

METABOLIC PROFILING FOR BIOMARKER DISCOVERY IN BIOCHEMICAL GENETICS

STEVEN LAWRENCE ROBINETTE

A thesis presented for the degree of
Doctor of Philosophy of Imperial College London

Department of Surgery and Cancer
Imperial College London

2013

All work presented here is the work of the author
unless otherwise appropriately cited or noted herein

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work

Abstract

Functional characterization of the phenotypic consequences of genetic variants increasingly constitutes the rate-limiting step in the study of biochemical genetics. This thesis presents the application of metabolic profiling technologies, including Nuclear Magnetic Resonance (NMR) spectroscopy and Mass Spectrometry (MS), to identify metabolic perturbations resulting from inborn errors of metabolism, congenital diseases affecting the kidney, and animal models of both genetic mutations and renal pathology. In the case of newborn screening for enzyme deficiencies, two mass spectrometric assays, traditional tandem mass spectrometry (MS/MS) with multiple reaction monitoring and direct injection nanospray high resolution mass spectrometry (ns-HR-MS), were applied to profile dried blood spot (DBS) samples of over 6,000 newborns and identify the metabolic perturbations resulting from 24 congenital disorders of metabolism.

To study genetic mutations with more complex phenotypic consequences affecting the kidney, urinary metabolic profiles were evaluated for four congenital kidney diseases. The natural history of cystinosis, showing changes in the urinary profile of cystinosis patients over time and with age, glomerular filtration rate, drug therapy, and transplantation, and a comparison of urinary perturbations seen in these human Mendelian disorders to the perturbations induced by region specific nephrotoxins characterize the urinary metabolic associations with these renal phenotypes. Finally, new computational approaches for analyzing metabolic profiling data are presented and evaluated, including for improving biomarker identification with two-dimensional NMR and increasing metabolomic coverage with high-resolution mass spectrometry. Integrating metabolic and genetic diagnostics should enhance understanding of the relationship between genes and health and mechanisms by which genetic variations manifest as inherited disease.

Contents

ABSTRACT	4
ACKNOWLEDGEMENTS	9
CONTRIBUTIONS OF OTHERS	10
ABBREVIATIONS	11
NOTATIONS	13
LIST OF FIGURES	14
LIST OF TABLES	17
LIST OF PUBLISHED PAPERS	18
1. INTRODUCTION AND BACKGROUND	20
1.1. Background.....	20
1.1.1. Characterizing the role of rare variants in human disease.....	21
1.1.2. Identifying genetic and metabolic covariation	22
1.1.3. Mendelian disease diagnostics: genotypes or metabotypes?	24
1.2. Scope of thesis	25
2. THEORY AND METHODS	28
2.1. Data acquisition – analytical chemistry of biological fluids	28
2.1.1. Nuclear Magnetic Resonance (NMR) Spectroscopy	28
2.1.1.1. Principles of NMR Spectroscopy.....	29
2.1.1.2. NMR Pulse Sequences	33
2.1.1.3. Sample preparation and acquisition considerations	36
2.1.2. Mass Spectrometry (MS) in Metabolomics.....	36
2.2. Dataset preparatory analysis.....	38
2.2.1. Data definition – resolving positional variation	39
2.2.2. Normalization.....	40

2.2.3. Centering and scaling	41
2.3. Statistical data analysis	43
2.3.1. Methods for unsupervised data analysis	44
2.3.2. Methods for supervised data analysis	47
2.3.2.1. Univariate tests	47
2.3.2.2. Multivariate supervised methods	48
2.3.3. Evaluating supervised models	49
2.3.3.1. Evaluating feature significance	50
2.3.3.2. Evaluating classification performance	50
2.4. Compound identification and verification	52
3. 2D NMR METABOLIC PROFILING IN BIOCHEMICAL GENETICS	55
3.1. Introduction and aims	55
3.2. Results – Method development	55
3.2.1. Automated crosspeak identification and integration in 2D NMR	56
3.2.2. Hierarchical Alignment of Two-Dimensional spectra (HATS)	57
3.2.3. Chemometrics applied to 2D NMR	58
3.2.3.1. Unfolding and refolding of full resolution data	61
3.2.3.2. Mapping integrated bins – multivariate Differential Analysis NMR Spectroscopy (mvaDANS)	62
3.3. Data sets	64
3.3.1. Nematode species differentiation: <i>P. pacificus</i> and <i>P. redivivus</i>	64
3.3.2. Animal model of peroxisomal disorder: <i>daf-22 C. elegans</i>	65
3.3.3. Genotype-metabotype association in a single pedigree: Autosomal Dominant Fanconi Synrome 1 (ADFS1)	66
3.4. Results and Discussion	67
3.4.1. Improved information recovery from 2D NMR spectra	67
3.4.2. Identification of β -oxidation intermediates in a <i>C. elegans</i> model of peroxisomal disorders	71
3.4.3. Identification of renal tubular transporter and TCA cycle substrates in Autosomal Dominant Fanconi Syndrome 1 probands	76
3.5. Conclusions	80

4. METABOLIC PROFILE SCREENING AND BIOMARKER IDENTIFICATION FOR INBORN ERRORS OF METABOLISM	81
4.1. Introduction and aims	81
4.2. Materials and Methods	83
4.2.1. Sample collection	83
4.2.2. Sample preparation	85
4.2.3. Mass spectrometry and data preparation	85
4.2.4. Data processing and analysis	86
4.3. Results – Characterizing the nS-HR-MS method	87
4.3.1. Isotopically labeled internal standards	87
4.3.2. Dried blood spot (DBS) metabolites	91
4.3.3. Comparison of nS-HR-MS and MS/MS measurements	94
4.4. Results and Discussion	95
4.4.1. Predictive modeling of PKU and MCADD	96
4.4.2. Biomarker identification	101
4.5. Conclusions	110
5. METABOLOMIC INVESTIGATION OF CONGENITAL DISEASES OF THE KIDNEY	111
5.1. Introduction and aims	111
5.2. Materials and Methods	111
5.2.1. Molecular biology and patients	112
5.2.1.1. Cystinosis	112
5.2.1.2. Autosomal Dominant Fanconi Syndrome 1 (ADFS1)	113
5.2.1.3. Autosomal Dominant Fanconi Syndrome 2 (ADFS2)	114
5.2.1.4. Enamel Renal Syndrome (ERS)	115
5.2.2. Samples and data acquisition	115
5.2.2.1. Comparison of congenital kidney diseases	116
5.2.2.2. Urinary metabolic natural history of cystinosis	116
5.2.2.3. Animal studies of nephrotoxin exposure	117
5.2.3. Data analysis	117
5.3. Results and Discussion	122
5.3.1. Comparison of congenital kidney diseases	122
5.3.2. Urinary metabolic natural history of cystinosis	128

5.3.3. Animal studies of nephrotoxin exposure	138
5.4. Conclusions	143
6. CONCLUSIONS AND FUTURE WORK	144
6.1. Conclusions from thesis.....	144
6.1.1. 2D NMR Metabolic profiling in biochemical genetics	144
6.1.2. Metabolic profile screening and biomarker identification for inborn errors of metabolism	145
6.1.3. Metabolomic investigation of congenital diseases of the kidney	146
6.1.4. General conclusions and scope for extending this work	146
6.2. Future directions of metabolic profiling in biochemical genetics	147
REFERENCES	151
APPENDIX A: SUPPLEMENTARY TABLES	179
APPENDIX B: PAPER I	187
APPENDIX B: PAPER II	194
APPENDIX C: PAPER III	204
APPENDIX D: PAPER IV	213
APPENDIX E: PAPER V	220
APPENDIX E: PAPER VI	228
APPENDIX E: PAPER VII	239
APPENDIX E: PAPER VIII	250
APPENDIX F: REPRODUCTION RIGHTS	260

Acknowledgements

I would like to thank my supervisor, Prof. Jeremy Nicholson, for his support, mentorship, and especially for his trust and flexibility in allowing me to find my own path to do the work I cared about. I would also like to thank my supervisor at the National Institutes of Health, Dr. William Gahl, for entertaining an unorthodox proposal from a student working in a new field and providing generous access to his lab, his patients, and his knowledge.

I owe a debt of gratitude to the many collaborators whose partnership and aid have made it possible to do this work, especially Dr. Robert Kleta, Prof. Frank Schroeder, Prof. Rafael Brüscheiler, Prof. Arthur Edison, and their research groups. Thought partnerships with Dr. Julia Dénes, Joram Posma, and Yevgeniy Izrayelit have been particularly productive.

I acknowledge the funding and support of the Marshall Aid Commemoration Commission, the National Science Foundation Graduate Research Fellowship, and the NIH-Oxford Cambridge Scholars Program.

I am happy to have had the chance to meet and get to know an assortment of amazing people it has been a pleasure to be friends with while here at Imperial. In the Department, Dr. Volker Behrends, Dr. Claire Merrifield, Dr. Silke Heinzmann, Dr. Judy Fonville, and Joram Posma come to mind. From the Marshall Scholarship, Austin McKinney, Matt Clawson, Aroop Mukharji, and Andrew Ehrich made these years both memorable and fun. To my wife, Lyra, thank you for listening to both the excitement and the frustration, and I look forward to our years together.

Steven L. Robinette

Contributions of Others

Chapter 3. Ramadan Ajredini (University of Florida) collected samples from *P. pacificus* and *P. redivivus* nematode strains which were obtained from the laboratory of Paul Sternberg (Caltech). Yevgeniy Izrayelit (Cornell) prepared and recorded NMR spectra from *N2* and *daf-22* strains of *C. elegans* and confirmed molecular structures by LC-MS. Robert Kleta (University College London) provided samples from ADFS1 patients.

Chapter 4. Julia Dénes (Imperial College London) developed the protocols for ns-HR-MS screening of newborn screening dried blood spot (DBS) and prepared and collected ns-HR-MS data for these samples. Screening centers in Giessen, Germany; Budapest, Hungary; and Copenhagen, Denmark provided DBS samples and MS/MS data.

Chapter 5. William Gahl (National Institutes of Health) provided samples and metadata from cystinosis patients from the archives prepared by Isa Bernardini (National Institutes of Health). Robert Kleta (University College London) provided samples from patients with both autosomal dominant forms of Fanconi Syndrome (ADFS1 and ADFS2) and enamel renal syndrome (ERS). His lab provided molecular characterization of these disorders. The Consortium for Metabonomic Toxicology (COMET) provided animal nephrotoxicity NMR data.

Abbreviations

ADFS1	Autosomal Dominant Fanconi Syndrome 1
ADFS2	Autosomal Dominant Fanconi Syndrome 2
ANOVA	Analysis of Variance
AUC	Area under curve [ROC]
BEA	2-bromoethanamine
COMET	Consortium for Metabonomic Toxicology
CRN	Creatinine
CYS	Cystinosis
DBS	Dried blood spot
dqfCOSY	double-quantum filtered Correlation Spectroscopy
ERS	Enamel Renal Syndrome
ESI	Electrospray Ionization
FDR	False Discovery Rate
FS	Fanconi Syndrome
FT	Fourier Transform
FT-ICR	Fourier Transform ion cyclotron resonance
GFR	Glomerular filtration rate
GWAS	Genome wide association study
HATS-PR	Hierarchical Alignment of Two-Dimensional Spectra - Pattern Recognition
HCA	Hierarchical cluster analysis
HPA	Hyperphenylalaninemia
HPLC	High Performance Liquid Chromatography
HSQC	Heteronuclear Single Quantum Correlation spectroscopy
IEM	Inborn error of metabolism
KEGG	Kyoto Encyclopedia of Genes and Genomes[kegg.jp]
LBP	L-Bifunctional Protein

LC-MS	Liquid Chromatography - Mass Spectrometry
LFC	Log-fold-change
<i>m/z</i>	mass-to-charge ratio
MCADD	Medium-chain acyl-CoA dehydrogenase deficiency
MCCV	Monte Carlo Cross-Validation
mQTL	metabolic Quantitative Trait Locus (loci)
MS	Mass Spectrometry
MS/MS	Tandem mass spectrometry
MSUD	Maple syrup urine disease
mvaDANS	multivariate Differential Analysis NMR Spectroscopy
NIPALS	Non-linear Iterative Partial Least Squares
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Effect Spectroscopy
nS-HR-MS	nanospray High Resolution Mass Spectrometry
OLS	Ordinary Least Squares
OMIM	Online Mendelian Inheritance in Man [omim.org]
PCA	Principal Component Analysis
Phe	Phenylalanine
PKU	Phenylketonuria
PLS	Partial Least Squares
PLS-DA	Partial Least Squares - Discriminant Analysis
ppm	parts-per-million
QTL	Quantitative trait locus (loci)
ROC	Receiver operating characteristic
SEM	Standard error of the mean
SNP	Single nucleotide polymorphism
STOCSY	Statistical Total Correlation Spectroscopy
TCA	Tricarboxylic acid [cycle]
TMAO	Trimethylamine-N-oxide
TOCSY	Total Correlation Spectroscopy
TSP	3-trimethylsilyl- 2,2,3,3-tetradeuteriopropionate
UPGMA	Unweighted pair group method with arithmetic mean
UPLC	Ultra-Performance Liquid Chromatography
VLCA	Very long chain ascarosides

Notations

The following notations are used throughout this thesis; where other variables or notations are introduced they are defined in the text.

β	Matrix of predictor coefficients, [p x c]
c	number of model components (scalar)
d	Distance (scalar)
δ	NMR chemical shift (ppm)
n	number of observations (scalar)
p	number of descriptor variables (scalar)
Q^2	test set goodness-of-fit (scalar)
R^2	training set goodness-of-fit (scalar)
r	Correlation coefficient (scalar)
T	Matrix of scores, [n x c]
U	Matrix of loadings, [p x c]
ω	NMR frequency domain
X	Matrix of descriptors, [n x p]
Y	Vector of responses, [n x 1]

List of Figures

1.1. Identification of genotype-phenotype associations in Mendelian disease is accelerating	21
1.2. Genes with known Mendelian phenotypes mapped onto KEGG human metabolic pathways.....	22
1.3. Three experimental designs integrating genomic and metabolomic analysis	24
2.1. Nuclear magnetic resonance spectroscopy	31
2.2. Pulse sequence diagrams for 1D and 2D NMR spectroscopy experiments.....	35
2.3. Schematic of preprocessing operations in metabolomics	39
2.4. Visualization of normalization effects	41
2.5. Visualization of scaling effects.....	43
2.6. Strategy for unknown compound identification using statistical spectroscopy and spectral databases	53
3.1. Overall strategy of Hierarchical Alignment of Two-Dimensional Spectra	60
3.2. Outline of mvaDANS operations.....	63
3.3. Alignment results for <i>P. redivivus</i> (Pre) and <i>P. pacificus</i> (Ppa) TOCSY spectra of flow-through fractions from C18 solid phase extractions	68
3.4. Comparison of normalization on full-resolution 2D NMR data on bin integrals	69
3.5. Effects of normalization on Principal Component Analysis of 2D NMR spectra	70

3.6. Human mitochondrial and peroxisomal β -oxidation with homologous <i>C. elegans</i> pathway	72
3.7. Statistical analysis of <i>C. elegans</i> mutant strain dqfCOSY spectra and confirmation of long chain ascarosides (VLCA) in <i>daf-22</i> exudates	73
3.8. Statistical analysis of urinary TOCSY spectra and identification of metabolic disruptions from a single pedigree Fanconi Syndrome.....	77
3.9. Elevated succinate and fumarate in ADFS1 urinary profiles	79
4.1. Detection frequencies and coefficients of variation for isotopically labeled internal standards by nS-HR-MS	88
4.2. Analysis of Variance (ANOVA) by batch for internal standards.....	89
4.3. Changes in batch median internal standard intensity over time	90
4.4. Detection frequencies for measured analytes in nS-HR-MS	91
4.5. Distributions of log-fold-change vs. median feature for set of one hundred control samples	92
4.6. Reduction of batch effects by median centering.....	93
4.7. Correlations between nS-HR-MS and MS/MS measurements for 34 metabolites detected by both techniques	94
4.8. nS-HR-MS vs. MS/MS measurements of Phenylalanine for PKU positive samples, HPA samples, and non-PKU/HPA positives	95
4.9. Receiver operating characteristic (ROC) curves for PKU diagnosis by phenylalanine	96
4.10. Classification of PKU test-set samples based on stepwise logistic regression model	98
4.11. Receiver operating characteristic (ROC) curves for MCADD diagnosis by octanoylcarnitine	99
4.12. Classification of MCADD test-set samples by stepwise logistic regression model	100
4.13. Classification of independent MCADD samples by stepwise logistic regression model	101
4.14. Biomarkers of PKU	103
4.15. Biomarkers of MCADD	105

4.16. Metabolites outside of normal ranges for 24 inborn errors by nS- HR-MS.....	106
4.17. Metabolites outside of normal ranges for 24 inborn errors by MS/MS.....	107
4.18. Biomarkers of MSUD mapped onto KEGG	109
5.1. Metabolic profiling of Fanconi Syndrome in cystinosis urine samples by NMR spectroscopy	119
5.2. Confirmation of tartrate in the urine of cystinosis patients	120
5.3. Clinical phenotypes observed in congenital kidney diseases	123
5.4. Urinary excretion of metabolites elevated in Fanconi Syndromes relative to controls	124
5.5. Metabolites differentiating congenital Fanconi Syndromes	125
5.6. Urinary citrate, acetate, and inositol excretion in ERS patients and unaffected controls.....	126
5.7. Metabolic perturbations observed in congenital kidney diseases.....	128
5.8. Analysis of cystinosis natural history metadata variables	129
5.9. Statistical analysis of metabolites from cystinosis urinary NMR data.....	131
5.10. Correlations between metabolites excreted by cystinosis patients.....	133
5.11. Significant associations between urinary metabolites and cystinosis patient metadata variables identified by regression analysis	134
5.12. Changes in key urinary metabolites by time from transplantation.....	136
5.13. Urinary NMR signals of glucose, TMAO and succinate over time from a single cystinosis patient.....	138
5.14. Urinary excretion of metabolites significantly altered in animals exposed to model nephrotoxins relative to controls	140
5.15. Trajectories of key metabolites over nephrotoxin exposure time- course.....	142

List of Tables

2.1. Frequently measured nuclei in biomolecular Nuclear Magnetic Resonance	29
2.2. Characteristics of commonly used mass analyzers.....	37
2.3. Data definition tools in NMR and MS.....	40
2.4. Binary classification confusion matrix	51
4.1. Inborn errors of metabolism profiled by nS-HR-MS and MS/MS dried blood spot analysis	84
4.2. Logistic regression models for prediction of PKU class by nS-HR-MS and MS/MS	97
4.3. Logistic regression models for prediction of MCADD class by nS-HR-MS and MS/MS	100
5.1. Cystinosis patient characteristics	113
5.2. Assignments of urinary NMR signals integrated for targeted analysis	118

List of Published Papers

Papers in Appendix B (I-II), C (III) and D (IV) are directly related to work presented in Chapters 3-5 in this thesis, respectively. Papers in Appendix E(V-VIII) were published during my tenure as a PhD candidate at Imperial and are related to the background theory, and perspectives reflected in Chapters 1, 2, and 6 but are not directly related to the work presented in Chapters 3-5. Reproduction permission for work from these papers is included in Appendix F and as a result figures and text from these papers appear throughout Chapters 1-6.

Appendix B

- I. Izrayelit Y[†], Robinette SL[†], Bose N, von Reuss SH, Schroeder FC. 2D NMR-based metabolomics uncovers interactions between conserved biochemical pathways in the model organism *Caenorhabditis elegans*. *ACS Chem Biol*, 2012. 8(2):314-319. doi:10.1021/cb3004644. [†]equal contribution
- II. Robinette SL, Ajredini R, Rasheed H, Zeinomar A, Schroeder FC, Dossey AT, Edison AS. Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology. *Anal Chem*, 2011. 83(5):1649-1657.

Appendix C

- III. Dénes J, Szabó E, Robinette SL, Szatmári I, Szőnyi L, Kreuder JG, Rauterberg EW, Takáts Z. Metabonomics of Newborn Screening Dried Blood Spot Samples: A Novel Approach in the Screening and Diagnostics of Inborn Errors of Metabolism. *Anal Chem*, 2012. 84(22):10113-10120. doi:10.1021/ac302527m

Appendix D

- IV. Jaureguiberry G, De la Dure-Molla M, Parry D, Quentric M, Himmerkus N, Koike T, Poulter J, Klootwijk E, Robinette SL, Howie AJ, Patel V, Figueres ML, Stanescu HC, Issler N, Nicholson JK, Bockenhauer D, Laing C, Walsh SB, McCredie DA, Povey S, Asselin A, Picard A, Coulomb A, Medlar AJ, Bailleul-Forestier I, Verloes A, Le Caignec C, Roussey G, Guiol J, Isidor B, Logan C, Shore R, Johnson C, Inglehearn C, Al-Bahlani S, Schmittbuhl M, Clauss F, Huckert M, Laugel V, Ginglinger E, Pajarola S, Spartà G, Bartholdi D, Rauch A, Addor MC, Yamaguti PM, Safatle HP, Acevedo AC, Martelli-Júnior H, E Dos Santos Netos P, Coletta RD, Gruessel S, Sandmann C, Ruehmann D, Langman CB, Scheinman SJ, Ozdemir-Ozenen D, Hart TC, Hart PS, Neugebauer U, Schlatter E, Houillier P, Gahl WA, Vikkula M, Bloch-Zupan A, Bleich M, Kitagawa H, Unwin RJ, Mighell A, Berdal A, Kleta R. Nephrocalcinosis (Enamel Renal Syndrome) Caused by Autosomal Recessive FAM20A Mutations. *Nephron Physiol*, 2012. 122(1-2):1-6. doi: 10.1159/000349989.

Appendix E

- V. Robinette SL, Lindon J, Nicholson JK. Statistical Spectroscopic Tools for Biomarker Discovery and Systems Medicine. *Anal Chem*, 2013. 85(11):5297-303.
- VI. Robinette SL, Holmes E, Nicholson JK, Dumas ME. Genetic Determinants of Metabolism in Health and Disease: From Biochemical Genetics to Genome-Wide Associations. *Genome Med*, 2012. 4:30. doi:10.1186/gm329
- VII. Robinette SL, Brüscheweiler R, Schroeder F, Edison AS. Metabolomics and Natural Products Research: Two Sides of the Same Coin. *Acc Chem Res*, 2011. 45(2):288-297.
- VIII. Veselkov KA, Vingara LK, Masson P, Robinette SL, Want E, Li JV, Barton RH, Claire BN, Walther B, Ebbels TM, Pelczar I, Holmes E, Lindon JC, Nicholson JK. Optimized pre-processing of UPLC-MS urinary metabolic profiles for improved information recovery. *Anal Chem*, 2011. 83(15):5864-5872.

CHAPTER 1

INTRODUCTION AND BACKGROUND

Rapid advances in the genomic sciences enabling the rapid acquisition of genome-wide gene-expression profiles, polymorphisms and, more recently, whole genome sequences have contributed to a dramatic increase in the number of known and suspected associations between genetic variations and human disease. As the costs of genomic sequencing fall, functional genomic analysis – the characterization of the phenotypic consequences of genetic variants – will increasingly constitute the rate-limiting step in the study of biochemical genetics. Genes - ‘**upstream**’ **blueprints** controlling dynamic biological processes - and metabolites - capturing both genetic **and environmental** ‘**downstream**’ effects - are intrinsically co-informational, shedding light on complementary biological processes. Thus, applying metabolic profiling technologies to identify the metabolic perturbations resulting from genetic mutations will likely help to annotate genes of unknown function, identify genetic controls of metabolism, and characterize mechanisms of disease.

1.1 Background

1.1.1 Characterizing the role of rare variants in human disease

The potential for genetic mutations to disrupt metabolic processes and result in human disease has been known since the discovery of alkaptonuria by Archibald Garrod in the early 20th Century [1]. In the following century, hundreds of these diseases, termed inborn errors of metabolism or Mendelian diseases due to their typically single-gene origins and Mendelian inheritance patterns, were identified. Following the introduction of molecular cloning and Sanger sequencing technologies, the molecular basis of these diseases began to be uncovered [2, 3]. Advances in

sequencing technology and drastic reductions in the cost of genome sequencing have accelerated the identification of the genetic background underlying these rare congenital diseases, and the molecular bases of inborn errors first described by Garrod are still being uncovered [4]. Statistics from the Online Mendelian Inheritance in Man (OMIM) database show an accelerating rate of growth in the number of known genes for which there is a described Mendelian phenotype (Fig.1.1).

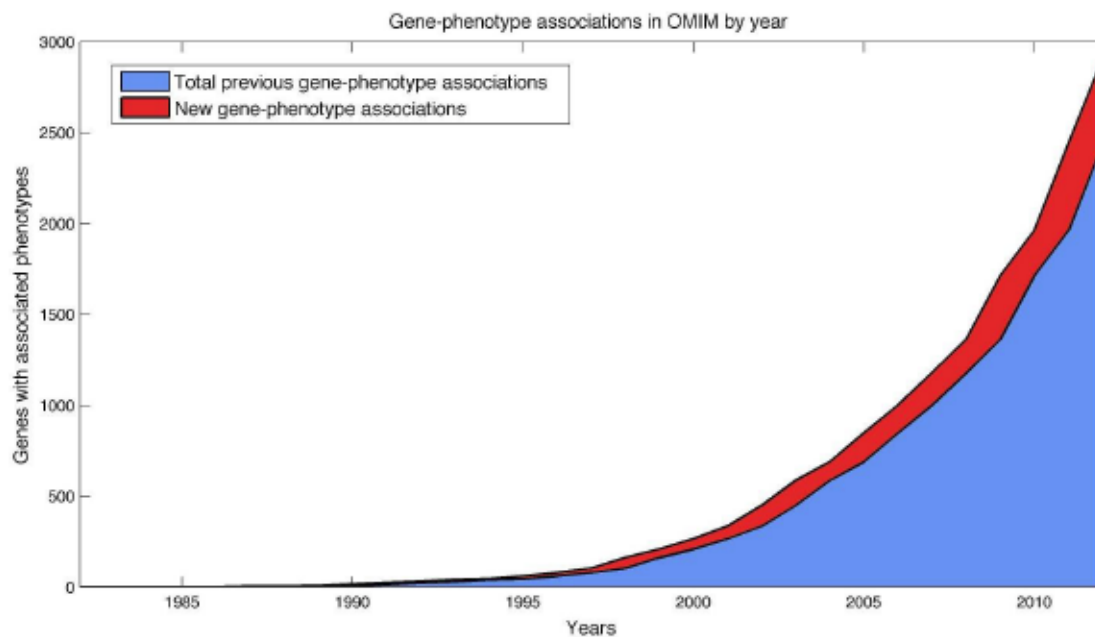


Figure 1.1. Identification of genotype-phenotype associations in Mendelian disease is accelerating. The rate of discovery for genotype-phenotype associations has increased year-on-year from the late 1980s to the present (red bar). This has resulting in a growing set of genes with known Mendelian phenotypes, which is now approaching 3,000. Data from omim.org.

Following the sequencing of the human genome [5] attention shifted from rare variants and rare diseases to common variants and common diseases. However, common single-nucleotide polymorphisms and genome-wide associations studies showed missing heritability in common diseases [6] and weak associations with a large number of genes [7]. As a result, there is once again growing appreciation for the importance of rare variants and alleles of large effect in both rare and common diseases, with some leading groups suggesting that common disease may in fact be overlapping phenotypes resulting from a large set of rare diseases [8]. Current statistics from OMIM (Fig.1.1) show that of known and suspected human Mendelian diseases, molecular bases have been identified for over 4,800 phenotypes, with ~3000

known genes associated with described Mendelian phenotypes (<http://omim.org/>) [9]. Mapping these genes onto human metabolic pathways (Fig.1.2) show that mutations that result in an identifiable human phenotype affect nearly all known metabolic pathways. These findings indicate the central importance of genetic variation, especially rare variation, in human disease.

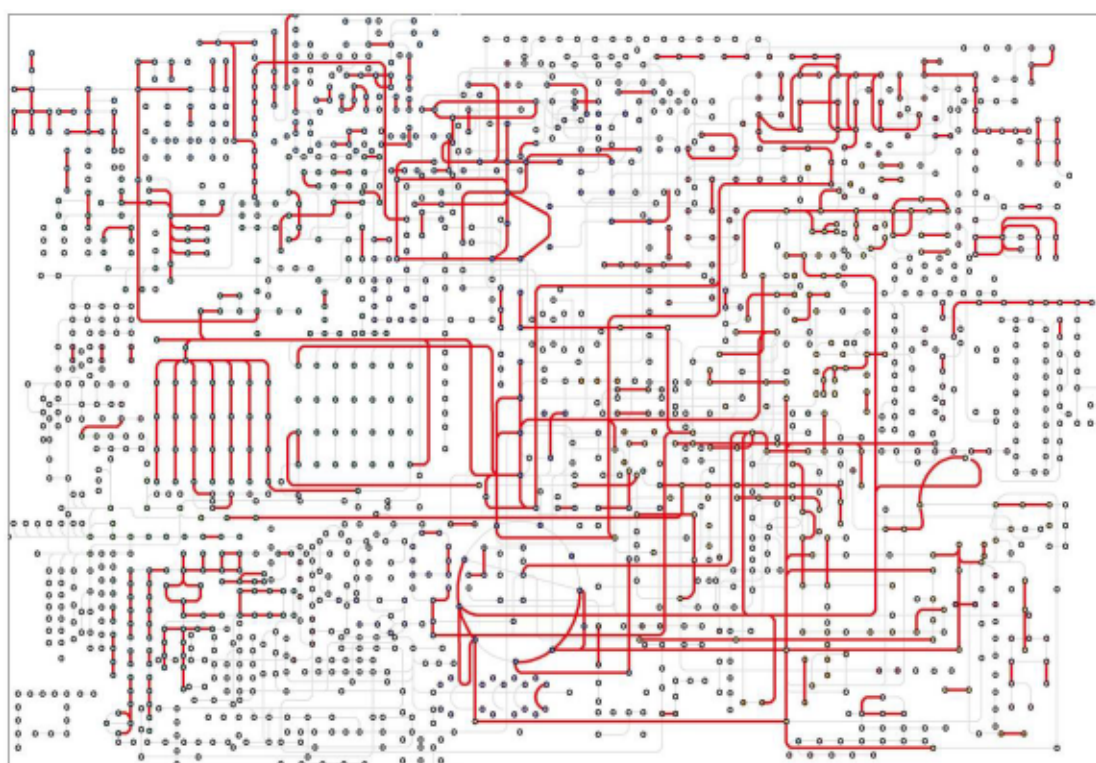


Figure 1.2. Genes with known Mendelian phenotypes mapped onto KEGG human metabolic pathways. Of the ~3,000 genes with known Mendelian phenotypes in OMIM, >450 of these matched genes in the KEGG global human metabolic pathway database (hsa01100). The distribution of these genes, plotted in red, indicates that known human Mendelian diseases affect the majority of human metabolic pathways. Data from omim.org and kegg.jp.

1.1.2 Identifying genetic and metabolic covariation.

Since the discovery of alkaptonuria by Archibald Garrod in the early 20th Century, the measurement of metabolites has been used as a proxy to identify human genetic diseases, especially inborn errors of metabolism. Uniform newborn screening for multiple inborn errors of metabolism, including urea cycle disorders and amino- and organic acidurias, using heel-prick testing with MS/MS exemplifies the power of

targeted metabolomic analysis to diagnose these diseases and enable the interruption of pathological processes resulting from genetic mutations [10]. Untargeted metabolic profiling of biofluids, especially urine and serum, is a powerful technique for diagnosing inborn errors of metabolism with often non-specific clinical presentation as metabolic intermediates accumulated in biofluid compartments can be easily identified [11, 12]. Untargeted NMR and MS have been used to diagnose known inborn errors [13-16], identify novel inborn errors using biofluid profiling [17, 18] and identify complex downstream metabolic consequences [19-21] and biomarkers of organ pathology resulting from genetic mutations [22-24].

While mutations in enzymes and transporters can often be readily diagnosed by biofluid profiling and a strong mechanistic link is easily inferred between the disruption of a metabolic pathway and resulting accumulation or depletion of metabolic intermediates, many Mendelian diseases result in more complex, progressive organ specific or multi-organ pathology [25]. In these cases, metabolic profiling has been applied to identify sites of lesions, describe progression, and attempt to identify proxy small molecular biomarkers of disease. Examples of this include the identification of markers of autosomal dominant polycystic kidney disease [24], comparison of the urinary profiles of several genetic forms of renal Fanconi syndrome [23], and description of abnormal brain metabolism in Smith-Lemli-Optiz syndrome [22]. Additionally, metabolite flux analysis using isotopically labeled metabolites suggests an additional way to apply metabolic profiling techniques to study the impact of genetic mutations [25].

The genomic corollary of using metabolic profiling to study Mendelian genes is quantitative trait loci (QTL) mapping of single metabolic traits. The advent of whole-genome SNP analysis and the use of QTL mapping for quantitative traits, such as height [26] led to interest in identifying genetic loci associated with quantitative variation in individual metabolite levels [27]. Examples of this include mapping of serum leptin levels to genes on human chromosome 2 in multiple human populations [28, 29], associations with plasma triglyceride levels [30, 31], and identification of genetic variants associated with high-density lipoprotein levels [32, 33].

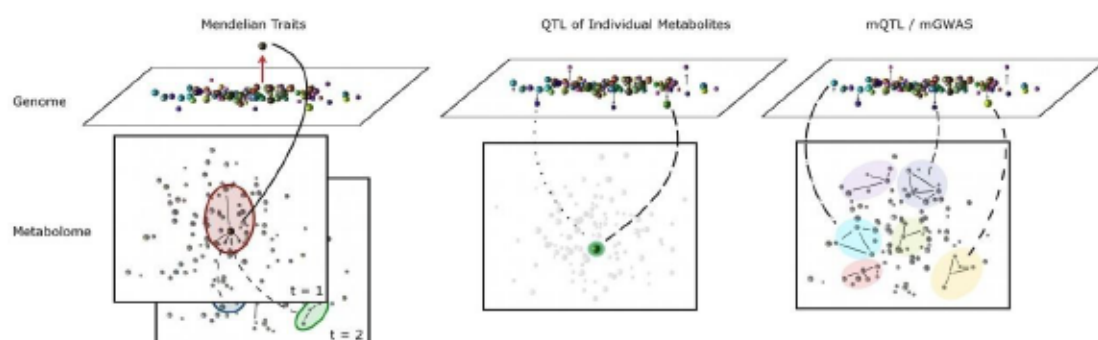


Figure 1.3. Three experimental designs integrating genomic and metabolomic analysis. Metabolic profiling applied to the diagnosis and study of human Mendelian diseases frequently identifies direct, causal relationships between genetic variants and downstream accumulation or deficiency of metabolic intermediates, which may vary or progress over time. QTL mapping of single quantified metabolites can identify strong associations between metabolite concentration and polymorphisms, though frequently additional, weaker associations with other alleles are discovered as well. mQTL and mGWAS studies are conceptually similar to QTL studies of individual metabolites, but search for associations between many metabolites and many genes, frequently yielding a larger set of associations between genetic polymorphisms and metabolite concentrations or ratios.

Recently, genome-wide single nucleotide polymorphism (SNP) measurements have been combined with metabolomics to identify a large number of gene-metabolite associations at once. In these studies, individuals are genotyped and metabotyped in parallel and the resulting genome-wide and metabolome-wide profiles are then quantitatively correlated. This approach, first termed metabolic Quantitative Trait Loci mapping (mQTL), was demonstrated in plants [34, 35] and then in animal models [36, 37]. Following these successes, the mQTL approach was quickly followed by the development of metabolomic or metabolome-wide GWAS in humans cohorts [38-45]. These approaches have led to the identification of a growing number of gene-metabolite associations, particularly between genes encoding enzymes and transporters and their metabolic substrates. Figure 1.3 outlines the associations identified by metabolic profiling of Mendelian disease, QTL analysis of metabolic traits, and mQTL/mGWAS.

1.1.3 Mendelian disease diagnostics: genotypes or metabotypes?

While inborn errors of metabolism were initially considered to be idiosyncratic "chemical individuality", growing understanding of neonatal diseases and the identification of pathological inborn errors such as phenylketonuria (PKU)

lead to efforts to implement population-wide newborn screening programs as early diagnosis enabled intervention, such as dietary restriction of phenylalanine in the case of PKU, that disrupted the progression of the disease. Early newborn screening programs, such as Robert Guthrie's PKU screening protocol [46], relied on bacterial inhibition assays, but in the 1990s the introduction of tandem mass spectrometry for the measurement of small molecule biomarkers of inborn errors [46-48] gave screening programs a much more sensitive and specific analytical platform. Today, targeted metabolic profiling based on MS/MS is the main primary screening technique for detection of the 31 diseases included in the US Recommended Uniform Screening Panel for neonatal screening [49].

As clinical genetics, including whole exome and genome sequencing [50-52], expands the set of known inborn errors it is likely that screening programs will include a larger number of diseases in the future. At the same time, new screening methods based on genotype rather than metabolite have appeared, including pre-implantation genetic diagnosis (PGD) [50-54], parental carrier testing [55, 56], and recently diagnostics based on fetal genome reconstructions from maternal sequencing [57]. These tests provide risk assessment and prenatal diagnostics, information that arguably has much greater utility than the neonatal diagnostics provided by metabolic testing. However, genetic backgrounds for even single gene disorders are highly heterogeneous, with multiple variants including point mutations, indels, and copy number variants of varying deleteriousness contributing to a spectrum of phenotypic consequences [58-60]. While a comprehensive catalog of these genotypes and their phenotypic consequences may enable purely genetic screening, both characterizing the spectrum of genotype-phenotype relationships and screening in the near term will require continued metabolic profiling.

1.2 Scope of thesis

Clearly, metabolic profiling has a continuing role to play in the study of biochemical genetics. This thesis presents the application of metabolic profiling technologies, including Nuclear Magnetic Resonance and Mass Spectrometry, to identify metabolic perturbations resulting from inborn errors of metabolism, congenital diseases affecting the kidney, and animal models of both genetic mutations

and renal pathology. While MS based newborn screening for inborn errors is a major application of metabolic profiling used in the clinic and would likely benefit from broader metabolomic coverage, these diseases are typically quite simple in their mechanisms and progression. Congenital diseases of the kidney, though also monogenic in origin, constitute a greater challenge and greater opportunity for contribution from metabolic profiling. Additionally, data from these studies are used to introduce and evaluate modified analytic methods for metabolic profiling that mitigate some of the challenges involved in biochemical genetics, such as small sample sizes, unknown or novel chemical structures, and the challenge of providing broader metabolomic coverage.

1D NMR is frequently used for metabolic profiling, but while 2D ^1H - ^1H NMR spectra require additional collection time they are well suited to small sample sizes and assignment challenges due to unexpected compounds and increased spectral complexity due to accumulation of metabolic intermediates or shunt metabolites. Data processing methods for 2D NMR spectral dataset are introduced, including for signal alignment, binning, and pattern recognition and evaluated using data collected from closely related nematode species, *C. elegans*, and humans. Increased structural information aids identification of accumulating long-chain intermediates produced by a peroxisomal β -oxidation defect in *C. elegans* and increased spectral resolution enhances analysis of Fanconi Syndrome (FS) resulting from a peroxisomal β -oxidation trafficking defect in a single human pedigree.

While tandem mass spectrometry (MS/MS) is used to diagnose inborn errors from dried blood spot samples in the majority of newborn screening programs, targeted analysis by MS/MS typically involves measurement of a small subset of metabolites present in human blood. As a result, diagnosis typically requires additional confirmatory metabolic measurements, and the identification of global metabolic perturbations resulting from Mendelian diseases is limited. To investigate whether broader metabolomic coverage could yield improved diagnostic accuracy and additional biomarkers, MS/MS data from ~6,000 newborns was compared with a recently introduced direct injection nanospray high resolution mass spectrometry (nS-HR-MS) assay providing coverage of >200 metabolites for newborns both unaffected and affected by 24 congenital disorders of metabolism.

Inborn errors of metabolism are typically considered to result from mutations in enzymes and transporters, and these can often be readily diagnosed by biofluid

profiling as a strong mechanistic link is easily inferred between the disruption of a metabolic pathway and resulting accumulation or depletion of metabolic intermediates. However, many Mendelian diseases result in more complex, progressive, organ specific or multi-organ pathology. The final chapter of this thesis presents urinary metabolic profiles from patients with one of four related congenital kidney diseases and the comparison of these profiles with each other and with unaffected family members. While these disorders result from distinct genetic mutations affecting different genes, pathways, and cellular and metabolic processes, they present with an overlapping set of clinical phenotypes. As a result, some of the **features of patients' metabolic profiles are shared, while others are distinct**, helping unravel the complex set of phenotypes observed in these disorders. Additionally, a time course analysis of the natural history of one of these disorders, cystinosis, is presented from samples collected over a 15-year period. The metabolic perturbations observed in these human diseases are compared to those observed in animal studies of model nephrotoxins affecting an overlapping set of kidney structures.

CHAPTER 2

THEORY AND METHODS

The work presented in this thesis is the result of applications of a diverse set of analytical and computational techniques which when applied in conjunction are typically referred to as metabolomics [61], metabonomics [62], or metabolic profiling. Details of experimental design, sample collection, data acquisition and analysis can be found in the individual chapters, as the tools used differ based on application. Protocols for metabolomics research for both major analytical platforms – Nuclear Magnetic Resonance (NMR) [63] and Mass Spectrometry (MS) [64-66] – as well as for data analysis [67, 68] and compound identification [69] are available. As a result this chapter focuses on introducing the theory of NMR and MS as well as data preparation, statistical analysis, and biomarker identification approaches applied in this thesis. The purpose of this chapter is to introduce the family of techniques frequently used in metabolomics research and to discuss their strengths and pitfalls to clarify the analytical and computational strategies employed in the following chapters.

2.1 Data acquisition – analytical chemistry of biological fluids

2.1.1 Nuclear Magnetic Resonance (NMR) Spectroscopy

Nuclear magnetic resonance (NMR) is a powerful and almost universal detector due to its ability to analyze essentially all types of molecules at atomic resolution. NMR spectroscopy has been applied continuously in organic chemistry, structural biology, and metabolic analysis for the past 50 years and as a result there are many comprehensive guides to both the theory and practice of NMR [70-72]. In small molecule research, NMR spectroscopy has a long history of application in organic chemistry for structural identification and is used extensively in

metabolomics. NMR spectroscopy is characterized by the following key properties: (i) high dynamic range, with several biological nuclei, such as ^1H , ^{13}C , ^{15}N or ^{31}P , being accurately measured over a large range of concentrations; (ii) high linearity of signal intensity with concentration; and (iii) high reproducibility.

While NMR and MS were used most frequently for the structural characterization of organic and inorganic chemical structures, beginning in the 1980s these techniques began to be used as biochemical sensors with changes in the patterns of complex mixtures measured to assess biological rather than structural changes[73]. In particular, ^1H NMR spectroscopy is robust, provides a high degree of structural information for both one-dimensional ^1H and two-dimensional ^1H - ^1H NMR, and is flexibly applied to extracts, biofluids and solid tissues using high-resolution magic angle spinning, and in vivo using magnetic resonance spectroscopy. Technological advancements in magnetic field strength with the introduction of 600, 800 and, recently, 1,000 MHz NMR spectrometers, pulse sequence experiments, and cryogenically cooled probes have increased the sensitivity and coverage for small-molecular-weight metabolites from urine, plasma, serum and tissue samples [63].

2.1.1.1 Principles of NMR spectroscopy

Nuclear spin is a fundamental quantum property of atomic nuclei, and nuclei with spin quantum number $1/2$ (Table 2.1) have two possible orientations corresponding to $m = +1/2$ and $m = -1/2$. While these spin states are degenerate in the absence of magnetic field, when an atomic nucleus with non-zero spin is placed in an ambient magnetic field B_0 the energetic consequences are determined by equation 2.1.

Nuclear spin quantum number, I			
$I = 1/2$	$I = 1$	$I = 3/2$	$I = 5/2$
^1H	^2H	^3H	^7Li
^{13}C	^{14}N	^{15}N	^{17}O
^{19}F	^{20}Ne	^{23}Na	^{25}Mg
^{29}Si	^{30}Si	^{31}P	^{33}S
^{35}Cl	^{36}Ar	^{39}K	^{41}K
^{41}K	^{40}Ca	^{43}Ca	^{45}Ca
^{45}Ca	^{46}Ca	^{47}Ti	^{49}Ti
^{49}Ti	^{50}Ti	^{51}V	^{53}Cr
^{53}Cr	^{54}Cr	^{57}Fe	^{59}Co
^{59}Co	^{60}Ni	^{61}Ni	^{63}Ni
^{63}Ni	^{64}Ni	^{67}Zn	^{69}Zn
^{69}Zn	^{70}Zn	^{71}Ga	^{73}Ga
^{73}Ga	^{74}Ge	^{75}Ge	^{77}Ge
^{77}Ge	^{78}Se	^{79}Se	^{81}Se
^{81}Se	^{82}Kr	^{83}Kr	^{85}Kr
^{85}Kr	^{86}Kr	^{87}Rb	^{89}Rb
^{89}Rb	^{90}Zr	^{91}Zr	^{93}Zr
^{93}Zr	^{94}Mo	^{95}Mo	^{97}Mo
^{97}Mo	^{98}Mo	^{99}Tc	^{101}Tc
^{101}Tc	^{102}Ru	^{103}Ru	^{105}Ru
^{105}Ru	^{106}Ru	^{107}Ag	^{109}Ag
^{109}Ag	^{110}Ag	^{111}Ag	^{113}Ag
^{113}Ag	^{114}Cd	^{115}Cd	^{117}Cd
^{117}Cd	^{118}Cd	^{119}Sn	^{121}Sn
^{121}Sn	^{122}Sn	^{123}Sn	^{125}Sn
^{125}Sn	^{126}Te	^{127}Te	^{129}Te
^{129}Te	^{130}Te	^{131}Xe	^{133}Xe
^{133}Xe	^{134}Xe	^{135}Xe	^{137}Xe
^{137}Xe	^{138}Ba	^{139}Ba	^{141}Ba
^{141}Ba	^{142}Ba	^{143}Ba	^{145}Ba
^{145}Ba	^{146}Ba	^{147}La	^{149}La
^{149}La	^{150}La	^{151}Eu	^{153}Eu
^{153}Eu	^{154}Eu	^{157}Gd	^{159}Gd
^{159}Gd	^{160}Gd	^{161}Gd	^{163}Gd
^{163}Gd	^{164}Gd	^{165}Dy	^{167}Dy
^{167}Dy	^{168}Dy	^{169}Er	^{171}Er
^{171}Er	^{172}Er	^{173}Er	^{175}Er
^{175}Er	^{176}Yb	^{177}Yb	^{179}Yb
^{179}Yb	^{180}Yb	^{181}Hf	^{183}Hf
^{183}Hf	^{184}Hf	^{187}Re	^{189}Re
^{189}Re	^{190}Os	^{191}Os	^{193}Os
^{193}Os	^{194}Os	^{195}Pt	^{197}Pt
^{197}Pt	^{198}Pt	^{199}Au	^{201}Au
^{201}Au	^{202}Hg	^{203}Hg	^{205}Hg
^{205}Hg	^{206}Hg	^{207}Pb	^{209}Pb
^{209}Pb	^{210}Pb	^{211}Pb	^{213}Pb
^{213}Pb	^{214}Po	^{215}Po	^{217}Po
^{217}Po	^{218}Po	^{219}Po	^{221}Po
^{221}Po	^{222}Rn	^{223}Rn	^{225}Rn
^{225}Rn	^{226}Rn	^{227}Rn	^{229}Rn
^{229}Rn	^{230}Th	^{231}Th	^{232}Th
^{232}Th	^{233}Th	^{234}Th	^{236}Th
^{236}Th	^{237}Pa	^{238}Pa	^{240}Pa
^{240}Pa	^{241}Pa	^{242}Pa	^{244}Pa
^{244}Pa	^{245}U	^{246}U	^{248}U
^{248}U	^{249}U	^{250}U	^{252}U
^{252}U	^{253}Np	^{254}Np	^{256}Np
^{256}Np	^{257}Np	^{258}Np	^{260}Np
^{260}Np	^{261}Pu	^{262}Pu	^{264}Pu
^{264}Pu	^{265}Pu	^{266}Pu	^{268}Pu
^{268}Pu	^{269}Pu	^{270}Pu	^{272}Pu
^{272}Pu	^{273}Am	^{274}Am	^{276}Am
^{276}Am	^{277}Am	^{278}Am	^{280}Am
^{280}Am	^{281}Cm	^{282}Cm	^{284}Cm
^{284}Cm	^{285}Cm	^{286}Cm	^{288}Cm
^{288}Cm	^{289}Cf	^{290}Cf	^{292}Cf
^{292}Cf	^{293}Cf	^{294}Cf	^{296}Cf
^{296}Cf	^{297}Cf	^{298}Cf	^{300}Cf
^{300}Cf	^{301}Es	^{302}Es	^{304}Es
^{304}Es	^{305}Es	^{306}Es	^{308}Es
^{308}Es	^{309}Fm	^{310}Fm	^{312}Fm
^{312}Fm	^{313}Fm	^{314}Fm	^{316}Fm
^{316}Fm	^{317}Md	^{318}Md	^{320}Md
^{320}Md	^{321}Md	^{322}Md	^{324}Md
^{324}Md	^{325}No	^{326}No	^{328}No
^{328}No	^{329}No	^{330}No	^{332}No
^{332}No	^{333}Lr	^{334}Lr	^{336}Lr
^{336}Lr	^{337}Lr	^{338}Lr	^{340}Lr
^{340}Lr	^{341}Lu	^{342}Lu	^{344}Lu
^{344}Lu	^{345}Lu	^{346}Lu	^{348}Lu
^{348}Lu	^{349}Y	^{350}Y	^{352}Y
^{352}Y	^{353}Y	^{354}Y	^{356}Y
^{356}Y	^{357}Zr	^{358}Zr	^{360}Zr
^{360}Zr	^{361}Zr	^{362}Zr	^{364}Zr
^{364}Zr	^{365}Nb	^{366}Nb	^{368}Nb
^{368}Nb	^{369}Nb	^{370}Nb	^{372}Nb
^{372}Nb	^{373}Mo	^{374}Mo	^{376}Mo
^{376}Mo	^{377}Mo	^{378}Mo	^{380}Mo
^{380}Mo	^{381}Tc	^{382}Tc	^{384}Tc
^{384}Tc	^{385}Tc	^{386}Tc	^{388}Tc
^{388}Tc	^{389}Ru	^{390}Ru	^{392}Ru
^{392}Ru	^{393}Ru	^{394}Ru	^{396}Ru
^{396}Ru	^{397}Rh	^{398}Rh	^{400}Rh
^{400}Rh	^{401}Rh	^{402}Rh	^{404}Rh
^{404}Rh	^{405}Pd	^{406}Pd	^{408}Pd
^{408}Pd	^{409}Pd	^{410}Pd	^{412}Pd
^{412}Pd	^{413}Ag	^{414}Ag	^{416}Ag
^{416}Ag	^{417}Ag	^{418}Ag	^{420}Ag
^{420}Ag	^{421}Cd	^{422}Cd	^{424}Cd
^{424}Cd	^{425}Cd	^{426}Cd	^{428}Cd
^{428}Cd	^{429}In	^{430}In	^{432}In
^{432}In	^{433}In	^{434}In	^{436}In
^{436}In	^{437}Sn	^{438}Sn	^{440}Sn
^{440}Sn	^{441}Sn	^{442}Sn	^{444}Sn
^{444}Sn	^{445}Sb	^{446}Sb	^{448}Sb
^{448}Sb	^{449}Sb	^{450}Sb	^{452}Sb
^{452}Sb	^{453}Te	^{454}Te	^{456}Te
^{456}Te	^{457}Te	^{458}Te	^{460}Te
^{460}Te	^{461}I	^{462}I	^{464}I
^{464}I	^{465}I	^{466}I	^{468}I
^{468}I	^{469}Xe	^{470}Xe	^{472}Xe
^{472}Xe	^{473}Xe	^{474}Xe	^{476}Xe
^{476}Xe	^{477}Ba	^{478}Ba	^{480}Ba
^{480}Ba	^{481}Ba	^{482}Ba	^{484}Ba
^{484}Ba	^{485}La	^{486}La	^{488}La
^{488}La	^{489}La	^{490}La	^{492}La
^{492}La	^{493}Ce	^{494}Ce	^{496}Ce
^{496}Ce	^{497}Ce	^{498}Ce	^{500}Ce
^{500}Ce	^{501}Pr	^{502}Pr	^{504}Pr
^{504}Pr	^{505}Pr	^{506}Pr	^{508}Pr
^{508}Pr	^{509}Nd	^{510}Nd	^{512}Nd
^{512}Nd	^{513}Nd	^{514}Nd	^{516}Nd
^{516}Nd	^{517}Pm	^{518}Pm	^{520}Pm
^{520}Pm	^{521}Pm	^{522}Pm	^{524}Pm
^{524}Pm	^{525}Sm	^{526}Sm	^{528}Sm
^{528}Sm	^{529}Sm	^{530}Sm	^{532}Sm
^{532}Sm	^{533}Eu	^{534}Eu	^{536}Eu
^{536}Eu	^{537}Eu	^{538}Eu	^{540}Eu
^{540}Eu	^{541}Gd	^{542}Gd	^{544}Gd
^{544}Gd	^{545}Gd	^{546}Gd	^{548}Gd
^{548}Gd	^{549}Ter	^{550}Ter	^{552}Ter
^{552}Ter	^{553}Ter	^{554}Ter	^{556}Ter
^{556}Ter	^{557}Dy	^{558}Dy	^{560}Dy
^{560}Dy	^{561}Dy	^{562}Dy	^{564}Dy
^{564}Dy	^{565}Ho	^{566}Ho	^{568}Ho
^{568}Ho	^{569}Ho	^{570}Ho	^{572}Ho
^{572}Ho	^{573}Er	^{574}Er	^{576}Er
^{576}Er	^{577}Er	^{578}Er	^{580}Er
^{580}Er	^{581}Tm	^{582}Tm	^{584}Tm
^{584}Tm	^{585}Tm	^{586}Tm	^{588}Tm
^{588}Tm	^{589}Yb	^{590}Yb	^{592}Yb
^{592}Yb	^{593}Yb	^{594}Yb	^{596}Yb
^{596}Yb	^{597}Lu	^{598}Lu	^{600}Lu
^{600}Lu	^{601}Lu	^{602}Lu	^{604}Lu
^{604}Lu	^{605}Hf	^{606}Hf	^{608}Hf
^{608}Hf	^{609}Hf	^{610}Hf	^{612}Hf
^{612}Hf	^{613}Ta	^{614}Ta	^{616}Ta
^{616}Ta	^{617}Ta	^{618}Ta	^{620}Ta
^{620}Ta	^{621}W	^{622}W	^{624}W
^{624}W	^{625}W	^{626}W	^{628}W
^{628}W	^{629}Re	^{630}Re	^{632}Re
^{632}Re	^{633}Re	^{634}Re	^{636}Re
^{636}Re	^{637}Os	^{638}Os	^{640}Os
^{640}Os	^{641}Os	^{642}Os	^{644}Os
^{644}Os	^{645}Ir	^{646}Ir	^{648}Ir
^{648}Ir	^{649}Ir	^{650}Ir	^{652}Ir
^{652}Ir	^{653}Pt	^{654}Pt	^{656}Pt
^{656}Pt	^{657}Pt	^{658}Pt	^{660}Pt
^{660}Pt	^{661}Au	^{662}Au	^{664}Au
^{664}Au	^{665}Au	^{666}Au	^{668}Au
^{668}Au	^{669}Hg	^{670}Hg	^{672}Hg
^{672}Hg	^{673}Hg	^{674}Hg	^{676}Hg
^{676}Hg	^{677}Tl	^{678}Tl	^{680}Tl
^{680}Tl	^{681}Tl	^{682}Tl	^{684}Tl
^{684}Tl	^{685}Pb	^{686}Pb	^{688}Pb
^{688}Pb	^{689}Pb	^{690}Pb	^{692}Pb
^{692}Pb	^{693}Bi	^{694}Bi	^{696}Bi
^{696}Bi	^{697}Bi	^{698}Bi	^{700}Bi
^{700}Bi	^{701}Po	$^{70210}\$	

$$E = -\gamma m \hbar B_0 \quad \square\square\square\square$$

Here, B_0 is the strength of the magnetic field, m is the nuclear spin, $\hbar$ is the reduced Planck constant ($\hbar/2 = 1.054571726 \times 10^{-34}$ J-s), and γ is the gyromagnetic ratio of the nucleus, values of which for typical NMR nuclei can be seen in Table 2.1. By convention, $m = +1/2$ is the lower energy orientation, and as a result the energy difference between the two spin 1/2 states is shown in eq. 2.2, and the energy differential between these two states will lead to net polarization of the population of nuclear spins as the equilibrium between spin +1/2 and spin -1/2 states shifts to favor the lower energy state. The ratio of these spin states is determined by equation 2.3.

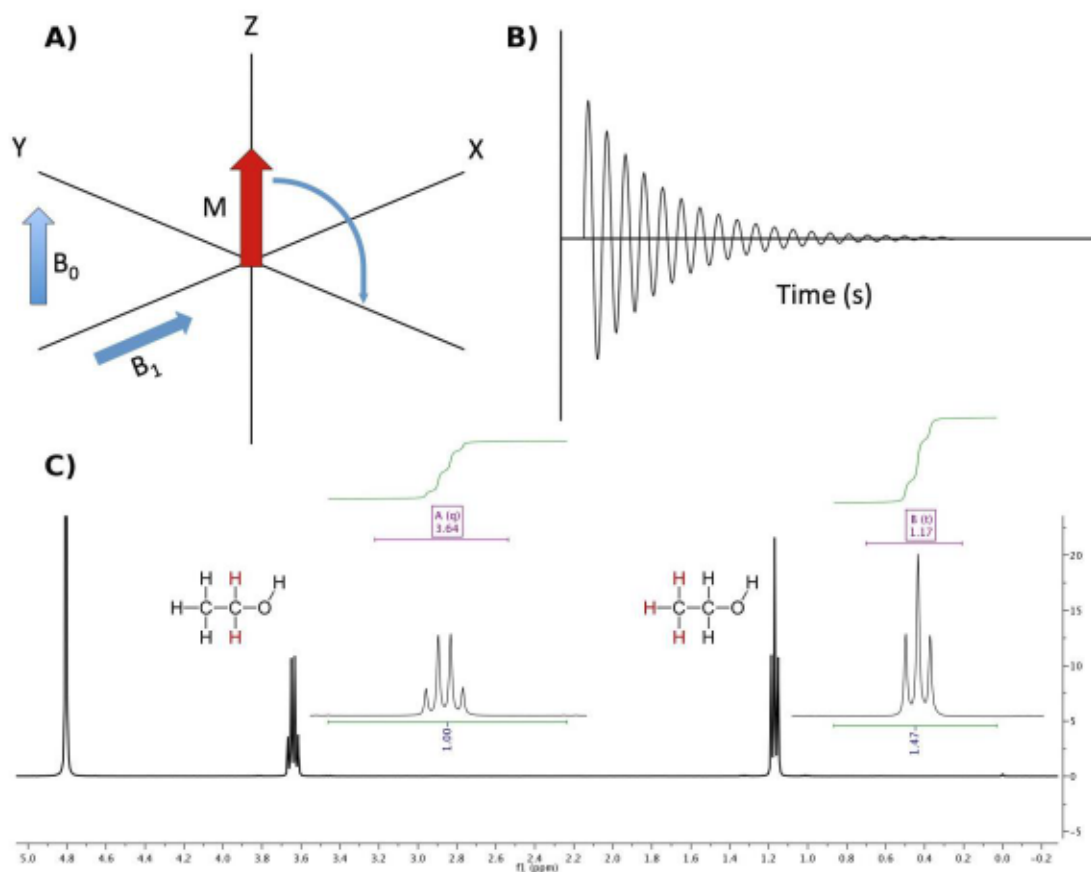
$$\Delta E = -\gamma \hbar B_0 \quad \square\square\square\square$$

$$\frac{N^-}{N^+} = e^{\frac{\Delta E}{kT}} \quad \square\square\square\square$$

Here, N^-/N^+ is the ratio of nuclei in the high energy to low energy spin state, ΔE is the energy difference from equation 2, k is Boltzmann's constant ($1.3806488 \times 10^{-23}$ J/K) and T is temperature in Kelvin. The sum of these magnetic moments yields a net or "bulk" magnetization aligned with the orientation of B_0 - typically referred to as the Z axis (Fig. 2.1A). In the absence of external sources of electromagnetic radiation, magnetic moments will precess (spin) around the orientation of B_0 at a frequency given by equation 2.4. This precession frequency (the Larmor frequency) is equal to the frequency which must be applied to equal the energy differential between spin states (eq. 2.5).

$$\omega_0 = \gamma B_0 \quad \square\square\square\square$$

$$\omega_0 = \Delta E / \hbar \quad \square\square\square\square$$



1. The NMR spectrum shows three main signals: a triplet for the methyl protons (CH_3) at approximately 1.2 ppm, a quartet for the methylene protons (CH_2) at approximately 3.7 ppm, and a singlet for the hydroxyl proton (OH) at approximately 4.7 ppm. The integration values are 1.00, 1.17, and 1.04 respectively, which are consistent with the expected 3:2:1 ratio of protons in ethanol.

2. The chemical shift of the methyl protons is approximately 1.2 ppm, the methylene protons are approximately 3.7 ppm, and the hydroxyl proton is approximately 4.7 ppm. The chemical structure of ethanol is shown above the spectrum.

3. The NMR spectrum is a 1D ^1H NMR spectrum of ethanol. The x-axis represents the chemical shift in ppm, ranging from 5.0 to -0.2. The spectrum shows three main signals: a triplet for the methyl protons (CH_3) at approximately 1.2 ppm, a quartet for the methylene protons (CH_2) at approximately 3.7 ppm, and a singlet for the hydroxyl proton (OH) at approximately 4.7 ppm. The integration values are 1.00, 1.17, and 1.04 respectively, which are consistent with the expected 3:2:1 ratio of protons in ethanol.

When electromagnetic radiation at the frequency ω_0 (radio frequency for the nuclei and field strengths typically used in NMR spectroscopy, 800 MHz for ^1H at a B_0 of 18.8 T) is applied to the population of nuclear spins, bulk magnetization is rotated in the XY plane, frequently in increments of 90 or 180 degrees, which are referred to as 90° or 180°pulses (Fig 2.1A).

$$B_1 = \cos(\alpha t) \quad \square\square\square\square$$

The component of bulk magnetization in the XY plane is given by equation 2.6 where B_1 is the strength of the radio frequency field, ω is the frequency of the pulse ($\omega \neq \omega_0$) and t is its duration. Thus, following a 90°pulse the entirety of the bulk magnetization is in the XY plane, while after a 180°pulse bulk magnetization is aligned along -Z. Following an RF pulse, magnetic moments will resume precession around B_0 at the Larmor frequency, but the oscillation of the bulk magnetization of the XY plane can now be detected as an induced current in the detector coil of an NMR probe, yielding a free induction decay (FID) which is the basis of NMR spectroscopy (Fig. 2.1B). The "decay" in FID refers to the gradual loss of signal due to spin-lattice (T_1) and spin-spin (T_2) relaxation which reduce the component of net magnetization in the XY plane as equilibrium returns to alignment with B_0 (T_1) and due to reduced precession coherence (T_2).

If all nuclei were to precess at exactly the same Larmor frequency, only one frequency would be detected in the FID and no chemical information would be present. However, electrons surrounding a given nucleus have the effect of shielding it from B_0 , and as a result the magnetic field and precession frequency experienced is slightly different for nuclei in differing chemical environments. As such, frequency differences given by equation 2.7 can be detected in the FID, yielding information about the chemical environment experienced by the atomic nuclei present in a sample.

$$\delta = \frac{\nu - \nu_{ref}}{\nu_{ref}} \cdot 10^6 \quad \square\square\square\square$$

Here δ is the chemical shift in parts-per-million (ppm), as the difference between ν and ν_{ref} is measured in Hz while the Larmor frequency ν_{ref} is measured in

MHz. This has the added benefit of freeing the chemical shift scale from dependence on field strength, so that measurements from spectrometers with differing B_0 can be compared. While the FID is in the time domain, chemical shifts are typically measured in the frequency domain, which is obtained by applying a Fourier transform to the FID (Fig. 2.1C).

The energy levels experienced by atomic nuclei are further complicated by J-coupling - changes in resonance frequency due to nearby nuclei with non-zero spin states. The presence of a two adjacent spin 1/2 nuclei will result in two distinct energy levels for each nucleus depending on the spin state of the adjacent nucleus, which results in "splitting" in the resulting NMR spectrum. These splittings, also referred to as multiplicity or fine structure, yield signals whose intensities are defined by Pascal's triangle, which contain additional information about the chemical environment experienced by a given nucleus and thus aid characterization of chemical structures.

2.1.1.2 NMR Pulse Sequences

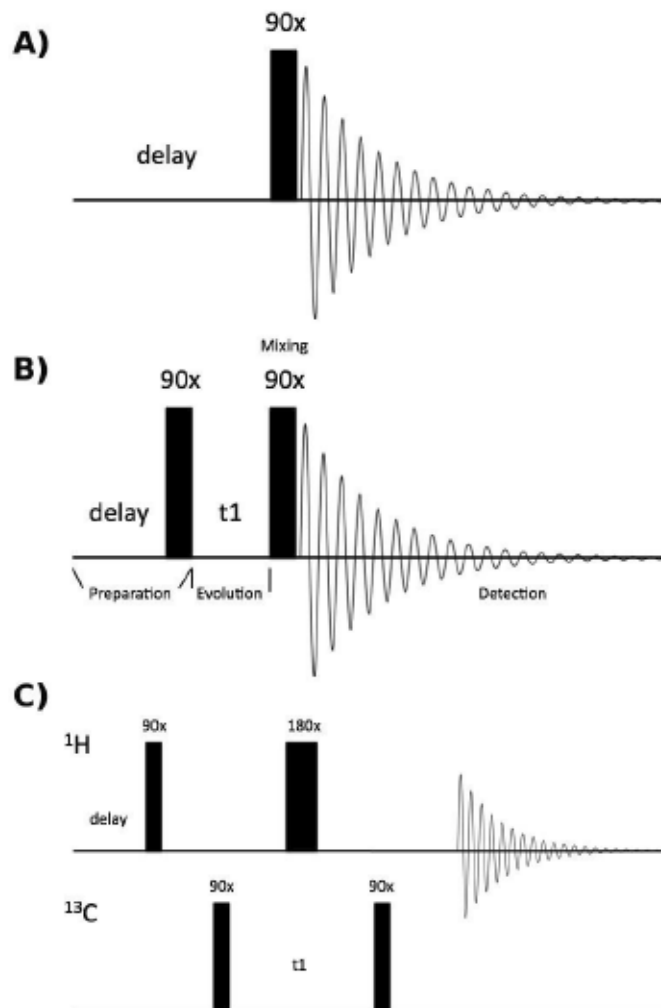
The number, timing, orientation, frequency, and sequence of RF pulses determine the information generated by an NMR experiment. Over the past forty years a large number of these experimental protocols, referred to as pulse sequences, have been developed and many have been implemented in commercial NMR platforms. In metabolomics research, pulse sequences for 1D ^1H , 2D ^1H - ^1H homonuclear, and 2D ^1H - ^{13}C heteronuclear experiments are widely used. Figure 2 shows pulse sequence diagrams for prototypical versions of these experiments.

The 1D "pulse-acquire" sequence (Fig. 2.2A) is arguably the simplest NMR pulse sequence. Here, a 90° pulse is followed by detection of the FID. A delay is typically added before the pulse to allow for T_1 relaxation processes to return bulk magnetization to the Z-axis. This sequence can be repeated in series as a number of "scans" and the resulting signals averaged in order to improve the signal-to-noise ratio. Urine and serum spectra typically are acquired in an aqueous medium, so a low-power, selective presaturation pulse is frequently added during the delay before the pulse to irradiate water and dampen the signal from the water protons. Variants of this experiment used frequently include use of the nuclear Overhauser effect [75] (1D Nuclear Overhauser Effect Spectroscopy, NOESY) for signal enhancement, T_2

filtering (Carr–Purcell–Meiboom–Gill, CPMG) to suppress signals from large molecules, and gradient pulses for improved solvent presaturation [63].

Two-dimensional NMR spectroscopy is arguably the most important spectroscopic technique for the elucidation of novel or unexpected structures [76]. Two-dimensional NMR spectra are unique in that they provide direct evidence for atom connectivity (through-bond connectivity) and spatial arrangements (through-space connectivity). Sequences for generalized ^1H - ^1H homonuclear experiments include preparation, evolution, mixing and detection phases, a schematic for which can be seen in Figure 2.2B. The second dimension is created by collecting multiple increments with varying values for the t_1 delay between pulses, and a second Fourier transform is typically applied to generate data in two frequency domain axes. Correlations are generated by magnetization transfer through bonds by J-coupling (CORrelation Spectroscopy, COSY [77]; Total CORrelation Spectroscopy, TOCSY [78]) or through space by dipolar coupling (2D Nuclear Overhauser Effect Spectroscopy, NOESY; Rotating-frame Nuclear Overhauser Effect Spectroscopy, ROESY). The double quantum filtered (dqf) COSY allows for greater visible fine structure in 2D crosspeaks as it is phase sensitive.

Heteronuclear experiments (Fig. 2.2C) allow for observation of couplings between ^1H and additional nuclei (typically ^{13}C in metabolomics) by pulsing on both ^1H and ^{13}C frequency channels. Frequently applied sequences include Heteronuclear Single Quantum Correlation spectroscopy (HSQC) and Heteronuclear Multiple-Bond quantum Correlation spectroscopy (HMBC) [79]. While 2D NMR experiments generally provide greater signal dispersion and increased chemical information, the main drawback of 2D Fourier transform (FT) NMR spectroscopy is the time required to collect the indirect second dimension at high digital resolution. Additionally, the use of natural abundance ^{13}C (Table 2.1) in heteronuclear NMR also limits sensitivity. As such, 2D NMR are frequently used for compound identification rather than metabolic profiling, though for limited sample experiments or mixtures with a high degree of chemical complexity, variation, or expected novel structures, 2D profiling can provide a substantial advantage (see Chapter 3).



1. The first pulse sequence (A) shows a delay followed by a 90-degree pulse and then a decaying oscillatory signal. This is a simple free induction decay (FID) experiment.

2. The second pulse sequence (B) shows a preparation period with a 90-degree pulse, followed by an evolution period (t1), and then a detection period with a mixing 90-degree pulse and a decaying oscillatory signal. This is a typical heteronuclear correlation experiment.

3. The third pulse sequence (C) shows two channels: ^1H and ^{13}C . The ^1H channel has a delay, a 90-degree pulse, a 180-degree pulse, and a decaying oscillatory signal. The ^{13}C channel has a 90-degree pulse, an evolution period (t1), and another 90-degree pulse. This is a more complex heteronuclear correlation experiment.

2.1.1.3 Sample preparation and acquisition considerations

Nuclear Magnetic Resonance requires relatively little sample preparation, especially as the sample does not come into contact with the instrument. High volume, low concentration samples can be dried under N_2 or freeze-dried to increase solute concentration, but mammalian biological fluids are typically analyzed raw or with the addition of a small amount of HPLC grade water in the case of limited volume samples. 5mm glass walled NMR tubes are frequently used, and require approximately 500 μ L of sample to achieve good magnetic field homogeneity across the sample, but 2.5 and 1.7 mm tubes can be used with smaller volumes. Aqueous samples typically require the addition of D_2O for a signal lock and a chemical reference, such as 3-trimethylsilyl- 2,2,3,3-tetradeuteriopropionate (TSP), to set the chemical shift axis. Dried samples can be resuspended in pure D_2O or another deuterated solvent such as methanol- d_4 (CD_3OD) or chloroform- d ($CDCl_3$). As temperature, pH, and osmolality all affect small molecular chemical shifts, urine samples are typically buffered at pH 7.4 by the addition of phosphate buffer and all experimental spectra are collected at the same temperature, typically 298 or 300K.

^1H NMR spectra are collected over a spectral width of at least 10 ppm, and up to 20 ppm in order to observe the full range of chemical shifts. Relaxation delays (RD) must be long enough to allow adequate T_1 relaxation and optimal $RD=5 \cdot T_1$ can be experimentally identified by the use of inversion recovery experiments. For TOCSY spectra, the amount of magnetization transfer and thus the intensity of crosspeaks is dependent on mixing time, though 60ms and 90ms are typical values and should be kept consistent to compare between spectra. Receiver gain affects signal intensities, but receiver gain values are typically set automatically which, along with concentration variance frequently necessitates normalization as a post-processing step. Fourier transforms, window functions, zero-filling, spectral phasing, baseline correction, and referencing, are typically required before spectra may be compared, though these may be performed on the console instrument.

2.1.2 Mass Spectrometry (MS) in Metabolomics

Mass spectrometry (MS), coupled to either liquid (LC) or gas (GC) chromatography, is also frequently applied to profile the metabolome [64, 65,80]. LC-

and GC-MS both boast high sensitivity to low-concentration analytes, as well as high-resolution chromatographic separation. While these techniques have faced challenges in compound identification, reproducibility and bias towards certain compound functional groups, the rapid pace of technological development in MS has sought to address many of these challenges. Advances in chromatography, including ultra-high-performance liquid chromatography (UPLC) and multidimensional gas chromatography (GC×GC, 3D-GC), have increased the speed and reproducibility of chromatography-coupled MS [81]. Additionally, advances in MS, including high-resolution time-of-flight (ToF) and quadrupole time-of-flight (qTOF) instruments, along with serial ion fragmentation (MS/MS, MSⁿ) have improved resolution, coverage and identification of low-molecular-weight species [82].

			
			
Resolution ($m/\Delta m_{50\%}$)			
			
			

resolution, typically calculated as peak width at half max in fractions of an atomic mass unit, and scan rate, the amount of time required between scan collection. A tradeoff typically exists between mass resolution and scan rate, as can be seen in the typical values for major mass analyzers shown in Table 2.2. As such, time-of-flight (ToF) detectors are frequently used in conjunction with chromatography, while Orbitrap and FT-ICR instruments offer greater resolution when scan rate is less critical, such as with direct injection [87]. Quadrupole instruments, including triple-quadrupoles and quadrupole-time-of-flight (qToF) instruments are generally used to fragment molecular ions to improve detection sensitivity, specificity, and provide additional structural information.

2.2 Dataset preparatory analysis

The basic premise of a metabolomics study is that spectral variables - be they NMR, MS, or generated by another detection technique, can be directly compared to identify qualitative and quantitative changes in the chemical composition of a biological mixture such as urine, serum, or an extract. The two assumptions that underlie this premise are 1) that signal intensity is related to compound abundance, and 2) that the spectral signal(s) arising from a compound are present in the same position in the spectrum. Were these assumptions not generally valid, spectroscopy would be of little use in analytical chemistry and the general principle that the same compound will give the same spectrum underpins the utility of the spectroscopic approach. Despite this, various forms of instrumental and experimental noise affect the spectra of biological mixtures and complicate the direct comparison of these spectra in metabolomics.

Preprocessing techniques have been developed to account for these complications and are now widely used in urinary metabolomics [88-90]. While the adoption and progressive development of preprocessing methodologies has advanced the art of data analysis in metabolomics, preprocessing techniques are still evolving and there is a high degree of heterogeneity in the preprocessing algorithms and the extent to which these techniques are applied. Despite variations in the workflow depending on instrumentation and experimental design, preprocessing in

metabolomics generally includes signal detection or alignment (Fig 2.3A), normalization (Fig 2.3B), and variance stabilization (Fig 2.3C).

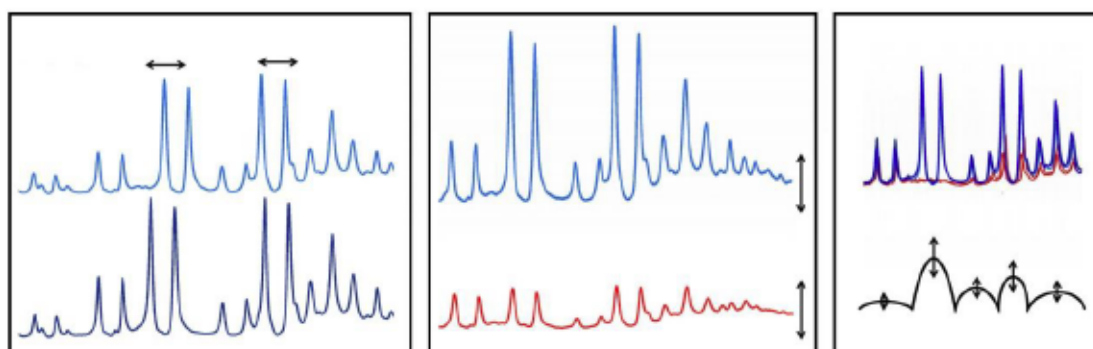
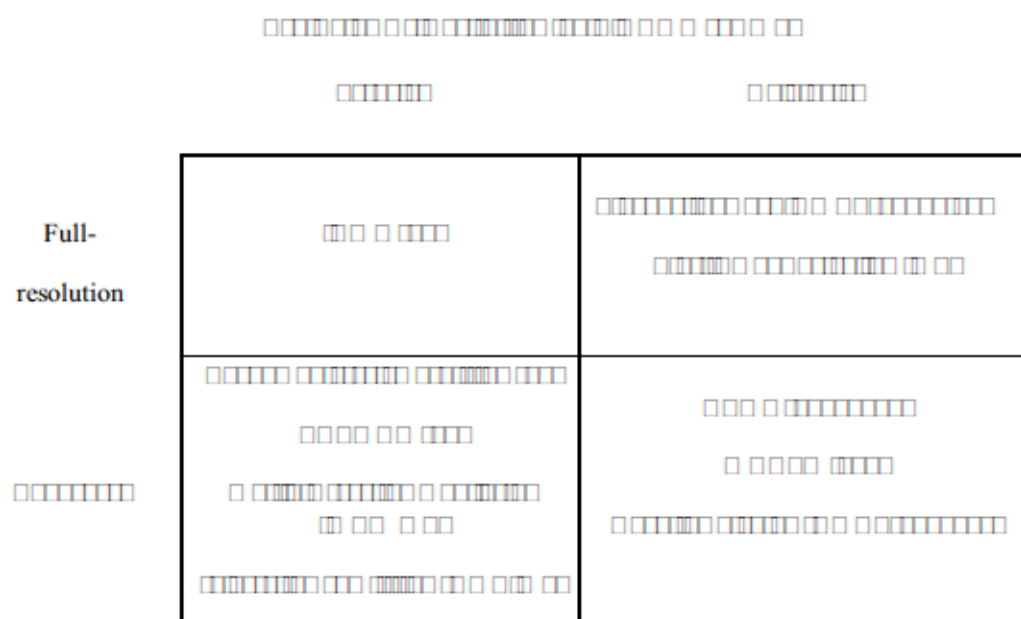


Figure 2.3 illustrates three key steps in metabolomics data processing: (A) Signal alignment, showing two traces (blue and red) with arrows indicating peak shifts; (B) Normalization, showing two traces (blue and red) with a vertical arrow indicating baseline adjustment; and (C) Variance stabilization, showing two traces (blue and red) with arrows indicating peak height adjustments.

2.2.1 Data definition – resolving positional variation

Data definition methods, including peak picking, binning, bucketing, signal alignment, integration, and fitting serve to ensure that measured variables reflect the abundance of the same metabolites across a set of samples. In NMR, “positional noise” is caused by variation in observed chemical shifts due to temperature, pH, osmolality, and the interaction of mixture components. In mass spectrometry, measured m/z values vary within a range related to the resolution of the mass detector, while in LC-MS retention time drift is frequently observed due to unpredictable changes in column conditioning. While many methods have been presented to resolve this positional variation, these methods typically fall into one of four categories combining either targeted or untargeted signal detection with full-resolution or condensed processing (Table 2.3).



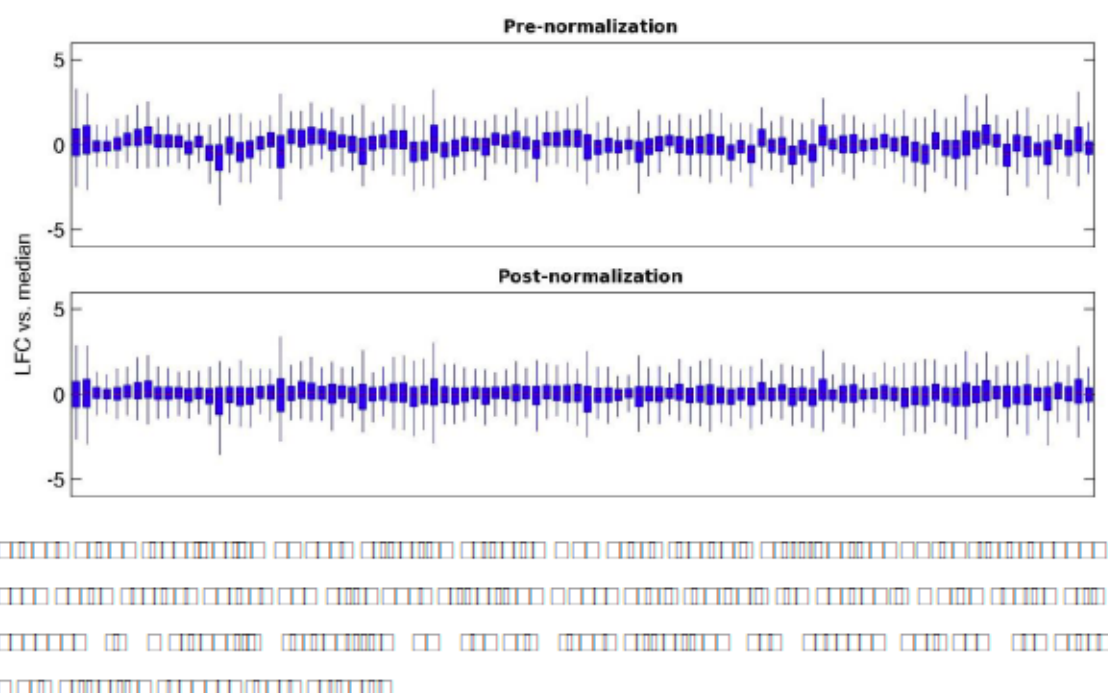
Targeted methods attempt to identify signals corresponding to a set of metabolites defined a priori and transform a set of spectra into a matrix of metabolite abundances, while untargeted methods consider as many spectral features as possible and reserve identification until after statistical analysis, usually for a subset of features. Full-resolution methods seek to preserve spectral line-shapes, and are more frequently used in NMR where lineshapes contain additional structural information, while condensed methods reduce a set of spectral datapoints to a single variable reflecting the level of a given feature. Table 2.3 shows the breakdown of frequently used techniques in both NMR and MS into these categories.

2.2.2 Normalization

Signal intensities in NMR and mass spectra are affected by a number of variables, including biological factors such as sample concentration and instrumental factors such as receiver gain, ion suppression, pulse angle, signal averaging, and more. While single compound measurements frequently use calibration curves to enable absolute quantification, in metabolomics these factors are frequently held constant where possible and comparison between metabolic variables is based on relative changes between measurements. In order to make quantitative comparisons between spectra, some form of data normalization is typically required. These include normalization to a chemical internal standard such as an isotopically labeled

standard or chemical reference, to a biological internal standard such as urinary creatinine, or a statistical method such as total area or probabilistic quotient normalization [88].

Each of these methods have strengths and weaknesses- the chemical standard can be added at a fixed concentration, but cannot correct for variable dilution in the sample. Biological “standards” are valuable so long as they are not affected by the experimental design – the equivalent of using a constitutively expressed “housekeeping” protein as a loading control in a western blot. Statistical normalization typically performs well, and while probabilistic quotient (median fold-change) normalization is more robust to small numbers of highly variable signals, these methods can still be compromised by coordinated changes in large numbers of metabolite abundances. Normalization effects can be evaluated by visualizing histograms of log-fold-change vs. median for all spectra (Figure 2.4) as these indicate samples for which a large percentage of datapoints are relatively high or relatively low compared to the dataset median.



2.2.3 Centering and scaling

Small molecules in human biofluids are present in a range of concentrations spanning multiple orders of magnitude from mM (urea in urine, glucose in serum and

CSF) to sub-nM (hormones). While the limit of detection in most NMR experiments is approximately 10 μM , the concentration range of NMR detectable small molecules still spans approximately 3-4 orders of magnitude [103], and for MS experiments dynamic range can be much greater. Despite the fact that changes in the concentration of all metabolites are potentially of interest, the much higher variance of high concentration compounds can lead to the signals arising from these compounds dominating multivariate models and masking more significant relative changes in low intensity peaks. This is compounded by analytical variance in which instrumental noise increases with signal intensity as well- a heteroscedastic noise structure [104]. Data centering is also typically applied to identify positive and negative associations, and this can be done by centering variables to either a mean or median value.

Recognition of this challenge has lead to the adoption of variance stabilization, or “scaling” methods that attempt to correct for this effect [89]. While unit variance scaling, or autoscaling, (eq 2.8) is frequently applied, this method penalizes variables with multimodal distributions, the identification of which is typically a goal of metabolomics research. As such, the log transformation may be better suited for both clinical metabolomics and the two-class experimental designs commonly used in the field. Equations 2.8, 2.9 and 2.10 show variance stabilization by unit variance scaling (eq. 2.8), log-fold-change vs. median (eq. 2.9) and offset log-fold-change vs. median (eq. 2.10). Here, X is a stabilized version of the spectrum S in a dataset of n spectra with m datapoints, σ_{S_m} is the standard deviation of each column m , and f is a scalar offset.

$$X_{mn} = \frac{(S_{mn} - \bar{S}_m)}{\sigma_{S_m}} \quad (2.8)$$

$$X_{mn} = \log\left(\frac{S_{mn}}{\text{median}(S_m)}\right) \quad (2.9)$$

$$X_{mn} = \log\left(\frac{S_{mn} + f \cdot \sigma_s}{\text{median}(S_m)}\right) \quad (2.10)$$

The effects of variance stabilization can be demonstrated by plotting peak variance against peak intensity rank (Figure 2.5). For un-scaled data, this produces a

characteristic “hockey stick” plot with variance dominated by high intensity variables. While log-fold change without an offset can lead to high variance for low intensity variables, inappropriate scaling of noise can be avoided by adding an offset (f) of approximately 10x the standard deviation of the noise (eq 2.10). By avoiding intensities near zero, the log fold change transformed noise values fall below those of true signals. The choice of the offset does make a difference - choice of offset values that are too low will fail to compensate for noise scaling, while offset values that are too high and approach signal intensities will not stabilize variances and scaling will have no effect. This method has been applied in both metabolomics and microarray analysis and strategies for offset selection reviewed previously [104-106]. Log-scaling also has the added benefit of transforming highly skewed distributions, as many frequentist statistical tests assume approximate normality [107].

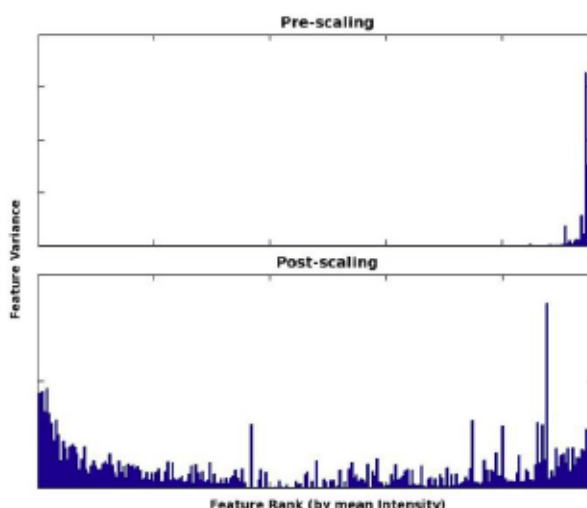


Figure 2.10: Feature variance vs. feature rank (by mean intensity) for a simulated dataset. The top plot shows the variance before scaling, which is dominated by high-intensity features. The bottom plot shows the variance after scaling, which is more uniform across the feature ranks.

2.3 Statistical data analysis

Data preparatory analysis, or preprocessing, transforms a set of biological NMR or mass spectra into a data matrix X_{np} in which N rows represent the metabolic profile of a single sample n and P columns represent the metabolic or spectroscopic

variable p . Relationships between samples and variables in X can then be analyzed using exploratory or "unsupervised" statistical methods. Alternately, relationships between X and a(n) independent variable(s) Y can be analyzed with "supervised" methods. In metabolomics research, Y is typically a biological response variable which can take the form of dummy variables such as $[0, 1]$ for control and affected samples, integers $[1, N]$ for classes such as individual, location, or response group, or quantitative response data $[-\infty, \infty]$ with data from an assay response or biological measurement. These supervised methods are typically used for regression or classification, predicting unknown Y on the basis of X , or for feature selection, identifying variables p significantly associated with Y .

2.3.1 Methods for unsupervised data analysis

Unsupervised methods attempt to identify relationships between variables and uncover sources of variation that may or may not be captured by the independent variables. Methods in this family do not make use of a Y vector, and as such there is typically no objective function or fit metric to evaluate their performance. Many unsupervised methods depend on the choice of a distance function that measures the similarity (or distance) between vectors X_n or X_p . In metabolomics research, the correlation coefficient is frequently used as a metric of similarity, and in statistical spectroscopy correlations are used to preprocess spectral datasets, identify biomarkers, and interpret metabolic co-regulation.

$$r_{mn} = \frac{\sum_{i=1}^j (S_{im} - \bar{S}_m)(S_{in} - \bar{S}_n)}{(j-1) \bullet \sigma_{S_m} \sigma_{S_n}} \quad (2.11)$$

$$d_{mn} = 1 - \frac{\sum_{i=1}^j (S_{im} - \bar{S}_m)(S_{in} - \bar{S}_n)}{(j-1) \bullet \sigma_{S_m} \sigma_{S_n}} \quad (2.12)$$

$$R = \frac{1}{j-1} X_1^T X_2 \quad (2.13)$$

Equation 2.11 shows how the correlation coefficient r_{mn} is calculated for two vectors x_{im} and x_{in} where m and n are column indices of X_p . In the context of

statistical spectroscopy, r_{mn} is the correlation coefficient between the intensities x of signal m and signal n across all spectra i through j , where these signals have mean intensity $\bar{x}_m$ and $\bar{x}_n$ and standard deviations σ_{x_m} and σ_{x_n} . When signals m and n arise from the same chemical structure, r_{mn} will tend to approach unity. When signals m and n arise from different chemical structures, the magnitude of r_{mn} will depend on the extent to which the metabolites giving rise to signals m and n are subject to the same metabolic perturbations (e.g. pathway connectivity) [108, 109]. As a result, the correlation coefficient r_{mn} can be used to assess both structural and pathway connectivity, though in the case of strong metabolic perturbations or poor signal resolution there may be ambiguity between the underlying origin of correlations. When x_{im} and x_{in} are vectors of intensities r_{mn} is the Pearson correlation coefficient, and when x_{im} and x_{in} are vectors of rank orders r_{mn} is the Spearman correlation coefficient. Correlation coefficients $[-1, 1]$ can also be converted to correlation distances d_{mn} $[0, 2]$ by equation 2.12.

Equation 2.13 is mathematically equivalent to equation 2.11, but the matrix notation emphasizes differences between datasets rather than signals. Here, $\mathbf{R}$ is a matrix of correlation coefficients between two autoscaled datasets $\mathbf{X}_1$ and $\mathbf{X}_2$, each of which is composed of j observations. When $\mathbf{X}_1$ and $\mathbf{X}_2$ take the form X_{np} , $\mathbf{R}$ will be the correlation matrix R_{pp} between all metabolic variables, and when $\mathbf{X}_1$ and $\mathbf{X}_2$ take the form X_{pn} $\mathbf{R}$ will be the correlation matrix R_{nn} between all samples. When $\mathbf{X}_1$ and $\mathbf{X}_2$ are the same matrix, $\mathbf{R}$ is homo-spectroscopic. When $\mathbf{X}_1$ and $\mathbf{X}_2$ are matrices derived from different analytical techniques, such as heteronuclear NMR or mass spectrometry, $\mathbf{R}$ is hetero-spectroscopic. The framework of statistical spectroscopy is flexibly extended in cases in which $\mathbf{X}_2$ is data derived from genomics, proteomics, or transcriptomics as well. In these cases $\mathbf{R}$ will reflect metabolite – non-metabolite associations as opposed to the traditional metabolite – metabolite associations.

