–	–	–	–	–	–	–
creatine	−0.23	3.67×10^{-12}	−0.28	1.4×10^{-6}	–	–	−0.20	2.77×10^{-11}	−0.20	2.21×10^{-7}	–	–
GAA	−0.16	3.79×10^{-6}	–	–	–	–	–	–	–	–	–	–
HMB	−0.18	2.14×10^{-6}	–	–	−0.23	1.63×10^{-7}	−0.26	1.31×10^{-19}	−0.28	5.19×10^{-13}	−0.28	5.19×10^{-13}
NMNA	–	–	–	–	–	–	+0.19	1.40×10^{-10}	+0.26	8.92×10^{-12}	–	–
NMND	–	–	–	–	–	–	+0.15	4.4×10^{-7}	+0.21	9.73×10^{-8}	–	–
4PY	–	–	–	–	–	–	+0.15	6.68×10^{-7}	–	–	–	–
scyllo-inositol	–	–	–	–	–	–	+0.21	1.29×10^{-12}	+0.28	3.12×10^{-13}	–	–
dimethyl sulfone	–	–	–	–	–	–	+0.14	1.17×10^{-6}	–	–	–	–
ascorbate	–	–	–	–	–	–	+0.18	4.47×10^{-10}	+0.25	1.42×10^{-10}	–	–
creatinine	–	–	–	–	–	–	−0.26	2.09×10^{-19}	−0.31	1.90×10^{-15}	−0.30	1.90×10^{-15}
glycine	–	–	–	–	–	–	−0.29	2.03×10^{-23}	−0.34	1.24×10^{-18}	−0.34	1.24×10^{-18}
alanine	–	–	–	–	–	–	−0.15	2.84×10^{-7}	–	–	–	–
lactate	–	–	–	–	–	–	−0.15	2.28×10^{-7}	−0.23	6.79×10^{-9}	–	–

[a]Correlation coefficients (R) and corresponding P-values are given for each metabolite significantly associated with age. Age-related variation is provided for all SEBAS and MIDUS participants and stratified by sex. 4CS, 4-cresyl-sulfate; 4PY, *N*-methyl-4-pyridone-3-carboxamide; GAA, guanidinoacetic acid; HMB, *β*-hydroxy-*β*-methylbutyrate; NMNA, *N*-methyl nicotinic acid; NMND, *N*-methyl nicotinamide; PAG, phenylacetylglutamine.

dx.doi.org/10.1021/pr400015z | J. Proteome Res. 2013, 12, 3166−3180

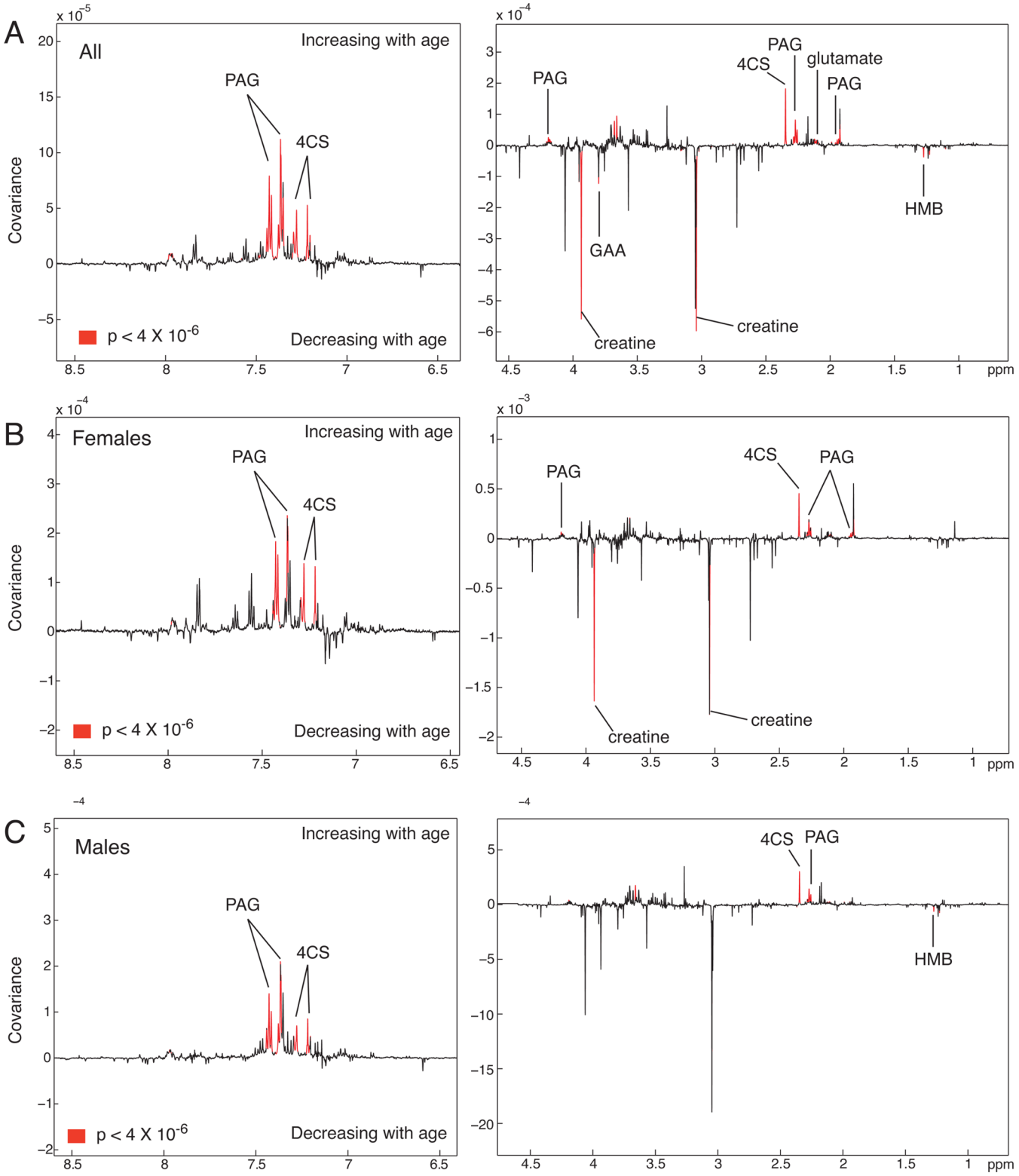


Figure 4. Age-related variation in SEBAS urinary metabolic profiles using linear regression. Covariance plots derived from linear regression analysis for (A) all SEBAS participants and stratified by sex ((B) females and (C) males). Covariance plots are colored by significance ($p < 4 \times 10^{-6}$). HMB, β-hydroxy-β-methylbutyrate; PAG, phenylacetylglutamine; 4CS, 4-cresyl-sulfate.