Underlying structure in X_{np} can then be identified by applying cluster analysis to distance matrix D_{pp} , D_{nn} , or both. While there are many techniques for cluster analysis, including k-means clustering, density or distribution based clustering, and related dimensionality reduction methods that can be applied to $\mathbf{D}$ such as multidimensional scaling, hierarchical clustering is among the more frequently used clustering tools in metabolomics research [110, 111].

$$d_{ab} = \frac{1}{|a| \bullet |b|} \sum \sum d_{mn} \quad (2.14)$$

Hierarchical cluster analysis (HCA) depends on the sequential (typically greedy) grouping of variables into clusters on the basis of distance. HCA can either build up clusters from individual variables (agglomerative HCA), or split clusters from the full set of variables (divisive HCA). The sequence of splits is then recorded in a tree, or dendrogram. The results of HCA depend on the choice of distance metric, such as the Pearson or Spearman correlation, and the linkage method, which determines how distances between clusters are calculated. The unweighted pair group method with arithmetic mean (UPGMA) algorithm for HCA applies agglomerative clustering with average-linkage, in which the distance between clusters is the average of the distance between all variables in the given clusters (eq. 2.14). Equation 2.14 shows the average-linkage distance between clusters a and b given distances d_{mn} . Other linkage methods include single, complete, and weighted linkages.

Alternately, Principal Component Analysis (PCA) [112] can project X into a lower dimensionality space to identify major sources of variance in X . When $p < n$, PCA is typically calculated based on the singular value decomposition of X (eq. 2.15). Here, U and V are eigenvector matrices of the covariance matrices $X^T X$ and XX^T respectively and S is a rectangular diagonal matrix. When X is autoscaled, $X^T X$ and XX^T are the Pearson correlation matrices R_{pp} and R_{nn} . While in singular value decomposition, elements of the diagonal matrix S go from low to high, in PCA the direction of S is reversed so that elements go from high to low, and U and V are reordered to match S . In the context of PCA, the coefficients of U are referred to as “loadings”, and the matrix T (eq. 2.16) as “scores”.

$$X = USV \quad (2.15)$$

$$T = VS \quad (2.16)$$

However, when $p > n$, the covariance matrix $X^T X$ may be singular and as a result singular value decomposition is not used. In this case, an iterative algorithm such as non-linear iterative partial least squares (NIPALS) [112] is typically used to calculate U and T . In this case only a subset of the eigenvectors of U may be used and as a result there is a residual matrix E such that

$$X = TU^T + E \quad (2.17)$$

The scores matrix T is then an $(n \times c)$ matrix and the loadings matrix U is a $(p \times c)$ matrix where c is the number of principal components used to reconstruct X to a certain level of precision (frequently 95% percent of variance). As such, for a given principal component (PC) c , there is a set of n scores and p loadings. In the context of metabolomics, the loadings p describe the metabolic pattern which give rise to the clustering pattern seen in the distances between the n scores. As such, PCA is used to identify the major sources of variation present in a dataset and the relationships between samples and metabolic variables induced by these sources.

2.3.2 Methods for supervised data analysis

Unlike unsupervised techniques, supervised methods make use of a biological response vector Y that can be categorical or continuous. These supervised methods perform two main functions in metabolomics data analysis - classification and feature selection. In the context of metabolomics in medicine, classification is related to diagnosis- for example, patients can be classified as either "healthy" or "diseased" (binary categorical Y), as a grade of disease severity (multiple categorical Y), or on a scale (continuous Y) on the basis of detected metabolite levels. Feature selection then provides a tool for "biomarker discovery"- the goal of which is to identify metabolites that are most associated with biological response captured in Y . Like unsupervised methods, supervised methods can be either univariate (i.e. analysis of a single variable at a time) or multivariate (analysis of multiple variables at once).

2.3.2.1 Univariate tests

Univariate tests can be applied to categorical and continuous Y . If Y is binary categorical, the t-statistic can be used to identify differences between the means of the two groups. Here, the t-statistic t is calculated from group means $\bar{x}_a, \bar{x}_b$ and variances $\frac{2}{x_a}, \frac{2}{x_b}$ where a and b are groups of size n_a, n_b with differing values Y . Typically, when Y is binary categorical it is represented by dummy variables $Y = 0$ and $Y = 1$

with a representing the group $Y = 1$ and b representing the group $Y = 0$. Thus, if t is positive, x is positively associated with Y and if t is negative x is negatively associated with Y .

$$t = \frac{(\bar{x}_a - \bar{x}_b)}{\sqrt{\frac{\sigma_{x_a}^2}{n_a} + \frac{\sigma_{x_b}^2}{n_b}}} \quad (2.19)$$

If Y is multiple categorical, analysis of variance (ANOVA) testing can be used to identify differences between group means across multiple groups. The F-statistic calculated in ANOVA compares between-group mean squares (MS_B) with within-group mean squares (MS_W), which are calculated from group means $\bar{x}_a$ and variances $\sigma_{x_a}^2$ while taking into account the number of groups a , the number of observations in each group n_a and the total number of observations N . When $a = 2$, ANOVA and Student's t-test are equivalent with $F = t^2$.

$$MS_B = \frac{\sum_a n_a (\bar{x}_a - \bar{x})^2}{a - 1} \quad (2.20)$$

$$MS_W = \frac{\sum_a \sigma_{x_a}^2 \cdot (n_a - 1)}{N - a} \quad (2.21)$$

$$F = MS_B / MS_W \quad (2.22)$$

Finally, if Y is continuous, correlation coefficients can be used in a supervised manner by taking the correlation between $\mathbf{X}$ and Y as in equation 2.23. As previously, $\mathbf{R}$ is a matrix - or vector if Y is $(n \times 1)$ - of correlation coefficients between datasets $\mathbf{X}$ and Y , each of which is composed of j observations and where both are autoscaled. Regression analysis can also be applied to single variables as well as multiple variables, as discussed below.

$$R = \frac{1}{j - 1} \mathbf{X}^T \mathbf{Y} \quad (2.23)$$

2.3.2.2 Multivariate supervised methods

Like univariate tests, multivariate supervised methods can be applied to both categorical and continuous Y . When $p < n$, generalized linear model regression techniques can be used, which estimate a set p of regression coefficients, or predictors, β based on equation 2.24. Generalized linear models are referred to as “generalized” because the choice of model is based on the distribution of Y . When Y is continuous, the corresponding model is ordinary least squares (OLS) regression, and predictions, denoted by $\hat{Y}$ by equation 2.25, where k is a constant intercept value. When Y is binary categorical, the corresponding model is multiple logistic regression, where $\hat{Y}$

$$\beta = (X^T X)^{-1} X^T Y \quad (2.24)$$

$$\hat{Y} = X\beta + k$$

$$\hat{Y} = \frac{1}{1 + e^{X\beta + k}}$$

As with singular value decomposition, generalized linear models depend on the inversion of the $X^T X$ covariance matrix and so are limited to cases where $p < n$. When $p > n$, iterative methods such as NIPALS [112] or SIMPLS [113] can be used to calculate β by Partial Least Squares (PLS). Like PCA, PLS calculates a number of components c where $c < p$ which are used to reconstruct X . The choice c must be optimized based on fit parameters (see 2.2.3), typically using cross-validation. Like generalized linear models, PLS can be applied to both continuous Y (termed partial-least-squares regression or PLS-R) and categorical Y (termed partial-least-squares discriminant analysis or PLS-DA) where predictions $\hat{Y}$

2.3.3 Evaluating supervised models

Interpreting feature selection methods and classification models typically depends on evaluating quality metrics such as significance, classification accuracy, and goodness-of-fit. Especially in the case of supervised methods where $p > n$, such as PLS, overfitting is concern and must be guarded against by various cross-validation and permutation testing strategies. While the methods discussed here are linear, parametric models, the application of non-linear or non-parametric models and machine learning techniques requires even greater vigilance in evaluating model performance.

2.3.3.1 Evaluating feature significance

When using univariate tests such as Student's t-test, ANOVA, and Pearson's correlation coefficient, significance levels can typically be ascribed to test statistics such as t-values or F-values by referring to corresponding t- and F-distributions. Here, p-values refer to the probability of observing a test statistic by chance given the degrees of freedom and the resulting probability density function. While the significance threshold α is typically set at $\alpha = 0.05$, since p-values refer to the probability of observing a test statistic by chance the occurrence of false-positive associations is expected to increase when multiple features (hypotheses) are evaluated by univariate methods. To correct for this multiple hypothesis testing, a correction is typically applied to the p-values calculated after multiple tests by the number of measured variables m . Methods for estimating a revised threshold p include Bonferroni correction (eq. 2.27), Šidák correction (eq. 2.28), and the Benjamini-Hochberg false discovery rate (FDR, eq. 2.29) [114, 115] where d is the largest index of sorted p-values for which equation 2.29 holds true.

$$p = \frac{\alpha}{m} \quad \square \square \square \square \square$$

$$p = 1 - (1 - \alpha)^{1/m} \quad \square \square \square \square \square$$

$$p \leq \frac{d}{m} \bullet \alpha \quad \square \square \square \square \square$$

2.3.3.2 Evaluating classification performance

Given a binary categorical Y and a binary categorical $\hat{Y}$ resulting from predictive modeling, it is possible to have four possible classification outcomes as shown in the confusion matrix in Table 2.4. These are referred to as true-positives, true-negatives, false-positives and false-negatives. Clearly, the goal of any classification method is to maximize the true predictions and minimize the false predictions. However, total positives and total negatives tend to be inversely related as they are based on the selection of a classification threshold. As a result, classification accuracy tends to be reported in terms of sensitivity (eq. 2.30) and specificity (eq. 2.31), which are percentage values.

	$Y = 0$	$Y = 1$
$\hat{Y} = 1$	n_{11}	n_{10}
$\hat{Y} = 0$	n_{01}	n_{00}

$$specificity = \frac{n_{TN}}{n_{TN} + n_{FP}} \quad (2.31)$$

In the case of overfitting, however, classification can be highly accurate for the dataset used to construct a model β , but essentially random for an independent dataset. To avoid this, datasets are frequently split into training and test sets where the training set is used to construct a predictive model and the test set is used to

evaluate the classification performance of the model. Training and test sets can also be defined iteratively, with multiple models constructed based on different permutations of training and test set assignment. These methods include leave-one-out cross-validation (LOOCV), in which a test set of $n = 1$ is used to evaluate model performance for all $n \in N$, k-fold cross validation in which k test-sets of size $n = k/N$ are used, and k-fold Monte-Carlo cross validation (MCCV) in which test-sets of size $n = k/N$ are used but membership in the test and train set is randomly permuted such that a given sample may be assigned to the train set multiple times [116, 117].

$$R^2 = 1 - \frac{\sum_{i \in \text{train}} (y_i - \hat{y}_i)^2}{\sum_{i \in \text{train}} (y_i - \bar{y})^2} \quad (2.32)$$

2.4 Compound identification and verification

statistical spectroscopy and spectral databases have reduced the complexity of the assignment workflow.

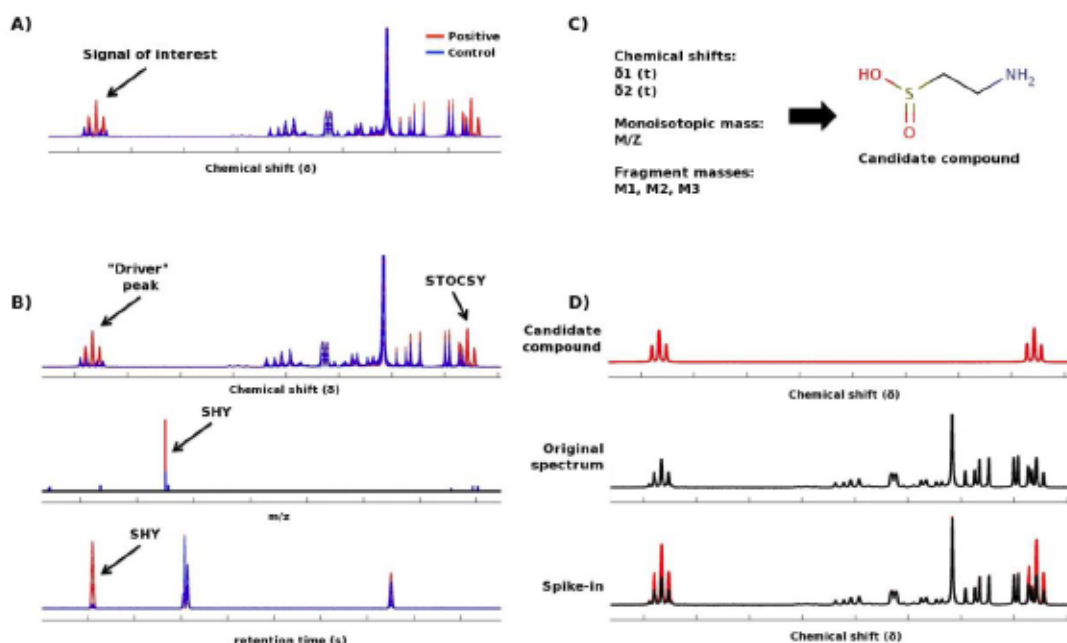


Figure 2.6 illustrates a four-stage approach to biomarker and compound identification. Panel A shows NMR spectra with a "Signal of Interest" highlighted. Panel B shows STOCSY and SHY correlations between peaks. Panel C shows the identification of a candidate compound (sulfonamide) based on chemical shifts, mass, and fragmentation. Panel D shows the original spectrum with a candidate compound spectrum and a spike-in for comparison.

In the case of unknown or unexpected compounds that are not novel compounds, a four-stage approach to biomarker and compound identification, represented schematically in Figure 2.6, may be used to assign these signals. In the first stage, spectral features are correlated with a given biological state or response by applying pattern recognition analysis to individual datasets generated by NMR spectroscopy or mass spectrometry (Fig 2.6A). Subsequently, STOCSY, SHY, and other statistical spectroscopy tools are then used to associate masses, fragments, and NMR spectral features (Fig 2.6B). In a third stage, the set of correlated features, including mass, retention time, fragmentation patterns, chemical shifts, and through-

bond correlations are searched in-silico for matching compounds using metabolite libraries and reference spectral data (Fig 2.6C).

While individual MS m/z values and single NMR chemical shifts are often non-specific, only a small number of structural isomers will give rise to similar NMR spectra and MS/MS fragments. To conclusively verify database matches, metabolite candidates of interest may be confirmed by direct spiking of pure chemical standards into the biological matrix for NMR, or comparison of parent masses, retention times, and fragmentation patterns of pure standards by mass spectrometric analysis (Fig 2.6D). Several web-based computational tools support biomarker and compound identification, including Metlin, XCMS Online, the BMRB and HMDB databases, and MetaboAnalyst, [69, 99, 121-123]. In cases where database matches or chemical standards are not available, 2D NMR can be used to conclusively identify structural isomers [120].

CHAPTER 3

2D NMR METABOLIC PROFILING IN BIOCHEMICAL GENETICS

3.1 Introduction and aims

Nuclear Magnetic Resonance (NMR) spectroscopy is a powerful analytical tool that provides detailed information about the chemical structure and dynamics of molecules. Its ability to analyze essentially all types of molecules at atomic resolution. Recently it has been applied extensively to metabolic profiling in studies of human and animal biofluids [124, 125], cellular biochemistry [126, 127], and non-mammalian chemical ecology [128, 129]. While one-dimensional (1D) ^1H NMR spectra have been used in the majority of NMR-based metabolic profiling studies, resonance overlap and lack of structural correlations can limit the utility of the 1D approach. In contrast, powerful multidimensional NMR methods are routinely used in structural biology and natural products studies but have been used much less frequently for metabolomics. Two dimensional (2D) homonuclear (such as TOCSY and COSY) and heteronuclear (such as HSQC and HMBC) spectra offer complementary information to 1D spectra. 2D NMR spectra provide a more comprehensive view of the molecular structure and dynamics, allowing for the identification of correlations between different nuclei in a molecule. This is particularly useful for identifying metabolites in complex mixtures, such as biofluids or cell extracts. The aim of this chapter is to provide a comprehensive overview of 2D NMR spectroscopy, its applications in metabolic profiling, and the challenges associated with its use in metabolomics.

The chapter is organized as follows. Section 3.2 discusses the basic principles of NMR spectroscopy, including the concepts of chemical shift, coupling constants, and relaxation times. Section 3.3 provides a detailed overview of 2D NMR spectroscopy, focusing on the most commonly used techniques: TOCSY, COSY, HSQC, and HMBC. Section 3.4 discusses the applications of 2D NMR spectroscopy in metabolic profiling, highlighting the strengths and limitations of each technique. Section 3.5 discusses the challenges associated with 2D NMR spectroscopy, including the need for specialized equipment and the complexity of data analysis. Finally, Section 3.6 provides a summary of the chapter and outlines the future prospects of 2D NMR spectroscopy in metabolomics.

The *a priori* information is used to guide the alignment process. This information can be obtained from various sources, such as chemical shift databases, reference spectra, or experimental data. The *a priori* information is used to constrain the alignment process, ensuring that the resulting spectra are aligned with the expected chemical shifts. This is achieved by incorporating the *a priori* information into the alignment algorithm, which then adjusts the alignment parameters to match the expected chemical shifts. The *a priori* information is also used to identify and correct any misalignments or distortions in the spectra. This is done by comparing the aligned spectra with the *a priori* information and identifying any discrepancies. These discrepancies are then used to adjust the alignment parameters, ensuring that the spectra are properly aligned.

The *a priori* information is used to guide the alignment process. This information can be obtained from various sources, such as chemical shift databases, reference spectra, or experimental data. The *a priori* information is used to constrain the alignment process, ensuring that the resulting spectra are aligned with the expected chemical shifts. This is achieved by incorporating the *a priori* information into the alignment algorithm, which then adjusts the alignment parameters to match the expected chemical shifts. The *a priori* information is also used to identify and correct any misalignments or distortions in the spectra. This is done by comparing the aligned spectra with the *a priori* information and identifying any discrepancies. These discrepancies are then used to adjust the alignment parameters, ensuring that the spectra are properly aligned.

3.2 Results – Method development

This section presents the theoretical framework of the new methods developed for preprocessing 2D ^1H - ^1H NMR spectra. These include guided alignment, presented as Hierarchical Alignment of Two-Dimensional Spectra (HATS), automated crosspeak identification and integration, presented as multivariate Differential Analysis of NMR Spectra (mvaDANS), and adapted chemometric

methods including normalization, scaling and pattern recognition for analysis of this data [94, 102].

3.2.1 Automated crosspeak identification and integration in 2D NMR

Spectral segments are detected from a representative pseudo-spectrum- here, an average spectrum calculated from the mean along the third dimension of the spectral stack (ω_1, ω_2, N). To identify segments of the pseudo-spectrum corresponding to crosspeak regions, a local noise surface is calculated using the AUTOPSY [134] method (equations 3.1-3.4). AUTOPSY calculates a point-wise estimation of underlying noise as follows.

$$d_x = \min \left(\sqrt{\frac{1}{n} \sum_{i=1}^n (\delta_{xi} - \bar{\delta}_x)^2} \right) \quad (3.1)$$

$$d_y = \min \left(\sqrt{\frac{1}{n} \sum_{i=1}^n (\delta_{iy} - \bar{\delta}_y)^2} \right) \quad (3.2)$$

$$b = \min(d_x, d_y) \quad (3.3)$$

$$p_{xy} = \sqrt{(\sqrt{d_x^2 - b^2}) + (\sqrt{d_y^2 - b^2}) + b^2} \quad (3.4)$$

Here, d_x and d_y are noise values for each slice x and y in the direct and indirect dimensions of the 2D NMR spectrum, respectively. The noise value for a given slice δ_x or δ_y is calculated by splitting the slice up into 16 regions of size $n=F1/16$ for the direct dimension and $n=F2/16$ for the indirect dimension. The noise for that slice is then the standard deviation of the region with the minimum standard deviation as shown in equations 3.1 and 3.2. Each point p in the spectrum is intersected by two slices x and y whose noise values are d_x and d_y . The noise at point p_{xy} combines the noise in slice x with the noise in slice y and with a baseline noise value b as in equations 3.3 and 3.4. This method of calculating noise is very well suited to the noise

structure of 2D NMR data, which is characterized by vertical noise such as T_1 and water noise.

Once the noise surface p_{xy} has been calculated, peaks can be detected by identifying local maxima above a certain multiple of the noise surface. While the exact threshold can vary and is a parameter that must be modified for appropriate processing of any individual dataset, values of 5-10 detect the large majority of cross-peaks. Segments are then defined as regions adjacent to detected peaks.

Initial spectral segments are identified as regions within 0.04 ppm of a peak in the ^1H dimension of homonuclear spectra. In the case of multiplicity, local maxima are grouped together into a single initial segment. Each segment is then labeled with a unique integer value and iterated minimum and maximum filtering expand these initial segments. Min/max filtering has the effect of growing the segments until they encounter another spectral segment or until they reach a maximum frequency range of approximately 0.2 ppm for ^1H . This procedure creates spectral segments with non-rectangular bounds that are bounded by nearby resonances yet avoid close cropping of cross-peak line shapes. These segments are then integrated using a trapezoidal numerical integration.

3.2.2 Hierarchical Alignment of Two-Dimensional Spectra (HATS)

The HATS alignment method presented here makes use of a guide tree to structure the alignment process. While many existing 1D alignment methods have attempted to align each spectrum in a dataset to a single consensus spectrum [90, 92, 135], the quality of these “star alignments” can suffer if there is wide chemical diversity between spectra, as any individual spectrum chosen as a star may not include the full complement of metabolites present in the other datasets. The use of guide trees was a significant advance for multiple sequence alignment in sequence bioinformatics, and methods such as CLUSTAL-W [136] have enjoyed widespread popularity in nucleic acid and protein sequence alignment. The use of a guide tree should be helpful in the alignment of 2D spectra as correlation spectra are often collected when significant metabolic diversity is expected and the assumptions underlying star alignments may not be appropriate.

The UPGMA agglomerative hierarchical clustering method was used to construct the initial guide tree for the subsequent pairwise alignment of spectral

groups. The squared distance of Pearson's correlation coefficient was applied here as a metric for clustering (equations 3.5 and 3.6), and the resulting dendrogram specifies the order in which alignment of the spectra proceeds.

$$r_{mn} = \frac{\sum_{i=1}^j (S_{im} - \bar{S}_m)(S_{in} - \bar{S}_n)}{(j-1) \cdot \sigma_{S_m} \sigma_{S_n}} \quad (3.5)$$

$$d_{mn} = 1 - r_{mn}^2 \quad (3.6)$$

Here, r_{mn} is the correlation coefficient between the reshaped vectors (1 x F1•F2) of spectra S_m and S_n with mean values $\bar{S}_m$ and $\bar{S}_n$, standard deviations σ_{S_m} and σ_{S_n} , and d_{mn} is a distance version of the correlation value. At each branch point in the dendrogram, the set of spectra specified by the branches are aligned. For the alignment illustrated in Figure 3.1, the two red spectra are aligned first, followed by the two blue spectra, then the two blue spectra are aligned with the black spectrum, and finally the two red spectra are aligned with the three blue and black spectra. If there are multiple spectra in an alignment group, positional adjustments are made equally to all spectra in the group.

The alignment at each branch point of the guide tree consists of two parts: a global alignment in which all spectral datapoints are shifted equally and a local alignment in which specific crosspeak regions are shifted individually. The global alignment step seeks to correct defects such as mis-calibration or referencing, which produce chemical shift differences across the entire spectrum. One of the groups of spectra is passed over the other group as a mask in a point-wise manner in both F1 and F2 dimensions, and at each positional step x and y the alignment quality is expressed as r_{xy} , the mean correlation between groups (equation 3.7).

$$r_{xy} = \frac{1}{a \cdot b} \sum_{m=1}^a \sum_{n=1}^b r_{mn} \quad (3.7)$$

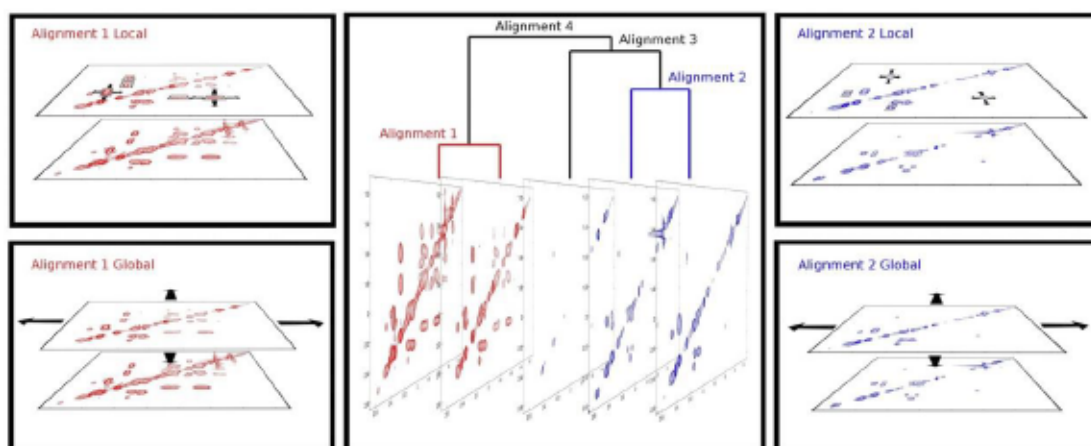


Figure 3.1. Overall strategy of Hierarchical Alignment of Two-Dimensional Spectra. Before alignment, the HATS algorithm produces a guide tree by the UPGMA agglomerative hierarchical clustering algorithm from a matrix of pairwise correlation distances. This guide tree is used to structure the alignment process. At each intersection of branches in the guide tree, the spectra specified by the two branches are aligned globally and locally. For the alignment tree shown here, the two red spectra are aligned first, followed by the two blue spectra. The two blue spectra are then aligned with the black spectrum, and finally the three blue and black spectra are aligned with the red spectra.

Here, a is the number of spectra in group 1 and b is the number in group 2, and r_{mn} is the correlation coefficient between the reshaped vectors of spectrum m in group 1 and spectrum n in group 2. Spectra in the mask group are shifted to the points m and n producing the maximal mean correlation which is identified by a simple gradient ascent algorithm. While global alignment can resolve factors affecting all resonances equally, it is not capable of resolving differences in specific resonances caused by conditions such as pH, temperature, metal ion concentration, and osmolality. To resolve local chemical shift variation, crosspeak regions must be aligned individually. To identify segments of the spectrum corresponding to crosspeak regions, a local noise surface is calculated for each spectrum, as outlined in the previous section [134]. This procedure creates spectral segments that are bounded by nearby resonances yet avoid close cropping of crosspeak lineshapes. Each crosspeak segment is then aligned individually by maximizing the mean cross-correlation of crosspeak segments between alignment groups in the same way as the global alignment but treating each crosspeak segment independently (equation 3.7, where r_{xy} is the lineshape vector for a single crosspeak region rather than the entire spectrum)

3.2.3 Chemometrics applied to 2D NMR

3.2.3.1 Unfolding and refolding of full resolution data

Following alignment, Principal Component Analysis (PCA) and Partial Least Squares Discriminant Analysis (PLS-DA) are applied for full resolution spectral pattern recognition. Each principal component consists of N scores and P loadings, where scores identify relationships between observations (here, TOCSY spectra collected from individual samples) and loadings represent the pattern of features (here, 2D NMR chemical shifts) underlying the relationships identified by scores. The NIPALS algorithm [112] was applied to calculate principal components given the impracticality of using singular value decomposition methods for $P \gg N$ matrices. While PCA avoids the problem of over-fitting by taking an unsupervised approach to pattern recognition, it is not optimized for class separation. Partial least squares methods such as PLS-DA are often the supervised method of choice for the analysis of metabolic differentials. The SIMPLS [113] PLS-DA algorithm was used to help clarify the metabolic differential identified by PCA while avoiding over-fitting by using 4-fold cross-validation and reporting performance statistics for the PLS-DA model.

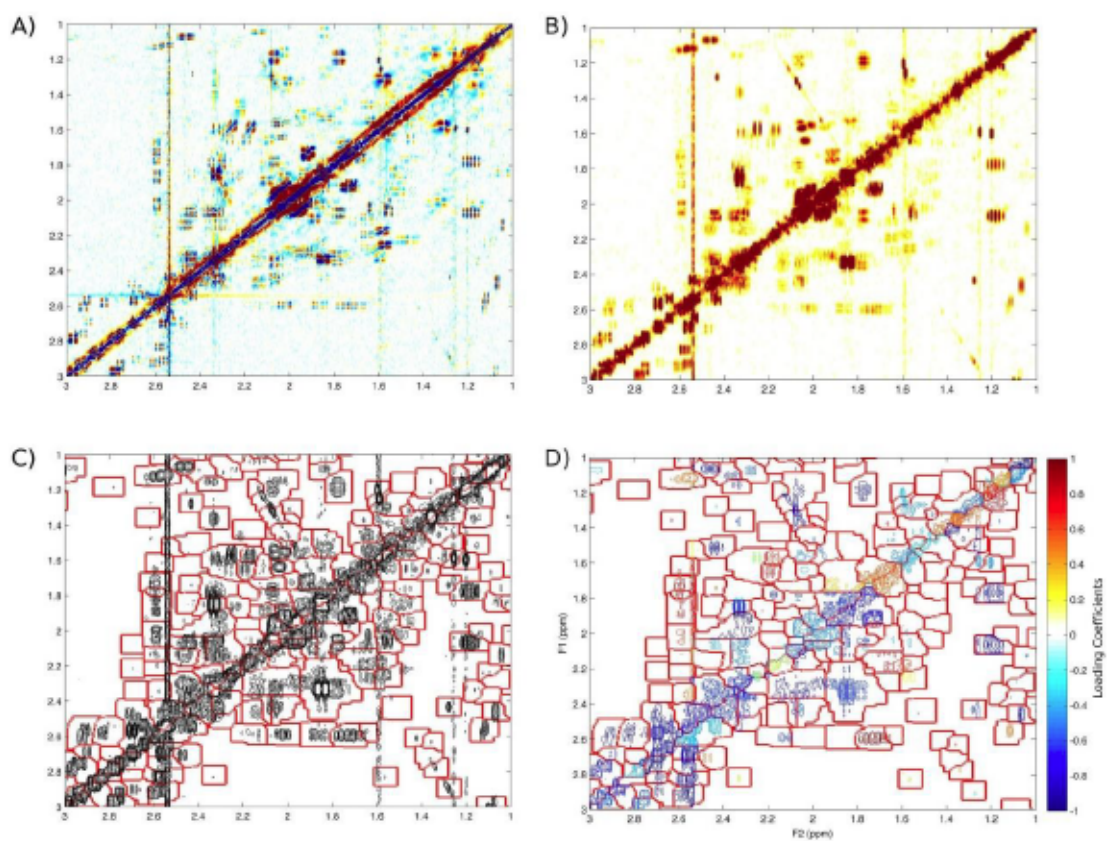
Each of the 4096×1024 TOCSY data matrices were reshaped to a 1×4194304 vector before the application of PCA or PLS. This is a necessary step because it creates an $N \times P$ matrix, where $P = F1 \bullet F2$ data points on which pattern recognition algorithms can operate. To visualize the resulting $1 \times P$ loadings, the loadings vectors for each principal component were reshaped back into a 4096×1024 data matrix. Contours were defined by multiplying the univariate scaled loadings by the standard deviation along the third dimension of the original TOCSY stack (ω_1, ω_2, N), a procedure known as back-scaling or back-projection [137]. Contour colors were then defined by the loading values themselves divided by the maximum absolute loading coefficient to produce a range of -1 to 1. This strategy produces loading spectra which have cross-peak line shapes similar to traditional TOCSY spectra but include sign and intensity information related to the relationships between spectra specified by the corresponding PCA scores. PLS-DA results are presented in a similar manner, but instead of specifying scores and loadings only the predictive

feature vector is shown, along with performance statistics for the model. This 2D NMR unfolding and refolding approach has recently been applied successfully to HSQC spectra and provides a simple methodology to apply statistical pattern recognition approaches to 2D NMR spectra [138, 139].

3.2.3.2 Mapping integrated bins – multivariate Differential Analysis NMR Spectroscopy (mvaDANS)

Integration of detected spectral segments transforms the 2D NMR dataset from a three-dimensional stack of full resolution 2D spectra (ω_1, ω_2, N) to a two-dimensional matrix of features (N, P) where P is the number of detected segments. Pattern recognition methods such as PCA, PLS, and derivatives of these methods can be applied directly to the bin matrix, though inclusion of intermediate normalization and stabilization steps applied to the bin matrix rather than the full resolution spectra improve pattern recognitions results [88, 104]. As in 3.2.3.1, the result of pattern recognition is, for each component in the model, a set of N scores showing the similarity of the samples to each other and a vector of P coefficients (referred to as loadings or predictors) that indicate the contribution of each bin to the separation between the samples. However, as a result of binning the number of variables (P) modeled is much lower than the full spectral resolution.

To interpret the model and identify important metabolites, coefficients are back projected onto the full resolution spectra by using the spectral segments. For all segments in the full resolution matrix, the data points within the segment are replaced with the value of the coefficient from pattern recognition of the bin extracted from that segment to produce a full resolution loadings matrix. For stabilized data, loading coefficients are independent of signal intensity and thus peak shape is not visible. To produce full-resolution back-projected loadings that resemble 2D NMR spectra, this matrix is then multiplied point-wise by the standard deviation along the third dimension of the 2D NMR stack (ω_1, ω_2, N). As with the full resolution data, contour lines are defined by the back-projected loadings, and the colors of these lines are defined by the coefficients themselves. Figure 3.2 outlines this approach, termed multivariate Differential Analysis NMR Spectroscopy (mvaDANS), as applied to a representative dqfCOSY spectrum of the *C. elegans* wild-type metabolome.



C. elegans daf-22

3.3 Data Sets

3.3.1 Nematode species differentiation: *P. pacificus* and *P. redivivus*

P. pacificus and *P. redivivus* were grown in 100 ml of 10% yeast extract in 250 ml Erlenmeyer flasks. Media were inoculated with 10 ml of a 10% yeast extract solution. The media were incubated at 25°C for 48 h. The media were then filtered through a 0.2 µm nylon syringe filter and stored at 4°C. The media were then used for the experiment.

P. pacificus 和 *P. redivivus* 的基因组大小和 GC 含量相似，但 *P. pacificus* 的基因组大小略大于 *P. redivivus*。在基因组注释方面，*P. pacificus* 的基因组注释包含更多的基因和蛋白质编码序列，这可能与 *P. pacificus* 的基因组大小和复杂性有关。在功能注释方面，*P. pacificus* 的基因组注释包含更多的与代谢和信号转导相关的基因和蛋白质编码序列，这可能与 *P. pacificus* 的生理功能有关。在系统发育分析方面，*P. pacificus* 和 *P. redivivus* 的基因组数据支持了它们在系统发育树上的位置，即 *P. pacificus* 属于 *P. pacificus* 属，而 *P. redivivus* 属于 *P. redivivus* 属。在比较基因组学方面，*P. pacificus* 和 *P. redivivus* 的基因组数据揭示了它们在基因组结构和功能上的相似性和差异性。在基因组组装方面，*P. pacificus* 的基因组组装质量略高于 *P. redivivus*，这可能与 *P. pacificus* 的基因组大小和复杂性有关。在基因组注释方面，*P. pacificus* 的基因组注释包含更多的基因和蛋白质编码序列，这可能与 *P. pacificus* 的基因组大小和复杂性有关。在功能注释方面，*P. pacificus* 的基因组注释包含更多的与代谢和信号转导相关的基因和蛋白质编码序列，这可能与 *P. pacificus* 的生理功能有关。在系统发育分析方面，*P. pacificus* 和 *P. redivivus* 的基因组数据支持了它们在系统发育树上的位置，即 *P. pacificus* 属于 *P. pacificus* 属，而 *P. redivivus* 属于 *P. redivivus* 属。在比较基因组学方面，*P. pacificus* 和 *P. redivivus* 的基因组数据揭示了它们在基因组结构和功能上的相似性和差异性。在基因组组装方面，*P. pacificus* 的基因组组装质量略高于 *P. redivivus*，这可能与 *P. pacificus* 的基因组大小和复杂性有关。

[illegible]

3.3.2 Animal model of peroxisomal disorder: *daf-22 C. elegans*

Wild-type *C. elegans* (N2) as well as *daf-22 (m130)* and *daf-22 (ok693)* mutant strains were obtained from the Caenorhabditis Genetics Center (CGC). Strains were maintained at 20 °C on NGM plates with bacteria (*E. coli* OP50) as food.

Liquid cultures were grown as previously reported [141]. Briefly, wild-type or *daf-22* mutant worms were grown for two generations on NGM plates seeded with OP50. Three crowded 6 cm NGM plates were washed into 100 mL solution of S-medium in a 500 mL Erlenmeyer flask and grown at 22°C and 220 pm. Concentrated OP50 from 1 L bacterial cultures, grown over-night, were given on day 1 and day 3. On day 5 the liquid culture was split into two 500 mL Erlenmeyer flasks and the growth media volume was maintained at 100 mL. An additional 1 L OP50 pellet was given as food upon splitting. The cultures were harvested on day 7 by centrifugation at 4750 rpm. The supernatant was lyophilized and the residue extracted with 95% ethanol (300 mL) at room temperature for 12 h. The resulting suspension was filtered and evaporated in vacuo at room temperature to yield the crude extracts. In this manner four extracts each of wild-type (N2), *daf-22 (m130)*, and *daf-22 (ok693)* cultures were prepared.

In preparation for NMR-spectroscopic analysis, the crude extracts were suspended in 1:8 methanol (MeOH):dichloromethane (DCM), using 7.5 mL of this solvent mixture per 1 gram of crude extract. Following centrifugation, 2 mL aliquots of supernatant of each sample were filtered over 1 g of silica (EMD Chemicals Inc.) using 10 mL of 1:8 MeOH:DCM for elution. The filtrate was evaporated to dryness in vacuo prior to analysis. For NMR spectroscopic analysis, 10-20 mg of the resulting metabolite extracts was used.

NMR samples were prepared by dissolving 10–20 mg portions of metabolite extract (see above) in 0.6 mL of CD₃OD. Samples were analyzed using a Varian INOVA 600 MHz NMR spectrometer. Nongradient phase-cycled dqfCOSY spectra were truncated at 0.4 s and acquired using the following parameters: 0.8 s acquisition time, 600 complex increments, 16 scans per increment. For processing with mvaDANS, spectra were zero-filled to 2048 × 1024 data points, a sine-square-shaped

window function was applied in both dimensions prior to Fourier transformation, and spectra were converted to magnitude mode prior to transfer to Matlab (Mathworks). For detailed analysis in phase-sensitive mode, dqfCOSY spectra were zero-filled to 8192×4096 data points and a cosine bell-shaped window function was applied in both dimensions before Fourier transformation. NMR spectra were processed using TopSpin (Bruker), VNMR (Varian), MestReC (MestreLabs), and MNova (MestreLabs) software packages.

3.3.3 Genotype-metabotype association in a single pedigree: Autosomal Dominant Fanconi Syndrome I (ADFS1)

ADFS1 is a rare autosomal dominant disorder characterized by Fanconi syndrome, which is a group of symptoms that affect the kidneys and can lead to bone disease, growth retardation, and other complications. The disorder is caused by a mutation in the *SLC12A3* gene. In this study, we performed a detailed analysis of the ADFS1 pedigree, focusing on the genotype-metabotype association. The analysis was performed using the following steps:

1. Identification of the affected individuals in the pedigree.
2. Determination of the mode of inheritance (autosomal dominant).
3. Identification of the mutation in the *SLC12A3* gene.
4. Analysis of the genotype-metabotype association.

250 μ L aliquots of filtered, centrifuged urine were combined with 150 μ L of D_2O and 100 μ L of NMR buffer of CDCl_3 . The mixture was then analyzed by NMR spectroscopy. The results of the analysis are shown in Figure 3.3.3.

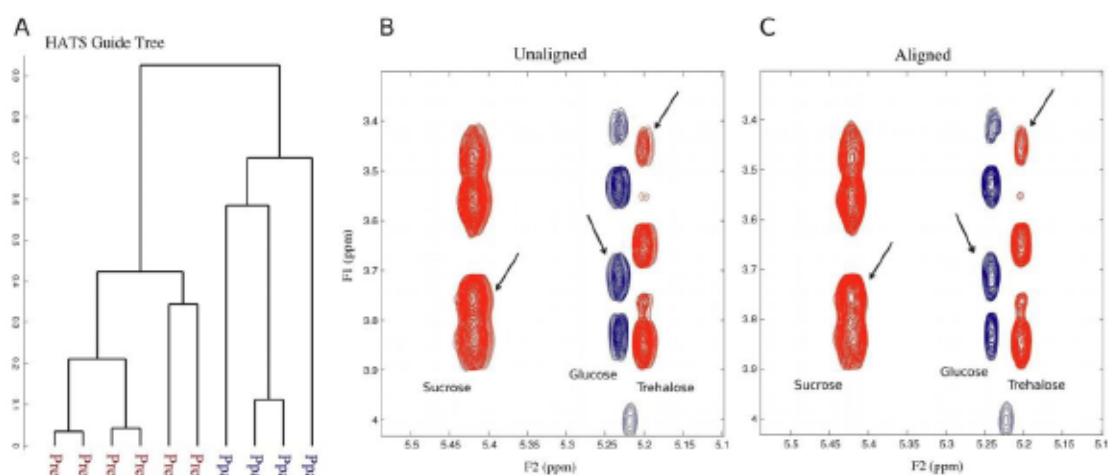
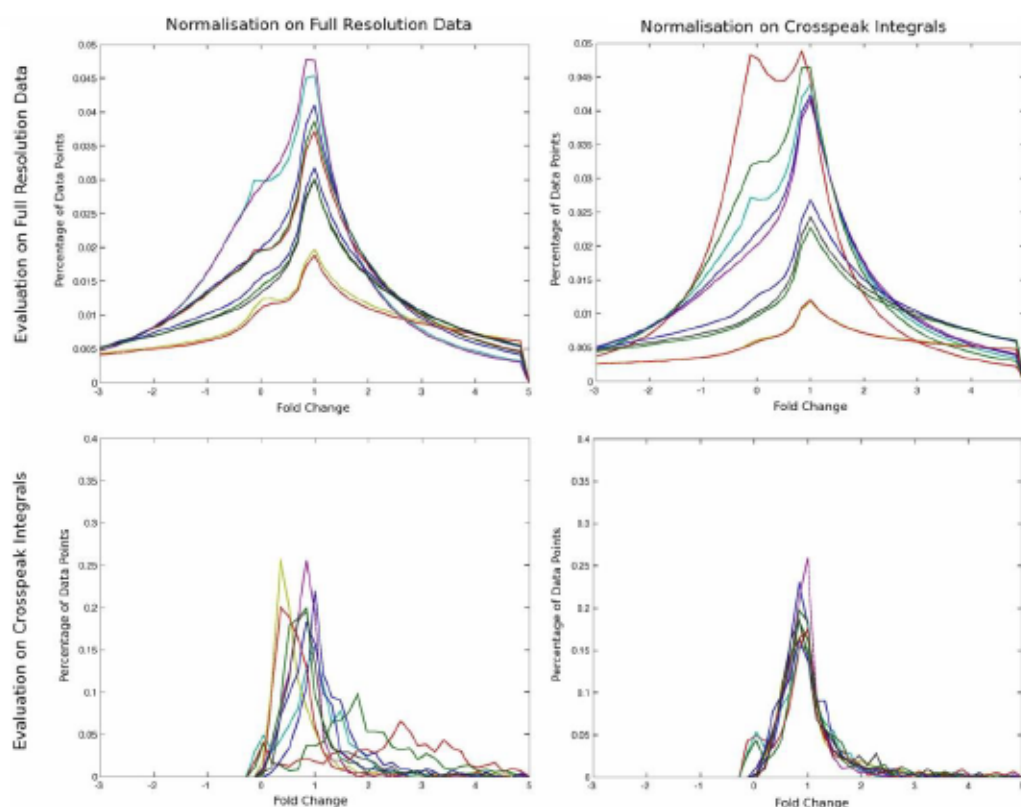


Figure 3.3. Alignment results for *P. redivivus* (Pre) and *P. pacificus* (Ppa) TOCSY spectra of flow-through fractions from C18 solid phase extractions. A) The guide tree produced by HATS groups spectra from the same species, resulting in intra-species alignment before inter-species alignment. The utility of a guide tree based alignment is exemplified by the anomeric crosspeaks of glucose, sucrose, and trehalose in the TOCSY spectra (B, C). By aligning spectra with similar composition first, false positive alignments of nearby but unrelated crosspeaks, such as the anomeric crosspeaks shown above, are avoided. Arrows indicate poorly aligned signals for which positional variation is correctly resolved by HATS.

As 2D NMR data are frequently used in a qualitative rather than quantitative manner, preprocessing steps such as normalization and variance stabilization or scaling are not frequently applied to 2D datasets. However, as in 1D NMR, signal intensity in 2D NMR reflects solute concentration in the sample being measured. While cross-peak integrals are not absolutely quantitative due to effects such as incomplete magnetization transfer and the number of bonds between the coupled nuclei giving rise to the crosspeak as well as other detection and processing parameters, integrals are relatively quantitative in that changes in solute concentration should result in linear changes in the integral value. Additionally, 2D NMR spectra, like 1D spectra, are affected by random dilution especially in the case of urinary spectra.

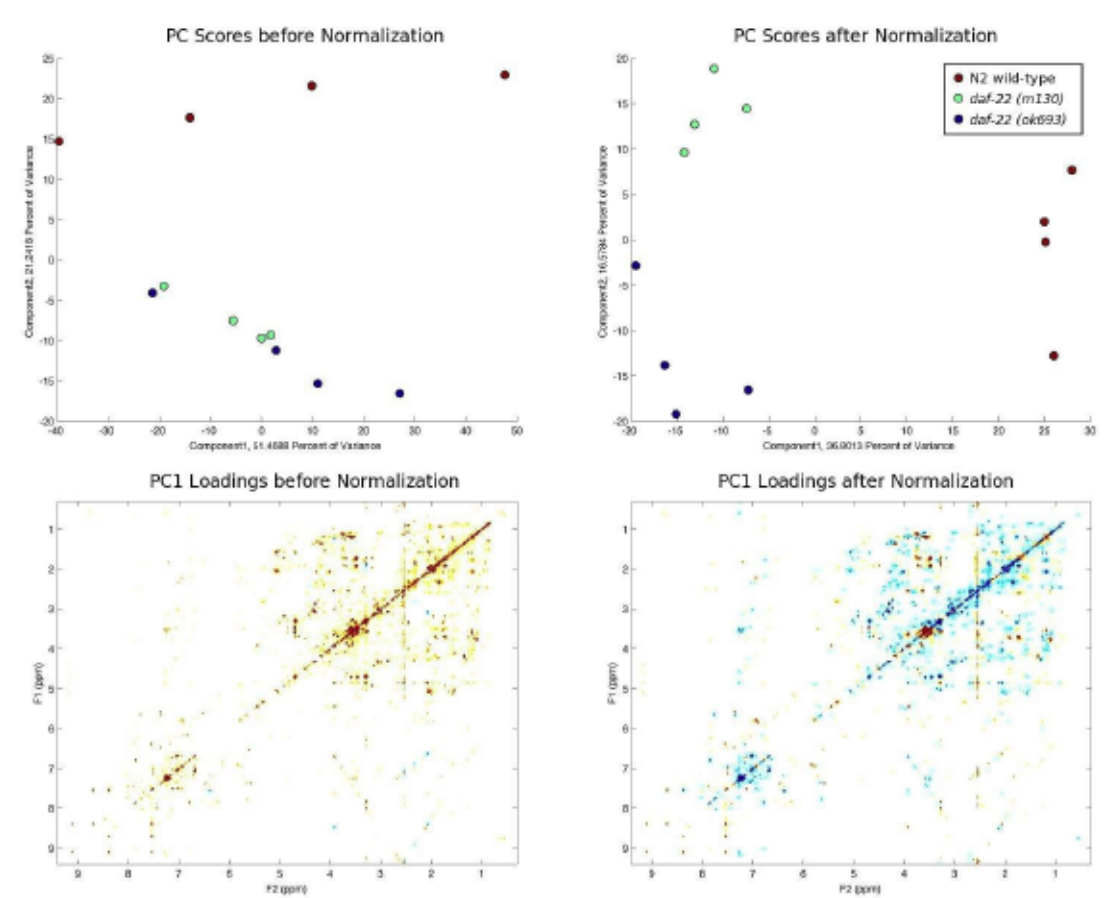
Unlike 1D data where a majority of the collected data points have a substantial contribution from signal with few variables dominated by noise, 2D correlation spectroscopy data is highly sparse. Even with the narrower spectral widths (10-12 ppm) common for TOCSY and COSY spectra, cross-peaks are present in less than 10% of spectral data points as determined by comparing detected spectral bins with

the full resolution data, even for crowded spectra such as the *C. elegans* extracts and human urine. As a result, normalization to the median fold change of all spectral data points does not rectify the linear changes in signal intensity caused by variable dilution. Figure 3.4 compares the effect of normalizing on the full resolution data and just the spectral bin integrals for the Autosomal Dominant Fanconi Syndrome 1 (ADFS1) human urinary data.



Here, normalizing on the full resolution data and evaluating the fold change distributions of the data show a distribution centered at 1 as expected. However, this distribution is very broad, as increased variance is to be expected when taking the ratio of noise values sampled randomly from a distribution centered at zero.

Estimating dilution factors from the full resolution data does not center the distribution of bin integrals. Normalizing the bin integrals directly does correct for the effects of random dilution on cross-peak intensity. Estimating dilution factors from the median of the distribution of bin integral fold changes for 2D correlation spectra is the equivalent of doing peak probabilistic quotient normalization for 1D spectra [88]. However, the increased sparseness of the 2D data makes normalization using signal values only even more important than for 1D data.



The following text is a placeholder for the main body of the document, which appears to be a corrupted or placeholder version of the text. The text is mostly illegible and contains many placeholder characters.

Dilution effects are well known in human urine, and the necessity of normalization to overcome variable dilution for human urinary data is established in

urinary metabolic profiling [104]. However, urine is not the only sample matrix that shows variable dilution effects, and these can be observed in extracts as well even when sample preparation attempts to standardize concentration by sample volume or dried mass. Dilution effects are shown in Figure 3.5 for the *C. elegans daf-22* and N2 spectra.

While final concentrations for NMR were roughly on the same order of magnitude due to the preparation of 10-20 mg of filtered extract in 600 μ L CD₃OD, this variation is still great enough that the variance in Principal Component Analysis is dominated by linear changes in signal intensity shown by the positive loading coefficients for nearly all spectral signals in PC1 of the un-normalized data, indicating that concentration effects are the primary source of variation in the raw data. While this would not be a significant complication to interpreting the data qualitatively, only following normalization is the dataset able to be interpreted in a quantitative manner using pattern recognition tools. These results indicate that the preprocessing methods developed for 2D NMR datasets presented here improve the interpretability of these data by adapting and extending the framework of the methodology applied to 1D NMR and mass spectral data in metabolomics.

3.4.2 Identification of β -oxidation intermediates in a *C. elegans* model of peroxisomal disorders

C. elegans is an established model organism in biochemical genetics, studies of which have yielded important insights into metabolic pathways and regulation [142-144]. Recent analyses of the *C. elegans* metabolome revealed a family of structurally diverse small molecules, the ascarosides, as products of conserved **peroxisomal** β -oxidation [131, 145,146]. Mass spectrometry-based comparative metabolomics characterized the functions of three enzymes in ascaroside biosynthesis, the acyl-CoA oxidase ACOX-1, the enoyl-CoA hydratase MAOC-1, and the 3-hydroxyacyl-CoA dehydrogenase DHS-28[146].

Figure 3.6 shows the *C. elegans* **peroxisomal** β -oxidation pathway responsible for ascaroside biosynthesis with the homologous human long-chain lipid catabolic pathway. The fourth enzyme in the *C. elegans* **peroxisomal** β -oxidation cascade, encoded by the gene *daf-22*, is a close homologue of mammalian 3-keto-acyl-CoA thiolases (human *SCP2*, 83% homology), but its predicted thiolase activity could not

be confirmed in previous work because no accumulating metabolic intermediates had been found.

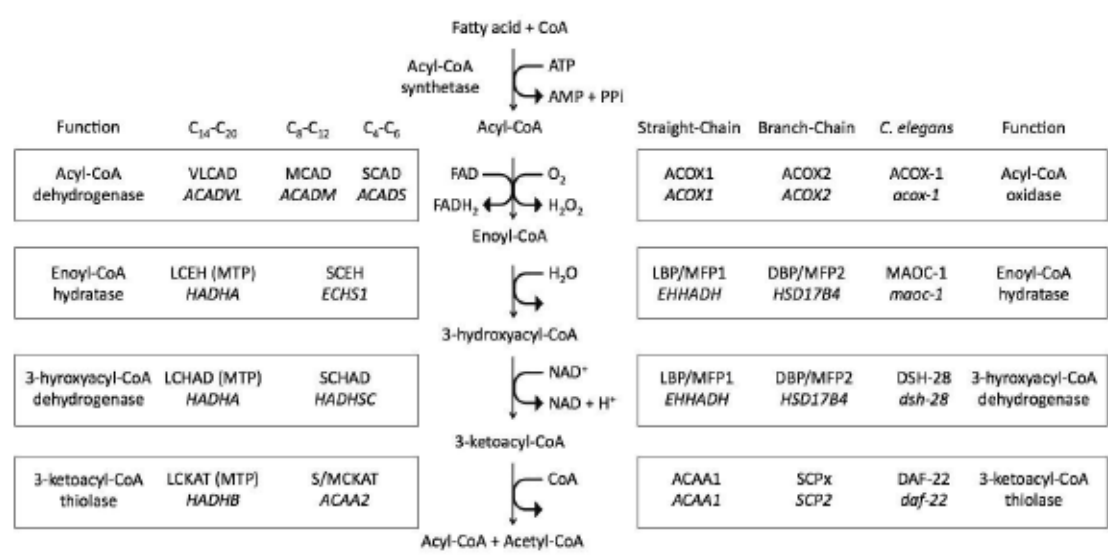


Figure 3.6. Human mitochondrial and peroxisomal β -oxidation with homologous *C. elegans* pathway. While the underlying biochemistry of β -oxidation is highly conserved between organelles and across substrates and species, a diverse set of enzymes in the mitochondrion catabolizes shorter chain lengths primarily for energetic purposes, while the peroxisome handles more structurally diverse compounds to supply processed fatty acids for both mitochondrial metabolism and diverse structural and metabolic roles.

Identification of shunt metabolites in *daf-22* mutant exo-metabolome would help confirm the biochemical function of this enzyme and provide insight into interactions with other biochemical pathways. To search for these metabolites, the *daf-22* metabolome was investigated using 2D NMR spectroscopy-based metabolomics using the multivariate Differential Analysis by 2D NMR Spectroscopy (mvaDANS) methodology outlined previously in section 3.2. These tools enabled the automated processing and comparative computational analysis of 2D NMR spectra from metabolome extracts obtained from two different *daf-22* alleles (*m130*, *ok693*) and wild-type *C. elegans*, including four independent biological replicates of each.

Automatic crosspeak identification and binning prepared the dqfCOSY spectra collected from wild-type and mutant *C. elegans* for statistical analysis via Principal Component Analysis, which clustered the data by genotype (Figure 3.7A). Coefficients from the PCA loadings were back-projected onto the dqfCOSY spectra as outlined previously (Figure 3.2), revealing a large number of crosspeaks that are up- or downregulated in *daf-22* mutants relative to wild-type (Figure 3.7B). Excluding crosspeaks also present in the *E. coli* food source, analysis of the remaining strongly *daf-22*-upregulated crosspeaks suggested a series of partial structures, **including methyl ketones, α -methyl branched fatty acids, and corresponding ethanolamides, as well as β -keto fatty acid derivatives** (Figure 3.7 F-I). Fractionation of the *daf-22* metabolome guided by the signals indicated in the PCA loadings led to the isolation and full characterization of very long-chain ascarosides (VLCA) that account for all of the predicted structural features (Figure 3.7K), and HPLC–MS analysis confirmed the absence of these components in wild-type samples (Figure 3.7J).

The accumulation of VLCAs in *daf-22* mutant *C. elegans* matches observations of accumulated very long chain fatty acids (VLCFAs) in human β -oxidation defects, including peroxisomal β -oxidation defects in the Zellweger spectrum of disorders and in mitochondrial β -oxidation defects such as VLCAD deficiency. The chain lengths (23- to 27-carbon) of the *daf-22* mutant-upregulated VLCAs in *C. elegans* are similar to those of very long-chain fatty acids that are oxidized solely in the human peroxisome [147] and accumulate in human disorders **involving defects in peroxisomal β -oxidation** [148]. While D-bifunctional protein and SCPx are considered the main biosynthetic β -oxidation enzymes in the human peroxisome, metabolizing phytanic acid, bile acid intermediates di- and trihydroxycholestanoic acid (DHCA and THCA), recent evidence suggests that these enzymes also serve as a backup system for oxidizing some straight-chain fatty-acids [148]. The co-occurrence of straight chain and methyl-branched derivatives among the *daf-22* mutant-specific VLCAs suggests that like SCPx, DAF-22 is involved in processing both branched and straight chain fatty acids.

The dual roles of β -oxidation in both energetics and biosynthesis can be observed in the differential pathology of defects to these systems. Mitochondrial β -oxidation defects (VLCAD, MCAD, and SCAD deficiency) are among the most common inborn errors of metabolism, and while these are considered leading causes

of sudden infant death syndrome (SIDS), effects are typically only observed upon fasting [149]. However, peroxisomal β -oxidation defects are typically much more complex. Classical Zellweger syndrome, caused by mutations in PTS1 and PTS2 targeting signals, results in defective peroxisomal biogenesis, the absence of peroxisomes in the liver and kidney, and a host of clinical consequences including cranial malformation, hypotonia, failure to thrive, and hepatic dysfunction [148].

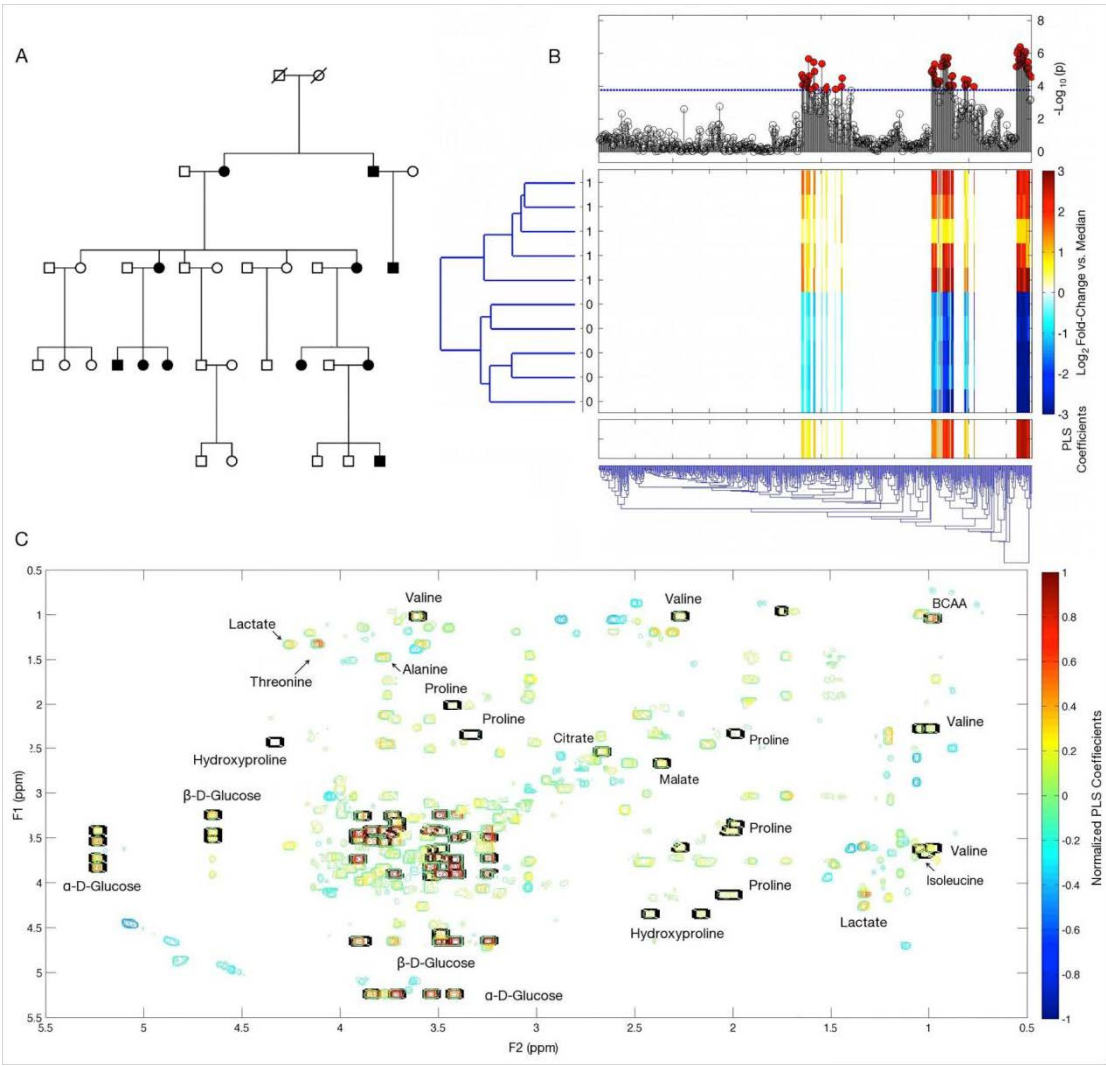
While several patients with LBP and peroxisomal 3-ketoacyl-CoA thiolase deficiency have been described [150, 151], reinvestigation of these patients following the discovery of the second peroxisomal β -oxidation pathway indicated that these patients were actually deficient in DBP [152, 153]. These patients show cerebro-hepato-renal phenotypes similar to those observed in Zellweger syndrome. There have been no further reported cases of true LBP deficiency, suggesting it may be asymptomatic as due to partial compensation by DBP. Investigation of knockout mice showed that while LBP $-/-$ mice showed no gross phenotypic abnormalities, DBP $-/-$ mice showed significant growth retardation and postnatal mortality [154]. As *daf-22* is homologous with human *SCP2* (BLAST e-value $4e-30$, 83% homology), it is unsurprising that *daf-22* plays a role in biosynthesis. While it is unknown whether *C. elegans* has a separate peroxisomal β -oxidation pathway for non-ascaroside fatty-acids, previous investigations of *maoc-1* and *dsh-28* mutant *C.elegans* did not show significant phenotypic effects such as those observed in DBP deficiency [146], suggesting this is a possibility.

Understanding of the metabolic role of the peroxisome, especially the enzymology of peroxisomal β -oxidation, has evolved significantly in the past decade. These results, together with analyses of the *C. elegans* peroxisomal β -oxidation pathway [146] expand the known roles of peroxisomal β -oxidation to include biosynthesis of signaling compounds. While *C. elegans* has been used as a model in genetics, relatively few attempts have been made to use common animal models such as *C. elegans* and *D. melanogaster* to study human inborn errors of metabolism, though what work has been done with these models indicates their potentially significant utility [143, 155-157]. Given the availability of large libraries of *C. elegans* mutant strains and mature RNAi protocols, and the increasing demands for functional genomic analysis resulting from increasing variant discovery in clinical genetics, *C. elegans* models may be useful tools in human biochemical genetics.

3.4.3 Identification of renal tubular transporter and TCA cycle substrates in Autosomal Dominant Fanconi Syndrome I probands

Fanconi Syndrome (FS) is a proximal tubular cell dysfunction that renders the renal tubule unable to recover small molecule solutes from the ultrafiltrate leading to the urinary loss of solutes including glucose, amino acids, and ions including phosphate, calcium, sodium, potassium, and bicarbonate. While the most common cause of Fanconi Syndrome in children is cystinosis [158], FS results from a range of Mendelian diseases [23], from toxicity to the proximal tubule [159], and from various idiopathic origins. As the active transport processes of the renal tubule are among the most energetically intensive processes in the body, mitochondrial stress and disruptions to energetic metabolism frequently result in Fanconi Syndrome as well. Fanconi Syndrome with an autosomal dominant inheritance pattern was identified in a five-generation African American family, (Figure 3.8A) [160]. Genetic linkage studies subsequently mapped this phenotype to chromosome 3 and sequencing subsequently identified a heterozygous missense mutation in *EHHADH* that created a de novo mitochondrial targeting motif leading to mistrafficking of L-Bifunctional Protein, an enzyme involved in peroxisomal fatty acid oxidation (Figure 3.6) to the mitochondrion.

To connect the renal phenotype with the *EHHADH* variant genotype, untargeted urinary metabolic profiling was performed using 1D and 2D NMR spectroscopy for (N=10) family members, with equal numbers of affected and unaffected individuals. While both 1D ^1H and 2D ^1H - ^1H TOCSY spectra of affected individuals clearly showed the glucosuria and generalized aminoaciduria characteristic of Fanconi Syndrome, the 1D ^1H spectra were observed to be highly complex and overlapped given the intense glucose and amino acid signals in these spectra. Additionally, the low number of samples complicated untargeted analysis by reducing the statistical power of significance testing given the large number of variables in full resolution NMR and the unsuitability of binning for overlapped signals.

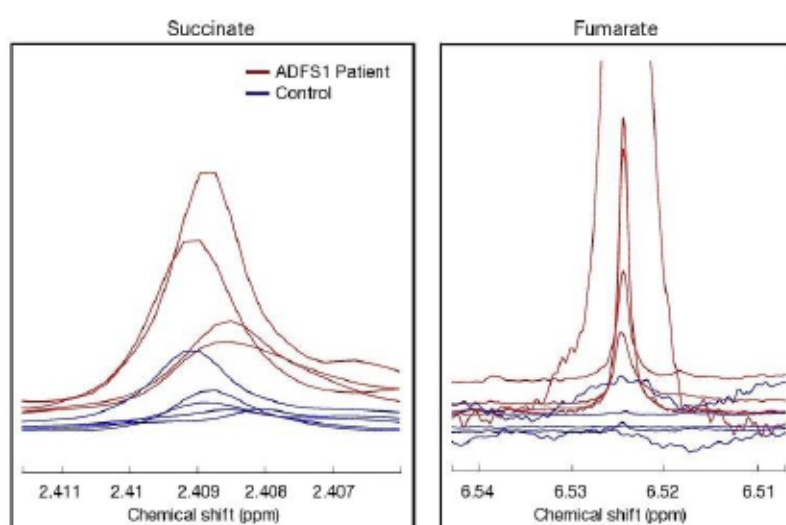


To mitigate these challenges and identify urinary signatures of ADFS1 from a single family, 2D TOCSY spectra were subjected to automatic crosspeak identification and binning using the mvaDANS algorithm presented above. To identify significantly variable metabolites, crosspeak bins were transformed by log₂ fold-change vs. median and analyzed using two-way hierarchical cluster analysis, PLS-DA, and significance testing using two-way, unpaired t-tests with Bonferroni correction for multiple hypothesis testing (Figure 3.8B). Hierarchical cluster analysis and PLS-DA clearly separate affected and unaffected individuals, with PLS-DA showing an R^2 of 0.94 and a Q^2 of 0.88 under four-fold Monte Carlo cross validation using a single PLS component explaining 61% of dataset variance. A manhattan plot identifies a set of 69 crosspeak bins with p-values less than the threshold set by Bonferroni correction ($p < 1.71 \times 10^{-04}$). The clustered heatmap of log-fold-change vs. median values showing these significant features indicate that, as expected from a urinary solute loss phenotype, all are more abundant in the urine of affected than unaffected individuals.

To identify the metabolites represented by these signals, PLS-DA coefficients were back-projected onto the full-resolution NMR data and bin boundaries of significant features outlined in black, which were assigned using reference data from spectra of pure compounds (Figure 3.8C). These results show a preponderance of signals from glucose and amino acids, especially branch-chain amino acids, proline, and hydroxyproline. While increased lactate, threonine, and alanine excretion were observed and indicated by positive PLS-DA coefficients for these signals, the presence of these metabolites in the urinary spectra of unaffected individuals meant they did not pass the threshold for significance testing by untargeted analysis. Additionally, citrate and malate were significantly increased in the urine of affected individuals.

While the observation of glucosuria and generalised aminoaciduria matches previous identification of Fanconi Syndrome in the family [160], the identification of significantly increased urinary citrate and malate suggest a more direct connection to the genotype. Increased urinary excretion of TCA cycle intermediates is known to accompany mitochondrial pathology [161-163]. Interpreting urinary excretion of these metabolites is complicated by the multiple roles of many of these metabolites, especially citrate, in the kidney [164]. However, evidence from screening clinics suggest that malate and fumarate are sensitive and specific indicators of mitochondrial

disease [163], and coordinated loss of multiple TCA cycle intermediates would support this assertion. While only citrate and malate were identified in the 2D NMR dataset, several TCA cycle intermediates are present as singlets in NMR spectra, which would preclude their identification by mvaDANS as the algorithm does not consider signals along the diagonal of TOCSY or COSY spectra. Of the TCA cycle intermediates that present as singlets in the corresponding 1D spectra (cis-Aconitate, succinate, fumarate, and oxaloacetate), signals from succinate and fumarate were observed to be elevated in ADFS1 urinary 1D spectra, with fumarate elevation especially notable (Figure 3.9).



These results connect the identified *EHHADH* variant genotype to the phenotypic observation of renal Fanconi Syndrome in a single family by providing additional evidence for energetic disruption in the renal proximal tubule. Proximal tubular cells are among the most energetically demanding in the body, relying on mitochondrial beta-oxidation to meet the requirements for ATP dependent active solute transport across the tubular lumen [165]. The observation of renal tubular transporter substrates and TCA cycle intermediates in the urine of ADFS1 patients matches observations of impaired oxidative phosphorylation in cells expressing

CHAPTER 4

METABOLIC PROFILE SCREENING AND BIOMARKER IDENTIFICATION FOR INBORN ERRORS OF METABOLISM

4.1 Introduction and aims

Inborn errors of metabolism (IEMs) have been studied since their identification by Sir Archibald Garrod in the early 1900s. While these defects were initially considered to be non-pathological idiosyncrasies of metabolism resulting from "chemical individuality", growing understanding of neonatal diseases and the identification of pathological inborn errors such as phenylketonuria (PKU) lead to efforts to implement population-wide newborn screening programs. Early screening programs, such as the PKU screening pioneered by Robert Guthrie, relied on bacterial inhibition assays [46], but in the 1990s screening programs were given a significant boost by the introduction of ESI-MS/MS for the measurement of small molecule biomarkers of inborn errors [47, 48]. Today, MS/MS is the main primary screening technique [86] for detection of the plurality of the 31 diseases included in the Recommended Uniform Screening Panel for neonatal screening [49] (<http://www.hrsa.gov/advisorycommittees/mchbadvisory/heritabledisorders/recommendedpanel/>).

While the introduction of ESI-MS/MS significantly improved the sensitivity and specificity of many screening assays, diagnostic algorithms for these assays still typically classify patients as positive or negative on the basis of cutoff values for a single metabolite or small number of metabolites and metabolite ratios [172]. Cutoff values are typically selected by individual screening labs in order to minimize false-

negatives, resulting in greater numbers of false-positive samples which must be followed up by secondary and sometimes tertiary tests [173]. Additionally, the measurement of these metabolites by MS/MS can require extensive sample preparation including derivitization, quantification by reference to isotopically labeled standards which represent a significant variable cost in screening, and the use of multiple reaction monitoring (MRM) involves a trade-off between sensitivity and acquisition time [174].

In addition to analytical challenges, screening labs will likely face challenges in the future as screening programs expand in scope. While the American College of Medical Genetics lists 31 diseases in the Uniform Screening Panel, current estimates suggest that as many as ~6,500 congenital diseases resulting from rare genetic mutations may exist (<http://omim.org/statistics/entry>) [9]. As clinical genetics, including whole exome and genome sequencing [50-52], expands the set of known inborn errors it is likely that screening programs will include a larger number of diseases in the future. An expanded panel will thus pose challenges to diagnostic sensitivity and specificity as more diseases are included.

Improving diagnostic sensitivity and specificity while reducing cost and complexity are important goals for newborn screening programs. Greater coverage of the metabolome may also provide important phenotypic context for the increased genomic knowledge provided by expanded sequencing programs. Identification of a broader set of metabolic markers of inborn errors could not only limit second-tier tests but also aid the study of biochemical genetics by identifying metabolic perturbations associated with mutations. Recent work simulating the metabolic consequences of enzyme deficiencies [175] and connecting serum and urinary metabolite concentrations with single nucleotide polymorphisms [38, 39, 44, 176] suggest expanded metabolic coverage in newborn screening may improve understanding of genotype-phenotype correlations for genetically heterogeneous conditions.

Expanding measurement to provide greater analytical coverage of the metabolome from dried blood spot (DBS) samples should allow many tests to be done in parallel, rather than in series. In addition, the use of multivariate regression to build models that predict disease on the basis of measurements of multiple metabolites should also provide improved sensitivity and specificity relative to cutoff values. While MS/MS is not incompatible with broader measurement and multivariate regression, newer methods such as direct injection nanospray high resolution mass

spectrometry (nS-HR-MS) [173] may be able to provide reduced time, cost, and complexity while expanding metabolome coverage and maintaining sensitivity and specificity.

This chapter compares MS/MS and single stage nS-HR-MS measurement and analysis of DBS samples for 6,050 unaffected and 102 newborns affected by 24 congenital disorders of metabolism (Table 4.1). For both MS/MS and nS-HR-MS measurements, sparse logistic regression models were built for identification of PKU and MCADD positive samples and the sensitivity and specificity of these models was evaluated for independent test sets. These regression based methods show comparable or greater sensitivity and specificity than canonical cutoff based algorithms. Finally, the expanded metabolomic coverage of nS-HR-MS was used to identify a broader range of perturbed metabolites for phenylketonuria (PKU), medium-chain acyl-CoA dehydrogenase deficiency (MCADD), and the 22 additional disorders. While re-designing screening methodologies will require analysis of a much larger affected population, expanded analytical coverage of the metabolome and multivariate regression has the potential to both increase diagnostic accuracy and provide important metabolic context for inborn errors.

4.2 Materials and Methods

4.2.1 Sample collection

Anonymous DBS samples were provided for 6050 healthy newborns by the Hesse Screening Center (Giessen, Germany). Ethical permission was provided by the clinical laboratory and by the National Research Ethics Service in the UK(12/SC/0634). 73 true positive samples were provided by the Hesse Screening Center and 29 samples were provided by the Hungarian Screening Laboratory (Budapest, Hungary). Characteristics of these samples are shown in Table 4.1. For all samples ESI-MS/MS data was collected in the original screening laboratories and provided along with codes matching the anonymous DBS samples. An additional 36 samples including 9 MCADD true-positive, 3 false-positive, and 24 true negative were provided by the National Institute for Health Data and Disease Control (Copenhagen, Denmark) as an independent test for the MCADD regression model.

All MS/MS data from screening facilities was provided as tab-delimited data files reporting metabolite levels rather than raw spectral data.

OMIM Phenotype ID	Inborn Error of Metabolism	Samples (N)
Inborn errors of fatty acid metabolism		
201450	Medium-chain acyl-coA dehydrogenase deficiency (MCADD)	20
201475	Very long-chain acyl-CoA dehydrogenase deficiency (VLCADD)	5
609016	Long-chain 3-hydroxyacyl-coA dehydrogenase deficiency (LCHADD)	2
212140	Carnitine transporter deficiency	2
201470	Short-chain acyl-coenzyme A dehydrogenase deficiency (SCADD)	1
Organic Acidurias		
606054	Propionic acidemia (PA)	4
243500	Isovaleric acidemia (IVA)	4
210200	3-Methylcrotonyl-CoA carboxylase deficiency (3MCC)	4
251000	Methylmalonic acidemia (MMA)	2
231670	Glutaric acidemia type 1 (GA I)	1
231680	Glutaric acidemia type 2 (GA II)	1
248600	Maple syrup urine disease (MSUD)	1
Amino acid metabolism defects		
261600	Phenylketonuria (PKU)	18
233910	Hyperphenylalaninemia (HPA)	16
276700, 613018, 276710	Tyrosinemia	2
250850	Methionine adenosyltransferase deficiency	1
Urea cycle defects		
215700	Citrullinemia	2
311250	Ornithine transcarbamylase deficiency	2
207900	Argininosuccinate lyase deficiency	1
608300	N-Acetylglutamate synthase deficiency (NAGS)	1
Carbohydrate metabolism disorders		

230400	Galactosemia	1
	Other	
193090	Cobalamin deficiency	5
275200	Hypothyroidism	4
	Non-ketotic hyperglycinemia (NKH)	2

4.2.2 Sample preparation

3.2 mm punches of dried blood spot samples were placed in 96-well filter plates (Millipore) and extracted with 100 μ L HPLC grade methanol (Sigma-Aldrich). An additional 100 μ L methanolic solution of 20 isotopically labeled internal standards (Cambridge Isotope Laboratories, Appendix A Table 1) was pipetted onto the samples and the plates were shaken for 35 minutes at room temperature to extract the blood spot samples. The resulting extract was filtered by centrifugation at 2000 rpm for 2 minutes into a second 96-well plate (Greiner Bio-One) and diluted with 50 μ L HPLC grade water (Sigma-Aldrich). Plates were sealed with heat seal foils (ABgene) and analyzed by mass spectrometry.

4.2.3 Mass spectrometry and data preparation

High-resolution mass spectrometry was performed using an Exactive instrument (Thermo) equipped with TriVersa NanoMate ion source (Advion). Chip-based infusion mode measurements were performed using 5 μ m nominal internal diameter nozzle chips, and a sample volume of 5 μ L was injected. Mass spectra were recorded in full scan mode over a mass range of 70-600 for both positive and negative ion modes at a nominal resolution of 100,000. Total data acquisition time was 2 min, 1 min in negative ion mode and 1 min in positive ion mode with automatic polarity switching. All mass spectra were recorded in Giessen, Germany with the exception of the samples provided by the National Institute for Health Data and Disease Control lab, which were recorded in London, UK.

Full scan mode .RAW files (Thermo) were converted to XML encoded text files using the MSConvert package of ProteoWizard. The resulting MS data, including intensities, m/z ratios, and detection mode, were extracted for each scan in

Matlab 2012b (Mathworks). Positive and negative scans were separated, interpolated with a resolution of 0.0001 mass units to generate consistent m/z domains between scans, and averaged for each data file. Intensities corresponding to isotopically labeled internal standards (Appendix A, Table 1) and 238 selected molecular ions assigned in previous work [173] (Appendix A, Table 2) in positive and negative detection modes were extracted by identifying local maxima within 0.0025 mass units of the expected ion m/z . Internal standard and analyte intensities for all detected $[M+H]^+$, $[M+Na]^+$, and $[M-H]^-$ adducts along with mass deviation in ppm were exported in CSV format for each processed spectrum.

4.2.4 Data processing and analysis

Ion intensities for each detection mode were imported from CSV format into Matlab and ion intensities were averaged according to eq. 4.1 to generate a consensus intensity for each compound across both detection modes.

$$I = \frac{\frac{\sum_{i=1}^N IS_i}{N} + \frac{\sum_{j=1}^M IS_j}{M}}{2} \quad (4.1)$$

Here, the intensities (IS) of N positive mode ions and M negative mode ions are averaged first by mode, and then across modes to give a single intensity (I) for a given compound. Samples were considered poor quality runs and excluded from further analysis if less than half of internal standards or less than half of metabolites were detected. 329 such samples were identified and excluded out of the set of 6050 control samples, while none of the affected samples were excluded.

nS-HR-MS data was preprocessed for relative quantitation prior to statistical analysis. Compound intensities were normalized using probabilistic quotient normalization [88], stabilized by an offset $\log_2$ fold-change vs. median with an offset of 10% of the dataset median value, and median centered by batch. Absolutely quantitative measurements from MS/MS were not modified and nS-HR-MS and MS/MS data were matched by anonymous sample ID. MS/MS data was available for 5627 of the 5721 high quality nS-HR-MS spectra.

Predictive models were developed for PKU and MCADD. Control and affected samples were split into training and test sets of equal size with equal allocation of affected samples to train and test sets. Significant variables were identified by a two-way, one-tailed, one (PKU or MCADD) vs. all (controls and non-PKU/MCADD affected) t-test with Bonferroni correction for multiple testing. Log ratios for significant variables vs. all other variables were then calculated according to eq. 4.2 for MS/MS and eq. 4.3 for nS-HR-MS.

$$\log_2\left(\frac{X}{Y}\right) \quad (4.2)$$

$$\log_2\left(\frac{X}{Z}\right) - \log_2\left(\frac{Y}{Z}\right) \quad (4.3)$$

For MS/MS, X and Y were the peak intensities of the precursor ion and the base peak, respectively. For nS-HR-MS, X and Y were the peak intensities of the precursor ion and the base peak, respectively. Z was the sum of the peak intensities of all ions in the spectrum. The log ratios were calculated according to eq. 4.2 for MS/MS and eq. 4.3 for nS-HR-MS. The log ratios were then used to identify significant variables using a two-way, one-tailed, one (PKU or MCADD) vs. all (controls and non-PKU/MCADD affected) t-test with Bonferroni correction using the number of ratios tested for each compound was included in the set of filtered variables.

Sparse logistic regression models were then built for the filtered set of compounds and compound ratios using multivariate logistic regression with forward stepwise variable selection. Models were selected on the basis of the deviance of an internal test set using 4-fold cross validation and evaluated using the external test set. Stepwise regression produces sparse models in which the coefficients for the majority of variables are set to 0 and prediction is based on a small number of compounds and ratios. For MCADD, these models were validated with an additional set of positive and control samples from the Danish screening lab, which were analyzed and preprocessed separately.

Compounds were considered to be markers for one of the 24 disorders if a Bonferroni correction for the samples of a single disorder vs. samples from all other disorders was significant ($p < 2.1 \times 10^{-4}$ for nS-HR-MS and $p < 0.0015$ for MS/MS) and if the median level of the compound in the affected samples was in the extreme 1% of the levels of all control samples.

4.3 Results – Characterizing the nS-HR-MS method

4.3.1 Isotopically labeled internal standards

As previously noted, a methanolic solution with a fixed concentration of twenty isotopically labeled standards (Appendix A, Table 1) was added to the DBS extracts. In a previous pilot study [173], these standards were used to calculate concentration values for endogenous metabolites. Preliminary results with an expanded set of 238 endogenous metabolites showed unexpected variation in these concentration values, and to identify the source of this variation the detection and variance of the internal standards was investigated.

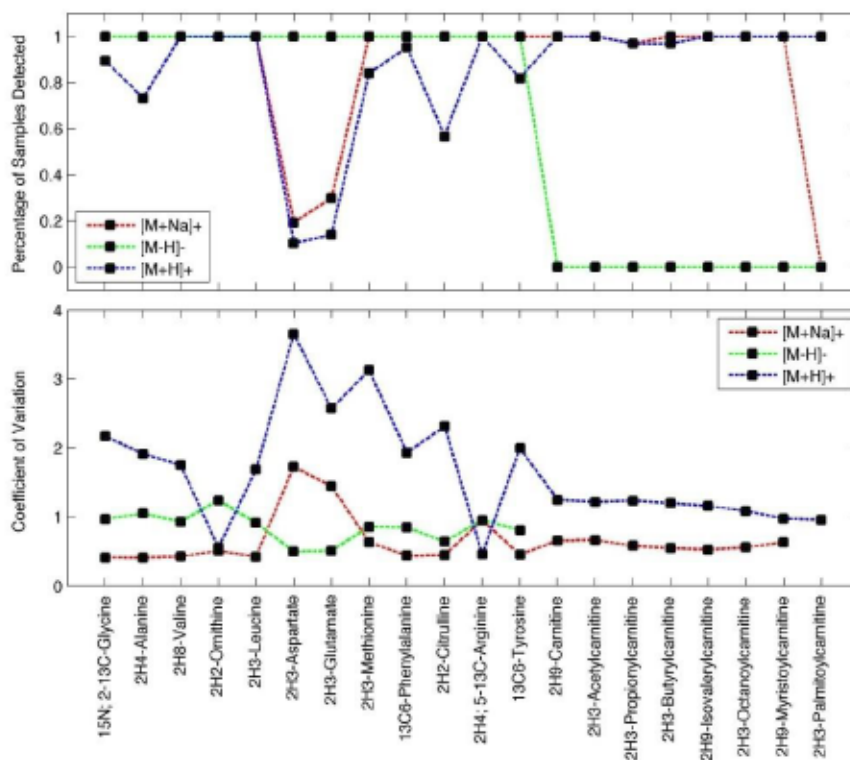
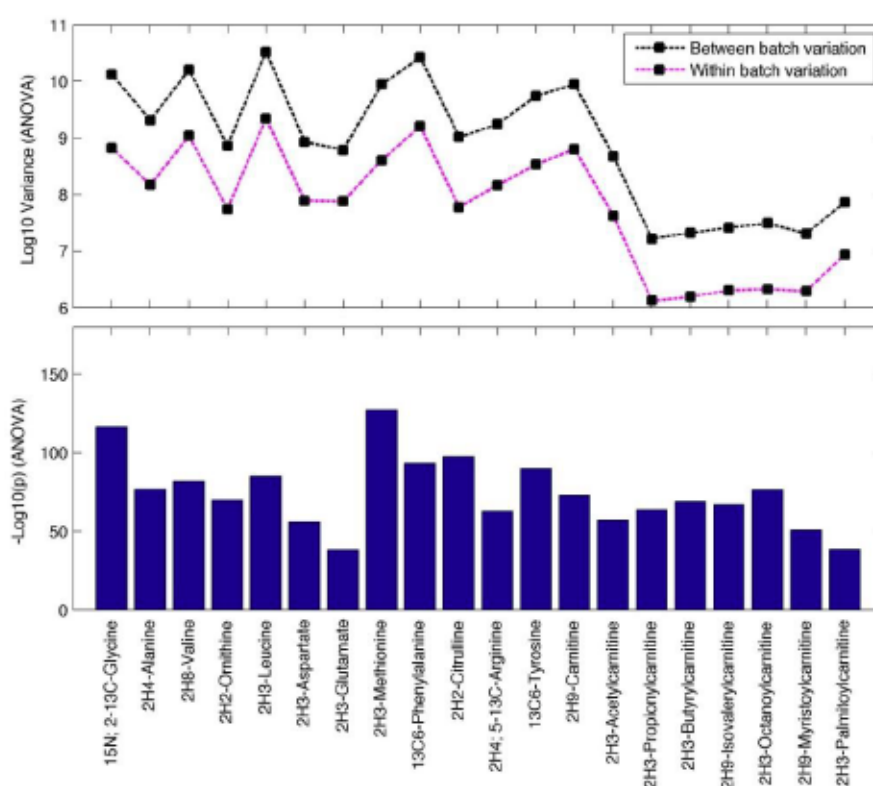
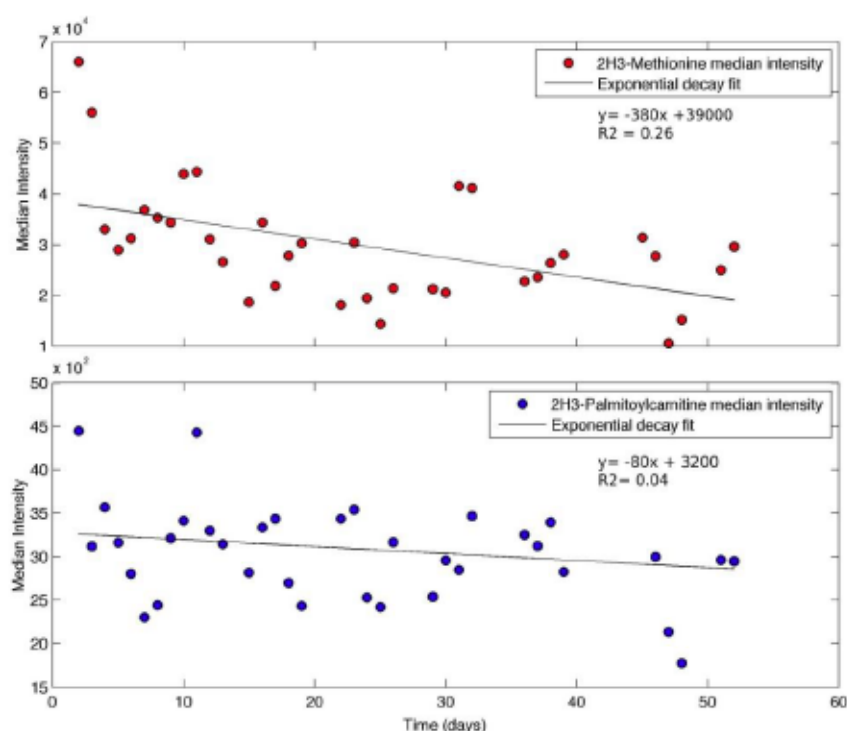


Figure 4.1 shows detection frequencies and coefficients of variation for the major $[M+H]^+$, $[M+Na]^+$, and $[M-H]^-$ adducts for each of the 20 isotopically labeled standards. Detection frequencies are typically high, especially for $[M+Na]^+$, and $[M-H]^-$, with the exception of acylcarnitine standards which are not detected in negative ionization mode and $^2H_3$ -Aspartate and $^2H_3$ -Glutamate, which appear to have low ionization efficiencies in positive ionization mode. However, coefficients of variation for the detected masses (excluding samples in which a given adduct was not detected) range from 40% to 360%. These values are high, considering that the standards should be present at a fixed concentration in DBS extracts.



Consensus intensities for internal standards and endogenous metabolites were calculated by averaging intensities of adducts in each mode as shown in equation 4.1. Using this method, consensus intensities were calculated for each of the 20 isotopically labeled internal standards and the variance of these intensities by batch

was investigated using balanced, one-way analysis of variance (ANOVA). For all standards, between-batch variance was greater than within-batch variance, as shown by the $\log_{10}$ variance and p-values calculated by ANOVA analysis (Figure 4.2). Notably, the difference between between-batch and within-batch variance does not seem to be related to the absolute level of variance.



As between-batch variance can be highly affected by a single or small number of batches, median intensities for a high variance internal standard and a low variance internal standard were calculated and plotted by the acquisition date of a given batch (Figure 4.3). For the high variance internal standard ($^2\text{H}_3$ -Methionine), median intensities appear to decay over time as batches run later in time show lower median intensities. Linear regression shows a slope of greater magnitude and a better fit for the high-variance internal standard. These results suggest that variance in internal standard intensities may introduce noise into the measurement of endogenous metabolites. As a result, evaluating preprocessing methods for the nS-HR-MS data

should compare both endogenous metabolite intensities relative to internal standards and the endogenous metabolite intensities alone.

4.3.2 Dried blood spot (DBS) metabolites

As with the internal standards, consensus intensities for the 238 metabolites listed in Appendix A, Table 2 were calculated from ions in both positive and negative mode according to eq. 4.1. Detection frequencies for metabolites approach 100% for over half of the metabolites targeted, but these frequencies drop to below 10% for a small number (Figure 4.4). While many mass spectrometry analysis packages exclude ions with low detection frequencies [98], some long-chain acylcarnitines that are important biomarkers for inborn errors have detection frequencies below 10%, as would be expected for sensitive and specific biomarkers of rare diseases.