OPLS-DA and linear regression analysis were used to establish that systematic differences in the metabolic phenotypes of men and women existed and to extract the sex-dependent metabolic characteristics. For the SEBAS specimen set (Supporting Information Figure S2A) a model with a predictive value (Q^2Y) of 0.236 for a 1 orthogonal, 1 aligned component model was obtained. As expected, the major discriminating metabolite between men and women was creatinine, which was found to be at systematically higher concentrations in male urine. Conversely, females excreted

dx.doi.org/10.1021/pr4000152 | J. Proteome Res. 2013, 12, 3166−3180

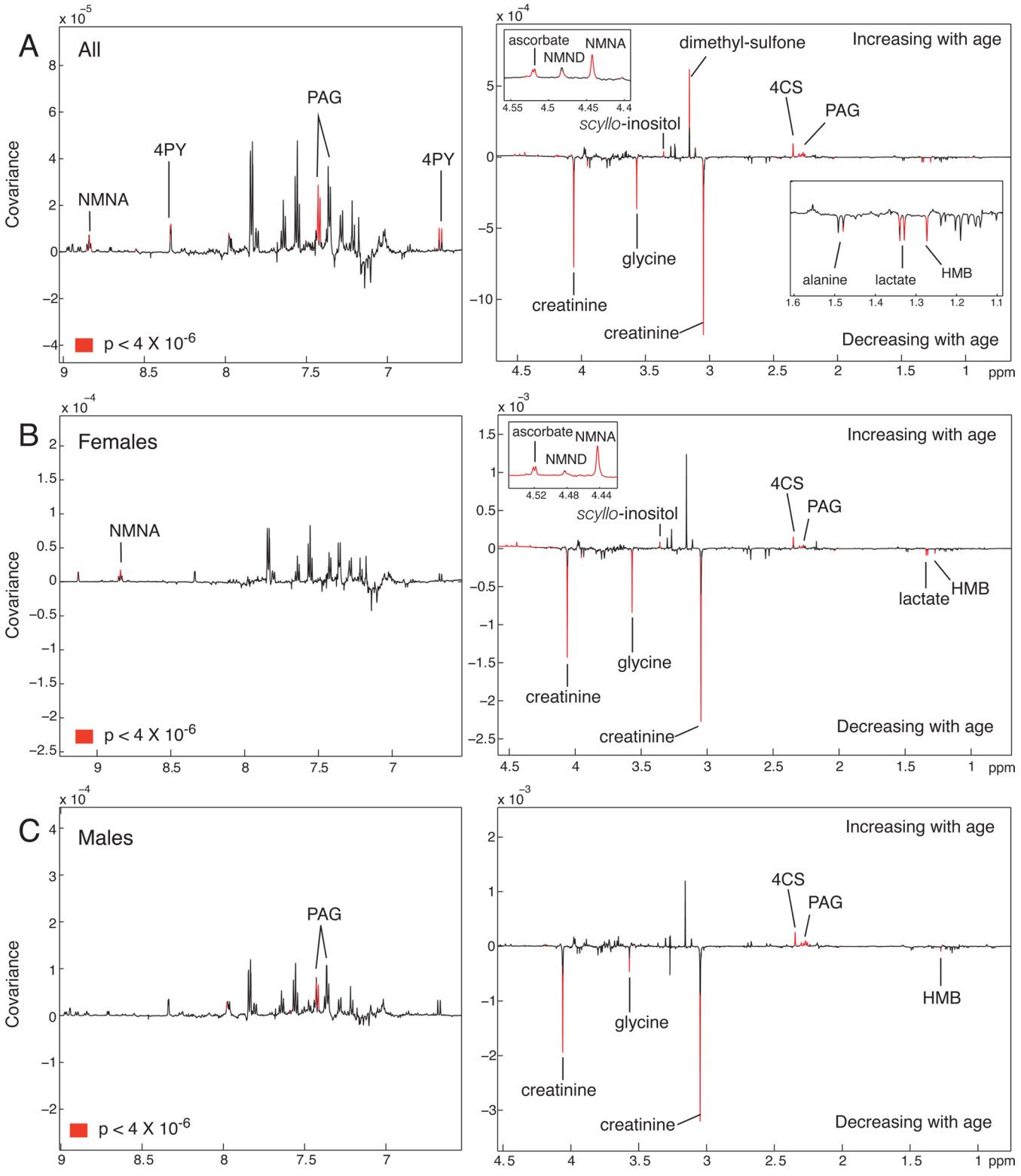


Figure 5. Age-related variation in MIDUS urinary metabolic profiles using linear regression. Covariance plots derived from linear regression analysis for (A) all MIDUS participants and stratified by sex ((B) females and (C) males). Covariance plots are colored by significance ($p < 4 \times 10^{-6}$). 4PY, *N*-methyl-4-pyridone-3-carboxamide; NMNA, *N*-methyl nicotinic acid; NMND, *N*-methyl nicotinamide; HMB, *β*-hydroxy-*β*-methylbutyrate; PAG, phenylacetylglutamine; 4CS, 4-cresyl-sulfate.

greater amounts of creatine and citrate than males. This difference is illustrated in the linear regression plot (Figure 3A). Men were also found to excrete greater amounts of a methylmalonate. Similar findings were noted in the OPLS-DA analysis between sexes in the MIDUS II specimen set (Supporting Information Figure S2B) with a $Q^2Y = 0.207$ for a 1 aligned and 1 orthogonal component model. As with the SEBAS cohort, men had higher urinary excretion of creatinine and methylmalonate and lower citrate and creatine than women. Additional sex-related differences in the US specimen

dx.doi.org/10.1021/pr4000152 | J. Proteome Res. 2013, 12, 3166−3180

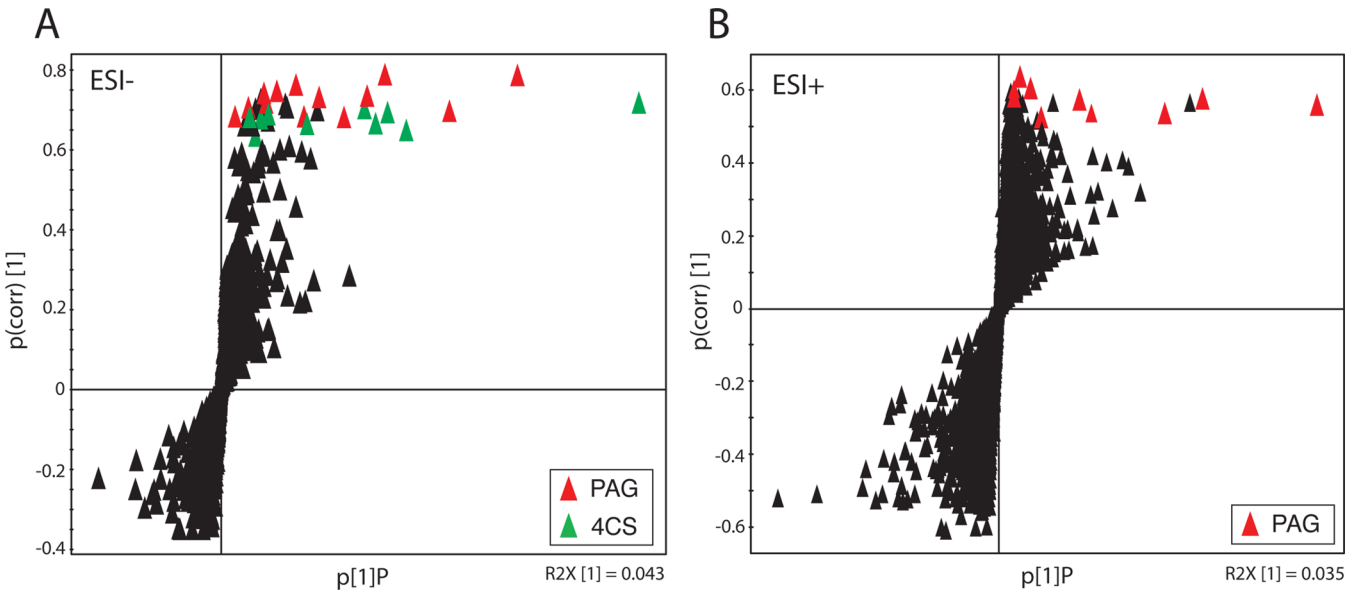


Figure 6. S-plots of the OPLS models identifying UPLC−MS derived-metabolic features associated with aging for (A) SEBAS and (B) MIDUS cohorts.

set included higher taurine in male urine and higher glycine and 4-cresyl sulfate concentrations in female urine (Figure 3B). The urinary concentration of creatinine was calculated from the CH_2 signal of creatinine at δ 4.06. The mean creatinine concentrations for men and women in the SEBAS population were 10.25 ± 5.83 mM and 7.26 ± 4.72 mM respectively and the values for the MIDUS participants were 11.07 ± 6.68 mM (men) and 10.55 ± 6.55 mM (women).

When the data sets were adjusted to align the age range for the SEBAS and MIDUS studies, some of the metabolites identified as being significantly different between men and women in the MIDUS II cohort were not sustained and the urinary metabolites differentiating between men and women were more similar for the two populations (Supporting Information Figure S3). Higher urinary concentrations of citrate and creatine were present in female urine from both SEBAS and MIDUS participants, whereas males excreted higher creatinine and methylmalonate. Additionally, for the MIDUS study, taurine was present in higher concentration in urine specimens collected from men, even after adjustment for age range.

Age-related Differences in Urinary Metabolic Phenotypes

PLS models were calculated for the SEBAS and MIDUS specimen sets independently for both the complete data sets and the age-restricted data sets as summarized in Supporting Information Table S1. Both the univariate linear regression and the OPLS regression models indicated that there was significant variation in the NMR metabolite profiles with age (summarized in Table 2). Mean signal intensities for each metabolite significantly associated with age have been calculated for youngest and oldest participants (n = 100) in the SEBAS and MIDUS studies and are provided in Supporting Information Table S2. Overall, for the SEBAS study, age was directly correlated with excretion of phenylacetylglutamine (PAG), 4-cresyl sulfate (4CS) and glutamate and was inversely correlated with excretion of creatine, β-hydroxy-β-methylbutyrate (HMB) and guanidinoacetate (GAA) (Figure 4). Further models were calculated for this data set after stratification by sex. For both sexes, the gut-microbially derived metabolites, PAG and 4CS, were directly correlated with age. There were also a few differences between the sex-specific models: HMB was inversely correlated with age for males, whereas females showed a similar trend in HMB with age but the age-related variation in urinary concentration was not significant. Women excreted lower amounts of creatine with age.

Similar patterns were observed in the MIDUS study, with PAG and 4CS excretion increasing and creatine, creatinine and HMB excretion inversely correlated with age (Figure 5A). In addition, *scyllo*-inositol, dimethyl-sulfone, *N*-methylnicotinamide (NMDA), *N*-methylnicotinic acid (NMNA), *N*-methyl-4-pyridone-3-carboxamide (4PY) and ascorbate excretion were also directly associated with age. Lower amounts of several amino acids (alanine, glycine and lactate) were excreted with increasing age. When stratified by sex, the females excreted higher PAG, 4CS, *scyllo*-inositol, NMNA, NMND and ascorbate as they aged and lower levels of HMB, creatine, creatinine, lactate and glycine (Figure 5B). Fewer metabolites were correlated with age in the male participants (Figure 5C), with PAG and 4CS positively correlated with age while HMB, creatinine and glycine were negatively correlated with age.

When the data sets were restricted to the same age range in both the MIDUS and SEBAS populations (Supporting Information Figures S4 and S5), the metabolites related to age in the complete data set persisted for SEBAS. For the MIDUS participants, the narrower age range reduced the sample size (females n = 365; males n = 297) and thus the predictive strength of the models. When male and female participants were considered together, PAG and 4CS were positively correlated with aging. In males, the higher concentration of urinary PAG was the metabolic feature most strongly associated with age. The analyses of urine from only MIDUS females yielded a model with poor predictive strength (Q^2Y = 0.008); the results from this linear regression are not shown in Supporting Information Figure S5.

UPLC−MS data indicated that the most discriminatory metabolite for both populations was PAG (Figure 6), followed by 4CS in the SEBAS population, confirming the results generated via NMR. These UPLC−MS metabolite findings

dx.doi.org/10.1021/pr4000152 | J. Proteome Res. 2013, 12, 3166−3180

were identified by comparison with authentic standards. For SEBAS, PAG was discriminatory in both the negative (p(corr) range 0.68−0.79) and positive (p(corr) range 0.72−0.82) ESI mode profiles with a mean coefficient of variation of 13 ± 2.8% and 15.5 ± 4.9%, respectively. For MIDUS, the CV values of PAG were similar (16.1 ± 6.3%) in ESI+, but as noted earlier, the ESI− data were of insufficient quality. 4CS was a discriminatory metabolite in urine samples of the SEBAS population analyzed in ESI− with a mean coefficient of variation of 19.1 ± 7.0%. The S-plots for the OPLS models constructed from the SEBAS (ESI−) and MIDUS (ESI+) UPLC−MS data are provided in Figure 6.

DISCUSSION

Human metabolism is influenced by a wide variety of genetic and environmental factors, giving rise to extensive variation in the composition of biological tissues and fluids. Understanding the nature of this variation both between individuals and across populations is critical to attributing systematic changes in metabolism to physiological processes or disease and remains a challenging aspect of biomarker research. In this study, we characterized metabolic signatures associated with sex and age in representative national populations from Taiwan (SEBAS) and the USA (MIDUS). A combination of NMR spectroscopy and UPLC−MS analysis was used to probe similarities and differences in urine specimens obtained from a large number of middle-aged and older participants. The most notable source of variation associated with age in both populations was attributed to metabolites derived from gut microbial transformation of aromatic amino acids, specifically PAG and 4CS.