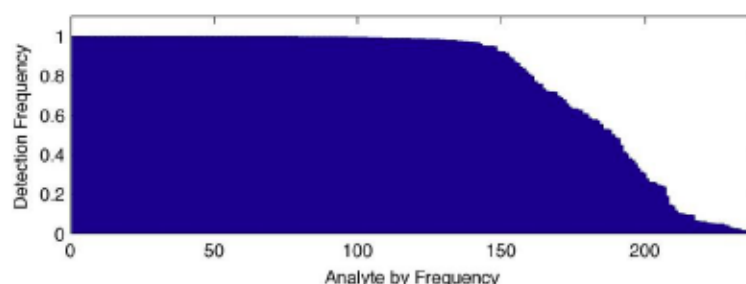
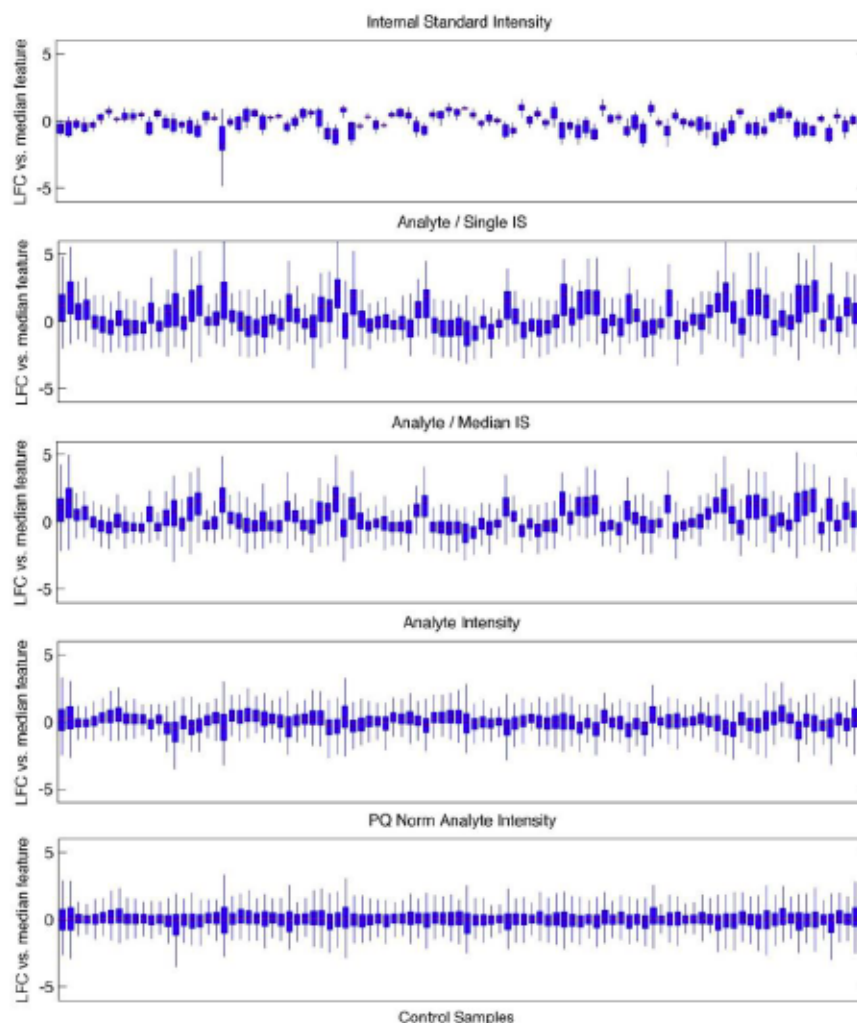


Figure 4.4: Detection frequencies for 238 metabolites. The graph shows that for the first 100 metabolites, the detection frequency is 1.0. For the next 100 metabolites, the frequency decreases to approximately 0.2, and for the remaining 38 metabolites, the frequency drops to 0.0.

Normalization is an important component of pre-processing metabolomics data, and to assess various normalization methods for nS-HR-MS data 100 samples were randomly selected from the 5721 good quality control samples. Distributions of log-fold-change vs. median were assessed for these 100 samples for the internal standard intensities alone, DBS metabolite / single internal standard ratios, where the reference internal standard was selected using the results from the pilot study [173], for metabolite / median internal standard, and for metabolite intensities alone. Box plots of these distributions are shown in Figure 4.5, applying the methodology to assess normalization discussed in Chapter 2. Here, sample distributions centered at 0

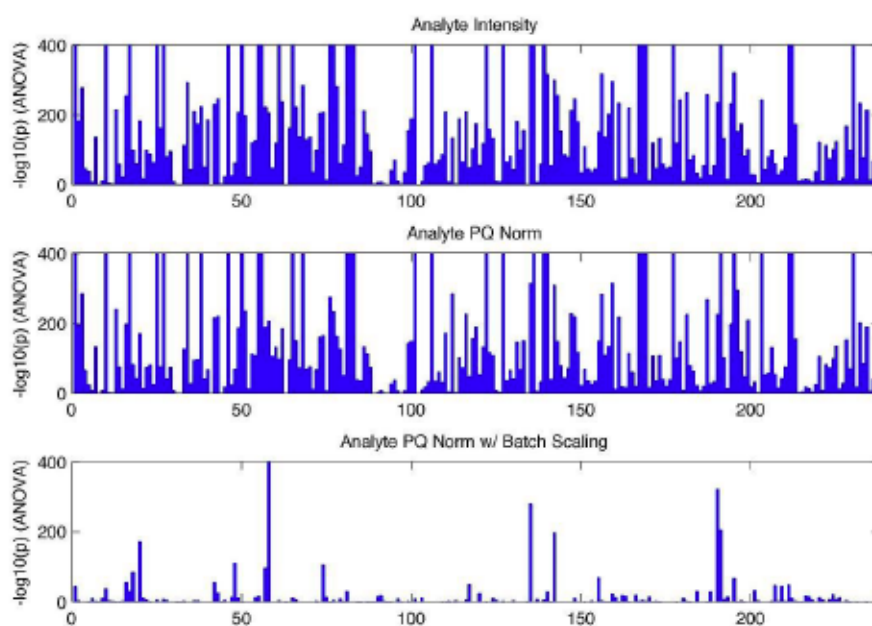
are roughly normalized, while samples with widely varying ranges show increased or decreased intensities for a large number of compounds.



These results show that while the internal standard levels tend to vary within a smaller range for each sample, there is large variance between samples. When DBS metabolite / internal standard ratios are calculated, the varying normalization of the internal standards impacts the normalization of the ratio. Strikingly, the DBS

metabolites themselves are roughly measured within the same ranges across the set of samples, and a standard probabilistic quotient normalization [88] applied to these intensities resolves the remaining normalization effects.

While processing the DBS metabolites without the internal standards does reduce variation between samples, batch effects are still a source of substantial variation, as ANOVA analysis by batch for the 238 compounds included in the targeted analysis show (Figure 4.6). To reduce these effects, the stabilized compound abundances were median centered by batch. Batch mean centering is an established technique in microarray analysis [177, 178]- batch median centering serves the same purpose while increasing the robustness of the method to the outlying observations expected in rare disease screening. Substantial reductions in between-group variation after median-centering indicates that while this correction does not resolve batch effects for all compounds, it does mitigate the effects of batch acquisition on the resulting data.

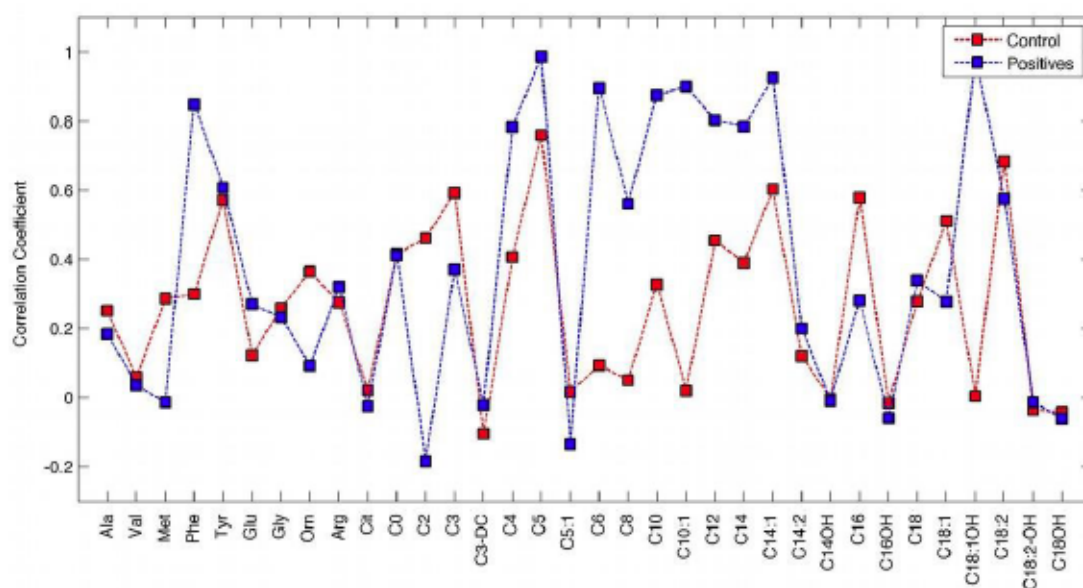


These results suggest that while the use of internal standards may allow a calculation of metabolite concentration in the original blood spot, high variance in the internal standard measurements reduce the utility of absolute quantification. Instead,

relative quantification as used in many metabolomics experiments yields more stable data.

4.3.3 Comparison of nS-HR-MS and MS/MS measurements

While the disparate approaches to the acquisition and processing of the nS-HR-MS and MS/MS data preclude direct, absolutely quantitative comparison, the nS-HR-MS and MS/MS data must both be sensitive to relative changes in compound levels for a sensitive and specific assay. Both MS/MS and nS-HR-MS methods detect a set of 34 metabolites, and correlations between the MS/MS and nS-HR-MS methods were assessed for all 5627 control and 102 positive samples with paired measurements. Figure 4.7 shows the correlations for control (red) and positive (blue) measurements by the two methods for each of the 34 compounds. Pearson product-moment correlations are calculated for the untransformed abundances – concentrations for MS/MS and compound intensities (equation 4.1) for nS-HR-MS.



Correlations between the reported MS/MS levels and nS-HR-MS levels vary from no correlation to near perfect correlation (0.99). Pearson correlations of raw

abundances are high when the variance in the metabolite concentration is high, such as when the metabolite level is affected by one of the conditions included in the dataset, and low when the variance is low. Relative correlations, including log-fold-change vs. median and rank correlations (not shown) are more stable, indicating that nS-HR-MS and MS/MS measurements agree on a relative rather than absolute scale.

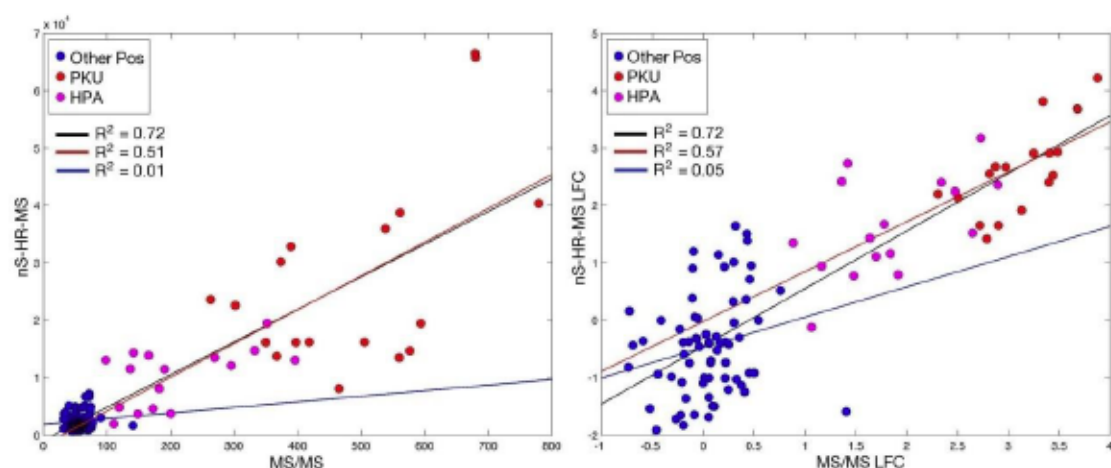


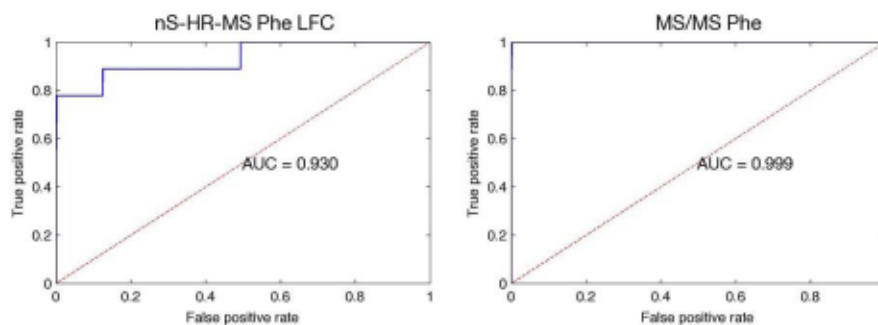
Figure 4.8 shows two scatter plots comparing nS-HR-MS and MS/MS data. The left plot shows nS-HR-MS (y-axis, scaled by 10^4) versus MS/MS (x-axis). The right plot shows nS-HR-MS LFC (y-axis) versus MS/MS LFC (x-axis). Both plots include data points for Other Pos (blue), PKU (red), and HPA (magenta), with regression lines and R^2 values.

To investigate the effects of increased metabolic variance on the correlation between MS/MS and nS-HR-MS data, Phe measurements were plotted for PKU (red), hyperphenylalaninemia (HPA, magenta), and other affected samples (blue) (Figure 4.8). While the correlation between Phe measurements tends to be lower for both PKU (red line) and non-PKU (blue line) samples alone, when PKU and non-PKU measurements are taken together (black line), agreement tends to be good. This means that both platforms reflect significant elevations or reductions in abundances due to metabolic perturbations. Notably, the slope for the fit line of the log-fold-change vs. median data was 1. While the raw abundances are on different scales (concentration vs. intensity), it appears that the MS/MS data separates the PKU, HPA, and other affected samples more accurately based on Phe levels alone than the nS-HR-MS data. This could be due to non-linearity in detection or the detection of other isobaric ions by nS-HR-MS.

4.4 Results and Discussion

4.4.1 Predictive modeling of PKU and MCADD

While IEM prevalence and the diseases included in screening programs vary between countries, phenylketonuria (PKU) and medium-chain acyl-CoA dehydrogenase deficiency (MCADD) are the most common inborn errors of amino acid metabolism and of mitochondrial fatty acid oxidation and screening programs for these diseases are in place in many countries internationally [179, 180]. These screening programs typically rely on MS/MS analysis of DBS samples combined with cutoff values for a small set of metabolites to identify affected individuals. These clinical algorithms are frequently set by individual screening centers, but published decision parameters include (Phe >120 $\mu\text{mol/L}$, Phe / tyr > 2) for phenylketonuria [180] and (C8>1 μM , C8/C10>1) for MCADD [181]. While some groups have proposed standardizing diagnostic algorithms and replacing cutoff values [172, 182], many labs still set these thresholds individually and modify them over time to capture previous positive and negative samples [172, 182, 183].

[illegible]

Newborn screening by serum metabolite measurement is arguably one of the first metabolomic tests to reach routine clinical use [184], but advances in metabolomics have paired advanced mass spectrometric measurement with multivariate data analysis to classify samples and identify biomarkers. Even recent

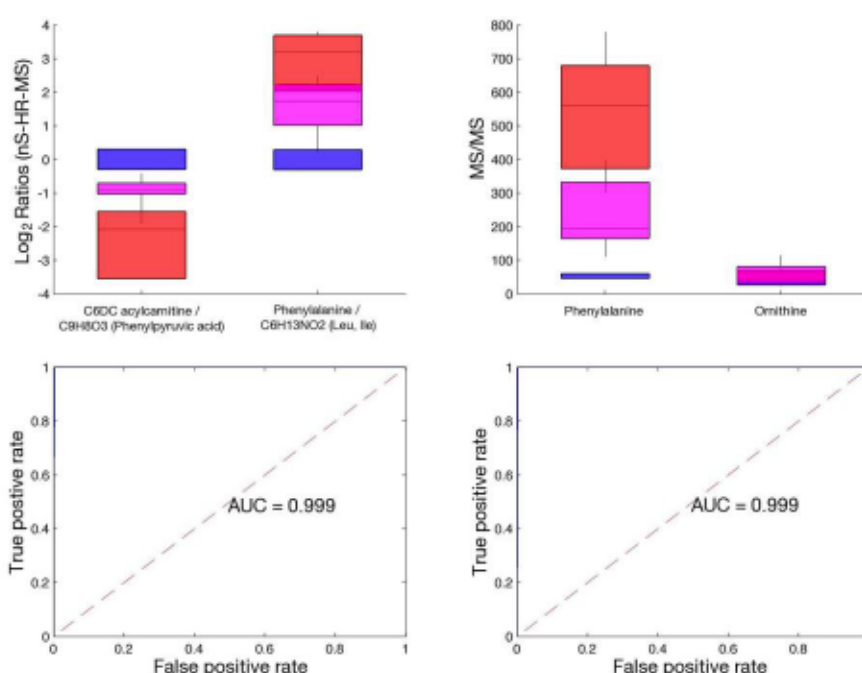
cutoff-free methods [182] have relied on percentile ranges and composite scores rather than pattern recognition. Here, sparse logistic regression methods are used to identify maximally predictive compounds and build models that predict the class of a sample based on the measurements of these markers by nS-HR-MS and MS/MS.

As Figure 4.8 shows, PKU, HPA, and other disorders are well separated by Phe concentration. As a result, receiver operating characteristic (ROC) curves for the classification of PKU samples by Phe concentration show high sensitivity and specificity across a range of Phe cutoff values resulting in an area under the curve (AUC) approaching unity (Figure 4.9). In ROC analysis, AUCs of 0.5 reflect random classification, while AUCs of 1 reflect perfect classification [185]. Phe measurements by both MS/MS and nS-HR-MS show high AUCs, though as noted previously the MS/MS measurements separate PKU, HPA, and unaffected samples more sensitively and specifically than nS-HR-MS data alone. High AUCs for PKU mean that a range of cutoff values result in high sensitivities and specificities, which allows for some choice in the precise value of a cutoff. Cutoff values suggested in the literature [180] yield a sensitivity and specificity of 100% and 99.8% for these samples. While the optimal cutoff value for these samples would be (Phe > 260 $\mu\text{mol/L}$), this estimation is only useful for illustration as it would have to be cross-validated using an independent test set.



Separating the set of 5627 control and 102 positive samples into test and train sets of equal size and applying forward logistic regression as outlined previously to training sets for nS-HR-MS and MS/MS yield the models shown in Table 4.2 and 4.3. Results of the application of these predictive models to test sets are shown in Figure 4.10 and 4.12. In Figure 4.10, box plots show the test set ranges for the variables

included in the forward variable selection process for true PKU (red), true HPA (magenta) and true controls and non-PKU/HPA affected (blue). ROC curves for classification by log-odds prediction for the test-set show that both nS-HR-MS and MS/MS yield highly sensitive and specific classification with AUC > 0.999. At 100% sensitivity these models both yield specificities of 99.8% for the test set. It is notable that both variables selected by forward logistic regression were ratios for nS-HR-MS and single compound abundances for MS/MS, and that the use of ratios in classification for nS-HR-MS closes the sensitivity and specificity gap relative to MS/MS based on Phe measurements alone.



While PKU is detected with high sensitivity and specificity on the basis of a single marker, medium-chain acyl-CoA dehydrogenase deficiency (MCADD) presents a more challenging case. Octanoylcarnitine (C8) is frequently measured by MS/MS to identify MCADD, with characteristic elevations in hexanoylcarnitine (C6), octanoylcarnitine (C8), decanoylcarnitine (C10), and decenoylcarnitine (C10:1) [186]. Alternate tests for MCADD have been proposed, including a test for leukocyte

MCAD enzymatic activity [187] and for urinary hexanoylglycine [188]. Although the primary genetic background in clinically diagnosed MCADD is c.985A>G homozygosity, the genetic background of MCADD is heterogeneous with the c.985A>G mutation making up only 55% percent of the screening population and less in non-European populations [189], yielding genetic screens potentially less sensitive than metabolic screens.

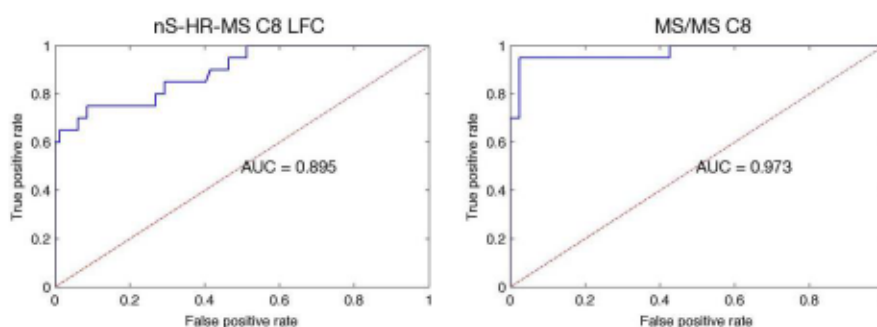
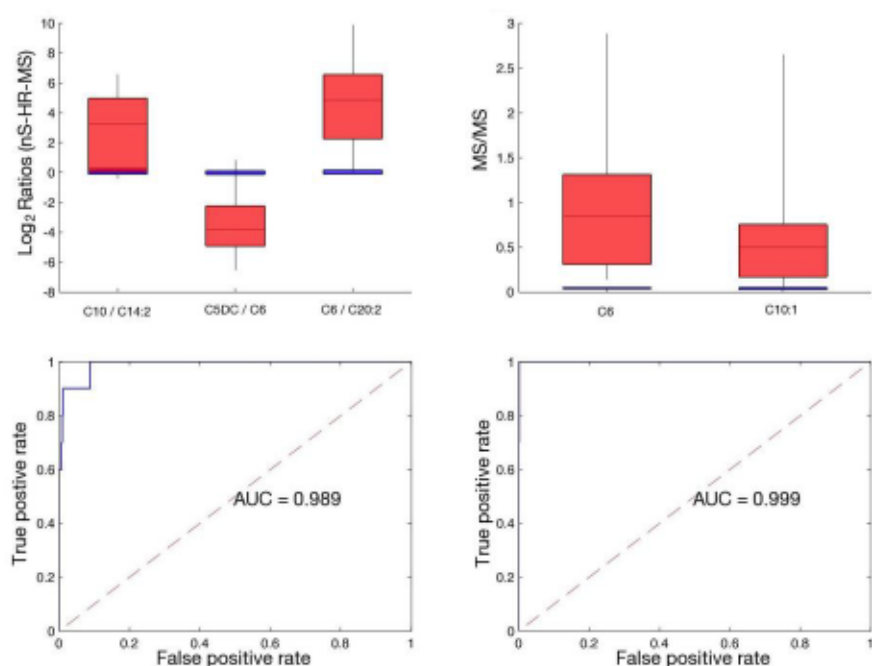


Figure 4.11 shows the ROC curves for the classification of MCADD DBS samples on the basis of C8 levels alone for nS-HR-MS and MS/MS. While MS/MS still does show a high AUC > 0.97, even the optimal cutoff value of (C8>0.1) for this dataset gives a false positive rate of 13.2% (specificity 86.8%) and this is biased by setting the cutoff without an independent test set. Published cutoff values (C8>1uM, C8/C10>1, [181]) yield a sensitivity of 65% and a specificity of 100% for the MS/MS data collected here. Applying forward logistic regression to test and train sets for nS-HR-MS and MS/MS data for MCADD generates the model shown in Table 4.3, and application of this predictive model to the test set yields the results shown in Figure 4.12. As with Figure 4.10, box plots show the test set ranges for variables in the logistic regression model for true MCADD (red) and true non-MCADD (blue). Both models yield AUCs well above those from classification based on C8 alone for their respective measurement platforms. At 100% sensitivity, the nS-HR-MS model yields a specificity of 91.1% and the MS/MS model yields an impressive 99.9%, indicating both models give fewer false positives than either published cutoff values or the optimal cutoff for C8 across the test and train sets.



The model was trained on a dataset of 100 samples, including 9 MCADD true-positive, 3 false-positive, and 24 true negative. The model was evaluated on an external test-set of 36 samples, including 9 MCADD true-positive, 3 false-positive, and 24 true negative. The model achieved a 100% correct classification rate on the external test-set.

The model was trained on a dataset of 100 samples, including 9 MCADD true-positive, 3 false-positive, and 24 true negative. The model was evaluated on an external test-set of 36 samples, including 9 MCADD true-positive, 3 false-positive, and 24 true negative. The model achieved a 100% correct classification rate on the external test-set.

Model	Intercept	Coefficients	Metabolites
Log ₂ Ratios (nS-HR-MS)	-7.9881	0.13	C10 / C14:2
		-1.42	C5DC / C6
		3.78	C6 / C20:2
MS/MS (μmol / L)	-11.4964	95.48	C6
		61.46	C10:1

An external test-set of 36 samples including 9 MCADD true-positive, 3 false-positive, and 24 true negative were provided by National Institute for Health Data and Disease Control (Copenhagen, Denmark). These samples constitute a truly independent test set as they represent different neonatal nationality, different sample collection locations, and the data was preprocessed independently from the set of control and affected samples from the German and Hungarian centers. Prediction of MCADD positives based on the model in Table 4.3 yields 100% correct classification

for both true positives and true negatives, as well as for samples which were identified as positives by MS/MS screening in Denmark but were later found to be unaffected. Box plots in Figure 4.13 indicate ranges for the predictive variables for MCADD true positive (red), MS/MS false positives (magenta), and true negatives (blue).

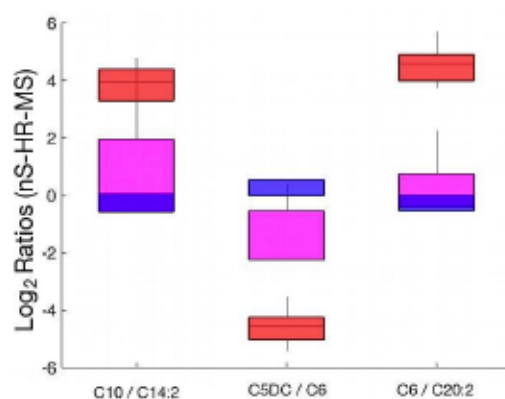


Figure 4.13: Box plots of Log₂ Ratios (nS-HR-MS) for three lipid ratios: C10 / C14:2, C5DC / C6, and C6 / C20:2. The y-axis ranges from -6 to 6. For each ratio, there are three box plots: red (top), magenta (middle), and blue (bottom). The red box plots represent true positives, the magenta box plots represent MS/MS false positives, and the blue box plots represent true negatives. The C10 / C14:2 ratio shows a clear separation between the groups, with true positives having the highest ratios and true negatives the lowest. The C5DC / C6 ratio shows a similar pattern, but with more overlap between the groups. The C6 / C20:2 ratio shows a clear separation between the groups, with true positives having the highest ratios and true negatives the lowest.

While models based on a small number of affected individuals – (n=18) for PKU and (n=19) for MCADD – must be validated by larger sample sizes and across a broader set of screening facilities and populations, these results indicate that predictive modeling for both nS-HR-MS and MS/MS measurements can improve upon the cutoff based methods that are currently standard in screening centers internationally. Additionally, though non-parametric pattern recognition methods such as random forest classification, neural networks, or support vector machines may improve upon logistic regression [190], logistic regression based on a small number of features was applied here as it is well understood, routinely applied in medicine [191] and does not greatly exceed the current measurement capabilities of many screening facilities. Multivariate methods based on predictive modeling should substantially aid the search for cutoff-free methods for newborn screening [182].

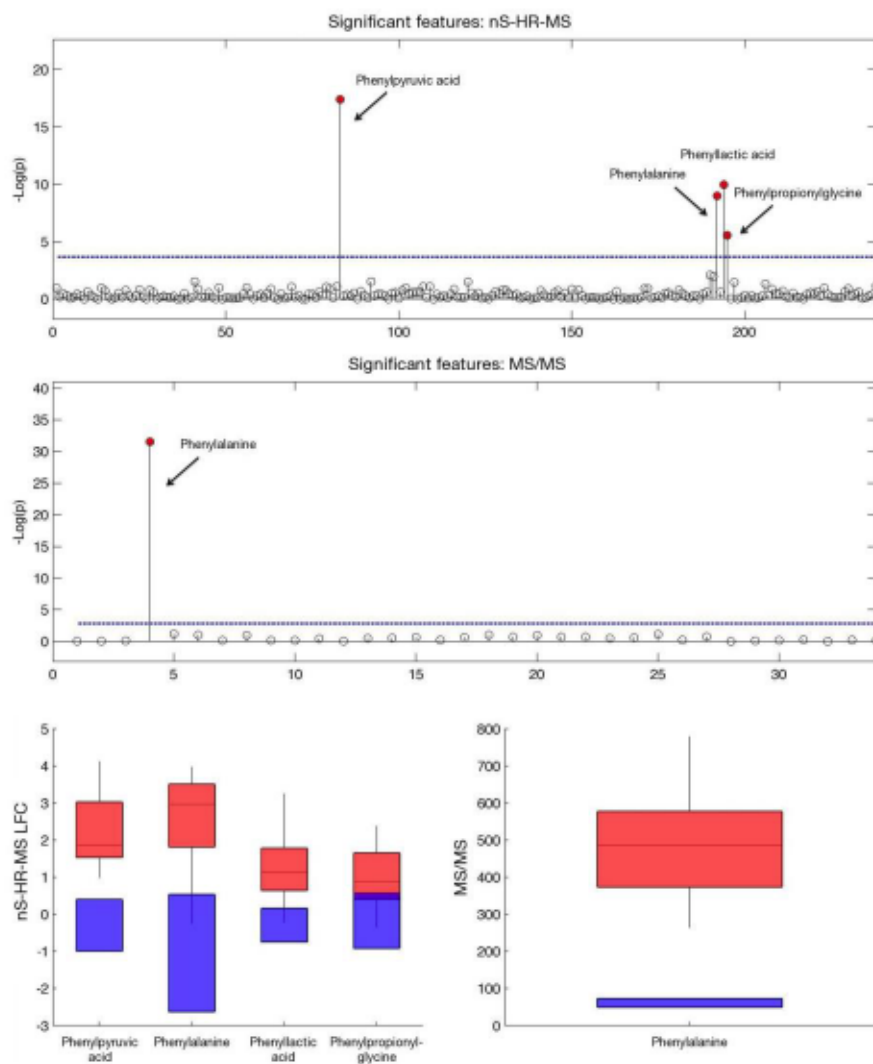
4.4.2 Biomarker identification

Broader metabolomic coverage in newborn screening has the potential to improve the sensitivity and specificity of diagnostic tests, but identification of a more comprehensive set of metabolic perturbations resulting from genetic defects could contribute greatly to the understanding of genetics, metabolism, and the relationships between genes and metabolites in health and disease [192]. Understanding the impact of genetic variation on health and disease is a critical task in modern medicine and life sciences. Mammalian metabolism is affected by genes, environment, nutrition, microbiota, and metabolite levels in biofluids are affected by all these factors [193]. **This complexity poses a significant challenge for “bottom up” prediction of the effects of genetic variation at the systems level [194], and as genomic advances increasingly shed light on human genetic diversity functional annotation of variants has become one of the main challenges in genetic research [154, 195].**

While several attempts have been made to connect genetic variants to metabolic perturbations through computational modeling [175] and regression of single-nucleotide polymorphisms (SNPs) against biofluid metabolite levels [38, 39, 44, 176], as of 2012 there were approximately 3,500 known Mendelian diseases for only ~20,000 known human genes [51]. Recent statistics discussed in Chapter 1 suggests this number will continue to rise. Even neglecting these new Mendelian diseases that will be discovered as sequencing efforts expand, potential functional analysis of nearly 20% of the human genome provides a strong rationale for studying the functional consequences of Mendelian disease. Additionally, while the term “Mendelian disease” suggests single genetic cause, **there is growing appreciation for the importance of “modifier” variations [196] and for the genetic heterogeneity of even straightforward single-gene diseases [58, 197].** While these effects complicate genetic screening for inborn errors, genetic follow up in conjunction with metabolic screening could help clarify the complex genetics of IEMs and aid functional assessment of rare variation.

The expanded set of metabolites measured by nS-HR-MS, the large number of control samples, and the number of diseases included in the newborn screening data presented here provides an opportunity to identify a broader set of perturbed metabolites specific to a range of inborn errors. **Figure 4.4.1 illustrates the potential for identifying a broader set of perturbed metabolites specific to a range of inborn errors by comparing the metabolite levels in newborn screening data to the metabolite levels in control samples.**

The results of the nS-HR-MS and MS/MS analysis are shown in Figure 2. The nS-HR-MS analysis identified four significant features (p < 2.1e-04 for nS-HR-MS and p < 0.0015 for MS/MS) - r

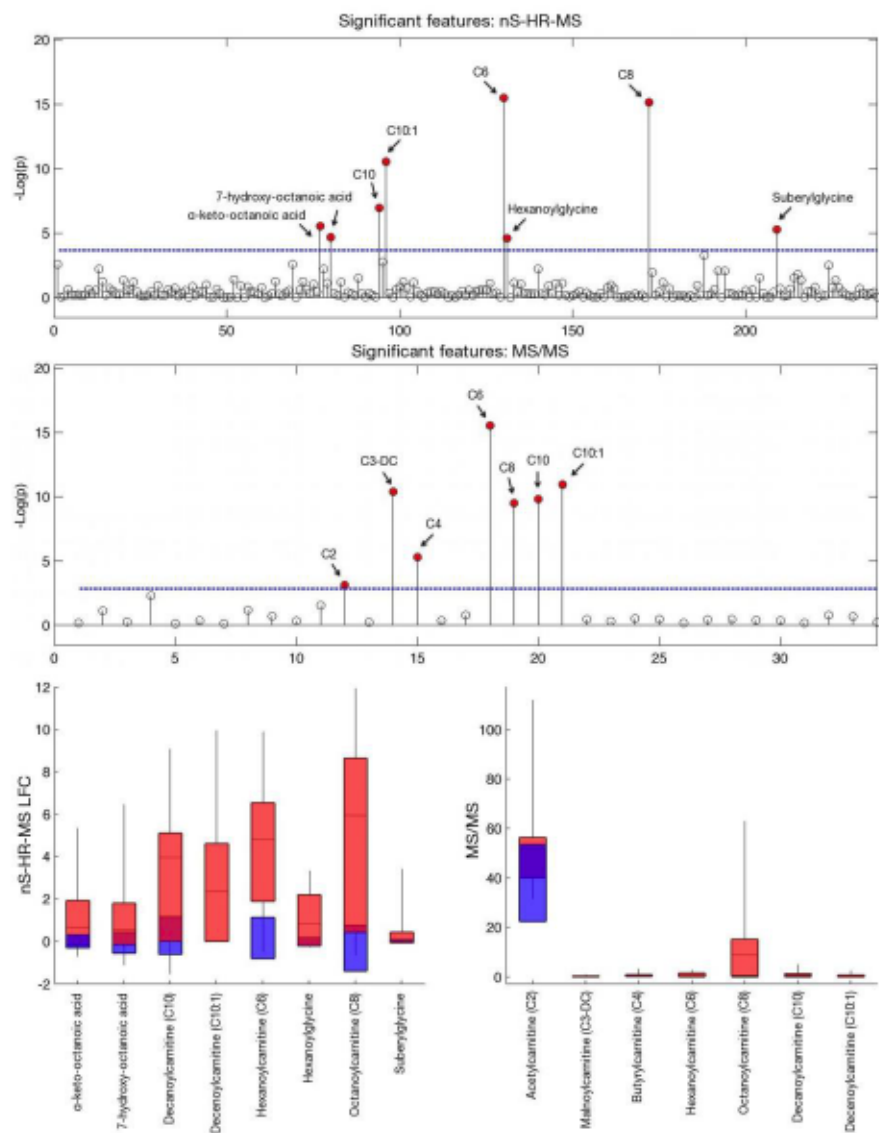


The results of the nS-HR-MS and MS/MS analysis are shown in Figure 2. The nS-HR-MS analysis identified four significant features (p < 2.1e-04 for nS-HR-MS and p < 0.0015 for MS/MS) - r

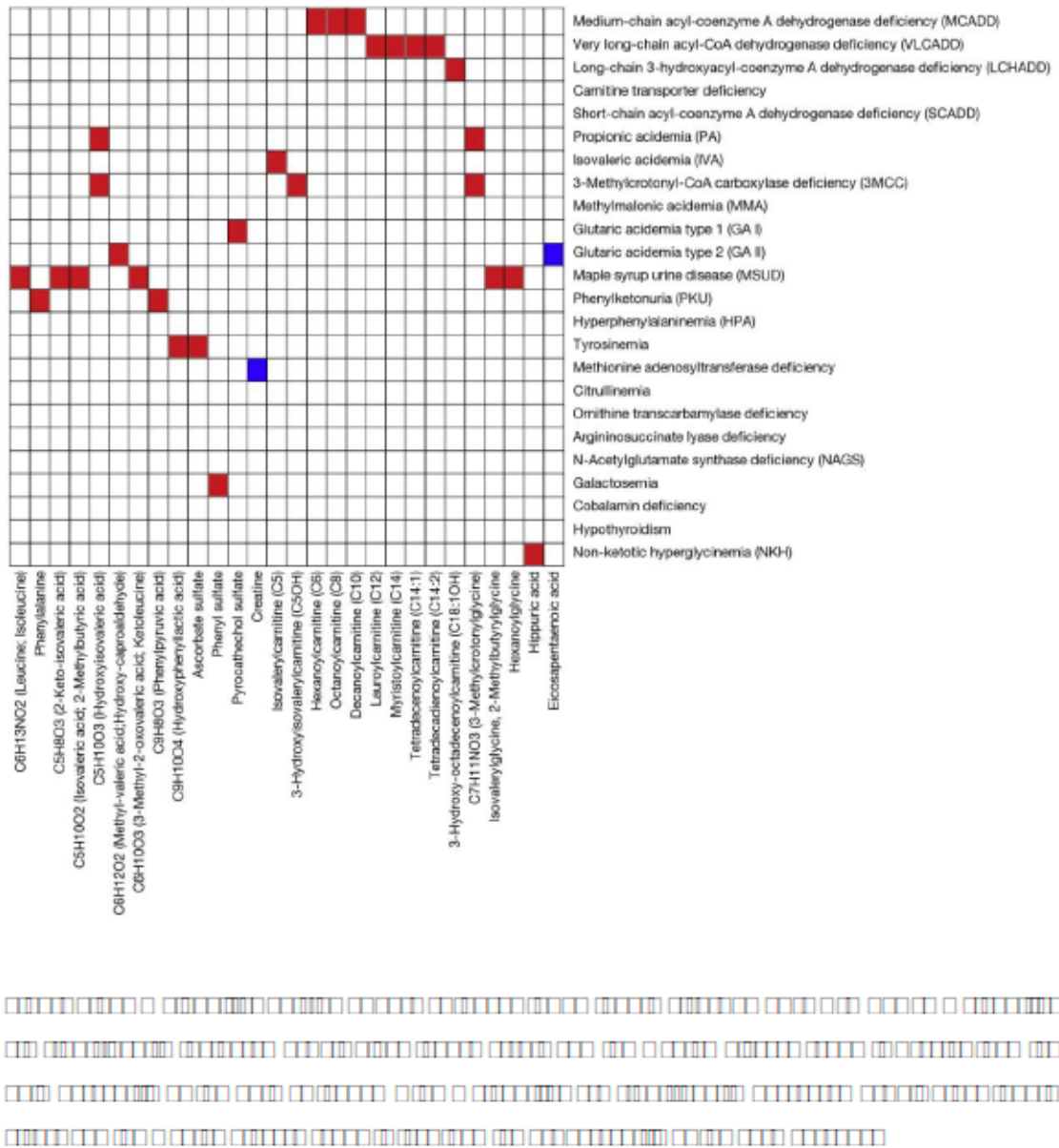
amphoteric metabolite, previous work has shown it is eliminated by renal filtration and as a result appears in large quantities in the urine of PKU patients [199].

Medium-chain acyl-CoA dehydrogenase (MCAD) catalyzes the rate-limiting step in mitochondrial beta-oxidation for fatty acids with chain lengths between 6 and 12 carbons [201]. MCAD deficiency can result in fatty acid oxidation failure following fasting, metabolic decompensation, and as a result is a major cause of sudden infant death syndrome [179]. Figure 4.15 shows the results of a two-way, two-tailed one vs. all t-test of MCADD positive samples vs. all other positive samples for nS-HR-MS and MS/MS measurements. MS/MS shows characteristic elevations of hexanoylcarnitine (C6), octanoylcarnitine (C8), decanoylcarnitine (C10), and decenoylcarnitine (C10:1) [179], as well as some shorter chain acylcarnitines.

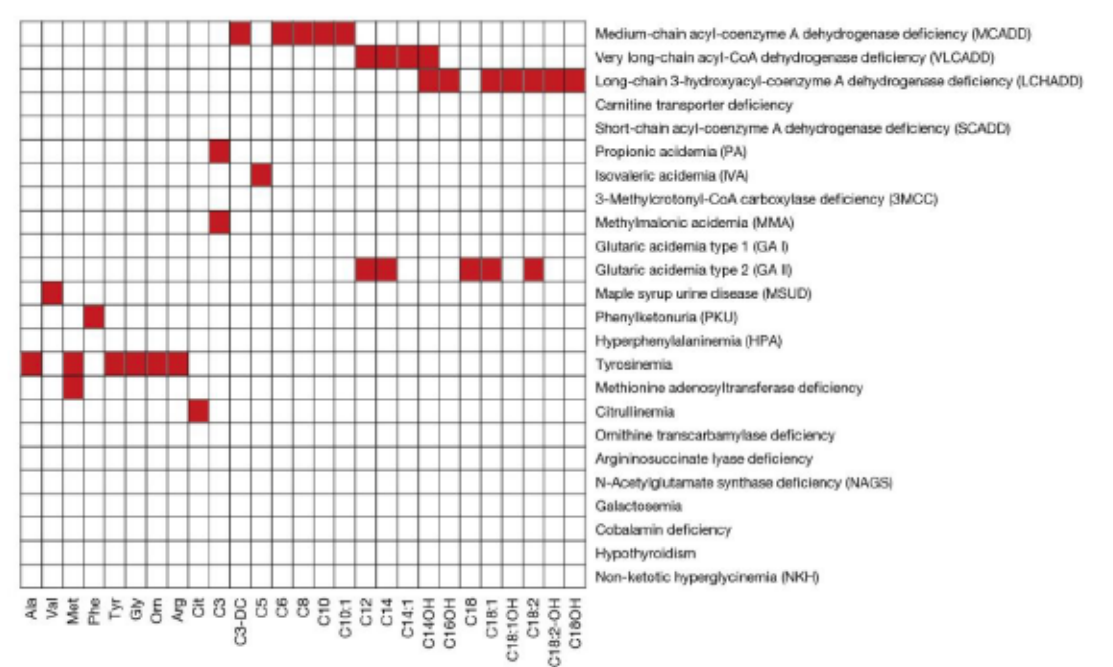
In addition to the primary acylcarnitines, nS-HR-MS detects acylglycines and fatty acids. Acylglycines are produced by glycine conjugation of acyl-CoAs and the C6 and C8 acylglycines suberylglycine and hexanoylglycine have been detected in the urine of MCADD patients [202], and these have been proposed as additional tests for MCADD screening [188]. Additionally, C8 fatty acids are detected, including hydroxy-octanoic acids and keto-octanoic acids. While the position of the hydroxyl and carbonyl cannot be verified based on parent mass, 3-, 6- and 7-hydroxyoctanoic acids have been identified as products of fatty acid metabolism [203] and have been identified as elevated in patients with abnormal fatty acid metabolism [204]. Previous reports identify increased medium-chain fatty acids, dicarboxylic acids, acylglycines, and medium-chain acylcarnitines in plasma and urine as shunt metabolites of elevated medium-chain acyl-CoA esters [205].



MCADD and PKU are among the most common inborn errors of metabolism, and the resulting availability of a greater number of samples yields greater statistical power to identify perturbed metabolites. To detect potential markers for the less common IEMs, a dual filtering strategy was used. Metabolites were considered a potential marker of an individual IEM if its p-value by one vs. all ≤ 0.001 if the median level of the compound in the affected samples was in the extreme 1% of the levels of all control samples. The markers identified for each of the IEMs in Figure 4.16 for nS-HR-MS and Figure 4.17 for MS/MS.



In these matrices, red elements denote an elevated metabolite for a given IEM, while blue elements denote a metabolite with decreased levels relative to controls and other IEM affected patients.



Results of note include the detection of perturbed mitochondrial fatty acid β -oxidation and resulting elevation of acylcarnitines for MCADD, very long-chain acyl-CoA dehydrogenase (VLCAD) deficiency, and long-chain 3-hydroxyacyl-CoA dehydrogenase (LCHAD) deficiency. For MCADD and VLCADD, chain-lengths of the perturbed metabolites correspond to the major activities of the MCAD and VLCAD enzymes. The presence of 3-hydroxylated acylcarnitines is characteristic of LCHAD deficiency, as LCHAD is a critical component of mitochondrial trifunctional enzyme (MTP) along with the long-chain enoyl-CoA hydratase (LCEH) and 3-ketoacyl-CoA thiolase. While *HADHA* encodes for the protein with both LCHAD and LCEH activities, the majority of affected patients show primarily LCHAD deficiency [206]. Markers of short-chain acyl-CoA dehydrogenase (SCAD) deficiency include C4 and C6, which were likely not detected due to low sample numbers.

Propionic academia (PA), resulting from propionyl-CoA carboxylase deficiency, and 3-methylcrotonyl-CoA carboxylase (3MCC) deficiency are both inborn errors of branch chain amino acid degradation of valine and isoleucine (PA) and leucine (3MCC deficiency). Both PA and 3MCC deficiency show elevated 3-methylcrotonylglycine and 3-hydroxyisovalerate. Both are direct shunt metabolites of 3-methylcrotonyl-CoA, where unmetabolised 3-methylcrotonyl-CoA is converted to 3-methylcrotonylglycine by glycine-N acetyltransferase and to 3-hydroxyisovaleryl-CoA by enoyl-CoA hydratase and then to 3-hydroxyisovalerate [207]. These and other metabolic intermediates of branch chain amino acid degradation have previously been observed in propionic academia [208]. Corresponding acylcarnitines – 3-hydroxyisovalerylcarnitine (C5OH) for 3MCC deficiency and propionylcarnitine (C3) for PA – were also elevated. Patients with isovaleric academia, another inborn error of branch-chain amino acid degradation, showed increased isovalerylcarnitine (C5) by both nS-HR-MS and MS/MS. Branched-chain keto acid dehydrogenase is the second step in the degradation of leucine, isoleucine and valine, and deficiencies in this enzyme cause Maple Syrup Urine Disease (MSUD) [209]. MSUD is characterized by urinary excretion of branch-chain amino acids and their keto-acid derivatives, all of which were identified as elevated by either MS/MS or nS-HR-MS. Figure 4.18 shows the accumulation of the compounds detected by MS/MS and nS-HR-MS due to branched-chain α -ketoacid dehydrogenase deficiency in MSUD mapped onto human metabolic pathways from KEGG (hsa01100).

Tyrosinemia is an inborn error caused by deficiencies in one of three enzymes in the tyrosine degradation pathway [210]. Tyrosinemia type 1, caused by fumarylacetoacetate (FAA) hydrolase deficiency, and tyrosinemia type 3, caused by 4-hydroxyphenylpyruvate dioxygenase deficiency, lead to increased urinary excretion of 4-hydroxyphenylpyruvate derivatives including hydroxyphenyllactic acid, while tyrosinemia type 2, which is caused by inability of tyrosine aminotransferase to catalyze the first step in tyrosine metabolism, does not [211, 212]. Hypertyrosinemia was detected by MS/MS, and elevated hydroxyphenyllactic acid was detected by nS-HR-MS. Interestingly, nS-HR-MS also identified elevated ascorbate sulfate, a metabolite of the vitamin C supplements used to treat neonatal tyrosinemia [213]. A similar finding was elevated citrate, administered to citrullinemia patients [214], detected by MS/MS.



While most of the perturbed metabolites identified here were elevated, a small number were decreased relative to the control population and the other inborn errors. Eicosapentaenoic acid (EPA) lower than 99% of the control population for the glutaric acidemia type 2 patient, which could be due to impaired fatty acid biosynthesis due to multiple acyl-CoA deficiency. Also potentially informative is the decreased creatine found by nS-HR-MS for the methionine adenosyltransferase deficiency patient. While hypermethioninemia, the classic marker of methionine adenosyltransferase deficiency, was identified by MS/MS, decreased creatine been postulated as a metabolic consequence of methionine adenosyltransferase deficiency as creatine synthesis by guanidinoacetate methyltransferase (GAMT) requires S-adenosylmethionine [215]. Though decreased creatine has not been previously detected in individuals with methionine adenosyltransferase deficiency, the identification of GAMT deficiency and impaired neurological development due to impaired creatine synthesis suggested a possibility for a role in impaired creatine

synthesis in the analogous neurological impairment seen in methionine adenosyltransferase patients [16, 215, 216].

4.5 Conclusions

Metabolic profiling is an important component of national newborn screening programs and is an increasingly important tool for the functional characterization of genetic variants. This chapter evaluates the applications of two mass spectrometric assays, traditional electrospray ionization tandem mass spectrometry (ESI-MS/MS) with multiple reaction monitoring and a new direct injection nanospray high resolution mass spectrometry (nS-HR-MS), to classification and biomarker discovery for inborn errors by dried blood spot (DBS) analysis. While a larger number of affected patient samples will be necessary to assess diagnostic accuracy at a level comparable to the studies performed by international newborn screening programs [172], results for phenylketonuria (PKU) and medium chain acyl-CoA dehydrogenase deficiency (MCADD) indicate that increased metabolic coverage together with predictive modeling can improve upon cutoff based diagnostics for these diseases.

Additionally, while the combination of expanded coverage metabolic profiling and predictive modeling holds promise for improving diagnostic tests for newborn screening, a second major benefit of expanding metabolic coverage in screening assays is the opportunity to identify metabolic changes incidental to and associated with the main metabolic lesion. Examples of these identified in this work include shunt metabolites for acyl-CoA dehydrogenase deficiencies and errors of amino acid degradation and metabolites with decreased blood levels relative to unaffected newborns. Identifying these secondary perturbations could improve understanding of mechanisms of pathology, help clarify spectrums of clinical and genetic variation, and aid the functional assessment of rare variants.

CHAPTER 5

METABOLOMIC INVESTIGATION OF CONGENITAL DISEASES OF THE KIDNEY

5.1 Introduction and aims

Mutations in enzymes and transporters can often be readily diagnosed by biofluid profiling and a strong mechanistic link is easily inferred between the disruption of a metabolic pathway and resulting accumulation or depletion of metabolic intermediates. However, many Mendelian diseases result in more complex, progressive organ specific or multi-organ pathology [25]. Examples include transporter deficiencies, organelle disorders such as lysosomal and peroxisomal storage diseases, congenital cancers, mistrafficking, or signalling disorders. While these complex phenotypes frequently have systemic effects, individual organs or tissue types may be affected to varying degrees. In these cases, metabolic profiling may be applied to identify sites of lesions, describe progression, and attempt to identify proxy small molecular biomarkers of the disease [22-24].

The utility of urinary metabolic profiling in nephrology has long been noted in both the metabolomics and nephrological communities [217-219]. Urinary metabolic profiling has been applied to study uremic toxins [220], acute kidney injury induced by exposure to nephrotoxins [161, 221-224], diabetic nephropathy [225-227], polycystic kidney disease [24, 228], renal Fanconi Syndrome [23], renal cell carcinoma [229, 230], and to monitor kidney transplantation [217, 231, 232]. In addition to using metabolic profiling to identify urinary compounds differentiating

nephrologically stressed humans and animals from controls, metabolic profiling has been used to identify characteristic urinary changes resulting from region specific acute toxicity [233, 234]. This work has shown that characteristic urinary metabolic profiles differentiate renal tubular toxins from renal medullary toxins, for example. While it has been suggested that studying inborn errors may aid toxicological research [235], metabolomics studies of nephrotoxins suggest the reverse is true as well.

This chapter presents a comparison of the urinary metabolic profiles of four congenital kidney diseases, a natural history study showing changes in the urinary profile of cystinosis patients over time and with age, glomerular filtration rate (GFR), drug therapy, and transplantation, and a comparison of the urinary perturbations seen in these human Mendelian disorders to the perturbations induced by region specific nephrotoxins. While the human renal disorders discussed here have quite distinct genetic causes, they result in an overlapping set of phenotypic consequences including Fanconi Syndrome (FS), progressive renal failure, nephrocalcinosis, and myopathy. These phenotypes suggest pathology affecting at least the renal proximal tubule, glomerulus, and renal interstitium. This chapter characterizes urinary metabolic associations with these phenotypes by comparing these conditions and investigating changes over time as these phenotypes develop.

5.2 Materials and Methods

5.2.1 Molecular biology and patients

5.2.1.1 Cystinosis

Cystinosis is an autosomal recessive lysosomal storage disease caused by loss-of-function mutations in the *CNTS* gene encoding cystinosin, the lysosomal membrane transport protein responsible for the efflux of cystine from lysosomes [158, 236-238]. As a consequence, intra-lysosomal cystine accumulates in many cell types, most of which contain crystals. Clinically, renal tubular cell death and FS develop in the first year of life, manifesting as polyuria, phosphaturia, glucosuria, and generalized aminoaciduria [239]. Cystinosis historically resulted in fatal end-stage

renal disease, but renal transplantation and treatment with the cystine-depleting agent, cysteamine, have significantly increased the lifespan of cystinosis patients [240-242].

Twenty-two cystinosis patients were investigated at the Clinical Center of the National Institutes of Health, USA, between 1995 and 2010 as part of clinical protocol 78-HG-0093, approved by the NICHD and NHGRI IRBs. Patients or their parents gave written, informed consent. Twenty-four hour urines were collected and aliquots stored at -30C. Sample collections and analyses were performed according to Imperial College Research Ethics Committee (ICREC) protocols approved by Imperial College London. Patient data, including transplantation status, are included in Table 5.1. Between 7 and 15 samples were collected from each patient over this time period, with urines collected from patients between the ages of 3 weeks and 49 years; the mean age at sample collection was 13 years. Of the 22 patients included in the study, 12 ultimately underwent renal transplantation during the period of collection. Following diagnosis, all patients were prescribed Cystagon^R (cysteamine bitartrate), generally at a dose of 60-90 mg/kg/day divided q6h. After transplantation, anti-rejection medications such as mycophenolic acid were given.

Table 5.1. Cystinosis patient characteristics

Patients	N=22
Age	3 wks – 49 years (mean 13 years)
Gender	11 Male, 11 Female
Samples per Patient	7 – 15 (mean 9 samples)
Patients with pre-transplant samples	N=19
Patients with post-transplant samples	N=12
Cystagon prescribed dose	75 – 850 mg (mean 437 mg)
Glomerular filtration rate	7– 176 mL/min (mean 59 mL/min)

5.2.1.2 Autosomal Dominant Fanconi Syndrome 1 (ADFS1)

Fanconi Syndrome is a proximal tubular cell dysfunction that renders the renal tubule unable to recover small molecule solutes from the ultrafiltrate leading to the urinary loss of solutes including glucose, amino acids, and ions including phosphate, calcium, sodium, potassium, and bicarbonate. While the most common cause of

pediatric Fanconi Syndrome is cystinosis [158], FS results from other known Mendelian diseases, including Oculocerebrorenal Syndrome of Lowe [243], Dent disease, Wilson disease, and tyrosinemia. While these diseases generally show an autosomal recessive inheritance pattern, Fanconi Syndrome with an autosomal dominant inheritance pattern was recently identified in a five-generation African American family [160]. Affected family members presented with FS but without the glomerular kidney failure observed in cystinosis. Genetic linkage studies subsequently mapped this phenotype to chromosome 3 and sequencing subsequently identified a heterozygous missense mutation in *EHHADH* that created a de novo mitochondrial targeting motif leading to mistrafficking of L-Bifunctional Protein, an enzyme originally involved in peroxisomal fatty acid oxidation, to the mitochondria.

Urine samples were collected from a total of (N=10) patients and family members from a five-generation African American family with isolated autosomal dominant FS. Seven members of a family with generalized FS were admitted to the NIH Clinical Center and enrolled in clinical protocol 76-HG-0238. Urine samples were collected from 5 of these patients. Eleven unaffected family members were seen locally, and urine samples were collected from 5 family members. All gave written, informed consent. The diagnosis of FS was established by routine laboratory investigations of urine and blood. All investigations, including genetic studies, were approved by the NHGRI Institutional Review Board and conducted according to the principles of the Declaration of Helsinki.

5.2.1.3 Autosomal Dominant Fanconi Syndrome 2 (ADFS2)

A second phenotype also presenting with renal Fanconi Syndrome and an autosomal dominant inheritance pattern was studied in members of four unrelated families, yet did not show the *EHHADH* mutations characteristic of ADFS1 patients. Clinical work-up identified abnormal looking mitochondria in affected family members, suggesting a mitochondrial dysfunction distinct from other observed forms of Fanconi Syndrome. As such, this as yet genetically uncharacterised form of Fanconi Syndrome was termed Autosomal Dominant Fanconi Syndrome type 2 (ADFS2).

Urine samples were collected from a total of 12 members of these families, of which (N=8) affected individuals provided samples. One affected individual provided

an additional sample at a later collection date. Participants were either admitted to the NIH Clinical Center in the USA and enrolled in clinical protocol 76-HG-0238 or enrolled in the protocol “Identification of genes involved in renal and electrolyte disorders” and seen in the United Kingdom. All gave written, informed consent. The diagnosis of FS was established by routine laboratory investigations of urine and blood. All investigations, including genetic studies, were approved by the NHGRI Institutional Review Board or Royal Free Hospital Research and Development and conducted according to the principles of the Declaration of Helsinki.

5.2.1.4 Enamel Renal Syndrome (ERS)

Enamel Renal Syndrome (ERS), resulting in amelogenesis imperfecta (AI) and nephrocalcinosis (NC) is an autosomal recessive disorder of calcium homeostasis caused by mutations in the *FAM20A* gene [244-247]. *FAM20A*, together with *FAM20B* and *FAM20C*, are involved in glycosaminoglycan synthesis, and experimental evidence suggests that loss of *FAM20A* function results in overabundance of chondroitin sulfate in the extracellular matrix. Clinically, ERS patients show abnormal enamel formation and tooth eruption as well as calcifications in the kidney, teeth, and gingival. Unlike other forms of nephrocalcinosis such as **Bartter’s syndrome, Dent’s Disease, and hypophosphataemia** with nephrolithiasis that involve defects in calcium transport from the renal tubule to the interstitium and result in hypercalciuria [248], the nephrocalcinosis in ERS results from increased sequestration of ionized calcium in the interstitium by chondroitin sulfate rich glycosaminoglycans [246].

Urine samples were collected from a total of (N=14) patients examined as part of the international genetic characterization of ERS [246]. Patients and families were identified in rare disease renal tubular or dental / craniofacial reference centers and gave informed consent. All patients had NC confirmed by either ultrasound, X-ray or CT, and all showed pathognomonic teeth findings with AI and delayed or missing tooth eruption. This study was approved by the institutional review boards and ethics committees of the various centers, details of which have been previously published [246].

5.2.2 Samples and data acquisition

5.2.2.1 Comparison of congenital kidney diseases

Spot urines were collected and aliquots stored at -30°C. Sample collections and analyses were performed according to Imperial College Research Ethics Committee (ICREC) protocols approved by Imperial College London. 250 µL aliquots of filtered, centrifuged urine were combined with 150 µL of HPLC grade water (Sigma) and 100 µL of D₂O NMR buffer containing 1 mM TSP (Sigma-Aldrich) and 3 mM NaN₃ in 5 mm NMR tubes (Wilmad). Aliquots were taken from urine samples of (N=10) cystinosis patients and prepared as above. 1D ¹H spectra were collected on a Bruker Avance III spectrometer operating at 800.32 MHz using a 5 mm cryoprobe. Spectra were acquired at 300K with a resolution of 64k data points and a spectral width of 20 ppm using a 1D pulse-acquire sequence with direct presaturation of residual water and a relaxation delay of 30s to accommodate full T₁ relaxation of urinary metabolites. All data were Fourier transformed, phased, and baseline corrected in XWinNMR (Bruker) before being transferred to Matlab R2010b (Mathworks) for preprocessing and data analysis.

5.2.2.2 Urinary metabolic natural history of cystinosis

For all 200 cystinosis urine samples, 500 µL aliquots of filtered, centrifuged urine were combined with 100 µL of NMR buffer containing D₂O and 1 mM TSP (3-trimethylsilyl- 2,2,3,3-tetradeuteriopropionate, Sigma-Aldrich) and 3 mM NaN₃ in 5 mm NMR tubes (Wilmad). 1D ¹H NMR spectra were collected for all samples on a Bruker Avance spectrometer operating at 600.44 MHz using a 5 mm BBI probe at 300K. Samples were run using an automated sample changer with a randomized run order. Spectra were acquired at a resolution of 64k data points and a spectral width of 20 ppm using a 1D pulse acquire sequence with gradient pulses and excitation sculpting for presaturation of water. For select samples, 1D ¹H spectra were collected using a pulse-acquire sequence with irradiation of residual water and TOCSY spectra were collected using a 90 ms mixing time DIPSI-2 sequence with gradient pulses and excitation sculpting on a Bruker Avance spectrometer operating at 800.32 MHz. All data were Fourier transformed, phased, and baseline corrected in XWinNMR (Bruker)

before being transferred to Matlab R2010b (Mathworks) for preprocessing and pattern recognition. One spectrum was left out of statistical evaluation due to spectral quality, leaving (N=199) for further analysis.

5.2.2.3 Animal studies of nephrotoxin exposure

1D ^1H NMR metabonomic datasets were collected from the urine samples of Sprague-Dawley rats exposed to the model nephrotoxins cisplatin, mercuric chloride (HgCl_2), and 2-bromoethanamine (BEA) studied as part of the Consortium for Metabonomic Toxicology, COMET [249, 250]. Animal selection and treatment followed the standard COMET protocol, as detailed elsewhere [250]. At the 0 hour time point, animals in each group were administered an intraperitoneal injection of vehicle (0.9% saline) with toxin dosages selected to induce overt toxicity without significant mortality (HD, 10 ml/kg cisplatin, 2 ml/kg HgCl_2 , 5 ml/kg BEA, N=10 each) or saline vehicle alone (C, N=10 for each study). In all experimental groups half of the animals were euthanized at 48 h and the remaining animals at 168 h after dosing. Urine samples were collected at -16, 0, 8, 24, 48, 72, 96, 120, 144, and 168 h postdosing and 1D ^1H NMR spectra were acquired and processed for each collected sample using standard COMET procedures [63, 251].

5.2.3 Data analysis

NMR spectra were referenced to TSP at 0.00 ppm before the removal of residual water signal from 4.5 – 5.2 ppm. Assignment of observed NMR peaks was achieved by reference to previously assigned urinary NMR spectra[110], to spectra of pure compounds and 2D ^1H - ^1H TOCSY spectra were acquired for selected samples for further confirmation of urinary metabolites. Figure 5.1 shows an aliphatic region of a 600 MHz NMR spectrum typical for a pre-transplant cystinosis urine sample. Tartrate was confirmed by LC-MS and spiking of an authentic standard, as it is not frequently observed at high levels in urinary NMR spectra (Figure 5.2).

Table 5.2. Assignments of urinary NMR signals integrated for targeted analysis. Abbreviations: s, singlet; d, doublet; dd, doublet of doublets; t, triplet; q, quartet; m, multiplet; AB, AB doublet.

Metabolite	Integrated signal	Other signals
Glucose	3.90 (dd)	3.26 (dd), 3.43 (m), 3.49 (m), 3.55 (dd), 3.73 (m), 3.85 (m), 4.66 (d), 5.24 (d)
Lactate	1.33 (d)	4.12 (q)
Glutamine	2.45 (m)	2.14 (m), 3.78 (t)
Lysine	1.73 (m)	1.48 (m), 1.93 (m), 3.05 (t), 3.78 (t)
Leucine	0.96 (t)	1.73 (m), 3.74 (m)
Isoleucine	1.02 (d)	0.94 (t), 1.26 (m), 1.47 (m), 1.99 (m)
Valine	0.99 (d)	1.04 (d), 2.30 (m), 3.63 (d)
Threonine	4.26 (m)	1.33 (d), 3.59 (d)
Alanine	1.48 (d)	3.78 (q)
Tyrosine	6.91 (d)	3.06 (dd), 3.2 (dd), 3.94 (dd), 7.2 (d)
TMAO	3.27 (s)	
Acetate	1.92 (s)	
Creatine	3.93 (s)	3.04 (s)
Creatinine	4.06 (s)	3.05 (s)
Citrate	2.54 (AB)	2.66 (AB)
Glycine	3.57 (s)	
Tartrate	4.34 (s)	
Formate	8.46 (s)	
Carnitine	3.23 (s)	2.44 (dd), 3.43 (m), 4.59 (m)
Dimethylamine	2.73 (s)	
Methylamine	2.61 (s)	
Dimethylglycine	2.93 (s)	
Hippurate	7.56 (t)	3.97 (d), 7.64 (t), 7.84 (d)
Acetone	2.33 (s)	
Succinate	2.40 (s)	
Fumarate	6.52 (s)	
Prolinebetaine	3.11 (s)	3.31 (s)
Hypotaurine	3.36 (t)	2.66 (t)
Myo-inositol	4.04 (t)	3.63 (dd), 3.53 (dd), 3.28 (t)

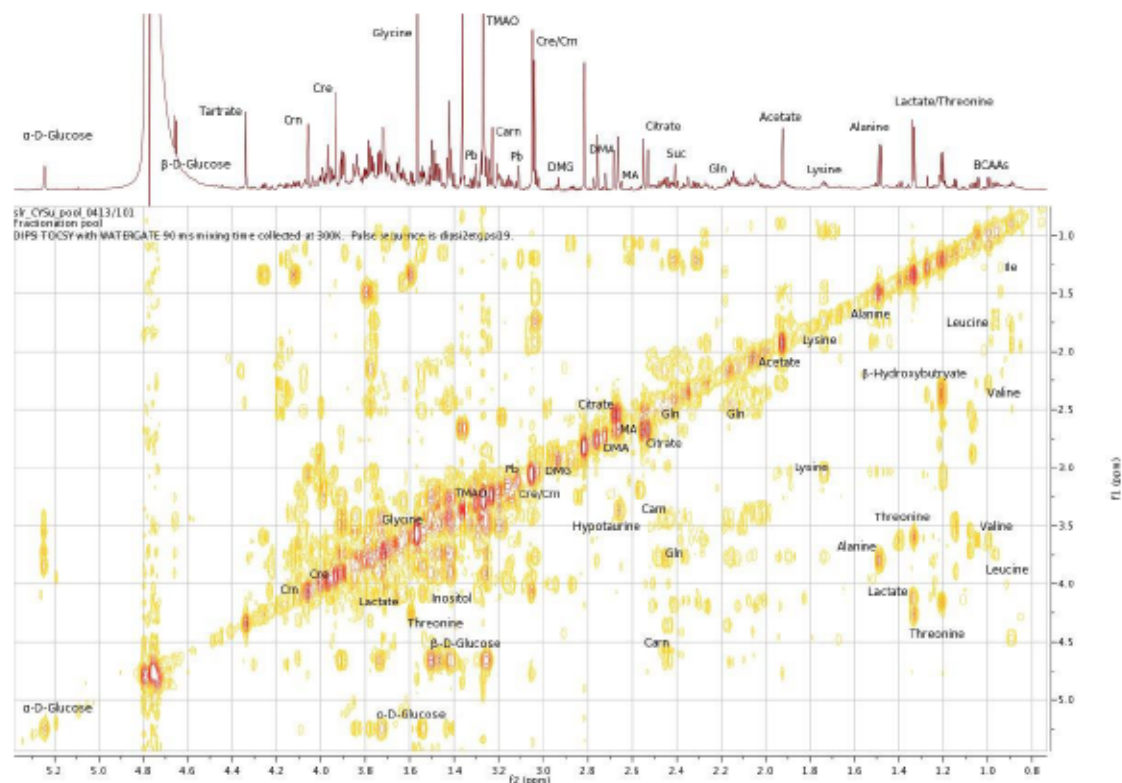
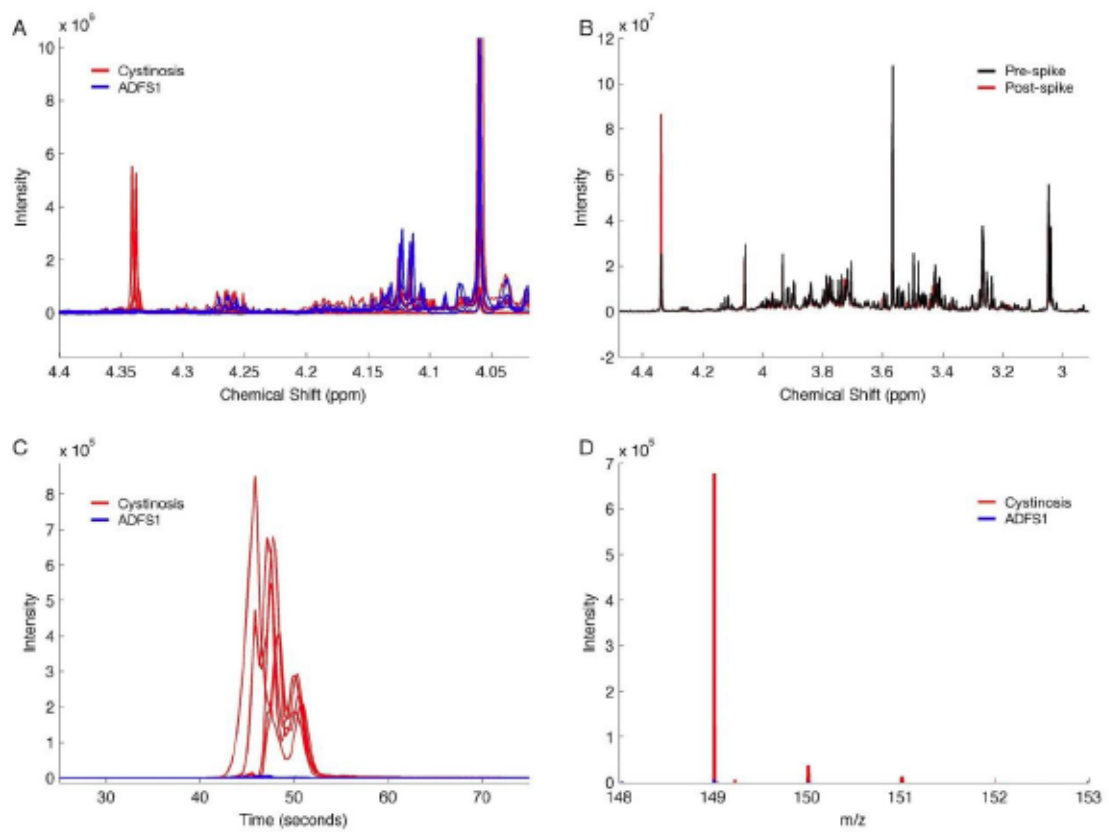


Figure 5.1. Metabolic profiling of Fanconi Syndrome in cystinosis urine samples by NMR spectroscopy. Assignments in both 1D ^1H (top) and 2D ^1H - ^1H TOCSY spectra show urinary loss of sugars and amino acids characteristic of renal Fanconi Syndrome, as well as increased excretion of lactate, acetate, and trimethylamine-N-oxide. 1D ^1H spectrum shows resonance frequencies of individual protons, while 2D TOCSY shows off-diagonal crosspeaks between coupled protons in the same metabolite. Abbreviations: BCAA, branch-chain amino acids; Gln, glutamine; Suc, succinate; MA, methylamine; DMA, dimethylamine; DMG, dimethylglycine; Pb, proline-betaine; Carn, carnitine; TMAO, trimethylamine-N-oxide; Cre, creatine; Crn, creatinine.



The data presented in this figure are derived from a series of experiments designed to investigate the metabolic profile of Cystinosis and ADFS1. Panel A displays the ^1H NMR spectra, highlighting the characteristic peaks for both conditions. Panel B shows the 2D NMR spectra, which provide insight into the interactions between different metabolites. Panel C illustrates the time-resolved NMR spectra, capturing the dynamic changes in the system over time. Finally, Panel D presents the mass spectrum, identifying the molecular ions and their corresponding m/z values.

Spectral signals representing 29 metabolites with well-resolved peaks were quantified by fitting Lorentzian lineshapes to the NMR multiplets indicated in Table 5.2 according to equation 5.1.

$$i_{\delta} = s \cdot \frac{w^2}{w^2 + (\delta_0 - \delta)^2} \quad (5.1)$$

Here, i_{δ} is the intensity of the Lorentzian lineshape at chemical shift position δ given peak half-width at half-max w and peak positions δ_0 , multiplied by a scaling factor s as the Lorentzian function gives values between 0 and 1. Peak positions δ_0 were detected by peak picking the spectral lineshapes and values for s and w were optimized by grid search between the ranges of $(0 < s < 1 \quad i_{\delta 0})$ and $(0.25 \text{ Hz} < w < 3 \text{ Hz})$ using the residual of the NMR lineshape after subtraction of i_{δ} as an objective function. Lorentzian fitting generates a matrix of N samples by M metabolites, which was normalized to the CRN singlet at 4.06 ppm for the spectra in 5.2.2.1 and 5.2.2.3 in order to compare across phenotypic groups and by probabilistic quotient normalization [88] for the spectra in 5.2.2.2 as CRN excretion is affected by the independent variables of age, gender, GFR, and transplantation status which vary in the natural history study. The normalized abundances were $\log_{10}$ scaled to meet the normality criterion of the Lilliefors test [252].

The data matrices of 5.2.2.1 and 5.2.2.3 were subjected to unpaired, two-sided t-tests between disease and toxicological groups and the resulting p-values were corrected using the Benjamini-Hochberg False Discovery Rate [114, 115] for multiple testing. Missing values (<10% of independent variables) were estimated using regression based multiple imputation [253]. PCA using the NIPALS algorithm was applied to the data matrix of 5.2.2.2 following mean centering and scaling to unit variance and partial correlations between metabolites were calculated independent of age, gender, GFR, Cystagon dose and transplantation status. One-way analysis of variance (ANOVA) was used to compare the variance between and within individual patients for both independent variables and metabolite levels. Associations between single independent variables and metabolite levels were identified by linear or logistic regression after correction for confounders by multivariable adjustment for other independent variables [254]. Associations were considered significant if regression p-values were less than a threshold calculated by Bonferroni correction. For time

courses in data sets 5.2.2.2 and 5.2.2.3 mean values and standard error of the means were calculated at each time point to show changes in metabolite levels over time. While animal samples were collected at regular intervals, human samples were collected at irregular intervals and to compare these measurements by time-to-transplant values were taken for the nearest year pre- or post-transplant.

5.3 Results and Discussion

5.3.1 Comparison of congenital kidney diseases

Renal Fanconi Syndrome is a consensus phenotype resulting from the loss of transporter activity in the renal proximal tubule that is caused by inherited conditions as well as chemical insult [255, 256]. This renders the renal tubule unable to recover small molecule solutes from the ultrafiltrate. Hence, FS is characterized by marked glucosuria and generalized aminoaciduria, as well as the loss of ions such as phosphate, calcium, sodium, potassium, and bicarbonate [23]. Renal Fanconi Syndrome can be isolated, as in the case of idiopathic FS, or it can be one aspect of a systemic condition as in the case of cystinosis.

In order to identify urinary perturbations resulting from FS and distinguish them from those resulting from systemic or non-tubule renal pathology, urinary profiles of cystinosis and three forms of congenital renal pathology with recently characterized genetic causes were compared. The presentation of these syndromes partially overlaps with that of cystinosis, allowing metabolites that discriminate between these similar phenotypes to be identified.

Figure 5.3 shows the clinical phenotypes associated with each of the diseases analyzed in this section. While autosomal dominant Fanconi Syndrome type I (ADFS1), autosomal dominant Fanconi Syndrome type II (ADFS2), and cystinosis all present with Fanconi Syndrome, there is no apparent renal failure in ADFS1 patients. Additionally, neither ADFS1 nor ADFS2 patients have the muscle weakness and progressive deterioration observed in cystinosis. Finally, comparison to controls and patients with a novel form of nephrocalcinosis (ERS) should provide controls for comparison against a normal urinary baseline and against a significantly differing renal phenotype.

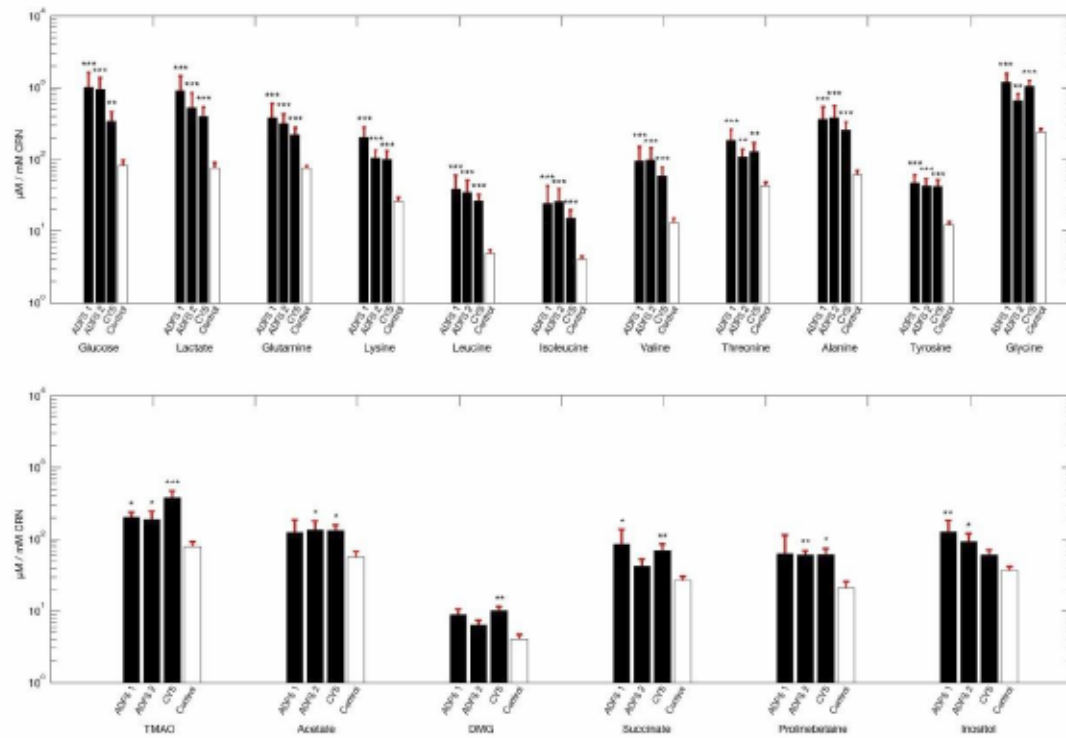


Figure 5.5 shows urinary abundances of metabolites relative to creatinine for each of the kidney diseases that present with Fanconi Syndrome and for unaffected family members for the metabolites that show significant differences not just between affected patients and controls, but between patients with individual diseases. These metabolites that discriminate between patient groups should indicate unique features of the pathology of these diseases.

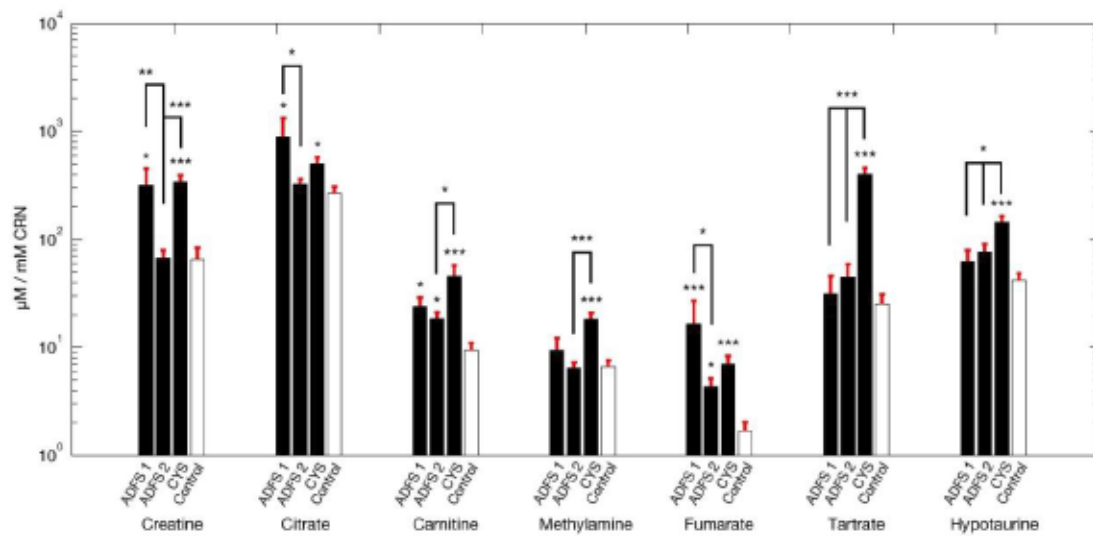
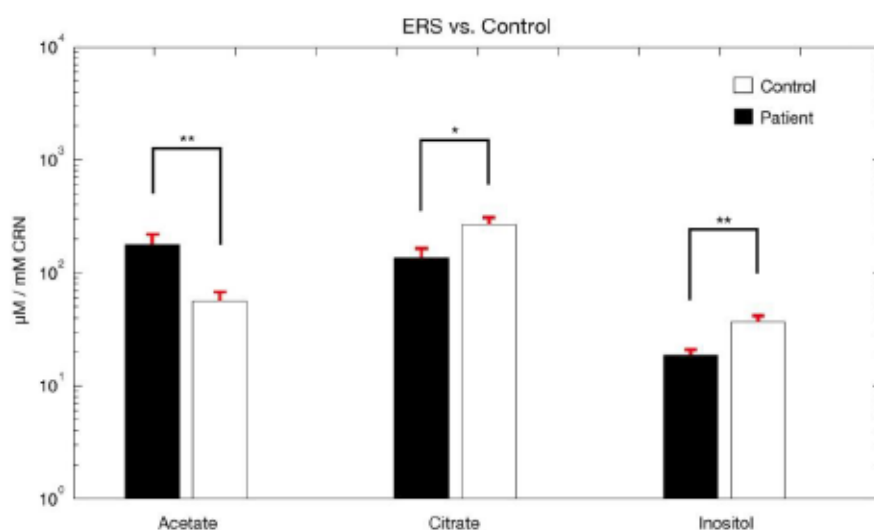


Figure 4: Urinary excretion of TCA cycle intermediates in ADFS1, ADFS2, CYS, and Control groups. The chart shows that ADFS1 and CYS patients have significantly elevated levels of citrate, fumarate, and tartrate compared to ADFS2 and controls. Creatine excretion is also elevated in ADFS1 and CYS patients. Statistical significance is indicated by asterisks (*, **, ***, ****) and brackets.

Citrate and fumarate are significantly increased in ADFS1 individuals relative to ADFS2 individuals (citrate, $p = 0.036$; fumarate, $p = 0.035$) as well as to controls (citrate, $p = 0.0066$; fumarate, $p = 1.91 \times 10^{-4}$). Elevated excretion of these TCA cycle intermediates, along with succinate (Figure 4) are likely indicative of the TCA cycle disruption caused by the underlying genetic defect in ADFS1 patients as increased urinary excretion of TCA cycle intermediates has frequently been observed to accompany mitochondrial pathology [161-163]. Retrospective analyses have suggested fumarate is a sensitive and specific indicators of mitochondrial disease [163], and coordinated loss of multiple TCA cycle intermediates in ADFS1 patients would support this assertion. Cystinosis patients also show elevations in citrate, fumarate, and succinate levels, and while citrate is also supplemented to alkalinise pre-transplant cystinosis patients, elevated fumarate and succinate suggest the energetics effects postulated as a cause renal Fanconi Syndrome in cystinosis patients [260-262] and sync with observations of altered mitochondrial appearance in proximal tubule cells of cystinosis patients [263].

The observation of elevated creatine excretion observed in ADFS1 and cystinosis patients relative to controls (ADFS1, $p = 0.015$; cystinosis, $p = 1.44 \times 10^{-4}$)

and ADFS2 individuals (ADFS1, $p = 0.0029$; cystinosis, $p = 4.70 \times 10^{-6}$) may also be a result of an energetic defect in these diseases. Cystinosis patients present with progressive myopathy [264, 265], which is absent in ADFS2 patients, and elevated urinary creatine/CRN ratios are an established biomarker of steroid induced myopathy [266]. While the muscle weakness and wasting characteristic of myopathy was not identified clinically in ADFS1 patients, there is no reason to suspect mistrafficking of LBP is limited to the kidney, and as a result pathology of tissues with significant energetic requirements, including both muscle and the renal tubule, is consistent.



ERS vs. Control

Elevated carnitine, hypotaurine, and tartrate and excretion by cystinosis patients are most likely the result of therapeutic interventions rather than a primary metabolic perturbation. L-Carnitine is frequently supplemented in cystinosis patients, and while absence of cysteamine in the urine of cystinosis patients confirms its pre-renal catabolism [267], hypotaurine has been identified as an excreted metabolite of cysteamine in animal models [268]. Tartrate is a minor metabolite known to be excreted following fruit consumption [269], however the extreme elevations in tartrate excretion by cystinosis patients relative to controls and to ADFS1 and ADFS2 patients indicated excretion of unmetabolised tartrate resulting from cysteamine

therapy under the cysteamine-bitartrate (Cystagon) formulation. While elevated methylamine excretion by cystinosis patients could be related to L-carnitine supplementation, studies of the origin of urinary methylamine have indicated a primarily endogenous origin with only minor contributions from dietary sources [270].

ERS urinary profiles, unlike those of the Fanconi Syndromes, show few differences relative to controls. The metabolites that are significantly modulated, citrate, acetate and myo-inositol, were also observed to be perturbed in the Fanconi Syndromes. However, while citrate excretion by ADFS1, ADFS2 and cystinosis patients was elevated relative to controls, citrate excretion by ERS patients is reduced across the cohort- especially so in six individuals for which the citrate signal is below the limit of detection. Given the complex role citrate plays in the kidney, it is unsurprising that its excretion is modulated by very different phenotypes. In ERS patients, reduced citrate excretion likely indicates the disruption of calcium homeostasis underlying the nephrocalcinosis in these patients rather than a TCA cycle defect. Hypocitraturia has been noted in instances of renal calcium precipitation, especially nephrolithiasis [271], and citrate supplementation is an established treatment for stone forming disorders, including hypercalciuric nephrolithiasis [272]. Hypocitraturia is also considered a major risk factor for nephrocalcinosis in neonates [273].

Figure 5.7 shows a summary of the urinary perturbations identified for the diseases investigated here. These findings both support and clarify work in the study of congenital kidney diseases, renal transplantation, and nephrotoxicity, and add important context to the following investigations of renal metabolism in cystinosis and pathology of structures of the kidney due to toxin exposure.

levels of excreted trimethylamine-N-oxide (TMAO) as well as the effects of oral cysteamine therapy and kidney transplantation.

The natural history of nephropathic cystinosis is characterized by a number of factors, changes in which are likely to affect urinary metabolite levels. These include factors which are specific features of cystinosis, including renal failure, drug therapy with cysteamine bitartrate, and transplantation, and general factors that are known to affect urinary metabolic profiles including age, gender and differences between individuals [277, 278]. In order to determine the relationships between the following variables- transplantation status, age, prescribed Cystagon dosage, gender, and glomerular filtration rate (GFR) and their relationships to urinary metabolite levels, correlations between these variables, fit to urinary metabolite levels, and differences between individuals were assessed (Figure 5.8).

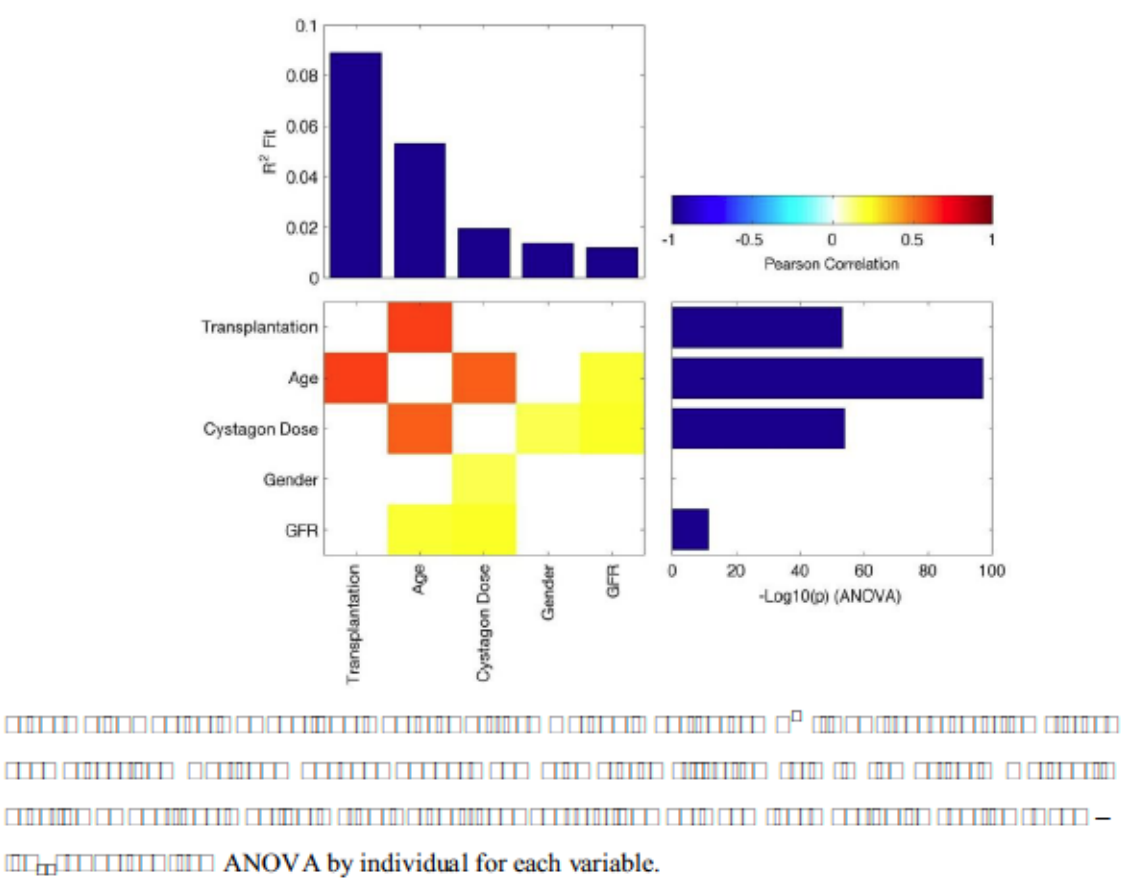
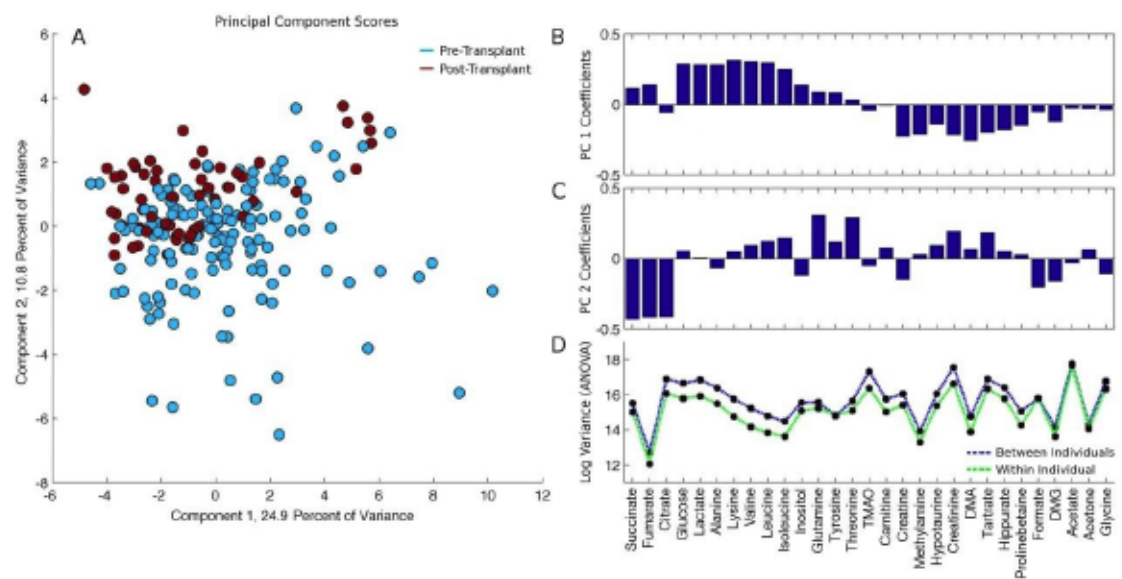


Figure 5.8 shows the R^2 fit of linear regression modeling of each variable to the full set of urinary metabolite levels, a heatmap showing significant correlations between variables, and a Manhattan representation of ANOVA p-values comparing variance between patients with variance within individual patients. These results

show that transplantation has the greatest effect on the urinary metabolic profiles of cystinosis patients, and while age has a substantial effect as well it is significantly correlated with all other variables except gender. One-way analysis of variance (ANOVA) by individual shows the effects of capturing metabolic profiles from individuals at varying points in the natural history of cystinosis. Despite an average of 9 samples per individual patient over the 15-year period in which samples were collected, these timepoints do not capture the entire progression of cystinosis for any individual patient and as a result ANOVA reports significant differences between individuals especially for age.

Statistical analysis of urinary metabolites over the set of pre- and post-transplant NMR spectra indicate while transplantation may be the dominant source of variation among the set of variables captured in Figure 5.8, Fanconi Syndrome accounts for a greater share of the variation in the cystinosis urinary profiles (Figure 5.9). Principal Component Analysis decomposes the data matrix of 199 samples and 29 metabolites into N scores (Figure 5.9A) and P loadings (Figure 5.9B, 5.9C) for a number of Principal Components (PCs) corresponding to major sources of variation in the overall dataset. Principal Component 1, representing 24.9% of the total variance in these metabolites, is dominated by the urinary abundances of glucose and amino acids characteristic of Fanconi Syndrome (Figure 5.9B).

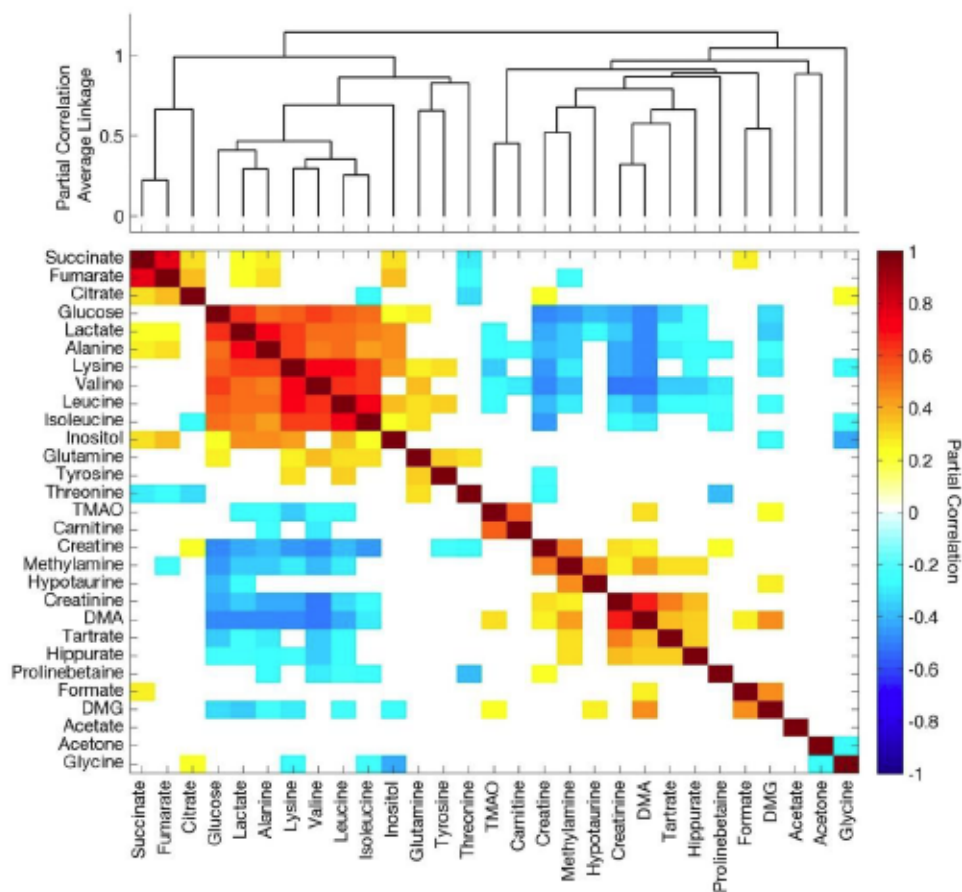
Principal Component 2 (10.8% of variance) shows some separation between pre-transplant (blue) and post-transplant (red) (Figure 5.9A), but the PC 2 loadings (Figure 5.9C) indicate a dominant contribution from TCA cycle intermediates succinate, fumarate, and citrate. This could indicate that after the development of Fanconi Syndrome, energetic stress as observed in section 5.3.1 is the major source of urinary metabolic variance in the natural history of cystinosis. Energetic stress to the native renal tubule would experience would result in correlations with transplantation, but may not mirror the direct effects of transplantation.



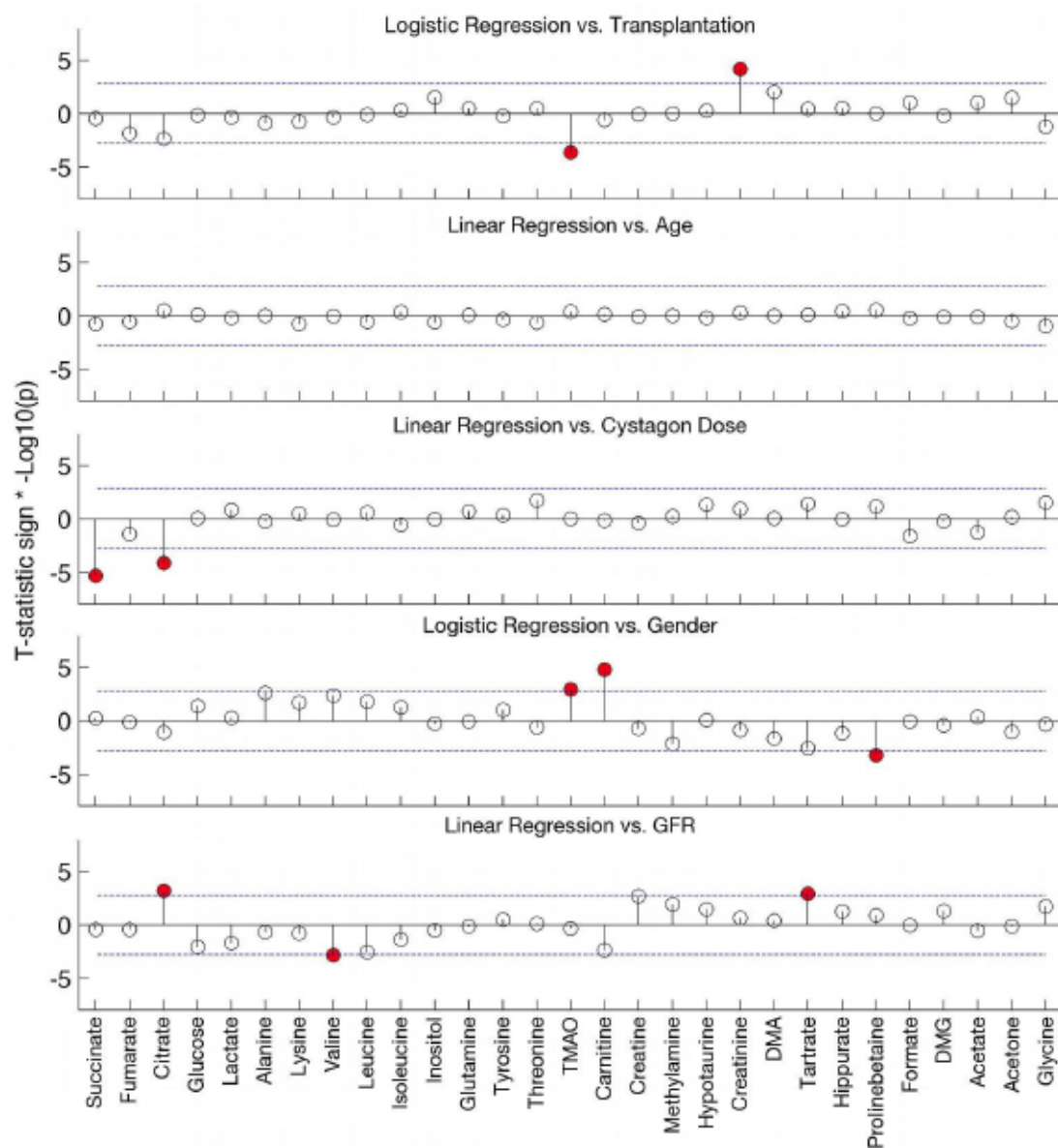
To identify whether Fanconi Syndrome shows greater variability from patient to patient or across the set of observations of a single patient, one-way analysis of variance (ANOVA) testing was performed using patient IDs as sub-groups after correction for confounding variables of transplantation status, age, prescribed Cystagon dosage, gender, and GFR which were shown to be significantly different between individuals (Figure 5.8). ANOVA results indicate that for the majority of characteristic Fanconi Syndrome metabolites (including glucose, lactate, alanine, lysine, valine, leucine, and isoleucine), variance between individual patients exceeds the variance within patients to a greater extent than most urinary metabolites (Figure 5.9D).

Multiple forms of membrane proteins in the solute carrier (SLC) family of transporters at the luminal and basolateral faces of the renal proximal tubule are responsible for the recovery of solutes, including separate transporters for positively, negatively, and neutrally charged amino acids [279]. While some congenital kidney diseases including Hartnup disease, iminoglycinuria, and cystinuria are caused by mutations in single solute carrier genes, renal Fanconi Syndrome is characterized by a generalized aminoaciduria that does not appear to be transporter-specific. Across the board reduction in transporter activity would be expected to generate high correlations between the excreted levels of actively transported solutes.

Clustered partial correlations between excreted metabolites independent of transplantation status, age, prescribed Cystagon dosage, gender, and GFR (Figure 5.10) show significant, positive correlations between the actively transported solutes glucose, lactate, myo-inositol, and uncharged, polar, branch-chain, and negatively charged amino acids. Lower baseline excretion of these amino acids and spectral crowding may be responsible for the reduced correlations of tyrosine, threonine, and glutamine, though significant positive correlations between these amino acids and other transporter substrates were still observed. The observation of significant positive correlations between the urinary loss of these metabolites supports observations of reduced intercellular ATP concentration in cystine-loaded cells and resulting coordinated loss of transporter activity [262]. The clustering of succinate, fumarate, and citrate, which are frequently observed to be excreted in the urine in mitochondrial disorders [280], suggest disrupted mitochondrial ATP production in renal tubular cells could result in the observed coordinated excretion of actively transported solutes.



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28



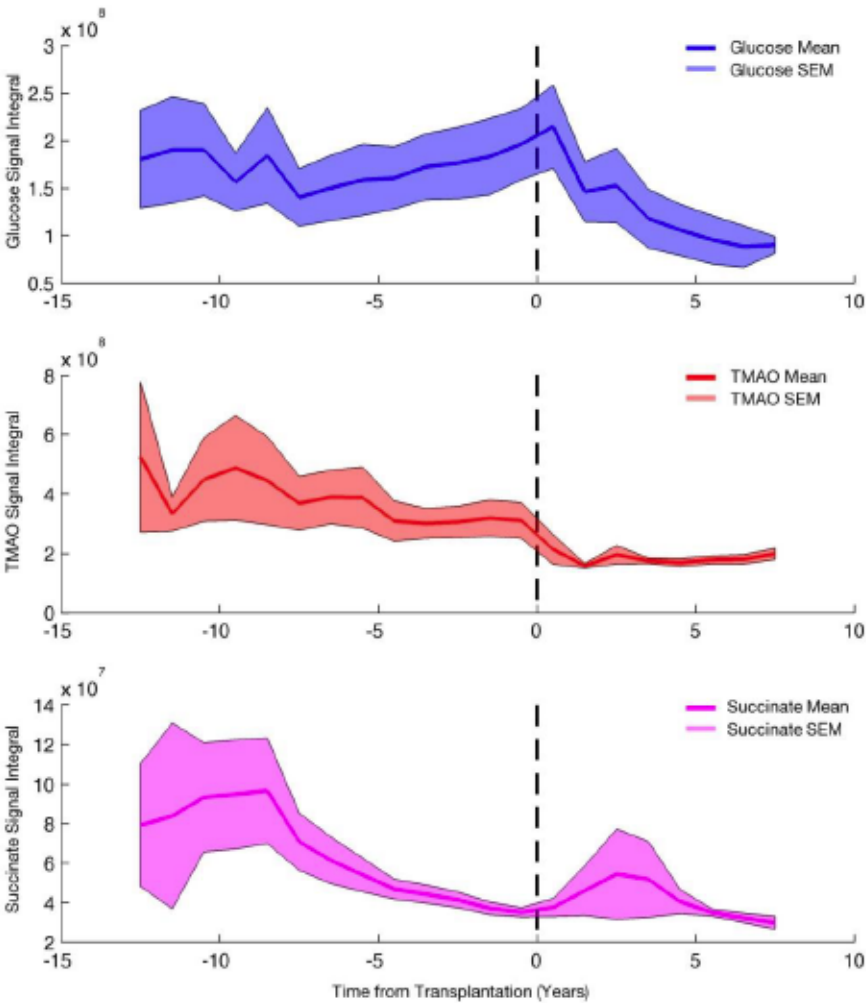
0.0017, Bonferroni correction) are indicated in red, and the ± 1.96

As indicated by Figure 5.8, the variables of transplantation status, age, prescribed Cystagon dosage, gender, and glomerular filtration rate (GFR) affect the urinary metabolic profile of cystinosis patients. In order to identify associations between these independent variables and metabolite levels, regression analysis was applied to model metabolite excretion as a function of each of these variables, correcting for all other variables as confounders. Associations were considered significant if regression p-values were less than a threshold calculated by Bonferroni correction ($p = 0.0017$). Figure 5.11 shows the results of linear regression vs. age, Cystagon dose, and GFR and logistic regression vs. dummy variables for transplantation status (pre-transplant, 0; post-transplant, 1) and gender (female, 0; male, 1). Here, the Manhattan representation of p-values for each metabolite from regression analysis and the Bonferroni threshold indicate significant associations between metabolites and transplantation status, age, Cystagon dosage, gender, and GFR.

Increases in creatinine excretion following transplantation match expectations of increased creatinine clearance, and reductions in TMAO levels relative to pre-transplant fit with observations of accumulating TMAO produced in end stage renal disease as a counter-osmolyte by the renal medulla [281] and changes in TMAO excretion during transplantation follow-up [217, 282]. Interestingly, TMAO and carnitine excretion was associated with gender as higher levels were observed in males. It is known that L-carnitine is metabolized to TMAO by the gut microflora [283-285], which could explain the association between carnitine and TMAO. The effect of L-carnitine supplementation on renal physiology and the cause of increased L-carnitine levels in males, however, is an area for further inquiry.

Hypotaurine and tartrate, excreted metabolites of cysteamine-bitartrate (Cystagon), would be expected to associate with prescribed Cystagon dosage. While there is a positive correlation between hypotaurine and tartrate excretion and Cystagon dosage, this association is not significant. Interestingly, it is the TCA cycle intermediates succinate and citrate that are significantly associated with Cystagon dosage. While citrate is supplemented to some cystinosis patients, lower succinate excretion with increased Cystagon dose suggests a reduction in mitochondrial stress in the renal tubule with cysteamine therapy. Citrate and tartrate are both significantly correlated with GFR, and cysteamine therapy has been shown on several occasions to promote maintenance of glomerular function and postpone transplantation [237,

240,242]. Interestingly, no single metabolites were significantly associated with age after correction for the other variables, despite the relatively high fit of age to urinary metabolic profiles overall (Figure 5.8). This, together with the significant correlations of age with transplantation status, Cystagon dose, and GFR suggest that the effects of age on the cystinosis urinary metabolic profile are due to changes in other factors correlated with age rather than growth and development alone.

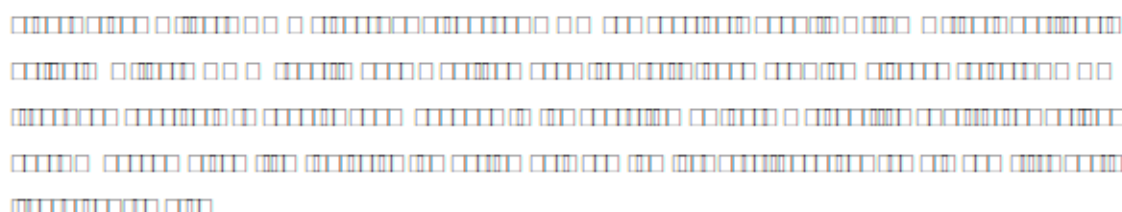


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

In order to investigate the changes in key metabolic parameters, mean glucose, TMAO and succinate levels were compared by time before and after transplantation

for the (N=9) cystinosis patients for whom both pre- and post-transplantation samples were available. As urine samples were collected during routine visits that were not necessarily scheduled at fixed intervals, time from transplantation was rounded to the nearest year to compare between individuals. Figure 5.12 shows mean (dark lines) and standard error of the mean (SEM, lighter regions) for glucose, TMAO, and succinate levels over the range of times from transplantation captured over the natural history study. Glucose levels show gradual increase from 8 years before transplantation and a slow decline following transplantation, and are characterized by high variation between individuals at all time points both before and after transplantation. TMAO excretion is high pre-transplant, but drops immediately and shows low variance following transplantation. Succinate excretion is high long before transplantation but drops in the years approaching renal failure and transplantation. Interestingly, succinate excretion climbs following transplantation before declining to the low levels seen immediately pre-transplant. Figure 5.13 shows these trends represented in the urinary NMR signals of these metabolites from a single patient preceding and following transplantation. Again, the notable characteristics include a gradual decline in glucose excretion, a rapid drop in TMAO, and pre-transplant declines in succinate excretion following by a post-transplant spike.

These results suggest a model for renal tubular damage in nephropathic cystinosis characterized by early mitochondrial stress leading to elevations in TCA cycle intermediates. Associations between TCA metabolite excretion and Cystagon dosage suggest that cysteamine therapy may retard or mitigate this process. As mitochondrial pathology leads to energetic stress in the renal tubule, the activity of transporters in the proximal renal tubule declines, resulting in increases in solute excretion including highly correlated losses of glucose, lactate, and amino acids. The decreases in succinate excretion observed approaching transplantation could suggest the majority of mitochondrial stress in the renal tubule occurs early, leaving lower levels of TCA cycle metabolites remaining to be lost by the tubule. TMAO levels, indicating counter-osmotic action by the kidney and potential metabolism from L-carnitine, drop following transplantation, but native kidneys not surgically removed continue to lose lower levels of transporter substrates into the urine. An interesting line of further investigation would be whether the post-transplant spike in succinate excretion indicates proximal tubular damage to the transplanted kidneys or residual excretion from the native kidneys.



Cisplatin is an alkylating antineoplastic agent [288] that is known to cause acute renal failure when exposed in high doses and has been shown to inhibit fatty acid oxidation and the pyruvate dehydrogenase complex [289]. Mercuric chloride exposure has been shown to increase reactive oxygen species [290, 291] and result in oxidative stress and nephrotoxicity [292]. Both cisplatin and mercuric chloride induce acute tubular necrosis and lesions in the proximal tubule [234] and have been shown to produce Fanconi-like urinary excretion of glucose and amino acids in animals [161, 233, 293]. Interestingly, while heavy-metal toxicity does cause renal tubular damage in humans, the major effect of cisplatin toxicity in humans is glomerular failure rather than tubular damage [294]. 2-bromoethanamine (BEA) induces necrosis of the renal papilla and has been shown to affect the epithelial cells of the collecting duct and papilla [295]. Figure 5.14 shows the urinary metabolic changes in animals exposed to these model nephrotoxins over the full time-course of sample collection in the COMET protocol (see 5.2.2.3).

These results show that both cisplatin and mercuric chloride induce significant urinary loss of glucose, lactate, and amino acids similar to that observed in the congenital Fanconi Syndromes. Interestingly, many of these solutes are also excreted by BEA exposed animals at a level significantly higher than controls, though less than cisplatin and HgCl₂ exposed animals. BEA exposed animals do not show significant differences from controls in glutamine and glycine excretion, and glycine excretion by cystinosis patients showed much lower correlations to glucose, lactate, and the other amino acids characteristic of FS. Similar patterns are observed for acetate and creatine excretion, with significant elevations especially by HgCl₂ exposed animals. Interestingly, while TMAO levels were elevated in cisplatin and HgCl₂ exposed animals, significant changes in TMAO levels were not observed in BEA exposed animals. Of the metabolites which showed significantly altered urinary excretion following toxin exposure, BEA exposed animals showed increased excretion of only methylamine and acetone relative to cisplatin and HgCl₂. Methylamine was observed to be significantly elevated in cystinosis patients relative to both ADFS1 and ADFS2 patients as well as to controls.

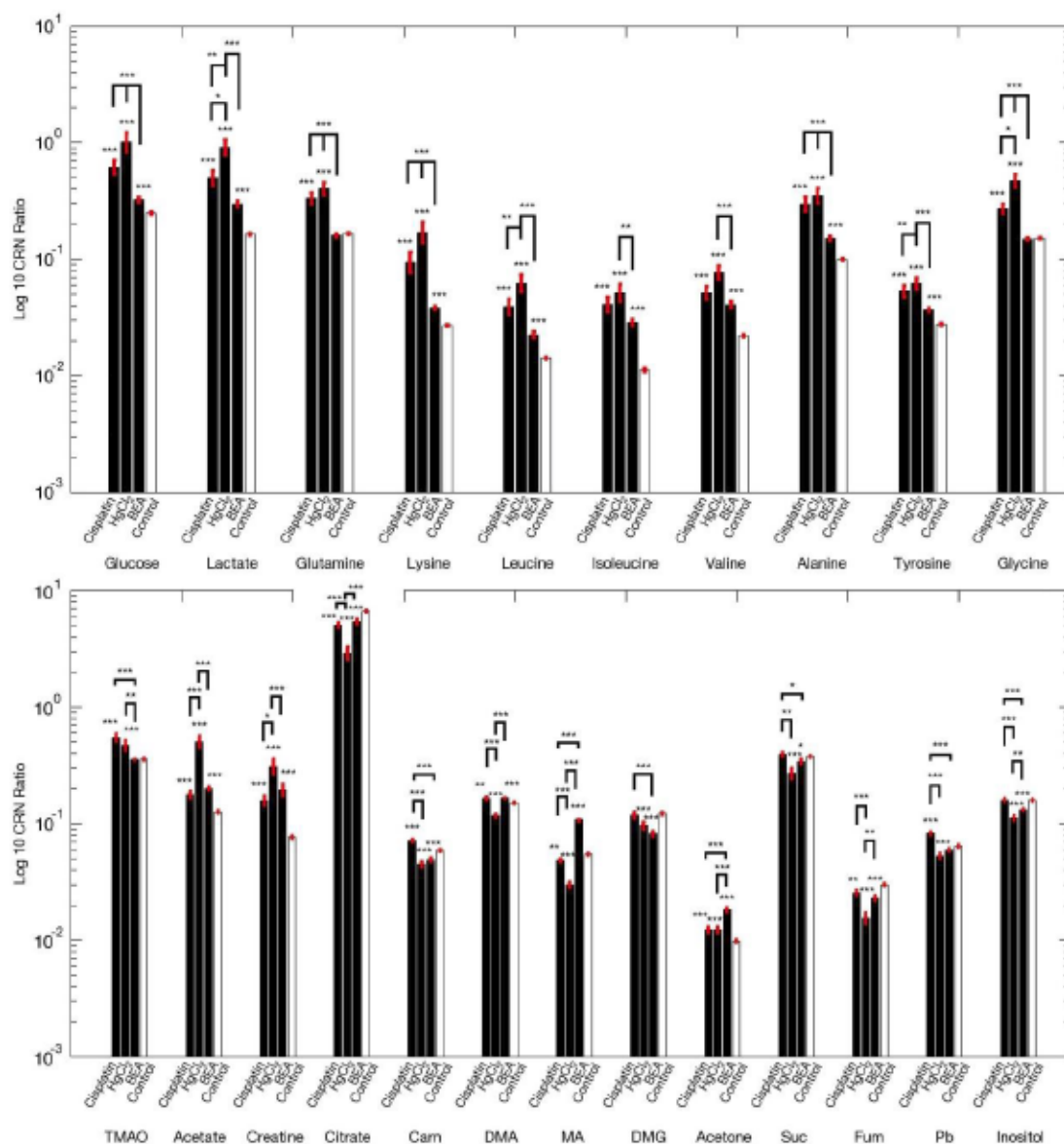


Figure 5.14. Bar charts showing the Log₁₀ CRN Ratio for various metabolites across four groups: Caspaltin, H₂O₂, BSA, and Control. The top chart displays Glucose, Lactate, Glutamine, Lysine, Leucine, Isoleucine, Valine, Alanine, Tyrosine, and Glycine. The bottom chart displays TMAO, Acetate, Creatine, Citrate, Carn, DMA, MA, DMG, Acetone, Suc, Fum, Pb, and Inositol. Each bar represents the mean Log₁₀ CRN Ratio, with a red dot indicating the standard deviation. Significance markers (***, **, *) indicate statistical significance relative to the control group.

The TCA cycle intermediates citrate, succinate and fumarate showed significantly greater reductions in excretion levels in HgCl₂ relative to cisplatin exposed animals. While overall reductions in both citrate and fumarate levels for cisplatin and HgCl₂ exposed animals does not fit with observations of increased excretion of TCA cycle intermediates at peak toxicity, analysis of the time-course excretion of glucose, TMAO, and succinate (Figure 5.15) indicates that both cisplatin and HgCl₂ exposed animals show increases in succinate excretion followed by declines, similar to the increase and decline seen in pre-transplant cystinosis patients. The timing of these changes, however, differs between cisplatin and HgCl₂ exposed animals. Cisplatin toxicity appears to result in an early increase in succinate excretion between 0-48 hours post-dose followed by a reduction that coincides with the maximal excretion of glucose and TMAO between 90-120 hours postdose, while peak succinate excretion resulting from HgCl₂ exposure coincides with peak glucose and TMAO excretion at 24 hours.

The observations of Fanconi-like excretion of glucose, lactate, and amino-acids as well as increased excretion of TMAO and increases and subsequent declines in the excretion of TCA cycle intermediates resulting from acute renal failure in cisplatin exposed animals match the urinary metabolic profile of the congenital Fanconi Syndromes and the changes in the cystinosis urinary profile over the natural history of the disease. Cisplatin exposure has been demonstrated to result in inhibition of medium chain acyl-CoA dehydrogenase and pyruvate dehydrogenase before the onset of acute tubular necrosis [289], which is ameliorated by treatment with L-carnitine and **PPAR α** ligands Wy-14643 and fibrate [289, 293]. The similarity of these metabolic profiles and previous evidence of energetic stress in models of cystinosis [260-262] suggest this may play a role in the development of FS in cystinosis patients.

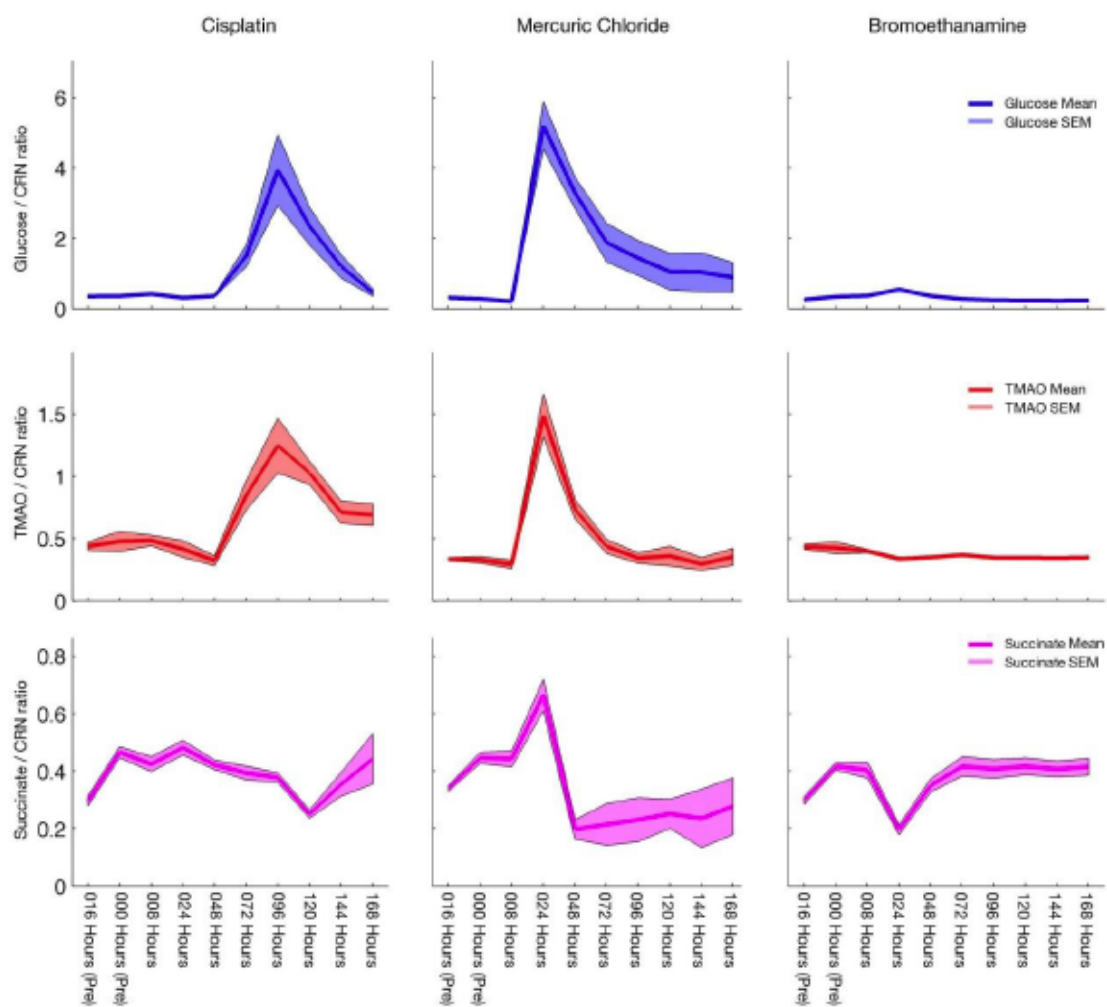


Figure 5.15. [Graph Description]

5.4 Conclusions

Urinary metabolic profiling has become an increasingly important tool in nephrology and metabolomics studies have made significant contributions, especially to the toxicological [161, 221-224] and transplantation literature [217, 231, 232]. **Metabolic profiling of Dent's disease and the Oculocerebrorenal Syndrome of Lowe** [23] and Autosomal Dominant Polycystic Kidney Disease (ADPKD) [24] have demonstrated characteristic urinary metabolic profiles resulting from genetic mutations affecting the kidney. While it has previously been suggested that metabolomics studies of inborn errors of metabolism may aid toxicological research [235] by helping identify metabolic profiles characteristic of disruption to individual metabolic pathways, the results shown here suggest that the study of model nephrotoxins may help interpret the urinary metabolic profiles of patients with congenital kidney diseases.

Unlike simple enzyme or transporter deficiencies for which the metabolic consequences of genetic mutations are easily interpreted, many of these disorders show a complex set of organ-specific and systemic phenotypes that both vary between individuals and change over time. As a result, comparisons between the metabolic profiles of animals exposed to nephrotoxins with known mechanisms of toxicity and multiple related forms of congenital disorders will likely help clarify mechanisms of pathology for these more complex genetic traits and may suggest interventions to mitigate the progression of these diseases. While these disorders are rare, the increasingly routine use of whole-exome and whole-genome sequencing in the clinic [296] together with more traditional GWAS and linkage analysis means that new genetic disorders are being discovered at a rapid pace. As a result, comparisons of the metabolic and biochemical changes observed in small numbers of patients over an increasingly broad set of overlapping clinical phenotypes will likely aid both the understanding of gene function and help diagnose and treat patients with genetic disorders.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions from thesis

The work presented in this thesis applies metabolic profiling technologies, including Nuclear Magnetic Resonance and Mass Spectrometry, to identify the metabolic effects of genetic mutations. This work focused primarily on identifying metabolic markers resulting from the biochemical genetics of human diseases, including inborn errors of metabolism and congenital diseases of the kidney, but also makes comparisons to data collected from model organisms. Additionally, new methods for applying metabolic profiling to clinical populations that mitigate some of the challenges involved in biochemical genetics, such as small sample sizes, unknown or novel chemical structures, and the challenge of providing broader metabolomic coverage, were developed and presented here. As a result, the work presented investigates several aspects of biomarker discovery by applying metabolic profiling in biochemical genetics.

6.1.1 2D NMR Metabolic profiling in biochemical genetics

In Chapter 3 the challenges of structural characterisation of unknown or unexpected metabolites, spectral resolution, and small sample sizes were addressed by applying 2D ^1H - ^1H NMR techniques to profile the metabolomes of two nematode species, *C. elegans* mutant strains, and congenital Fanconi syndrome in a single human pedigree. Due to the lack of computational approaches to the analysis of 2D NMR data, methods for signal detection, alignment, pattern recognition and visualization were developed and presented. These methods, including Hierarchical Alignment of Two-Dimensional Spectra and Pattern Recognition (HATS-PR) [94]

and multivariate Differential Analysis by 2D NMR Spectroscopy (mvaDANS) [102] were applied to identify accumulating very long-chain ascarosides as a result of a peroxisomal β -oxidation defect in *C. elegans* and elevated urinary excretion of proximal tubular transporter substrates and TCA cycle intermediates resulting from mistrafficking of a peroxisomal β -oxidation protein to the mitochondria in a single human pedigree. These results suggest 2D NMR will be an important tool for metabolic profiling when a premium is placed on structural characterization in complex metabolic mixtures.

6.1.2 Metabolic profile screening and biomarker identification for inborn errors of metabolism

Chapter 4 compared tandem mass-spectrometry (MS/MS) and nanospray high resolution mass spectrometry (nS-HR-MS) analysis of DBS samples for 6,050 unaffected and 102 newborns affected by 24 congenital disorders of metabolism. While broader coverage of the metabolome may offer increased diagnostic sensitivity and specificity in newborn screening programs, the trade off between analytical specificity and coverage between MS/MS and nS-HR-MS meant that for the population of phenylketonuria (PKU) and medium-chain-acyl-CoA dehydrogenase deficiency (MCADD) patients screened in this work diagnostic sensitivity and specificity were roughly comparable. However, for both methods logistic regression models outperformed cutoff-based diagnostics. It is possible that more advanced non-linear, non-parametric machine learning methods such as random forests, neural networks, or support-vector machines may perform even better, but these methods tend to be unstable with low sample sizes and as a result much larger affected populations would need to be screened to investigate the application of these techniques. The broader coverage provided by nS-HR-MS did improve biomarker identification however, as shunt metabolites in the phenylalanine metabolic pathway and fatty acids and glycine conjugated fatty acids were identified in the PKU and MCADD populations by nS-HR-MS that were not measured by MS/MS. While low numbers of affected neonates limited the investigation of the metabolic consequences for the other 20 inborn errors for which samples were available, outlier identification relative to the control population and the other diseases suggested that greater

metabolic coverage and comparative investigation of a large set of inborn errors could provide important insight into the mechanisms of these diseases.

6.1.3 Metabolomic investigation of congenital diseases of the kidney

While biochemical genetics in clinical populations is typically associated with enzyme deficiencies, many Mendelian diseases result in complex, progressive pathology that can have individual organ or systemic effects. The work presented in Chapter 5 compares the urinary metabolic profiles of four congenital diseases of the kidney with overlapping phenotypes and investigates the metabolic natural history of one of these diseases, cystinosis. Additionally, the human urinary perturbations are compared to data collected from previous animal studies of exposure to region specific model nephrotoxins. These identify the main characteristics of proximal tubular pathology in both humans and animals, in which TCA cycle disruption is followed by loss of proximal tubular transporter substrates and an increase in markers of renal stress. These results both associate urinary metabolic profiles with region specific renal pathology and suggest connections with the mechanisms of pathology. The metabolic associations with transplantation, glomerular filtration-rate (GFR), and cysteamine therapy in cystinosis may also have ramifications for patient care, especially as TCA cycle disruption appears associated with decreased GFR and is mitigated by cysteamine therapy. Comparative metabolic profiling between diseases with overlapping phenotypes such as those presented here may play an important role in the study of complex Mendelian traits.

6.1.4 General conclusions and scope for extending this work

The work presented in this thesis, together with the rich literature on NMR and MS profiling of inborn errors and QTL analysis suggests an important and growing role for metabolic profiling in biochemical genetics. The identification of accumulating very-long-chain ascarosides in *C. elegans* mutants and elevated TCA cycle intermediates in Fanconi syndrome patients both show how metabolic profiling can help annotate the function of genes or mutations of unknown function and provide context on the mechanism and development of distinct phenotypes. Metabolic profiling, typically by targeted quantification of a small number of metabolites in

urine or blood, already plays a large role in medicine, but expanding the analytical coverage of the metabolome in these tests could enhance diagnostics and selection of therapeutics in stratified or personalized medicine in the future. The identification of an expanded set of metabolic biomarkers of inborn errors by an assay with broader metabolomic coverage will at least provide a better perspective on metabolic processes in health and disease. It is likely that larger studies will also show diagnostic improvements resulting from broader coverage.

While the studies presented here provide useful data for the phenotypes investigated and broader context about the uses of metabolic profiling in biochemical genetics, these results are limited in their scope by the small numbers of affected individuals included in these studies. This is unsurprising given how rare these conditions are, and the cystinosis results in particular indicate the additional power available to larger cohort studies. While comparisons across phenotypes and with controls enable improved biomarker discovery and interpretation, larger sample sets such as those available to national newborn screening programs and clinical trials could provide greater statistical power. Greater numbers of affected patient samples would be particularly beneficial in the analysis of inborn errors. Large, international collaborations between screening centers that allow a large number of cases of individually rare diseases to be collected may be promising.

Besides an increased number of affected samples, investigating a broader set of diseases would be an interesting area of further exploration as well. In the case of **congenital kidney diseases, comparison with related diseases such as Dent's disease**, the Oculocerebrorenal Syndrome of Lowe, polycystic kidney disease, and diseases related to individual transporters such as cystinuria, iminoglycinuria, and Hartnup disease could help resolve the effects of loss of tubular transporter activity from broader renal stress and allow greater resolution of the metabolic profiles associated with individual renal structures. The scope for an expanded set of inborn errors of metabolism is potentially even greater, as the map of known genes associated with Mendelian phenotypes (Fig. 1.2) indicates. These expanded studies will require adoption of screening technologies with broader coverage or improved bio-banking in rare disease research and medicine.

6.2 Future directions of metabolic profiling in biochemical genetics

Metabolomics, despite several decades of development, is still a relatively young field with substantial room for growth in terms of methodology, standardization, and adoption. Routine application of metabolic profiling in a non-research setting is hindered by the absence of a single analytical platform for metabolome analysis, the specialized knowledge required to interpret spectral data and identify biomarker chemical structures, and uncertainty regarding both interpretation of metabolome-wide data and the protection and management of resulting information. Advances in instrumentation, informatics, and analytics software will be necessary to resolve these outstanding challenges [120].

While the development of a highly sensitive and specific analytical platform that accurately measures a substantial fraction of the metabolome would solve many of these challenges, in the near term NMR and MS will likely continue to be applied in parallel and in conjunction. Metabolic profiling studies would benefit from improved integration of NMR and MS data as these techniques are complementary and both are often required for metabolite identification. Identification of metabolites will depend on the development of improved bioinformatic resources, including spectral databases. However, targeted assays and improved databases will still miss the presence of novel or unexpected metabolites, such as shunt metabolites. Acceleration and automation of the identification process remains the major bottleneck, both because of limitations described above, and perhaps more importantly, because of the great chemical diversity in metabolism. Protocols for rapid identification of metabolite structures will enhance untargeted metabolic pr

As a result, continued development of both analytical and computational methods in metabolomic research are critical for the widespread adoption of these techniques. New institutional programs will play an important role in translating the methodology of metabolic phenotyping and developing improved tools and **procedures for the collection and analysis of metabolic profile and related 'omics data** to support clinical decision-making through the framework of systems medicine. Large scale metabotyping for populations will yield increasing power to the identification of metabolic perturbations in disease and allow more sensitive and

specific classification of disease states on the basis of metabolic changes in biofluids. Just as the increased availability of human genome sequence data is providing important context for identifying disease associated variants, the availability of metabolic data on large populations will increasingly characterize normal ranges for reference in identifying outliers and abnormal metabolic profiles. Fulfilling the promise of metabolic profiling in systems medicine awaits the day when comprehensive measurement of biological variables, including gene sequences, epigenetic modifications, expression and translation levels, and metabolite abundances are routine and this data is widely available.

Genomics is currently undergoing yet another revolution as next-generation sequencing technologies increase accuracy, coverage, read-length, and drastically decrease the cost of whole exome sequencing (WES) and whole genome sequencing (WGS). The introduction of third-generation sequencing technologies in the near future promises to continue this trend [297]. Consequently, the near term promises a dramatic expansion in the availability of sequence data both in the lab and clinic. The relevance of increased sequence data to the continued integration of metabonomic and genomic data is twofold: first, an opportunity for metabonomics to contribute to the **increased clinical presence of ‘omics sciences led by genome sequencing**, and second, a challenge to develop methods of integrating metabolic profiles with sequences rather than polymorphisms.

Despite the significant opportunity represented by low-cost sequencing for the integration of genomic and metabonomic data and the identification of gene-metabolite associations, several challenges stand in the way of routine paired analysis of sequences and metabolic profiles. The first of these is the challenge of discovering significant associations with low sample numbers. Many of the successes reported in clinical sequencing have made use of data from a small number of patients, and sometimes only a sole patient. In these cases, potentially disease-causative variants are often identified using filtering strategies rather than statistical analysis [130, 298]. While the diagnosis of human inborn errors of metabolism from single patient biofluid NMR spectra demonstrates the potential of metabolic profiling to work with **low sample numbers**, **lack of statistical validation means that the “biological signal”** in these cases must be quite marked. A second challenge is a dearth of tools for statistical analysis of sequence data. While QTL mapping using single nucleotide polymorphisms is well established, statistical techniques for QTL mapping with both

rare and common variants are just beginning to be introduced [299]. As in metabolomics, the continued development of analytical and computational methods, combined with institutional initiatives that expand data collection and access will be necessary to overcome these challenges.

Of the ‘omics sciences, **genomics and metabolomics are uniquely** complementary, the strengths of each addressing weaknesses of the other. Genes are (mostly) static, an “upstream” blueprint controlling dynamic biological processes. **The identities and quantities of “downstream” metabolites capture both genetic and** environmental influences, and can be measured serially to assess variation through time. Genomic studies often struggle to establish a firm link between genetic variants and phenotypic observations, and while metabolomics provides a closer proxy to phenotype it is often difficult to infer underlying causality from variations in metabolism. Together, the integrated application of genomics and metabolomics promises a bridging of the gap between genotype and phenotype through intermediate metabolism to help annotate genes of unknown function, genetic controls of metabolism, and mechanisms of disease.

The first step in the process of whole-genome sequencing (WGS) is the preparation of a DNA library. This involves fragmenting the DNA into small pieces, typically 150-300 base pairs in length, and ligating them to sequencing adapters. The resulting library is then sequenced using a high-throughput sequencing platform, such as Illumina's Next-Generation Sequencing (NGS) technology. The raw sequencing data is then processed and aligned to a reference genome, allowing for the identification of genetic variants and structural changes.

Initial sequencing and analysis of the human genome. *Nature* 2001;409:908-921.

- 22 The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* 2010;109:20832-20837.
- 22 Heritability, weak effects, and rare variants in genomewide association studies. *Clin Chem* 2012;57:200-205.
- 22 Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2011;11:246-257.
- 22 Charting a course for genomic medicine from base pairs to bedside. *Nature* 2012;470:206-211.
- 22 Screening newborns for inborn errors of metabolism by tandem mass spectrometry. *N Engl J Med* 2003;348:1057-1068.
- 22 Proton nuclear magnetic resonance spectroscopy of body fluids in the field of inborn errors of metabolism. *Ann Clin Biochem* 2004;40:200-205.
- 22 Diagnosing inborn errors of lipid metabolism with proton nuclear magnetic resonance spectroscopy. *Clin Chem* 2005;52:200-205.
- 22 ¹H-NMR analysis of trimethylamine in urine for the diagnosis of fish-odour syndrome. *Clin Chim Acta* 2002;263:200-205.
- 22 Prolidase deficiency diagnosed by ¹H NMR spectroscopy of urine. *Inherit Metab Dis* 2002;24:200-205.
- 22 Proton nuclear magnetic resonance spectroscopic detection of sialic acid storage disease. *Clin Chem* 2002;48:200-205.
- 22 Guanidinoacetate

The quantitative trait locus on chromosome 2 for serum leptin levels is confirmed in African-Americans. *Diabetes* 2000;48:2000-2000

A major susceptibility locus influencing plasma triglyceride concentrations is located on chromosome 15q in Mexican Americans. *Am J Hum Genet* 2000;66:2000-2000

Genome-wide linkage analysis replicates susceptibility locus for fasting plasma triglycerides: NHLBI Family Heart Study. *Hum Genet* 2000;115:2000-2000

Linkage of high-density lipoprotein-cholesterol concentrations to a locus on chromosome 9p in Mexican Americans. *Nat Genet* 2000;30:2000-2000

Locus for quantitative HDL-cholesterol on chromosome 10q in Finnish families with dyslipidemia. *J Lipid Res* 2000;45:2000-2000

The genetics of plant metabolism. *Nat Genet* 2000;38:2000-2000

Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat Biotechnol* 2000;24:2000-2000

Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models. *Nat Genet* 2000;39:2000-2000

Untargeted metabolome quantitative trait locus mapping associates variation in urine glycerate to mutant glycerate kinase. *J Proteome Res* 2000;1:2000-2000

Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet* 2000;4:2000-2000

A genome-wide perspective of genetic variation in human metabolism. *Nat Genet* 2000;42:2000-2000

- Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLoS Genet* 2022;18(1):e1009577.
- A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet* 2022;18(1):e1009577.
- Human metabolic profiles are stably controlled by genetic and environmental variation. *Mol Syst Biol* 2022;18(1):e1009577.
- Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 2022;604(7821):477-482.
- A genome-wide association study of metabolic traits in human urine. *Nat Genet* 2022;54(1):1-10.
- Genome-wide association studies with metabolomics. *Genome Med* 2022;14(1):1-10.
- A Simple Phenylalanine Method for Detecting Phenylketonuria in Large Populations of Newborn Infants. *Pediatrics* 2022;150(1):e20211232.
- Use of tandem mass spectrometry for multianalyte screening of dried blood specimens from newborns. *Clin Chem* 2022;68(1):1-10.
- A review of clinical diagnostic applications of liquid chromatography-tandem mass spectrometry. *Mass Spectrom Rev* 2022;41(1):1-10.
- Newborn screening: toward a uniform screening panel and system. *Genet Med* 2022;24(1):1-10.

2022 2022 202220222022 2022 20222022 2022 20222022 2022 20222022 2022 20222022
 2022 20222022 2022 20222022 2022 Clinical application of exome sequencing in
 undiagnosed genetic conditions.*J Med Genet* 2022249:20222022
 2022 20222022 2022 2022 2022 2022 2022 2022 2022 2022 2022 2022 2022 2022
 20222022 202220222022 2022 202220222022 202220222022 2022 2022 202220222022 2022
 202220222022 2022 2022 202220222022 202220222022 2022 20222022 2022 20222022 2022
 2022 2022 20222022 202220222022 2022202220222022 2022 202220222022 2022
 Rapid whole-genome sequencing for genetic disease diagnosis in neonatal
 intensive care units.*Sci Transl Med* 202224:20222022
 2022 20222022 2022 20222022 2022 20222022 2022 20222022 2022 2022 20222022 2022
 20222022 2022 20222022 2022 20222022 202220222022 20222022 202220222022 2022
 202220222022 20222022 202220222022 202220222022 2022 20222022 202220222022 2022
 Clinical diagnosis by
 whole-genome sequencing of a prenatal sample.*N Engl J Med* 20222
 367:20222022
 2022 20222022 202220222022 202220222022 2022 20222022 2022 20222022 2022 20222022
 Preimplantation genetic
 diagnosis.*Nat Rev Genet* 202223:20222022
 2022 2022 2022 202220222022 2022 202220222022 2022 202220222022 2022 202220222022
 Preimplantation genetic
 diagnosis.*Lancet* 20222363:20222022
 2022 20222022 202220222022 2022 2022 20222022 2022 202220222022 2022 202220222022
 Attitudes towards carrier
 testing in minors: a systematic review.*Genet Couns* 2022216:20222022
 2022 20222022 2022 20222022 2022 20222022 2022 202220222022 2022 202220222022 2022
 202220222022 2022 20222022 2022 20222022 2022 20222022 2022 20222022 2022 20222022
 An
 empirical estimate of carrier frequencies for 400+ causal Mendelian
 variants: results from an ethnically diverse clinical sample of 23,453
 individuals.*Genet Med* 2022215:20222022
 2022 2022 2022 20222022 2022 20222022 2022 2022 2022 2022 2022 2022 2022 2022
 2022 2022 20222022202220222022 2022 20222022 202220222022 20222022 2022 2022
 20222022 202220222022 2022 202220222022 2022 202220222022 2022 202220222022
 Noninvasive whole-genome sequencing of a
 human fetus.*Sci Transl Med* 202224:20222022
 2022 20222022 2022 2022 20222022 2022 20222022 2022 202220222022 2022 202220222022
 Interallelic complementation in an
 inborn error of metabolism: genetic heterogeneity in
 argininosuccinate lyase deficiency.*Proc Natl Acad Sci USA* 20222
 81:20222022
 2022 20222022 202220222022 20222022 2022 20222022 20222022 2022 202220222022 2022
 20222022 2022 20222022202220222022 2022 20222022 2022 202220222022 20222022 2022
 202220222022 20222022 2022 2022 20222022 20222022 2022 202220222022 2022 20222022
 202220222022 20222022 2022 20222022 202220222022 202220222022 2022 20222022 2022
 202220222022 2022 20222022 202220222022 202220222022 2022 20222022 202220222022
 2022 20222022 2022 20222022 2022 20222022 2022 20222022 2022 202220222022 2022
 Phenotypic heterogeneity of
 genomic disorders and rare copy-number variants.*N Engl J Med* 20222
 367:20222022
 2022 2022 2022202220222022 20222022 20222022 2022 20222022 20222022 202220222022
 Applying Genomic Analysis to Newborn Screening.*Mol Syndromol*
 202223:20222022
 2022 20222022 2022 20222022 20222022202220222022 20222022 2022 2022 2022 2022
 20222022 202220222022 2022 2022202220222022 2022 2022 202220222022 20222022 2022 2022
 20222022 2022A functional genomics strategy that uses metabolome data

- Coherence Transfer by Isotropic Mixing - Application to Proton Correlation Spectroscopy. *J Magn Reson* 2000;53:222-232
- Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chem Phys Lett* 1999;69:222-232
- Mass spectrometry-based metabolomics. *Mass Spectrom Rev* 2000;26:222-232
- Mass spectrometry strategies in metabolomics. *J Biol Chem* 2000;286:222-232
- XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. *Anal Chem* 2000;80:222-232
- The emergence of mass spectrometry in biochemical research. *Proc Natl Acad Sci USA* 2000;91:222-232
- The basics of mass spectrometry in the twenty-first century. *Nat Rev Drug Discov* 2000;2:222-232
- MALDI imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology. *Histochem Cell Biol* 2000;130:222-232
- Altered metabolism and newborn screening using tandem mass spectrometry: lessons learned from the bench to bedside. *Curr Pharm Biotechnol* 2000;12:222-232
- High-resolution mass spectrometers. *Annu Rev Anal Chem* 2000;1:222-232
- Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabolomics. *Anal Chem* 2000;78:222-232
- Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 2000;7:222-232
- Recursive segment-wise peak alignment of biological (1)h NMR spectra for improved metabolic biomarker recovery. *Anal Chem* 2000;81:222-232
- rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Mag Reson Chem* 2000;47 Suppl 1:222-232
- icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *J Magn Reson* 2000;202:222-232
- Enhancing metabolomic data analysis with Progressive Consensus Alignment of NMR Spectra (PCANS). *BMC Bioinf* 2000;1:222-232
- Hierarchical alignment and full resolution pattern recognition of 2D NMR spectra: application to nematode chemical ecology. *Anal Chem* 2000;83:222-232

1. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* 2014;53:200-202.
2. The metabolic footprint of aging in mice. *Sci Rep* 2011;1:1.
3. Principal component analysis. *Chemometr Intell Lab* 2002;2:1-17.
4. SIMPLS: An alternative approach to partial least squares regression. *Chemometr Intell Lab* 2000;18:1-14.
5. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B* 2002;57:1-27.
6. Quantitative trait Loci analysis using the false discovery rate. *Genetics* 2002;171:1-11.
7. Omics-based clinical trial designs. *Curr Opin Oncol* 2007;25:1-11.
8. Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Stat Methods Med Res* 2007;23:1-11.
9. Assessment of PLS-DA cross validation. *Metabolomics* 2004;1:1-11.
10. Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. *J Proteome Res* 2007;6:1-11.
11. Robinette SL, Bruschweiler R, Schroeder FC, Edison AS: NMR in Metabolomics and Natural Products Research: Two Sides of the Same Coin. *Acc Chem Res* 2012;45(2):288-297.
12. METLIN: a metabolite mass spectral database. *Ther Drug Monit* 2007;27:1-11.
13. HMDB: the Human Metabolome Database. *Nucleic Acid Res* 2007;35:1-11.
14. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acid Res* 2009;37:1-11.
15. High-resolution proton magnetic-resonance spectroscopy of biological fluids. *Prog Nucl Magn Reson Spectrosc* 2000;21:1-11.
16. Metabonomics: a platform for studying drug toxicity and gene function. *Nat Rev Drug Discov* 2001;1:1-11.

- 111111 11111111111111111111 11 111111111111 11111111111111 11111111 111111111111
1111111111 11 Differential analysis of 2D NMR spectra: new natural products from a pilot-scale fungal extract library. *Angew Chem Int Ed Engl* 111111116:1111111111
- 11111 111111111 11 1111111 11 11 1 11111 1 111 111111 111 Time-resolved metabolic footprinting for nonlinear modeling of bacterial substrate utilization. *Appl Environ Microbiol* 111111175:1111111111
- 11111 111111 11111111 11 11 11111 111111111 11 111 111 11111111111 11 A metabolic signature of long life in *Caenorhabditis elegans*. *BMC Biol* 11111118:1111
- 11111 111111111 11 111111111111 11111 11111 11111 11 1111 111111 11111111111 11 11 1111111111 11 111111111111 111 Targeted metabolomics reveals a male pheromone and sex-specific ascaroside biosynthesis in *Caenorhabditis elegans*. *Acs Chem Biol* 11111117:111111111111
- 11111 111 11111 1 111 1 11111111111 11 11111111 111111 111 1 111111111111 1111 1111111111 11 Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 111111112:1111111111
- 11111 1111111111 11111111 111111 11111 11 11 1111 11111111111 11 11111111111111 11 1 111111111111 1111 A shortcut to identifying small molecule signals that regulate behavior and development in *Caenorhabditis elegans*. *Proc Natl Acad Sci USA* 1111111106:111111111111
- 11111 111 11 11111111 1 11111111 11111111 111111 11 11 1111111111111111 11111111 11 1 11111 111111111111 11 11 Comparison of 1D and 2D NMR spectroscopy for metabolic profiling. *J Proteome Res* 11111117:1111111111
- 11111 1 111111 1 11111111111111 11111111 11 Chemometrics in metabolomics--a review in human disease diagnosis. *Anal Chim Acta* 1111111659:11111111
- 11111 111111111111111111 11 1111111111 11111111111111 11111111 11 Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J Magn Reson* 1111111135:1111111111
- 11111 1 111 11 11 111111111 11 111111 11111 11 11 Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal Chem* 111111177:111111111111
- 11111 11111 11111 111 1 111111 111 11111111 111 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acid Res* 111111122:111111111111
- 11111 111111111 111 11 11 1 111111111 111111111 11 11111111 111 111111 111 111111111 11 Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ¹H NMR data sets. *Anal Chem* 111111177:111111111111
- 11111 11111111111 1 11 111111 1111111111111 111111111 11 Visualization and interpretation of OPLS models based on 2D NMR data. *Chemometr Intell Lab* 111111192:1111111111
- 11111 11111111111 1 11 11111111111111111 11111 11 11111 1111111111 1111111111 11 1111111111 11 Identification of lignin and polysaccharide modifications in *Populus* wood by chemometric analysis of 2D NMR spectra from dissolved cell walls. *Mol Plant* 11111112:1111111111

- Iterative Schemes for Bibear Operators; Application to Spin Decoupling. *J Magn Reson* 2007;7:222-2222
- Caenorhabditis elegans as an emerging model for studying the basic biology of obesity. *Dis Model Mech* 2002;2:2222222
- C elegans: a model for exploring the genetics of fat storage. *Dev Cell* 2004;2:2222222
- Propionyl-CoA and adenosylcobalamin metabolism in Caenorhabditis elegans: evidence for a role of methylmalonyl-CoA epimerase in intermediary metabolism. *Mol Genet Metab* 2003;39:222222
- The genetics of ageing. *Nature* 2002;464:2222222
- Biosynthesis of the Caenorhabditis elegans dauer pheromone. *Proc Natl Acad Sci USA* 2007;106:2222222222
- Comparative metabolomics reveals biogenesis of ascarosides, a modular library of small-molecule signals in C. elegans. *J Am Chem Soc* 2007;134:2222222222
- Lignoceric acid is oxidized in the peroxisome: implications for the Zellweger cerebro-hepato-renal syndrome and adrenoleukodystrophy. *Proc Natl Acad Sci USA* 2002;81:2222222222
- Biochemistry of mammalian peroxisomes revisited. *Annu Rev Biochem* 2007;75:22222222
- Medium chain acyl-CoA dehydrogenase deficiency human genome epidemiology review. *Genet Med* 2007;1:22222222
- Pseudo-Zellweger syndrome: deficiencies in several peroxisomal oxidative activities. *J Pediatr* 2002;108:2222222
- Peroxisomal bifunctional enzyme deficiency. *J Clin Invest* 2002;83:22222222
- Peroxisomal bifunctional protein deficiency revisited: resolution of its true enzymatic and molecular basis. *Am J Hum Genet* 2002;64:2222222
- Reinvestigation of peroxisomal 3-ketoacyl-CoA thiolase deficiency: identification of the true defect at the level of d-bifunctional protein. *Am J Hum Genet* 2002;70:2222222222
- Developmental changes of bile acid composition and conjugation in L- and D-bifunctional protein single and double knockout mice. *Developmental changes of bile acid*

- composition and conjugation in L- and D-bifunctional protein single and double knockout mice *Proc Natl Acad Sci U S A* 2007;104:11111-11116
- 2007 2007;104:11111-11116 **Finding function in novel targets: C. elegans as a model organism.** *Nat Rev Drug Discov* 2007;6:277-287
- 2007 2007;6:277-287 **Drosophila provides rapid modeling of renal development, function, and disease.** *Am J Physiol Renal Physiol* 2007;299:F299-F307
- 2007 2007;299:F299-F307 **Towards a platform for the metabonomic profiling of different strains of Drosophila melanogaster using liquid chromatography-Fourier transform mass spectrometry.** *FEBS J* 2007;276:276-287
- 2007 2007;276:276-287 **Complications of nephropathic cystinosis after renal failure.** *Pediatr Nephrol* 2007;61:200-207
- 2007 2007;61:200-207 **Glycine attenuates Fanconi syndrome induced by maleate or ifosfamide in rats.** *Kidney Int* 2007;71:249-257
- 2007 2007;71:249-257 **Idiopathic Fanconi syndrome in a family. Part I. Clinical aspects.** *J Am Soc Nephrol* 2007;18:222-231
- 2007 2007;18:222-231 **Proton NMR spectra of urine as indicators of renal damage. Mercury-induced nephrotoxicity in rats.** *Mol Pharmacol* 2007;72:27-35
- 2007 2007;72:27-35 **Mutation of the fumarase gene in two siblings with progressive encephalopathy and fumarase deficiency.** *J Clin Invest* 2007;117:293-301
- 2007 2007;117:293-301 **Metabolomic approaches to mitochondrial disease: correlation of urine organic acids.** *Mitochondrion* 2007;7:24-31
- 2007 2007;7:24-31 **Renal handling of citrate.** *Kidney Int* 2007;71:38-47
- 2007 2007;71:38-47 **Metabolic substrate utilization by rabbit proximal tubule. An NADH fluorescence study.** *Am J Physiol* 2007;294:254-261
- 2007 2007;294:254-261 **Deconvolution of Two-Dimensional NMR Spectra by Fast Maximum Likelihood Reconstruction (FMLR): Application to Quantitative Metabolomics.** *Anal Chem* 2007;79:2007-2015
- 2007 2007;79:2007-2015 **Principles and features of single-scan two-dimensional NMR spectroscopy.** *J Am Chem Soc* 2007;129:125-134
- 2007 2007;129:125-134 **Single-scan 2D Hadamard NMR spectroscopy.** *Angew Chem Int Ed Engl* 2007;46:248-251
- 2007 2007;46:248-251 **Theory of covariance nuclear magnetic resonance spectroscopy.** *J Chem Phys* 2007;126:121-131
- 2007 2007;126:121-131 **Covariance nuclear magnetic resonance spectroscopy.** *J Chem Phys* 2007;126:120-131
- 2007 2007;126:120-131 **Spectral reconstruction methods in fast NMR: reduced dimensionality, random sampling and maximum entropy.** *J Magn Reson* 2007;218:2-11
- 2007 2007;218:2-11 **Fast NMR: reduced dimensionality, random sampling and maximum entropy.** *J Magn Reson* 2007;218:2-11

- Predicting metabolic biomarkers of human inborn errors of metabolism.***Mol Syst Biol* 2022;5:2022
- Metabonomic, transcriptomic, and genomic variation of a population cohort.***Mol Syst Biol* 2022;6:2022
- The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis.***BMC Med Genomics* 2022;1:22
- Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods.***PLoS ONE* 2022;6:222222
- Screening for MCAD deficiency in newborns.***BMJ* 2022;38:2022
- Phenylketonuria.***Lancet* 2022;376:222222
- Outcome of neonatal screening for medium-chain acyl-CoA dehydrogenase deficiency in Australia: a cohort study.***Lancet* 2022;369:222222
- Enhanced interpretation of newborn screening results without analyte cutoff values.***Genet Med* 2022;14:222222
- Use of the phenylalanine:tyrosine ratio to test**

- Studies on the mechanism for renal elimination of N-acetylphenylalanine: its pathophysiologic significance in phenylketonuria. *J Lab Clin Med* 105:222-227
- Urinary excretion of N-acetyl amino acids in patients with some inborn errors of amino acid metabolism. *Scand J Clin Lab Invest Suppl* 184:222-227
- A role for Sp and nuclear receptor transcription factors in a cardiac hypertrophic growth program. *Proc Natl Acad Sci USA* 94:222-227
- General (medium-chain) acyl-CoA dehydrogenase deficiency (non-ketotic dicarboxylic aciduria): quantitative urinary excretion pattern of 23 biologically significant organic acids in three cases. *Clin Chim Acta* 132:222-227
- 3-, 6- and 7-hydroxyoctanoic acids are metabolites of medium-chain triglycerides and excreted in urine as glucuronides. *J Mass Spectrom* 31:222-227
- Gas chromatographic-mass spectrometric characterisation of unsaturated dicarboxylic acids in urine. *J Chromatogr* 534:222-227
- Rapid diagnosis of MCAD deficiency: quantitative analysis of octanoylcarnitine and other acylcarnitines in newborn blood spots by tandem mass spectrometry. *Clin Chem* 43:222-227
- Long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency. *Acta Paediatr* 88:222-227
- KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acid Res* 28:222-227
- Increased excretion of lactate, glutarate, 3-hydroxyisovalerate and 3-methylglutaconate during clinical episodes of propionic acidemia. *Clin Chim Acta* 123:222-227
- Variant maple syrup urine disease (MSUD)--the entire spectrum. *J Inher Metab Dis* 29:222-227
- Animal models reveal pathophysiologies of tyrosinemias. *J Nutr* 133:222-227
- Tyrosinemia type III: diagnosis and ten-year follow-up. *Acta Paediatr* 86:222-227
- Clinical, biochemical, and genetic analysis of a Korean neonate with hereditary tyrosinemia type 1. *Clin Chem Lab Med* 47:222-227
- Chronic tyrosinemia associated with 4-hydroxyphenylpyruvate dioxygenase deficiency with acute intermittent ataxia and without visceral and bone involvement. *Pediatr Res* 17:222-227

- Treatment of adult-type citrullinemia with administration of citrate.***Tohoku J Exp Med* 2000;134:223-227
- Demyelination of the brain is associated with methionine adenosyltransferase I/III deficiency.***J Clin Invest* 2000;98:223-227
- Guanidinoacetate methyltransferase deficiency: the first inborn error of creatine metabolism in man.***Am J Hum Genet* 2000;58:223-227
- NMR spectroscopy as a novel approach to the monitoring of renal transplant function.***Kidney Int* 2000;43:223-227
- Uroscopy in the 21st century: high-field NMR spectroscopy.***Nephrol Dial Transplant* 2000;12:223-227
- Metabolomics in the study of kidney diseases.***Nature Rev Nephrol* 2000
- Metabolomic analysis of uremic toxins by liquid chromatography/electrospray ionization-tandem mass spectrometry.***J Chromatogr B Analyt Technol Biomed Life Sci* 2000;378:223-227
- Acute renal failure following accidental cutaneous absorption of phenol: application of NMR urinalysis to monitor the disease process.***Hum Toxicol* 2000;8:223-227
- Mapping the biochemical trajectory of nephrotoxicity by pattern recognition of NMR urinalysis.***Nmr Biomed* 2000;5:223-227
- Integrated pathway analysis of rat urine metabolic profiles and kidney transcriptomic profiles to elucidate the systems toxicology of model nephrotoxins.***Chem Res Toxicol* 2000;21:223-227
- Discovery of metabolomics biomarkers for early detection of nephrotoxicity.***Toxicol Pathol* 2000;37:223-227
- Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting.***PLoS ONE* 2000;5:223-227
- ¹H-NMR-based metabolomic analysis of metabolic profiling in diabetic nephropathy rats induced by streptozotocin.***Am J Physiol Renal Physiol* 2000;300:223-227
- Metabolic profiling reveals new serum biomarkers for differentiating diabetic nephropathy.***Anal Bioanal Chem* 2000;404:223-227
- A metabolomics approach using juvenile cystic mice to**

identify urinary biomarkers and altered pathways in polycystic kidney disease. *Am J Physiol Renal Physiol* 2007;298:R1111-R1117

Urine metabolomic analysis identifies potential biomarkers and pathogenic pathways in kidney cancer. *OMICS* 2007;15:1111-1117

Urinary acylcarnitines are altered in human kidney cancer. *Int J Cancer* 2007;130:1111-1117

Metabolomics in monitoring kidney transplants. *Curr Opin Nephrol Hy* 2007;15:1111-1117

Metabolomics: a complementary tool in renal transplantation. *Contrib Nephrol* 2007;160:1111-1117

Investigations into the biochemical effects of region-specific nephrotoxins. *Mol Pharmacol* 2007;35:1111-1117

Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data: the consortium on metabonomic toxicology screening approach. *J Proteome Res* 2007;6:1111-1117

Biomarkers, metabonomics, and drug development: can inborn errors of metabolism help in understanding drug toxicity? *AAPS J* 2007;9:1111-1117

Cystine transport is defective in isolated leukocyte lysosomes from patients with cystinosis. *Science* 2007;217:1111-1117

Cystinosis. *N Engl J Med* 2007;347:1111-1117

Molecular pathogenesis of cystinosis: effect of CTNS mutations on the transport activity and subcellular localization of cystinosin. *Hum Mol Genet* 2007;13:1111-1117

Early development of the renal lesions in infantile cystinosis. *Pediatr Nephrol* 2007;15:1111-1117

Pharmacological treatment of nephropathic cystinosis with cysteamine. *Expert Opin Pharmacother* 2007;5:1111-1117

Twice-daily cysteamine bitartrate therapy for children with cystinosis. *J Pediatr* 2007;156:1111-1117

Cysteamine therapy delays the progression of nephropathic cystinosis in late adolescents and adults. *Kidney Int* 2007;31:1111-1117

Clinical and laboratory findings in the oculocerebrorenal syndrome of Lowe, with special reference to growth and renal function. *N Engl J Med* 2007;324:1111-1117

- Whole-Exome sequencing identifies FAM20A mutations as a cause of amelogenesis imperfecta and gingival hyperplasia syndrome. *Am J Hum Genet* 2022;88:2022-2032
- Novel FAM20A mutations in hypoplastic amelogenesis imperfecta. *Hum Mutat* 2022;33:2022-2032
- Nephrocalcinosis (Enamel Renal Syndrome) Caused by Autosomal Recessive FAM20A Mutations. *Nephron Physiol* 2022;122:2022-2032
- FAM20A mutations can cause enamel-renal syndrome (ERS). *PLoS Genetics* 2022;19:2022-2032
- Nephrocalcinosis: molecular insights into calcium precipitation within the kidney. *Clin Sci* 2022;106:2022-2032
- Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicol Appl Pharmacol* 2022;187:2022-2032
- The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. *Pharmacogenomics* 2022;6:2022-2032
- Temporal metabonomic modeling of l-arginine-induced exocrine pancreatitis. *J Proteome Res* 2022;7:2022-2032
- On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *J Am Stat Assoc* 2022;62:2022-2032
- Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2022;338:2022-2032

- Methods to adjust for bias and confounding in critical care health services research involving observational data. *J Crit Care* 2022;21:0000
- The nature of the renal lesion in the Fanconi syndrome. *J Pathol Bacteriol* 2000;65:00000000
- Evaluation of the proximal tubular function in hereditary renal Fanconi syndrome. *Nephrol Dial Transplant* 2000;23:00000000
- Experimental Fanconi's syndrome resulting from 4-pentenoate infusion in the dog. *Am J Physiol* 2000;257:00000000
- Renal amino acid transport in adults with oxidative phosphorylation diseases. *Kidney Int* 2000;47:00000000
- Urinary L-lactate excretion is increased in renal Fanconi syndrome. *Nephrol Dial Transplant* 2000;19:00000000
- Renal cells in culture as a model for cystinosis. *J Basic Clin Physiol Pharmacol* 2000;01:00000000
- Intracellular cystine loading inhibits transport in the rabbit proximal convoluted tubule. *J Clin Invest* 2000;85:00000000
- Role of adenosine triphosphate (ATP) and NaK ATPase in the inhibition of proximal tubule transport with intracellular cystine loading. *J Clin Invest* 2000;87:00000000
- The Fanconi syndrome with cystinosis. Electron microscopy of renal biopsy specimens from five patients. *Am J Med* 2000;33:00000000
- Distal myopathy in nephropathic cystinosis. *Pediatr Nephrol* 2000;14:00000000
- Nephropathic cystinosis associated with cardiomyopathy: a 27-year clinical follow-up. *BMC Nephrol* 2000;3:00
- Steroid myopathy in connective tissue disease. *Am J Med* 2000;61:00000000
- Negligible urinary cysteamine loss in cystinosis patients with Fanconi syndrome. *Clin Nephrol* 2000;57:00000000
- Nuclear magnetic resonance (NMR)-based metabolomic studies on urine and serum biochemical profiles after chronic cysteamine supplementation in rats. *J Agr Food Chem* 2000;59:00000000
- Metabolic profiling strategy for discovery of nutritional biomarkers: proline betaine as a marker of citrus consumption. *Am J Clin Nutr* 2000;92:00000000
- Methylamine in human urine. *Clin Chim Acta* 2000;312:00000000
- Low urinary citrate excretion in nephrolithiasis. *Urology* 2000;21:00000000

- Correction of hypocitraturia and prevention of stone formation by combined thiazide and potassium citrate therapy in thiazide-unresponsive hypercalciuric nephrolithiasis. *Am J Med* 2000;79:222-227
- Hypocitraturia is one of the major risk factors for nephrocalcinosis in very low birth weight (VLBW) infants. *Kidney Int* 2000;63:222-227
- The pathogenesis of cystinosis: mechanisms beyond cystine accumulation. *Am J Physiol Renal Physiol* 2000;279:222-227
- Use of high-resolution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine. *Clin Chem* 2000;30:222-227
- Rapid screening of metabolic diseases by proton NMR. *Lancet* 2000;355:222-227
- Indexing glomerular filtration rate to suit children. *J Nucl Med* 2000;44:222-227
- Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. *Anal Chem* 2000;79:222-227
- Amino acid transport across mammalian intestinal and renal epithelia. *Physiol Rev* 2000;88:222-227
- Biochemical diagnosis of mitochondrial disorders. *J Inher Metab Dis* 2000;34:222-227
- Accumulation of trimethylamine and trimethylamine-N-oxide in end-stage renal disease patients undergoing haemodialysis. *Nephrol Dial Transplant* 2000;21:222-227
- Proton nuclear magnetic resonance spectroscopy of urine and plasma in renal transplantation follow-up. *Nephron* 2000;65:222-227
- Choline and L-carnitine as precursors of trimethylamine. *Biochem Soc Trans* 2000;25:222-227
- Dietary precursors of trimethylamine in man: a pilot study. *Food Chem Toxicol* 2000;37:222-227
- Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat Med* 2000;19:222-227
- Cyclosporin A-induced changes in endogenous metabolites in rat urine: a metabolomic investigation using high field 1H NMR spectroscopy, HPLC-TOF/MS and chemometrics. *J Pharmaceut Biomed* 2000;35:222-227
- Toxicometabolomics approach to urinary biomarkers for

- mercuric chloride (HgCl₂)-induced nephrotoxicity using proton nuclear magnetic resonance (¹H NMR) in rats.*Toxicol Appl Pharmacol* 2009;249:222-232
- Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells.*PLoS Comput Biol* 2011;7:e1002002
- PPAR alpha ligand protects during cisplatin-induced acute renal failure by preventing inhibition of renal FAO and PDC activity.*Am J Physiol Renal Physiol* 2009;286:F1111-F1119
- Mercuric chloride-induced reactive oxygen species and its effect on antioxidant enzymes in different regions of rat brain.*J Environ Sci Health B* 2003;32:111-120
- Low mercury concentrations cause oxidative stress and endothelial dysfunction in conductance and resistance arteries.*Am J Physiol Heart Circ Physiol* 2002;295:H1111-H1118
- Understanding renal toxicity of heavy metals.*Toxicol Pathol* 2007;26:22-33
- Metabolomic study of cisplatin-induced nephrotoxicity.*Kidney Int* 2009;269:111-120
- Effects of cisplatin on different measures of glomerular function in the human kidney with special emphasis on high-dose.*Cancer Chemother Pharmacol* 2002;21:111-120
- 2-Bromoethanamine nephrotoxicity in the nude mouse: an atypical targeting for the renal cortex.*Int J Exp Pathol* 2007;71:111-120
- Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time.*Genet Med* 2013;13:111-120
- Exome sequencing identifies the cause of a mendelian disorder.*Nat Genet* 2012;42:111-120
- Strategies for exome and genome sequence data analysis in disease-gene discovery projects.*Clin Genet* 2010;80:111-120
- Region-based and pathway-based QTL mapping using a p-value combination method.*BMC proceedings* 2015 Suppl 9:111-120

Appendix A: Supplementary Tables

Appendix Table 1. Internal standards detected by nS-HR-MS

Internal standard	Original conc. ($\mu\text{mol/l}$)	Working conc. ($\mu\text{mol/l}$)	m/z $[\text{M}+\text{H}]^+$	m/z $[\text{M}+\text{Na}]^+$	m/z $[\text{M}-\text{H}]^-$
-------------------	--	---	-----------------------------	---------------------------------	-----------------------------

¹⁵ N; 2- ¹³ C-Glycine	25	10	78.04024	100.02218	76.02459
² H ₄ -Alanine	5	2	94.08061	116.06256	92.06496
² H ₈ -Valine	5	2	126.13702	148.11896	124.12137
² H ₂ -Ornithine	5	2	135.11026	157.09220	133.09461
² H ₃ -Leucine	5	2	135.12128	157.10323	133.10563
² H ₃ -Aspartate	5	2	137.06416	159.04611	135.04851
² H ₃ -Glutamate	5	2	151.07981	173.06176	149.06416
² H ₃ -Methionine	5	2	153.07771	175.05965	151.06206
¹³ C ₆ -Phenylalanine	5	2	172.10690	194.08885	170.09125
² H ₂ -Citrulline	5	2	178.11607	200.09802	176.10042
² H ₄ ; 5- ¹³ C-Arginine	5	2	180.14796	202.12990	178.13231
¹³ C ₆ -Tyrosine	5	2	188.10182	210.08376	186.08617
² H ₉ -Camitine	1.52	0.608	171.16951	193.15145	
² H ₃ -Acetylcarnitine	0.38	0.152	207.14241	229.12436	
² H ₃ -Propionylcarnitine	0.076	0.0304	221.15806	243.14001	
² H ₃ -Butyrylcarnitine	0.076	0.0304	235.17371	257.15566	
² H ₉ -Isovalerylcarnitine	0.076	0.0304	255.22703	277.20897	
² H ₃ -Octanoylcarnitine	0.076	0.0304	291.23631	313.21826	
² H ₉ -Myristoylcarnitine	0.076	0.0304	381.36788	403.34982	
² H ₃ -Palmitoylcarnitine	0.152	0.0608	403.36151	425.34346	

Appendix Table 2. Masses targeted in nS-HR-MS assay

m/z	Compound Name	Adduct
73.0290	Propionic acid	[M-H] ⁻

74.0242	Glycine	[M-H] ⁻
75.0082	Glycolic acid	[M-H] ⁻
76.0399	Glycine	[M+H] ⁺
83.0221	Urea	[M+Na] ⁺
85.0290	2-Deoxytetronic acid	[M-H] ⁻
87.0082	Pyruvic acid	[M-H] ⁻
88.0399	Alanine	[M-H] ⁻
88.9875	Oxalic acid	[M-H] ⁻
89.0239	Lactic acid	[M-H] ⁻
90.0555	Alanine	[M+H] ⁺
91.0395	Glycerol	[M-H] ⁻
93.0340	Phenol	[M-H] ⁻
96.9691	Phosphoric acid	[M-H] ⁻
98.0218	Glycine	[M+Na] ⁺
101.0239	C ₄ H ₆ O ₃ (Acetoacetic acid; Ketobutyric acid)	[M-H] ⁻
101.0603	C ₅ H ₁₀ O ₂ (Pivalic acid; Valeric acid; Isovaleric acid; 2-Methylbutyric acid)	[M-H] ⁻
102.0191	C ₃ H ₅ NO ₃ (Pyruvatoxime; 3-Oxoalanine)	[M-H] ⁻
102.0555	C ₄ H ₉ NO ₂ (Aminobutyric acid; Dimethylglycine)	[M-H] ⁻
103.0031	Malonic acid	[M-H] ⁻
103.0144	Urea-1-carboxylate	[M-H] ⁻
103.0395	C ₄ H ₈ O ₃ (Hydroxybutyric acid; Hydroxyisobutyric acid)	[M-H] ⁻
104.0348	Serine	[M-H] ⁻
105.0188	Glyceric acid	[M-H] ⁻
106.0504	Serine	[M+H] ⁺
108.0449	Aminophenol	[M-H] ⁻
111.0082	3-Furoic acid	[M-H] ⁻
111.0195	Uracil	[M-H] ⁻
111.0446	C ₆ H ₈ O ₂ (6-Hydroxy-2,4-hexadienal; Hexenedial; 4-Oxo-2-hexenal; etc)	[M-H] ⁻
111.0534	N,N'-Dimethylurea	[M+Na] ⁺
112.0374	Alanine	[M+Na] ⁺
112.0511	Creatinine	[M-H] ⁻

113.0239	C ₅ H ₆ O ₃ (2-Hydroxy-2,4-pentadienoate; 3-Oxo-4-pentenoic acid)	[M-H] ⁻
114.0555	Proline	[M-H] ⁻
114.0667	Creatinine	[M+H] ⁺
115.0031	C ₄ H ₄ O ₄ (Maleic acid; Fumaric acid)	[M-H] ⁻
115.0395	C ₅ H ₈ O ₃ (Keto-isovaleric acid; Keto-valeric acid; Methyl acetoacetate; Levulinic acid; 2-Ketovaline; 5-Oxo-pentanoic acid)	[M-H] ⁻
115.0759	C ₆ H ₁₂ O ₂ (Methyl-valeric acid; Diethyl-acetic acid; Dimethyl-ethyl-acetic acid; 3,3-Dimethyl-butanoic acid; etc)	[M-H] ⁻
116.0348	C ₄ H ₇ NO ₃ (Aceturic acid; L-Aspartic acid b-semialdehyde; 2S-Amino-3-oxo-butanoic acid)	[M-H] ⁻
116.0712	Proline	[M+H] ⁺
116.0712	Valine	[M-H] ⁻
117.0188	C ₄ H ₆ O ₄ (Methylmalonic acid; Succinic acid)	[M-H] ⁻
117.0552	C ₅ H ₁₀ O ₃ (Hydroxyisovaleric acid; Hydroxyvaleric acid; 2-Methyl-3-hydroxybutyric acid)	[M-H] ⁻
118.0140	C ₃ H ₅ NO ₄ (Aminomalonic acid; 3-Nitropropionic acid)	[M-H] ⁻
118.0504	Threonine	[M-H] ⁻
118.0868	Valine	[M+H] ⁺
119.0344	C ₄ H ₈ O ₄ (Dihydroxybutyric acid; Dihydroxyisobutyric acid; Erythrose; Erythrulose; etc)	[M-H] ⁻
120.0661	Threonine	[M+H] ⁺
121.0290	Benzoic acid	[M-H] ⁻
124.0068	Taurine	[M-H] ⁻
124.9909	Hydroxyethanesulfonate	[M-H] ⁻
128.0324	Serine	[M+Na] ⁺
128.0348	C ₅ H ₇ NO ₃ (Pyroglutamic acid; N-Acryloylglycine; 4-Oxoproline)	[M-H] ⁻
129.0188	C ₅ H ₆ O ₄ (Mesaconic acid; Glutaconic acid; Citraconic acid; etc)	[M-H] ⁻
129.0552	C ₆ H ₁₀ O ₃ (Mevalonolactone; 2-Methyl-3-ketovaleric acid; 2-Ketohexanoic acid; etc)	[M-H] ⁻
130.0504	C ₅ H ₉ NO ₃ (Propionylglycine; Hydroxyproline)	[M-H] ⁻
130.0868	C ₆ H ₁₃ NO ₂ (Leucine; Isoleucine)	[M-H] ⁻
131.0344	C ₅ H ₈ O ₄ (Ethylmalonic acid; Methylsuccinic acid; Glutaric acid)	[M-H] ⁻
131.0457	Asparagine	[M-H] ⁻
131.0708	C ₆ H ₁₂ O ₃ (2-Hydroxyisocaproic acid; 2-Hydroxy-3-methylvaleric acid; 5-Hydroxyhexanoic acid)	[M-H] ⁻

131.0821	Ornithine	[M-H] ⁻
132.0297	Aspartic acid	[M-H] ⁻
132.0661	Hydroxyproline	[M+H] ⁺
132.0773	Creatine	[M+H] ⁺
132.1025	C ₆ H ₁₃ NO ₂ (Leucine; Isoleucine)	[M+H] ⁺
133.0137	Malic acid	[M-H] ⁻
133.0613	Asparagine	[M+H] ⁺
133.0977	Ornithine	[M+H] ⁺
134.0453	Aspartic acid	[M+H] ⁺
135.0294	Erythronic acid; Threonic acid	[M-H] ⁻
135.0446	Phenylacetic acid	[M-H] ⁻
137.0239	4-Hydroxybenzoic acid	[M-H] ⁻
137.0351	Urocanic acid	[M-H] ⁻
138.0531	Proline	[M+Na] ⁺
140.0687	Valine	[M+Na] ⁺
141.0188	5-Hydroxymethyl-2-furoic acid	[M-H] ⁻
142.0480	Threonine	[M+Na] ⁺
143.0344	C ₆ H ₈ O ₄ (3-Methylglutaconic acid; 2-Hexenedioic acid; 3-Hydroxyadipic acid 3,6-lactone)	[M-H] ⁻
143.1072	C ₈ H ₁₆ O ₂ (Valproic acid; Caprylic acid)	[M-H] ⁻
144.0661	C ₆ H ₁₁ NO ₃ (Isobutyrylglycine; Butyrylglycine)	[M-H] ⁻
145.0137	2-Ketoglutaric acid	[M-H] ⁻
145.0501	C ₆ H ₁₀ O ₄ (3-Methylglutaric acid; Adipic acid)	[M-H] ⁻
145.0613	Glutamine	[M-H] ⁻
145.0977	Lysine	[M-H] ⁻
146.0453	Glutamic acid	[M-H] ⁻
147.0294	Hydroxyglutaric acid	[M-H] ⁻
147.0657	Mevalonic acid	[M-H] ⁻
147.0770	Glutamine	[M+H] ⁺
147.1134	Lysine	[M+H] ⁺
148.0432	Methionine	[M-H] ⁻
148.0610	Glutamic acid	[M+H] ⁺

148.9909	Thiodiglycolic acid	[M-H] ⁻
149.0086	Tartaric acid	[M-H] ⁻
149.0450	Arabinose	[M-H] ⁻
150.0589	Methionine	[M+H] ⁺
151.0256	Xanthine	[M-H] ⁻
151.0395	Hydroxyphenylacetic acid	[M-H] ⁻
152.0712	Dopamine	[M-H] ⁻
153.0188	Gentisic acid	[M-H] ⁻
154.0480	Hydroxyproline	[M+Na] ⁺
154.0617	Histidine	[M-H] ⁻
154.0844	C ₆ H ₁₃ NO ₂ (Leucine; Isoleucine)	[M+Na] ⁺
155.0093	Orotic acid	[M-H] ⁻
155.0433	Asparagine	[M+Na] ⁺
156.0661	C ₇ H ₁₁ NO ₃ (Tiglylglycine; 3-Methylcrotonylglycine)	[M-H] ⁻
156.0773	Histidine	[M+H] ⁺
157.0362	Allantoin	[M-H] ⁻
157.0501	Succinylacetone	[M-H] ⁻
157.0865	C ₈ H ₁₄ O ₃ (2-Propyl-3-ketopentanoic acid; Alpha-ketooctanoic acid)	[M-H] ⁻
158.0817	Isovalerylglycine, 2-Methylbutyrylglycine	[M-H] ⁻
159.0294	Oxadipic acid	[M-H] ⁻
159.0657	C ₇ H ₁₂ O ₄ (3-Methyladipic acid; Pimelic acid)	[M-H] ⁻
159.1021	C ₈ H ₁₆ O ₃ (2-Propyl-3-hydroxypentanoic acid; 2-Propyl-5-hydroxypentanoic acid; 7-Hydroxyoctanoic acid)	[M-H] ⁻
161.0450	C ₆ H ₁₀ O ₅ (3-Hydroxy-3-methylglutaric acid; 2-Hydroxyadipic acid)	[M-H] ⁻
162.1130	Camitine (C0)	[M+H] ⁺
163.0395	C ₉ H ₈ O ₃ (Phenylpyruvic acid; Coumaric acid)	[M-H] ⁻
164.0712	Phenylalanine	[M-H] ⁻
165.0552	Phenyllactic acid	[M-H] ⁻
166.0868	Phenylalanine	[M+H] ⁺
167.0205	Uric acid	[M-H] ⁻
167.0344	C ₈ H ₈ O ₄ (Vanillic acid; Homogentisic acid)	[M-H] ⁻
168.0661	Norepinephrine	[M-H] ⁻

169.0589	Glutamine	[M+Na] ⁺
171.0657	Octenedioic acid	[M-H] ⁻
172.0408	Methionine	[M+Na] ⁺
172.0974	Hexanoylglycine	[M-H] ⁻
172.9909	Phenyl sulfate	[M-H] ⁻
173.0086	Aconitic acid	[M-H] ⁻
173.0814	C ₈ H ₁₄ O ₄ (2-Propylglutaric acid; Suberic acid)	[M-H] ⁻
173.1039	Arginine	[M-H] ⁻
174.0402	N-Acetylaspartic acid	[M-H] ⁻
174.0555	Indole-3-acetic acid	[M-H] ⁻
174.0879	Citrulline	[M-H] ⁻
175.0243	Ascorbic acid	[M-H] ⁻
175.1195	Arginine	[M+H] ⁺
176.1035	Citrulline	[M+H] ⁺
178.0504	Hippuric acid	[M-H] ⁻
178.0592	Histidine	[M+Na] ⁺
179.0344	Hydroxyphenylpyruvic acid	[M-H] ⁻
179.0556	C ₆ H ₁₂ O ₆ (Galactose; Fructose)	[M-H] ⁻
180.0661	Tyrosine	[M-H] ⁻
181.0501	C ₉ H ₁₀ O ₄ (Homovanillic acid; Hydroxyphenyllactic acid)	[M-H] ⁻
181.0712	Galactitol	[M-H] ⁻
182.0817	Tyrosine	[M+H] ⁺
182.0817	Epinephrine	[M-H] ⁻
184.0950	Camitine (C0)	[M+Na] ⁺
185.1178	C ₁₀ H ₁₈ O ₃ (Hydroxydecanoic acid; Oxo capric acid)	[M-H] ⁻
187.0970	Azelaic acid	[M-H] ⁻
188.0018	Aminophenol sulfate	[M-H] ⁻
188.0687	Phenylalanine	[M+Na] ⁺
188.9858	Pyrocatechol sulfate	[M-H] ⁻
189.0763	3-Hydroxysuberic acid	[M-H] ⁻
191.0192	Citric acid	[M-H] ⁻
194.0453	2-Hydroxyhippuric acid	[M-H] ⁻

195.0505	C ₆ H ₁₂ O ₇ (Gluconic acid; Galactonic acid)	[M-H] ⁻
196.0610	L-3,4-Dihydroxyphenylalanine	[M-H] ⁻
196.0974	Metanephrine	[M-H] ⁻
197.0450	Vanilmandelic acid	[M-H] ⁻
198.0855	Citrulline	[M+Na] ⁺
199.0970	Decenedioic acid	[M-H] ⁻
201.1127	Sebacic acid	[M-H] ⁻
203.0532	C ₆ H ₁₂ O ₆ (Galactose; Fructose)	[M+Na] ⁺
203.0821	Tryptophan	[M-H] ⁻
204.0637	Tyrosine	[M+Na] ⁺
204.0661	Indolelactic acid	[M-H] ⁻
204.1236	Acetylcamitine (C2)	[M+H] ⁺
205.0348	Methyicitric acid	[M-H] ⁻
205.0688	Galactitol	[M+Na] ⁺
205.0977	Tryptophan	[M+H] ⁺
206.0817	Phenylpropionylglycine	[M-H] ⁻
215.1032	Diacetylornithine	[M-H] ⁻
215.1647	3-Hydroxydodecanoic acid	[M-H] ⁻
216.9807	Hydroxybenzoic acid sulfate	[M-H] ⁻
217.1076	Hydroxysebacic acid	[M-H] ⁻
218.1028	Pantothenic acid	[M-H] ⁻
218.1392	Propionylcarnitine (C3)	[M+H] ⁺
220.0821	N-Acetylhexosamine	[M-H] ⁻
225.0875	Porphobilinogen	[M-H] ⁻
226.1055	Acetylcamitine (C2)	[M+Na] ⁺
227.0796	Tryptophan	[M+Na] ⁺
227.2011	Myristic acid	[M-H] ⁻
230.1028	Suberylglycine	[M-H] ⁻
232.1549	Butyrylcarnitine (C4)	[M+H] ⁺
232.9756	Gentisic acid sulfate	[M-H] ⁻
238.9862	Oxoadipic acid sulfate	[M-H] ⁻
239.0160	Cystine	[M-H] ⁻

239.9967	Dihydroxyquinoline sulfate	[M-H] ⁻
240.1212	Propionylcarnitine (C3)	[M+Na] ⁺
242.0123	Methyldioxyindole sulfate	[M-H] ⁻
242.9963	Coumaric acid sulfate	[M-H] ⁻
243.0617	Uridine	[M-H] ⁻
244.1549	Tiglylcarnitine (C5:1)	[M+H] ⁺
245.1389	3-Hydroxydodecanedioic acid	[M-H] ⁻
246.1705	Isovalerylcarnitine (C5)	[M+H] ⁺
248.1134	Malonylcarnitine (C3DC)	[M+H] ⁺
249.0069	3,4-Dihydroxyphenylglycol O-sulfate	[M-H] ⁻
254.1368	Butyrylcarnitine (C4)	[M+Na] ⁺
254.9811	Ascorbate sulfate	[M-H] ⁻
255.2324	Palmitic acid	[M-H] ⁻
259.0219	Galactose-1-phosphate	[M-H] ⁻
260.1862	Hexanoylcarnitine (C6)	[M+H] ⁺
262.1291	Succinylcarnitine (C4DC)	[M+H] ⁺
262.1654	3-Hydroxyisovalerylcarnitine (C5OH)	[M+H] ⁺
263.1032	Phenylacetylglutamine	[M-H] ⁻
266.1368	Tiglylcarnitine (C5:1)	[M+Na] ⁺
268.1525	Isovalerylcarnitine (C5)	[M+Na] ⁺
269.0661	Phenol glucuronide	[M-H] ⁻
269.2481	Heptadecanoic acid	[M-H] ⁻
270.0954	Malonylcarnitine (C3DC)	[M+Na] ⁺
273.0069	Ferrulic acid sulfate	[M-H] ⁻
275.2011	Stearidonic acid	[M-H] ⁻
276.1447	Glutarylcarnitine (C5DC)	[M+H] ⁺
277.2168	Alpha-linolenic acid	[M-H] ⁻
279.2324	Linoleic acid	[M-H] ⁻
281.2481	Oleic acid	[M-H] ⁻
282.1681	Hexanoylcarnitine (C6)	[M+Na] ⁺
283.2637	Stearic acid	[M-H] ⁻
284.1110	Succinylcarnitine (C4DC)	[M+Na] ⁺

284.1474	3-Hydroxyisovalerylcarnitine (C5OH)	[M+Na] ⁺
285.0610	Pyrocatechol	[M-H] ⁻
288.2175	Octanoylcarnitine (C8)	[M+H] ⁺
291.1305	Argininosuccinate	[M+H] ⁺
297.0974	Phenylethanol glucuronide	[M-H] ⁻
298.1267	Glutaryl carnitine (C5DC)	[M+Na] ⁺
300.0389	N-Acetylhexosamine sulfate	[M-H] ⁻
301.2168	Eicosapentaenoic acid	[M-H] ⁻
302.1604	C6DC acylcarnitine	[M+H] ⁺
303.2324	Arachidonic acid	[M-H] ⁻
305.2481	Eicosatrienoic acid	[M-H] ⁻
306.0760	Glutathione	[M-H] ⁻
307.2637	Eicosadienoic acid	[M-H] ⁻
308.0916	Glutathione	[M+H] ⁺
309.2794	Gondoic acid	[M-H] ⁻
310.1994	Octanoylcarnitine (C8)	[M+Na] ⁺
311.2950	Arachidic acid	[M-H] ⁻
312.2175	Decadienoylcarnitine (C10:2)	[M+H] ⁺
314.2331	Decenoylcarnitine (C10:1)	[M+H] ⁺
316.2488	Decanoylcarnitine (C10)	[M+H] ⁺
327.2324	Docosahexaenoic acid	[M-H] ⁻
329.2481	Docosapentaenoic acid	[M-H] ⁻
333.2794	Adrenic acid	[M-H] ⁻
334.1994	Decadienoylcarnitine (C10:2)	[M+Na] ⁺
335.2950	Docosadienoic acid	[M-H] ⁻
336.2151	Decenoylcarnitine (C10:1)	[M+Na] ⁺
337.3107	Erucic acid	[M-H] ⁻
338.2307	Decanoylcarnitine (C10)	[M+Na] ⁺
339.3263	Behenic acid	[M-H] ⁻
344.2801	Lauroylcarnitine (C12)	[M+H] ⁺
355.2637	Tetracosahexaenoic acid	[M-H] ⁻
355.3212	2-Hydroxydocosanoic acid	[M-H] ⁻

357.2794	Tetracosapentaenoic acid	[M-H] ⁻
366.2620	Lauroylcarnitine (C12)	[M+Na] ⁺
367.3576	Lignoceric acid	[M-H] ⁻
368.2801	Tetradecadienoylcarnitine (C14:2)	[M+H] ⁺
370.2957	Tetradecenoylcarnitine (C14:1)	[M+H] ⁺
372.3114	Myristoylcarnitine (C14)	[M+H] ⁺
373.2743	Hydroxy-cholenoic acid	[M-H] ⁻
388.3063	3-Hydroxy-tetradecanoylcarnitine (C14OH)	[M+H] ⁺
389.2692	Dihydroxy-cholenoic acid	[M-H] ⁻
390.2620	Tetradecadienoylcarnitine (C14:2)	[M+Na] ⁺
391.2848	C24H40O4 (Deoxycholic acid; Chenodeoxycholic acid; Ursodeoxycholic acid)	[M-H] ⁻
392.2777	Tetradecenoylcarnitine (C14:1)	[M+Na] ⁺
394.2933	Myristoylcarnitine (C14)	[M+Na] ⁺
395.3889	Cerotic acid	[M-H] ⁻
398.3270	Palmitoleoylcarnitine (C16:1)	[M+H] ⁺
400.3427	Palmitoylcarnitine (C16)	[M+H] ⁺
410.2882	3-Hydroxy-tetradecanoylcarnitine (C14OH)	[M+Na] ⁺
413.2668	Dihydroxy-cholenoic acid	[M+Na] ⁺
415.3212	Hydroxy-cholestenoic acid	[M-H] ⁻
416.3376	3-Hydroxy-hexadecanoylcarnitine (C16OH)	[M+H] ⁺
417.3369	Hydroxy-cholestanoic acid	[M-H] ⁻
420.3090	Palmitoleoylcarnitine (C16:1)	[M+Na] ⁺
422.3246	Palmitoylcarnitine (C16)	[M+Na] ⁺
423.2512	PA (17:0)	[M-H] ⁻
424.3427	Linoleylcarnitine (C18:2)	[M+H] ⁺
426.3583	Oleylcarnitine (C18:1)	[M+H] ⁺
428.3740	Stearoylcarnitine (C18)	[M+H] ⁺
431.3161	Dihydroxy-cholestenoic acid	[M-H] ⁻
433.3318	Dihydroxy-cholestanoic acid	[M-H] ⁻
438.3195	3-Hydroxy-hexadecanoylcarnitine (C16OH)	[M+Na] ⁺
439.3188	Hydroxy-cholestenoic acid	[M+Na] ⁺

440.3376	3-Hydroxy-linoleylcarnitine (C18:2OH)	[M+H] ⁺
441.3345	Hydroxy-cholestanoic acid	[M+Na] ⁺
442.3532	3-Hydroxy-octadecenoylcarnitine (C18:1OH)	[M+H] ⁺
444.3689	3-Hydroxy-octadecanoylcarnitine (C18OH)	[M+H] ⁺
445.2954	Dihydroxy-oxocholestenoic acid	[M-H] ⁻
446.3246	Linoleylcarnitine (C18:2)	[M+Na] ⁺
447.3111	Trihydroxy-cholestenoic acid	[M-H] ⁻
448.3063	Glycochenodeoxycholic acid	[M-H] ⁻
448.3403	Oleoylcarnitine (C18:1)	[M+Na] ⁺
449.3267	Trihydroxy-cholestanoic acid	[M-H] ⁻
450.3559	Stearoylcarnitine (C18)	[M+Na] ⁺
450.3583	C20:3 acylcarnitine	[M+H] ⁺
452.3740	C20:2 acylcarnitine	[M+H] ⁺
455.3137	Dihydroxy-cholestenoic acid	[M+Na] ⁺
457.3294	Dihydroxy-cholestanoic acid	[M+Na] ⁺
461.2903	Trihydroxy-oxocholestenoic acid	[M-H] ⁻
462.3195	3-Hydroxy-linoleylcarnitine (C18:2OH)	[M+Na] ⁺
463.3060	Tetrahydroxy-cholestenoic acid	[M-H] ⁻
464.3012	Glycocholic acid	[M-H] ⁻
464.3352	3-Hydroxy-octadecenoylcarnitine (C18:1OH)	[M+Na] ⁺
465.3039	Cholesterol-sulfate	[M-H] ⁻
466.3508	3-Hydroxy-octadecanoylcarnitine (C18OH)	[M+Na] ⁺
469.2930	Dihydroxy-oxocholestenoic acid	[M+Na] ⁺
471.3086	Trihydroxy-cholestenoic acid	[M+Na] ⁺
473.3243	Trihydroxy-cholestanoic acid	[M+Na] ⁺
478.2934	C23H46NO7P (Lyso-PE (18:1); Lyso-PC (15:1))	[M-H] ⁻
481.2988	26-Hydroxycholesterol-3-sulfate	[M-H] ⁻
485.2879	Trihydroxy-oxocholestenoic acid	[M+Na] ⁺
487.3036	Tetrahydroxy-cholestenoic acid	[M+Na] ⁺
496.3385	C24H50NO7P (Lyso-PC (16:0); Lyso-PE (19:0))	[M+H] ⁺
498.2889	Taurochenodeoxycholic acid	[M-H] ⁻
500.2777	Lyso-PE (20:4)	[M-H] ⁻

514.2839	Taurocholic acid	$[M-H]^-$
518.3205	Lyso-PC (18:3)	$[M+H]^+$
549.3556	PA (13:0/12:0)	$[M-H]^-$

Appendix B: Paper I

Izrayelit Y[†], Robinette SL[†], Bose N, von Reuss SH, Schroeder FC. 2D NMR-based metabolomics uncovers interactions between conserved biochemical pathways in the model organism *Caenorhabditis elegans*. *ACS Chem Biol*, 2012. 8(2):314-319. doi:10.1021/cb3004644.