Global Sources of Metabolic Variation

Major sources of variation within each data set were found to be similar and comprised a mixture of endogenous, dietary, gut-microbial and xenobiotic signatures from human metabolite profiles. The general overview of the metabolic profiles provided by principal components analysis identified metabolites of dietary origin contributing to variation in the metabolic profiles and differing across the two samples. In SEBAS, the excretion of methylamines was a strong source of variation while hippurate concentrations were highly variable in the MIDUS II data set. Urinary dimethylamine (DMA) and trimethylamine (TMA) are predominantly gut microbial products of dietary choline metabolism.[23] The high concentration of TMA in fish is responsible for the characteristic odor. The significant findings in the Taiwanese data may be indicative of greater variation in fish/choline consumption across this cohort, although TMAO is also known to be a component of foods that are high in phytoestrogens such as soy and miso. This interpretation is reasonable given that no dietary restriction was required prior to specimen collection and that fish, seafood and soy are major components of the Taiwanese diet. Alternatively, choline biotransformation capacity encoded in the microbiome may vary widely in this sample. TMAO is a hepatic oxidation product of dietary amines, specifically TMA, and was noted to vary across SEBAS participants in a similar manner to its metabolic precursor. Recent work has demonstrated an association between gut microbial-produced TMA and TMAO and cardiovascular disease risk in humans,[24] where TMAO was demonstrated to be pro-atherogenic.

A further indication that gut microbial capacity may differ between the American and Taiwanese populations is the difference in the urinary variation and concentration of hippurate, a gut microbial−mammalian cometabolite, which is formed from glycine conjugation of dietary or microbially produced benzoic acid in the liver mitochondria. Hippurate was found in higher concentrations in the MIDUS cohort than the SEBAS cohort (SEBAS mean hippurate 1.4 ± 1.51 mM; MIDUS 2.15 ± 1.71 mM) and was also responsible for a large part of the variation in the PCA scores plot in the MIDUS but not the SEBAS data set (Figures 1, 2). Typical urinary concentrations of hippurate in a predominantly Caucasian population have been reported as 1.83 ± 1.24 mM.[25] Differences in the excretion pattern of hippurate and methylamines may simply reflect dietary variation—for example in the consumption of fish, coffee and other sources of benzoic acid (a precursor of hippurate)—or may partially relate to population differences in the gut microbiota and/or their activities. It has been shown that gut microbial transformations can be influenced or entrained by diet. For example, certain porphyranases from marine Bacteroidetes have been acquired by the gut microbiota of Japanese populations where sushi is a stable part of the diet but are absent from the metagenome of Americans.[26]

From the principal components analysis, creatinine was identified as the metabolite with the greatest variation across both the Taiwanese and US samples. Creatinine is known to differ between sexes, with age, with meat consumption, and to be proportional to muscle mass. It is expected, therefore, that creatinine might vary widely across these two large-scale sets of specimens. Urinary creatinine was also strongly influenced by sex, with higher concentrations found in men, in keeping with the known influence of muscle mass.

Other metabolites that exhibited a high degree of variation across the two data sets included xenobiotics such as acetaminophen metabolites, namely acetaminophen-glucuronide and acetaminophen-sulfate, an interesting reflection of prevailing medical practice and medication use across two nations. Acetaminophen metabolites (predominantly glucuronide and sulfate) emerged as strong contributors to the coefficients of the first principal component of the MIDUS PCA model and the second principal component of the SEBAS model.

Sex-dependent Metabolites in the SEBAS and MIDUS Samples

Variation attributable to sex was a major component of both the SEBAS and the MIDUS data sets. On the whole the sex-dependent urinary signature was similar for both data sets. As expected, differences in urinary creatinine proved to be the strongest discriminator with higher levels of urinary creatinine excretion in men, reflecting their greater muscle mass. Creatinine has also been shown to be directly correlated with body weight.[27] Metabolic profiling studies in Swiss (n = 84 women and 66 men),[28] American (n = 30 women and 30 men)[29] and Greek (n = 61 women and 61 men)[30] populations using ^{1}H NMR spectroscopy and multivariate statistics have also reported that creatinine dominates the models. Metabolic profiling studies in rats and mice have also reported higher urinary creatinine concentrations in male animals.[31]

Urinary citrate levels were higher in women than men, in both the SEBAS and MIDUS samples, a finding also reported in prior studies of Swiss, American and Greek populations.[28−30] Higher urinary citrate levels in females have also been found in animal studies, and it is known that urinary citrate excretion increases during pregnancy along with 2-oxoglutarate and

lactate.[32] Urinary citrate excretion in women rises during ovulation and following the administration of estrogens.[33] A comparison of the age-restricted samples suggested that the citrate variation between men and women was stronger in SEBAS ($r = 0.24$; $p = 1.21 \times 10^{-12}$) than in MIDUS ($r = 0.19$; $p = 5.99 \times 10^{-7}$). The higher levels of urinary citrate in women is thought to account for their lower risk of kidney stone formation due to citrate's inhibitory influence on calcium salt crystallization. Conversely, hypocitraturia is an important risk factor for kidney stone formation.[34]

Amino acid excretion was found to differ between sexes in the MIDUS sample only. Greater taurine excretion was observed in male participants while higher glycine excretion was noted in females. Taurine is an amino acid associated with meat intake and could thus reflect dietary preferences for meat consumption,[35] but increased excretion is also a consequence of increased tissue catabolism and protein turnover, which is known to be higher in men. Glycine is required for the biosynthesis of creatine, which was also observed to be greater in females than males. The higher excretion of glycine may therefore reflect a greater requirement for creatine synthesis in these females.

Methylmalonate (MMA) was present in greater amounts in male than in female urine. This sex effect was consistent across both the Taiwanese and US samples. This malonic acid derivative is a precursor for succinyl-CoA and its synthesis requires the cofactor, cobalamin (vitamin B_{12}). Hence, urinary MMA is known to be elevated in cobalamin-deficient individuals. Cobalamin deficiency is most common in elderly white males[36] and has been associated with cognitive impairment, anemia and peripheral neuropathy.[37]

Characterization of Age-associated Metabolites in the SEBAS and MIDUS Samples

Age-related variation was apparent in both data sets. Two notable metabolites—phenylacetylglutamine (PAG) and 4-cresyl sulfate (4CS)—were positively correlated with age, even when the samples were stratified by sex. Another variation that was consistent across both samples was lower excretion of β-hydroxy-β-methylbutyrate (HMB) and creatine in older participants.