2D NMR-Based Metabolomics Uncovers Interactions between Conserved Biochemical Pathways in the Model Organism *Caenorhabditis elegans*

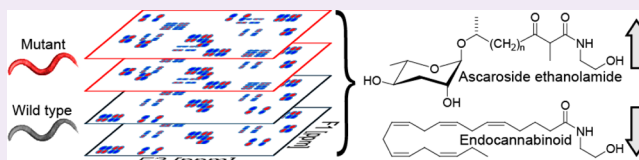
Yevgeniy Izrayelit,^{†,⊥} Steven L. Robinette,^{‡,⊥} Neelanjan Bose,[†] Stephan H. von Reuss,[†] and Frank C. Schroeder^{*,†}

[†]Boyce Thompson Institute and Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, United States

[‡]Department of Surgery and Cancer, Imperial College London, London, United Kingdom

Supporting Information

ABSTRACT: Ascarosides are small-molecule signals that play a central role in *C. elegans* biology, including dauer formation, aging, and social behaviors, but many aspects of their biosynthesis remain unknown. Using automated 2D NMR-based comparative metabolomics, we identified ascaroside ethanolamides as shunt metabolites in *C. elegans* mutants of *daf-22*, a gene with homology to mammalian 3-ketoacyl-CoA thiolases predicted to function in conserved peroxisomal lipid β -oxidation. Two groups of ethanolamides feature β -keto functionalization confirming the predicted role of *daf-22* in ascaroside biosynthesis, whereas α -methyl substitution points to unexpected inclusion of methylmalonate at a late stage in the biosynthesis of long-chain fatty acids in *C. elegans*. We show that ascaroside ethanolamide formation in response to defects in *daf-22* and other peroxisomal genes is associated with severe depletion of endocannabinoid pools. These results indicate unexpected interaction between peroxisomal lipid β -oxidation and the biosynthesis of endocannabinoids, which are major regulators of lifespan in *C. elegans*. Our study demonstrates the utility of unbiased comparative metabolomics for investigating biochemical networks in metazoans.



The free-living nematode *Caenorhabditis elegans* is used extensively as a model system for the study of aging and metabolism.^{1–3} Recent analyses of the *C. elegans* metabolome revealed a family of structurally diverse small molecules, the ascarosides, as products of conserved peroxisomal β -oxidation (Figure 1).^{4–6} The ascarosides form a modular compound library, integrating building blocks from lipid, carbohydrate and amino acid metabolism into small molecule signals that regulate key aspects of *C. elegans* biology, including dauer formation and social behaviors.^{7–10} Previous work used mass spectrometry-based comparative metabolomics to characterize the functions of three enzymes in ascaroside biosynthesis, the acyl-CoA oxidase ACOX-1, the enoyl-CoA hydratase MAOC-1, and the 3-hydroxyacyl-CoA dehydrogenase DHS-28 (Figure 1b).⁴ However, the predicted role of the fourth enzyme in the peroxisomal β -oxidation cascade, which is encoded by the gene *daf-22*, a close homologue of mammalian 3-keto-acyl-CoA thiolases, could not be confirmed because no characteristic shunt metabolites were found. Given the central role of *daf-22* in the proposed peroxisomal β -oxidation pathway, we hypothesized that other, yet unidentified metabolites accumulate in *daf-22* mutants, which may help confirm the biochemical function of this enzyme and provide insight into interactions with other biochemical pathways.

Therefore, we reinvestigated the *daf-22* metabolome using unbiased NMR spectroscopy-based metabolomics. Previously,

we had employed manual comparisons of 2D NMR spectra of wild-type and *daf-22* metabolomes to identify additional members of the ascaroside family of signaling molecules.^{6,9,11} These studies highlighted the utility of 2D NMR spectra for comparative metabolomics but also demonstrated the need for computational tools that would enable statistical analysis of series of spectra from different mutants. For this purpose, we developed multivariate Differential Analysis by 2D NMR Spectroscopy (mvaDANS), which enables automated processing and comparative computational analysis for sets of 2D NMR spectra.

mvaDANS builds on previous efforts to use 2D NMR for metabolomic analysis by incorporating dynamic signal detection and pattern recognition (see Methods).^{12–18} Figure 2 summarizes the use of this approach for identifying compounds that are upregulated in *daf-22* mutants and thus may represent shunt metabolites or biosynthetic intermediates. We acquired dqfCOSY spectra for metabolome extracts obtained from two different *daf-22* alleles (*m130*, *ok693*) and wild-type *C. elegans*, including four independent biological replicates (Figure 2a, Supplementary Figures S1 and S2). Automatic crosspeak identification and binning (Supplemen-

Received: August 31, 2012

Accepted: November 16, 2012

Published: November 19, 2012

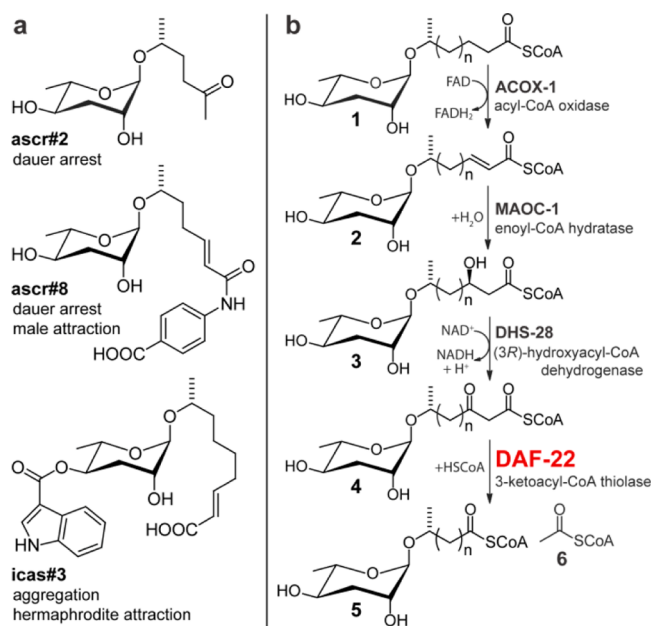


Figure 1. Ascarosides are signaling molecules in *C. elegans*. (a) Example structures and biological functions of ascaroside-based signaling molecules in *C. elegans*. (b) Model for peroxisomal biosynthesis of ascaroside side chains and roles of peroxisomal β -oxidation enzymes ACOX-1, MAOC-1, DHS-28, and the putative 3-ketoacyl-CoA thiolase DAF-22.

tary Figure S3) prepared the dqfCOSY spectra for statistical analysis via Principal Component Analysis (PCA, see Methods),¹⁹ which clustered the data by genotype (Figures 2a). Coefficients from the PCA loadings were back-projected onto the dqfCOSY spectra, which revealed a large number of crosspeaks that are up- or downregulated in *daf-22* mutants relative to wild-type (Figures 2a, Supplementary Figures S4 and S5).

For this study, we restricted analysis to crosspeaks that were strongly featured in spectra from both *daf-22* alleles but were very weak or undetectable in wild-type. Furthermore, we excluded crosspeaks also present in the *E. coli* food source. Manual analysis of the remaining strongly *daf-22*-upregulated crosspeaks and inputs from additional HMBC spectra suggested a series of partial structures, including methyl ketones, α -methyl branched fatty acids, and corresponding ethanolamides, as well as β -keto fatty acid derivatives (Figure 2b). Considering the proposed role of *daf-22* in ascaroside biosynthesis (Figure 1b), we hypothesized that these structural features belong to accumulated shunt metabolites. Subsequent fractionation of the *daf-22* metabolome guided by the *daf-22*-specific NMR signals led to isolation of representative members of six different families of very long-chain ascarosides (VLCA) that account for all of the predicted structural features (Figure 3, see Supporting Information for full spectroscopic data). None of these *daf-22* mutant-specific compounds had previously been reported. Most prominent are a series of (ω -1)-*O*-ascarosyl 2-ketones (**12**), as well as (ω -1)-*O*-ascarosyl 3-keto fatty acid ethanolamides, which include the 2-methyl pentacosanoic acid derivative **7** and the straight-chain heptacosanoic acid derivative **8**. Compounds **7** and **8** are accompanied by smaller amounts of the corresponding bis-norhomologues. Similarly, the 2-methyl pentacos-2-enoic acid ethanolamide **9** and the two fatty acids **10** and **11** are

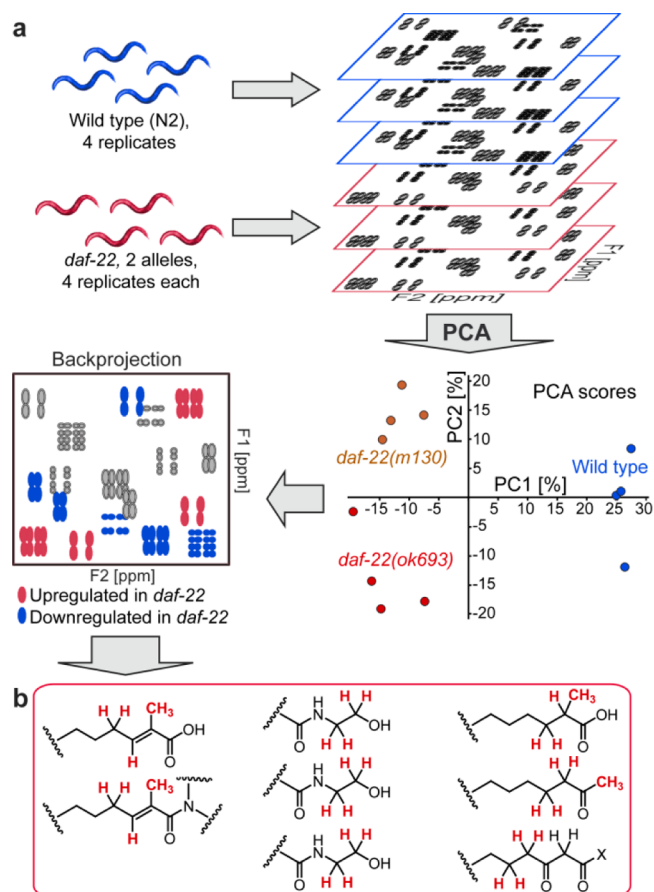


Figure 2. mvaDANS reveals structural motifs upregulated in *daf-22* mutants. (a) Schematic overview of mvaDANS for identification of up- and downregulated metabolites in *daf-22* mutants. In the principal component analysis (PCA), PC 1 separates the four wild-type data sets (blue dots) from both *daf-22* mutant data sets (red and brown dots), whereas PC2 separates the two *daf-22* alleles. In the back projection (lower left), the coloring of cross peaks corresponds to the PC 1 loading coefficients; blue peaks represent signals downregulated and red peaks signals upregulated in the two *daf-22* mutant strains. See Supplementary Figure S4 for actual spectra. (b) Partial structures inferred from manual analysis of *daf-22*-specific crosspeaks suggesting α -methyl branched fatty acids, methyl ketones, and β -keto fatty acid derivatives.

accompanied by bis-norhomologues. In contrast, chain lengths of the (ω -1)-*O*-ascarosyl 2-ketones (**12**) vary more broadly and are not preferentially odd-numbered (Figure 3c). HPLC–MS analysis confirmed the absence of these components in wild-type samples (Figure 3a).

The incorporation of ethanolamine in the *daf-22* specific VLCAs was unexpected and suggested the possibility of interactions between peroxisomal β -oxidation and ethanolamide-based signaling pathways. Recently, *N*-acyl-ethanolamines (NAEs) were identified as endocannabinoid-like signaling molecules in *C. elegans* and linked to dietary restriction-dependent lifespan regulation.²⁰ We asked whether the upregulation of VLCAs in *daf-22* mutants affects NAE levels. We found that NAE production, including the most active endocannabinoid in *C. elegans*,²⁰ eicosapentaenoyl ethanolamide (**13**, EPEA), and the ligand of mammalian cannabinoid receptors, arachidonoyl ethanolamide (AEA, or anandamide), was dramatically downregulated in both *daf-22* alleles (Figure 3d and Supplementary Figure S6), suggesting a shift of

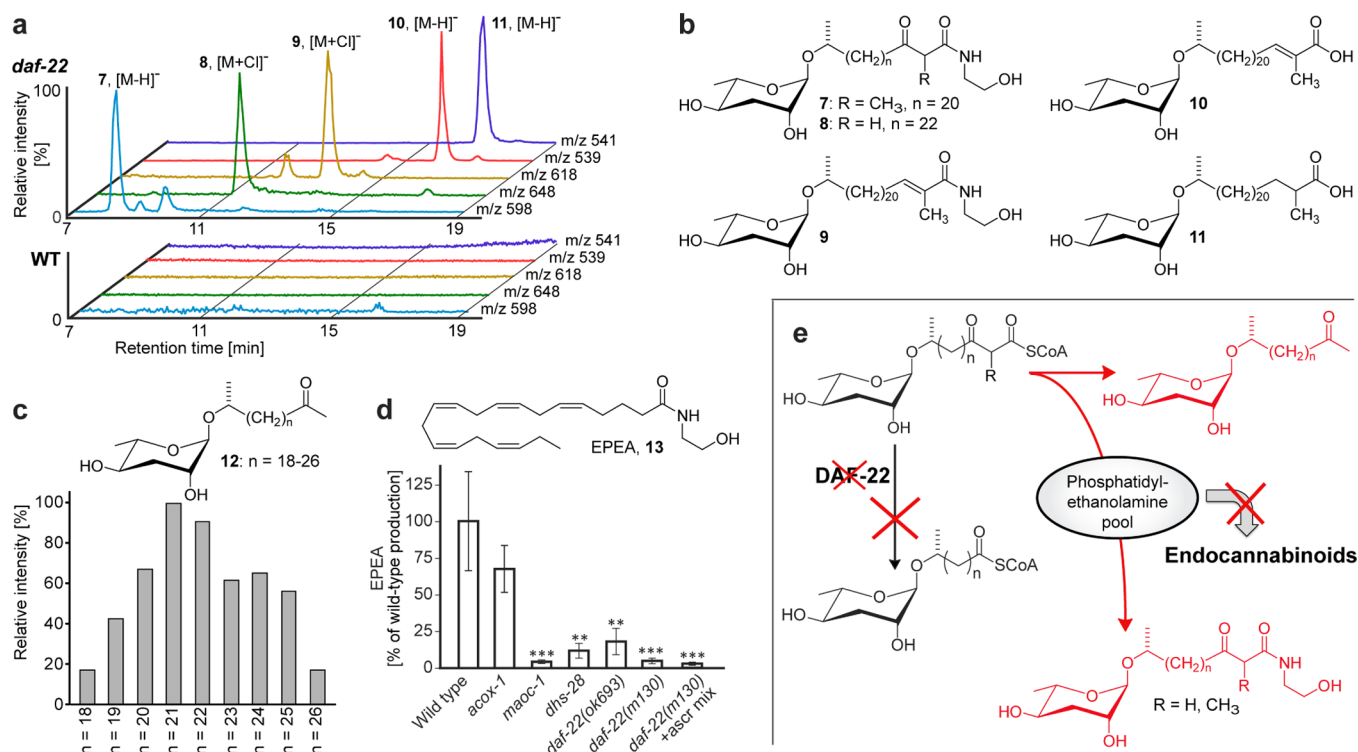


Figure 3. Long-chain ascarosides identified in the *daf-22* mutant metabolome and interaction with NAE biosynthesis. (a) HPLC–MS analysis confirms long-chain ascarosides and ethanolamides in the *daf-22* metabolome that are absent in wild-type. (b) *daf-22* mutant-specific VLCAs identified via *mvaDANS* and subsequent HPLC–MS analysis. (c) Quantitative distribution of *daf-22*-upregulated ascarosyl methylketones. (d) Mutation of *maoc-1*, *dhs-28*, and *daf-22* greatly reduces EPEA production and addition of ascarosides (2.5 μ M ascr#3 and ascr#9, see ref 4 for structures) to *daf-22(m130)* cultures does not rescue EPEA production. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. (e) Interactions of peroxisomal β -oxidation and endocannabinoid biosynthesis. Mutation of *daf-22* abolishes processing of long-chain ascaroside CoA esters, whose conversion into ethanolamides is associated with reduced endocannabinoid production,¹³ likely due to depletion of phosphatidylethanolamine pools.

ethanolamine utilization from the NAEs to VLCAs. Similarly, we found greatly reduced EPEA and AEA production in mutants of *maoc-1* and *dhs-28* (Figure 3d and Supplementary Figure S6), which act directly upstream of *daf-22* (Figure 1b). EPEA-levels in *acox-1* mutant worms were not significantly affected, and AEA levels only slightly decreased, corresponding to the much weaker effect of the *acox-1* mutation on peroxisomal β -oxidation.⁴ EPEA and AEA production was not rescued by growing *daf-22* worms with addition of synthetic ascarosides (Figure 3d and Supplementary Figure S6), suggesting that the observed reduction of NAE levels in peroxisomal β -oxidation mutants is not an indirect effect of the lack of ascaroside pheromone. These findings indicate that in *C. elegans* specific defects in peroxisomal β -oxidation result in depletion of endocannabinoid pools (Figure 3e). It is unknown whether conditions that reduce expression of peroxisomal β -oxidation genes in healthy animals trigger down-regulation of endocannabinoids, which could serve as a regulatory mechanism tying fat metabolism and ascaroside signaling to dietary restriction pathways. Ethanolamides of short-chain ascarosides serve as pheromones in a parasitic nematode species,²¹ suggesting that interactions of peroxisomal β -oxidation and ethanolamide metabolism also exist in other nematode genera.

Mammalian homologues of *daf-22* are thought to be specifically required for chain shortening of branched-chain fatty acids such as phytanic acid.^{22,23} However, the co-occurrence of straight-chain and methyl-branched derivatives among the *daf-22* mutant-specific VLCAs suggests that *daf-22* is involved in processing of both branched and straight-chain fatty

acids. The chain lengths of the *daf-22* mutant-specific VLCAs indicate that 23- to 27-carbon fatty acids may represent end points of *C. elegans* fatty acid biogenesis, in which the last chain-extension step incorporates either malonate or methylmalonate. Notably, our metabolomic analyses provided no evidence for incorporation of methylmalonate at earlier stages in the biosynthesis of these fatty acid derivatives. The chain lengths of the *daf-22* mutant-upregulated VLCAs in *C. elegans* are similar to those of very long-chain fatty acids that accumulate in human disorders involving defects in peroxisomal β -oxidation such as Zellweger syndrome.^{22,23} However, human fatty acid biosynthesis is not known to selectively incorporate propionate or methylmalonate in the last chain-extension step to form α -methyl-branched variants.

In a biosynthetic model (Figures 1b, 3e), the identified *daf-22* mutant-specific compounds plausibly represent shunt metabolites derived from fatty-acid and β -keto-acid CoA esters, in which the α,β -unsaturation and β -keto functionalities result from upstream action of ACOX-1, MAOC-1, and DHS-28. The *daf-22*-upregulated long-chain methyl ketones may derive from corresponding β -keto fatty acids, which are highly prone to decarboxylation. Although described in an earlier study,⁵ we could not find any evidence for β -keto fatty acids in our metabolite extracts by either NMR spectroscopy or HPLC–MS. The odd-numbered side-chain lengths of the *daf-22*-upregulated ascarosyl fatty acids are consistent with the structures of the known end products of peroxisomal β -oxidation, the short-chain ascaroside pheromones, most of which include side chains with 3, 5, 7, or 9 carbons (Figure

1a).⁴ These odd-numbered side chains further suggest that ascaroside side-chain biosynthesis starts with a three-carbon (methylmalonate-derived) template or that *C. elegans* elongates odd-numbered fatty acids obtained from food.

In conclusion, mvaDANS-based comparison of *daf-22* and wild-type metabolomes revealed long-chain ascarosyl ethanolamides as an unexpected class of shunt metabolites in *daf-22* mutant worms. The abundance of β -keto derivatives and methyl ketones among the identified long-chain ascarosides supports that *daf-22* acts as a thiolase in ascaroside pheromone biosynthesis, as had been suspected on the basis of homology to mammalian peroxisomal thiolases. Our results will motivate more detailed inspection of similarities of nematode and mammalian fatty-acid metabolism, especially with regard to possible interactions between endocannabinoid signaling, phosphatidylethanolamine utilization, and peroxisomal β -oxidation. In *C. elegans*, reduced NAE levels promote dietary-restriction-induced lifespan extension,²⁰ suggesting that defects in peroxisomal β -oxidation may affect lifespan via changes in NAE levels.

Untargeted comparative metabolomics will form an important tool for reinvestigating metazoan biochemistry. mvaDANS extends the scope of 2D NMR-based metabolomic analysis^{24–26} by introducing automated cross-peak identification and integration. Whereas pattern recognition techniques have been applied previously to 2D NMR spectra, these approaches have either relied on manual peak selection,²⁷ parametric approaches,²⁸ or used spectral unfolding/refolding techniques.^{10,11} We expect that advances in statistical spectroscopy will continue to accelerate the structural and functional characterization of biogenic small molecules.

METHODS

C. elegans Strains and General Culture Methods. Wild-type *C. elegans* (N2, Bristol), *daf-22(m130)*, *daf-22(ok693)*, *acox-1(ok2257)*, *maoc-1(hj13)*, and *dhs-28(hj8)* mutant strains were obtained from the *Caenorhabditis* Genetics Center (CGC). Strains were maintained at 20 °C on NGM plates with bacteria (*E. coli* OP50) as food.

Preparation of C. elegans Metabolite Extracts. Liquid cultures were grown as previously reported.³ Briefly, wild-type or *daf-22* mutant worms were grown for two generations on NGM plates seeded with OP50. Three crowded 6 cm NGM plates were washed into 100 mL solution of S-medium in a 500 mL Erlenmeyer flask and grown at 22 °C and 220 rpm. Concentrated OP50 from 1 L bacterial cultures, grown overnight, were given on day 1 and day 3. On day 5 the liquid culture was split into two 500 mL Erlenmeyer flasks, and the growth media volume was maintained at 100 mL. An additional 1 L OP50 pellet was given as food upon splitting. The cultures were harvested on day 7 by centrifugation at 4750 rpm. The supernatant was lyophilized, and the residue was extracted with 95% ethanol (300 mL) at RT for 12 h. The resulting suspension was filtered and evaporated *in vacuo* at RT to yield the crude extracts. In this manner four extracts each of wild-type (N2), *daf-22(m130)*, and *daf-22(ok693)* cultures were prepared. For NAE quantification, mutant strains were grown as above in duplicate. Liquid cultures for ascaroside mixture experiments were prepared as described above with addition of 2.5 μ M of ascr#3 and 2.5 μ M of ascr#9 at day 0.

In preparation for NMR spectroscopic analysis, the crude extracts were suspended in 1:8 methanol (MeOH)/dichloromethane (DCM), using 7.5 mL of this solvent mixture per 1 g of crude extract. Following centrifugation, 2 mL aliquots of supernatant of each sample were filtered over 1 g of silica (EMD Chemicals Inc.) using 10 mL of 1:8 MeOH/DCM for elution. The filtrate was evaporated to dryness *in vacuo* prior to analysis. For NMR spectroscopic analysis, 10–20 mg of the resulting metabolite extracts was used.

Preparation of Bacterial Metabolite Extracts. *E. coli* OP50 was grown overnight at 37 °C and 220 rpm to an OD of 1.2 in 2 L of Lysogeny broth (LB). The bacterial broth was spun for 20 min at 4750 rpm. The supernatant was then lyophilized, crushed with mortar and pestle, and extracted with 150 mL 95% ethanol (2 times). The extract was filtered through a sintered crucible and concentrated *in vacuo* prior to analysis.

Mass Spectrometric Analysis. Metabolite extracts prepared as described above were resuspended in 1.5 mL methanol and centrifuged, and 30 μ L of this sample was injected into HPLC–MS for analysis. HPLC–MS analysis was performed using the Agilent 1100 Series HPLC system with an Agilent Eclipse XDB-C18 column (9.4 mm $\times$ 250 mm, 5 μ m particle diameter) connected to the Quattro II mass spectrometer using a 10:1 split. For short chain ascarosides, a 0.1% aqueous acetic acid–acetonitrile gradient was used. The gradient started with 5% acetonitrile for 5 min and then increased to 100% acetonitrile over 32 min. For long chain ascarosides, an acetonitrile–isopropanol gradient was used, starting with 5% isopropanol in acetonitrile for 5 min; then isopropanol was increased to 75% over 32 min. Samples were analyzed by HPLC–ESI–MS in negative and positive ion modes using a capillary voltage of 3.5 kV and a cone voltage of –40 V and +20 V, respectively. The elemental composition of the long chain ascarosides was performed by high resolution MS and HPLC–MS using the Xevo G2 QTof.

EPEA and AEA (14) were quantified by HPLC–MS (using the Quattro II mass spectrometer) using a water (0.1% acetic acid)–acetonitrile gradient as described above for short chain ascarosides. The retention time of EPEA and AEA were established by injection of an authentic standard (Enzo Life Sciences and Cayman Chemicals, respectively). Relative endocannabinoid content in wild-type and mutant strains was quantified by integration of LC–MS signals from corresponding ion traces for (M – H)[–] or (M + Cl)[–]. All integration values were normalized by the liquid culture pellet dry weight. All quantitative data reported in Figure 3 were from at least two independent biological repeats.

NMR Spectrometric Analysis. NMR samples were prepared by dissolving 10–20 mg portions of metabolite extract (see above) in 0.6 mL of CD₃OD. Samples were analyzed using a Varian INOVA 600 MHz NMR spectrometer. Nongradient phase-cycled dqfCOSY spectra were truncated at 0.4 s and acquired using the following parameters: 0.8 s acquisition time, 600 complex increments, 16 scans per increment. For processing with mvaDANS, spectra were zero-filled to 2048 $\times$ 1024 data points, a sine-square-shaped window function was applied in both dimensions prior to Fourier transformation, and spectra were converted to magnitude mode prior to transfer to MATLAB.

For detailed analysis in phase-sensitive mode, dqfCOSY spectra were zero-filled to 8192 $\times$ 4096 data points and a cosine bell-shaped window function was applied in both dimensions before Fourier transformation. Gradient and nongradient HSQCAD and HMBC spectra of enriched fractions of *daf-22*-upregulated metabolites were acquired using 8–32 scans, 0.25 s acquisition time, and 256–500 complex increments. NMR spectra were processed using Bruker TopSpin, Varian VNMR, MestreLabs' MestReC, and MNova software packages.

Isolation of Major *daf-22*-Upregulated Metabolites. To 8 g of Celite (prewashed with ethyl acetate) was added a solution of *daf-22* metabolite extract from 16 100-mL liquid cultures. After evaporation of the solvent, the material was dry-loaded into an empty 25 g RediSep Rf loading cartridge. Fractionation was performed using a Teledyne ISCO CombiFlash system over a RediSep Rf GOLD 40 g HP Silica Column using a normal phase dichloromethane–methanol solvent system, starting with 6 min of 100% dichloromethane, followed by a linear increase of methanol content up to 10% at 24 min, followed by another linear increase of methanol content up to 40% at 40 min, followed by another linear increase of methanol content up to 95% at 45 min, which was then continued to 48 min. This yielded 70 fractions (~20 mL each), which were individually evaporated *in vacuo* and analyzed by NMR spectroscopy (¹H NMR, dqfCOSY) and HPLC–MS. ¹H NMR spectroscopic signals identified via mvaDANS as

representing *daf-22*-upregulated metabolites were detected in fractions 35–45. Isolation of major *daf-22*-upregulated metabolites was achieved by preparative HPLC using the same HPLC method described above for HPLC–MS analysis of long chain ascarosides. Fractions containing *daf-22*-upregulated metabolites were characterized further by dqfCOSY, HSQC, and HMBG as well as high-resolution HPLC–MS using the Xevo G2 QTof.

Statistical Analysis. Figure 3d and Supplementary Figure S6: we used unpaired student's *t* tests with Welch's correction for comparing the relative abundance of EPEA and AEA in wild type and mutant strains. **P* < 0.05, ***P* < 0.01, ****P* < 0.001.

mvaDANS Analysis. For identification of *daf-22*-upregulated metabolites via mvaDANS, we acquired dqfCOSY spectra for metabolome extracts obtained from two different *daf-22* alleles (*m130*, *ok693*) and wild-type *C. elegans* (Figure 2, Supplementary Figures S1 and S2). dqfCOSY spectra from four independent biological replicates were processed in magnitude mode using the mvaDANS algorithm in Matlab, which incorporated dynamic binning and multivariate pattern recognition analysis (Supplementary Figure S3). Automatic cross peak identification and binning converted each of the dqfCOSY spectra into a set of crosspeak integrals, which were corrected for dilution using Probabilistic Quotient Normalization,²⁹ mean-centered, and scaled to unit variance. This matrix of bin integrals was then subjected to statistical analysis including two-way hierarchical clustering, Principal Component Analysis (PCA) (Figure 2a)¹⁹ and Partial Least Squares-Discriminant Analysis (PLS-DA) (Supplementary Figure S7).³⁰ Coefficients from PCA loadings and PLS-DA predictors were back-projected onto the dqfCOSY spectra to enable the detection of NMR spectroscopic signals representing compounds differentially expressed in the three *C. elegans* strains (Supplementary Figures S4, S8).

mvaDANS Automation and Spectral Interpretation. Please see Supporting Information.

■ ASSOCIATED CONTENT

● Supporting Information

This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: schroeder@cornell.edu.

Author Contributions

[†]These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by the National Institutes of Health (GM088290 to F.C.S. and T32GM008500 to Y.I.), and the Vassar College Mary Landon Sague Fellowship (to Y.I.). We thank M. Kukula (BTI) for help with high-resolution MS.

■ REFERENCES

- (1) Fielenbach, N., and Antebi, A. (2008) *C. elegans* dauer formation and the molecular basis of plasticity. *Genes Dev.* 22, 2149–2165.
- (2) Kenyon, C. J. (2010) The genetics of ageing. *Nature* 464, 504–512.
- (3) Jones, K. T., and Ashrafi, K. (2009) *Caenorhabditis elegans* as an emerging model for studying the basic biology of obesity. *Dis. Models & Mech.* 2, 224–229.
- (4) von Reuss, S. H., Bose, N., Srinivasan, J., Yim, J. J., Judkins, J. C., Sternberg, P. W., and Schroeder, F. C. (2012) Comparative metabolomics reveals biogenesis of ascarosides, a modular library of small molecule signals in *C. elegans*. *J. Am. Chem. Soc.* 134, 1817–1824.
- (5) Butcher, R. A., Ragains, J. R., Li, W., Ruvkun, G., Clardy, J., and Mak, H. Y. (2009) Biosynthesis of the *Caenorhabditis elegans* dauer pheromone. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1875–1879.
- (6) Pungaliya, C., Srinivasan, J., Fox, B. W., Malik, R. U., Ludewig, A. H., Sternberg, P. W., and Schroeder, F. C. (2009) A shortcut to identifying small molecule signals that regulate behavior and development in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7708–7713.
- (7) Butcher, R. A., Fujita, M., Schroeder, F. C., and Clardy, J. (2007) Small-molecule pheromones that control dauer development in *Caenorhabditis elegans*. *Nat. Chem. Biol.* 3, 420–422.
- (8) Srinivasan, J., Kaplan, F., Ajredini, R., Zachariah, C., Albhorn, H. T., Teal, P. E., Malik, R. U., Edison, A. S., Sternberg, P. W., and Schroeder, F. C. (2008) A blend of small molecules regulates both mating and development in *Caenorhabditis elegans*. *Nature* 454, 1115–1118.
- (9) Srinivasan, J., von Reuss, S. H., Bose, N., Zaslawer, A., Mahanti, P., Ho, M. C., O'Doherty, O. G., Edison, A. S., Sternberg, P. W., and Schroeder, F. C. (2012) A modular library of small molecule signals regulates social behaviors in *Caenorhabditis elegans*. *PLoS Biol.* 10, e1001237.
- (10) Macosko, E. Z., Pokala, N., Feinberg, E. H., Chalasani, S. H., Butcher, R. A., Clardy, J., and Bargmann, C. I. (2009) A hub-and-spoke circuit drives pheromone attraction and social behaviour in *C. elegans*. *Nature* 458, 1171–1175.
- (11) Robinette, S. L., Bruschweiler, R., Schroeder, F. C., and Edison, A. S. (2011) NMR in metabolomics and natural products research: Two sides of the same coin. *Acc. Chem. Res.* 45, 288–297.
- (12) Hedenstrom, M., Wiklund-Lindstrom, S., Oman, T., Lu, F., Gerber, L., Schatz, P., Sundberg, B., and Ralph, J. (2009) Identification of lignin and polysaccharide modifications in populus wood by chemometric analysis of 2D NMR spectra from dissolved cell walls. *J. Mol. Plant* 2, 933–942.
- (13) Hedenstrom, M., Wiklund, S., Sundberg, B., and Edlund, U. (2008) Visualization and interpretation of OPLS models based on 2D NMR data. *Chemom. Intell. Lab. Syst.* 92, 110–117.
- (14) Robinette, S. L., Ajredini, R., Rasheed, H., Zeinomar, A., Schroeder, F. C., Dossey, A. T., and Edison, A. S. (2011) Hierarchical alignment and full resolution pattern recognition of 2D NMR Spectra: Application to nematode chemical ecology. *Anal. Chem.* 83, 1649–1657.
- (15) Robinette, S. L., Zhang, F., Bruschweiler-Li, L., and Bruschweiler, R. (2008) Web server based complex mixture analysis by NMR. *Anal. Chem.* 80, 3606–3611.
- (16) Zhang, F., Dossey, A. T., Zachariah, C., Edison, A. S., and Bruschweiler, R. (2007) Strategy for automated analysis of dynamic metabolic mixtures by NMR. application to an insect venom. *Anal. Chem.* 79, 7748–7752.
- (17) Zhang, F., and Bruschweiler, R. (2007) Robust deconvolution of complex mixtures by covariance TOCSY spectroscopy. *Angew. Chem., Int. Ed.* 46, 2639–2642.
- (18) Zhang, F., Bruschweiler-Li, L., Robinette, S. L., and Bruschweiler, R. (2008) Self-consistent metabolic mixture analysis by heteronuclear NMR. Application to a human cancer cell line. *Anal. Chem.* 80, 7549–7553.
- (19) Wold, S., Esbensen, K., and Geladi, P. (1987) Principal component analysis. *Chemom. Intell. Lab. Syst.* 2, 37–52.
- (20) Lucanic, M., Held, J. M., Vantipalli, M. C., Kiang, I. M., Graham, J. B., Gibson, B. W., Lithgow, G. J., and Gill, M. S. (2010) N-Acylethanolamine signalling mediates the effect of diet on lifespan in *Caenorhabditis elegans*. *Nature* 473, 226–229.
- (21) Noguez, J. H., Conner, E. S., Zhou, Y., Ciche, T. A., Ragains, J. R., and Butcher, R. A. (2012) A novel ascaroside controls the parasitic life cycle of the entomopathogenic nematode *Heterorhabditis bacteriophora*. *ACS Chem. Biol.* 7, 961–966.
- (22) Wanders, R. J. A., and Waterham, H. R. (2006) Biochemistry of mammalian peroxisomes revisited. *Annu. Rev. Biochem.* 75, 295–332.
- (23) Wanders, R. J. A. (2004) Peroxisomes, lipid metabolism, and peroxisomal disorders. *Mol. Genet. Metab.* 83, 16–27.

- (24) Wishart, D. S. (2007) Current progress in computational metabolomics. *Brief. Bioinform.* 8, 279–293.
- (25) Nicholson, J. K., Connelly, J., Lindon, J. C., and Holmes, E. (2002) Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discovery* 1, 153–161.
- (26) Gartland, K. P., Beddell, C. R., Lindon, J. C., and Nicholson, J. K. (1991) Application of pattern recognition methods to the analysis and classification of toxicological data derived from proton nuclear magnetic resonance spectroscopy of urine. *Mol. Pharmacol.* 39, 629–642.
- (27) Van, Q. N., Issaq, H. J., Jiang, Q., Li, Q., Muschik, G. M., Waybright, T. J., Lou, H., Dean, M., Uitto, J., and Veenstra, T. D. (2008) Comparison of 1D and 2D NMR spectroscopy for metabolic profiling. *J. Proteome Res.* 7, 630–639.
- (28) Chylla, R. A., Hu, K., Ellinger, J. J., and Markley, J. L. (2011) Deconvolution of two-dimensional NMR spectra by fast maximum likelihood reconstruction: application to quantitative metabolomics. *Anal. Chem.* 83, 4871–4880.
- (29) Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ¹H NMR metabonomics. *Anal. Chem.* 78, 4281–4290.
- (30) de Jong, S. (1993) SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* 18, 251–263.

Appendix B: Paper II

□□□□□□□□ □□□□□□□□ □□□ □□□□ □□□□□ □□□□ □Robinette SL, Ajredini R, Rasheed H, Zeinomar A, Schroeder FC, Dossey AT, Edison AS. Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology. *Anal Chem*, 2011. 83(5):1649-1657.□□ □□□□□□□□ □□□□□□□□ □□□□□□□□ □□□□ □□□□ □□□□ □□□□□□□□□□□□

Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology

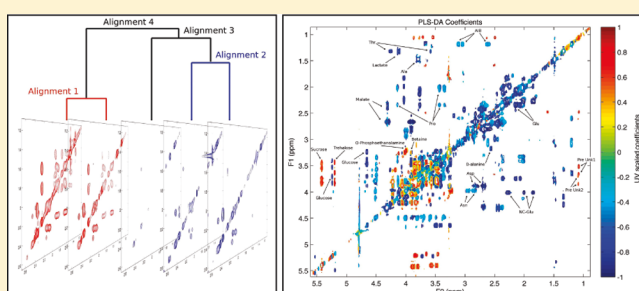
Steven L. Robinette,^{†,⊥} Ramadan Ajredini,[†] Hasan Rasheed,[†] Abdulrahman Zeinomar,[†] Frank C. Schroeder,[§] Aaron T. Dossey,^{†,||} and Arthur S. Edison^{*,†,‡}

[†]Department of Biochemistry and Molecular Biology and [‡]National High Magnetic Field Laboratory, University of Florida, Gainesville, Florida 32610-0245, United States

[§]Boyce Thompson Institute, Cornell University, Ithaca, New York 14853, United States

 Supporting Information

ABSTRACT: Nuclear magnetic resonance (NMR) is the most widely used nondestructive technique in analytical chemistry. In recent years, it has been applied to metabolic profiling due to its high reproducibility, capacity for relative and absolute quantification, atomic resolution, and ability to detect a broad range of compounds in an untargeted manner. While one-dimensional (1D) ¹H NMR experiments are popular in metabolic profiling due to their simplicity and fast acquisition times, two-dimensional (2D) NMR spectra offer increased spectral resolution as well as atomic correlations, which aid in the assignment of known small molecules and the structural elucidation of novel compounds. Given the small number of statistical analysis methods for 2D NMR spectra, we developed a new approach for the analysis, information recovery, and display of 2D NMR spectral data. We present a native 2D peak alignment algorithm we term HATS, for hierarchical alignment of two-dimensional spectra, enabling pattern recognition (PR) using full-resolution spectra. Principle component analysis (PCA) and partial least squares (PLS) regression of full resolution total correlation spectroscopy (TOCSY) spectra greatly aid the assignment and interpretation of statistical pattern recognition results by producing back-scaled loading plots that look like traditional TOCSY spectra but incorporate qualitative and quantitative biological information of the resonances. The HATS-PR methodology is demonstrated here using multiple 2D TOCSY spectra of the exudates from two nematode species: *Pristionchus pacificus* and *Panagrellus redivivus*. We show the utility of this integrated approach with the rapid, semiautomated assignment of small molecules differentiating the two species and the identification of spectral regions suggesting the presence of species-specific compounds. These results demonstrate that the combination of 2D NMR spectra with full-resolution statistical analysis provides a platform for chemical and biological studies in cellular biochemistry, metabolomics, and chemical ecology.



Nuclear magnetic resonance (NMR) is a powerful and almost universal detector due to its ability to analyze essentially all types of molecules at atomic resolution. Recently, it has been applied extensively to metabolic profiling in studies of human and animal biofluids,^{1,2} cellular biochemistry,^{3,4} and nonmammalian chemical ecology.^{5–8} While one-dimensional (1D) ¹H NMR spectra have been used in the majority of NMR-based metabolic profiling studies, resonance overlap and lack of structural correlations can limit the utility of the 1D approach. In contrast, powerful multidimensional NMR methods are routinely used in structural biology and natural products studies⁹ but have been used much less frequently for metabolomics. Two dimensional (2D) homonuclear (such as total correlation spectroscopy (TOCSY) and correlation spectroscopy (COSY)) and heteronuclear (such as heteronuclear single quantum coherence (HSQC) and heteronuclear multiple-bond correlation (HMBC)) spectra offer increased dispersion of signals in two dimensions that can

overcome the problem of signal overlap, which often limits the use of 1D ¹H NMR spectra. Additionally, the information provided by these 2D experiments through specific atomic correlations can assist in the identification and assignment of known molecules and the structural elucidation of unknown small molecules that may be significant to the underlying biology. The presence of novel, uncharacterized small molecules in complex biological mixtures is a frequent occurrence, especially in nonmammalian metabolomics where a priori knowledge of the metabolome is often incomplete. Even when complex biological mixtures are composed mainly of known metabolites, the assignment of a large number of spectral peaks is a significant complication for metabolic profiling. The use of 2D correlation

Received: October 14, 2010

Accepted: January 19, 2011

Published: February 11, 2011

spectra such as TOCSY simplify the use of semiautomated assignment platforms such as MetaboMiner¹⁰ and COLMAR.^{11–13}

While it is clear that metabolic profiling would benefit from more extensive use of 2D NMR spectra, preprocessing methods such as signal alignment and visualization tools for multiple 2D spectra have lagged behind similar methods for 1D spectra. The importance of signal alignment for statistical analysis of spectral data sets is widely appreciated by the metabolomics community for both NMR^{14,15} and mass spectrometry.^{16,17} The development of 1D alignment methods has allowed these spectra to be analyzed with increasing accuracy and resolution. However, the relative paucity of 2D spectral alignment¹⁸ and analysis methods has meant that metabolomics studies making use of 2D spectra have relied on low-resolution bins or crosspeak integrals to avoid the problem of peak shift.^{19–21} While often quite successful in mitigating the effects of chemical shift variation, the use of integrals or bins for pattern recognition causes significant loss of valuable information from spectral resolution, complicating subsequent model interpretation and assignment of relevant metabolites. Here, we present a method for hierarchical alignment of sets of 2D NMR spectra, thereby enabling pattern recognition using full-resolution data; we call this overall approach HATS-PR for hierarchical alignment of two-dimensional spectra-pattern recognition. We demonstrate that statistical representations of full-resolution spectra such as scaled principle component analysis (PCA) loadings enhance the interpretation of metabolic variation in the data set and assist in the identification and assignment of statistically relevant metabolites.

We have applied HATS-PR to an experimental comparison of the exudates (i.e., exometabolomes) from two nematode species, *Pristionchus pacificus* and *Panagrellus redivivus*. Nematode chemical ecology, especially applied to the model organism *Caenorhabditis elegans*, has become the subject of a number of recent studies in development,^{22–25} aggregation behavior,²⁶ mating behavior,^{27,28} aging,²⁹ chemical ecology,^{7,26} and functional genomics.^{6,29,30,31} The discovery and structural elucidation of a class of pheromones called ascarosides has expanded the domain of *C. elegans* research well into chemical ecology.³² In a combination of chemical ecology and genetics, Pungaliya et al. used comparative 2D NMR spectral analysis to discover new ascarosides and their biological roles in *C. elegans*.²⁷ By applying standard statistical pattern recognition methods to 2D NMR spectra, we demonstrate the utility of HATS-PR to distinguish and characterize chemical differences between two different nematode species.

EXPERIMENTAL SECTION

Collection of *P. pacificus* and *P. redivivus* Exudates. *P. pacificus* and *P. redivivus* strains were obtained from the Sternberg laboratory at Caltech.

P. pacificus. Ten nematode growth medium (NGM) plates seeded with *Escherichia coli* (OP50 strain) were inoculated with worms and incubated at room temperature for 3 days. Worms were washed off the plates with 25 mL of S-complete medium and transferred to a 250 mL Erlenmeyer flask containing 3% (0.75 g) *E. coli* (HB101 strain). Worms were incubated at 22 °C for 3 days and then transferred to a 2 L Erlenmeyer flask containing 250 mL of S-complete medium and 3% (7.5 g) HB101. Once the majority of worms were adults with eggs, sucrose flotation was used to remove bacteria and other life-stage worms. Adults with eggs were then incubated for 48 h at 22 °C at

250 rpm without food to allow the adult worms to lay the eggs and the eggs to hatch. J2 *P. pacificus* were separated from adult worms using a 20 μ m nylon filter (Sefar Inc.) to obtain a semisynchronous culture. Semisynchronized J2 *P. pacificus* (2.5 million) were grown to adult stage on S-complete medium supplemented with 3% HB101 and incubated in a shaker at 250 rpm and 22 °C. Worms from this culture were then put onto a sucrose gradient to remove bacteria. Clean *P. pacificus* were placed in M9 buffer for 30 min to clear their gut. Worms were washed three times with double-distilled water (ddH₂O) and then incubated for 1 h in ddH₂O with a density of $\sim$ 30 000 worms/mL to collect worm exudates. Worm exudates were filtered with a 0.2 μ m nylon syringe filter and stored at -80 °C for further analysis.

P. redivivus. *P. redivivus* were cultivated similarly to *P. pacificus* and transferred to liquid culture. A mixed population of *P. redivivus* with worm density of 10 000 worms/mL was incubated in 250 mL of S-complete medium supplemented with 3% HB101. *P. redivivus* were cultured for 12 days, and 3% HB101 was added every third day. After 12 days of incubation, the worms were separated from bacteria via sucrose flotation. The *P. pacificus* protocol described above was also used to collect *P. redivivus* exudates.

Sample Fractionation. *P. pacificus* and *P. redivivus* exudates were each fractionated by C18 solid phase extraction using a Varian Mega BE C18, which has a 5 mL void volume. A 50 mL syringe was attached to the column using a rubber adapter. Teflon tape was placed around the rubber stopper so that rubber pieces would not contaminate the solution. Approximately constant pressure was applied to the column by gently pressing the syringe. Each C18 column was first washed with 50 mL of 90% methanol, followed by 50 mL of ddH₂O for equilibration. *P. pacificus* and *P. redivivus* exudates were then added to individual columns, and the flow-through was collected. The columns were washed with 25 mL of ddH₂O and then eluted with 25 mL of 50% methanol and 25 mL of 90% methanol. All samples were collected in glass vials with Teflon caps and stored at -80 °C.

NMR Spectroscopy. The flow-through fractions were lyophilized, dissolved in 500 μ L of 99.9% D₂O (Cambridge Isotope Laboratories), and put into 5 mm NMR tubes. The 50% MeOH fractions were lyophilized, dissolved in 110 μ L of 99.95% methanol-*d*₄ (Cambridge Isotope Laboratories), and put into 2.5 mm NMR tubes. 1D ¹H and 2D TOCSY spectra were collected on a Bruker Avance II spectrometer operating at 600.23 MHz using a 5 mm TXI cryoprobe. NMR data were collected with spectral widths of 11 ppm at a sample temperature of about 300 K. 1D ¹H spectra were collected using a simple pulse-acquire sequence with presaturation of residual water. TOCSY spectra were collected with the DIPSI-2 mixing sequence³³ with water presaturation and 60 or 90 ms mixing times for the 50% MeOH or flow-through fractions, respectively. TOCSY data were collected with 2048 and 256 complex data points in the direct and indirect dimensions, respectively. TOCSY data were zero filled to 2048 F2 and 1024 F1 data points, Fourier transformed, phased, and baseline corrected in Topspin 2.0 (Bruker) before being transferred to Matlab 2009b for alignment and pattern recognition analysis. As an additional application to a more complex mixture, TOCSY spectra were acquired from human urine (details provided in Supporting Information), and data were processed similarly to the worm exudate spectra.

Hierarchical Alignment of Two-Dimensional Spectra. The HATS alignment method presented here makes use of a guide tree to structure the alignment process. While many existing 1D

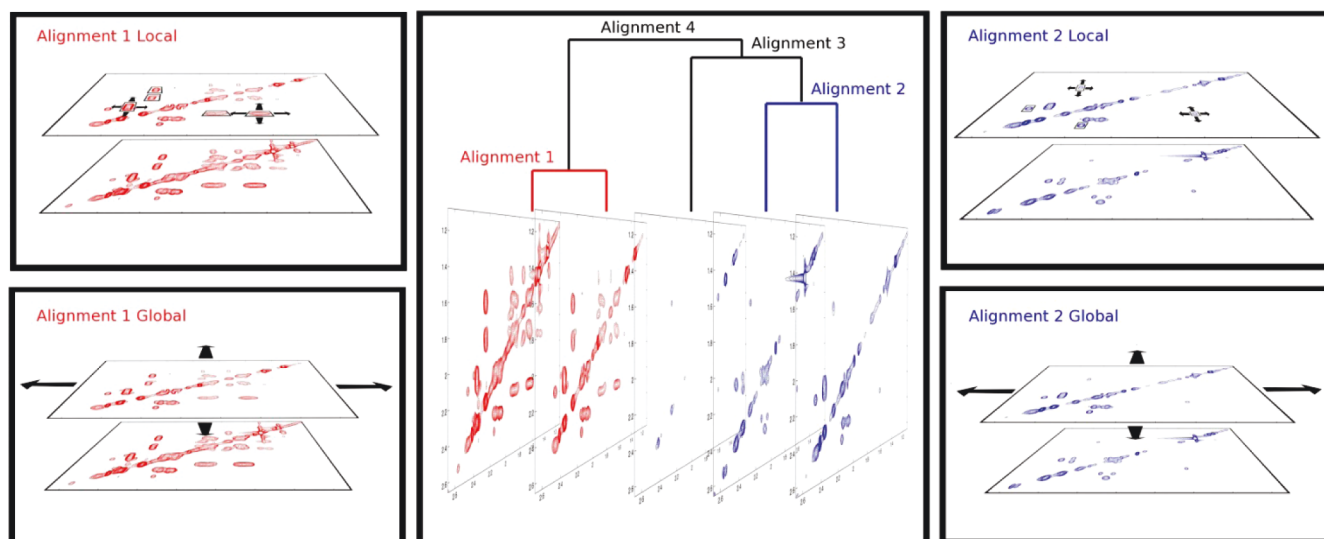


Figure 1. Overall strategy of hierarchical alignment of two-dimensional spectra (HATS). Before alignment, the HATS algorithm produces a guide tree by the UPGMA agglomerative hierarchical clustering algorithm from a matrix of pairwise correlation distances. This guide tree is used to structure the alignment process. At each intersection of branches in the guide tree, the spectra specified by the two branches are aligned globally and locally. For the alignment tree shown here, the two red spectra are aligned first, followed by the two blue spectra. The two blue spectra are then aligned with the black spectrum, and finally, the three blue and black spectra are aligned with the red spectra.

alignment methods have attempted to align each spectrum in a data set to a single consensus spectrum, the quality of these “star alignments” can suffer if there is wide chemical diversity between spectra, as any individual spectrum chosen as a star may not include the full complement of metabolites present in the other data sets. The use of guide trees was a significant advance for multiple sequence alignment in sequence bioinformatics, and methods such as CLUSTAL-W³⁴ have enjoyed widespread popularity in nucleic acid and protein sequence alignment. We suggest that the use of a guide tree will be especially important in the alignment of 2D spectra as correlation spectra are often collected when significant metabolic diversity is expected and the assumptions underlying star alignments may not be appropriate.

The UPGMA agglomerative hierarchical clustering method³⁵ was used to construct the initial guide tree for the subsequent pairwise alignment of spectral groups. We have used the squared distance of Pearson’s correlation coefficient as a metric for clustering (eqs 1 and 2), and the resulting dendrogram specifies the order in which alignment of the spectra proceeds.

$$r_{mn} = \frac{\sum_{i=1}^j (S_{im} - \bar{S}_m)(S_{in} - \bar{S}_n)}{\sigma_{S_m} \sigma_{S_n}} \quad (1)$$

$$d_{mn} = 1 - r_{mn}^2 \quad (2)$$

Here, r_{mn} is the correlation coefficient between the reshaped vectors ($1 \times F1 \cdot F2$) of spectra S_m and S_n with mean values $\bar{S}_m$ and $\bar{S}_n$ and standard deviations σ_{S_m} and σ_{S_n} and d_{mn} is a distance version of the correlation value. At each branch point in the dendrogram, the set of spectra specified by the branches are aligned. For the alignment illustrated in Figure 1, the two red spectra are aligned first, followed by the two blue spectra; then, the two blue spectra are aligned with the black spectrum, and finally, the two red spectra are aligned with the three blue and black spectra. If there are multiple spectra in an alignment group, positional adjustments are made equally to all spectra in the group.

The alignment at each branch point of the guide tree consists of two parts: a global alignment in which all spectral data points are shifted equally and a local alignment in which specific crosspeak regions are shifted individually. The global alignment step seeks to correct defects such as miscalibration or referencing, which produce chemical shift differences across the entire spectrum. One of the groups of spectra is passed over the other group as a mask in a point-wise manner in both F1 and F2 dimensions, and at each positional step x and y , the alignment quality is expressed as r_{xy} , the mean correlation between groups (eq 3). Here, a is the number of spectra in group 1, and b is the number in group 2, and r_{mn} is the correlation coefficient between the reshaped vectors of spectrum m in group 1 and spectrum n in group 2. Spectra in the mask group are shifted to the points m and n producing the maximal mean correlation which is identified by a simple gradient ascent algorithm.

$$r_{xy} = \frac{1}{a \cdot b} \sum_{m=1}^a \sum_{n=1}^b r_{mn} \quad (3)$$

While global alignment can resolve factors affecting all resonances equally, it is not capable of resolving differences in specific resonances caused by conditions such as pH, temperature, metal ion concentration, and osmolality. To resolve local chemical shift variation, crosspeak regions must be aligned individually. To identify segments of the spectrum corresponding to crosspeak regions, a local noise surface is calculated for each spectrum.³⁶ Initial spectral segments are identified as regions within 0.04 ppm of a peak, defined as a local maximum with intensity at least $10 \times$ the noise value at that point. Each segment is then labeled with a unique integer value, and these initial segments are expanded by iterated minimum and maximum filtering. Min/max filtering has the effect of growing the segments until they encounter another spectral segment or until they reach a maximum frequency range of approximately 0.2 ppm (Figure 1S, Supporting Information). This procedure creates spectral segments that are bounded by

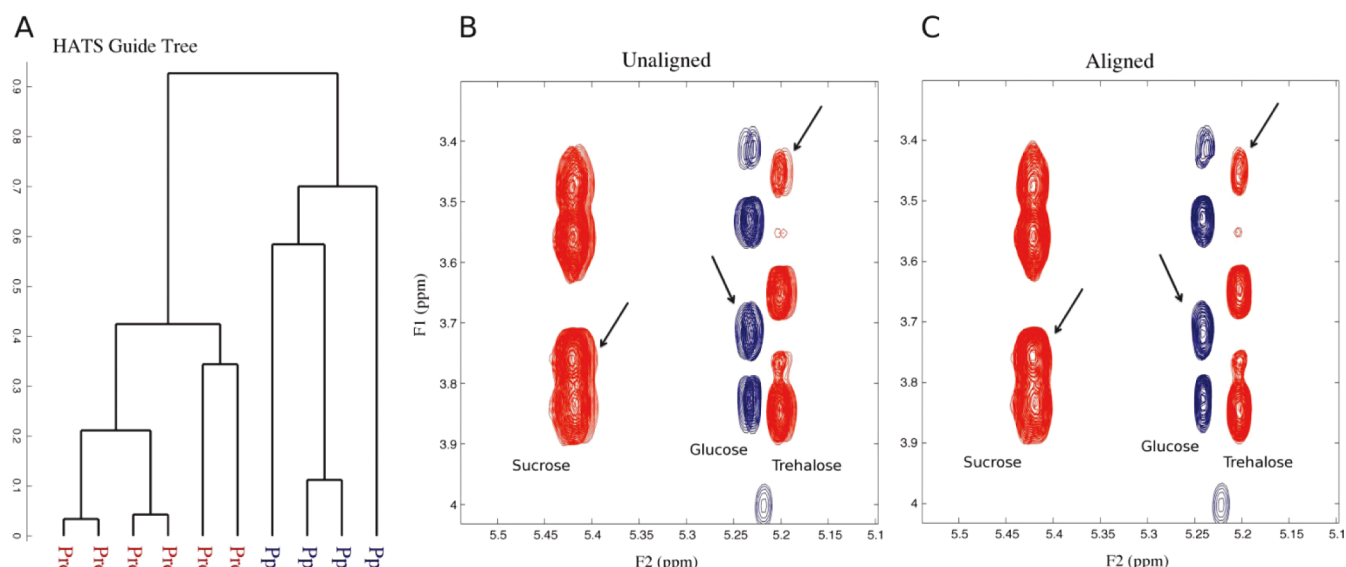


Figure 2. Alignment results for *P. redivivus* (Pre) and *P. pacificus* (Ppa) TOCSY spectra of flow-through fractions from C18 solid phase extractions. (A) The guide tree produced by HATS groups spectra from the same species, resulting in intraspecies alignment before interspecies alignment. The utility of a guide tree-based alignment is exemplified by the anomeric crosspeaks of glucose, sucrose, and trehalose in the TOCSY spectra (B, C). By aligning spectra with similar composition first, false positive alignments of nearby but unrelated crosspeaks, such as the anomeric crosspeaks shown above, are avoided.

nearby resonances yet avoid close cropping of crosspeak line-shapes. Each crosspeak segment is then aligned individually by maximizing the mean cross-correlation of crosspeak segments between alignment groups in the same way as the global alignment but treating each crosspeak segment independently (eq 3, where r_{xy} is the line shape vector for a single crosspeak region rather than the entire spectrum).

Full Resolution TOCSY Pattern Recognition. Following alignment, HATS-PR makes use of principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) for full resolution spectral pattern recognition. PCA is an unsupervised dimensionality reduction technique that is widely used in metabolomics. PCA transforms a data set of N observations with P features into a set of principal components that are orthogonal linear combinations of features accounting for as much of the variation in the original data set as possible. Each principal component consists of N scores and P loadings, where scores identify relationships between observations (here, TOCSY spectra) and loadings represent the pattern of features (here, 2D NMR chemical shifts) underlying the relationships identified by scores. We used the NIPALS algorithm³⁷ to identify the first five principal components of the mean-centered and univariate scaled flow-through ($N = 10$) and 50% MeOH ($N = 8$) TOCSY spectra. While PCA avoids the problem of overfitting by taking an unsupervised approach to pattern recognition, it is not optimized for class separation. Partial least-squares methods such as PLS-DA are often the supervised method of choice for the analysis of metabolic differentials. We apply SIMPLS³⁸ PLS-DA to the data set to help clarify the metabolic differential identified by PCA while avoiding overfitting by the use of 4-fold cross-validation and reporting performance statistics for the PLS-DA model.

Pattern recognition of NMR spectra was done in Matlab 2009b following Probabilistic Quotient Normalization³⁹ to account for random dilution effects and alignment using the HATS method for the TOCSY spectra and the RSPA¹⁵ algorithm for

the 1D ^1H spectra. Each of the $N \ 2048 \times 1024$ TOCSY data matrices was reshaped to a $1 \times 2\ 097\ 152$ vector before the application of PCA or PLS. This is a necessary step because it creates an $N \times P$ matrix, where $P = F1 \cdot F2$ data points on which pattern recognition algorithms can operate. To visualize the resulting $1 \times P$ loadings, the loadings vectors for each principal component were reshaped back into a 2048×1024 data matrix. Contours were defined by multiplying the univariate scaled loadings by the standard deviation along the third dimension of the original TOCSY stack, a procedure known as back-scaling. Contour colors were then defined by the loading values themselves divided by the maximum loading coefficient to produce a range of -1 to $+1$. This strategy produces loading spectra which have crosspeak lineshapes similar to traditional TOCSY spectra but include sign and intensity information related to the relationships between spectra specified by the corresponding PCA scores. PLS-DA results are presented in a similar manner, but instead of specifying scores and loadings, only the predictive feature vector is shown, along with performance statistics for the model. This 2D NMR unfolding and refolding approach has recently been applied successfully to HSQC spectra and provides a simple methodology to apply statistical pattern recognition approaches to 2D NMR spectra.^{40,41}

Semi-Automated Assignment of TOCSY Loadings. To assign the compounds represented in our back-scaled TOCSY loadings plots, we applied COLMAR query¹¹ to correlated chemical shifts present in the loadings spectra. When COLMAR query identified a match to a compound reference spectrum present in the BMRB,^{42,43} HMDB,⁴⁴ or MMCD⁴⁵ databases, the reference TOCSY spectrum for the suggested compound was downloaded from the database web servers, processed in Topspin 2.0, and overlaid onto an experimental spectrum for confirmation. We considered the compound to be assigned if all reference chemical shifts and experimental chemical shifts differ by less than 0.02 ppm and all through-bond correlations (TOCSY crosspeaks) in the

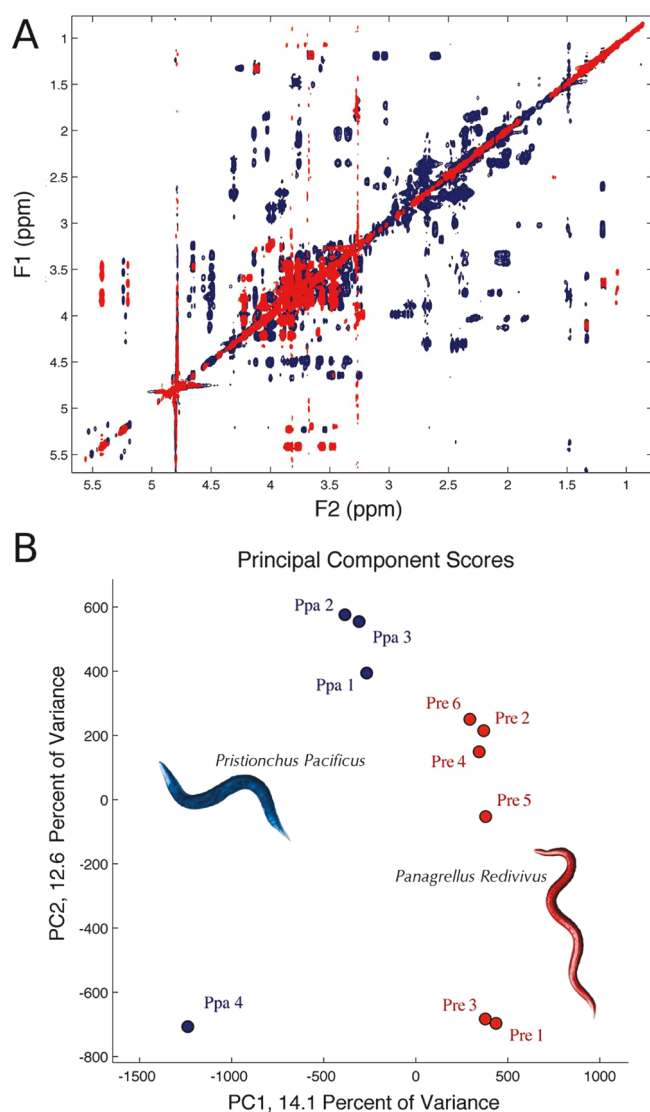


Figure 3. (A) Overlay of *P. redivivus* (Pre) and *P. pacificus* (Ppa) flow-through TOCSY spectra and (B) scores plot from PCA of the flow-through TOCSY data set. *P. redivivus* and *P. pacificus* spectra are clearly differentiated by PC1, which allows the loadings of PC1 to be interpreted as relative quantitative spectral differences between the species.

experimental spin-system were accounted for by the reference spectrum. It should be emphasized that for this study we did not verify assignments by spiking natural samples with authentic commercial or synthetic standards, a common practice for final verification of compound identity by NMR.

RESULTS

Previous work from our group and others has shown that *C. elegans* releases a large number of ascarosides that have been shown to have multiple functions.^{22–25,27,28,32} These ascarosides typically partition into the 50% MeOH fraction of a C18 solid phase extraction. We have also shown that *C. elegans* releases an abundance of small polar molecules including amino acids, sugars, and organic acids, many of which elute from a C18 column in the hydrophilic flow-through fraction.⁷ Here, we analyzed 2D TOCSY NMR data of the flow-through and 50% MeOH C18 elution fractions from *P. redivivus* and *P. pacificus* exudates in order to demonstrate our new methods and to show the utility of

this approach in metabolomics and natural product comparative studies. In order to demonstrate the general applicability of our approach, we have also applied it to the alignment of 2D TOCSY spectra from human urine (Figure 1S, Supporting Information).

Alignment of TOCSY Spectra. Despite the high reproducibility of NMR spectra, resonance variation, sometimes referred to as positional noise, resulting from slight differences in sample conditions hinder comparative studies and complicate the interpretation of statistical analyses such as univariate regression and significance testing, principal component analysis (PCA), and partial least-squares (PLS). Previous research with 1D ¹H NMR spectra has shown that peak shift can result in both false positives and negatives when attempting to identify metabolite biomarkers.⁴⁶ While binning is often used to reduce positional variation, loss of high-resolution information complicates metabolite assignment and limits the added value of 2D experiments. The HATS alignment methodology presented here compensates for the effects of resonance variation, thereby increasing the accuracy and interpretability of comparative 2D NMR analyses.

The HATS-PR approach to eliminating positional variation is to first cluster all of the spectra through a guide tree, as shown in Figure 2A. The problem of positional variation in our nematode TOCSY spectra is easily seen in the expansion of the anomeric region of the sugars from the flow-through fractions (Figure 2B, C). In the unaligned data (Figure 2B), local differences in resonance frequency of the anomeric protons affect multiple spectra of both *P. redivivus* and *P. pacificus* samples and result in a general “blurriness” that significantly impacts subsequent statistical analyses. Removing this variation is clearly desirable, but the close proximity (~ 0.03 ppm) and similar patterns and lineshapes of the *P. pacificus* glucose and *P. redivivus* trehalose crosspeaks raise the possibility of inappropriate alignment of the glucose to the trehalose crosspeaks. The use of the HATS-PR guide tree (Figure 2A) helps avoid this possibility by aligning the most similar before the less similar spectra. Using this approach, the glucose and trehalose resonances are not compared for alignment until they have been well aligned separately. As shown in Figure 2C, this results in the correction of positional variation within both glucose and trehalose crosspeaks and rejection of the inappropriate alignment of the two sugars.

***P. redivivus* and *P. pacificus* Release Different Polar Molecules.** Next, using the aligned TOCSY spectra (Figure 3A) of C18 flow-through fractions, we conducted PCA and PLS analysis. The PCA scores show clear, unambiguous separation of the *P. redivivus* spectra from the *P. pacificus* spectra in the first principal component (Figure 3B). This separation allows the PC1 loadings to be interpreted as a pattern of TOCSY crosspeaks differentiating the two species (Figure 4). The back-scaling procedure outlined above produces contours reflecting the crosspeak lineshapes in the original spectra but colored according to the loading coefficients. As the *P. redivivus* and *P. pacificus* PC1 scores have opposite signs, we represent features with greater intensity in the *P. redivivus* spectra in red (loadings between 0 and 1) and with greater *P. pacificus* intensity in blue (loadings between -1 and 0). The PLS predictors also agree with the PC1 loadings (Figure 2S, Supporting Information). The assignment of the crosspeaks in the PC1 loadings spectrum enables the identification of compounds unique to each species as well as quantitative differences between compound levels.

To interpret this spectral information in a chemical context, it is necessary to assign NMR resonances to their molecules (Table 1S, Supporting Information). While assignment has traditionally

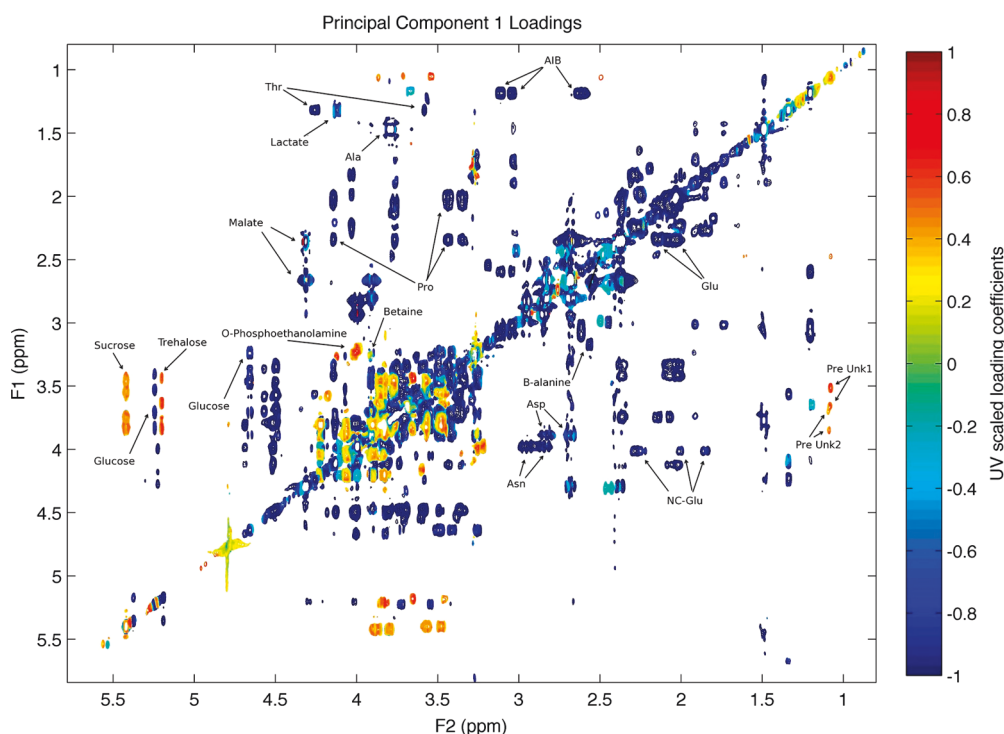


Figure 4. Spectral back-projection of PC1 loadings. Here, contours are defined by the back-projected intensities, while colors are defined by the unit variance-scaled loading coefficients. Crosspeaks with positive loading coefficients (red) represent compounds overexpressed by *P. redivivus* relative to *P. pacificus*. Blue crosspeaks represent negative loading coefficients indicating *P. pacificus* overexpressed compounds. Assignment of crosspeaks in the back-projected loadings places the chemical information contained in the TOCSY spectra in an immediate biological context. Abbreviations: AIB, 3-aminoisobutyrate; Ala, alanine; Asn, asparagine; Asp, aspartate; Glu, glutamate; NC-Glu, N-carbamylglutamate; Pre Unk1, *P. redivivus* unknown 1; Pre Unk2, *P. redivivus* unknown 2; Pro, L-proline; Thr, threonine.

been one of the more difficult steps in metabolomic analysis due to the large chemical complexity in most crude biological matrices, the establishment of reference spectra databases for common metabolites such as the BMRB,^{41–43} HMDB,⁴⁴ and MMCD⁴⁵ has facilitated the development of database searching algorithms to match sets of correlated chemical shifts to a compound ID. For example, PRIME SpinAssign,^{47,48} NMRshifDB,⁴⁹ MetaboMiner,¹⁰ and COLMAR¹² are freely available web utilities that offer automated database matching capabilities to aid the assignment process. The assignments shown in Figure 4 were generated by a combination of COLMAR query¹¹ to identify possible matches and manual overlay of reference spectra from the BMRB to confirm identifications, as described in the Experimental Section. While correlated chemical shifts were identified manually for input into COLMAR query, the back-scaled loadings matrix produced by HATS-PR should be compatible with spectral decomposition methods such as DemixC¹³ and local peak clustering.⁵⁰

***P. redivivus* and *P. pacificus* Appear to Produce Different Ascarosides.** *C. elegans* releases at least 10 different ascarosides important in the nematode's chemical ecology,³² and methods to efficiently compare other species for ascaroside-like compounds would be useful. The most diagnostic NMR signatures for ascarosides are typically resonance peaks from anomeric, methyl, and methylene protons (Figure 5) present in the eluate of C18 SPE chromatography matrix with 50% MeOH C18. As with the flow-through fraction, PC1 separates the *P. redivivus* and *P. pacificus* 50% MeOH fraction spectra (Figure 3S, Supporting Information). The PC1 loadings from the 50% MeOH fractions show numerous crosspeaks in the ascaroside signature region.

These peaks are difficult to resolve using 1D ¹H NMR, and loadings from pattern recognition on the 1D spectra are difficult to interpret (Figure 4S, Supporting Information). Interestingly, both *P. redivivus* and *P. pacificus* show signals representing ascaroside-like compounds; however, differences in chemical shift and signal intensity suggest different sets of ascarosides produced by these species. We tried database matching the 50% MeOH, as we did with the flow-through fractions, but we were unsuccessful in finding any matches. Identification of these compounds is in progress using traditional approaches and will be reported elsewhere.

DISCUSSION AND CONCLUSIONS

We have presented an efficient method to align, statistically analyze, and identify compounds in 2D NMR spectra. Although we demonstrated the approach with TOCSY spectra, all of the algorithms used for alignment, PCA, and PLS-DA are general and should work without modification for other types of 2D NMR data sets. In addition to problems with alignment, which we have addressed in this study, the increased acquisition time per sample relative to 1D experiments remains a barrier to adoption of 2D NMR methods in metabolic profiling. However, many groups are developing rapid methods for 2D data acquisition^{51–53} and designing more sensitive probes,^{54–56} making 2D NMR analysis of relatively large numbers of samples increasingly feasible.

The additional structural information provided by 2D NMR methods is often helpful for structural characterization of known metabolites and is necessary for assignment of unknown or novel compounds. Our use of COLMAR query for assignment of compounds present in multiple TOCSY spectra constitutes a

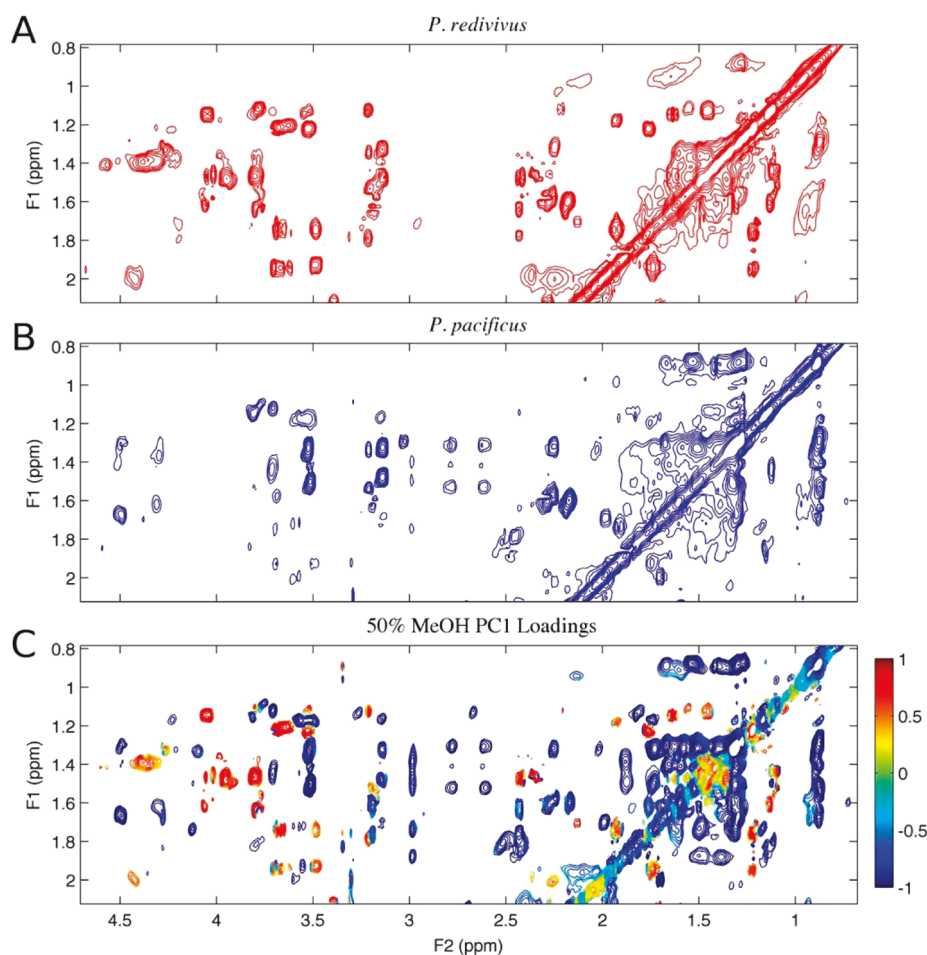


Figure 5. Representative spectra and PC1 loadings for *P. redivivus* and *P. pacificus* 50% MeOH fractions. The region selected is useful for ascaroside differentiation, as crosspeaks here indicate correlations from the methylene protons on the ascaroside side chain to protons near the terminal functional groups. PC1 loadings suggest that *P. redivivus* and *P. pacificus* produce different mixtures of ascaroside-like compounds.

proof-of-principle demonstration of how metabolomic analysis can be streamlined by integrating multi-2D statistical pattern recognition with informatics tools for compound assignment. While not applied in this study, we suggest that the application of automated spectral deconvolution methods, which isolate signals arising from single molecules, such as local clustering⁵⁰ and DemixC,¹³ to back-scaled 2D NMR loadings matrices will result in further streamlining of metabolomic data analysis. While automated assignment is currently restricted to compounds for which reference data is present in spectral databases, progress is being made on methods to identify probable chemical structures directly from chemical shift calculations.⁵⁷ The polar metabolites present in the flow-through fraction of the nematode exudate C18 SPE extractions were particularly well suited for semiautomated assignment by COLMAR query as the reference databases contain large numbers of sugars, amino acids, small organic acids, and other common biological metabolites that tend to partition in the aqueous phase.

It is common practice with 2D NMR to overlay spectra with different colors, as shown for example in Figure 3A. There are several advantages that HATS-PR has oversimple spectral overlay. First, with spectral overlay, it is very difficult or impossible to see quantitative changes in a particular compound. Using HATS-PR, quantitative information is clearly represented by changes in the color of the crosspeaks in the loadings plots. Second, using overlays is not particularly amenable to evaluating statistical

replicates and comparing multiple conditions. However, with HATS-PR, like any PCA or PLS analysis, statistical replicates enhance the analysis and are easily visualized through the corresponding scores plots and regression statistics. Additionally, quantitative covariance between 2D NMR crosspeaks may provide a means by which to identify connectivity in metabolites with multiple spin systems.

In principle, it is possible to do PCA analysis on 2D NMR data without alignment. However, alignment of the spectra aids in the interpretation of individual compounds by improving the quality of the pattern recognition analysis. Pattern recognition on unaligned 1D spectra often results in line shapes resembling first derivatives rather than Lorentzian line shapes.¹⁵ In 2D NMR spectra, misalignment often results in the duplication of peaks (Figure S5, Supporting Information) in the back-scaled loadings or attenuation of loading coefficients (Figure 6S, Supporting Information) due to peak position differences and can result in compound misidentification. Despite making small changes in the chemical shifts of the crosspeaks, alignment of crosspeak lineshapes to their maximum correlation results in more interpretable loadings for both the flow-through and 50% MeOH fraction, as shown by supplemental Figures S4 and S5, Supporting Information. HATS-PR represents, to our knowledge, the first usage of a guide tree in a spectral alignment method, which removes the requirement that all compounds be present in a

single spectrum selected as the alignment reference. We believe guide trees will also be generally useful for peak alignment in 1D ^1H NMR, chromatographic profiles, and LC-MS spectra as well.

HATS-PR is a relatively simple way to chemically discriminate two (or more) biological states, such as treated vs untreated, diseased vs normal, one species vs another, etc. Using existing database search algorithms and rapidly growing NMR metabolomic databases, it is straightforward using HATS-PR not only to compare biological states but also to identify and assign the relevant compounds that differentiate these states. In this study, we compared exudates of two nematode species. The samples were prepared under identical conditions, yet revealed surprisingly large differences, particularly in the polar flow-through fractions. In previous work, we found that *C. elegans* produces at least 40 common, mostly polar metabolites,⁷ and we expected to find a similar group of compounds in *P. redivivus* and *P. pacificus*. HATS-PR very efficiently showed us that the two species produce significantly different sets of polar small molecules, including sugars. *P. redivivus* in general releases a less diverse set of polar molecules but produces large amounts of the disaccharide trehalose and their exudates also contain a large quantity of sucrose, yet no glucose was observed. While *P. pacificus* releases a very diverse set of polar small molecules including glucose, the two disaccharides sucrose and trehalose were not detected. All of the nematodes for this study were separated from their bacterial food using a sucrose gradient, so it is possible that some or all of the sucrose detected in *P. redivivus* originates from the sample preparation. However, we find a comparable amount of trehalose in the same samples, and more importantly, we find no sucrose in *P. pacificus*, which were prepared with an identical sucrose gradient.

Despite the fact that spectra collected on the flow-through fractions of *P. redivivus* were much simpler than those from *P. pacificus*, the TOCSY spectra of 50% MeOH fractions from the same worm preparations exhibited similar levels of complexity but also some clear differences. Although, on the basis of the data presented here, we are unable to determine unambiguously whether these fractions contain ascarosides, the spectral signatures of the 50% MeOH fractions for both species suggest the presence of complex ascaroside mixtures, similar to those found in *C. elegans*. The identification and assignment of these compounds requires additional work and additional NMR spectroscopic and mass spectrometric analyses and is beyond the scope of this study. However, it is clear that HATS-PR is a powerful method to compare small molecules from one organism to another and as such provides a significant bridge between metabolomics and natural products analysis. As our use of HATS-PR for alignment of a set of urine spectra demonstrates, the method is broadly applicable and can successfully align arrays of 2D spectra even in situations where sample variation leads to significant chemical shift differences between spectra. We expect that this work will contribute significantly to informatics approaches to metabolic profiling as well as to biological and structural characterization of metabolic phenotypes in both mammalian and nonmammalian models.

■ ASSOCIATED CONTENT

S **Supporting Information.** Additional information as noted in text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: aedison@ufl.edu.

Present Addresses

[†]Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, London, UK SW7 2AZ.

[‡]United States Department of Agriculture, Agricultural Research Service.

■ ACKNOWLEDGMENT

S.L.R. acknowledges financial support from the Barry M. Goldwater Foundation and from the Howard Hughes Medical Institute Science for Life program at UF. We thank James R. Rocca, Paul Sternberg, and Andrea Choe for helpful discussions. Burt Singer and Greg Stupp kindly read the manuscript and provided us with stimulating discussions. Nematode NMR data were collected in the University of Florida AMRIS Facility, which is part of the NSF-funded National High Magnetic Field Laboratory. Human urine NMR data were collected at Imperial College, and we are grateful to Jeremy Nicholson for the use of the NMR spectrometer. This work was supported by NIH 1R01GM085285-01 and 3R01GM085285-01A1S1 to A.S.E.

■ REFERENCES

- (1) Nicholson, J. K.; Wilson, I. D. *Prog. Nucl. Magn. Reson. Spectrosc.* **1989**, *21*, 449–501.
- (2) Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. *Nat. Rev. Drug Discovery* **2002**, *1*, 153–161.
- (3) Behrends, V.; Ebbels, T. M.; Williams, H. D.; Bundy, J. G. *Appl. Environ. Microbiol.* **2009**, *75*, 2453–2463.
- (4) Schroeder, F. C.; Gibson, D. M.; Churchill, A. C.; Sojikul, P.; Wursthorn, E. J.; Krasnoff, S. B.; Clardy, J. *Angew. Chem., Int. Ed. Engl.* **2007**, *46*, 901–904.
- (5) Bundy, J. G.; Sidhu, J. K.; Rana, F.; Spurgeon, D. J.; Svendsen, C.; Wren, J. F.; Sturzenbaum, S. R.; Morgan, A. J.; Kille, P. *BMC Biol.* **2008**, *6*, 25.
- (6) Blaise, B. J.; Giacomotto, J.; Elena, B.; Dumas, M. E.; Toulhoat, P.; Segalat, L.; Emsley, L. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19808–19812.
- (7) Kaplan, F.; Badri, D. V.; Zachariah, C.; Ajredini, R.; Sandoval, F. J.; Roje, S.; Levine, L. H.; Zhang, F.; Robinette, S. L.; Alborn, H. T.; Zhao, W.; Stadler, M.; Nimalendran, R.; Dossey, A. T.; Bruschweiler, R.; Vivanco, J. M.; Edison, A. S. *J. Chem. Ecol.* **2009**, *35*, 878–892.
- (8) Bundy, J. G.; Spurgeon, D. J.; Svendsen, C.; Hankard, P. K.; Osborn, D.; Lindon, J. C.; Nicholson, J. K. *FEBS Lett.* **2002**, *521*, 115–120.
- (9) Rae, R.; Iatsenko, I.; Witte, H.; Sommer, R. J. *Environ. Microbiol.* **2010**, *12*, 3007–3021.
- (10) Xia, J.; Bjorn Dahl, T. C.; Tang, P.; Wishart, D. S. *BMC Bioinformatics* **2008**, *9*, 507.
- (11) Robinette, S. L.; Zhang, F.; Bruschweiler-Li, L.; Bruschweiler, R. *Anal. Chem.* **2008**, *80*, 3606–3611.
- (12) Zhang, F.; Robinette, S. L.; Bruschweiler-Li, L.; Bruschweiler, R. *Magn. Reson. Chem.* **2009**, *47* (Suppl 1), S118–122.
- (13) Zhang, F.; Dossey, A. T.; Zachariah, C.; Edison, A. S.; Bruschweiler, R. *Anal. Chem.* **2007**, *79*, 7748–7752.
- (14) Lee, G. C.; Woodruff, D. L. *Anal. Chim. Acta* **2004**, *513*, 413–416.
- (15) Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K. *Anal. Chem.* **2009**, *81*, 56–66.
- (16) Wong, J. W. H.; Cagney, G.; Cartwright, H. M. *Bioinf.* **2005**, *21*, 2088–2090.

- (17) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779–787.
- (18) Zheng, M.; Lu, P.; Liu, Y.; Pease, J.; Usuka, J.; Liao, G.; Peltz, G. *Bioinformatics* **2007**, *23*, 2926–2933.
- (19) Lewis, I. A.; Schommer, S. C.; Markley, J. L. *Magn. Reson. Chem.* **2009**, *47* (Suppl 1), S123–126.
- (20) Ludwig, C.; Ward, D. G.; Martin, A.; Viant, M. R.; Ismail, T.; Johnson, P. J.; Wakelam, M. J.; Gunther, U. L. *Magn. Reson. Chem.* **2009**, *47* (Suppl 1), S68–73.
- (21) Van, Q. N.; Issaq, H. J.; Jiang, Q.; Li, Q.; Muschik, G. M.; Waybright, T. J.; Lou, H.; Dean, M.; Uitto, J.; Veenstra, T. D. *J. Proteome Res.* **2008**, *7*, 630–639.
- (22) Butcher, R. A.; Ragains, J. R.; Clardy, J. *Org. Lett.* **2009**, *11*, 3100–3103.
- (23) Butcher, R. A.; Ragains, J. R.; Kim, E.; Clardy, J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14288–14292.
- (24) Butcher, R. A.; Fujita, M.; Schroeder, F. C.; Clardy, J. *Nat. Chem. Biol.* **2007**, *3*, 420–422.
- (25) Jeong, P. Y.; Jung, M.; Yim, Y. H.; Kim, H.; Park, M.; Hong, E.; Lee, W.; Kim, Y. H.; Kim, K.; Paik, Y. K. *Nature* **2005**, *433*, 541–545.
- (26) Macosko, E. Z.; Pokala, N.; Feinberg, E. H.; Chalasani, S. H.; Butcher, R. A.; Clardy, J.; Bargmann, C. I. *Nature* **2009**, *458*, 1171–1175.
- (27) Pungaliya, C.; Srinivasan, J.; Fox, B. W.; Malik, R. U.; Ludewig, A. H.; Sternberg, P. W.; Schroeder, F. C. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 7708–7713.
- (28) Srinivasan, J.; Kaplan, F.; Ajredini, R.; Zachariah, C.; Alborn, H. T.; Teal, P. E.; Malik, R. U.; Edison, A. S.; Sternberg, P. W.; Schroeder, F. C. *Nature* **2008**, *454*, 1115–1118.
- (29) Fuchs, S.; Bundy, J. G.; Davies, S. K.; Viney, J. M.; Swire, J. S.; Leroy, A. M. *BMC Biol.* **2010**, *8*, 14.
- (30) Blaise, B. J.; Giacomotto, J.; Triba, M. N.; Toulhoat, P.; Piotto, M.; Emsley, L.; Segalat, L.; Dumas, M. E.; Elena, B. J. *Proteome Res.* **2009**, *8*, 2542–2550.
- (31) Atherton, H. J.; Jones, O. A.; Malik, S.; Miska, E. A.; Griffin, J. L. *FEBS Lett.* **2008**, *582*, 1661–1666.
- (32) Edison, A. S. *Curr. Opin. Neurobiol.* **2009**, *19*, 378–388.
- (33) Shaka, A. J.; Lee, C. J.; Pines, A. J. *Magn. Reson.* **1988**, *77*, 274–293.
- (34) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. *Nucleic Acids Res.* **1994**, *22*, 4673–4680.
- (35) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*; Freeman: San Francisco, CA, 1973.
- (36) Koradi, R.; Billeter, M.; Engeli, M.; Guntert, P.; Wuthrich, K. *J. Magn. Reson.* **1998**, *135*, 288–297.
- (37) Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
- (38) Dejong, S. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251–263.
- (39) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. *Anal. Chem.* **2006**, *78*, 4281–4290.
- (40) Hedenstrom, M.; Wiklund-Lindstrom, S.; Oman, T.; Lu, F. C.; Gerber, L.; Schatz, P.; Sundberg, B.; Ralph, J. *Mol. Plant* **2009**, *2*, 933–942.
- (41) Hedenstrom, M.; Wiklund, S.; Sundberg, B.; Edlund, U. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 110–117.
- (42) Markley, J. L.; Anderson, M. E.; Cui, Q.; Eghbalnia, H. R.; Lewis, I. A.; Hegeman, A. D.; Li, J.; Schulte, C. F.; Sussman, M. R.; Westler, W. M.; Ulrich, E. L.; Zolnai, Z. *Pac. Symp. Biocomput.* **2007**, 157–168.
- (43) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L. *Nucleic Acids Res.* **2008**, *36*, D402–408.
- (44) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncz, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* **2007**, *35*, D521–526.
- (45) Cui, Q.; Lewis, I. A.; Hegeman, A. D.; Anderson, M. E.; Li, J.; Schulte, C. F.; Westler, W. M.; Eghbalnia, H. R.; Sussman, M. R.; Markley, J. L. *Nat. Biotechnol.* **2008**, *26*, 162–164.
- (46) Dumas, M. E.; Maibaum, E. C.; Teague, C.; Ueshima, H.; Zhou, B. F.; Lindon, J. C.; Nicholson, J. K.; Stampler, J.; Elliott, P.; Chan, Q.; Holmes, E. *Anal. Chem.* **2006**, *78*, 2199–2208.
- (47) Akiyama, K.; Chikayama, E.; Yuasa, H.; Shimada, Y.; Tohge, T.; Shinozaki, K.; Hirai, M. Y.; Sakurai, T.; Kikuchi, J.; Saito, K. *In Silico Biol.* **2008**, *8*, 339–345.
- (48) Chikayama, E.; Sekiyama, Y.; Okamoto, M.; Nakanishi, Y.; Tsuboi, Y.; Akiyama, K.; Saito, K.; Shinozaki, K.; Kikuchi, J. *Anal. Chem.* **2010**, *82*, 1653–1658.
- (49) Steinbeck, C.; Kuhn, S. *Phytochemistry* **2004**, *65*, 2711–2717.
- (50) Robinette, S. L.; Veselkov, K. A.; Bohus, E.; Coen, M.; Keun, H. C.; Ebbels, T. M.; Beckonert, O.; Holmes, E. C.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2009**, *81*, 6581–6589.
- (51) Shrot, Y.; Frydman, L. *J. Chem. Phys.* **2008**, *128*, 052209.
- (52) Sommer, R. J.; Tautz, D. *Nature* **1993**, *361*, 448–450.
- (53) Freeman, R.; Kupce, E. J. *Biomol. NMR* **2003**, *27*, 101–113.
- (54) Zhang, F. L.; Bruschweiler, R. *J. Am. Chem. Soc.* **2004**, *126*, 13180–13181.
- (55) Brey, W. W.; Edison, A. S.; Nast, R. E.; Rocca, J. R.; Saha, S.; Withers, R. S. *J. Magn. Reson.* **2006**, *179*, 290–293.
- (56) Olson, D. L.; Peck, T. L.; Webb, A. G.; Magin, R. L.; Sweedler, J. V. *Science* **1995**, *270*, 1967–1970.
- (57) Kuhn, S.; Egert, B.; Neumann, S.; Steinbeck, C. *BMC Bioinf.* **2008**, *9*, 400.

Appendix C: Paper III

□□□□□□□□ □□□□□□□□ □□□□ □□□□□□□□ □□□□ □Dénes J, Szabó E, Robinette SL, Szatmári I, Szőnyi L , Kreuder JG, Rauterberg EW, Takáts Z. Metabonomics of Newborn Screening Dried Blood Spot Samples: A Novel Approach in the Screening and Diagnostics of Inborn Errors of Metabolism. *Anal Chem*, 2012. 84(22):10113-10120.)□□□□□□□□□□□□□□ □□□□□□ □□□□ □□□□□□□□□□□□

Metabonomics of Newborn Screening Dried Blood Spot Samples: A Novel Approach in the Screening and Diagnostics of Inborn Errors of Metabolism

Júlia Dénes,[†] Eszter Szabó,[‡] Steven L. Robinette,[†] Ildikó Szatmári,[§] László Szőnyi,[§] Joachim G. Kreuder,[⊥] Ernst W. Rauterberg,[⊥] and Zoltán Takáts^{*,†,§}

[†]Biomolecular Medicine, Department of Surgery and Cancer, Imperial College London, United Kingdom

[‡]Massprom Ltd., Budapest, Hungary

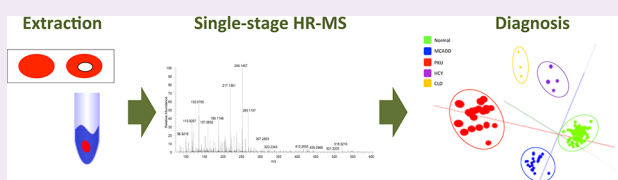
[§]Metabolic Diagnostic Laboratory, 1st Department of Pediatrics, Semmelweis University, Budapest, Hungary

[⊥]Hesse Child Health Centre, Johann Wolfgang Goethe University Hospital, Giessen, Germany

S Supporting Information

ABSTRACT: A novel, single stage high resolution mass spectrometry-based method is presented for the population level screening of inborn errors of metabolism. The approach proposed here extends traditional electrospray tandem mass spectrometry screening by introducing nanospray ionization and high resolution mass spectrometry, allowing the selective detection of more than 400 individual metabolic constituents

of blood including acylcarnitines, amino acids, organic acids, fatty acids, carbohydrates, bile acids, and complex lipids. Dried blood spots were extracted using a methanolic solution of isotope labeled internal standards, and filtered extracts were electrosprayed using a fully automated chip-based nanospray ion source in both positive and negative ion mode. Ions were analyzed using an Orbitrap Fourier transformation mass spectrometer at nominal mass resolution of 100 000. Individual metabolic constituents were quantified using standard isotope dilution methods. Concentration threshold (cutoff) level-based analysis allows the identification of newborns with metabolic diseases, similarly to the traditional electrospray tandem mass spectrometry (ESI-MS/MS) method; however, the detection of additional known biomarkers (e.g., organic acids) results in improved sensitivity and selectivity. The broad range of detected analytes allowed the untargeted multivariate statistical analysis of spectra and identification of additional diseased states, therapeutic artifacts, and damaged samples, besides the metabolic disease panel.



Since its introduction, electrospray tandem mass spectrometry (ESI-MS/MS) has become the most important and widely applied method in newborn screening (NBS) programs.^{1–3} Currently, there are more than 70 different metabolic diseases included in neonatal screening, with approximately 90% of them screened by the ESI-MS/MS method. The standard mass spectrometry newborn screening procedure comprises the extraction of dried blood spot (DBS) samples with a methanolic solution of stable isotope labeled internal standards. Amino acids and acylcarnitines are semi-quantitatively determined by detecting their *n*-butyl ester derivatives in positive ion mode tandem mass spectrometry. Amino acids are traditionally detected by neutral loss (NL 102) scan, and acylcarnitines are detected by precursor ion (PS 85) scan, with quantification based generally on multiple reaction monitoring (MRM) data.

Ionization efficiency and hence sensitivity is routinely increased via butyl esterification by eliminating the acidic carboxyl moiety from the analyte molecules. Derivatization is not an absolute prerequisite; however, only 30% of screening laboratories currently use derivatization-free methods.^{4,5} Elimination of butyl esterification is considered to yield more

accurate results by avoiding the formation of free amino acids via acidic hydrolysis of protein content.

Improving the sensitivity and specificity of screening, with a consequent reduction in the false-positive rate, is a key area of development in the newborn screening field. Cut-off values in the case of the tandem mass spectrometry-based technique are established to minimize the false negative rate to zero level. As a result, the number of the false-positive cases can occasionally be high, especially in the cases of certain diseases, such as homocystinuria, methylmalonic acidemia, and propionic acidemia (methionine and propionylcarnitine are used as diagnostic markers for all three diseases). Since suspected positive cases require confirmatory testing, false-positive cases involve unnecessary stress for the family of affected newborns and increased analytical burden on the metabolic diagnostic laboratories.^{6,7} Screening laboratories have used two main strategies to reduce the false-positive rate: excluding certain diseases from the screening (especially those, which otherwise

Received: September 11, 2012

Accepted: October 24, 2012

Published: October 24, 2012

do not fulfill the criteria for screening postulated by Wilson and Jungner⁸) and introducing so-called second-tier tests⁹ that use the original DBS sample for confirmation of the positive result. In these cases, the initial positive screening result is not reported until the confirmatory test validates this result, decreasing the unnecessary stress on families.¹⁰

Development of second-tier tests by the Mayo Clinic group provided a sufficient solution for the problem by minimizing overall false-positive rates to 0.09%.⁹ These tests are highly specific compared to the screening methods,^{11–16} but their introduction requires additional instrumentation, trained personnel, and an extension of the turnover time of screening, which may be critical in the case of several diseases.

Although most of the screened metabolic disorders are caused by the complete or partial absence of a single enzymatic function, the missing functionality results not only in the increase of the level of the specific substrate. Multiple metabolic constituents, both up- and downstream in the metabolic pathways, show changes in their concentrations; furthermore, elevation of the concentration of certain metabolites may also activate alternative pathways. As a result, serum profiles should reflect broad perturbations in the metabolome characteristic of individual diseases. While metabolomic studies traditionally employ GC/MS or LC/MS for the untargeted analysis of biological fluids,^{17–19} the feasibility of direct infusion mass spectrometry analysis has been demonstrated by a number of recent publications.^{20–22} Advantages of the elimination of chromatographic separation include shorter analysis times, simplified data analysis, and improved reproducibility. Since the specificity of unit resolution mass analyzers is poor, either high resolution (HR) or tandem mass spectrometry was used, to ensure selective detection of thousands of spectral features. Nevertheless, these improvements do not compensate for the suppression effects of electrospray ionization, which represent the most serious drawback for MS-only approaches.²³ Besides employing chromatographic separation prior to MS detection, suppression effects can partially be eliminated by minimizing the diameter of electrospray emitter tips, i.e., using nano-electrospray (or nanospray) ionization.²⁴ Nanospray has been demonstrated to suffer less from suppression effects, both in the case of large proton affinity differences between different analytes and the presence of inorganic salts in the sample.

The combination of nanospray ionization with high resolution mass spectrometry (nS-HR-MS) provides metabolome coverage comparable to that of LC/MS approaches as it is demonstrated by the current study. The concept was implemented by developing a high-throughput nanospray-HR-MS method for mapping the metabolic profile of DBS samples and establishing correlation between metabolite concentrations determined using the new approach and the traditional ESI-MS/MS technique. Resulting metabolic profiles were also correlated with the presence of various diseased statuses of the subjects using multivariate statistical methods. The developed method was tested for the most widely screened metabolic diseases including phenylketonuria (PKU) and medium-chain acyl-coenzyme A dehydrogenase deficiency (MCADD).

METHODS AND MATERIALS

Samples. Anonymous DBS samples were obtained from the Screening Laboratory at Semmelweis University, Budapest, Hungary, and Hesse Screening Centre, Gießen, Germany. 500 healthy and 66 abnormal samples were analyzed; for further details, see Supporting Information.

Sample Preparation. Dried blood spot samples ($d = 0.6$ mm, ~ 10 μL of whole blood) were placed in 96-well filter plates (Millipore, Billerica, MA, USA) and extracted with 100 μL of methanolic (HPLC grade, Sigma-Aldrich, St. Louis, MO, USA) solution, containing known concentrations of stable isotope labeled standards (Cambridge Isotope Laboratories, Andover, MA, USA); see Table S-1, Supporting Information, for details. The samples were extracted for 30 min at room temperature. The methanolic extracts were subsequently filtered into 96-well plates (Greiner, Frickenhausen, Germany), diluted with water (HPLC grade, Sigma-Aldrich, St. Louis, MO, USA) in 1:1 ratio, and analyzed by mass spectrometry.

Mass Spectrometry. High resolution mass spectrometry was performed using LTQ Orbitrap Discovery and Exactive instruments (Thermo Scientific GmbH, Bremen, Germany). A high-throughput screening method was developed using the Exactive instrument equipped with TriVersa NanoMate ion source (Advion, Ithaca, NY, USA). Chip-based infusion mode measurements were performed using 5 μm nominal internal diameter nozzle chips, and a sample volume of 5 μL was injected. Total data acquisition time was 2 min, 1 min in negative ion mode and 1 min in positive ion mode with automatic polarity switching. Details about the data analysis are provided in the Supporting Information.

RESULTS AND DISCUSSION

Detection and Identification of Metabolites. Nano-spray–high resolution mass spectrometry analysis of the methanolic extracts of newborn DBS samples results in spectra featuring hundreds of different metabolic constituents in both positive and negative mode. Parameters of the instruments were optimized to achieve the best intensity, sensitivity, and mass resolution in the lower mass range of 70–600 m/z (parameters are summarized in Table S-2, Supporting Information). Since the mass spectrometry device allows one to alternate polarity acquisition, positive and negative ion spectra were recorded in the same experiment (Figure S-1, Supporting Information). Positive ion mode data features ~ 450 individual ion signals (excluding isotopes and adduct peaks) in the mass range of 70–1000 m/z , being present in all 500 samples obtained from healthy newborns and analyzed by high resolution mass spectrometry. Negative ion mode data contains information about the presence of ~ 300 molecular constituents in the identical sample set. In addition to the core set of frequently observed ions, several hundred ion signals are present only in a fraction of the samples. Taking all samples into account, a total of 1150 ion signals were registered (without isotope and adduct signals) using a 10:1 signal-to-noise threshold.

Identification of ion signals was performed using accurate masses of ions determined at 50 000–100 000 fwhm (depending on actual m/z of the ion) nominal resolution with mass accuracy window of 1.0 ppm and tandem mass spectrometry. The accurate mass in the given mass range at the given precision generally allows the unequivocal determination of the molecular formula of species detected. Given the constraints set by the human biological origin of the species, one ion signal can be associated with only a few isomeric metabolic constituents. Tandem MS experiments were carried out to clarify the isomeric composition of ion signals corresponding to multiple known metabolites. Proper identification was achieved for 185 species in negative ion mode and 93 species in positive ion mode, corresponding to an overall 278 identified molecular

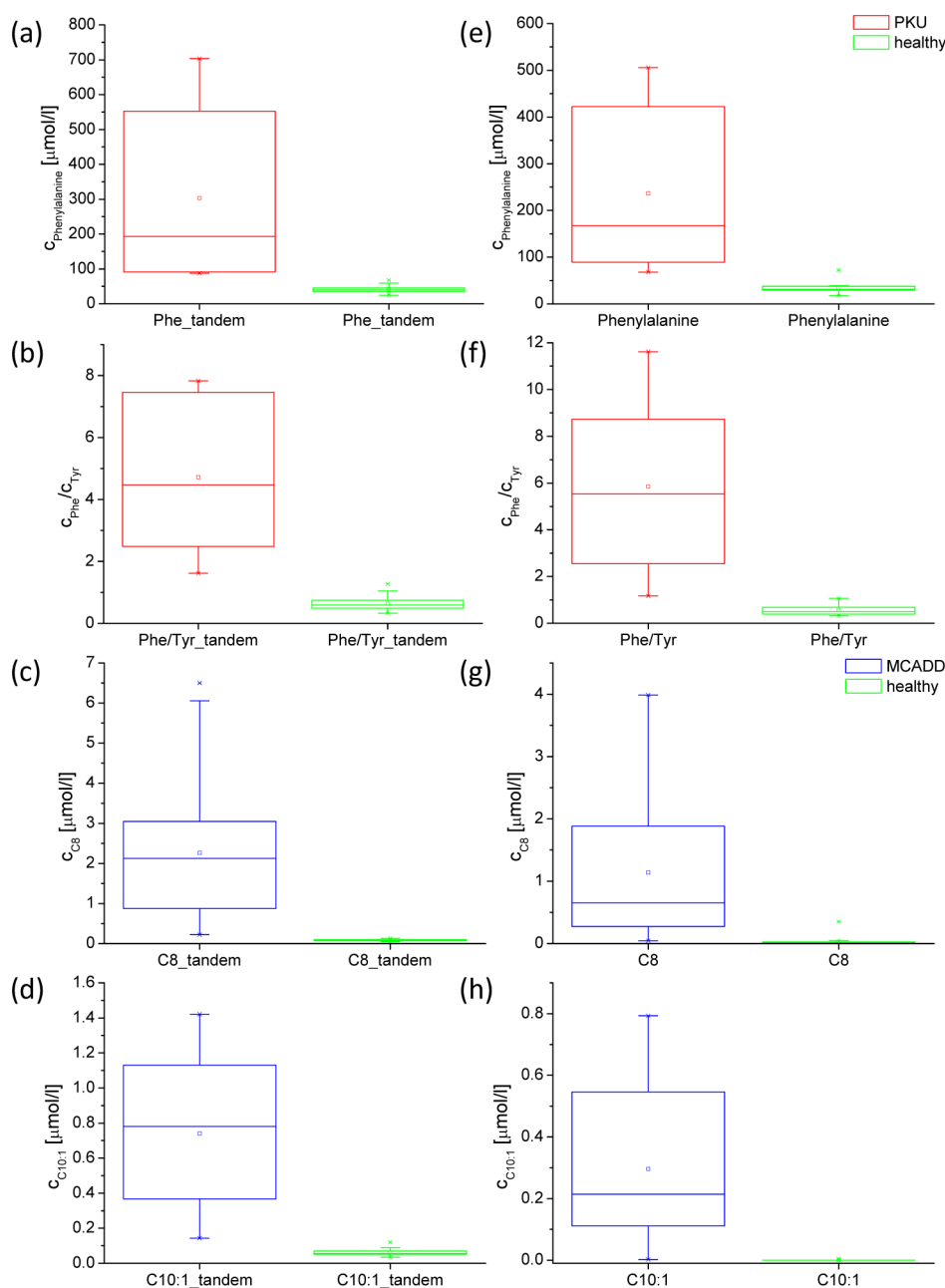


Figure 1. Box charts for PKU (a,b,e,f) and MCADD markers (c,d,g,h). (a, e) Phenylalanine, (b, f) phenylalanine and tyrosine ratio, (c, g) octanoylcarnitine (C8), and (d, h) decenoylcarnitine (C10:1). Tandem mass spectrometry screening results are shown on the left; nS-HR-MS results are on the right, from the same data set (500 healthy, 21 PKU, 21 MCADD).

constituents up to date; however, the peak identification process is still ongoing. Results of the identification are summarized in Table S-3, Supporting Information.

The observed metabolites include practically all amino acid and acylcarnitine metabolites relevant to newborn screening in addition to metabolites utilized for confirmatory diagnostics of metabolic diseases, including organic acids, carbohydrates, carbohydrate phosphates, bile acids, fatty acids, eicosanoids, phospholipids, and sphingolipids. Although single stage high resolution mass spectrometry does not allow the differentiation of isomers, the use of instruments with automated fragmentation of selected peaks (LTQ Orbitrap) enables these distinctions to be made. The resulting MS/MS spectra often provide sufficient information on the isomeric composition of the corresponding peak, as it is demonstrated in the

cases of 4-hydroxyphenyllactic and homovanillic acids in Figure S-2, Supporting Information.

One of the main concerns regarding quantitative determination using FT-MS technique is the narrow dynamic range provided by these instruments. Limit of detection (LOD) and linear range for a number of key metabolites was determined by extracting parallel samples (discs punched from the same NBS card, containing DBS from a single subject taken at the same time) with methanol containing various concentrations of isotope labeled internal standards. Identical DBS sample is expected to contain a reproducible amount of the analyte, so in this study, the native metabolite was used as internal standard. LOD values were defined as the concentration which yields 3:1 signal-to-noise ratio. Linear range was defined as the widest logarithmic concentration range where linear regression to the

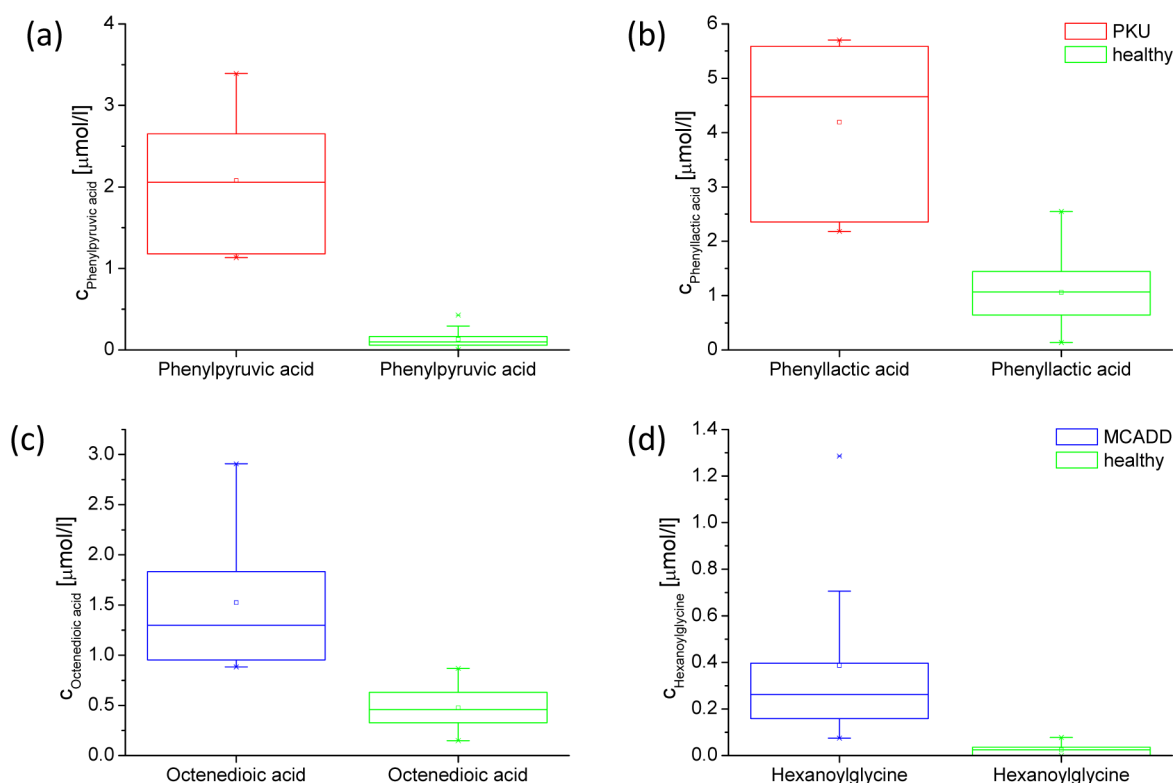


Figure 2. Box charts for possible markers of PKU, (a) phenylpyruvic acid and (b) phenyllactic acid. Measurements were performed in negative ion mode for 500 healthy, 21 PKU, and 12 false-positive PKU samples. Possible markers of MCADD, (c) octenedioic acid and (d) hexanoylglycine. Measurements were performed in negative ion mode for 500 healthy and 21 MCADD samples.

data gives $R^2 > 0.990$. Results are shown in Table S-4, Supporting Information, for 12 amino acids and 8 acylcarnitines generally used in newborn screening programs. As it is demonstrated, the LOD values are always lower than the physiological (or pathological if disease results in lower than physiological concentration) levels of the metabolites. Calibration curves exceeding beyond the linear range always show saturation phenomenon at higher concentrations (i.e., no breakdown phenomenon was observed). This guarantees that all pathological values will be detected as diseased, even if the determined level falls outside of the linear range.

Validation against Traditional ESI-MS/MS Method.

Both the traditional ESI-MS/MS method and the nS-HR-MS method presented here were able to quantify a set of 20 amino acid and 20 acylcarnitine species. These analytes were used as a benchmark to validate the nS-HR-MS method against the ESI-MS/MS techniques for a set of 500 newborn samples analyzed with both methods. Correlation plots generally reveal a linear relationship with correlation coefficients (Pearson product-moment correlation coefficient, r) in the range of 0.4–0.92. The slope of functions obtained by linear regression was <1 when nS-HR-MS data is plotted as a function of values measured by ESI-MS/MS, indicating that the former technique tends to yield equal or lower values compared to the latter one (see Figure S-3, Supporting Information). The systematic deviation was tentatively associated with the improved selectivity of the high resolution method, but we could not confirm this assumption due to the inability of the Orbitrap mass spectrometer to isolate parent ions at arbitrarily narrow m/z windows. Furthermore, while the HR-MS method detects the unmodified molecule, the traditional method detects its butyl-ester, raising further complications for the comparison.

The butyl-esterification procedure may also induce the hydrolysis of proteins, which shifts the determined amino acid concentrations further toward higher values. Nevertheless, both nS-HR-MS and ESI-MS/MS were validated against a quantification of amino acids by LC-MS/MS, and the results for a limited number of samples clearly showed that nS-HR-MS results gives better correlation with those of LC-MS/MS (0.87 vs 0.68 for Phe, 50 samples). These results indicate that the traditional method gives systematically higher concentrations caused by either interferences with similar molecular weight and fragmentation pattern or identical metabolite produced during sample preparation. However, this systematic shift does not necessarily compromise the diagnostic value of concentrations obtained by the traditional method; while the healthy and diseased range may be different for the nS-HR-MS and the traditional ESI-MS/MS approach, the separation of the ranges is comparable or better for metabolites determined by both methods.

Screening for PKU and MCADD. Although the set of diseases included in newborn screening programs varies from country to country (in some cases even between regions of the same country), phenylketonuria and medium-chain acyl-coenzyme A dehydrogenase deficiency are included in all programs where mass spectrometry is used for screening. Phenylketonuria is generally screened using the phenylalanine concentration as primary parameter and Phe/Tyr ratio (and occasionally Fischer quotient) as secondary/confirmatory parameters.²⁵ Box plots representing the distribution of key metabolite concentrations for healthy population and diseased population using the two methods are shown in Figure 1 for MCAD deficiency and PKU, since samples from a sufficiently large number of freshly diagnosed patients were available in

these two cases. The box plot diagrams indicate that these analytes differentiate between healthy and diseased population using the nS-HR-MS method.

While both sensitivity and specificity of phenylalanine level-based phenylketonuria screening is excellent, the rare false-positive cases are generally associated with a transient high phenylalanine level caused by total parenteral nutrition or immature metabolism of prematurely born babies. Both of these statuses are indicated in principle on the screening cards; however, on one hand, this information is often missing, and on the other hand, these babies require further examination anyway, according to current existing protocols. In contrast, the novel method offers the utilization of a few additional parameters which are analytically not associated with the phenylalanine concentration. Phenyllactic acid and phenylpyruvic acid are metabolites along an alternative catabolic pathway of phenylalanine. Since phenylalanine is normally converted to tyrosine by the phenylalanine hydroxylase (PAH) enzyme, loss of this enzymatic function (i.e., in case of phenylketonuria) dramatically increases the concentration of the metabolites of the alternative pathway, e.g., that of phenyllactic and phenylpyruvic acid.²⁶ These two metabolites along with a few others are readily detected by the nS-HR-MS method in negative ion mode, using the same nanospray emitter. Separation of healthy and diseased population is even better in the case of phenylpyruvic acid than in the case of phenylalanine. DBS samples from 12 false-positive cases (babies screened out with Phe levels higher than 99 percentile, but having normal PAH activity) were analyzed using nS-HR-MS method, and the phenylpyruvic acid levels were found to be $<0.25 \mu\text{M}$ (cf. Figure 2a,b) in all cases. In 5 of these cases, the elevated Phe level was associated with amino acid infusion, while in 3 cases the gestational age of babies was <36 weeks. Although our results do not fully validate phenylpyruvic acid as a selective marker for PKU in cases of premature babies and babies receiving amino acid infusion, the observations strongly support its testing at the population level as an additional/confirmatory marker for PKU.

Samples of patients suffering from other aminoacidopathies including tyrosinemia I, citrullinemia, and homocystinuria have also been tested with similar results. Markers found in the case of each disease are summarized in Table S-5, Supporting Information.

MCAD deficiency is one of the most abundant mitochondrial fatty acid beta oxidation disorders, and it is included in most of the newborn screening programs worldwide. The deficiency of medium-chain acyl-coenzyme A dehydrogenase activity interrupts the stepwise beta oxidation of long- and medium-chain fatty acids, causing problems in the general energy (ATP) production of the body. The condition also results in the accumulation of fatty acid metabolites produced by alternative (i.e., to mitochondrial beta oxidation) metabolic pathways, e.g., peroxysomal beta oxidation or omega oxidation by cytochrome P450. The disease is traditionally screened by ESI-MS/MS method, monitoring the elevated level of octanoylcarnitine (C8). While an elevated octanoylcarnitine level is a hallmark of the disease, the sole use of this metabolite results in fairly high (up to 3%) false-positive rates. False positives are generally associated with the overlap of healthy (comprising also heterozygous cases) and diseased population and administration of valproic acid to infants with epileptic symptoms. Since valproic acid is an isomer of caprylic acid and conjugates with carnitine, the resulting valproylcarnitine cannot be

distinguished from C8. A further problem, pertaining to the specificity of the method, is that other disorders of mitochondrial fatty acid beta oxidation (e.g., Glutaric acidemia type 2) also involve the elevation of C8 levels. Hence, in addition to C8, other acylcarnitine concentrations (C2, C6, C10, C10:1 or their concentration ratios) are used as additional or confirmatory markers.^{27,28} The nS-HR-MS method is also able to quantify these acylcarnitine levels and ratios, as it is demonstrated by Figure 1.

The standard, generally accepted confirmatory test for newborns screened with suspected MCAD deficiency is urinary organic acid analysis, followed by genetic testing for the G985 hotspot mutation. In the case of urinary GC/MS analysis, dicarboxylic acids, acylglycines, and fatty acid metabolites are determined for the confirmation of the diagnosis. Unlike the standard positive ion ESI-MS/MS method, nS-HR-MS technique is able to detect and quantify these metabolites from the dried blood spot extracts. Since generally the urinary concentration (normalized to creatinine level) of these metabolites is used for diagnostics,^{29,30} it was a question if whether the concentration determined in DBS samples also serves as an acceptable marker for the disease. As the box charts depicted in Figure 2 demonstrate, octenedioic acid and hexanoylglycine levels determined by the nS-HR-MS technique can serve as analytically independent markers for the MCAD deficiency. Similarly to PKU, we have studied the utility of these markers for improving the specificity of the results. Twenty-one newborn DBS samples were selected which were found to be positive for MCAD deficiency; however, the results of the confirmatory test did not support the diagnosis. For the traditional ESI-MS/MS method, C8, C6, C10:1-acylcarnitines and C10:1/C2-acylcarnitine ratios were used as screening parameters, and the result was accepted as positive where at least 3 of these parameters were higher than a predefined cutoff value. The nS-HR-MS analysis of these samples yielded confirmatory metabolite concentrations lower than the first percentile of diseased population, between 3rd and 44th percentile of healthy population for octenedioic acid, between 65th and 89th percentiles for hexanoylglycine, and between 40th and 51st percentile for suberic acid.

The results indicate that, using these sets of markers as confirmatory markers, the false-positive cases found by acylcarnitine analysis can be correctly classified as healthy newborns. Using octanoylcarnitine level combined with these markers and accepting a case positive if all of the parameters exceed the estimated 1 percentile value of diseased population, 100% selectivity and sensitivity was achieved for the detection of MCAD deficiency in the test set; however, the tested population (500 healthy and 21 diseased) is not sufficiently large to draw far-fetched conclusions regarding specificity and sensitivity of the method in a population level screening program.

As it has already been pointed out, the nS-HR-MS method reveals the presence of more than a thousand molecular species, with currently 278 compounds already identified. The group of identified compounds includes a number of species which are known markers of metabolic and nonmetabolic diseases as it is discussed in detail in the Supporting Information.

Statistical Data Analysis. While the detection and inclusion of alternative biomarkers into the mass spectrometry screening method already provides an incremental improvement in the specificity of the process, the untargeted nature of the presented nS-HR-MS method also allows a multivariate

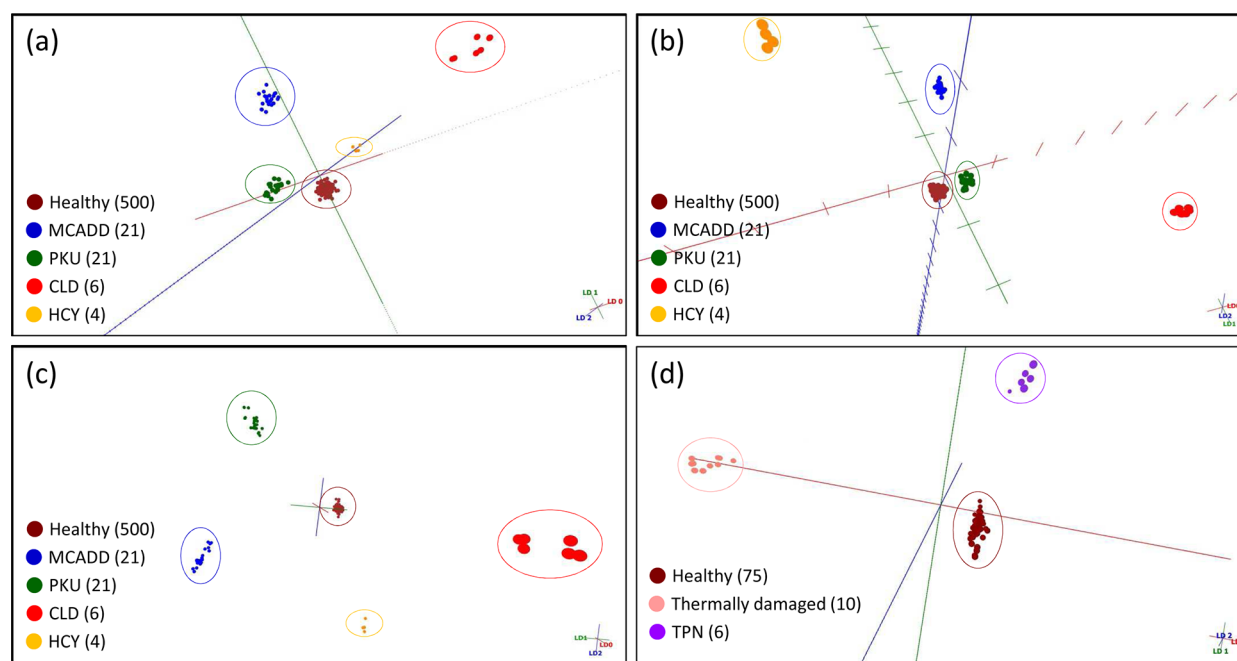


Figure 3. Three dimensional LDA models representing the 60 dimensional PCA/LDA analyses of the sample sets. (a) Model of low resolution (1 Da bin size) measurements; the first 60 components represent 96.52% of the total variance. (b) Model of high resolution (0.01 Da bin size) measurements; the first 60 components represent 98.22% of the total variance. (c) Model for identified peaks; the first 60 components represent 98.81% of the total variance. (d) Model for low resolution measurements on different types of false-positive samples.

statistical data interpretation approach, similar to the method recently described for ESI-MS/MS.³¹ Since the removal of a specific enzymatic function of key importance from the metabolic machinery is expected to result in a global change in the metabolome (vide supra), this has to be traceable by pattern recognition approaches.³² Mass spectra can be treated as n -dimensional data vectors where n equals the number of possible m/z values, i.e., the peak capacity of the spectrum. This value in the mass range of 70–600 m/z is 106 000 at high resolution (0.01 Da bin size) or 1060 at low resolution (1 Da bin size) for the combined positive/negative ion mode data. Individual spectra appear as points in this 106 000 dimensional space, potentially forming groups according to the specific conditions. However, since in most of these dimensions the data cloud has no extension and many of these dimensions are not linearly independent, it is reasonable to reduce the dimensions by appropriate mathematical tools, e.g., principal component analysis (PCA). Dimension reduction by PCA results in a still orthogonal data space determined by the linearly independent principal components. Three or two-dimensional projection of the data in the PCA space is generally used for the demonstration of similarities and dissimilarities. Separation of assigned data groups (i.e., condition specific data groups) can further be improved by linear discriminant analysis (LDA), which is a supervised linear transformation resulting in a nonorthogonal data space where the distances among data groups are maximized. Similarly to PCA, three-dimensional projection of LDA is used for the demonstration of dissimilarity among assigned data groups. Besides demonstration of the similarities and dissimilarities, an assigned LDA space can also be used for the identification of an unknown data point by localizing it in the multidimensional data space and determining its distance from the assigned data groups. In the optimal case, the unknown falls into or close to one of the clearly separated condition-specific data clouds.

Since the peaks of the important biomarker molecules have typically low absolute intensity values, application of normalization is necessary during the data analysis. While a number of multivariate statistical tools suitable for the analysis of this type of data exist, in the present study, we employed only consecutive PCA and LDA analyses. Peaks in mass spectra were reduced to the list of the identified metabolite molecules shown in the Table S-3, Supporting Information, and peak-specific PCA/LDA was performed on this reduced data set. The influence of the intensive contaminant peaks, especially in positive ion mode, and nonmetabolite peaks can be avoided using this approach.

The results of PCA and LDA analysis of the first 60 principal components corresponding to a newborn cohort comprising 500 healthy newborns, 21 MCADD, 21 PKU, 21 homocystinuria (HCY) patients, and 6 cholestatic liver diseases (CLD) is shown in Figure 3. Figure 3a represents the data obtained from low resolution measurements; Figure 3b shows the high resolution analysis, and Figure 3c demonstrates the peak-specific data analysis method. Each plot reveals complete separation of the healthy population and individual metabolic disorder-specific data groups.

In the model presented here, it is seemingly not critical to use high resolution mass spectrometry since unambiguous identification can be obtained even using data binned to unit resolution. However, when a statistically relevant number of healthy samples (in the order of some 10 000s vs 500 in the current study) and more conditions (all currently screened inborn errors of metabolism including sample mistreatment conditions etc.) are examined, a considerable spreading of the data within individual subject groups is expected. Thus, the compactness of data groups will become more critical, and a chance of unambiguous identification of the conditions will be improved by high mass resolution.

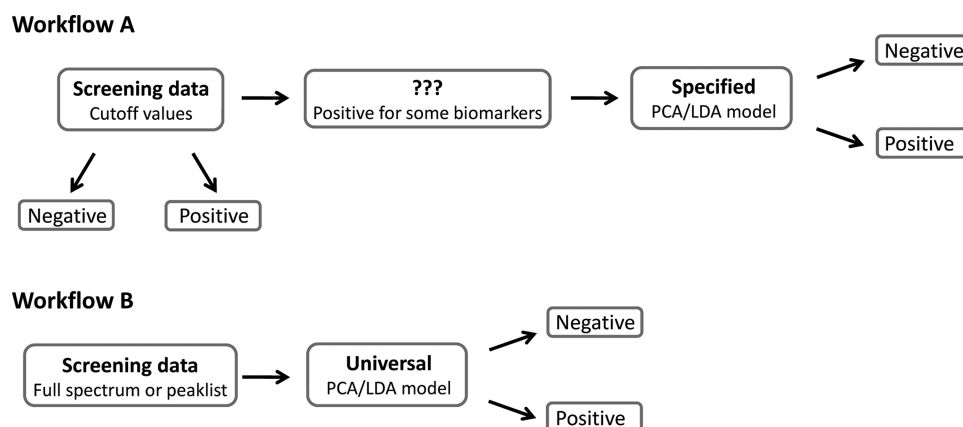


Figure 4. Different workflows of the newborn screening process suggested for the application of the nS-HR-MS approach.

Sample conditions associated with false-positive cases have also been studied using the multivariate statistical approach. Total parenteral nutrition (TPN) cases or thermally damaged samples obtained from healthy subjects have shown complete separation, already in the three-dimensional projection of the 60 dimensional LDA space (Figure 3d). Although single stage mass spectrometry is not suitable for identification of associated metabolites, the differences of the spectral pattern enable the distinction between disorders and these conditions. Selective identification of these conditions is extremely important, since most of the false positives are associated with these or other similar causes.

The approach offers two potential workflows depicted in Figure 4. In case of workflow A, the screening data is evaluated using cutoff values established for a number of parameters in the case of each condition. If the results are unambiguous (all markers are either in the healthy or in the diseased range), then the result is established. However, if some values are in the healthy range while others are in the diseased range, then PCA/LDA analysis is performed using a model which comprises all relevant conditions. For instance, if octanoylcarnitine is found to be elevated, then a model comprising MCADD, Glutaric acidemia type 2 cases, false positives, babies receiving valproate, and damaged cards would be used.

In case of workflow B, the data points corresponding to screened newborns are localized in the universal model (containing all true positives, normal healthy newborns, and nonmetabolic conditions) and evaluation is performed in a single step. While workflow B seems to be simpler, the large number of assigned data groups may compromise selectivity, which makes the utilization of workflow A safer.

Due to the limited number of positive samples, the system was tested only in cases of PKU, MCADD, and HCY using both workflows and also the ESI-MS/MS method used for population level screening in western and central Hungary. The cutoff values used in the national screening program are harmonized with those established by the international collaborative study coordinated by Rinaldo et al. at Mayo Clinic.³³ Performance of the three approaches was compared, and it was concluded that both workflows A and B operate flawlessly using these limited numbers of samples and conditions, while the traditional method yields a low, but still significant, number of false positives ($\sim 0.65\%$). The required analysis time is comparable for each method, but significant increase of overall turnaround time occurs, when second-tier tests are included. The necessity of second-tier tests in the case

of the ESI-MS/MS method is well expressed through the reduced false-positive rate (0.09%), which is expected to reach even lower values in the herein reported method (estimated value: 0.05%).

CONCLUSIONS

The presented nS-HR-MS method offers a promising alternative to the traditionally used ESI-MS/MS approach at the population level screening of newborn babies for inborn errors of metabolism. Thereby, the chip-based nanoelectrospray system offers several advantages compared to an ESI autosampler, most notably the lack of cross-contamination resulting in a significantly reduced number of false-positive results. Besides the detection of amino acids and acylcarnitines, the novel method is also capable of the semiquantitative determination of highly specific disease markers including various organic acids, acylglycines, or even carbohydrates. The untargeted nature of the novel method also allows multivariate pattern recognition analysis of the resulting spectral information, which results in improved specificity at the recognition of screened conditions, and also the identification of interferences caused by the medical treatment of babies or damaged DBS samples. Routine application of the method is expected to decrease the number of second samples requested as well as the number of second-tier or confirmatory tests (amino acid chromatography, GC/MS profiling of urinary organic acids, succinylacetone, etc).

In principle, the method also has potential in the screening of galactosemia (monitoring hexose and hexose-phosphate), biotinidase deficiency, congenital hypothyroidism, lysosomal storage disorders, certain hemoglobinopathies, and a number of other diseases involving global changes in the metabolome. The sample set used for this pilot study was a limited collection of healthy and diseased neonates, which represents neither the statistical variance of healthy population nor the prevalence of the examined disorders yet. The data presented here already shows significant differences between normal and abnormal samples, indicating the suitability of the proposed methodology for the neonatal screening of DBS samples at proof-of-principle level. The study is being extended for a larger, statistically relevant set of samples, and the validation at population level is already in progress.

■ ASSOCIATED CONTENT

■ Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Address: Biomolecular Medicine, Department of Surgery and Cancer, Imperial College London, South Kensington Campus, Sir Alexander Fleming Building, SW7 2AZ, London, United Kingdom. Phone: +44 20 7594 2760. E-mail: z.takats@imperial.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The Exactive instrument was generously provided by Thermo Fisher Scientific for the project and the TriVersa NanoMate ion source was kindly provided by Advion for the high-throughput analysis of the samples. The work was funded by European Research Council under Starting Grant scheme (Contract No: 210356) and Hessisches Ministerium für Wissenschaft und Kunst (LOEWE Schwerpunkt "AmbiProbe"), and the NIHR Biomedical Research Centre.

■ REFERENCES

- (1) Rashed, M. S. *J. Chromatogr., B* **2001**, 758, 27–48.
- (2) Chace, D. H.; Kalas, T. A.; Naylor, E. W. *Clin. Chem.* **2003**, 49, 1797–1817.
- (3) Chace, D. H. *J. Mass Spectrom.* **2009**, 44, 163–170.
- (4) Piraud, M.; Vianey-Saban, C.; Petritis, K.; Elfakir, C.; Steghens, J. P.; Morla, A.; Bouchu, D. *Rapid Commun. Mass Spectrom.* **2003**, 17, 1297–1311.
- (5) De Jesus, V. R.; Chace, D. H.; Lim, T. H.; Mei, J. V.; Hannon, W. H. *Clin. Chim. Acta* **2010**, 411, 684–689.
- (6) Waisbren, S. E.; Albers, S.; Amat, S.; Ampola, M.; Brewster, T. G.; Demmer, L.; Eaton, R. B.; Greenstein, R.; Korson, M.; Larson, C.; Marsden, D.; Msall, M.; Naylor, E. W.; Pueschel, S.; Seashore, M.; Shih, V. E.; Levy, H. L. *JAMA-J. Am. Med. Assoc.* **2003**, 290, 2564–2572.
- (7) Hewlett, J.; Waisbren, S. E. *J. Inherited Metab. Dis.* **2006**, 29, 677–682.
- (8) Wilson, J. M. G.; Jungner, G. *WHO Chronicle* **1968**, 22, 473.
- (9) Matern, D.; Tortorelli, S.; Oglesbee, D.; Gavrilov, D.; Rinaldo, P. *J. Inherited Metab. Dis.* **2007**, 30, 585–592.
- (10) Chace, D. H.; Hannon, W. H. *Clin. Chem.* **2010**, 56, 1653–1655.
- (11) Oglesbee, D.; Sanders, K. A.; Lacey, J. M.; Magera, M. J.; Casetta, B.; Strauss, K. A.; Tortorelli, S.; Rinaldo, P.; Matern, D. *Clin. Chem.* **2008**, 54, 542–549.
- (12) Chace, D. H.; Lim, T.; Hansen, C. R.; De Jesus, V. R.; Hannon, W. H. *Clin. Chim. Acta* **2009**, 407, 6–9.
- (13) Schwarz, E.; Liu, A. P.; Randall, H.; Haslip, C.; Keune, F.; Murray, M.; Longo, N.; Pasquali, M. *Pediatr. Res.* **2009**, 66, 230–235.
- (14) Forni, S.; Fu, X. W.; Palmer, S. E.; Sweetman, L. *Mol. Genet. Metab.* **2010**, 101, 25–32.
- (15) Turgeon, C. T.; Magera, M. J.; Cuthbert, C. D.; Loken, P. R.; Gavrilov, D. K.; Tortorelli, S.; Raymond, K. M.; Oglesbee, D.; Rinaldo, P.; Matern, D. *Clin. Chem.* **2010**, 56, 1686–1695.
- (16) Shigematsu, Y.; Hata, I.; Tajima, G. *J. Inherited Metab. Dis.* **2010**, 33, S283–S288.
- (17) Pasikanti, K. K.; Ho, P. C.; Chan, E. C. Y. *J. Chromatogr., B* **2008**, 871, 202–211.
- (18) Theodoridis, G. A.; Gika, H. G.; Want, E. J.; Wilson, I. D. *Anal. Chim. Acta* **2012**, 711, 7–16.
- (19) Dettmer, K.; Aronov, P. A.; Hammock, B. D. *Mass Spectrom. Rev.* **2007**, 26, 51–78.
- (20) Koulman, A.; Tapper, B. A.; Fraser, K.; Cao, M. S.; Lane, G. A.; Rasmussen, S. *Rapid Commun. Mass Spectrom.* **2007**, 21, 421–428.
- (21) Lokhov, P. G.; Kharybin, O. N.; Archakov, A. I. *Int. J. Mass Spectrom.* **2012**, 309, 200–205.
- (22) Schuhmann, K.; Almeida, R.; Baumert, M.; Herzog, R.; Bornstein, S. R.; Shevchenko, A. *J. Mass Spectrom.* **2012**, 47, 96–104.
- (23) Tang, L.; Kebarle, P. *Anal. Chem.* **1993**, 65, 3654–3668.
- (24) Wilm, M.; Mann, M. *Anal. Chem.* **1996**, 68, 1–8.
- (25) Chace, D.; Millington, D.; Terada, N.; Kahler, S.; Roe, C.; Hofman, L. *Clin. Chem.* **1993**, 39, 66–71.
- (26) Kaufman, S. *J. Pediatr.* **1989**, 114, 895–900.
- (27) Chace, D. H.; Hillman, S. L.; Van Hove, J. L. K.; Naylor, E. W. *Clin. Chem.* **1997**, 43, 2106–2113.
- (28) Clayton, P. T.; Doig, M.; Ghafari, S.; Meaney, C.; Taylor, C.; Leonard, J. V.; Morris, M.; Johnson, A. W. *Arch. Dis. Child.* **1998**, 79, 109–115.
- (29) Bennett, M. J.; Coates, P. M.; Hale, D. E.; Millington, D. S.; Pollitt, R. J.; Rinaldo, P.; Roe, C. R.; Tanaka, K. *J. Inherited Metab. Dis.* **1990**, 13, 707–715.
- (30) Bonafe, L.; Troxler, H.; Kuster, T.; Heizmann, C. W.; Chamoles, N. A.; Burlina, A. B.; Blau, N. *Mol. Genet. Metab.* **2000**, 69, 302–311.
- (31) Marquardt, G.; Currier, R.; McHugh, D. M. S.; Gavrilov, D.; Magera, M. J.; Matern, D.; Oglesbee, D.; Raymond, K.; Rinaldo, P.; Smith, E. H.; et al. *Genet. Med.* **2012**, 14, 648–655.
- (32) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, 29, 1181–1189.
- (33) McHugh, D. M. S.; Cameron, C. A.; Abdenur, J. E.; Abdulrahman, M.; Adair, O.; Al Nuaimi, S. A.; Ahlman, H.; Allen, J. J.; Antonozzi, I.; Archer, S.; et al. *Genet. Med.* **2011**, 13, 230–254.

Appendix D: Paper IV

Nephrocalcinosis (Enamel Renal Syndrome) Caused by Autosomal Recessive *FAM20A* Mutations

Graciana Jaureguiberry¹ Muriel De la Dure-Molla² David Parry³ Mickael Quentric⁴ Nina Himmerkus⁵ Toshiyasu Koike⁶ James Poulter³ Enriko Klootwijk¹ Steven L. Robinette⁷ Alexander J. Howie¹ Vaksha Patel¹ Marie-Lucile Figueres⁸ Horia C. Stanescu¹ Naomi Issler¹ Jeremy K. Nicholson⁷ Detlef Bockenhauer¹ Christopher Laing¹ Stephen B. Walsh¹ David A. McCredie⁹ Sue Povey¹⁰ Audrey Asselin¹¹ Arnaud Picard¹¹ Aurore Coulomb¹² Alan J. Medlar¹ Isabelle Bailleul-Forestier¹³ Alain Verloes¹⁴ Cedric Le Caignec¹⁵ Gwenaëlle Roussey¹⁶ Julien Guiol¹⁷ Bertrand Isidor¹⁵ Clare Logan³ Roger Shore¹⁸ Colin Johnson³ Christopher Inglehearn³ Suhaila Al-Bahlani¹⁹ Matthieu Schmittbuhl²⁰ François Clauss²⁰ Mathilde Huckert²⁰ Virginie Laugel²¹ Emmanuelle Ginglinger²² Sandra Pajarola²³ Giuseppina Spartà²⁴ Deborah Bartholdi²³ Anita Rauch²³ Marie-Claude Addor²⁵ Paulo M. Yamaguti²⁶ Heloisa P. Safatle²⁷ Ana Carolina Acevedo²⁶ Hercilio Martelli-Júnior²⁸ Pedro E. dos Santos Netos²⁸ Ricardo D. Coletta²⁹ Sandra Gruessel⁵ Carolin Sandmann⁵ Denise Ruehmann⁵ Craig B. Langman³⁰ Steven J. Scheinman³¹ Didem Ozdemir-Ozenen³² Thomas C. Hart³³ P. Suzanne Hart³⁴ Ute Neugebauer³⁵ Eberhard Schlatter³⁵ Pascal Houillier⁸ William A. Gahl³⁴ Miikka Vikkula⁴ Agnès Bloch-Zupan²⁰ Markus Bleich⁵ Hiroshi Kitagawa⁶ Robert J. Unwin¹ Alan Mighell¹⁸ Ariane Berdal¹¹ Robert Kleta¹

¹Centre for Nephrology, University College London, London, UK; ²Rothschild Dental Hospital Service, Paris, France; ³Molecular Medicine, University of Leeds, Leeds, UK; ⁴de Duve Institute, Université catholique de Louvain, Brussels, Belgium; ⁵Physiology, University of Kiel, Kiel, Germany; ⁶Biochemistry, Kobe Pharmaceutical University, Kobe, Japan; ⁷Biomolecular Medicine, Imperial College London, London, UK; ⁸Cordeliers Research Center, Paris-Descartes University, Paris, France; ⁹Royal Children's Hospital, Melbourne, Vic., Australia; ¹⁰Department of Genetics, Evolution and Environment, UCL, London, UK; ¹¹INSERM, UMRs 872, University Paris-Diderot, and ¹²Pathology service, Armand Trousseau Hospital, Paris, ¹³Toulouse Hospital, Sabatier University, Toulouse, ¹⁴Department of Genetics, APHP – Robert Debré University Hospital, Paris, and Services de ¹⁵Génétique, ¹⁶Pédiatrie and ¹⁷Stomatologie, CHU de Nantes, Nantes, France; ¹⁸Leeds Dental Institute, University of Leeds, Leeds, UK; ¹⁹Al-Nahda Hospital, Muscat, Sultanate of Oman; ²⁰University of Strasbourg, Strasbourg, ²¹IGBMC, INSERM, U964, Illkirch, and ²²Service de Génétique, Hôpital Emile Muller, Mulhouse, France; ²³Medical Genetics, University of Zurich, and ²⁴Nephrology Unit, University Children's Hospital, Zurich, and ²⁵Service de Génétique Médicale, Lausanne, Switzerland; ²⁶Health Sciences School and ²⁷Department of Medical Genetics, University of Brasília, Brasília, ²⁸State University of Montes Claros, Minas Gerais, and ²⁹Dental School, State University of Campinas, São Paulo, Brazil; ³⁰Pediatric Nephrology, Northwestern University, Chicago, Ill., and ³¹The Commonwealth Medical College, Scranton, Pa., USA; ³²Pedodontics, Yeditepe University, Istanbul, Turkey; ³³Periodontics, University of Illinois at Chicago, Chicago, Ill., and ³⁴NHGR, NIH, Bethesda, Md., USA; ³⁵Internal Medicine D, University of Muenster, Muenster, Germany

G. Jaureguiberry, M. De la Dure-Molla, D. Parry, M. Quentric, N. Himmerkus, T. Koike, M. Bleich, H. Kitagawa, R.J. Unwin, A. Mighell, A. Berdal, and R. Kleta contributed equally to this work.

Key Words

Nephrolithiasis · Urolithiasis · Amelogenesis imperfecta · FAM20B · FAM20C

Abstract

Background/Aims: Calcium homeostasis requires regulated cellular and interstitial systems interacting to modulate the activity and movement of this ion. Disruption of these systems in the kidney results in nephrocalcinosis and nephrolithiasis, important medical problems whose pathogenesis is incompletely understood. **Methods:** We investigated 25 patients from 16 families with unexplained nephrocalcinosis and characteristic dental defects (amelogenesis imperfecta, gingival hyperplasia, impaired tooth eruption). To identify the causative gene, we performed genome-wide linkage analysis, exome capture, next-generation sequencing, and Sanger sequencing. **Results:** All patients had biallelic *FAM20A* mutations segregating with the disease; 20 different mutations were identified. **Conclusions:** This autosomal recessive disorder, also known as enamel renal syndrome, of *FAM20A* causes nephrocalcinosis and amelogenesis imperfecta. We speculate that all individuals with biallelic *FAM20A* mutations will eventually show nephrocalcinosis.

Copyright © 2013 S. Karger AG, Basel

Introduction

Nephrocalcinosis (NC), diagnosed by radiographs, CT, or increased echogenicity on ultrasound, represents an important renal complication because it can accompany progressive deterioration of glomerular function or nephrolithiasis [1]. In some cases, NC provides a clue to an underlying genetic disorder such as hyperoxaluria or distal renal tubular acidosis with or without deafness [2, 3]. In other cases, NC is a side effect of chronic treatment with various drugs, including loop diuretics and vitamin D. In general, the pathogenesis of NC has not been adequately elucidated, although hypercalciuria appears to be a common finding.

Often, rare genetic diseases reveal previously unrecognized mechanisms of action regarding physiology, cell biology, and metabolism. Here, we present genetic studies into a rare human disease of NC combined with amelogenesis imperfecta (AI), a disorder of abnormal enamel formation and impaired tooth eruption.

Methods

Patients and families were identified in our rare disease renal tubular or dental/craniofacial reference centers and gave informed consent. This study was approved by the institutional review boards and ethics committees of the various centers. All patients had NC confirmed by either ultrasound, X-ray or CT, and all showed characteristic teeth findings with AI and delayed or missing tooth eruption.

In short, multipoint parametric linkage analysis utilizing 2,000 highly polymorphic markers (DeCode, Iceland) across the whole genome in 4 informative families as well as homozygosity mapping for another consanguineous family was used to determine the locus linked to this trait, as published before [4, 5]. Next-generation sequencing using exome capture (Perkin Elmer, USA; Leeds Translational Genomics Unit, UK) was performed on 5 patients from 5 unrelated families; 4 were part of the linkage analysis. Subsequent data analysis was restricted to novel sequence variants within the linked region. The frequency of each variant was examined in >100 ethnically matched alleles available in public databases (1000 Genomes, release 12 – May 2012). Sanger sequencing was performed as described, and mutations were sequenced in family members for segregation analysis.

In detail, genomic DNA was isolated from peripheral blood lymphocytes for all subjects using standard protocols. Genotypes from polymorphic markers for 4 families were generated by DeCode, Iceland. Analyses were carried out as published before with modifications [4]. In short, genotypes were examined using a multipoint parametric linkage analysis and haplotype reconstruction was performed via Allegro and Genehunter for an autosomal recessive model with complete penetrance, disease allele frequency of 0.001 (DeCode map with appropriate allele frequencies). The data were formatted using Mega2 (version 4.0) through Alohomo (version 0.30, Win32); non-informative markers were filtered out. Mendelian inconsistencies were checked using PedCheck (version 1.1); unlikely genotypes were identified and filtered using Merlin (version 1.1 alpha 3). The Allegro haplotype output files were visualized with HaploPainter. In addition and in parallel, whole-genome SNP microarray analysis was performed on genomic DNA for another consanguineous family by AROS Applied Technology; resulting data were analyzed using IBD finder software [5]. Whole-exome sequencing was performed using 3 µg of genomic DNA, which was sheared and ligated to Illumina adapters, according to Agilent's SureSelect Library Prep protocol. The sample was then size selected (200–300 bp) by agarose gel electrophoresis and enriched for 12 cycles, using PCR prior to hybridization to the SureSelect reagent for 24 h at 65°C. The library was denatured using NaOH and diluted to a concentration of 12 pM, of which 120 µl was hybridized onto a v5 single-read flow cell (Illumina, San Diego, Calif., USA). Samples were prepared for sequencing according to Illumina's standard amplification, linearization, blocking and primer hybridization protocols. The flow cell was then loaded onto an Illumina GAIIx and sequencing performed for 80 cycles after which the raw data were processed using the Illumina pipeline. The data (qseq) files generated were aligned to the human reference sequence (hg19/GRCh37) using Novoalign short-read alignment software (Novocraft Technologies, Selangor, Malaysia). Duplicate reads and reads mapping to multiple locations were excluded from any analysis. SAMtools and the Genome

Table 1. *FAM20A* mutations in patients with NC and AI

Family	Age, years	Gender	<i>FAM20A</i> mutations
1	21	male	c.915–918delCTTT; p.F305fsX380
2	27	female	IVS2+1G>A/c.913–914delTTT; p.F305fsX378
	31	male	IVS2+1G>A/c.913–914delTTT; p.F305fsX378
3	23	male	IVS4+1G>C/c.1348–1349delTTC; p.S450fsX469
	25	female	IVS4+1G>C/c.1348–1349delTTC; p.S450fsX469
4	59	male	c.1475–1482dupAACCCAC; p.L495fsX509
	64	female	c.1475–1482dupAACCCAC; p.L495fsX509
5	12	female	c.406C>T; p.R136X
6	20	male	c.34–35delCT; p.L12fsX78
7	16	female	c.1513delA; p.I505fsX506
	22	male	c.1513delA; p.I505fsX506
8	20	male	c.1432C>T; p.R478X
9	13	male	c.518T>G; p.L173R
10	29	female	c.727C>T/c.1228–1229delGA; p.R243X/p.D410fsX414
11	19	female	c.217C>T/c.727C>T; p.R73X/p.R243X
	20	male	c.217C>T/c.727C>T; p.R73X/p.R243X
12	18	female	c.1369A>T; p.K457X
13	14	female	c.755–757delTCT/c.641–719del79bp; p.F252del/p.I214fsX259
	16	male	c.755–757delTCT/c.641–719del79bp; p.F252del/p.I214fsX259
14	21	female	IVS5+2T>G
15	24	male	c.907–908delAG; p.S303fsX378
	31	male	c.907–908delAG; p.S303fsX378
	37	female	c.907–908delAG; p.S303fsX378
16	17	female	c.34–35delCT/c.612delC; p.L12fsX78/p.A204fsX215
	18	female	c.34–35delCT/c.612delC; p.L12fsX78/p.A204fsX215

Mutations are described on the cDNA and predicted protein levels. Listing of one allele indicates homozygosity; two alleles indicate compound heterozygosity. Every patient had biallelic mutations involving insertions, deletions, essential splice sites, missense changes or nonsense changes.

Analysis Toolkit were used to further process the alignment files for variant calling [6, 7].

For confirmation of variants detected by exome capture/next-generation sequencing and for mutation detection in additional cases, we amplified all coding exons and exon-intron boundaries of *FAM20A* using standard PCR methodology with intronic (genomic) primers. PCR products were separated on 1% agarose gels with ethidium bromide using electrophoresis and visualized under UV light. Specific bands were cut and DNA was isolated and purified using standard procedures. Bi-directional sequencing of all exons and exon-intron boundaries were performed using a Beckman Coulter CEQ8000 or an Applied Biosystems 3130xl capillary sequencer per the manufacturer's protocol. Sequencing data was analyzed and compared with the published reference sequence for *FAM20A* (NG_029809, April 2012, NCBI build 37.3).

Results

We ascertained 25 patients (12 males, 13 females; age 12–64) in 16 families with NC and characteristic dental findings, i.e., the triad of AI, gingival thickening and impairment of tooth eruption (table 1). The diagnosis of NC was made predominantly by nephrologists based upon characteristic imaging findings. Generalized hypoplastic AI was evident from eruption of the deciduous teeth early in childhood with subsequent impaired eruption of the permanent teeth and development of gingival enlargement. Clinical details of some patients have been published [8–11]. None of the parents or offspring of our patients had AI.

From the 16 families with this disorder, we selected four informative families (fig. 1A) for whole-genome

Discussion

Calcium plays many critical roles in human physiology, serving as an intracellular messenger, an extracellular neuromuscular excitatory ion, and a structural component of bone and teeth. For example, tooth enamel (calcium hydroxyapatite) is the hardest human tissue and can function into old age, despite being the only mineralized tissue with no capacity for cellular repair. Hence, the tempo-spatial concentration of calcium must be exquisitely regulated in different compartments, bound to albumin within the circulation, sequestered by calbindin within cells, including teeth [12], and excreted under tight control by the kidney [13]. Within the vasculature, calcium availability is closely modulated by an intricate interplay between bones and regulatory hormones that involves positive and negative feedback mechanisms (e.g., vitamin D, PTH, calcitonin) [14–15]. Consequences of dysregulated calcium homeostasis include nephrolithiasis and NC, i.e., precipitation of calcium in the urinary collecting system and renal interstitium, respectively.

The processes that normally maintain calcium balance in renal tissues can be delineated by studying genetic disorders involving NC. A prime example lies in an autosomal recessive disorder of NC combined with AI. Linkage analysis and focused exome sequencing performed on affected families identified the causative gene as *FAM20A*, previously associated only with a disorder of AI [16, 17].

Previous causes of NC have involved epithelial and paracellular disturbances in calcium transport, predominantly caused by mutations in calcium-specific channels and proteins [18–20]. These systems either reabsorb filtered calcium from the filtrate (urine) across the renal tubular cell or release calcium from the tubular cell into the interstitial compartment. When they malfunction, increased urinary calcium precipitates within the renal tubule (leading to nephrolithiasis and urolithiasis) or within the interstitium (leading to NC); this is invariably accompanied by hypercalciuria.

Our 25 patients exhibited NC and AI, but ascertainment of other patients with biallelic *FAM20A* mutations

will expand the phenotype and define the entire spectrum of this rare disease. Also, two different *Fam20a* knock-out mouse models exhibited different findings; in one, no kidney findings were reported, whereas the other showed arterial calcification without NC [21, 22]. The patients previously reported with *FAM20A* mutations and AI may also prove to have NC.

Conclusions

Our findings have implications for diagnosis and treatment. *FAM20A* is a locally secreted protein with low abundance in saliva [23] and blood [24], suggesting that replacement therapy could provide an option for a treatment or prophylaxis.

Acknowledgements

The authors thank all participating patients and families. We dedicate this work to the late Professor Oliver Wrong, who remained actively involved in these studies from their inception in the 1990s until his passing in February 2012, and whom we miss dearly. Funding for this study was kindly provided by the David and Elaine Potter Charitable Foundation (to R.K.), St Peter's Trust for Kidney, Bladder and Prostate Research (to D. Bockenhauer, R.J.U., R.K.), Kids Kidney Research (to D. Bockenhauer, R.K.), the Oxalosis & Hyperoxaluria Foundation (to R.K.), the European Union, FP7 EURenOMICS grant agreement No. 305608 (to D. Bockenhauer, R.J.U., R.K.), the Intramural Research Program of the National Human Genome Research Institute (to W.A.G.), the MEXT-Supported Program for the Strategic Research Foundation at Private Universities, 2012–2016 (to H.K.), the Wellcome Trust (to A.M., C.I.), the Sir Jules Thorn Charitable Trust (to C.J., C.I.), the University of Strasbourg (to A.B.-Z.), the French Ministry of Health National Program for Clinical Research (to A.B.-Z.), the Hôpitaux Universitaires de Strasbourg (to A.B.-Z.), the Institut Français pour la Recherche Odontologique (IFRO) (to A.B.-Z.), INSERM (to M.Q., A.B.), the Paris-Descartes University (to M.Q.), the French Ministry of Health and AHP Rothschild Hospital (to M.D.M.), the Paris-Diderot University (to M.D.M., A.A., A.B.), the F.R.S.-FNRS (Fonds de la Recherche Scientifique) (to M.V.), the Minas Gerais State Research Foundation (FAPEMIG) (to H.M.-J.), CAPES Brazil (to A.C.A.), and Deutsche Forschungsgemeinschaft (Ci 107/4-2) (to E.S.).

References

- 1 Evan AP, Unwin RJ, Williams JC Jr: Renal stone disease: a commentary on the nature and significance of Randall's plaque. *Nephron Physiol* 2011;119:p49–p53.
- 2 Kleit R, Bockenhauer D: Bartter syndromes and other salt-losing tubulopathies. *Nephron Physiol* 2006;104:p73–p80.
- 3 Vivante A, Lotan D, Pode-Shakked N, Landau D, Svec P, Nampoothiri S, Verma I, Abu-Libdeh A, Bockenhauer D, Dekel B, Anikster Y: Familial autosomal recessive renal tubular acidosis: importance of early diagnosis. *Nephron Physiol* 2011;119:p31–p39.

- 4 Bockenbauer D, Feather S, Stanescu HC, Bandulik S, Zdebik AA, Reichold M, Tobin J, Lieberer E, Sterner C, Landouze G, Arora R, Sirimanna T, Thompson D, Cross JH, van't Hoff W, Al Masri O, Tullus K, Yeung S, Anikster Y, Klootwijk E, Hubank M, Dillon MJ, Heitzmann D, Arcos-Burgos M, Knepper MA, Dobbie A, Gahl WA, Warth R, Sheridan E, Kleta R: Epilepsy, ataxia, sensorineural deafness, tubulopathy, and KCNJ10 mutations. *N Engl J Med* 2009;360:1960–1970.
- 5 Carr IM, Sheridan E, Hayward BE, Markham AF, Bonthron DT: IBDfinder and SNPsetter: tools for pedigree-independent identification of autozygous regions in individuals with recessive inherited disease. *Hum Mutat* 2009;30:960–967.
- 6 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
- 7 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
- 8 Dellow EL, Harley KE, Unwin RJ, Wrong O, Winter GB, Parkins BJ: Amelogenesis imperfecta, nephrocalcinosis, and hypocalciuria syndrome in two siblings from a large family with consanguineous parents. *Nephrol Dial Transplant* 1998;13:3193–3196.
- 9 Paula LM, Melo NS, Silva Guerra EN, Mes-trinho DH, Acevedo AC: Case report of a rare syndrome associating amelogenesis imperfecta and nephrocalcinosis in a consanguineous family. *Arch Oral Biol* 2005;50:237–242.
- 10 Martelli-Junior H, dos Santos Neto PE, de Aquino SN, de Oliveira Santos CC, Borges SP, Oliveira EA, Lopes MA, Coletta RD: Amelogenesis imperfecta and nephrocalcinosis syndrome: a case report and review of the literature. *Nephron Physiol* 2011;118:p62–p65.
- 11 Hall RK, Phakey P, Palamara J, McCredie DA: Amelogenesis imperfecta and nephrocalcinosis syndrome. Case studies of clinical features and ultrastructure of tooth enamel in two siblings. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 1995;79:583–592.
- 12 Berdal A, Hottot D, Pike JW, Mathieu H, Dupret JM: Cell- and stage-specific expression of vitamin D receptor and calbindin genes in rat incisor: regulation by 1,25-dihydroxyvitamin D3. *Dev Biol* 1993;155:172–179.
- 13 Bleich M, Shan Q, Himmerkus N: Calcium regulation of tight junction permeability. *Ann NY Acad Sci* 2012;1258:93–99.
- 14 Goldsmith DJ, Cunningham J: Mineral metabolism and vitamin D in chronic kidney disease – more questions than answers. *Nat Rev Nephrol* 2011;7:341–346.
- 15 Loupy A, Ramakrishnan SK, Wootla B, Chambrey R, de la Faille R, Bourgeois S, Bruneval P, Mandet C, Christensen EI, Faure H, Cheval L, Laghmani K, Collet C, Eladari D, Dodd RH, Ruat M, Houillier P: PTH-independent regulation of blood calcium concentration by the calcium-sensing receptor. *J Clin Invest* 2012;122:3355–3367.
- 16 O'Sullivan J, Bitu CC, Daly SB, Urquhart JE, Barron MJ, Bhaskar SS, Martelli-Junior H, dos Santos Neto PE, Mansilla MA, Murray JC, Coletta RD, Black GC, Dixon MJ: Whole-exome sequencing identifies FAM20A mutations as a cause of amelogenesis imperfecta and gingival hyperplasia syndrome. *Am J Hum Genet* 2011;88:616–620.
- 17 Cho SH, Seymen F, Lee KE, Lee SK, Kweon YS, Kim KJ, Jung SE, Song SJ, Yildirim M, Bayram M, Tuna EB, Gencay K, Kim JW: Novel FAM20A mutations in hypoplastic amelogenesis imperfecta. *Hum Mutat* 2012;33:91–94.
- 18 Shan Q, Himmerkus N, Hou J, Goodenough DA, Bleich M: Insights into driving forces and paracellular permeability from claudin-16 knockdown mouse. *Ann NY Acad Sci* 2009;1165:148–151.
- 19 Dimke H, Hoenderop JG, Bindels RJ: Molecular basis of epithelial Ca²⁺ and Mg²⁺ transport: insights from the TRP channel family. *J Physiol* 2011;589:1535–1542.
- 20 Breiderhoff T, Himmerkus N, Stuijver M, Mutig K, Will C, Meij IC, Bachmann S, Bleich M, Willnow TE, Muller D: Deletion of claudin-10 (Cldn10) in the thick ascending limb impairs paracellular sodium permeability and leads to hypermagnesemia and nephrocalcinosis. *Proc Natl Acad Sci USA* 2012;109:14241–14246.
- 21 An C, Ide Y, Nagano-Fujii M, Kitazawa S, Shoji I, Hotta H: A transgenic mouse line with a 58-kb fragment deletion in chromosome 11E1 that encompasses part of the Fam20a gene and its upstream region shows growth disorder. *Kobe J Med Sci* 2010;55:E82–E92.
- 22 Vogel P, Hansen GM, Read RW, Vance RB, Thiel M, Liu J, Wronski TJ, Smith DD, Jeter-Jones S, Brommage R: Amelogenesis imperfecta and other biomineralization defects in Fam20a and Fam20c null mice. *Vet Pathol* 2012;49:998–1017.
- 23 Gonzalez-Begne M, Lu B, Han X, Hagen FK, Hand AR, Melvin JE, Yates JR: Proteomic analysis of human parotid gland exosomes by multidimensional protein identification technology (MudPIT). *J Proteome Res* 2009;8:1304–1314.
- 24 Q an WJ, Kaleta DT, Petritis BO, Jiang H, Liu T, Zhang X, Mottaz HM, Varnum SM, Camp DG, 2nd, Huang L, Fang X, Zhang WW, Smith RD: Enhanced detection of low abundance human plasma proteins using a tandem IgY12-SuperMix immunoaffinity separation strategy. *Mol Cell Proteomics* 2008;7:1963–1973.

Appendix E: Paper V

□□□□□□□□ □□□□□□□□ □□ □□□□ □□□□□ □□□□ □Robinette SL, Lindon J, Nicholson JK.
Statistical Spectroscopic Tools for Biomarker Discovery and Systems Medicine.*Anal
Chem*, 2013. 85(11):5297-303.) □□□□□□□□□□□□□□ □□□□□ □□□□ □□□□□□□□□□

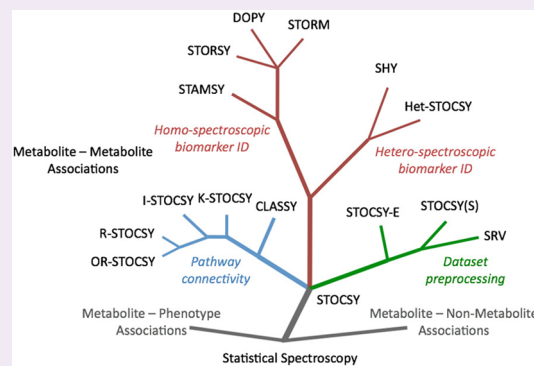
Statistical Spectroscopic Tools for Biomarker Discovery and Systems Medicine

Steven L. Robinette, John C. Lindon, and Jeremy K. Nicholson*

Computational and Systems Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, U.K.

ABSTRACT: Metabolic profiling based on comparative, statistical analysis of NMR spectroscopic and mass spectrometric data from complex biological samples has contributed to increased understanding of the role of small molecules in affecting and indicating biological processes. To enable this research, the development of statistical spectroscopy has been marked by early beginnings in applying pattern recognition to nuclear magnetic resonance data and the introduction of statistical total correlation spectroscopy (STOCSY) as a tool for biomarker identification in the past decade. Extensions of statistical spectroscopy now compose a family of related tools used for compound identification, data preprocessing, and metabolic pathway analysis. In this Perspective, we review the theory and current state of research in statistical spectroscopy and discuss the growing applications of these tools to medicine and systems biology. We also

provide perspectives on how recent institutional initiatives are providing new platforms for the development and application of statistical spectroscopy tools and driving the development of integrated “systems medicine” approaches in which clinical decision making is supported by statistical and computational analysis of metabolic, phenotypic, and physiological data.



The connection between the chemistry of tissues and biological fluids and health has been recognized since the beginning of medicine as a discipline. The color, smell, and taste of urine, for example, were used as diagnostics by physicians from ancient Greece through to the Middle Ages.¹ While metabonomics,² and the related field of metabolomics,³ have replaced human chemosensation with modern analytical chemistry techniques, the basic insight that chemical patterns are indicative of biological processes underlying both physiological and pathological states remains the same. Several terms are used interchangeably and employ essentially the same technologies and procedures: metabonomics, metabolomics, metabolic profiling, and metabolic phenotyping. Metabonomics employs a so-called top down systems biology philosophy and is defined as “the quantitative measurement of the dynamic multi-parametric metabolic response of living systems to pathophysiological stimuli or genetic modification”.²

Technological developments in nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) have played a key role in facilitating the large-scale measurement of low-molecular weight metabolites.⁴ While these techniques feature varying strengths in coverage, sensitivity, selectivity for various chemical classes, reproducibility, provision of structural information, and sample preparation requirements, both stand out in their capacity to measure a large number of small molecule analytes in an untargeted fashion from complex biological mixtures such as human biofluids.⁵ While the application of these technologies to characterize human biochemical matrices including biofluids and tissues is not new, the combination of untargeted profiling with ‘omics experimental designs and

multivariate statistical pattern recognition in metabonomics and metabolomics has led to significant improvements in understanding of the role of small molecules in controlling and marking biological processes in toxicology,⁶ microbiology,⁷ and in human health and disease.⁸ Protocols for these techniques, including sample preparation, acquisition parameters, and resulting information content have been published previously.^{9–13}

The development of statistical spectroscopy¹⁴ and the ongoing refinement and extension of statistical spectroscopy tools are significantly increasing the utility of the metabonomics approach as a general-purpose tool in analytical chemistry. Over 20 years ago, Noda developed the concept and theory of statistically correlating peak intensities across sets of spectra and applied it to IR data.^{15–17} In a later development, Brüschweiler and co-workers published the application to the rows or columns of a 2D NMR spectrum and termed it covariance spectroscopy.^{18–21} While chemical identification has historically been dependent on purification and separation sciences, the combination of biological replicates and correlation based approaches for NMR and MS together with improved compound databases are increasingly enabling the identification of unknown compounds and even novel chemical structures without prerequisite isolation.²² These approaches are being extended to support the integration of metabolic profiles with parallel data sets from genomics and proteomics, yielding a framework for “global” systems biology in which organism level

Received: March 9, 2013

Accepted: April 24, 2013

Published: April 24, 2013



human biological networks are mapped by identifying associations between human genes, gut microflora, proteins, and metabolites.²³ Additionally, the combination of metabolic profiling with statistical spectroscopy and predictive modeling is increasingly being incorporated into clinical practice, both as diagnostics, to monitor patients, and to support clinical decision-making.^{24–26}

To date, metabonomics has significantly impacted the practice of toxicology,⁶ microbiology,⁷ drug discovery,²⁷ and the study of metabolic regulation and disease.⁸ In this Perspective, we review the origins and theory of statistical spectroscopy, major milestones in the development of the field, and survey the landscape of statistical spectroscopic tools currently available. We then discuss how the framework of statistical spectroscopy is being applied to global systems biology through the identification of gene-metabolite and other systems level associations. Just as the development of systems biology has advanced research in the life sciences, we are currently seeing an analogous development of “systems medicine”, the integrated study of system level metabolic, phenotypic, and physiological changes in response to disease processes or therapies by application of computational and statistical approaches to support clinical decisions. We discuss applications of statistical spectroscopy toward the development of a systems medicine, especially in surgical metabonomics and patient journey monitoring. We conclude with a perspective on how statistical spectroscopy will continue to contribute to analytical chemistry and major institutional initiatives that will advance this science in the near term.

Statistical Spectroscopy As a Tool for Biomarker Discovery and Disease Modeling. Early metabonomics studies in which pattern recognition was applied to NMR spectra frequently noted complex, coordinated multianalyte responses between biological states and through time.²⁸ The application of dimensionality reduction techniques such as principal component analysis (PCA) and later supervised learning methods including partial least-squares (PLS) and orthogonal projection onto latent

structures (OPLS)²⁹ spurred an interest in using metabolic profiles as a multivariate descriptor of an underlying biological state, resulting in now widespread application of chemometric tools in metabonomics.^{30,31} While individual metabolites were often assigned in representative spectra, discussions involving the application of pattern recognition in metabonomics frequently highlighted the ability to distinguish spectra from two or more biological states based on differences between metabolic profiles and of “metabolic trajectories” over time.³²

While these methods focused on identifying correlations between signal and state (phenotype), the introduction of new methods in the past decade significantly enhanced metabolite identification by identifying correlations between signals. 2D Fourier transform NMR, 2D-correlation IR spectra,^{15–17} and covariance NMR^{18–21} identify correlations between spectroscopic signals by making use of systematic variation in a sequence of one-dimensional spectra to improve chemical characterization, but these methods all make use of measurements collected from a single sample. With the introduction of statistical total correlation spectroscopy (STOCSY),^{14,33} this approach was extended to identify covariance in peak intensities across multiple, independent samples that can generate associations between signals rising from the same molecular structure independent of through-bond or through-space connectivity (Figure 1). While STOCSY is typically applied to identify signal-signal rather than signal-state correlations, like most statistical spectroscopy tools it can be combined with chemometrics as a two stage process as was noted in the original presentation.¹⁴ The theoretical framework of STOCSY (see Box 1) was recognized to be flexibly extendable beyond ¹H NMR both to other metabolic profiling techniques and to multiple ‘omics technologies. This recognition has led to sustained development of a diverse set of tools in statistical spectroscopy (Table 1). Statistical spectroscopy is now composed of a family of related tools that perform a range of functions in metabonomic research (Figure 2). While initial applications of STOCSY sought to

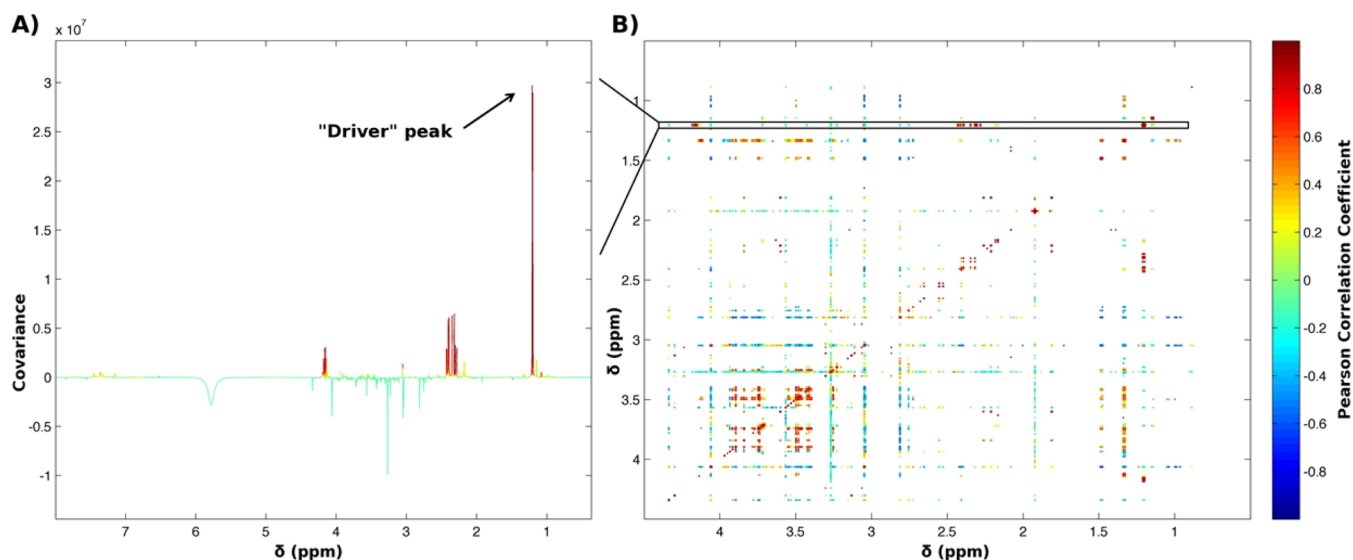


Figure 1. Statistical total correlation spectroscopy (STOCSY): (A) one-dimensional (1D) and (B) two-dimensional (2D) STOCSY correlation/covariance plots of a urinary NMR data set. Here, lineshapes are calculated by covariance ($[-\infty, \infty]$) while colors are set by correlation ($[-1, 1]$). One-dimensional STOCSY plots (A) are traces of the two-dimensional STOCY (B, black box) at a given chemical shift taken from the “driver” peak. For the maxima of the doublet of 3-hydroxybutyric acid at 1.2 ppm (“driver” peak), high positive correlation coefficients are both sensitive and specific indicators of structural connectivity, and “pathway” connectivities show mostly negative correlations. For other metabolite signals seen in the 2D STOCSY, high positive correlations such as those between lactate (1.33 ppm), alanine (1.48 ppm), and glucose (~ 3.5 –4 ppm) indicate coordinated excretion rather than structural relationships.

Box 1. Theory of Statistical Spectroscopy

$$r_{mn} = \frac{\sum_{i=1}^j (x_{im} - \bar{x}_m)(x_{in} - \bar{x}_n)}{(j-1) \cdot \sigma_{x_m} \sigma_{x_n}} \quad (1)$$

$$R = \frac{1}{j-1} X_1^t X_2 \quad (2)$$

$$R = \frac{1}{j-1} X_1^t Y \quad (3)$$

In eq 1, r_{mn} is the correlation coefficient between the intensities x of signal m and signal n across all spectra i through j , where these signals have mean intensity $\bar{x}_m$ and $\bar{x}_n$ and standard deviations σ_{x_m} and σ_{x_n} . When signals m and n arise from the same chemical structure, r_{mn} will tend to approach unity. When signals m and n arise from different chemical structures, the magnitude of r_{mn} will depend on the extent to which the metabolites giving rise to signals m and n are subject to the same metabolic perturbations (e.g., pathway connectivity). As a result, the correlation coefficient r_{mn} can be used to assess both structural and pathway connectivity, though in the case of strong metabolic perturbations or poor signal resolution there may be ambiguity between the underlying origin of correlations.

Equation 2 is mathematically equivalent to eq 1, but the matrix notation emphasizes differences between data sets rather than signals. Here, R is a matrix of correlation coefficients between two autoscaled data sets X_1 and X_2 , each of which is composed of j observations. When X_1 and X_2 are the same matrix, R is homospectroscopic. When X_1 and X_2 are matrices derived from different analytical techniques, such as heteronuclear NMR or mass spectrometry, R is heterospectroscopic. The framework of statistical spectroscopy is flexibly extended in cases in which X_2 is data derived from genomics, proteomics, or transcriptomics as well. In these cases R will reflect metabolite–nonmetabolite associations as opposed to the traditional metabolite–metabolite associations. When X_2 is not a second data set but is instead a vector of phenotypic traits Y (eq 3), R reflects metabolite–phenotype associations.

identify structural relationships between 1D ^1H NMR signals, offshoots quickly appeared applying and modifying the STOCSY methodology to preprocess spectroscopic data sets, to reveal structural relationships between multiple analytical platforms and to identify coregulation resulting from biological and pathway relationships between metabolites. Figure 2 maps out a statistical spectroscopy “family tree” outlining these various techniques and their primary applications.

One of the major drivers of the development of new tools in statistical spectroscopy is the ability of these tools to detect both associations across spin-systems that are not visible using traditional correlation NMR spectroscopy and weaker “pathway” correlations between compounds of interest and related compounds whose expression is modulated by associated biological pressures.³³ One of the initial key challenges in statistical spectroscopy involved differentiating between these two sources of correlations. As a result, substantial effort has been applied to characterizing the range of correlations from within-compound “structural” associations and between-compound “biological” correlations.³⁴ Extensions of statistical spectroscopy beyond biomarker identification have sought to

make use of these properties to (1) improve differentiation of structural correlations by local clustering, subset selection, and stoichiometric relationships,^{35–37} (2) query networks of biological coregulation of metabolites,^{35,38–40} and (3) improve preparatory processing of spectroscopic data for subsequent statistical analysis.^{41–43}

Statistical spectroscopy is flexibly applied to nearly any analytical technique that results in quantitative or semiquantitative measurements of molecular features. Because of this flexibility, statistical spectroscopy versions of many correlation spectroscopy experiments have been formulated, including Het-STOCSY (heteronuclear STOCSY) for ^1H – ^{31}P ⁴⁴ and ^1H – ^{19}F ⁴⁵ correlations, diffusion-ordered projection spectroscopy (DOPY),⁴⁶ and statistical heterospectroscopy (SHY) to identify associations between NMR and MS data,⁴⁷ which often prove challenging to identify using hyphenation techniques such as LC–NMR–MS. Advantages of these correlation techniques include higher resolution and sensitivity relative to direct experimental observation and an extended range of correlations, both positive and negative.