Associations with age that were unique to the SEBAS population included a positive relationship between urinary glutamate and age and an inverse relationship with guanidinoacetic acid (GAA). For MIDUS participants, ascorbate, *N*-methylnicotinamide (NMND), *N*-methylnicotinic acid (NMNA), *N*-methyl-4-pyridone-3-carboxamide (4PY), dimethyl-sulfone and *scyllo*-inositol were directly associated with age, while creatinine, lactate, alanine and glycine were inversely correlated with age.

Through this molecular epidemiology approach we have identified potential metabolic windows into multiple age-related processes and diseases. These have great potential for understanding the biochemical basis of disease processes, early diagnostics and health implications of such diseases. Specifically, the results are relevant to the biochemical events associated with sarcopenia, neurological dysfunction and the susceptibility to gastrointestinal infection.

Creatinine, creatine and HMB are likely to be associated with muscle turnover, which declines with age. As discussed with respect to sex differences in creatinine excretion, creatinine is an index of muscle mass[27] and aging is associated with progressive loss of muscle performance and lean mass.[38] In a metabolic profiling study of aging in Labrador retriever dogs, the level of urinary creatinine rose during development through young adulthood, reached a maximum at 5−9 years old and then declined in later life.[39] Differences in creatinine concentration with age can also arise from the age-dependent decrease in renal plasma flow and glomerular filtration rate.[40] However, since the proximal tubules are responsible for the excretion of 10% of creatinine then although reduced glomerular filtration rate may contribute to the association between age and declining creatinine, it is unlikely to be the main factor influencing this event. Muscle holds a vital role in whole-body protein metabolism serving as a repository for protein and amino acids and maintaining systemic protein synthesis. Reasons for the decline in muscle mass with age include reduced exercise, poor nutrition and loss of muscle integrity. However, a definitive mechanism for muscle loss with age has not yet been established. Maintenance of muscle mass can protect against various pathologies and diseases. Age-related muscle mass atrophy (sarcopenia) can have adverse effects on protein metabolism, immune function, organ function and wound healing.[41] Proposed reasons for sarcopenia stem from a host of intrinsic and extrinsic factors including decreased hormonal activity.[42] The inverse association between HMB and age is also consistent with the progressive loss of muscle mass with age and has previously been reported as characteristic of differences between young (19−40 years) and old (41−69) in a metabolic profiling study in a small cohort of Americans.[29] HMB is a metabolite of the amino acid leucine and has a protective effect on muscle loss. It can serve as a precursor for cholesterol synthesis in muscle tissue, which can then have an important role in strengthening the cellular membrane of muscle cells. Furthermore, HMB can attenuate protein degradation and up-regulate protein synthesis in muscle tissue. Research has shown that supplementing the elderly with HMB can decrease muscle damage and increase lean body mass.[43]

Elevations in the excretion of several metabolites in the nicotinic acid pathway—*N*-methylnicotinic acid (trigonelline or NMNA), *N*-methylnicotinamide (NMND) and *N*-methyl-4-pyridone-3-carboxamide (4PY)—were positively associated with age in the American cohort. This type of metabolic dysregulation may be associated with age-related neurodegenerative conditions and cognitive dysfunction associated with aging e.g. Parkinson's and Alzheimer's disease.[6] Lower urinary 4-PY concentrations have been found in stressed rats compared with controls, and those exhibiting fatigue have perturbed nicotinate and nicotinamide metabolism.[44] Increased NMND excretion has also been observed in individuals with Parkinson's disease[45,46] and has been implicated as a mechanism mediating the death of dopamine-generating cells.[47] Similarly, brain concentrations of inositol metabolites have been linked to neurodegenerative diseases, specifically Alzheimer's dementia, and are present in greater amounts in elderly than in young individuals,[48] suggesting that the regulatory integrity for maintaining intracellular inositol concentrations may weaken with age.

Indices of Age-associated Variation in the Gut Microbiome

Mammals are now considered to be "superorganisms" or "metaorganisms" whose processes represent the sum of both genomic and microbiomic contributions. It is reasonable, therefore, to consider how aging affects the symbiotic relationship between the host and resident microbiota. Such age-associated changes are likely to be reciprocal in nature with

dx.doi.org/10.1021/pr4000152 | J. Proteome Res. 2013, 12, 3166−3180

microbial modulations being both a cause and consequence of structural and biochemical changes in the gastrointestinal tract, immunosenescence and alterations in food consumption caused by changes in appetite, taste and digestion. In addition, host factors, including reduced physical activity, oropharangeal dysphagia and changes in gut motility and immune competence in the elderly can all impact on health and the microbiota.[49] Conditions such as constipation and slow gut transit times are also more prevalent in the elderly and may lead to increased usage of various medications for chronic symptoms.[50] Elderly people are more likely than younger people to be the recipients of drug therapy of many classes, including ones that affect the gut microbiome (e.g., elderly, defined as >65 years, comprise approximately 13% of the U.S.A. population, but are the recipients of >40% of all prescription drugs[51]). Laxatives, antibiotics, and calcium channel blockers commonly lead to side-effects such as diarrhea, malabsorption and constipation.[52]

PAG and 4CS showed the strongest association with age for both populations with a correlation coefficient (r) of 0.32 ($p = 1.2 \times 10^{-21}$) and 0.32 ($p = 1.53 \times 10^{-21}$), respectively, for SEBAS and 0.29 ($p = 6.55 \times 10^{-23}$) and 0.23 ($p = 9.83 \times 10^{-16}$) for MIDUS (Figures 4 and 5). PAG and 4CS are formed from protein putrefaction of phenylalanine and tyrosine by the gut microbiota. Phenylalanine is converted to phenylacetate in the colon and subsequently conjugated with glutamine in the liver and the gut mucosa,[53] whereas 4CS is a product of microbial tyrosine breakdown *via* hydroxyphenylacetate to *4-cresyl*, followed by conjugation with sulfate.[54] Age-related variations were also observed in the bacterial fermentation product, lactic acid, being negatively associated with aging in the American sample.

The marked age-associated alteration of PAG and 4CS concentrations are consistent with known shifts in the composition of the microbiome, including increased representation from enterobacteria and decreasing proportions of anaerobes and Bifidobacteria.[55] The ratio of Firmicutes to Bacteroidetes has also been found to be lower in the elderly.[56] Decreases in anaerobes and *Bifidobacterium* spp. and increases in enterobacteria may increase susceptibility to gastrointestinal infections, and changes in the composition of gut microbiota have been implicated in many diseases such as Irritable Bowel Syndrome (IBS), Ulcerative Colitis (UC) and Crohn's disease (CD).[57] Moreover nosocomial infections such as *Clostridium difficile* are known to have greater morbidity in the elderly. The diversity of species comprising the dominant fecal microbiota increase with aging.[58] In addition to the composition changes, the interaction between the microbiota and intestinal functions likely shift with age. He et al. demonstrated that certain Bifidobacterium strains isolated from healthy adults aged 30–40 were able to bind better to the intestinal mucus than were the same bacterial strains isolated from healthy seniors (>70 years of age).[59] However, not all researchers have consistently found these age-related differences. Other studies have shown that there is a tendency for stability in the gut microbiome throughout adulthood,[60] and several studies suggest that age-related alterations in microbial composition may be dependent upon the population and geographic location.[61] Aging has been associated with an increase in enterobacteria and Clostridia in particular, while health-promoting bacteria such as the Bifidobacteria have been reported to decline in abundance and diversity of species with age.[58] Several bacteria can synthesize 4CS such as members of the Clostridia including *Clostridium difficile*.[62]