It is our perspective that many applications of analytical chemistry will benefit from recent advances in statistical spectroscopy. By combining the direct observation of complex mixtures with postprocessing to highlight and correlate signals of interest, statistical spectroscopy suggests a new, four-stage approach to biomarker and compound identification, represented schematically in Figure 3. In the first stage, spectral features are correlated with a given biological state or response by applying pattern recognition analysis to individual data sets generated by NMR spectroscopy or mass spectrometry (Figure 3A). Subsequently, STOCSY, SHY, and other statistical spectroscopy tools are then used to associate masses, fragments, and NMR spectral features (Figure 3B). In a third stage, the set of correlated features, including mass, retention time, fragmentation patterns, chemical shifts, and through-bond correlations are searched in-silico for matching compounds using metabolite libraries and reference spectral data (Figure 3C). While individual MS m/z values and single NMR chemical shifts are often non-specific, only a small number of structural isomers will give rise to similar NMR spectra and MS-MS fragments. To conclusively verify database matches, metabolite candidates of interest may be confirmed by direct spiking of pure chemical standards into the biological matrix for NMR or comparison of parent masses, retention times, and fragmentation patterns of pure standards by mass spectrometric analysis (Figure 3D). Several web-based computational tools to support biomarker and compound identification, including Metlin, XCMS Online, the BMRB and HMDB databases, and MetaboAnalyst, have been introduced in recent years,^{48–52} and additional support will be provided by continued integration of statistical spectroscopy and pattern recognition with spectral databases.

Applications of Statistical Spectroscopy in Systems Biology and in Medicine. High-throughput measurement and analysis of genes, transcripts, proteins, and others in addition to metabolites have enabled the simultaneous monitoring of many elements of biological networks. It is our perspective that statistical spectroscopy will play an important role in the development of top-down and “global” systems biological approaches.²³ These expand the framework of statistical spectroscopy to identify associations across multiple omics sciences (see Box 1). Examples include the development of metabolomic quantitative trait loci mapping (mQTL)^{53–55} and related metabolomic

Table 1

statistical spectroscopic tool	analysis platform	purpose	ref(s)
statistical total correlation spectroscopy (STOCSY)	$^1\text{H}/^1\text{H}$ NMR	assign biomarker chemical structure by identifying structural correlations in 1D NMR	14, 33
subset optimization by reference matching (STORM)	$^1\text{H}/^1\text{H}$ NMR	identify structural correlations from minor peaks in large NMR data sets	36
diffusion ordered projection spectroscopy (DOPY)	$^1\text{H}/^1\text{H}$ DOSY NMR	improve resolution of correlated signals	46
statistical total regression spectroscopy (STORSY)	$^1\text{H}/^1\text{H}$ JRES NMR	identify stoichiometric relationships between NMR signals	37
statistical total correlation mass spectrometry (STAMSY)	LC–MS/LC–MS	identify isotope/adduct correlations in mass spectrometry	
statistical heterospectroscopy (SHY)	^1H NMR/LC–MS	identify corresponding signals in NMR and MS	47
heteronuclear STOCSY (Het-STOCSY)	$^1\text{H}/^{19}\text{F}$, $^1\text{H}/^{31}\text{P}$ NMR	correlation of heteronuclear NMR signals	44, 45
cluster analysis statistical spectroscopy (CLASSY)	$^1\text{H}/^1\text{H}$ NMR	distinguish structural/pathway correlations and identify metabolic perturbations	35
kinetic STOCSY (K-STOCSY)	$^1\text{H}/^1\text{H}$ NMR	follow sequences of biotransformations on short time scales	76
iterative STOCSY (I-STOCSY)	$^1\text{H}/^1\text{H}$ NMR	trace network of pathway correlations	38
recoupled STOCSY (R-STOCSY)	$^1\text{H}/^1\text{H}$ NMR	identify sets of pathway relationships in perturbed metabolic networks	40
orthogonal filtered recoupled STOCSY (OR-STOCSY)	$^1\text{H}/^1\text{H}$ NMR	identify metabolites in weakly perturbed networks	39
STOCSY-editing (STOCSY-E)	$^1\text{H}/^1\text{H}$ NMR	remove signals from exogenous compounds	42
STOCSY-scaling (STOCSY(S))	$^1\text{H}/^1\text{H}$ NMR	scale statistical variance of compound signals	43
statistical recoupling of variables (SRV)	$^1\text{H}/^1\text{H}$ NMR	adaptive binning of signals to reduce dimensionality	41

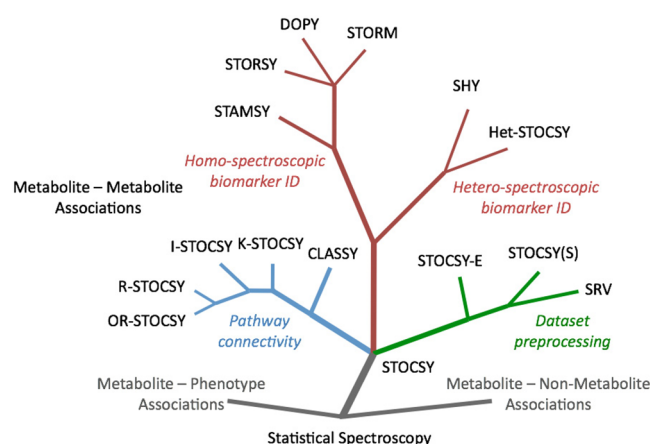


Figure 2. Statistical spectroscopy family tree. Statistical spectroscopy is based on the quantitative correlation of spectral signals derived from analytical chemistry data and can be applied to identify metabolite–metabolite associations, metabolite–nonmetabolite associations, and metabolite–phenotype associations. STOCSY and its extended family of methodologies are typically applied to assign chemical structures, prepare spectroscopic data sets for pattern recognition analysis, and identify coregulation between metabolites. Branches in the tree show different statistical spectroscopic tools and indicate the major applications of these tools. Additional information about individual tools and relevant citations are shown in Table 1.

genome wide association studies (mGWAS),^{56–59} metabolome–microbiome associations, and work integrating genome, transcriptome, and metabolome variation to identify functional networks governing human homeostasis.⁶⁰

Developments in genomics and metabonomics have now enabled the large-scale profiling of both genetic variants and metabolic profiles necessary to draw firm statistical associations between genes and metabolites. As statistical power grows with sample sizes, the relatively small set of known associations between single nucleotide polymorphisms (SNPs) and biofluid metabolite concentrations identified by NMR or MS is growing rapidly. New mQTL/mGWAS publications are both replicating and expanding the set of known gene–metabolite associations.⁵⁸

Many of these discovered associations make immediate biological sense. Many “cis” associations between polymorphisms in metabolic enzymes and adjacent metabolites (urine glycerate and glycerate kinase⁵⁴) and transporters and transporter substrates (serum urate and SLC2A9 urate transporter⁵⁸) are immediately comprehensible. Other “pathway” associations can be more difficult to explain at first but are no less informative. These associations are likely to annotate unknown gene functions and help explain human metabolic networks in both physiological and pathological states. The nearly simultaneous discovery of a statistical association between the *NTSE* gene, encoding a vascular epithelial enzyme, and serum inosine concentration⁵⁸ and the identification of a human vascular calcification phenotype cause by rare mutations in the *NTSE* gene and resulting disruption of adenosine metabolism⁶¹ indicates the potential power of this approach.

In addition to mammalian genomic determinants of metabolism, the global systems biology approach enables the identification of gut microbiota impacts on host metabolism. Sensitivity to intestinal microbes is especially important for organism-level networks as our microbiota includes over a thousand species of symbionts with a total genetic contribution of probably more than 100 times the ~20 000 predicted human genes. This “second genome” coordinates extensively with host metabolic and physiological systems, especially the liver and immune system.⁶² Associations of microbe populations with metabolites have been enabled by semiquantitative characterizations of bacterial species using techniques such as fluorescent in situ hybridization (FISH) and denaturing gradient gel electrophoresis (DGGE) and have been used to identify associations between bacterial populations, related metabolites, and obesity.⁶³

While extensions of statistical spectroscopy to nonmetabolite data sets is a promising area of research, weaker associations between levels of biological organization (e.g., gene regulation and metabolite abundance) require increased statistical power to detect. Population-level studies are increasingly necessary to identify significant associations in top-down systems biology. The recently established MRC-NIHR National Phenome Centre in the U.K. will play an important role in developing the scale necessary to do this kind of science.

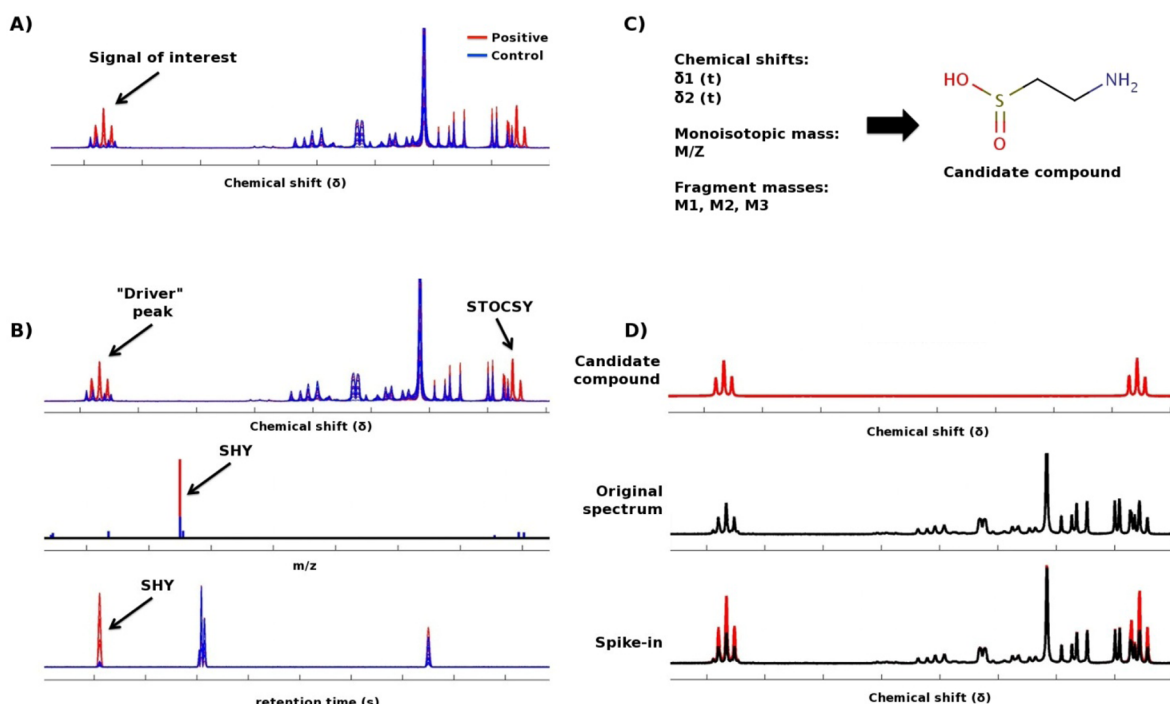


Figure 3. Strategy for unknown compound identification using statistical spectroscopy. Signals of interest are identified from an experimental data set (a). Statistical spectroscopy tools, such as STOCSY and SHY, are used to identify correlations between the signal of interest and additional signals in the same or other spectral data sets collected from the same set of samples (b). Candidate compounds are identified using database search tools to match the set of correlated spectral features with structures (c). A candidate compound is confirmed by spiking a solution of the candidate into the original mixture (d).

While metabolomics continues to promise a significant potential for translating biomarkers identified by association studies into clinical tests, the advent of “surgical metabolomics” has seen significant contributions in identifying downstream biochemical consequences of surgical interventions and in tracking metabolic perturbations and subsequent biochemical journeys toward recovery.^{64,65} Metabolomics has been applied to study experimental models of bariatric surgery,⁶⁶ laparotomy,⁶⁷ ischemia/reperfusion,⁶⁸ and the identification of complex impacts on gut microflora from these interventions. Growing appreciation for the utility of metabolomics in patient monitoring following surgery has led to efforts toward applying metabolic profiling to analyze “patient journeys” from disease through surgical interventions and subsequent recovery, applications and prospects of which have been recently reviewed.^{65,69}

Application of metabolomic monitoring real-time during surgery is also a growing area of interest. Development of “intelligent knife” technology coupling molecular ion generating surgical techniques, such as electrocautery, laser surgery, or ultrasonic aspiration, with mass spectrometry has shown significant promise for the rapid identification and guided excision of pathological tissues.^{26,70} Here, mass spectrometry acts as a biochemical sensor, sampling tissues *in situ* to provide real time information guiding surgical decisions. Ionization conditions under electrosurgery or laser surgery generally lead to loss of chemical stability and the mass spectrometry instrumentation suitable for *in-theater* use may not provide sufficient resolution or separation to conclusively identify individual biomolecules.⁷¹ However, rather than attempting to identify individual markers and make tissue classification decisions based on levels of these markers, classification of real-time mass spectra is done on the

basis of pattern recognition of spectral features, biochemical “fingerprints”.^{72,73}

Proof-of-principle demonstrations for the intelligent-knife technology began with sampling and analysis of nonhuman, *ex vivo* tissues.^{70–72} This early work demonstrated classification of tissue types and various tumors is possible on the basis of mass spectrometric fingerprints. Subsequent work identified brain tumors as a good candidate for early application of the intelligent knife, and several groups have reported *ex vivo* sampling and classification of human brain tissue and tumor types.^{25,74,75} Recent work has demonstrated surgical mass spectrometric differentiation of both gray matter and white matter from meningioma and glioblastoma⁷⁴ and distinguishing of multiple gliomas, including oligodendroglioma, astrocytoma, and oligoastrocytoma.²⁵

The recently established Imperial Clinical Phenotyping Centre will play an important role in extending surgical metabolomics, intelligent knife technology, and the framework of patient journey monitoring into clinical practice. Combining metabolic profile measurement over the course of the patient journey including diagnosis, intervention, and postintervention with statistical spectroscopy and predictive modeling will provide important information to guide clinical decision-making.

Future Directions for the Field. Direct, quantitative comparison of spectra generated by analytical chemistry techniques is a basic cornerstone of current practice in metabolomics/metabolomics. The application of pattern recognition techniques to spectral data has driven over 20 years of metabolomic science that has led to a renewed appreciation for the role of small molecules and transformed our approach to research in toxicology, functional genomics, microbiology, and pathology. In the past decade, the introduction and continued development

of statistical spectroscopy has further extended the reach of this approach to compound identification, providing rapid diagnostics in a clinical setting, and the study of biochemical organization at the systems level. Despite these advances, routine application of metabolic profiling in a nonresearch setting is hindered by the absence of a single analytical platform for metabolome analysis, the specialized knowledge required to interpret spectral data and identify biomarker chemical structures, and uncertainty regarding both interpretation of metabolome-wide data and the protection and management of resulting information. Advances in instrumentation, informatics, and analytics software will be necessary to resolve these outstanding challenges.

These challenges are not insurmountable, and in the U.K. the recently established MRC-NIHR National Phenome Centre and the Imperial Clinical Phenotyping Centre at St. Mary's Hospital in London will play an important role in translating the methodology of metabolic phenotyping and developing improved tools and procedures for the collection and analysis of metabolic profiles and related 'omics data to support clinical decision-making through the framework of systems medicine. We anticipate the day when comprehensive measurement of biological variables, including gene sequences, epigenetic modifications, expression and translation levels, and metabolite abundances are routine and this data is widely available. Developments in statistical spectroscopy and integration with the related fields of genomics and proteomics will lead to an important role for analytical chemistry in paving the way to this future.

AUTHOR INFORMATION

Corresponding Author

*E-mail: j.nicholson@imperial.ac.uk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

S.L.R. acknowledges support from the NSF Graduate Research Fellowship Program and from the Marshall Aid Commemoration Commission.

REFERENCES

- (1) Nicholson, J. K.; Lindon, J. C. *Nature* **2008**, *455*, 1054–1056.
- (2) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *29*, 1181–1189.
- (3) Raamsdonk, L. M.; Teusink, B.; Broadhurst, D.; Zhang, N.; Hayes, A.; Walsh, M. C.; Berden, J. A.; Brindle, K. M.; Kell, D. B.; Rowland, J. J.; Westerhoff, H. V.; van Dam, K.; Oliver, S. G. *Nat. Biotechnol.* **2001**, *19*, 45–50.
- (4) Dunn, W.; Ellis, D. *TrAC, Trends Anal. Chem.* **2005**, *24*, 285–294.
- (5) Dunn, W.; Bailey, N.; Johnson, H. *Analyst* **2005**, *130*, 606–625.
- (6) Nicholson, J. K.; Keun, H.; Ebbels, T. J. *Proteome Res.* **2007**, *6*, 4098–4099.
- (7) Kell, D. *Curr. Opin. Microbiol.* **2004**, *7*, 296–307.
- (8) Nicholson, J. K.; Holmes, E.; Elliott, P. J. *Proteome Res.* **2008**, *7*, 3637–3638.
- (9) Xia, J.; Wishart, D. S. *Nat. Protoc.* **2011**, *6*, 743–760.
- (10) 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.; Consortium THSMH. *Nat. Protoc.* **2011**, *6*, 1060–1083.
- (11) Chan, E. C. Y.; Pasikanti, K. K.; Nicholson, J. K. *Nat. Protoc.* **2011**, *6*, 1483–1499.
- (12) Want, E. J.; Wilson, I. D.; Gika, H.; Theodoridis, G.; Plumb, R. S.; Shockcor, J.; Holmes, E.; Nicholson, J. K. *Nat. Protoc.* **2010**, *5*, 1005–1018.
- (13) Beckonert, O.; Keun, H.; Ebbels, T.; Bundy, J.; Holmes, E.; Lindon, J.; Nicholson, J. K. *Nat. Protoc.* **2007**, *2*, 2692–2703.
- (14) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J.; Holmes, E.; Nicholson, J. *Anal. Chem.* **2005**, *77*, 1282–1289.
- (15) Noda, I. *Appl. Spectrosc.* **1993**, *47*, 1329–1336.
- (16) Noda, I. *Appl. Spectrosc.* **1990**, *44*, 550–561.
- (17) Noda, I. *J. Am. Chem. Soc.* **1989**, *111*, 8116–8118.
- (18) Zhang, F.; Brüsweiler, R. *J. Am. Chem. Soc.* **2004**, *126*, 13180–13181.
- (19) Trbovic, N.; Smirnov, S.; Zhang, F.; Brüsweiler, R. *J. Magn. Reson.* **2004**, *171*, 277–283.
- (20) Brüsweiler, R. *J. Chem. Phys.* **2004**, *121*, 409–414.
- (21) Brüsweiler, R.; Zhang, F. *J. Chem. Phys.* **2004**, *120*, 5253–5260.
- (22) Robinette, S. L.; Brüsweiler, R.; Schroeder, F. C.; Edison, A. S. *Acc. Chem. Res.* **2012**, *45*, 288–297.
- (23) Nicholson, J. K.; Wilson, I. D. *Nat. Rev. Drug Discovery* **2003**, *2*, 668–676.
- (24) Takats, Z.; Denes, J.; Kinross, J. *Future Oncol.* **2012**, *8*, 113–116.
- (25) Eberlin, L. S.; Norton, I.; Dill, A. L.; Golby, A. J.; Ligon, K. L.; Santagata, S.; Cooks, R. G.; Agar, N. Y. R. *Cancer Res.* **2012**, *72*, 645–654.
- (26) Schäfer, K. C.; Dénes, J.; Albrecht, K.; Szaniszló, T.; Balog, J.; Skoumal, R.; Katona, M.; Tóth, M.; Balogh, L.; Takáts, Z. *Angew. Chem., Int. Ed.* **2009**, *48*, 8240–8242.
- (27) Nicholson, J.; Connelly, J.; Lindon, J.; Holmes, E. *Nat. Rev. Drug Discovery* **2002**, *1*, 153–161.
- (28) Gartland, K. P.; Beddell, C. R.; Lindon, J. C.; Nicholson, J. K. *J. Pharm. Biomed. Anal.* **1990**, *8*, 963–968.
- (29) Bylesjo, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. *J. Chemom.* **2006**, *20*, 341–351.
- (30) Madsen, R.; Lundstedt, T.; Trygg, J. *Anal. Chim. Acta* **2010**, *659*, 23–33.
- (31) van der Greef, J.; Smilde, A. K. *J. Chemometr.* **2005**, *19*, 376–386.
- (32) Keun, H.; Ebbels, T.; Bollard, M.; Beckonert, O.; Antti, H.; Holmes, E.; Lindon, J.; Nicholson, J. *Chem. Res. Toxicol.* **2004**, *17*, 579–587.
- (33) Holmes, E.; Cloarec, O.; Nicholson, J. K. *J. Proteome Res.* **2006**, *5*, 1313–1320.
- (34) Alves, A. C.; Rantalainen, M.; Holmes, E.; Nicholson, J. K.; Ebbels, T. M. D. *Anal. Chem.* **2009**, *81*, 2075–2084.
- (35) Robinette, S. L.; Veselkov, K. A.; Bohus, E.; Coen, M.; Keun, H. C.; Ebbels, T. M. D.; Beckonert, O.; Holmes, E. C.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2009**, *81*, 6581–6589.
- (36) Poma, J. M.; Garcia-Perez, I.; De Iorio, M.; Lindon, J. C.; Elliott, P.; Holmes, E.; Ebbels, T. M. D.; Nicholson, J. K. *Anal. Chem.* **2012**, *84*, 10694–10701.
- (37) Fonville, J. M.; Maher, A. D.; Coen, M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2010**, *82*, 1811–1821.
- (38) Sands, C. J.; Coen, M.; Ebbels, T. M. D.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2011**, *83*, 2075–2082.
- (39) Blaise, B. J.; Navratil, V.; Emsley, L.; Toulhoat, P. J. *Proteome Res.* **2011**, *10*, 4342–4348.
- (40) Blaise, B. J.; Navratil, V.; Domange, C.; Shintu, L.; Dumas, M. E.; Elena-Herrmann, B.; Emsley, L.; Toulhoat, P. J. *Proteome Res.* **2010**, *9*, 4513–4520.
- (41) Blaise, B. J.; Shintu, L.; Elena, B.; Emsley, L.; Dumas, M. E.; Toulhoat, P. *Anal. Chem.* **2009**, *81*, 6242–6251.
- (42) Sands, C. J.; Coen, M.; Maher, A. D.; Ebbels, T. M. D.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2009**, *81*, 6458–6466.
- (43) Maher, A. D.; Fonville, J. M.; Coen, M.; Lindon, J. C.; Rae, C. D.; Nicholson, J. K. *Anal. Chem.* **2012**, *84*, 1083–1091.

- (44) Coen, M.; Hong, Y. S.; Cloarec, O.; Rhode, C.; Reily, M.; Robertson, D.; Holmes, E.; Lindon, J.; Nicholson, J. *Anal. Chem.* **2007**, *79*, 8956–8966.
- (45) Keun, H.; Athersuch, T.; Beckonert, O.; Wang, Y.; Saric, J.; Shockcor, J.; Lindon, J.; Wilson, I.; Holmes, E.; Nicholson, J. *Anal. Chem.* **2008**, *80*, 1073–1079.
- (46) Smith, L. M.; Maher, A. D.; Cloarec, O.; Rantalainen, M.; Tang, H.; Elliott, P.; Stamler, J.; Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Anal. Chem.* **2007**, *79*, 5682–5689.
- (47) Crockford, D.; Holmes, E.; Lindon, J.; Plumb, R.; Zirah, S.; Bruce, S.; Rainville, P.; Stumpf, C.; Nicholson, J. *Anal. Chem.* **2006**, *78*, 363–371.
- (48) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27*, 747–751.
- (49) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem.* **2012**, *84*, S035–S039.
- (50) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L. *Nucleic Acids Res.* **2008**, *36*, D402–408.
- (51) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeronci, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; MacInnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* **2007**, *35*, D521–D526.
- (52) Xia, J.; Psychogios, N.; Young, N.; Wishart, D. S. *Nucleic Acids Res.* **2009**, *37*, W652–W660.
- (53) Dumas, M. E.; Wilder, S. P.; Bihoreau, M. T.; Barton, R. H.; Fearnside, J. F.; Argoud, K.; D'Amato, L.; Wallis, R. H.; Blancher, C.; Keun, H. C.; Baunsgaard, D.; Scott, J.; Sidemann, U. G.; Nicholson, J. K.; Gauguier, D. *Nat. Genet.* **2007**, *39*, 666–672.
- (54) Cazier, J. B.; Kaisaki, P. J.; Argoud, K.; Blaise, B. J.; Veselkov, K.; Ebbels, T. M. D.; Tsang, T.; Wang, Y.; Bihoreau, M. T.; Mitchell, S. C.; Holmes, E. C.; Lindon, J. C.; Scott, J.; Nicholson, J. K.; Dumas, M. E.; Gauguier, D. *J. Proteome Res.* **2012**, *11*, 631–642.
- (55) Nicholson, G.; Rantalainen, M.; Li, J. V.; Maher, A. D.; Malmudin, D.; Ahmadi, K. R.; Faber, J. H.; Barrett, A.; Min, J. L.; Rayner, N. W.; Toft, H.; Krestyaninova, M.; Viksna, J.; Neogi, S. G.; Dumas, M. E.; Sarkans, U.; MolPAGE Consortium; Donnelly, P.; Illig, T.; Adamski, J.; Suhre, K.; Allen, M.; Zondervan, K. T.; Spector, T. D.; Nicholson, J. K.; Lindon, J. C.; Baunsgaard, D.; Holmes, E.; McCarthy, M. I.; Holmes, C. C. *PLoS Genet.* **2011**, *7*, e1002270.
- (56) Gieger, C.; Geistlinger, L.; Altmaier, E.; de Angelis, M. H.; Kronenberg, F.; Meitinger, T.; Mewes, H. W.; Wichmann, H. E.; Weinberger, K. M.; Adamski, J.; Illig, T.; Suhre, K. *PLoS Genet.* **2008**, *4*, e1000282.
- (57) Illig, T.; Gieger, C.; Zhai, G.; Römisch-Margl, W.; Wang-Sattler, R.; Prehn, C.; Altmaier, E.; Kastenmüller, G.; Kato, B. S.; Mewes, H. W.; Meitinger, T.; de Angelis, M. H.; Kronenberg, F.; Soranzo, N.; Wichmann, H. E.; Spector, T. D.; Adamski, J.; Suhre, K. *Nat. Genet.* **2010**, *42*, 137–141.
- (58) Suhre, K.; Shin, S. Y.; Petersen, A. K.; Mohney, R. P.; Meredith, D.; Wägele, B.; Altmaier, E.; CARDIoGRAM; Deloukas, P.; Erdmann, J.; Grundberg, E.; Hammond, C. J.; de Angelis, M. H.; Kastenmüller, G.; Köttgen, A.; Kronenberg, F.; Mangino, M.; Meisinger, C.; Meitinger, T.; Mewes, H. W.; Milburn, M. V.; Prehn, C.; Raffler, J.; Ried, J. S.; Römisch-Margl, W.; Samani, N. J.; Small, K. S.; Wichmann, H. E.; Zhai, G.; Illig, T.; Spector, T. D.; Adamski, J.; Soranzo, N.; Gieger, C. *Nature* **2011**, *477*, 54–60.
- (59) Suhre, K.; Wallaschofski, H.; Raffler, J.; Friedrich, N.; Haring, R.; Michael, K.; Wasner, C.; Krebs, A.; Kronenberg, F.; Chang, D.; Meisinger, C.; Wichmann, H. E.; Hoffmann, W.; Völzke, H.; Völker, U.; Teumer, A.; Biffar, R.; Kocher, T.; Felix, S. B.; Illig, T.; Kroemer, H. K.; Gieger, C.; Römisch-Margl, W.; Nauck, M. *Nat. Genet.* **2011**, *43*, 565–569.
- (60) Inouye, M.; Kettunen, J.; Soininen, P.; Silander, K.; Ripatti, S.; Kumpula, L. S.; Hämmäläinen, E.; Jousilahti, P.; Kangas, A. J.; Männistö, S.; Savolainen, M. J.; Jula, A.; Leiviskä, J.; Palotie, A.; Salomaa, V.; Perola, M.; Ala-Korpela, M.; Peltonen, L. *Mol. Syst. Biol.* **2010**, *6*, 441.
- (61) St Hilaire, C.; Ziegler, S. G.; Markello, T. C.; Brusco, A.; Groden, C.; Gill, F.; Carlson-Donohoe, H.; Lederman, R. J.; Chen, M. Y.; Yang, D.; Siegenthaler, M. P.; Arduino, C.; Mancini, C.; Freudenthal, B.; Stanescu, H. C.; Zdebik, A. A.; Chaganti, R. K.; Nussbaum, R. L.; Kleta, R.; Gahl, W. A.; Boehm, M. *N. Engl. J. Med.* **2011**, *364*, 432–442.
- (62) Nicholson, J. K.; Holmes, E.; Wilson, I. D. *Nat. Rev. Microbiol.* **2005**, *3*, 431–438.
- (63) Waldram, A.; Holmes, E.; Wang, Y.; Rantalainen, M.; Wilson, I. D.; Tuohy, K. M.; McCartney, A. L.; Gibson, G. R.; Nicholson, J. K. *J. Proteome Res.* **2009**, *8*, 2361–2375.
- (64) Kinross, J. M.; Holmes, E.; Darzi, A. W.; Nicholson, J. K. *Lancet* **2011**, *377*, 1817–1819.
- (65) Mirnezami, R.; Kinross, J. M.; Vorkas, P. A.; Goldin, R.; Holmes, E.; Nicholson, J.; Darzi, A. *Ann. Surg.* **2012**, *255*, 881–889.
- (66) Li, J. V.; Reshat, R.; Wu, Q.; Ashrafi, H.; Bueter, M.; le Roux, C. W.; Darzi, A.; Athanasiou, T.; Marchesi, J. R.; Nicholson, J. K.; Holmes, E.; Gooderham, N. J. *Front. Microbiol.* **2011**, *2*, 183.
- (67) Kinross, J. M.; Alkhamesi, N.; Barton, R. H.; Silk, D. B.; Yap, I. K. S.; Darzi, A. W.; Holmes, E.; Nicholson, J. K. *J. Proteome Res.* **2011**, *10*, 277–287.
- (68) Kinross, J.; Warren, O.; Basson, S.; Holmes, E.; Silk, D.; Darzi, A.; Nicholson, J. K. *Biomarkers Med.* **2009**, *3*, 175–192.
- (69) Nicholson, J. K.; Holmes, E.; Kinross, J. M.; Darzi, A. W.; Takats, Z.; Lindon, J. C. *Nature* **2012**, *491*, 384–392.
- (70) Guenther, S.; Schäfer, K. C.; Balog, J.; Dénes, J.; Majoros, T.; Albrecht, K.; Tóth, M.; Spengler, B.; Takáts, Z. *J. Am. Soc. Mass. Spectrom.* **2011**, *22*, 2082–2089.
- (71) Hinz, K. P.; Gelhausen, E.; Schäfer, K. C.; Takats, Z.; Spengler, B. *Anal. Bioanal. Chem.* **2011**, *401*, 3165–3172.
- (72) Balog, J.; Szaniszló, T.; Schaefer, K. C.; Denes, J.; Lopata, A.; Godorhazy, L.; Szalay, D.; Balogh, L.; Sasi-Szabo, L.; Toth, M.; Takats, Z. *Anal. Chem.* **2010**, *82*, 7343–7350.
- (73) Schäfer, K. C.; Szaniszló, T.; Günther, S.; Balog, J.; Dénes, J.; Keserü, M.; Dezső, B.; Tóth, M.; Spengler, B.; Takáts, Z. *Anal. Chem.* **2011**, *83*, 1632–1640.
- (74) Schäfer, K. C.; Balog, J.; Szaniszló, T.; Szalay, D.; Mezey, G.; Dénes, J.; Bognár, L.; Oertel, M.; Takáts, Z. *Anal. Chem.* **2011**, *83*, 7729–7735.
- (75) Agar, N. Y. R.; Golby, A. J.; Ligon, K. L.; Norton, I.; Mohan, V.; Wiseman, J. M.; Tannenbaum, A.; Jolesz, A. *Neurosurgery* **2011**, *68*, 280–289.
- (76) Johnson, C. H.; Athersuch, T. J.; Wilson, I. D.; Iddon, L.; Meng, X.; Stachulski, A. V.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2008**, *80*, 4886–4895.

Appendix E: Paper VI

This publication was originally printed in Genome Medicine by BioMed Central.

REVIEW

Steven L Robinette^{1,2}, Elaine Holmes¹, Jeremy K Nicholson^{1,*} and Marc E Dumas^{1,*}

Abstract

Increasingly sophisticated measurement technologies have allowed the fields of metabolomics and genomics to identify, in parallel, risk factors of disease; predict drug metabolism; and study metabolic and genetic diversity in large human populations. Yet the complementarity of these fields and the utility of studying genes and metabolites together is belied by the frequent separate, parallel applications of genomic and metabolomic analysis. Early attempts at identifying co-variation and interaction between genetic variants and downstream metabolic changes, including metabolic profiling of human Mendelian diseases and quantitative trait locus mapping of individual metabolite concentrations, have recently been extended by new experimental designs that search for a large number of gene-metabolite associations. These approaches, including metabolomic quantitative trait locus mapping and metabolomic genome-wide association studies, involve the concurrent collection of both genomic and metabolomic data and a subsequent search for statistical associations between genetic polymorphisms and metabolite concentrations across a broad range of genes and metabolites. These new data-fusion techniques will have important consequences in functional genomics, microbial metagenomics and disease modeling, the early results and implications of which are reviewed.

Keywords Metabonomics/metabolomics, Quantitative Trait Locus Mapping, Biochemical Genetics, NMR, MS

Introduction

The last few decades have witnessed a radical change in the biological sciences, as the advent of the 'omics' era has brought large-scale measurement of genes [1,2], transcripts [3], proteins [4,5] and metabolites [6-8]. Genetics has undergone a high-throughput revolution, with genomic sciences enabling the rapid acquisition of genome-wide gene-expression profiles, polymorphisms and, more recently, whole genome sequences [9]. These advances have been matched by advances in the measurement of small-molecule metabolites in the associated fields of metabonomics [10,11] and metabolomics [8,12]. Like genomics, these fields aim for comprehensive measurement and analysis of variations, but in the complement of low-molecular-weight compounds within a cell, tissue or biofluid.

For the past 20 years, developments in genomics and metabonomics have progressed on parallel tracks, exchanging experimental designs, data-analysis techniques and applications in basic biology and medicine [13-16]. Yet genes and metabolites are intrinsically co-informational, each shedding light on complementary biological processes. The genome encodes the metabolic capacities of the cell (the microbiome also influences mammalian metabolism), and changes in the activity and function of enzymes, transporters and transcription factors resulting from genetic variations have a direct impact on the identities and quantities of both intracellular and extracellular metabolites. Metabolite concentrations are ultimately quantitative, phenotypic traits, the genetics of which are described by the quantitative trait locus (QTL) - a DNA sequence controlling the phenotypic outcome of the quantitative trait, such as a metabolite concentration. Since the origins of biochemical genetics over a century ago, the integrated study of genetics and metabolism has produced significant advances in the understanding of basic biological processes and in the diagnosis and treatment of human disease [17].

Metabolic profiling of single gene mutations [18] and QTL mapping of single metabolic traits [19] both

*Correspondence: j.nicholson@imperial.ac.uk; m.dumas@imperial.ac.uk

¹Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, Exhibition Road, South Kensington, London SW7 2AZ, UK

Full list of author information is available at the end of the article

represent early attempts at identifying gene-metabolite associations through omic sciences, by regressing one gene against many metabolites or one metabolite against many genes. In more recent studies, metabolome-wide profiling of biofluids by nuclear magnetic resonance (NMR) spectroscopy or mass spectrometry (MS) is combined with whole-genome profiling of single nucleotide polymorphisms (SNPs) to identify many gene-metabolite associations simultaneously, by regressing many metabolite levels against many polymorphisms [20]. While these studies, termed metabolomic quantitative trait locus (mQTL) mapping, or metabolomic genome-wide association studies (mGWAS), are often applied to human genes and human biofluid profiles, these techniques hold the additional promise of investigating interactions between gut microflora genes and biofluid metabolites. By regressing metagenomic sequences against metabolic profiles, metagenome-wide metabolome-wide associations provide insight into the metabolic cross-talk of bacterial and human genomes in the larger human superorganism [21,22]. This review discusses the highly analogous methods and applications of genomics and metabolomics, as well as recent attempts at integrating the two fields towards a more comprehensive and holistic understanding of gene function and the control of metabolic processes.

Developments in instrumentation and experimental design in metabonomics parallel those in genomics

Increases in the speed, accuracy and coverage of genomic analysis have been mirrored by technological developments in the large-scale measurement of low-molecular-weight metabolites, using the two major analytical platforms of NMR spectroscopy and MS [23]. While these techniques feature varying strengths in coverage, sensitivity, selectivity for various chemical classes, reproducibility, provision of structural information, and sample-preparation requirements, both stand out in their capacity to measure a large number of small-molecule analytes in an untargeted fashion from complex biological mixtures, such as human biofluids [24].

One of the most popular analytical chemistry techniques, NMR spectroscopy has a long history of application in organic chemistry for structural identification and is used extensively in metabonomics. NMR spectroscopy is characterized by the following key properties, which are fit for purpose: (i) high dynamic range, with several biological nuclei, such as ^1H , ^{13}C , ^{15}N or ^{31}P , being accurately measured over a large range of concentrations; (ii) high linearity of signal intensity with concentration; and (iii) high reproducibility. In particular, ^1H NMR spectroscopy is robust, provides a high degree of structural information for both one-dimensional ^1H and

two-dimensional ^1H - ^1H NMR, and is flexibly applied to extracts, biofluids and solid tissues using high-resolution magic angle spinning, and *in vivo* using magnetic resonance spectroscopy. Technological advancements in magnetic field strength with the introduction of 600, 800 and, recently, 1,000 MHz NMR spectrometers, pulse sequence experiments, and cryogenically cooled probes have increased the sensitivity and coverage for small-molecular-weight metabolites and lipid components from urine, plasma, serum and tissue samples [11].

MS, coupled to either liquid (LC) or gas (GC) chromatography, is also frequently applied to profile the metabolome [25-27]. LC- and GC-MS both boast high sensitivity to low-concentration analytes, as well as high-resolution chromatographic separation. While these techniques have faced challenges in compound identification, reproducibility and bias towards certain compound functional groups, the rapid pace of technological development in MS has sought to address many of these challenges. Advances in chromatography, including ultra-high-performance liquid chromatography (UPLC) and multidimensional gas chromatography (GC $\times$ GC, 3D-GC), have increased the speed and reproducibility of chromatography-coupled MS [28]. Additionally, advances in MS, including high-resolution time-of-flight (ToF) and quadrupole time-of-flight (qTOF) instruments, along with serial ion fragmentation (MS-MS, MS ^{n}) have improved resolution, coverage and identification of low-molecular-weight species [29].

While many comparisons have been made between the utility of these techniques for metabolite analysis [30], advances in both have paralleled advances in speed, accuracy and coverage in genomics. As a result, broad-coverage metabolite profiling using NMR spectroscopy and MS has been increasingly used to answer the same questions as genomics, especially in the search for risk factors for disease at the population level and predictors of drug metabolism in personalized medicine [13-16]. These are mirrored in parallel developments in experimental design and data analysis.

Just as the introduction of the genome-wide association study (GWAS) began the search for associations between genome-wide polymorphisms with disease phenotypes in large population cohorts [15,31], the metabolome-wide association study (MWAS) has used large numbers of biofluid spectra and statistical regression to search for associations between metabolites present in human biofluids and both quantitative and binary physiological and pathological traits [16]. Two-class experimental designs, in which metabolite associations with disease are identified by statistical regression of metabolic profiles against binary variables for affected and control individuals, are common, and associations between metabolites and disease have been reported for obesity

and insulin resistance [32], prostate cancer [33], autism [34], ulcerative colitis [35] and more [36]. However, the recent application of metabolic profiling to large population cohorts, quantitative traits and population differences (termed MWAS) has revealed metabolite associations with diet and blood pressure [37], region and cardiovascular risk [38] and ethnicity [39].

In addition to studies of disease risk, both metabolites and genes have been queried to predict drug metabolism. Paralleling previous work in pharmacogenomics, the introduction of pharmaco-metabonomics has demonstrated that drug metabolism can be predicted from the metabolite composition of urinary biofluids before drug administration [13,40]. Recent applications of pharmaco-metabonomics have highlighted metabolic predictors of acetaminophen toxicity in animals [13] and humans [41,42], capecitabine toxicity [43] and microbial influences on drug detoxification [41].

Early integration: metabolic profiling of Mendelian traits and QTL mapping of single metabolites

Since the discovery of alkaptonuria by Archibald Garrod in the early 20th century, the measurement of metabolites has been used as a proxy to identify human genetic diseases, especially inborn errors of metabolism. Uniform newborn screening for multiple inborn errors of metabolism, including urea-cycle disorders and amino- and organic-acidurias, using heel-prick testing with GC-MS-MS exemplifies the power of targeted metabolomic analysis to diagnose these diseases and enable the interruption of pathological processes resulting from genetic mutations [44]. More recently, untargeted NMR spectroscopy and MS have been used to diagnose known inborn errors [45–48], identify novel inborn errors using biofluid profiling [49,50] and identify complex downstream metabolic consequences [51–53] and biomarkers of organ pathology resulting from genetic mutations [54–56].

Untargeted metabolic profiling of biofluids, especially urine and serum, is a powerful technique for diagnosing inborn errors of metabolism with often non-specific clinical presentation, as metabolic intermediates accumulated in biofluid compartments can be easily identified [18,57,58]. As a result, diagnosis of suspected inborn errors has been reported for many Mendelian diseases [45–50], especially using NMR spectroscopy. In some cases, metabolic profiling of biofluids from patients with suspected inborn errors has led to the discovery of previously undescribed diseases, with the identification of causal genes following the description of metabolic perturbations, as occurred with aminoacylase 1 deficiency and beta-ureidopropionase deficiency [49,50].

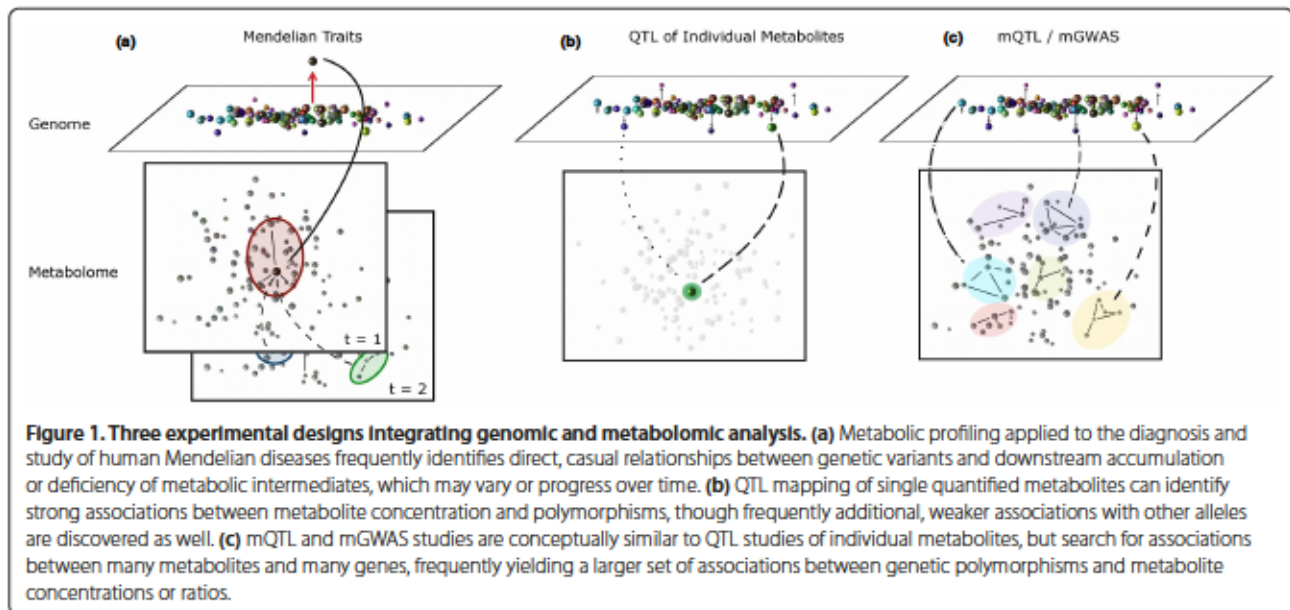
While mutations in enzymes and transporters can often be readily diagnosed by biofluid profiling, and a strong mechanistic link is easily inferred between the

disruption of a metabolic pathway and resulting accumulation or depletion of metabolic intermediates, many Mendelian diseases result in more complex, progressive organ-specific or multi-organ pathology [59]. In these cases, metabolic profiling has been applied to identify sites of lesions, describe progression and attempt to identify proxy small-molecular biomarkers of the disease. Examples of this include the identification of markers of autosomal dominant polycystic kidney disease [55], comparison of the urinary profiles of several genetic forms of renal Fanconi syndrome [56] and description of abnormal brain metabolism in Smith-Lemli-Optiz syndrome [54]. Additionally, metabolite flux analysis using isotopically labeled metabolites suggests an additional way to apply metabolic profiling techniques to study the impact of genetic mutations [59].

The genomic corollary of using metabolic profiling to study Mendelian genes is QTL mapping of single metabolic traits. The advent of whole-genome SNP analysis and the use of QTL mapping for quantitative traits, such as height [60], led to interest in identifying genetic loci associated with quantitative variation in individual metabolite levels [19]. Examples of this include mapping of serum leptin levels to genes on human chromosome 2 in multiple human populations [61,62], associations with plasma triglyceride levels [63,64] and identification of genetic variants associated with high-density lipoprotein (HDL) levels [65,66]. Recent studies have investigated associations between serum lipid fractions and polymorphisms, constituting an intermediate between traditional metabolite QTL mapping and mQTL/mGWAS [67]. Like many QTL mapping studies, attempts to identify loci significantly associated with biofluid levels of single metabolites frequently indicate a large number of genetic associations. This is almost certainly an indication of complex, multigenic control processes regulating energy metabolism and homeostasis, and the identification of large numbers of multiple loci provides an important documentation of genes involved in complex metabolic pathways. However, a large number of loci each contributing to a potentially small percentage of observed variance in metabolite levels complicates direct interpretation of genotype-phenotype relationships in these cases. Figure 1 shows a schematic illustration of gene-metabolite correlations in biochemical genetics, traditional QTL mapping, and mQTL/mGWAS.

Identifying the genetic determinants of the metabolome: mQTL and mGWAS

GWAS currently requires increasingly large cohorts to ensure discovery of new genes associated with disease phenotypes [68]. Although this approach is very efficient, the biological relevance of these associations can be difficult to assess. The identification of phenotypes



related to disease mechanism, onset and progression represents a promising research avenue.

The systematic search for molecular endophenotypes (that is, internal phenotypes) that can be mapped onto the genome began with the quantitative genetic analysis of gene-expression profiles, referred to as genetical genomics [69] or expression QTL (eQTL) mapping [70]. Treating genome-wide gene-expression profiles as quantitative traits was originally developed in model organisms and applied to humans [70,71]. In eQTL mapping, *cis*-regulatory associations between genomic variations and gene-expression levels are discovered by integrated analysis of quantitative gene-expression profiles and SNPs. The identification of a SNP at a gene locus affecting its own expression represents a powerful self-validation. However, eQTL mapping presents a series of drawbacks: (i) frequently analyzed cell lines often have altered gene expression, and access to biopsy samples from organs directly relevant to pathology is often impossible; and (ii) due to the gene-centric nature of eQTL mapping, this approach bypasses the biological consequences of the endophenotypes generating the association.

Immediately following the success of the eQTL mapping approach [70], in which *cis*-regulatory associations between genomic variations and gene-expression levels are discovered by integrated analysis of quantitative gene-expression profiles and SNPs, metabolic profiles were included as endophenotypic quantitative traits. This led to mapping of multiple quantitative metabolic traits directly onto the genome to identify mQTLs in plants [72,73], then in animal models [74,75]. In mQTL mapping, individuals are genotyped and phenotyped in

parallel and the resulting genome-wide and metabolome-wide profiles are then quantitatively correlated (Box 1). mQTL mapping presents a significant advantage over gene-expression products such as transcripts [70] or proteins [76]: the ever-increasing coverage of the metabolome allows a glimpse at the real molecular endpoints, which are closer to the disease phenotypes of interest. Following the success of mQTL mapping in plants [72,73] and then in mammalian models [75], this approach was quickly followed by the development of mGWAS in humans cohorts ([77-83], see also the review by J Adamski [84]).

One of the distinctive features of mGWAS is the intrinsically parallel identification of associations between monogenetically determined metabolic traits and their causative gene variants (see Table 1 for a list of human mQTL-metabolite associations).

The mechanistic explanation of gene/metabolite associations identified by mQTL mapping can be difficult. The simplest case corresponds to associations between genes encoding enzymes and metabolites, which are either substrates or products of the enzyme they are associated with [74,75] (Figure 2). This corresponds to a direct *cis*-acting mechanism. Also, one of the interesting discoveries from results obtained by Suhre *et al.* is that a number of gene variants causing metabolic variation correspond to solute transporter genes, as the majority of the genes in this category belong to the solute carrier (SLC) family [78,81,82]. Again, this corresponds to a direct mechanistic link. In other cases, the link between gene variants and their associated metabolites can demonstrate pathway, rather than direct, connectivity, such as polymorphisms in enzymes associated with

Box 1. Mathematical modeling for mQTL identification

The statistical analysis involved in mQTL mapping and mGWAS does not currently differ substantially from the statistical methods used to identify genetic loci associated with single quantitative traits. mQTL and mGWAS involve independent QTL mapping of each metabolite identified by metabolic profiling, though accurate analysis is dependent upon proper preprocessing of both genomic and metabolomic data. Associations are identified using techniques such as Haley-Knott regression implemented in the R/QTL package, which uses local information about surrounding markers [103], or typical univariate association tests such as χ^2 or Cochran-Armitage trend tests implemented in PLINK [104]. The results of mQTL and association mapping are typically displayed using a logarithm of odds (LOD, $-\log_{10}(P \text{ value})$) score, which allows establishment of genome/metabolome LOD score maps [74,75], or more classical Manhattan plots [77,78,81,82] (Figure 2).

The main challenge in mQTL data modeling is multiple correlation testing. Assuming the use of high-resolution metabolic profiles (1,000 to 10,000 features) and genome-wide SNP coverage (600,000 SNPs), a typical metabolome-wide GWAS can apply between 600,000,000 and 6,000,000,000 univariate tests. Given the number of tests involved, there are numerous opportunities for false discoveries and multiple testing corrections are required to account for this. Genome-wide significance levels can be estimated using Bonferroni correction [77], but also using Benjamini and Hochberg or Benjamini and Yekutieli corrections [105]. Finally, permutation and resampling methods also provide empirical estimates for false discovery thresholds [74,79].

metabolites several reactions downstream of the compound directly acted upon by the enzyme itself (as observed with *NTSE* polymorphisms and inosine).

More opaque associations may be *trans*-acting in a broader sense: the causative gene variant can be a molecular switch, and the metabolite it is associated with is in fact regulated indirectly by this molecular switch (further down in the regulation events). This is particularly the case when the causative gene variant encodes a transcription factor, inducing the medium- to long-term expression of entire gene networks, or when the gene variant encodes a kinase or a phosphatase regulating entire pathways on much shorter time-scales. Unlike *cis*-acting mQTL/metabolite associations, which can be seen as self-validation of the causative gene at the locus, *trans*-acting mQTL associations present the challenge of identification of the most relevant causative gene at the locus. If a SNP is associated with a metabolite, the closest gene at the locus is not necessarily the most relevant candidate, and further investigation of a larger biological network, such as protein-protein interactions [85], may be necessary to identify mechanistic relationships between genetic variants and downstream metabolism. Despite these challenges, which are familiar to

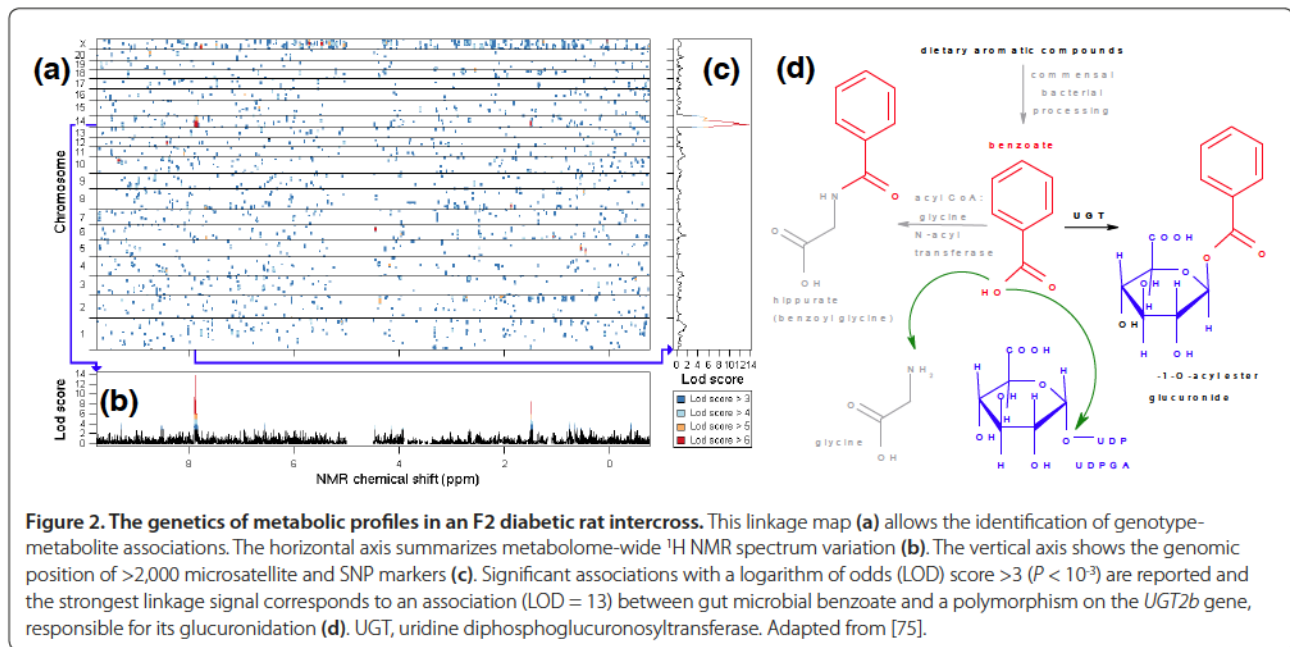
practitioners of biochemical genetics, statistical identification of gene-metabolite associations by mQTL and mGWAS promises to significantly advance current understandings of gene function, metabolic regulation and mechanisms of pathology.

A glimpse of our extended genome with microbiome-metabolome associations

The functional genomic association studies and data-integration approaches described above rely predominantly on mammalian genome sequences and their annotation (excluding MWAS, which makes use only of metabolite profiling data and does not, as such, require genomic data). However, human phenotypes result from the interaction of several sets of genes: the karyome, the chondriome and the microbiome, respectively corresponding to eukaryotic chromosomes, mitochondrial chromosomes and, finally, gut bacterial prokaryotic chromosomes. The latest human gut microbiome gene catalogue identified 3.3 million non-redundant genes [86], which was dubbed 'our other genome', and the bacterial species composition of the gut microbiome varies from one individual to another, but this variation is stratified, not continuous, and suggests the existence of stable bacterial communities, or 'enterotypes' [87].

The classical identification of associations between gut bacteria and metabolites has been performed on a case-by-case basis for decades. However, the correlation of metabolic profiles with multiple gut bacterial abundance profiles was initiated a few years ago with the introduction of bacteria/metabolite association networks [21]. Semi-quantitative characterizations of microbial populations using denaturing gradient gel electrophoresis (DGGE) and fluorescent *in situ* hybridization (FISH) have yielded associations with obesity and related metabolites [88]. Recently, the introduction of high-throughput sequencing of bacterial 16S rDNA profiles and correlation with metabolic profiles has greatly increased the coverage and quantification of microbial species [89]. The correlation of metabolic profiles with 16S rDNA microbiome profiles provides a strategy for the identification of co-variation between metabolites and bacterial taxa, and such associations point to the production or regulation of metabolic biosynthesis by these microbes.

Given these early successes, the integration of metabolome-wide experimental profiles with metagenome-wide metabolic reconstruction models obtained from full microbiome sequencing should provide a clear insight into the functional role of the gut microbiome, especially the synthesis of metabolites and resultant impacts on human metabolism. This critical need for a marriage between metabolomics/metabonomics and metagenomics has been clearly identified for several years [90]. How



new experimental data change our understanding of our commensal microflora remains to be seen.

Future directions - the rise of sequencing and consequences for genome-metabolome data fusion

Genomics is currently undergoing yet another revolution, as next-generation sequencing technologies increase the accuracy, coverage and read-length, and drastically decrease the cost of whole-exome sequencing (WES) and whole-genome sequencing (WGS). The introduction of third-generation sequencing technologies in the near future promises to continue this trend [91]. Consequently, the near term promises a dramatic expansion in the availability of sequence data both in the laboratory and in the clinic. The relevance of the explosion of sequence data to the continued integration of metabonomic and genomic data is twofold: first, an opportunity for metabonomics to contribute to the increased clinical presence of omics sciences led by genome sequencing; and second, a challenge to develop methods of integrating metabolic profiles with sequences rather than polymorphisms.

The introduction of WES and WGS into the clinic is already well underway, with success stories including discoveries of new Mendelian disorders [92,93] and successful therapy designed on the basis of mutation discovery [94]. Of known and suspected human Mendelian diseases, molecular bases have been identified for over 3,000, with another approximately 3,700 phenotypes suspected of having a Mendelian basis [95,96]. As sequencing identifies an increasing number of variants with associations to disease, the rate-limiting step in

genomic medicine will move from discovery to functional annotation of sequence variants. Metabolite profiling, along with other high-throughput measurement and data-analysis technologies, may find increasing acceptance in medicine, as investigators rush to keep up with a deluge of sequence data. Increasingly, routine genome sequencing will create a significant resource for large-scale population studies like those currently used to identify gene-metabolite associations, and will critically include rare variants not captured by polymorphism data.

Recent results demonstrate the potential power of integrating WES/WGS with metabolic profiling and mQTL/mGWAS. In 2011, a publication in the *New England Journal of Medicine* reported the discovery by WES of a novel human Mendelian disease in which rare mutations in the *NT5E* gene result in arterial calcifications due to loss of CD73 (encoded by *NT5E*) function, which converts AMP to adenosine in the vasculature [97]. Within the year, an independent metabolome-wide GWAS study published in *Nature* reported a statistically significant association between a SNP near the *NT5E* locus (rs494562) and inosine concentration in human serum, as part of a much larger set of gene-metabolite associations (Table 1) [81]. While in this case the publication of the statistical association followed the description of the human phenotype, future genetic studies will be greatly aided by gene-metabolite co-variation discovered by association studies.

Despite the significant opportunity represented by low-cost sequencing for the integration of genomic and metabonomic data and the identification of gene-metabolite

Table 1. Human gene-metabolite associations identified by mQTL/mGWAS

Metabolite	Biofluid	SNP ID	Local gene	P-value	Reference(s)
Trimethylamine	Urine	rs7072216	PYROXD2 (C10orf33)	7.90E-15	[79]
N-acetylated compound(s)	Urine	rs9309473	ALMS1, NAT8, TPRKB, DUSP11	1.40E-11	[79]
3-Amino-isobutyrate	Urine	rs37369	AGXT2	1.1E-06	[79]
				3.17E-75	[82]
2-Hydroxyisobutyrate	Urine	rs830124	WDR66, HPD	1.59E-15	[82]
Dimethylamine	Plasma	rs6584194	PYROXD2 (C10orf33)	8.10E-03	[79]
Sphingomyelin SM C14:10	Serum	rs9309413	PLEK	1.95E-09	[77]
Lysine	Serum	rs992037	PARK2	1.20E-07	[77]
Sphingomyelin SM(OH,COOH) C18:2	Serum	rs1148259 (rs1200826)	ANKRD30A	3.04E-09	[77]
Phosphatidylcholine PC aa C36:4	Serum	rs174548	FADS1	4.52E-08	[77]
Phosphatidylethanolamine PE aa C38:6	Serum	rs4775041	LIPC	9.66E-08	[77]
C0	Serum	rs7094971	SLC16A9	3.80E-20	[78]
N-Acetylmethionine	Serum	rs13391552	NAT8	5.40E-252	[81]
5-Oxoproline	Serum	rs6558295	OPLAH	1.50E-59	[81]
Androsterone sulfate	Serum	rs17277546	CYP3A4	8.70E-40	[81]
Urate	Serum	rs4481233	SLC2A9	5.50E-34	[81]
Glycine	Serum	rs2216405	CPS1	1.60E-27	[81]
		rs7422339	CPS1	2.12E-24	[83]
Succinylcarnitine	Serum	rs2652822	LACTB	7.20E-27	[81]
Isobutyrylcarnitine	Serum	rs662138	SLC22A1	7.30E-25	[81]
Aspartylphenylalanine	Serum	rs4329	ACE	8.20E-20	[81]
Serine	Serum	rs477992	PHGDH	2.60E-14	[81]
Inosine	Serum	rs494562	NT5E	7.40E-13	[81]
Proline	Serum	rs2023634	PRODH	2.00E-22	[81]
α-Hydroxyisovalerate	Serum	rs2403254	HP55	1.00E-20	[81]
Bradykinin, des-arg(9)	Serum	rs4253252	KLKB1	6.60E-18	[81]
Glutamine	Serum	rs2657879	GLS2	3.10E-17	[81]
Isovalerylcarnitine	Serum	rs272889	SLC22A4	7.40E-16	[81]
Decanoylcarnitine	Serum	rs8396	ETFDH	5.50E-15	[81]
Carnitine	Serum	rs7094971	SLC16A9	3.40E-14	[81]

118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000

associations, several challenges stand in the way of routine paired analysis of sequences and metabolic profiles. The first of these is the challenge of discovering significant associations with low sample numbers. Many of the successes reported in clinical sequencing have made use of data from a small number of patients, and sometimes only a sole patient. In these cases, potentially disease-causative variants are often identified using filtering strategies rather than statistical analysis [98,99]. While the diagnosis of human inborn errors of metabolism from single-patient biofluid NMR spectra demonstrates the potential of metabolic profiling to work with low sample numbers, lack of statistical validation

means that the 'biological signal' in these cases must be quite marked. A second challenge is a dearth of tools for statistical analysis of sequence data. While QTL mapping using SNPs is well established, statistical techniques for QTL mapping with both rare and common variants are just beginning to be introduced [100]. It is likely that increased availability of large-scale population sequence data from initiatives such as the 1000 Genomes Project [101,102] and ClinSeq [103] will spur the development of statistical methods that can be deployed to identify gene-metabolite associations.

Of the omics sciences, genomics and metabolomics are uniquely complementary, the strengths of each addressing

weaknesses of the other. Genes are (mostly) static, an ‘upstream’ blueprint controlling dynamic biological processes. The identities and quantities of ‘downstream’ metabolites capture both genetic and environmental influences, and can be measured serially to assess variation through time. Genomic studies often struggle to establish a firm link between genetic variants and phenotypic observations, and while metabonomics provides a closer proxy to phenotype, it is often difficult to infer underlying causality from variations in metabolism. Together, the integrated application of genomics and metabonomics promises a bridging of the gap between genotype and phenotype through intermediate metabolism, to help annotate genes of unknown function, genetic controls of metabolism, and mechanisms of disease.

Abbreviations

DGGE, denaturing gradient gel electrophoresis; FISH, fluorescent *in situ* hybridization; GC, gas chromatography; GWAS, genome-wide association study; HDL, high density lipoprotein; LC, liquid chromatography; mGWAS, metabolomic genome-wide association study; mQTL, metabolomic quantitative trait locus; MS, mass spectrometry; MWAS, metabolome-wide association study; NMR, nuclear magnetic resonance; QTL, quantitative trait locus; qToF, quadrupole time-of-flight; SNP, single nucleotide polymorphism; ToF, time-of-flight; UPLC, ultra-performance liquid chromatography; WES, whole exome sequencing; WGS, whole genome sequencing.

Acknowledgements

SLR acknowledges support from the NSF Graduate Research Fellowship Program and from the Marshall Aid Commemoration Commission. (M-ED is funded by Nestlé (RDLS015375), Agence Nationale de la Recherche (ANR-08-GENO-030-02), and EU-FP7 EURATRANS (HEALTH-F4-2010-241504).

Competing interests

The authors declare that they have no competing interests.

Author details

¹Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, Exhibition Road, South Kensington, London SW7 2AZ, UK. ²Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA.

g3y00 " "(B: (Q000A: ? A

References

1. Hood L, Galas D: k " " " ' ' 00000 " 0'(T e Q8Nature 2003, CA?I444-448.
2. Collins FS, Green ED, Guttmacher AE, Guyer MS, Institute UNHGR: Q0 000('00 0 " ('0000 0'(' "000 z0000 x0z 8Nature 2003, CAAI835-847.
3. Wang Z, Gerstein M, Snyder M: i e Q7 00(x0 000000x00000 00 000000000 z08Nat Rev Genet 2009, ? : I57-63.
4. Liotta LA, Ferrari M, Petricoin E: S00 z00000 00 z00 000 0(0y000" 8Nature 2003, CAD905.
5. Tyers M, Mann M: V000 (' "0000 z0000000 00 z08Nature 2003, CAAI193-197.
6. Nicholson J, Connelly J, Linton J, Holmes E: d "0ky0000 z0(x000000 ('00 000" 00" (' "00 000 z 00(x0" (' "00"('00z0008Nat Rev Drug Discov 2002, ?I153-161.
7. Nicholson JK, Wilson ID: I 0 " " "0000" 0' (w00yx000000 0 y 0000 0(0 "0ky0000 z0000 0 " (z0000000 0' 0 "0ky0000 8Nat Rev Drug Discov 20036A1668-676.
8. Fiehn O: d "0ky0000 z070 "000y"00 "0"(' "00000"0x0" 0 "00000"08 Plant Mol Biol 2002, CG155-171.
9. Diamandis EP: e "000 "0"000000'00"0z 0' I(x0"0 0 0000000(00 00 z0000 " x' 0000z00Clin Chem 2009, DDI2088-2092.
10. Nicholson J, Linton J, Holmes E: d "0ky0000 z01000 "0000" 0' 0 " (" 0 "0ky00z0'00000 00 000' 0000 0 0000x0 00 0000' z0000 000x(

0 00000x00' 00000z0000x0000 0(y 0000 z00e d i 00" z000000 z(" x08 Xenobiotica 1999, AH1181-1189.
11. Beckonert O, Keun H, Ebbels T, Bundy J, Holmes E, Linton J, Nicholson J: d "0ky00z000f000' 00 "0ky0000 z0(x0" 0 "0ky0000 z0000z" " "00'000e d i (00" z00000000' 0000' 00000 x00' 000 (x0" 00000"(' "000z008Nat Protocols 2007, AI2692-2703.
12. Oliver SG, Winson MK, Kell DB, Baganz F: j 0000 x00z('000z000x000x00000' (0 " 0' x000 "0000 "8Trends Biotechnol 1998, ?EI373-378.
13. Andrew Clayton T, Linton JC, Cloarec O, Antti H, Charuel C, Hanton G, Provost J-P, Le Net J-L, Baker D, Walley RJ, Everett JR, Nicholson JK: g x00 xz07 0 "0ky0000 z0 "00000 0' (x0" 00"0000x00" " (" 00' 00' x00 "008Nature 2006, CC: I1073-1077.
14. Evans WE, McLeod HL: g x00 xz07 "0000 z070"00" (" 00000000' 00' 000" 000 x0" 00" " ("ff" z008N Engl J Med 2003, BCG538-549.
15. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN: X"0000 " 70 " " (x000z x00000000 " 0"00z000 00 0000 0(z0000"000000z"000 0000x00" (z x000 0 "08Nat Rev Genet 2008, H356-369.
16. Nicholson JK, Holmes E, Elliott P: k " 0 "0ky0000 " 70 " " (x000z x00000000 0(x0"0 00000x0 00 x0"(" 0' x00' 0000' xz00008J Proteome Res 2008, F1637-3638.
17. Rinaldo P, Hahn S, Matern D: S00 z00y 0z " 0 z007 "0"0z0 00" (00 "0007 f000z"00008Acta Paediatr Suppl 2004, HBI22-26; discussion 27.
18. Constantinou M, Papakonstantinou E, Spraul M, Sevastiadou S, Costalos C, Koupparis M, Shulpis K, Tsantili-Kakoulidou A, Mikros E: [7(e d i 7yx0" (" 0 "0ky0000 z01000 " (" x' 000000" (0y0000" 000000' 0 "0ky0000 (0000" 8 Anal Chim Acta 2005, DCAI169-177.
19. Almasy L, Blangero J: [00 x0(hkc0000" ' 00 x000 0' 8Genetica 2009, ?BEI333-340.
20. Ala-Korpela M, Kangas AJ, Inouye M: X"0000 " 70 " " (x000z x00000000 " 0x0" (0000 0 0y 0000' 0000 " 0 "0x0008Trends Genet 2011, AFI493-498.
21. Martin F-PJ, Dumas M-E, Wang Y, Legido-Quigley C, Yap IK, Tang H, Zirah S, Murphy GM, Cloarec O, Linton JC, Sprenger N, Fay LB, Kochhar S, van Bladeren P, Holmes E, Nicholson JK: Q00007"00 0(00000 0 y 0000' 00" 0 0' (0 z00y 00 " 70 x0 0 x0000 "0ky00z(000 0z00000 0(x0 000 00" 0 "08Mol Syst Biol 2007, BI112.
22. Nicholson JK, Holmes E, Wilson ID: X0000 z00000 x0 00 000 x0 0 x00(0 "0ky0000 (x0" 00" 0000x00" (" x00 (zx07 8Nat Rev Microbiol 2005, BI431-438.
23. Dunn W, Ellis D: d "0ky0000 z0(z0000 00x0000z00000000 0x0" (0 " 0' 0000 "08Trac-Trend Anal Chem 2005, ACI285-294.
24. Dunn W, Bailey N, Johnson H: d " x00000' 0 " 0 "0ky0000 "I(z0000 00 x0000z000 z 00000 "08Analyst 2005, ?B: I606-625.
25. Chan ECY, Pasikanti KK, Nicholson JK: X00yx0000x0000 "0ky00z000f000" (0000z" 00"00000" (' x0z z00 x000' 000 00 x0000" z0000 "0008Nat Protoc 2011, EI1483-1499.
26. Dettmer K, Aronov PA, Hammock BD: d x00000" z0000 " 0007yx0" (" 0 "0ky0000 z08Mass Spectrom Rev 2007, AEI51-78.
27. Want EJ, Wilson ID, Gika H, Theodoridis G, Plumb RS, Shockcor J, Holmes E, Nicholson JK: X00yx00 "0ky00z000f000" 0000z" "00'0'000000 000" (I gc57d j 8Nat Protoc 2010, DI1005-1018.
28. Lei Z, Huhman DV, Sumner LW: d x00000" z0000 " 00000x0" ' " 0 0(0 "0ky0000 z08J Biol Chem 2011, AGEI25435-25442.
29. Benton HP, Wong DM, Trauger SA, Siuzdak G: oS d j AI(000z"000' 000" " 0 (0 x0000" z0000 "0000" x00('0000 "0ky0000 (" "00fiz0000(x00" 0000z0000 z x00z0' 0000008Anal Chem 2008, G: I6382-6389.
30. Fernie AE, Trethewey R, Krotzky A, Willmitzer L: 00000x0000(70 "0ky0000 (000f000" I('000 (" x' 0000z0000000 0 y 0000 08Nat Rev Mol Cell Biol 2004, DI763-769.
31. Consortium WTCC: X"0000 " 70 " " (x000z x00000000 00' ?C6 :: (zx00 00' (0' 0'0000 0 00" 0' x00' 0x0" (B6 :: 0 x0" (z0000008Nature 2007, CCFI661-678.
32. Newgard CB, An J, Bain JR, Muehlbauer MJ, Stevens RD, Lien LF, Haqq AM, Shah SH, Arlotto M, Slentz CA, Rochon J, Gallup D, Ilkayeva O, Wenner BR, Yancy WS, Eisenberg H, Musante G, Surwit RS, Millington DS, Butler MD, Svetkey LP: Qy000z " " 7z x 0(x0 00(xz " 70 00" 0 "0ky00z0' 0x000' 0 x0 " ff"0'000x0 00y"0' (x0" 00 x0(00 x00x00" (z0000y00 000(00000(0'0000z" 8Cell Metab 2009, H131-326.
33. Sreekumar A, Poisson LM, Rajendiran TM, Khan AP, Cao Q, Yu J, Laxman B, Mehra R, Lonigro RJ, Li Y, Nyati MK, Ahsan A, Kalyana-Sundaram S, Han B, Cao X, Byun J, Omenn GS, Ghosh D, Pennathur S, Alexander DC, Berger A, Shuster JR, Wei JT, Varambally S, Beecher C, Chinnaiyan AM: d "0ky0000 z000f00 0 " "00" x0 0000 00x0000 ('000000000" (00000000 (zx00z" 0000' 00 0008 Nature 2009, CDFI910-914.

34. Yap IKS, Angley M, Veselkov KA, Holmes E, Lindon JC, Nicholson JK: I 000x00 0 "0ky00z0 "00000 0" (" ff"000x0 z 0 000 0 (x0000 ('000 0 " 0 000ff" z0 " 0y00 "0x00" (x' " 0 x0 " " (z0000000/ Proteome Res 2010, H2996-3004.
35. Bjerrum JT, Nielsen OH, Hao F, Tang H, Nicholson JK, Wang Y, Olsen J: d "0ky0000 z0 000z" 0000" (z00000" x' 0000z00y 00 x00" 0 " "00fizz0000 x00" (000 " 00000 " 0x0 00 00000 00/ Proteome Res 2010, H954-962.
36. Vinayavekhin N, Homan EA, Saghatelian A: U000000 (" 0" x00 0 0000' (0 "0ky0000 z08Acs Chem Biol 2010, D191-103.
37. Holmes E, Loo RL, Stamler J, Bictash M, Yap IKS, Chan Q, Ebbels T, de Iorio M, Brown IJ, Veselkov KA, Daviglus ML, Kesteloot H, Ueshima H, Zhao L, Nicholson JK, Elliott P: [00 x00 "0ky00z0 "00000 (" 0" 0000x00" (00 x000z x00000 0 (" " "0x0" (y0000 000 0000 8Nature 2008, CDB1396-400.
38. Yap IKS, Brown IJ, Chan Q, Wijeyesekera A, Garcia-Perez I, Bictash M, Loo RL, Chadeau-Hyam M, Ebbeis T, de Iorio M, Maibaum E, Zhao L, Kesteloot H, Daviglus ML, Stamler J, Nicholson JK, Elliott P, Holmes E: d "0ky0000 " 0 " " (x000z x0000000 " 0 " "0x0" (y 00 x000 000 x00 " 0000 0x0 00000 (x00 00000 (z 00 00 00000000x00 ff"00 00000 (zx00 00000000 0" x00" (Je kU d Qgij 000 00/ Proteome Res 2010, H6647-6654.
39. Bictash M, Ebbels TM, Chan Q, Loo RL, Yap IKS, Brown IJ, de Iorio M, Daviglus ML, Holmes E, Stamler J, Nicholson JK, Elliott P: f 00 00 000 0 " 00z00y0000 0 "0ky00z0 "00000 0" (x00 0 "0ky0000 " 0 " (x000z x00000000 " 00 (" 0 " " 000 00/ Clin Epidemiol 2010, EB1970-979.
40. Nicholson JK, Wilson ID, Lindon JC: g x00 xz00 "0ky0000 z00x000" ff" z000 '0000 "000x0000 " 0 " " z 00 8Pharmacogenomics 2011, ? A1103-111.
41. Clayton TA, Baker D, Lindon JC, Everett JR, Nicholson JK: g x00 xz00 "0ky0000 z (" " "00fizz00000' (x00' 0 fizz000 0000 z00y 00 (" " "0ky00z0 (00z0000xf" z000" (00 x00" 00 " 0ky0000 8Proc Natl Acad Sci U S A 2009, ? : E114728-14733.
42. Winnike JH, Li Z, Wright FA, Macdonald JM, O'Connell TM, Watkins PB: I 00 0' (0 x00 xz00 0 "0ky0000 z0 000" x00000 " z0000' (xz"0k 000 " 0700 0z" " ("0x00000 z 00 (00 x0000Clin Pharmacol Ther 2010, GG45-51.
43. Backshall A, Sharma R, Clarke SJ, Keun HC: g x00 xz00 "0ky0000 z000f000' (x00x000 " z00000 000 z 00 (00x00 0000 0 " 0000 "0ky00 z0000 z000z00z" 0 00 x00 " 0 (zx00 z 0ky 00 8Clin Cancer Res 2011, ? F13019-3028.
44. Wilcken B, Wiley V, Hammond J, Carpenter K: j z00 " 0 0' 00 0y000000 0y0000 ("000000 0 "0ky0000 (y0000 " 0 0 x0000 " z0000 "0000V Engl J Med 2003, BCG2304-2312.
45. Engelke UFH, Tassini M, Hayek J, de Vries M, Bilos A, Vivi A, Valensin G, Buoni S, Zannolli R, Brussel W, Kremer B, Salomons GS, Veendrick-Meekes MJBM, Kluijtmans LAJ, Morava E, Wevers RA: X00x00 " 00xz"000 0 " 0 000000 "00" (2Q0d k3 " fiz " 0z00" x' 0000 " (y00000000 (e d i 000 z0000z00000' (y00 00f00 " 08 NMR Biomed 2009, AA1538-544.
46. Maschke S, Wahl A, Azaroual N, Boulet O, Crunelle V, Imbenotte M, Foulard M, Vermeersch G, Lhermitte M: [7 7e d i (x00x00000' 000 " 0 000 0 " (00000" ('0000 " (" x' 00000000 000 000 "8Clin Chim Acta 1997, AEB1139-146.
47. Moolenaar S, Engelke U, Abeling N, Mandel H, Duran M, Wevers R: g000z" x00' (" " fiz " 0z00" x' 0000 " (y00 7 (e d i 000 z0000z00000' 0000 8J Inherit Metab Dis 2001, AC1843-850.
48. Sewell AC, Murphy HC, Iles RA: g00000000 z0 x00 x' 00 0z00 000x00z" (000 z0000z00 z (" 0" z00000' 0 x00z (" 00000' " (" 0" x00 8Clin Chem 2002, CG357-359.
49. Engelke UFH, Sass JO, Van Coster RN, Gerlo E, Olbrich H, Krywawych S, Calvin J, Hart C, Omrans H, Wevers RA: e d i 000 z0000z00000' (x00 000xz0000' ? (" " fiz " 0z00x0000 0y0000 ("00000 0 "0ky0000 8NMR Biomed 2008, A21138-147.
50. Moolenaar SH, Göhlich-Ratmann G, Engelke UF, Spraul M, Humpfer E, Dvortsak P, Voit T, Hoffmann GF, Bräutigam C, van Kuilenburg AB, van Gennip A, Vreken P, Wevers RA: y"007 0 " 000000 00x00 (" " fiz " 0z00(x0000 0y0000 ("00000 0 "0ky0000 (" 00000 0 " 00000 (e d i 000 z0000z0000000000 8Magn Reson Med 2001, CE11014-1017.
51. Engelke U, Kremer B, Kluijtmans L, van der Graaf M, Morava E, Loupatty F, Wanders R, Moskau D, Loss S, van den Bergh E, Wevers R: e d i 000 z0000z000 z (000 " " 000000 0 " 000 0000 0' 0000 0 (000 " 0 00 0000z00 z (xz " 000x00000' (0x00' (000 000 " " " z000 000 0z 00 0 "0ky0000 8NMR Biomed 2006, ? H271-278.
52. Holmes E, Foxall P, Spraul M, Farrant R, Nicholson J, Lindon J: FD: (d [00 7 (e d i 000 z0000z0000z x00z00 000000' 0 " (z00 00 000 "0ky00z0x000 000' (0000" (000 0x00 00000 0 (0y0000 (" 000000 " 0ky0000 l(A7 00 0000 0000z (xz" 000x0000 0 x000 000000000 (" 0" x00 8J Pharmaceut Biomed 1997, ? D1647-1659.
53. Wikoff WR, Gangoiti JA, Barshop BA, Siuzdak G: d "0ky0000 z0 " " 000f" 000 00000x00000 0 (00 x00 (" 0000 "0000 00000 x000 "0ky0000 8Clin Chem 2007, DB12169-2176.
54. Caruso P, Poussaint T, Tzika A, Zurakowski D, Astrakas L, Elias E, Bay C, Irons M: d i]0x00 (" 7 (d i j (f00 0' 00 (0 00 0 0 00000 000 " 8Neuroradiology 2004, CE13-14.
55. Gronwald W, Klein MS, Zeltner R, Schulze B-D, Reinhold SW, Deutschmann M, Immervoll A-K, Böger CA, Banas B, Eckardt K-U, Oefner PJ: T "0 z00000' (x000000 x00 00 0x000000z000z00 " 00 00 " 00 x00' (y (e d i 000 z0000z00 z (f00 " 00000000' 00' 0000 " 8Kidney Int 2011, FH11244-1253.
56. Vilasi A, Cutillas PR, Maher AD, Zirah SFM, Capasso G, Norden AWG, Holmes E, Nicholson JK, Unwin RJ: S000 y 00 " " 00000 00 z (x00 0 "0ky00000 z0000 " 00 (00 00 (" " 00 0z0000 000 0 " 00x00W0z000 0000 000 " 8Am J Physiol Renal Physiol 2007, AHB1F456-F467.
57. Moolenaar S, Engelke U, Wevers RA: g00000000 z0 x00 x' 00 0z00 000x00z" (000 z0000z00000' (y00 00f00 " 00 00 " (f" 00 00 " 0y0000 (" 000000 0 "0ky0000 8 Ann Clin Biochem 2003, C: I16-24.
58. Oostendorp M, Engelke U, Willemsen M, Wevers R: T x' 00000' (0y0000 (" 00000 0' 000 " 0 "0ky0000 0 0 00000000z00 x00 x' 00 0z00 000x00z" (000 z0000z00000Clin Chem 2006, DA11395-1405.
59. Lanpher B, Brunetti-Pierri N, Lee B: 0y0000 (" 0000000' 0 "0ky0000 l00 " (f000 (0000 (d " 00 " 00x000 (z000 000 00 " 0" x00 00000 00000 " 00 00 00 00 (0' 000 0 00000 00' 00x00 ("x00 x00 0000z00 " 000 (000 x00 (z 000 00000 " (A8Nat Genet 1997, ? D1273-276.
60. Willemsen G, Boomsma DI, Beem AL, Vink JM, Slagboom PE, Posthuma D: h k00' 000 " " 000' 00000' (x' 0000 " 0000 " 00x00 (0T 000 0y0000 0x 000Eur J Hum Genet 2004, ? A1820-828.
61. Comuzzie AG, Hixson JE, Almasly L, Mitchell BD, Mahaney MC, Dyer TD, Stern MP, MacCluer JW, Blangero J: Q00 x00000x00000000 000 000z000 " 00 00 00 (0' 000 0 00000 00' 00x00 ("x00 x00 0000z00 " 000 (000 x00 (z 000 00000 " (A8Nat Genet 1997, ? D1273-276.
62. Rotimi CN, Comuzzie AG, Lowe WL, Luke A, Blangero J, Cooper RS: k " (00x0000000 000 000z00000 (z 000 00000 " (A' 0000 000 0 00000 00 00 0 z000f000 " " (0(Q0zx007Q0 " 0zx000Diabetes 1999, CG643-644.
63. Arnett DK, Miller MB, Coon H, Ellison RC, North KE, Province M, Leppert M, Eckfeldt JH: X" 0000 " 0 " " 0000x' " (x00x0000 000z00 0000z" 00y 000000z00 '000' x0000 00000 x000 00z" 00 " 0[e [cR] (W00 00 (x000 000 08Hum Genet 2004, ?? D1468-474.
64. Duggirala R, Blangero J, Almasly L, Dyer TD, Williams KL, Leach RJ, O'Connell P, Stern MP: Q00 x000000z" 00y 00000z000 00f00 0z 00' 00000 x000 00z" 00 " (z000z "00000000 0000z00 " 000 (z 000 00000 " ? (D10(d " 0zx00Q0 " 0zx0008Am J Hum Genet 2000, EE11237-1245.
65. Arya R, Duggirala R, Almasly L, Rainwater DL, Mahaney MC, Cole S, Dyer TD, Williams K, Leach RJ, Hixson JE, MacCluer JW, O'Connell P, Stern MP, Blangero J: c 0000 " 00' (' 7 " 000000000000 00z 00 00 0000z00z" 000000000000 (x 00z00000 (z 000 00000 " (H0 (0(d " 0zx00Q0 " 0zx0008Nat Genet 2002, B: 1102-105.
66. Lilja HE, Suviolahti E, Soro-Paavonen A, Hiekkalinna T, Day A, Lange K, Sobel E, Taskinen M-R, Peltonen L, Perola M, Pajukanta P: c0z00' 00000x00000000' ([T c7z 0000 00000 (z 000 00000 " ? : 00 (W000 00 (x00 00 00 0 (" 0000 " " 0 x8/ Lipid Res 2004, CD1876-1884.
67. Petersen A-K, Stark K, Musameh MD, Nelson CP, Römisch-Margl W, Kremer W, Raffler J, Krug S, Skurk T, Rist MJ, Daniel H, Hauner H, Adamski J, Tomaszewski M, Döring A, Peters A, Wichmann H-E, Kaess BM, Kalbitzer HR, Huber F, Pfahler V, Samani NJ, Kronenberg F, Dieplinger H, Illig T, Hengstenberg C, Suhre K, Gieger C, Kastenmüller G: X" 00 00z (x000z x00000 0 00000000 0 (00y' 00z0000000000 " (00 000 x00000000 " 0y 0000 zx00x0000 8Hum Mol Genet 2012, A211433-1443.
68. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park J-H, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Mägi R, Pastinen T, Liang L, Heid IM, Luan Ja, Thorleifsson G, et al: [000 0 " 000 000x000z00000 0 " (0 " " 0000 z000 (x00' (y 0000 zx00x00 x000xf" z00 00 x00 (" " 08Nature 2010, CEF1832-838.
69. Jansen RC, Nap JP: X" 00 0zx00 " 0000 z000 " (x' " " " 00000' (000 00' 0' x0008 Trends Genet 2001, ? F1388-391.
70. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH: X" 00 00z00 00' (" 00" (" 0000 00000000 00 " (00 x 00 0000 (x00 0 x008Nature 2003, CAA1297-302.
71. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KCC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WOC: Q00 " 0000 " 0 " " (x0000z x00000000 000' (" 00y x00 " 00" (" 0000 0000008Nat Genet 2007, BH11202-1207.

72. Keurentjes JJB, Fu J, de Vos CHR, Lommen A, Hall RD, Bino RJ, van der Plas LHW, Jansen RC, Vreugdenhil D, Koornneef M:
Nat Genet 2006, 34:842-849.
73. Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, Kopka J, Willmitzer L, Zamir D, Fernie AR:
Nat Biotechnol 2006, 24:447-454.
74. Cazier J-B, Kaisaki PJ, Argoud K, Blaise BJ, Veselkov K, Ebbels TMD, Tsang T, Wang Y, Bihoreau M-T, Mitchell SC, Holmes EC, Lindon JC, Scott J, Nicholson JK, Dumas M-E, Gauguier D:
J Proteome Res 2012, 11:631-642.
75. Dumas M-E, Wilder SP, Bihoreau M-T, Barton RH, Fearnside JF, Argoud K, D'Amato L, Wallis RH, Blancher C, Keun HC, Baunsgaard D, Scott J, Sidelmann UG, Nicholson JK, Gauguier D:
Nat Genet 2007, 39:666-672.
76. Klose J, Nock C, Herrmann M, Stühler K, Marcus K, Blüggel M, Krause E, Schalkwyk LC, Rastan S, Brown SDM, Büsow K, Himmelbauer H, Lehrach H:
Nat Genet 2002, 32:385-393.
77. Gieger C, Geistlinger L, Altmaier E, Hrabé de Angelis M, Kronenberg F, Meitinger T, Mewes H-W, Wichmann H-E, Weinberger KM, Adamski J, Illig T, Suhre K:
PLoS Genet 2008, 4:e1000282.
78. Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, Pohn C, Altmaier E, Kastenmüller G, Kato BS, Mewes H-W, Meitinger T, de Angelis MH, Kronenberg F, Soranzo N, Wichmann H-E, Spector TD, Adamski J, Suhre K:
Nat Genet 2010, 42:137-141.
79. Nicholson G, Rantalainen M, Li JV, Maher AD, Malmödin D, Ahmadi KR, Faber JH, Barrett A, Min JL, Rayner NW, Toft H, Krestyaninova M, Viksna J, Neogi SG, Dumas M-E, Sarkans U, Consortium M, Donnelly P, Illig T, Adamski J, Suhre K, Allen M, Zondervan KT, Spector TD, Nicholson JK, Lindon JC, Baunsgaard D, Holmes E, McCarthy MI, Holmes CC:
PLoS Genet 2011, 7:e1002270.
80. Nicholson G, Rantalainen M, Maher AD, Li JV, Malmödin D, Ahmadi KR, Faber JH, Hallgrimsdóttir IB, Barrett A, Toft H, Krestyaninova M, Viksna J, Neogi SG, Dumas M-E, Sarkans U, Consortium TM, Silverman BW, Donnelly P, Nicholson JK, Allen M, Zondervan KT, Lindon JC, Spector TD, McCarthy MI, Holmes E, Baunsgaard D, Holmes CC:
Mol Syst Biol 2011, 7:525.
81. Suhre K, Shin S-Y, Petersen A-K, Mohny RP, Meredith D, Wägele B, Altmaier E, CARDIOGRAM, Deloukas P, Erdmann J, Grundberg E, Hammond CJ, de Angelis MH, Kastenmüller G, Köttgen A, Kronenberg F, Mangino M, Meisinger C, Meitinger T, Mewes H-W, Milburn MV, Pohn C, Raffler J, Ried JS, Römisch-Margl W, Samani NJ, Small KS, Wichmann H-E, Zhai G, Illig T, et al:
Nature 2011, 475:54-60.
82. Suhre K, Wallaschowski H, Raffler J, Friedrich N, Haring R, Michael K, Wasner C, Krebs A, Kronenberg F, Chang D, Meisinger C, Wichmann H-E, Hoffmann W, Völzke H, Völker U, Teumer A, Biffar R, Kocher T, Felix SB, Illig T, Kroemer HK, Gieger C, Römisch-Margl W, Nauck M:
Nat Genet 2011, 43:565-569.
83. Mittelstrass K, Ried JS, Yu Z, Krumsiek J, Gieger C, Pohn C, Römisch-Margl W, Polonikov A, Peters A, Theis FJ, Meitinger T, Kronenberg F, Weidinger S, Wichmann HE, Suhre K, Wang-Sattler R, Adamski J, Illig T:
PLoS Genet 2011, 7:e1002215.
84. Adamski J:
Genome Med 2012, 4:1-10.
85. Davidovic L, Navratil V, Bonaccorso CM, Catania MV, Bardoni B, Dumas M-E:
Genome Res 2011, 21:2190-2202.
86. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, et al:
Nature 2010, 464:59-65.
87. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, et al:
Nature 2011, 473:174-180.
88. Waldram A, Holmes E, Wang Y, Rantalainen M, Wilson ID, Tuohy KM, McCartney AL, Gibson GR, Nicholson JK:
J Proteome Res 2009, 8:2361-2375.
89. Claus SP, Ellero SL, Berger B, Krause L, Bruttin A, Molina J, Paris A, Want EJ, de Waziers I, Cloarec O, Richards SE, Wang Y, Dumas M-E, Ross A, Rezzi S, Kochhar S, van Bladeren P, Lindon JC, Holmes E, Nicholson JK:
MBio 2011, 2:e00271-00210.
90. Turnbaugh PJ, Gordon JL:
Cell 2008, 134:708-713.
91. Blow N:
Nat Methods 2008, 5:267-274.
92. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ:
Nat Genet 2010, 42:30-35.
93. St Hilaire C, Ziegler SG, Markello TC, Brusco A, Groden C, Gill F, Carlson-Donohoe H, Lederman RJ, Chen MY, Yang D, Siegenthaler MP, Arduino C, Mancini C, Freudenthal B, Stanescu HC, Zdebik AA, Chaganti RK, Nussbaum RL, Kleta R, Gahl WA, Boehm M:
N Engl J Med 2011, 364:432-442.
94. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, Serpe JM, Dasu T, Tschannen MR, Veith RL, Basehore MJ, Broeckel U, Tomita-Mitchell A, Arca MJ, Casper JT, Margolis DA, Bick DP, Hessner MJ, Routes JM, Verbsky JW, Jacob HJ, Dimmock DP:
Genet Med 2011, 23:255-262.
95. [http://omim.org/statistics/entry]
96. Green ED, Guyer MS, Institute NHGR:
Nature 2011, 473:204-213.
97. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J:
Nat Rev Genet 2011, 12:745-755.
98. Robinson PN, Krawitz P, Mundlos S:
Clin Genet 2011, 89:127-132.
99. Yang H-C, Chen C-W:
BMC Proc 2011, 5:S43.
100. Consortium GP:
Nature 2010, 464:1061-1073.
101. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Jakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HYK, Leng J, Li R, Li Y, Lin C-Y, Luo R, et al:
Nature 2011, 470:59-65.
102. Biesecker LG, Mullikin JC, Facio FM, Turner C, Cherukuri PF, Blakesley RW, Bouffard GG, Chines PS, Cruz P, Hansen NF, Teer JK, Maskeri B, Young AC, Program NCS, Manolio TA, Wilson AF, Finkel T, Hwang P, Arai A, Remaley AT, Sachdev V, Shamburek R, Cannon RO, Green ED:
Genome Res 2009, 19:1665-1674.
103. Feenstra B, Skovgaard IM, Broman KW:
Genetics 2006, 172:2269-2282.
104. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC:
Am J Hum Genet 2007, 81:559-575.
105. Blaise BJ, Shintu L, Elena B, Emsley I, Dumas M-E, Toulhoat P:
Anal Chem 2009, 81:6242-6251.