Other studies have reported associations between age and mammalian-microbial urinary cometabolites. One ^{1}H NMR-based profiling study investigating lifelong changes in the urinary metabolome of dogs under caloric restricted and nonrestricted conditions found that hippurate and 3-HPPA concentrations increased with age.[39] Urinary levels of amines, resulting from degradation of dietary choline by gut microbiota, also changed with age. This increase in gut microbial metabolites was enhanced by dietary restriction. Similar results have been shown in a study in which rats fed with chow diets were compared with rats fed with casein-rich diets.[63] Moreover, in both humans and nonhumans, clear differences in microbially derived metabolites have been shown in the urinary, fecal and plasma profiles from obese individuals with metabolites such as hippurate and PAG being associated with leaner phenotypes. Thus, it is possible that variation in the excretion of 4CS and PAG seen with age in both the SEBAS and MIDUS surveys reflect a general reduction in caloric intake by the older participants.

CONCLUSIONS

In summary, this work reinforces the great potential of applying metabolome-wide association studies to large-scale epidemiology studies. Through this application we have identified potential metabolic windows into later life diseases. These windows point to an underpinning dysregulation of the microbiota that may relate to increased susceptibility to GI infection in the elderly. Additionally some of the changes are suggestive of a decline in muscle mass. Specifically, we have shown significant age-related differences in the urinary metabolite profiles of Taiwanese and American populations, with the strongest effects being attributed to 4-cresyl sulfate and phenylacetylglutamine. These metabolite differences were significant in both males and females and revealed a marked shift in the functionality of the gut microbiome with age. In addition, the bacterial fermentation product, lactic acid, was negatively correlated with age in Americans. The age-related variation in these gut microbial metabolites may reflect increasing enterobacterial numbers and warrants further investigation to directly link metabolic profiles to fecal microbial composition. The appearance of functional aging observed in the microbiome was consistent across both national populations in spite of some cultural features.

ASSOCIATED CONTENT

Supporting Information

Supplemental figures and tables. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel: 020 7594 3220. Fax: 020 7594 3226. E-mail: elaine.holmes@imperial.ac.uk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Institute on Aging (grant numbers R01AG16790, R01AG16661, P01-AG020166); and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grant number

R24HD047879). SEBAS was funded by the Demography and Epidemiology Unit of the Behavioral and Social Research Program of the National Institute on Aging [grant numbers R01 AG16790, R01 AG16661]. The Bureau of Health Promotion (Department of Health, Taiwan) provided additional financial support for SEBAS 2000. We acknowledge the hard work and dedication of the staff at the Center for Population and Health Survey Research (BHP), who were instrumental in the design and implementation of the SEBAS and supervised all aspects of the fieldwork and data processing. The MIDUS longitudinal follow-up was supported by the National Institute on Aging [grant number P01-AG020166]. The specimen collection was also facilitated by the General Clinical Research Centers Program [grant numbers M01-RR023942 to Georgetown University; M01-RR00865 to UCLA] and by the Clinical and Translational Science Award program of the National Center for Research Resources, National Institutes of Health [grant number 1UL1RR025011 to University of Wisconsin-Madison].

REFERENCES

(1) Piper, M. D.; Selman, C.; McElwee, J. J.; Partridge, L. Separating cause from effect: how does insulin/IGF signalling control lifespan in worms, flies and mice? *J. Intern. Med.* **2008**, *263* (2), 179−91.

(2) Wijeyesekera, A.; Selman, C.; Barton, R. H.; Holmes, E.; Nicholson, J. K.; Withers, D. J. Metabotyping of long-lived mice using (1)H NMR spectroscopy. *J. Proteome Res.* **2012**, *11* (4), 2224−35.

(3) Robert, L.; Labat-Robert, J.; Robert, A. M. Genetic, epigenetic and posttranslational mechanisms of aging. *Biogerontology* **2010**, *11* (4), 387−99.

(4) Walston, J.; Hadley, E. C.; Ferrucci, L.; Guralnik, J. M.; Newman, A. B.; Studenski, S. A.; Ershler, W. B.; Harris, T.; Fried, L. P. Research agenda for frailty in older adults: toward a better understanding of physiology and etiology: summary from the American Geriatrics Society/National Institute on Aging Research Conference on Frailty in Older Adults. *J. Am. Geriatr. Soc.* **2006**, *54* (6), 991−1001.

(5) Coe, C. L.; Love, G. D.; Karasawa, M.; Kawakami, N.; Kitayama, S.; Markus, H. R.; Tracy, R. P.; Ryff, C. D. Population differences in proinflammatory biology: Japanese have healthier profiles than Americans. *Brain Behav. Immun.* **2011**, *25* (3), 494−502.

(6) Muangpaisan, W.; Mathews, A.; Hori, H.; Seidel, D. A systematic review of the worldwide prevalence and incidence of Parkinson's disease. *J. Med. Assoc. Thai.* **2011**, *94* (6), 749−55.

(7) van den Bussche, H.; Koller, D.; Kolonko, T.; Hansen, H.; Wegscheider, K.; Glaeske, G.; von Leitner, E. C.; Schafer, I.; Schon, G. Which chronic diseases and disease combinations are specific to multimorbidity in the elderly? Results of a claims data based cross-sectional study in Germany. *BMC Public Health* **2011**, *11*, 101.

(8) Regitz-Zagrosek, V.; Lehmkuhl, E.; Weickert, M. O. Gender differences in the metabolic syndrome and their role for cardiovascular disease. *Clin. Res. Cardiol.* **2006**, *95* (3), 136−47.

(9) Holmes, E.; Loo, R. L.; Stamler, J.; Bictash, M.; Yap, I. K.; Chan, Q.; Ebbels, T.; De Iorio, M.; Brown, I. J.; Veselkov, K. A.; Daviglus, M. L.; Kesteloot, H.; Ueshima, H.; Zhao, L.; Nicholson, J. K.; Elliott, P. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* **2008**, *453* (7193), 396−400.

(10) Yap, I. K.; Brown, I. J.; Chan, Q.; Wijeyesekera, A.; Garcia-Perez, I.; Bictash, M.; Loo, R. L.; Chadeau-Hyam, M.; Ebbels, T.; De Iorio, M.; Maibaum, E.; Zhao, L.; Kesteloot, H.; Daviglus, M. L.; Stamler, J.; Nicholson, J. K.; Elliott, P.; Holmes, E. Metabolome-wide association study identifies multiple biomarkers that discriminate north and south Chinese populations at differing risks of cardiovascular disease: INTERMAP study. *J. Proteome Res.* **2010**, *9* (12), 6647−54.

(11) Nicholson, J. K.; Lindon, J. C.; Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **1999**, *29* (11), 1181−9.

(12) Mishur, R. J.; Rea, S. L. Applications of mass spectrometry to metabolomics and metabonomics: detection of biomarkers of aging and of age-related diseases. *Mass Spectrom. Rev.* **2012**, *31* (1), 70−95.

(13) Houtkooper, R. H.; Argmann, C.; Houten, S. M.; Canto, C.; Jeninga, E. H.; Andreux, P. A.; Thomas, C.; Doenlen, R.; Schoonjans, K.; Auwerx, J. The metabolic footprint of aging in mice. *Sci. Rep.* **2011**, *1*, 134.

(14) Okuda, Y.; Kawai, K.; Yamashita, K. Age-related change in ketone body metabolism: diminished glucagon effect on ketogenesis in adult rats. *Endocrinology* **1987**, *120* (5), 2152−7.

(15) Takiyama, N.; Matsumoto, K. Age-and sex-related differences of serum carnitine in a Japanese population. *J. Am. Coll. Nutr.* **1998**, *17* (1), 71−4.

(16) Goldman, N.; Lin, I. F.; Weinstein, M.; Lin, Y. H. Evaluating the quality of self-reports of hypertension and diabetes. *J. Clin. Epidemiol.* **2003**, *56* (2), 148−54.

(17) Marmot, M. G.; Fuhrer, R.; Ettner, S. L.; Marks, N. F.; Bumpass, L. L.; Ryff, C. D. Contribution of psychosocial factors to socio-economic differences in health. *Milbank Q.* **1998**, *76* (3), 403−48.

(18) Beckonert, O.; Keun, H. C.; Ebbels, T. M.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protoc.* **2007**, *2* (11), 2692−703.

(19) Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K. Recursive segment-wise peak alignment of biological (1)h NMR spectra for improved metabolic biomarker recovery. *Anal. Chem.* **2009**, *81* (1), 56−66.

(20) Chadeau-Hyam, M.; Ebbels, T. M.; Brown, I. J.; Chan, Q.; Stamler, J.; Huang, C. C.; Daviglus, M. L.; Ueshima, H.; Zhao, L.; Holmes, E.; Nicholson, J. K.; Elliott, P.; De Iorio, M. Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. *J. Proteome Res.* **2010**, *9* (9), 4620−7.

(21) Want, E. J.; Wilson, I. D.; Gika, H.; Theodoridis, G.; Plumb, R. S.; Shockcor, J.; Holmes, E.; Nicholson, J. K. Global metabolic profiling procedures for urine using UPLC-MS. *Nat. Protoc.* **2010**, *5* (6), 1005−18.

(22) Veselkov, K. A.; Vingara, L. K.; Masson, P.; Robinette, S. L.; Want, E.; Li, J. V.; Barton, R. H.; Boursier-Neyret, C.; Walther, B.; Ebbels, T. M.; Pelczer, I.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Anal. Chem.* **2011**, *83* (15), 5864−72.

(23) Zhang, A. Q.; Mitchell, S. C.; Smith, R. L. Dimethylamine in human urine. *Clin. Chim. Acta* **1995**, *233* (1−2), 81−8.

(24) Wang, Z.; Klipfell, E.; Bennett, B. J.; Koeth, R.; Levison, B. S.; Dugar, B.; Feldstein, A. E.; Britt, E. B.; Fu, X.; Chung, Y. M.; Wu, Y.; Schauer, P.; Smith, J. D.; Allayee, H.; Tang, W. H.; DiDonato, J. A.; Lusis, A. J.; Hazen, S. L. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **2011**, *472* (7341), 57−63.

(25) Saude, E. J.; Adamko, D.; Rowe, B. H.; Marrie, T.; Sykes, B. D. Variation of metabolites in normal human urine. *Metabolomics* **2007**, *3*, 439−51.

(26) Hehemann, J. H.; Correc, G.; Barbeyron, T.; Helbert, W.; Czjzek, M.; Michel, G. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **2010**, *464* (7290), 908−12.

(27) Davies, K. M.; Heaney, R. P.; Rafferty, K. Decline in muscle mass with age in women: a longitudinal study using an indirect measure. *Metabolism* **2002**, *51* (7), 935−9.

(28) Kochhar, S.; Jacobs, D. M.; Ramadan, Z.; Berruex, F.; Fuerholz, A.; Fay, L. B. Probing gender-specific metabolism differences in humans by nuclear magnetic resonance-based metabonomics. *Anal. Biochem.* **2006**, *352* (2), 274−81.

(29) Slupsky, C. M.; Rankin, K. N.; Wagner, J.; Fu, H.; Chang, D.; Weljie, A. M.; Saude, E. J.; Lix, B.; Adamko, D. J.; Shah, S.; Greiner, R.; Sykes, B. D.; Marrie, T. J. Investigations of the effects of gender,

diurnal variation, and age in human urinary metabolomic profiles. *Anal. Chem.* **2007**, *79* (18), 6995−7004.
(30) Psihogios, N. G.; Gazi, I. F.; Elisaf, M. S.; Seferiadis, K. I.; Bairaktari, E. T. Gender-related and age-related urinalysis of healthy subjects by NMR-based metabonomics. *NMR Biomed* **2008**, *21* (3), 195−207.
(31) Stanley, E. G.; Bailey, N. J.; Bollard, M. E.; Haselden, J. N.; Waterfield, C. J.; Holmes, E.; Nicholson, J. K. Sexual dimorphism in urinary metabolite profiles of Han Wistar rats revealed by nuclear-magnetic-resonance-based metabonomics. *Anal. Biochem.* **2005**, *343* (2), 195−202.
(32) Yasuda, M.; Tsunoda, S.; Nagasawa, H. Comparison of urinary component levels in 4 strains of mice with different physiological characteristics. *In Vivo* **1997**, *11* (2), 109−13.
(33) Hammar, M. L.; Berg, G. E.; Larsson, L.; Tiselius, H. G.; Varenhorst, E. Endocrine changes and urinary citrate excretion. *Scand J Urol Nephrol* **1987**, *21* (1), 51−3.
(34) Welshman, S. G.; McGeown, M. G. Urinary citrate excretion in stone-formers and normal controls. *Br J Urol* **1976**, *48* (1), 7−11.
(35) Laidlaw, S. A.; Shultz, T. D.; Cecchino, J. T.; Kopple, J. D. Plasma and urine taurine levels in vegans. *Am. J. Clin. Nutr.* **1988**, *47* (4), 660−3.
(36) Carmel, R.; Green, R.; Jacobsen, D. W.; Rasmussen, K.; Florea, M.; Azen, C. Serum cobalamin, homocysteine, and methylmalonic acid concentrations in a multiethnic elderly population: ethnic and sex differences in cobalamin and metabolite abnormalities. *Am. J. Clin. Nutr.* **1999**, *70* (5), 904−10.
(37) Clarke, R.; Refsum, H.; Birks, J.; Evans, J. G.; Johnston, C.; Sherliker, P.; Ueland, P. M.; Schneede, J.; McPartlin, J.; Nexo, E.; Scott, J. M. Screening for vitamin B-12 and folate deficiency in older persons. *Am. J. Clin. Nutr.* **2003**, *77* (5), 1241−7.
(38) Hurley, B. F. Age, gender, and muscular strength. *J. Gerontol. A: Biol. Sci. Med. Sci.* **1995**, *50*, 41−4.
(39) Wang, Y.; Lawler, D.; Larson, B.; Ramadan, Z.; Kochhar, S.; Holmes, E.; Nicholson, J. K. Metabonomic investigations of aging and caloric restriction in a life-long dog study. *J. Proteome Res.* **2007**, *6* (5), 1846−54.
(40) Perrone, R. D.; Madias, N. E.; Levey, A. S. Serum creatinine as an index of renal function: new insights into old concepts. *Clin. Chem.* **1992**, *38* (10), 1933−53.
(41) Wolfe, R. R. The underappreciated role of muscle in health and disease. *Am. J. Clin. Nutr.* **2006**, *84* (3), 475−82.
(42) Cannon, J. G. Intrinsic and extrinsic factors in muscle aging. *Ann. N.Y. Acad. Sci.* **1998**, *854*, 72−7.
(43) Vukovich, M. D.; Stubbs, N. B.; Bohlken, R. M. Body composition in 70-year-old adults responds to dietary beta-hydroxy-beta-methylbutyrate similarly to that of young adults. *J. Nutr.* **2001**, *131* (7), 2049−52.
(44) Zhang, F.; Jia, Z.; Gao, P.; Kong, H.; Li, X.; Lu, X.; Wu, Y.; Xu, G. Metabonomics study of urine and plasma in depression and excess fatigue rats by ultra fast liquid chromatography coupled with ion trap-time of flight mass spectrometry. *Mol. Biosyst.* **2010**, *6* (5), 852−61.
(45) Willets, J. M.; Lunec, J.; Williams, A. C.; Griffiths, H. R. Neurotoxicity of nicotinamide derivatives: their role in the aetiology of Parkinson's disease. *Biochem. Soc. Trans.* **1993**, *21* (3), 299S.
(46) Williams, A.; Sturman, S.; Steventon, G.; Waring, R. Metabolic biomarkers of Parkinson's disease. *Acta Neurol. Scand. Suppl.* **1991**, *136*, 19−23.
(47) Fukushima, T.; Kaetsu, A.; Lim, H.; Moriyama, M. Possible role of 1-methylnicotinamide in the pathogenesis of Parkinson's disease. *Exp. Toxicol. Pathol.* **2002**, *53* (6), 469−73.
(48) Kaiser, L. G.; Schuff, N.; Cashdollar, N.; Weiner, M. W. Scyllo-inositol in normal aging human brain: 1H magnetic resonance spectroscopy study at 4 T. *NMR Biomed.* **2005**, *18* (1), 51−5.
(49) Dean, M.; Raats, M. M.; Grunert, K. G.; Lumbers, M. Factors influencing eating a varied diet in old age. *Public Health Nutr.* **2009**, *12* (12), 2421−7.
(50) Wald, A. Constipation in elderly patients. Pathogenesis and management. *Drugs Aging* **1993**, *3* (3), 220−31.
(51) Cho, S.; Lau, S. W.; Tandon, V.; Kumi, K.; Pfuma, E.; Abernethy, D. R. Geriatric drug evaluation: where are we now and where should we be in the future? *Arch. Intern. Med.* **2011**, *171* (10), 937−40.
(52) Triantafyllou, K.; Vlachogiannakos, J.; Ladas, S. D. Gastrointestinal and liver side effects of drugs in elderly patients. *Best Pract. Res. Clin. Gastroenterol.* **2010**, *24* (2), 203−15.
(53) Ramakrishna, B. S.; Gee, D.; Weiss, A.; Pannall, P.; Roberts-Thomson, I. C.; Roediger, W. E. Estimation of phenolic conjugation by colonic mucosa. *J. Clin. Pathol.* **1989**, *42* (6), 620−3.
(54) Ramakrishna, B. S.; Roberts-Thomson, I. C.; Pannall, P. R.; Roediger, W. E. Impaired sulphation of phenol by the colonic mucosa in quiescent and active ulcerative colitis. *Gut* **1991**, *32* (1), 46−9.
(55) Hebuterne, X. Gut changes attributed to ageing: effects on intestinal microflora. *Curr. Opin. Clin. Nutr. Metab. Care* **2003**, *6* (1), 49−54.
(56) Mariat, D.; Firmesse, O.; Levenez, F.; Guimaraes, V.; Sokol, H.; Dore, J.; Corthier, G.; Furet, J. P. The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiol.* **2009**, *9*, 123.
(57) Partridge, L.; Thornton, J.; Bates, G. The new science of ageing. *Philos. Trans. R. Soc. Lond., B: Biol. Sci.* **2011**, *366* (1561), 6−8.
(58) Hopkins, M. J.; Sharp, R.; Macfarlane, G. T. Variation in human intestinal microbiota with age. *Dig. Liver Dis.* **2002**, *34* (Suppl 2), S12−8.
(59) He, F.; Ouwehand, A. C.; Isolauri, E.; Hosoda, M.; Benno, Y.; Salminen, S. Differences in composition and mucosal adhesion of bifidobacteria isolated from healthy adults and healthy seniors. *Curr. Microbiol.* **2001**, *43* (5), 351−4.
(60) Tiihonen, K.; Ouwehand, A. C.; Rautonen, N. Human intestinal microbiota and healthy ageing. *Ageing Res. Rev.* **2010**, *9* (2), 107−16.
(61) O'Sullivan, O.; Coakley, M.; Lakshminarayanan, B.; Claesson, M. J.; Stanton, C.; O'Toole, P. W.; Ross, R. P. Correlation of rRNA gene amplicon pyrosequencing and bacterial culture for microbial compositional analysis of faecal samples from elderly Irish subjects. *J. Appl. Microbiol.* **2011**, *111* (2), 467−73.
(62) Elsden, S. R.; Hilton, M. G.; Waller, J. M. The end products of the metabolism of aromatic amino acids by Clostridia. *Arch. Microbiol.* **1976**, *107* (3), 283−8.
(63) Bell, J. D.; Sadler, P. J.; Morris, V. C.; Levander, O. A. Effect of aging and diet on proton NMR spectra of rat urine. *Magn. Reson. Med.* **1991**, *17* (2), 414−22.