Cite this article as: Robinette SL, et al:

Genome Medicine 2012, 30.

Appendix E: Paper VII

□□□□□□□□ □□□□□□□□ □□□□ □□□□ □□□□□□ □□□□ □Robinette SL, Brüscheiler R, Schroeder F, Edison AS. Metabolomics and Natural Products Research: Two Sides of the Same Coin. *Acc Chem Res*, 2011. 45(2):288–297.) □□□□□□□□ □□□□□□ □□ □□□□□□ □□□□ □□□□□□□□□□□□

NMR in Metabolomics and Natural Products Research: Two Sides of the Same Coin

STEVEN L. ROBINETTE,[†] RAFAEL BRÜSCHWEILER,[‡]
FRANK C. SCHROEDER,[§] AND ARTHUR S. EDISON^{*,‡}

[†]Department of Surgery and Cancer, Faculty of Medicine, Imperial College London,
Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK,

[‡]Department of Chemistry & Biochemistry and National High Magnetic Field
Laboratory, Florida State University, Tallahassee, Florida 32306, United States,

[§]Boyce Thompson Institute and Department of Chemistry and Chemical Biology,
Cornell University, Ithaca, New York 14853, United States, and ^{*}Department of
Biochemistry & Molecular Biology and National High Magnetic Field Laboratory,
University of Florida, PO Box 100245, Gainesville, Florida 32610-0245, United States

RECEIVED ON JUNE 14, 2011

CONSPECTUS

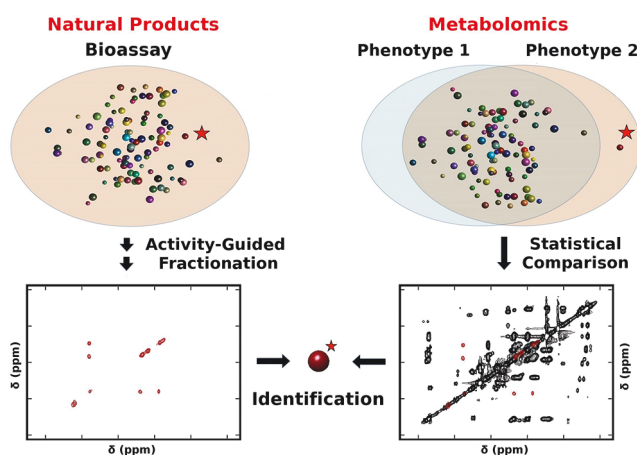
Small molecules are central to biology, mediating critical phenomena such as metabolism, signal transduction, mating attraction, and chemical defense. The traditional categories that define small molecules, such as metabolite, secondary metabolite, pheromone, hormone, and so forth, often overlap, and a single compound can appear under more than one functional heading. Therefore, we favor a unifying term, biogenic small molecules (BSMs), to describe any small molecule from a biological source.

In a similar vein, two major fields of chemical research, natural products chemistry and metabolomics, have as their goal the identification of BSMs, either as a purified active compound (natural products chemistry) or as a biomarker of a

particular biological state (metabolomics). Natural products chemistry has a long tradition of sophisticated techniques that allow identification of complex BSMs, but it often fails when dealing with complex mixtures. Metabolomics thrives with mixtures and uses the power of statistical analysis to isolate the proverbial “needle from a haystack”, but it is often limited in the identification of active BSMs. We argue that the two fields of natural products chemistry and metabolomics have largely overlapping objectives: the identification of structures and functions of BSMs, which in nature almost inevitably occur as complex mixtures.

Nuclear magnetic resonance (NMR) spectroscopy is a central analytical technique common to most areas of BSM research. In this Account, we highlight several different NMR approaches to mixture analysis that illustrate the commonalities between traditional natural products chemistry and metabolomics. The primary focus here is two-dimensional (2D) NMR; because of space limitations, we do not discuss several other important techniques, including hyphenated methods that combine NMR with mass spectrometry and chromatography.

We first describe the simplest approach of analyzing 2D NMR spectra of unfractionated mixtures to identify BSMs that are unstable to chemical isolation. We then show how the statistical method of covariance can be used to enhance the resolution of 2D NMR spectra and facilitate the semi-automated identification of individual components in a complex mixture. Comparative studies can be used with two or more samples, such as active vs inactive, diseased vs healthy, treated vs untreated, wild type vs mutant, and so on. We present two overall approaches to comparative studies: a simple but powerful method for comparing two 2D NMR spectra and a full statistical approach using multiple samples. The major bottleneck in all of these techniques is the rapid and reliable identification of unknown BSMs; the solution will require all the traditional approaches of both natural products chemistry and metabolomics as well as improved analytical methods, databases, and statistical tools.



1. The Common Theme: Biogenic Small Molecules

The identification and functional analysis of biogenic small molecules (BSMs) form the primary objectives of both natural products chemistry and metabolomics: natural products chemists are traditionally interested in the identification of new molecular structures with biological activity, whereas metabolomics and the related field of metabonomics have focused on correlating known BSMs with specific biological properties. Despite similar objectives, application of similar techniques, and frequent study of the same organisms, natural products and metabolomics research have remained largely separate fields, lacking regular interaction and exchange of ideas.

BSMs control intracellular processes (metabolism), intercellular processes (nervous system and hormonal regulation), intraspecific communication (e.g., via pheromones), and interspecific interactions (e.g., via defensive compounds); as a result, many drugs are derived from BSMs.¹ Many biogenic small molecules function at multiple levels, making it difficult and often artificial to differentiate between “metabolites”, that is, compounds that are part of primary metabolism, and “secondary metabolites”, that is, compounds not necessarily required for metabolic functioning. Similarly, functional categorizations such as “hormone”, “pheromone”, “biosynthetic intermediate”, or “catabolite” can be problematic. For example, citric acid cycle intermediates are ligands of GPCRs that function in intercellular signaling.² Therefore, for the purpose of this review, we use “biogenic small molecules” (BSMs) as a unifying term to refer to all “metabolites”, “secondary metabolites”, and “natural products”. Small molecules generally have molecular weights less than 1500 Da, but it should be noted that BSMs are further distinguished from larger biomolecules such as proteins and nucleic acids in that they are not strictly derived from a small number of known building blocks. As a result, the molecular structures of BSMs can be highly diverse and irregular, and their identification and characterization can present great analytical challenges.

This separation of natural products research and metabolomics appears ultimately rooted in differences of experimental design. The basic goal of both fields is to take complex mixtures of BSMs and identify a subset of compounds that describe a biological process or possess intrinsic biological activity. Whereas natural products research has relied on activity guided fractionation to isolate simple mixtures or individual compounds, metabolomics researchers have replaced chemical isolation with direct comparative

analysis of complex mixtures to statistically identify subsets of BSMs relevant in a specific biological context.

We think that the fields of metabolomics and natural products are essentially “two sides of the same coin” and would both benefit from an increased exchange of expertise and approach. In natural products research, the process of isolating individual components can lead to chemical modification or degradation of the BSMs of interest. Furthermore, activity-guided fractionation necessarily risks losing important biological information that was encoded in the original BSM mixture, because most of the sample is not analyzed. The use of activity-guided fractionation therefore often fails in cases where several BSMs act in synergy. For example, a family of BSMs recently identified in the model organism *Caenorhabditis elegans* act synergistically as a mate-attracting pheromone, but most of this activity is lost when individual components of the pheromone are separated during fractionation.³

On the other hand, metabolomics uses statistical analysis of NMR and mass spectra of complex BSM mixtures to detect spectral features that correlate to a phenotype or biological property of interest and their response to stimuli. One of the most pressing challenges in metabolomics is the subsequent chemical identification of the BSMs represented by the detected spectral features. This problem of assigning peaks to BSMs is a major bottleneck in the typical metabolomics workflow. Structure elucidation techniques commonly applied in natural products research, especially integrated use of 2D NMR spectroscopy and mass spectrometry, may hold the key to increasing the number of identifiable compounds in metabolomic analyses.

2. Experimental Techniques That Bridge Natural Products and Metabolomics

In this Account, we focus on several approaches to mixture analysis by NMR. Each uses different sampling strategies, data sets, and computational algorithms, but they all illustrate methods that bridge natural products and metabolomics. We start with the most chemically intuitive approach of simply using 2D NMR spectra to analyze unfractionated mixtures for unstable BSMs. Next, we show that the statistical method of covariance can be used to enhance the resolution of 2D NMR spectra and thus enable *in silico* separation of BSMs in a mixture for database or *ab initio* identification using individual spectra. We then describe a comparative approach in which 2D NMR spectra from two different organisms with different genetic backgrounds are

compared to highlight unique BSMs. The final section highlights some of the powerful methods of statistically extracting specific chemical information from 1D and 2D NMR data of a particular biological state using tools of multivariate analysis. Specific technical aspects of NMR spectroscopy,⁴ hyphenated techniques,⁵ statistical analysis,^{6,7} and overall protocol development⁸ have been reviewed elsewhere.

2.1. Direct Observation of Mixtures. Two-dimensional NMR spectroscopy is arguably the most important spectroscopic technique for the elucidation of novel or unexpected structures.^{4,9} Two-dimensional NMR spectra are unique in that they provide direct evidence for atom connectivity and spatial arrangements. In contrast to metabolomics, which routinely uses 1D NMR spectra for the characterization of complex, often entirely unfractionated, BSM mixtures,¹⁰ natural product researchers traditionally relied on pure, isolated samples for analysis, whereas definite identification of novel compounds from complex mixtures was usually not pursued or deemed possible. However, recent examples have demonstrated that identification of new compounds via 2D NMR spectroscopy of complex mixtures not only is feasible but frequently offers significant advantages over fractionation-based approaches.

Development of methods for identifying novel structures via 2D NMR analyses of mixtures was motivated by several instances in which traditional activity-guided fractionation failed to reveal the BSMs of interest. In one of the first examples, the poison gland secretion of *Myrmecaria* ants was suspected to contain highly toxic alkaloids; however, chromatographic fractionation revealed only nontoxic monoterpene hydrocarbons in some samples. Subsequent 2D NMR analyses of unfractionated ant venom then revealed the highly unstable heptacyclic alkaloid myrmecarin 430A, representing one of the first examples for natural products based on the oligomeric assembly of several similarly functionalized fatty-acid chains.^{11,12} Myrmecarin 430A, whose intriguing structure and biogenesis has spawned several efforts toward its total synthesis,¹³ represents one of the first examples of BSMs that were identified without ever having been isolated. Similarly, a 2D NMR spectroscopic screen of a library of spider venom samples revealed that sulfated nucleosides, which despite their structural simplicity had not previously been found in nature, form major venom components in several spider species, including the infamous brown recluse, *Loxosceles reclusa* (Figure 1).^{14,15} These arthropod examples showed that high-resolution 2QF-COSY spectra are well suited for detecting and characterizing unknown BSMs from complex

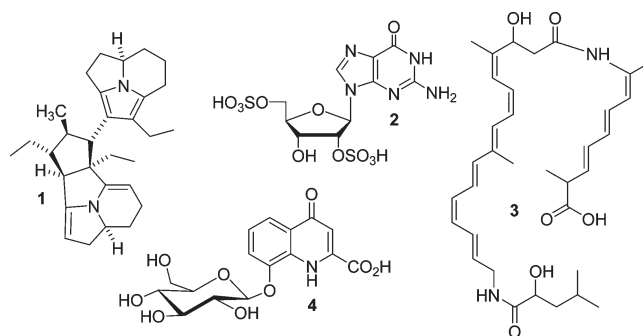


FIGURE 1. Structures of BSMs (biogenic small molecules) identified via NMR-spectroscopic analysis from largely unfractionated metabolite mixtures. Myrmecarin 430A (**1**) and bacillaene (**3**) represent members of a small but growing class of natural products that have never been isolated in pure form.^{11,12,15,35} The sulfated nucleoside **2** was identified from spider venom.¹⁵ The xanthurenic acid derivative **4** is a natriuretic identified from partially fractionated human urine samples.⁴⁹

mixtures because they provide detailed structural information and often permit interpretation of overlapping signals, one primary challenge for using NMR spectroscopy for mixtures.

2.2. Statistical Methods for Identification of BSMs from Single Biological Mixtures. The popularity of 1D ¹H NMR in metabolomics and metabonomics applications is primarily founded on the simplicity and efficiency of data collection. Natural product research, on the other hand, has adopted from early on the powerful repertoire of 2D NMR techniques. Due to the narrow NMR linewidths of small molecules, 1D ¹H NMR spectra of mixtures can provide some rather specific information about the mixture components. They are particularly useful in situations where the pool of component candidates is limited, for example in the case of a urine sample.¹⁶ In such cases, the observation of a single non-overlapping peak multiplet of a mixture component permits identification of the component and determination of its relative concentration with high confidence. On the other hand, 1D ¹H NMR spectra of mixtures containing unknown components permit neither the determination of the 1D spectral traces that belong to individual components nor the assignment of peaks to specific atoms. In such cases, the use of 2D NMR methods becomes indispensable. *J*-resolved NMR spectra¹⁷ are useful for simplifying overlap in 1D ¹H NMR but lack important atomic correlations that are the cornerstone of structure determination via 2D NMR; therefore, this approach is not included in this Account.

The main drawback of 2D Fourier transform (FT) NMR spectroscopy is the time required to collect the indirect second dimension at high digital resolution. This dimension is acquired in the time domain by repeating the same pulse

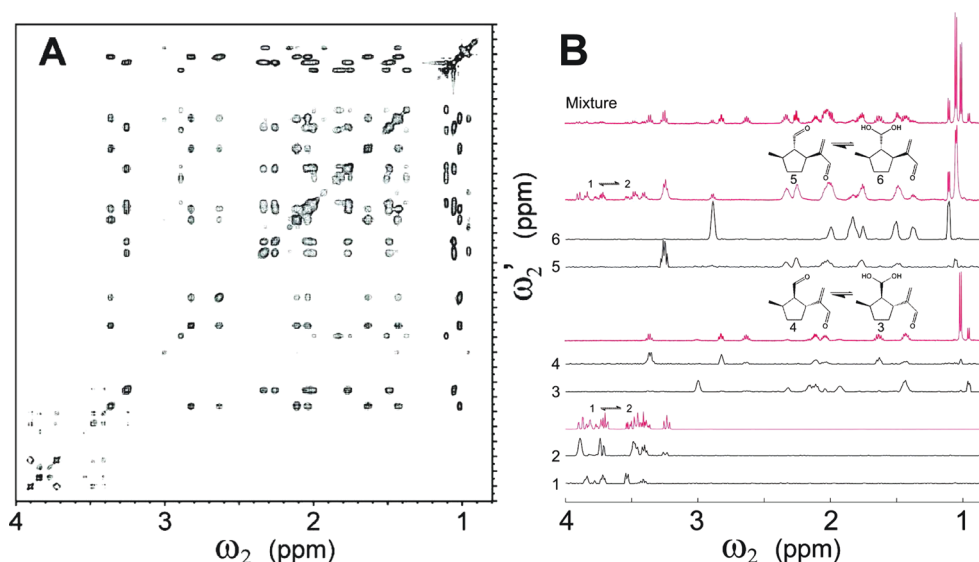


FIGURE 2. (A) Aliphatic section of covariance proton TOCSY spectrum of defensive secretion of a single walking stick insect. (B) One-dimensional ^1H NMR spectrum of the mixture. The six black spectra are covariance TOCSY traces extracted from covariance TOCSY of panel A using the DemixC approach. The bottom three red spectra are reference 1D spectra of purified components. Each reference spectrum contains two species, α -glucose (trace 1) and β -glucose (trace 2); dialdehyde and diol forms of the anisomorphal (traces 4 and 3, respectively); and the peruphasmal (traces 5 and 6, respectively) monoterpenes. Chemical structures of the anisomorphal and peruphasmal and their corresponding geminal diols are shown as insets. Figure is adapted from ref 25.

sequence N_1 times with the evolution time t_1 incremented from experiment to experiment. Because the digital spectral resolution is determined according to the Nyquist theorem by $\Delta\nu = \text{SW}/N_1$, where SW is the spectral width, a relatively large number of increments, $N_1 > 256$, is required to obtain acceptable spectral resolution along the indirect dimension. This sampling requirement, which is independent of sensitivity, becomes more severe at higher magnetic field strengths $\mathbf{B}_0$ as $\text{SW} \propto \mathbf{B}_0$.

Covariance NMR^{18,19} represents an alternative approach to 2D FT that overcomes some of these limitations. Instead of applying FT along the indirect dimension, the 2D correlation information is reconstructed by statistical means through covariance yielding a symmetric spectrum **C** that has the same high resolution along the indirect and direct dimensions. In this way, the minimal number of t_1 increments required can be reduced to $N_1 < 100$.²⁰ It can be shown that **C** is mathematically related to the 2D FT spectrum **F** through $\mathbf{C} = \text{sqr}(\mathbf{F}^T \cdot \mathbf{F})$ ¹⁹ where *sqr* denotes the matrix square root, which can be efficiently computed via singular value decomposition (SVD).²¹ The covariance method can be directly applied to TOCSY and NOESY-type spectra, whereas for COSY an additional regularization step should be applied.²²

While for short mixing times the information content of TOCSY and COSY spectra are similar, for longer mixing times

(~100 ms) typically all protons that belong to the same molecule or spin system show 2D cross-peaks to each other. This property conveniently permits the spectral deconvolution of the mixture into 1D NMR traces that correspond to individual components. Such a decomposition can be achieved by non-negative matrix factorization (NMF)²³ or bottom-up clustering of all 1D cross sections of a covariance TOCSY spectrum using the DemixC method.²⁴ Application of the method to the analysis of a single defensive spray milking from an adult female walking-stick insect *Anisomorpha buprestoides* using a 1-mm high-temperature superconducting NMR probe revealed that it contains, in addition to glucose, two stereoisomeric terpenes, each with a dialdehyde and a diol in slow chemical exchange (Figure 2).²⁵

DemixC traces can be treated like 1D NMR spectra, and hence, they are well suited for compound identification by database searching. For this purpose, a peak list is generated for a given DemixC trace, which is then queried against the peak lists of the components of an NMR database, such as the metabolomics BMRB²⁶ or HMDB,²⁷ as implemented in the COLMAR suite of web servers^{28,29} (<http://spinportal.magnet.fsu.edu>). NMR databases^{26,27,30} are improving and expanding rapidly, and correspondingly, their importance for metabolomics and natural products research is growing. Detailed descriptions of these databases are available at the listed Web sites.

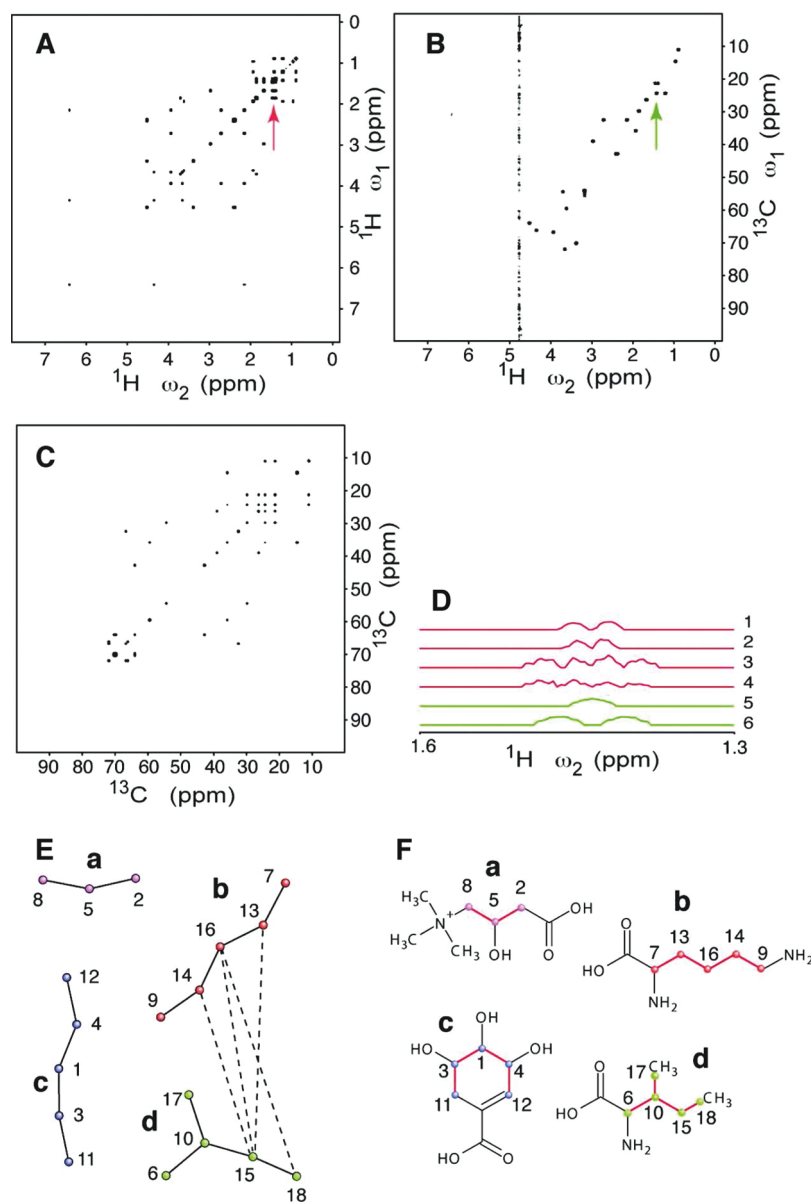


FIGURE 3. Doubly indirect covariance spectroscopy of model mixture containing carnitine, lysine, isoleucine, and shikimate: (A) 2QF-COSY spectrum; (B) ^{13}C – ^1H HSQC spectrum; (C) doubly indirect covariance spectrum; (D) absolute value cross sections along ^1H dimension of COSY and HSQC for the overlap at $\omega_2 = 1.45$ ppm (traces 1–4 belong to COSY and traces 5 and 6 to HSQC); (E) carbon–carbon connectivity graphs of mixture components. Dashed lines indicate extraneous connectivities to nodes 15 and 16 due to ^1H overlap at 1.45 ppm as indicated by arrows in panels A and B. They are identified by filtering according to differential peak positions and shapes (1st and 2nd moments) (panel E) or, alternatively, by diffusion order spectroscopy (DOSY). The labels a, b, c, and d in panels E and F belong to carnitine, lysine, shikimate, and isoleucine, respectively. Figure is adapted from ref 33.

Database matching has also been demonstrated for heteronuclear 2D ^{13}C – ^1H HSQC³¹ and HSQC-TOCSY NMR³² spectra of mixtures. Although at natural ^{13}C abundance the intrinsic sensitivity of HSQC is lower than for TOCSY, along the ^{13}C dimension the fully decoupled HSQC spectrum displays very sharp peaks with low probability of cross-peak overlap. Moreover, HSQC spectra only yield correlations between directly bonded ^1H – ^{13}C nuclei and are therefore unsuitable to trace the carbon backbones of individual

mixture components. This objective can be achieved by the doubly indirect covariance NMR method combining a 2D COSY spectrum **Y** with a 2D HSQC spectrum **H** via basic matrix operations: $\mathbf{D} = \mathbf{H} \cdot \mathbf{Y} \cdot \mathbf{H}^T$.³³ **D** is a symmetric ultra-high resolution ^{13}C – ^{13}C correlation spectrum that can be analyzed using graph theory to identify the carbon skeletons of individual mixture components, as shown in Figure 3. In case of proton overlap, additional connectivities do occur. They can be suppressed by an unsupervised “moment”

filtering method³⁴ that removes false correlations based on differential peak positions and peak widths (Figure 3). Besides the carbon backbone topology, the method also provides ¹³C and, via HSQC, ¹H chemical shifts, which are useful reporters on the nature of additional chemical groups, for example, a phosphate or amino group, attached to these carbons that are not directly visible in the COSY and HSQC spectra. The doubly indirect covariance approach is designed for studying samples from uncharted territories where not only the concentrations of mixture components but also their chemical structures are not *a priori* known, thereby bridging the fields of natural product research and metabolomics.

2.3. Comparative Analysis of Biological Mixtures. Statistical techniques used in comparative metabolomics will be particularly useful for identifying new BSMs associated with specific phenotypes or genotypes, perhaps the largest looming challenge in chemical biology and natural products chemistry. For most known BSMs in nonmammalian systems, their biological functions and biosynthetic heritage are not known, and correspondingly, many genes believed to play a role in small molecule biosynthesis have remained orphans. The recent identification of the antibiotic bacillaene from *Bacillus subtilis* using differential analyses of 2D NMR spectra (DANS) illustrates the potential for adapting comparative approaches to natural products research.^{35,36} Previous studies had shown that the very large pksX gene cluster (~2% of the *B. subtilis* genome) encodes an unusual hybrid PKS/NRPS (polyketide synthase/nonribosomal peptide synthetase) that produced a small molecule with antibiotic properties. Despite copious production of this metabolite, its structure had remained undetermined because all isolation attempts had failed as a result of chemical instability. Ultimately, bacillaene was identified via 2D NMR spectroscopic comparison of the unfractionated metabolomes of a pksX knockout and a pksX-expressing strain. DANS analysis of 2QF-COSY spectra obtained for these two metabolite samples enabled straightforward identification of spectral features that were present in the pksX-expressor but absent in the knockout, which formed the basis for subsequent identification of bacillaene's complete structure (Figure 1).³⁵

A similar approach was used for the identification of a component of the male-attracting pheromone of the nematode *Caenorhabditis elegans*. Earlier studies based on activity-guided fractionation of worm metabolite extracts had shown that *C. elegans* hermaphrodites produce a male-attracting blend of derivatives of the dideoxysugar ascarylose.³

However, due to strong synergism, not all components of this pheromone blend could be identified via fractionation. Based on the observation that *daf-22* mutant worms do not produce the male attractant, Pungaliya et al. employed DANS to compare the *C. elegans* wild-type and *daf-22* mutant metabolomes, which revealed the unanticipated *p*-amino-benzoic acid derivative ascr#8 as an important component of the male-attracting pheromone (Figure 4).³⁷ The identification of ascr#8 from *C. elegans* and bacillaene from *B. subtilis* illustrate that using comparative approaches for the analysis of whole-metabolome 2D NMR spectra offers significant opportunities for natural products chemistry and chemical biology.

2.4. Statistical Methods for Identifying BSMs from Multiple Biological Mixtures. The combination of untargeted spectral analysis of complex small molecule mixtures with statistical pattern recognition techniques for direct spectral comparison has transformed the study of metabolite associations with disease, gene function, and drug metabolism.^{38,39} Quantitative comparisons of spectra from two or more distinct biological states has been a defining feature of an approach that has been referred to as "metabolomics", "metabonomics", and "metabolic profiling".^{38,39} As such, the statistical data analysis methodology for comparing spectra is as significant a part of metabolomics as the underlying analytical chemistry technique that is used to generate data sets. Two major classes of statistical tools are used in metabolomics: those that identify relationships between biological states and spectral signals and those that look for correlations between spectral signals themselves.

Most metabolomics studies make use of statistical techniques to look for correlations between spectral signals and the biological states of the analyzed samples. Examples of these tools include principal component analysis (PCA), partial least squares (PLS), orthogonal projection onto latent structures (O-PLS), and discriminant analysis (PCA-DA, PLS-DA, O-PLS-DA).⁶ In the general implementation of a metabolomics study, two biological states are chosen such that the majority of BSMs are unchanged, but certain BSMs that are sensitive to differences in the two biological states are quantitatively or qualitatively different. Pattern recognition tools then act as a filter to remove the spectroscopic signals from the shared metabolites from consideration while highlighting the signals arising from the metabolites present in one state or the other (Figure 5).⁶

Although the use of 2D NMR spectra in metabolomics has generally lagged behind the use of 1D NMR or mass

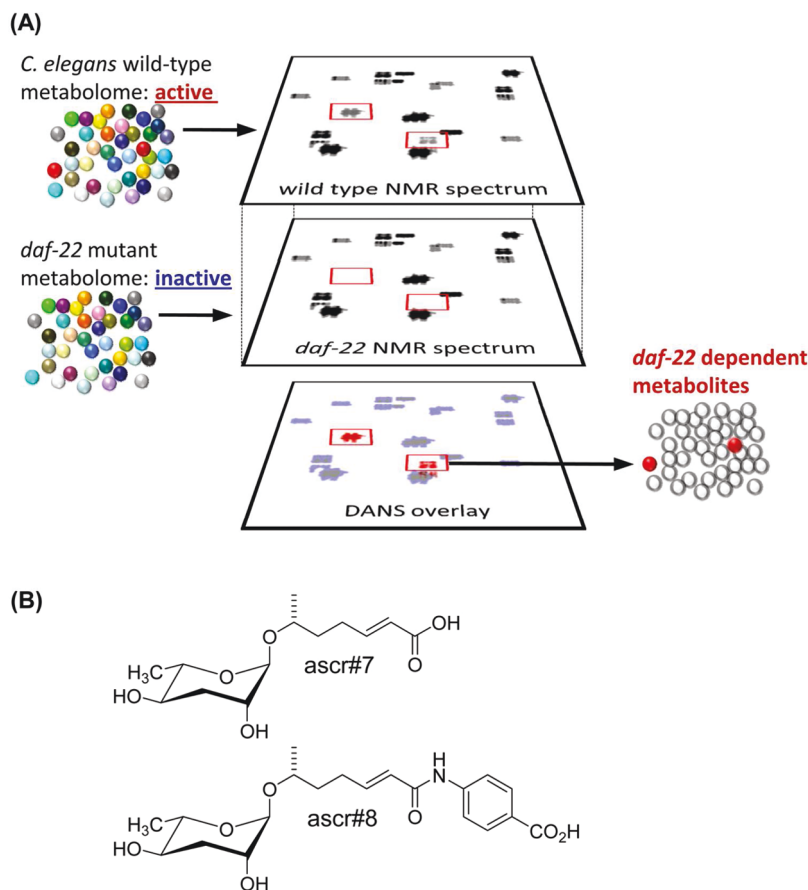


FIGURE 4. Identification of male-attracting pheromones in *C. elegans* via differential analysis of 2D NMR spectra (DANS): (A) overlay of COSY spectra from *C. elegans* wild-type and *daf-22* mutant metabolomes reveals *daf-22*-dependent signals; (B) structures of two new metabolites, ascr#7 and ascr#8, that were identified from additional analysis of the *daf-22*-dependent signals.³⁷

spectra,⁴⁰ a recent example demonstrates that full-resolution 2D NMR spectra can be used for statistical analysis (e.g., via PCA) following proper alignment. Using a peak alignment algorithm termed HATS-PR (hierarchical alignment of two-dimensional spectra - pattern recognition), Robinette et al. compared sets of TOCSY spectra derived from metabolome samples of two nematode species, *Pristionchus pacificus* and *Panagrellus redivivus*.⁴¹ The method combined the strengths of traditional 1D statistical methods and 2D NMR data, illustrating how a combination of natural products and metabolomics approaches can be used together and providing a framework for applying comparative statistical analysis to BSM discovery.

While the majority of the statistical tools in metabolomics seek to identify correlations between signals and biological states, in the past 5 years there have been a number of methods developed to identify correlations of spectroscopic signals to other signals. The best known of these methods is statistical total correlation spectroscopy (STOCSY).^{42,43} While covariance TOCSY^{18,19,21} (see section 2.2) identifies

correlations between signals by assessing their covariances across multiple t_1 -increments collected from a single sample, STOCSY uses 1D spectra of multiple biological replicates in the place of t_1 -increments. Because all the signals from a given compound increase and decrease together with the concentration of the compound, peaks arising from the same compound exhibit high positive statistical correlations. Because STOCSY does not depend on physical phenomena such as magnetization transfer, it can correlate signals even across multiple spin systems that could not be correlated by TOCSY.

Though STOCSY was applied to identify correlations between protons within the same spectrum, it was quickly realized that statistical correlations could be identified across multiple types of spectra. Heteronuclear STOCSY, or Het-STOCSY, has been applied to identify ^1H – ^{19}F and ^1H – ^{31}P correlations from multiple biological replicates of 1D ^1H and 1D ^{19}F or ^{31}P spectra.^{44,45} Additionally, peaks from NMR and mass spectra have been correlated using statistical heterospectroscopy (SHY).⁴⁶ This approach is particularly promising given that while ascertaining both molecular formula and the

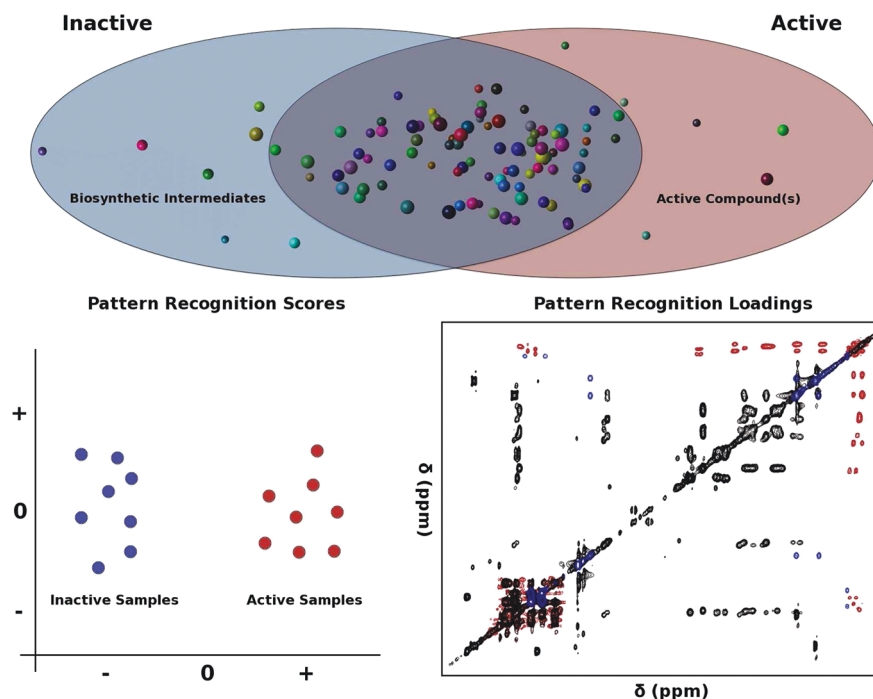


FIGURE 5. When two biological mixtures (blue and red ovals) share most compounds (colored spheres) but differ in biological activity, pattern recognition techniques applied to NMR spectra of the mixtures can identify the spectral signatures of the differential compounds. Multivariate statistical tools such as PCA and PLS decompose a set of mixture spectra into a score for each sample spectrum and loadings for each chemical shift in the NMR data set. When pattern recognition produces scores that separate the active (red circles) and inactive mixtures (blue circles), the loadings indicate which signals arise from compounds specific to the active (red cross-peaks) and inactive (blue cross-peaks) mixtures.

structural restraints encoded in NMR spectra are critical for structural elucidation, there are few methods for correlating nuclear resonances with molecular weights in mixtures.

One drawback of these techniques is that any small molecules that covary in concentration due to biological factors such as coregulation will exhibit statistical correlations as well, which can lead to ambiguity about whether a given correlation between two peaks represents a structural or biological relationship.⁴⁷ A recent extension of STOCSY termed cluster analysis statistical spectroscopy (CLASSY) has focused on using the network structure of correlations to resolve local clusters of tightly correlated peaks.⁴⁸ In the future, clustering both signals and states using two-way or bidustering techniques should provide a means of both identifying signals specific to an active biological state and grouping correlated signals as a first step in the identification of BSMs. While “signal–state” correlation tools are useful for identifying signals arising from a biologically active small molecule, they also could be extremely useful in structure elucidation.

3. Conclusion/Future Perspectives

Long-established NMR techniques in natural product research and the more recent statistical techniques of metabolomics

both strive to identify BSMs and correlate them to activity. As described in this Account, several approaches using NMR take advantage of the strengths of both approaches in solving important biological problems. Advances in analytical instrumentation in NMR and mass spectrometry as well as improved databases and computational resources are fueling these new capabilities. However, a great amount of work remains. Some important challenges include the following:

- *Improved integration of NMR and MS data.* The two techniques are complementary and both are often required for BSM identification. However, they have very different sensitivities and dynamic ranges. MS is very sensitive, but not all compounds are easily detected. NMR is almost a universal detector, but it suffers from low sensitivity. Despite advances, such as SHY,⁴⁶ this is still a difficult problem that limits many studies.
- *Improved Databases.* This has improved significantly in the past few years, but the size of NMR databases is still relatively small, and many NMR chemical shifts are dependent on solvent, sample concentration, and other acquisition parameters, which makes exact matching a challenge.

- **Rapid Identification of BSMs.** Ultimately, the goal in virtually every natural products or metabolomics study is the identification of active BSMs. Acceleration and automation of the identification process remains the major bottleneck, both because of limitations described above, and perhaps more importantly, because of the great chemical diversity in BSMs. Although comparison to experimental reference spectra may enable identification of common abundant metabolites, a large number of minor and trace components, especially organism-specific metabolites, usually remain unidentified in automated analyses. There is no universal identification algorithm or protocol that can be robustly applied to all classes of molecules, and thus exists a major need for the further development of methods that can build on the recent progress made, including the work described in this Account.

This work was supported by the National Institutes of Health (Grant R01 GM066041 to R.B., Grant R01 GM088290 to F.C.S., and Grant R01 GM085285 to A.S.E.). S.L.R acknowledges support from the NSF Graduate Research Fellowship Program and from the Marshall Aid Commemoration Commission.

BIOGRAPHICAL INFORMATION

Steven L. Robinette received his B.S. in Biochemistry and Molecular Biology from the University of Florida in 2010 and is currently a Ph.D. student in Jeremy Nicholson's group at Imperial College London. He is the recipient of a Marshall Scholarship and an NSF predoctoral fellowship and is a NIH Oxford–Cambridge Scholar.

Rafael Brünschweiler studied Physics at ETH Zürich and obtained his Ph.D. in 1991 at the Laboratory for Physical Chemistry at ETH under the direction of Prof. Richard R. Ernst. He then joined the Scripps Research Institute, La Jolla, California, as a postdoctoral fellow before returning to ETH as an Oberassistent. In 1998, he accepted the Carlson Chair position of Chemistry and Biochemistry at Clark University, Worcester, Massachusetts, before joining Florida State University as George M. Edgar Professor of Chemistry and Biochemistry and the National High Magnetic Field Laboratory as Associate Director for Biophysics in 2004. His interests in Biophysical Chemistry focus on the development and application of NMR and computational methods for the analysis of the structure, dynamics, interactions, and function of biological molecules.

Frank C. Schroeder received undergraduate degrees in chemistry and physics and conducted his Ph.D. research in natural product chemistry with Wittko Francke at the University of Hamburg, Germany. He continued with postdoctoral studies in Jerrold Meinwald's laboratory at Cornell University and

subsequently joined the laboratory of Jon Clardy at Harvard Medical School, where he developed NMR spectroscopic approaches for the identification and functional characterization of small molecules in model organisms. In 2007, he joined the faculty of the Boyce Thompson Institute and Department of Chemistry and Chemical Biology at Cornell University.

Arthur S. Edison obtained a B.S. in chemistry from the University of Utah where he conducted research with William Epstein and David Grant. He completed his Ph.D. in biophysics from the University of Wisconsin—Madison under the supervision of John L. Markley and Frank Weinhold. In 1993, Dr. Edison joined the laboratory of Anthony O. W. Stretton at the University of Wisconsin—Madison as a Jane Coffin Childs postdoctoral fellow, and he joined the faculty at the University of Florida and the National High Magnetic Field Laboratory in 1996 where he is currently Professor of Biochemistry & Molecular Biology and Director of Chemistry & Biology at the NHMFL.

FOOTNOTES

*Corresponding author. E-mail address: aedison@ufl.edu.

REFERENCES

- Newman, D. J.; Cragg, G. M. Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod.* **2007**, *70*, 461–477.
- He, W.; Miao, F. J.; Lin, D. C.; Schwandner, R. T.; Wang, Z.; Gao, J.; Chen, J. L.; Tian, H.; Ling, L. Citric acid cycle intermediates as ligands for orphan G-protein-coupled receptors. *Nature* **2004**, *429*, 188–193.
- Srinivasan, J.; Kaplan, F.; Ajredini, R.; Zachariah, C.; Alborn, H. T.; Teal, P. E.; Malik, R. U.; Edison, A. S.; Sternberg, P. W.; Schroeder, F. C. A blend of small molecules regulates both mating and development in *Caenorhabditis elegans*. *Nature* **2008**, *454*, 1115–1118.
- Edison, A. S.; Schroeder, F. C.; Lew, M.; Hung-Wen, L. NMR - Small Molecules and Analysis of Complex Mixtures. In *Comprehensive Natural Products II*; Elsevier: Oxford, 2010; pp 169–196.
- Wolfender, J. L.; Ndjoko, K.; Hostettmann, K. Liquid chromatography with ultraviolet absorbance-mass spectrometric detection and with nuclear magnetic resonance spectroscopy: a powerful combination for the on-line structural investigation of plant metabolites. *J. Chromatogr. A* **2003**, *1000*, 437–455.
- Madsen, R.; Lundstedt, T.; Trygg, J. Chemometrics in metabolomics—a review in human disease diagnosis. *Anal. Chim. Acta* **2010**, *659*, 23–33.
- Steuer, R.; Morgenthal, K.; Weckwerth, W.; Selbig, J. A gentle guide to the analysis of metabolomic data. *Methods Mol. Biol.* **2007**, *358*, 105–126.
- Kim, H. K.; Choi, Y. H.; Verpoorte, R. NMR-based metabolomic analysis of plants. *Nat. Protoc.* **2010**, *5*, 536–549.
- Forseth, R. R.; Schroeder, F. C. NMR-spectroscopic analysis of mixtures: from structure to function. *Curr. Opin. Chem. Biol.* **2011**, *15*, 38–47.
- 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*, 2692–2703.
- Schroder, F.; Sinnwell, V.; Baumann, H.; Kaib, M.; Francke, W. Myrmicarin-663: A new decacyclic alkaloid from ants. *Angew. Chem., Int. Ed. Engl.* **1997**, *36*, 77–80.
- Schroder, F.; Sinnwell, V.; Baumann, H.; Kaib, M. Myrmicarin 430A: A new heptacyclic alkaloid from Myrmecaria ants. *Chem. Commun.* **1996**, 2139–2140.
- Snyder, S. A.; Elsohly, A. M.; Kontes, F. Synthetic and theoretical investigations of myrmicarin biosynthesis. *Angew. Chem., Int. Ed. Engl.* **2010**, *49*, 9693–9698.
- Taggi, A. E.; Meinwald, J.; Schroeder, F. C. A new approach to natural products discovery exemplified by the identification of sulfated nucleosides in spider venom. *J. Am. Chem. Soc.* **2004**, *126*, 10364–10369.
- Schroeder, F. C.; Taggi, A. E.; Gronquist, M.; Malik, R. U.; Grant, J. B.; Eisner, T.; Meinwald, J. NMR-spectroscopic screening of spider venom reveals sulfated nucleosides as major components for the brown recluse and related species. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14283–14287.
- Liao, W. I.; Lin, Y. Y.; Chu, S. J.; Hsu, C. W.; Tsai, S. H. Bradyarrhythmia caused by ginseng in a patient with chronic kidney disease. *Am. J. Emerg. Med.* **2010**, *28*, 538.e5–538.e6.

- 17 Viant, M. R. Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochem. Biophys. Res. Commun.* **2003**, *310*, 943–948.
- 18 Brüschweiler, R.; Zhang, F. Covariance nuclear magnetic resonance spectroscopy. *J. Chem. Phys.* **2004**, *120*, 5253–5260.
- 19 Brüschweiler, R. Theory of covariance nuclear magnetic resonance spectroscopy. *J. Chem. Phys.* **2004**, *121*, 409–414.
- 20 Chen, Y.; Zhang, F.; Bermel, W.; Brüschweiler, R. Enhanced covariance spectroscopy from minimal datasets. *J. Am. Chem. Soc.* **2006**, *128*, 15564–15565.
- 21 Trbovic, N.; Smirnov, S.; Zhang, F.; Brüschweiler, R. Covariance NMR spectroscopy by singular value decomposition. *J. Magn. Reson.* **2004**, *171*, 277–283.
- 22 Chen, Y.; Zhang, F.; Snyder, D.; Gan, Z.; Brüschweiler-Li, L.; Brüschweiler, R. Quantitative covariance NMR by regularization. *J. Biomol. NMR* **2007**, *38*, 73–77.
- 23 Snyder, D. A.; Zhang, F.; Robinette, S. L.; Brüschweiler-Li, L.; Brüschweiler, R. Non-negative matrix factorization of two-dimensional NMR spectra: application to complex mixture analysis. *J. Chem. Phys.* **2008**, *128*, No. 052313.
- 24 Zhang, F.; Brüschweiler, R. Robust deconvolution of complex mixtures by covariance TOCSY spectroscopy. *Angew. Chem., Int. Ed.* **2007**, *46*, 2639–2642.
- 25 Zhang, F.; Dossey, A. T.; Zachariah, C.; Edison, A. S.; Brüschweiler, R. Strategy for automated analysis of dynamic metabolic mixtures by NMR. Application to an insect venom. *Anal. Chem.* **2007**, *79*, 7748–7752.
- 26 Markley, J. L.; Anderson, M. E.; Cui, Q.; Eghbalnia, H. R.; Lewis, I. A.; Hegeman, A. D.; Li, J.; Schulte, C. F.; Sussman, M. R.; Westler, W. M.; Ulrich, E. L.; Zolnai, Z. New bioinformatics resources for metabolomics. *Pac. Symp. Biocomput.* **2007**, *2007*, 157–168.
- 27 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.; Nazirova, A.; Shaykhtudinov, R.; Li, L.; Vogel, H. J.; Forsythe, I. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Res.* **2009**, *37*, D603–D610.
- 28 Robinette, S. L.; Zhang, F.; Brüschweiler-Li, L.; Brüschweiler, R. Web server based complex mixture analysis by NMR. *Anal. Chem.* **2008**, *80*, 3606–3611.
- 29 Zhang, F.; Robinette, S. L.; Brüschweiler-Li, L.; Brüschweiler, R. Web server suite for complex mixture analysis by covariance NMR. *Magn. Reson. Chem.* **2009**, *47* (Suppl 1), S118–S122.
- 30 Cui, Q.; Lewis, I. A.; Hegeman, A. D.; Anderson, M. E.; Li, J.; Schulte, C. F.; Westler, W. M.; Eghbalnia, H. R.; Sussman, M. R.; Markley, J. L. Metabolite identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol.* **2008**, *26*, 162–164.
- 31 Lewis, I. A.; Schommer, S. C.; Hodis, B.; Robb, K. A.; Tonelli, M.; Westler, W. M.; Sussman, M. R.; Markley, J. L. Method for determining molar concentrations of metabolites in complex solutions from two-dimensional ^1H – ^{13}C NMR spectra. *Anal. Chem.* **2007**, *79*, 9385–9390.
- 32 Zhang, F.; Brüschweiler-Li, L.; Robinette, S. L.; Brüschweiler, R. Self-consistent metabolic mixture analysis by heteronuclear NMR. Application to a human cancer cell line. *Anal. Chem.* **2008**, *80*, 7549–7553.
- 33 Zhang, F.; Brüschweiler-Li, L.; Brüschweiler, R. Simultaneous de novo identification of molecules in chemical mixtures by doubly indirect covariance NMR spectroscopy. *J. Am. Chem. Soc.* **2010**, *132*, 16922–16927.
- 34 Bingol, K.; Salinas, R. K.; Brüschweiler, R. Higher-rank correlation NMR spectra with spectral moment filtering. *J. Phys. Chem. Lett.* **2010**, *1*, 1086–1089.
- 35 Butcher, R. A.; Schroeder, F. C.; Fischbach, M. A.; Straight, P. D.; Kolter, R.; Walsh, C. T.; Clardy, J. The identification of bacillaene, the product of the PksX megacomplex in *Bacillus subtilis*. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1506–1509.
- 36 Moldenhauer, J.; Chen, X. H.; Borris, R.; Piel, J. Biosynthesis of the antibiotic bacillaene, the product of a giant polyketide synthase complex of the trans-AT family. *Angew. Chem., Int. Ed.* **2007**, *46*, 8195–8197.
- 37 Pungaliga, C.; Srinivasan, J.; Fox, B. W.; Malik, R. U.; Ludewig, A. H.; Sternberg, P. W.; Schroeder, F. C. A shortcut to identifying small molecule signals that regulate behavior and development in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 7708–7713.
- 38 Fiehn, O. Metabolomics—the link between genotypes and phenotypes. *Plant Mol. Biol.* **2002**, *48*, 155–171.
- 39 Nicholson, J. K. L. J. C.; Holmes, E. 'Metabonomics': Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis in biological NMR spectroscopic data. *Xenobiotica* **2000**, *9*, 1181–1189.
- 40 Zheng, M.; Lu, P.; Liu, Y.; Pease, J.; Usuka, J.; Liao, G.; Peltz, G. 2D NMR metabolomic analysis: a novel method for automated peak alignment. *Bioinformatics* **2007**, *23*, 2926–2933.
- 41 Robinette, S. L.; Ajredini, R.; Rasheed, H.; Zeinomar, A.; Schroeder, F. C.; Dossey, A. T.; Edison, A. S. Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology. *Anal. Chem.* **2011**, *83*, 1649–1657.
- 42 Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^1H NMR data sets. *Anal. Chem.* **2005**, *77*, 1282–1289.
- 43 Holmes, E.; Cloarec, O.; Nicholson, J. K. Probing latent biomarker signatures and in vivo pathway activity in experimental disease states via statistical total correlation spectroscopy (STOCSY) of biofluids: Application to HgCl_2 toxicity. *J. Proteome Res.* **2006**, *5*, 1313–1320.
- 44 Coen, M.; Hong, Y. S.; Cloarec, O.; Rhode, C. M.; Reilly, M. D.; Robertson, D. G.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Heteronuclear ^1H – ^{31}P statistical total correlation NMR spectroscopy of intact liver for metabolic biomarker assignment: Application to galactosamine-induced hepatotoxicity. *Anal. Chem.* **2007**, *79*, 8956–8966.
- 45 Keun, H.; Athersuch, T.; Beckonert, O.; Wang, Y.; Saric, J.; Shockcor, J.; Lindon, J.; Wilson, I.; Holmes, E.; Nicholson, J. Heteronuclear ^{19}F – ^1H Statistical Total Correlation Spectroscopy as a Tool in Drug Metabolism: Study of Flucloxacillin Biotransformation. *Anal. Chem.* **2008**, *80*, 1073–1079.
- 46 Crockford, D.; Holmes, E.; Lindon, J.; Plumb, R.; Zirah, S.; Bruce, S.; Rainville, P.; Stumpf, C.; Nicholson, J. Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: Application in metabolomic toxicology studies. *Anal. Chem.* **2006**, *78*, 363–371.
- 47 Alves, A. C.; Rantalainen, M.; Holmes, E.; Nicholson, J. K.; Ebbels, T. M. D. Analytic properties of statistical total correlation spectroscopy based information recovery in ^1H NMR metabolic data sets. *Anal. Chem.* **2009**, *81*, 2075–2084.
- 48 Robinette, S. L.; Veselkov, K. A.; Bohus, E.; Coen, M.; Keun, H. C.; Ebbels, T. M. D.; Beckonert, O.; Holmes, E. C.; Lindon, J. C.; Nicholson, J. K. Cluster analysis statistical spectroscopy using nuclear magnetic resonance generated metabolic data sets from perturbed biological systems. *Anal. Chem.* **2009**, *81*, 6581–6589.
- 49 Cain, C. D.; Schroeder, F. C.; Shankel, S. W.; Mitchnick, M.; Schmertzler, M.; Bricker, N. S. Identification of xanthurenic acid 8-O-beta-D-glucoside and xanthurenic acid 8-O-sulfate as human natriuretic hormones. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17873–17878.

Appendix E: Paper VIII

□□□□□□□□ □□□□□□□□□ □□□ □□□□ □□□□□ □□□□ □Veselkov KA, Vingara LK, Masson P, Robinette SL, Want E, Li JV, Barton RH, Claire BN, Walther B, Ebbels TM, Pelczer I, Holmes E, Lindon JC, Nicholson JK. Optimized pre-processing of UPLC-MS urinary metabolic profiles for improved information recovery.*Anal Chem*, 2011. 83(15):5864-5872.) □□□□□□□□□□□□□□□ □ □□□□□ □□□□ □□□□□□□□□□□□

Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery

Kirill A. Veselkov,[†] Lisa K. Vingara,[†] Perrine Masson,[§] Steven L. Robinette,[†] Elizabeth Want,[†] Jia V. Li,[†] Richard H. Barton,[†] Claire Boursier-Neyret,[§] Bernard Walther,[§] Timothy M. Ebbels,[†] István Pelczer,^{||} Elaine Holmes,[†] John C. Lindon,[†] and Jeremy K. Nicholson^{*,†}

[†]Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, United Kingdom

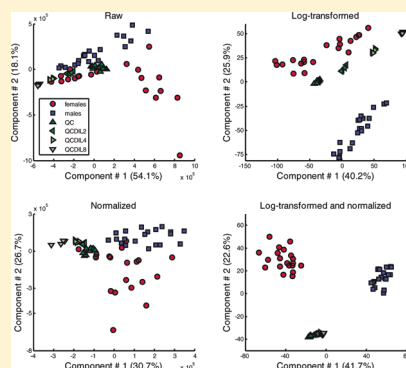
[‡]Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland 97239, United States

[§]Technologie Servier, 27 Rue Eugène Vignat, Orleans 45000, France

^{||}Department of Chemistry, Princeton University, Princeton, New Jersey 08544-1014, United States

S Supporting Information

ABSTRACT: Ultra-performance liquid chromatography coupled to mass spectrometry (UPLC/MS) has been used increasingly for measuring changes of low molecular weight metabolites in biofluids/tissues in response to biological challenges such as drug toxicity and disease processes. Typically samples show high variability in concentration, and the derived metabolic profiles have a heteroscedastic noise structure characterized by increasing variance as a function of increased signal intensity. These sources of experimental and instrumental noise substantially complicate information recovery when statistical tools are used. We apply and compare several preprocessing procedures and introduce a statistical error model to account for these bioanalytical complexities. In particular, the use of total intensity, median fold change, locally weighted scatter plot smoothing, and quantile normalizations to reduce extraneous variance induced by sample dilution were compared. We demonstrate that the UPLC/MS peak intensities of urine samples should respond linearly to variable sample dilution across the intensity range. While all four studied normalization methods performed reasonably well in reducing dilution-induced variation of urine samples in the absence of biological variation, the median fold change normalization is least compromised by the biologically relevant changes in mixture components and is thus preferable. Additionally, the application of a subsequent log-based transformation was successful in stabilizing the variance with respect to peak intensity, confirming the predominant influence of multiplicative noise in peak intensities from UPLC/MS-derived metabolic profile data sets. We demonstrate that variance-stabilizing transformation and normalization are critical preprocessing steps that can benefit greatly metabolic information recovery from such data sets when widely applied chemometric methods are used.



Detection of a large number of low molecular weight metabolites in complex mixtures has become an important tool to study human disease and the effects of drug therapy, to identify biochemical functions of unknown genes and assess the impact of mutations, and to investigate the interactions of genes and environment to produce observable phenotypes at the cell and organism level.^{1–3} This approach, often referred to as metabonomics,⁴ metabolomics,¹ or metabolic profiling, aims to improve understanding of physiology and metabolism by using analytical chemistry techniques to assess metabolic changes in biofluids, tissues, and cell extracts following an experimental perturbation. Unlike microarray based “omics” platforms, metabonomics relies on established analytical chemistry techniques such as liquid chromatography/mass spectrometry (LC/MS), gas chromatography/mass spectrometry (GC/MS) and nuclear magnetic resonance spectroscopy (NMR) to obtain metabolic

signatures.^{1–3} The recent introduction of ultraperformance LC (UPLC/MS) has greatly enhanced chromatographic performance, increasing the sensitivity and throughput of LC/MS measurements. Although the determination of the exact number of metabolites which can be measured by untargeted UPLC/MS is complicated by sample preparation, chemical diversity of the matrix, and the presence of isotopes and fragment and adduct ions, UPLC/MS routinely detects thousands of features representative of hundreds to thousands of metabolites in biological mixtures.⁵

While these analytical chemistry techniques boast high sensitivity and reproducibility and are capable of untargeted detection

Received: February 23, 2011

Accepted: April 28, 2011

Published: April 28, 2011

of a large number of diverse chemical species, the collection and comparison of a large number of NMR or mass spectra pose formidable data analysis challenges regardless of whether NMR, GC/MS, or LC/MS is applied. While the capabilities of these techniques have been compared elsewhere, spectra generated by each require similar preprocessing steps before comparative analysis.⁶ The signatures of thousands of metabolites generated by UPLC/MS require extensive preprocessing prior to comparative statistical analysis.^{7–10} A key step in the analysis of UPLC/MS data sets is the transformation of ion intensities resolved by elution time into a matrix of features (m/z –retention time pairs; columns) present in each sample (rows) by peak detection, alignment, and area extraction algorithms. A number of open source tools such as XCMS,¹⁰ MAVEN,¹¹ MZmine,^{12,13} OpenMS,^{14,15} and XAlign,¹⁶ as well as proprietary software packages such as MarkerLynx (Waters Corp), Mass Profiler (Agilent), and others from Thermo Scientific and Applied Biosystems, are currently available to achieve these objectives. Subsequent statistical analysis tools operate on this matrix of ion intensities.

Despite the significant attention paid to peak identification methods, instrumental and experimental challenges complicating the direct interpretation of ion intensities as metabolite concentrations have been largely neglected in the experimental literature.^{17,18} This is especially true with respect to the urinary matrix commonly used in metabolomics studies. While urine is an ideal biofluid for metabolomics studies since it is obtained noninvasively, and its composition is affected by genetic and environmental factors reflecting the physiology of multiple organs, urine dilution effects and instrumental variation from the analytical method play a significant confounding role when one attempts to characterize these biological and physiological factors through NMR and MS measurements of small molecule concentrations.¹⁹ Ensuring that NMR- and MS-generated urinary profiles are directly comparable by reducing the impact of variance associated with dilution and noise by data normalization and transformation are critical steps in MS data preprocessing prior to multivariate statistical analysis.^{18,20}

Normalization is used to identify and remove sources of systematic variation between sample profiles due to factors that are irrelevant with regard to biological processes, such as sample dilution or variation in instrument detector sensitivity, in order to ensure that spectra are comparable across runs and across related sample sets.²⁰ Up to 15-fold changes in urine volume are commonly observed under normal conditions, resulting in significant variation in dilution of metabolite concentrations across many samples.²¹ Additionally, disease or toxic effects may cause further changes in overall urine volume, further complicating interpretation of changes in observed metabolite concentrations. Technical variation in UPLC/MS arises from sources such as differences in LC separation efficiency (e.g., due to column aging), and drifts in ionization and detector efficiencies (e.g., due to source contamination). In order to obtain repeatable and interpretable LC/MS metabolic profiling data, the chromatographic retention time, ion signal intensity, and mass accuracy must be stable both within a sample batch and between multiple batches. It is well-known that the first few injections of sample on an LC/MS system give unrepresentative results for global metabolic analysis, due to small changes in chromatographic retention time and/or signal intensity. Usually after 5–10 injections of the matrix, retention times stabilize as the column becomes “conditioned” and the system then shows little variability through the remainder of the run. Additionally, the source

of the mass spectrometer can become contaminated, leading to gradual changes in instrument sensitivity over time. Providing that these changes are not major, the subsequent data treatment will not be too adversely affected, provided that careful randomization of the samples has been performed to ensure that all of the experimental groups are affected to the same extent. Quality control samples, ideally a pooled sample consisting of aliquots from all samples in a study batch, can be used to condition the column at the start of the run and are then injected periodically, for example, every 10 samples throughout the run to assess instrument stability. These QC samples can also be used to assess stability across multiple instruments.

In order to ensure that samples are comparable within a batch and between multiple sets of samples, a variety of methods can be employed to estimate normalization factors.^{17,20,22–24} Each technique relies on a set of assumptions about the analytical measurements and the biological variability of a data set in question. The normalization approaches applied to LC/MS metabolic data sets have been thus far largely limited to basic techniques that scale all spectra by a common factor, usually the total metabolite peak intensity or physiological variables such as urine volume, osmolality, or creatinine level.²⁴ However, both the total metabolite concentration and physiological variables of urine can vary considerably with reference to a variety of factors such as age, gender, diet, and time of urine collection. The instrumental effects of ion suppression and detector saturation can also cause highly abundant metabolite peaks to respond differentially to dilution, and thus the same scaling factor may not necessarily be required across the peak intensity range.²⁰ More recently, a normalization method was developed to correct for UPLC/MS peak intensity attenuation of serum profiles with respect to the order of injection based on QC data.²⁵ It is a robust method to account for technical variance due to factors such as loss of instrumental sensitivity, but it is not directly applicable to urine studies where additional and often more substantial variation of total urinary volume (and hence metabolite concentrations) is expected.

In addition, the UPLC/MS peak intensities of metabolic profiles are subject to noise from various sources and hence of different types resulting in a heteroscedastic noise structure, characterized by increasing technical variance as a function of increased signal intensity.^{17,18} This can adversely affect standard statistical and pattern recognition tools,^{26,27} which assume the data noise to be consistent across the whole intensity range (i.e., homoscedastic). This issue can be addressed either by modifying a statistical model to accommodate heteroscedastic noise structure^{26,28} or by validating an appropriate variance-stabilization transformation to make peak intensity consistent across the whole intensity range.²⁷ Where possible, the last is often a preferred option. Here, we introduce a statistical model for UPLC/MS peak intensities of metabolite levels that comprises data normalization and quantification of measurement noise. The objective is to remove sources of systematic variation between sample profiles due to factors that are irrelevant with regard to biological processes such as sample dilution (normalization) and to stabilize the technical variance as a function of signal intensity (variance-stabilizing transformation). Four different normalization techniques [locally weighted scatter plot smoothing (LOESS), quantile, median fold change, and total intensity] are compared for adjusting peak intensities between samples to a common scale. The choice is made to cover a range of assumptions regarding UPLC/MS metabolic measurements and biological variability of a data set. The utility of logarithmic-based variance-stabilizing transformations is evaluated.

The proposed approach could be adapted for use in other types of biochemical mixtures and is thus generally applicable to a wide range of analytical chemistry applications.

MATERIALS AND METHODS

Study Details. Urine samples analyzed as part of a study on gender differences in rat metabolic profiles were used to assess the different preprocessing procedures. After 1 week of acclimatization, 40 Wistar rats (20 males and 20 females), aged 9 weeks, were housed individually in metabolism cages for 24 h under a conventional environment with controlled conditions [temperature 20–24 °C, relative humidity 40–70%, 12-h light cycle (7 a.m. to 7 p.m.)]. Filtered tap water was available ad libitum. Food [sterilized A04C-10 feed pellets from SAFE (Villemoisson-sur-Orge, France)] was available ad libitum for the first 8 h, and then the animals were fasted for the subsequent 16 h. Urine was collected over the intervals 0–8 h and 8–24 h into refrigerated tubes protected from light containing sodium azide (500 μ L of 1% solution). After collection, all samples were centrifuged and 225 μ L aliquots were stored at –40 °C until UPLC/MS analysis.

Sample Preparation. Water (150 μ L) was added to 50 μ L of urine. Samples were vortexed and centrifuged (16000g, 10 min), and 140 μ L of supernatant was transferred into MS 96-well plates. A quality control (QC) sample, representative of the sample batch, was prepared by combining 40 μ L of each sample. Dilutions 1/2, 1/4, and 1/8 of this QC sample in water were also prepared.

UPLC/MS Analysis. Analyses were performed on an Acquity UPLC system (Waters, Elstree, U.K.) coupled to a LCT Premier time-of-flight mass spectrometer (Waters, Manchester, U.K.), operated in the positive electrospray ionization (ESI+) mode. Chromatography was carried out at 40 °C with a 0.5 mL/min flow rate on a Waters Acquity UPLC HSS T3 column (1.8 μ m, 2.1 $\times$ 100 mm), with the following solvent system: A = 0.1% formic acid in water, B = 0.1% formic acid in acetonitrile. A 10 min gradient was used, followed by a 2 min re-equilibration phase (0–1 min, hold at 0% B; 1–3 min, 0–15% B (linear); 3–6 min, 15–50% B (linear); 6–9 min, 50–95% B (linear); 9–10 min, hold at 95% B; 10.1–12 min, hold at 0% B). Gradient type = 6 (Waters). Samples were kept at 4 °C in the autosampler during the analysis. The injection volume was 5 μ L. ESI conditions were as follows: source temperature 120 °C, desolvation temperature 350 °C, cone gas flow 25 L/h, desolvation gas flow 900 L/h, capillary voltage 3200 V, cone voltage 35 V. The instrument was operated in V optics mode and set to acquire data over the m/z range 50–1000 with scan time of 0.2 s and an interscan delay of 0.01 s. Data were collected in centroid mode. Leucine enkephalin (MW = 555.62) (200 pg/ μ L in acetonitrile/water 50:50) was used as a lock mass for accurate mass measurements. The instrument was calibrated before analysis by use of 0.5 mM sodium formate solution. The run order was randomized to minimize its correlation with the factors of the experimental design (gender and collection time point). The QC sample was injected regularly throughout the run (after every 10 samples) to monitor the stability of the analytical platform.^{5,9,29} This QC sample was also used to condition the column at the beginning of the UPLC/MS run,^{5,9} because the first injections of a batch are often not reproducible.

Raw Data Preprocessing. UPLC/MS raw data files were converted to netCDF format by use of the DataBridge tool implemented in MassLynx4.1 software (Waters). They were then preprocessed by use of the freely available XCMS software.¹⁰

The CentWave algorithm³⁰ was used for peak picking with a peak width window of 3–15 s; the m/z width for the grouping step was changed to 0.1 Da; and the bandwidth parameter was kept to default (30 s) for the first grouping and then determined from the retention time deviation profile after retention time correction. This preprocessing step produced an output table of time-aligned detected features with their retention time, m/z ratio, and intensity in each sample. Isotope peaks, fragments, and adducts were treated as separate metabolite features. For the work carried out in this paper, the urine samples of the collection interval 8–24 h only were used, resulting in 40 urine samples (20 males and 20 females) plus the QC samples.

Statistical Model. An UPLC/MS data set following the above routine preprocessing stages can be represented by a rectangular table (**X**) of real number elements (x_{ik}). The rows and columns of **X** represent, respectively, samples and metabolite peak intensities. Due to a variety of experimental, instrumental, environmental, and signal processing influences, the resulting peak intensities are a combination of real signal and noises of varying characteristics, which are generally of additive or multiplicative nature^{18,31} and thus can be modeled by

$$x_{ik} = \beta_{ik} + n_{ik}s_{ik}e^{\eta_{ik}} \quad (1)$$

where x_{ik} is the measured peak intensity (level) of the i th metabolite in the k th sample, s_{ik} is the expected peak intensity (level), β_{ik} is the random background (electronic) noise, η_{ik} is the multiplicative random noise (for example, due to sample preparation variation or fluctuations in the ion source or the sample-introduction equipment), and n_{ik} is the normalization scaling factor of signal intensities of the k th sample as a result of variable sample dilution assumed to be either fixed (n_k) or peak-intensity-dependent (n_{ik}). Additive noise is characterized by random fluctuations in the baseline, irrespective of the presence of metabolite signals, and is derived from electronic noise in the equipment. Conversely, multiplicative noise grows with the signal intensity of a metabolite and is often proportional to it. As a result of the influence of multiplicative noise, metabolites with higher peak intensities would exhibit larger variability when repeatedly measured, and thus weak signals can be buried in the noise of strong signals. The objective of the preprocessing strategies considered here is to remove the systematic variation due to differences in overall sample concentration (normalization) and to convert multiplicative noise into additive noise (variance-stabilizing transformation):

$$\text{vst}\left(\frac{x_{ik}}{n_{ik}}\right) \approx \mu_{ik} + \varepsilon_{ik} \quad (2)$$

where vst denotes a variance-stabilizing transformation, μ_{ik} is the transformed peak intensity, and ε_{ik} is the random noise. The overall procedure is quite straightforward. The fixed or peak-intensity-dependent normalization factor is initially estimated for each sample profile. The peak intensity is then divided by the sample-dependent normalization factor. The generalized logarithm or logarithm variance-stabilizing transformation is finally applied to convert multiplicative noise into additive noise.

Normalization: Accounting for Systematic Bias between Samples. Four different normalization techniques were selected for adjusting peak intensities between samples to a common scale. The choice was made to cover a range of assumptions regarding UPLC/MS metabolic measurements and biological variability of a data set. The first two approaches considered here

are the total intensity and median fold change normalizations. Both methods assume that measured peak intensities are directly proportional to concentrations of metabolites in solution. Under this assumption, the change in intensity of a profile due to variable sample dilution and/or inconsistency of UPLC/MS acquisition parameters is expected to be uniform across all peaks and thus a fixed scaling factor is employed.

Total intensity normalization forces all samples in a set of experiments to have equal total intensity:

$$n_k^{\text{TI}} = \sum_i x_{ik} \quad (3)$$

where n_k^{TI} is the total intensity of the k th sample.²⁴ This technique assumes that the total concentration for all metabolites in a sample does not vary across samples in a data set. This assumption is not generally fulfilled in urinary metabolic studies, due to changes in the number and concentration of small molecules excreted.³² Changes in the peak intensities of a few high-concentration metabolites can compromise noticeably the normalization performance because of their substantial contribution to the total peak intensity.

Median fold change normalization adjusts the median of log fold changes of peak intensities between samples in a set of experiments to be approximately zero:

$$n_k^{\text{MFC}} = \text{median} \left(\frac{x_{ik}}{x_{ir}} \right) \quad (4)$$

where n_k^{MFC} is the median fold change of peak intensities between the k th sample and the target sample. This normalization is based on the reasonable premise that metabolite peaks affected purely by dilution will exhibit the same fold changes between two sample profiles. This should be the case if peak intensities are directly proportional to metabolite levels in solution. The median value of fold change between sample and target profiles has been shown to be a robust estimate of this dilution factor.³² The performance of this method is robust against at least 50% peak intensities exhibiting asymmetrical increase or decrease in response to biological factors such as disease or toxic challenge. The choice of a target profile has been shown not to be crucial for the method performance; it can be any sample of a given data set or, as typically selected, the median profile.

The other two approaches considered here are quantile and LOESS normalizations. These methods assume that systematic sample-to-sample variation is peak-intensity-dependent, that is, not necessarily the same scaling factor is used over the peak intensity range. This variation may arise due to the effects of ion suppression and detector saturation on measured metabolite levels.

Quantile normalization enforces all samples in a set of experiments to have identical peak intensity distribution:²²

$$n_{ik}^{\text{QN}} = \frac{x_{ik}}{x_{r(i)t}} \quad (5)$$

where n_{ik}^{QN} is the quantile-intensity-dependent normalization factor of the k th sample and the subscript (i) enclosed in parentheses indicates the order statistic of the i th peak intensity of the k th sample and x_t is a target sample. Each peak intensity of a sample by the quantile normalization is simply substituted by the peak intensity with the same order statistic $r(i)$ of a target sample. For example, the highest intensity in a given sample takes the value of the highest peak intensity in the target sample, the second highest peak intensity takes the second highest in the

target sample, and so on. The target profile is typically selected to be the mean or median profile of a data set. The method assumes that the distribution of peak intensity across a data set is nearly the same for all samples and can be problematic with high-intensity values since they typically vary noticeably from sample to sample. It is also assumed that there is a similar number of metabolites with increased and decreased signals across the peak intensity range.

Locally weighted scatter plot smoothing (LOESS) normalization adjusts the local median of log fold changes of peak intensities between samples in a data set to be approximately zero across the whole peak intensity range:^{23,33}

$$n_{ik}^{\text{LOESS}} = f(x_{ik}) \quad (6)$$

where n_{ik}^{LOESS} is the LOESS-intensity-dependent normalization factor of the k th sample calculated by nonlinear mapping of sample intensities into those of a reference sample. It is assumed that the proportion of metabolic changes across biological samples is relatively small or that there is a similar number of metabolites with increased and decreased signals across the peak intensity range.

Variance-Stabilizing Transformation: Accounting for Heteroscedastic Noise Structure. To account for heteroscedastic noise structure observed in UPLC/MS metabolite measurements, two logarithm-based transformations are chosen on the basis of the model (eq 1). It is first assumed that the noise is mainly multiplicative in nature, that is, the influence of background random noise is negligible (i.e., $\beta_{ki} \approx 0$). This assumption has been previously used in modeling UPLC/MS data¹⁷ and other types of high-throughput data generated by, for example, microarray technology.^{26,34} In this noise model, the standard deviation of peak intensity is proportional to its expected value. The logarithm is an appropriate variance-stabilizing transformation, as it makes peak intensity variance independent of its expected value by transforming multiplicative error into additive error:

$$\log \left(\frac{x_{ik}}{n_{ik}} \right) \approx \log(s_{ik}) + \eta_{ik} \quad (7)$$

It is second assumed that the error structure of UPLC/MS measurements has both additive and multiplicative noises (eq 1). The additive random component may arise from the integration of count-based signal inherent in the majority of MS instrumentation and/or the presence of a small basal unspecific background signal component. The variance of peak intensity under this model is a quadratic function of its expected value.²⁶ The same model has been previously considered in modeling LC/MS data¹⁸ and is commonly used in gene expression studies.^{26,27,31} The inverse generalized logarithm transformation has been derived as an appropriate variance-stabilizing transformation in this case:²⁷

$$\text{glog} \left(\frac{x_{ik}}{n_{ik}} \right) \approx \mu_{ik} + \varepsilon_{ik} \quad (8)$$

$$\text{glog}(x) = \log \left(\frac{x + \sqrt{x^2 + c}}{2} \right) \quad (9)$$

where c is a fitted parameter. This transformation (eq 9) coincides with the logarithm for large intensities but is approximately linear for low intensities and interpolates smoothly in between. The scripts for

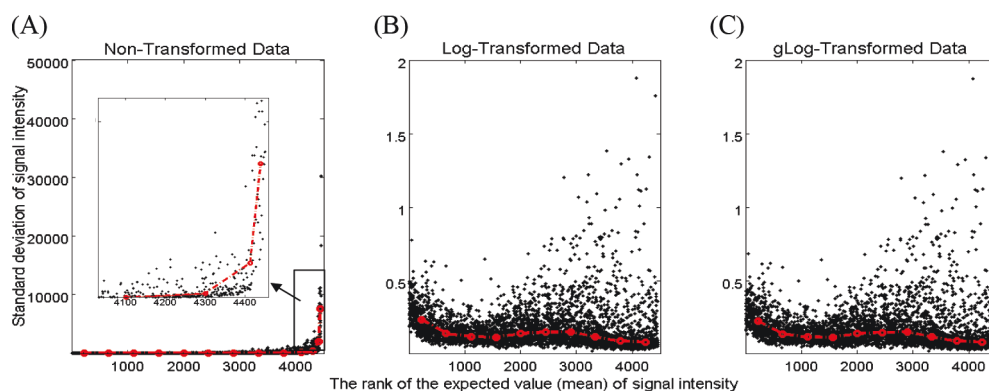


Figure 1. Standard deviation as a function of the rank of mean signal intensity for QC samples: (A) nontransformed data, (B) logarithm- (log-) transformed data, and (C) generalized logarithm- (glog-) transformed data. The running median of the standard deviation is shown as a dotted red line in the plot. In the absence of heteroscedastic noise structure, it should be approximately a horizontal line with minor fluctuations but no overall trend.

performing variance-stabilizing normalizations described above are available, upon request, in the open source statistical language R.

Principal Component Analysis. Formally, principal component analysis (PCA) summarizes the data matrix in terms of scores and loadings:³⁵

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \quad (10)$$

where $\mathbf{X}$ is a data matrix of sample profiles, $\mathbf{T}$ is a matrix of scores, $\mathbf{P}$ is a matrix of weights (loadings) of spectral variables into the scores, $\mathbf{E}$ is a residual, and the prime (') denotes a transpose operation. PC scores summarize interrelationships between observations, for example, groupings according to physiological traits or gender and dose- or time-dependent changes. A subset of metabolites correlated with PC scores can be identified by use of the corresponding PC loadings. PCA was performed by using the built-in function of Matlab for eigenvalue decomposition (<http://www.mathworks.com/>).

RESULTS AND DISCUSSION

Performance of Logarithmic-Based Variance Stabilizing Transformations. To evaluate the changes in variation of peak intensities across the entire measurement intensity range, replicate injections of a quality control (QC) sample, obtained by pooling equal aliquots of each urine sample in the study, were used, as this QC sample represents all metabolites present in the experimental set and can serve as a measure of technical variability. The QC sample was measured eight times in the experimental run. The standard deviation as a function of the rank of the mean of peak intensity for replicate injections of the QC sample is shown in Figure 1A. In the absence of heteroscedastic noise structure, the running median of the standard deviation should approximate a horizontal line with minor fluctuations but no overall trend. This is not the case for the UPLC/MS data, where the variation of peak intensity grows with the rank of mean intensity, that is, as intensity increases. As a consequence for replicate measurements, larger peak intensities are progressively more variable. This means that assumptions underlying many widely applied statistical methods are not valid, and this in turn could lead to incorrect statistical inferences, having consequences for metabolic information recovery. To correct for heteroscedastic noise structure, the log and glog transformations were applied. Both transformations performed comparatively well in stabilizing the variance of transformed peak

intensities across the entire measurement intensity range (Figure 1B,C). This indicates the negligible influence of background additive noise and the predominant influence of multiplicative noise in the UPLC/MS peak intensity data. Such a dominating influence of multiplicative noise has been previously observed in LC/MS measurements of protein levels.¹⁸

Performance of Normalization Methods in the Absence of Biological Variation. To assess the ability of normalization to account for dilution effects, the QC sample was systematically diluted, ranging from 2- to 8-fold. An 8-fold dilution of urine samples is quite common in metabolic studies. The objective here is to evaluate whether sample dilution is likely to induce uniform or peak-intensity-dependent changes in metabolic profiles. The uniform change due to sample dilution occurs when measured peak intensities are directly proportional to metabolite levels in solution (e.g., 2-fold sample dilution would result in 2-fold decrease in all peak intensities). However, various instrumental influences can potentially introduce peak-intensity-dependent biases between samples. As a result of ion suppression or detector saturation, for example, the peak intensities of highly abundant peaks may respond differently to sample dilution or not be affected by it at all. One way to test this is to visualize the log fold changes of peak intensities between diluted and undiluted samples as a function of log peak intensity; in microarray analysis this referred to as the M versus A plot. Typical behavior of peak intensities in response to 2- and 4-fold dilutions is shown in Figure 2 panels A and B, where red curves (calculated by the nonlinear LOESS regression) capture the adjustment required for a particular intensity range between diluted and undiluted samples. Clearly, the dilution-induced adjustment of peak intensities is comparable across the entire peak intensity range (e.g., the scaling factor of around 2 is necessary to bring all peak intensities of the 2-fold diluted sample to a common scale with the undiluted sample). The majority of peak intensities are thus likely to respond linearly to sample dilution. This is further confirmed by the fact that the reduction of the percentage of dilution-induced variation for individual peak intensities (calculated by one minus ratio of variance of transformed peak intensity after and prior to application of normalization) is comparable (on average 70–95%) over the peak intensity range between four studied normalization methods (Figure 2C). It should be noted that high sample dilution might limit the detection of low-abundance metabolites in biological samples by UPLC/MS.

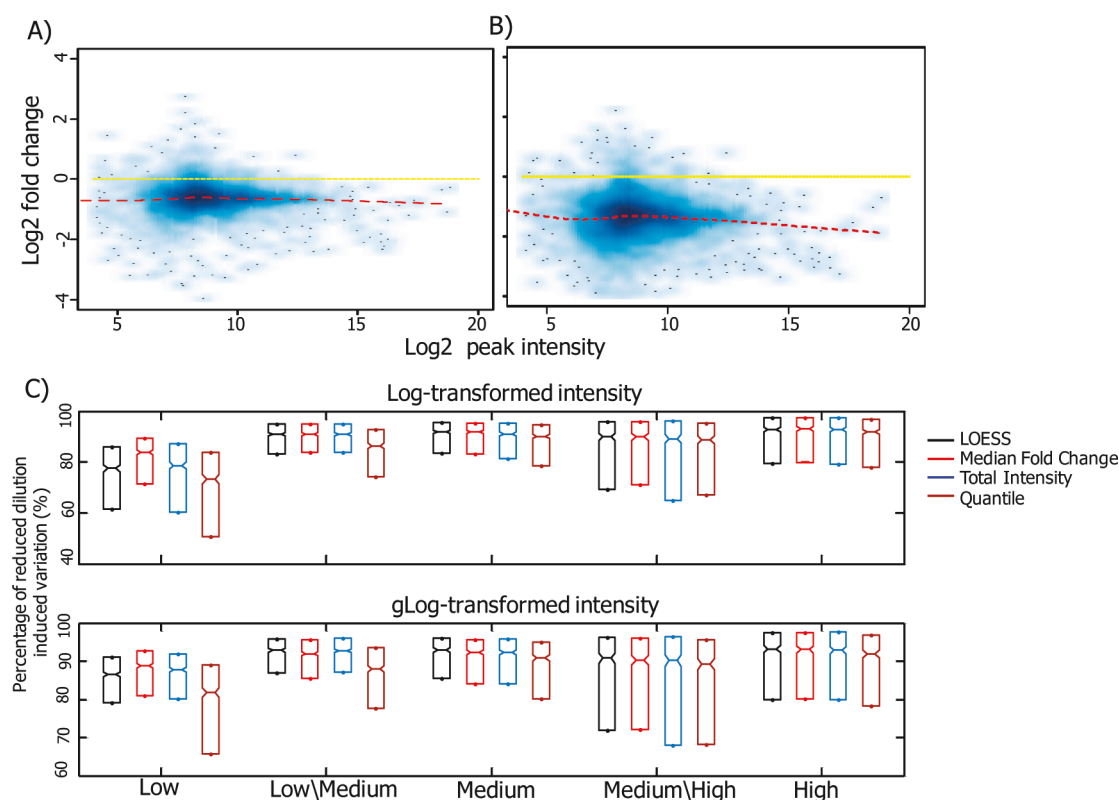


Figure 2. Assessment of peak-intensity-dependent bias due to variable sample dilution in UPLC/MS metabolic measurements. The log fold change in peak intensities between diluted and undiluted samples as a function of log peak intensity is shown for (A) 2-fold and (B) 4-fold sample dilutions. Red curves (calculated by nonlinear LOESS regression) capture the adjustment required for a particular intensity range between diluted and undiluted samples. (C) Reduction of the percentage of dilution-induced variation for individual log- or glog-transformed peak intensities across peak intensity range. Reduction of dilution-induced variation is calculated by $1 - \text{var}_{\text{norm}}/\text{var}_{\text{non-norm}}$, where var_{norm} and $\text{var}_{\text{non-norm}}$ are variances of transformed peak intensities after and prior to application of normalization, respectively. The sorted mean intensity values for 4453 peaks were divided into five approximately equal-sized data subsets, denoted by low, low/medium, medium, medium/high, and high peak intensity ranges. The upper edge of the box is the 75th percentile and the lower edge is the 25th percentile of values.

However, this issue relates only to the instrumental data acquisition and not to the subsequent data normalization.

Comparison of Normalization Methods in the Presence of Biological Variation. A robust normalization method not only must account for dilution effects in samples of the same composition but also must be applicable to experimental data where changes in mixture composition are to be expected. Certain assumptions with regard to the nature of biological variance are inherent to all normalization methods (see the Materials and Methods section for a detailed description). As all four studied methods performed comparatively well in the absence of biological variation, a method whose performance is robust to biologically induced changes in peak intensity would be of comparative advantage. The total intensity, quantile, and LOESS normalizations rely heavily on the assumption that there are similar numbers of metabolites with increased and decreased signals.^{22,23} This assumption may not generally be fulfilled in metabolic studies since the increases in the level of one metabolite in response to stimuli may not necessarily be accompanied by the decreased level of another metabolite. The application of the LOESS and quantile methods is thus limited to cases in which a low percentage of metabolites have asymmetrically increasing or decreasing abundances. The performance of total intensity normalization can be compromised by a single large metabolite peak that varies substantially from sample to sample.³² By

contrast, the median fold change normalization allows at least half of metabolite peak intensities to exhibit asymmetrical changes between samples. This is a relaxed assumption with regard to the nature and proportion of differentially abundant metabolite peaks, which is generally applicable to data from metabolic studies.

The performance of normalization methods in the presence of biological variance can be diagnosed by use of summary statistics of sample peak intensity distributions via box plots. If it is assumed that up to a quarter of peak intensities are asymmetrically increased or decreased due to biological effects, the log ratios of nondifferentially abundant metabolite peaks should have a comparably small spread around zero across samples (noted by the box in Figure 3) and a median of zero (a black diamond in the box in Figure 3). This is illustrated on a urine data set of control Wistar rats, in which the majority of peak intensities are expected not to be differentially abundant. The peak intensities of nonnormalized samples exhibit large variation in overall concentration within sample types of male (red) and female (blue) rats (Figure 3A). This variation is within the 8-fold dilution range of the QC sample (green). The effect of this dilution-induced variation is that any informational changes in metabolite concentration between sample types would need to be very large to be detected. As a result, a large set of biomarker candidates would simply not be identified by statistical

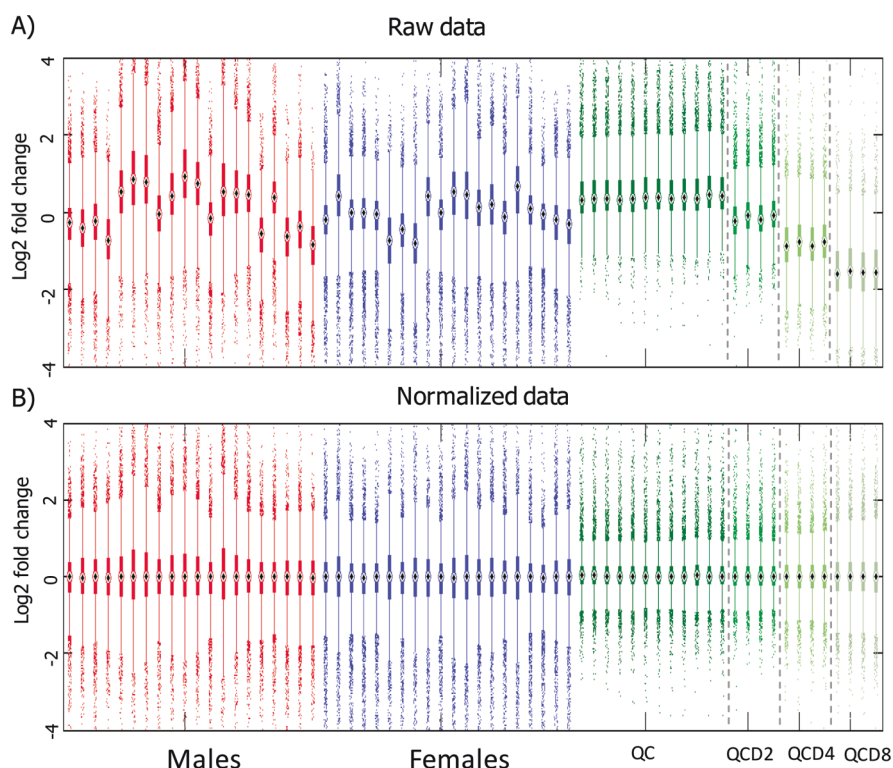


Figure 3. Box plots of the log 2-fold change in peak intensities relative to the median profile calculated across a data set: (A) nonnormalized and (B) median fold change normalized data. The upper edge of the box is the 75th percentile and the lower edge is the 25th percentile of relative intensity values and points noted by ‘•’ are metabolite features which exhibit relative changes. The median is shown as a black diamond in the box. The boxes are colored by sample types. QC, quality control urine samples of Wistar rats; QCD2, 2-fold diluted quality control urine samples; QCD4, 4-fold diluted quality control urine samples; QCD8, 8-fold diluted quality control urine samples.

techniques. The median fold change, quantile, and LOESS normalizations successfully reduce dilution-induced variance, as shown by the near zero log ratios of nondifferentially abundant peaks, indicating the negligible influence of peak intensity biases between samples (Figure 3B and Figure S1, Supporting Information). Total intensity normalization, however, shows inferior performance as indicated by the remaining sample-to-sample variation. This is most likely due to small changes in total metabolite excretion between samples. We suggest that while quantile and LOESS show similar performance to median fold change, the median fold change normalization should be preferred because of its relaxed assumption with the proportion of asymmetric metabolic marker changes.

Impact of Variance-Stabilizing Normalization on Exploratory Data Analysis via PCA. Principal component analysis (PCA) is the most commonly used method to explore relationships between samples in metabolic profiling studies. It aims to extract a small number of latent components that summarize the measured data with minimal information loss by taking advantage of the correlation structure of peak intensities. The information content in PCA is measured as a sum of variances of individual peak intensities, referred to as the total variance. In the presence of multiplicative noise, metabolites with higher peak intensities are progressively more variable when measured repeatedly and therefore constitute a larger share of the total variation. The resulting PC scores are impacted heavily by the random variation of these high-intensity metabolites and therefore can poorly represent the complete variation structure of a data set. This effect is illustrated on the urine data set of control

Wistar rats in Figure 4. The sample scores of the first two PCs (raw peak intensities, Figure 4) are scattered, indicating a dominant influence of random variation. As a consequence, the assessment of overall similarities and differences between samples is quite poor; the grouping of technical replicates with respect to the degree of dilution is not even clear. The normalization of raw peak intensities by median fold change results in only slight changes in the PC scores because of the dominant impact of multiplicative noise in high-concentration mixture components (normalized peak intensities, Figure 4). After application of log-based variance-stabilizing transformations, the first two components clearly capture the systematic variation due to the difference in sample composition and variable sample dilution (log/glog-transformed peak intensities, Figure 4). The dispersion in the scores of both male and female Wistar rats along PC1 and PC2 parallel to the QC dilution series indicates a range in concentration similar to the 8-fold dilution changes of the QC sample. Normalization successfully removes this dilution-induced variation, leading to tight clustering of technical replicates and improved clustering of biological replicates. In addition, all clusters of sample types are clearly distinguishable (log/glog-transformed and normalized peak intensities, Figure 4). The sample scores of the first two PCs between log- and glog-based transformations across four normalization methods are somewhat comparable (Figure S2, Supporting Information). This confirms the results of the previous sections: that multiplicative noise is the most significant source of background variation in MS metabolic profiling and that all four normalization methods remove the majority of dilution-induced variation from the data.

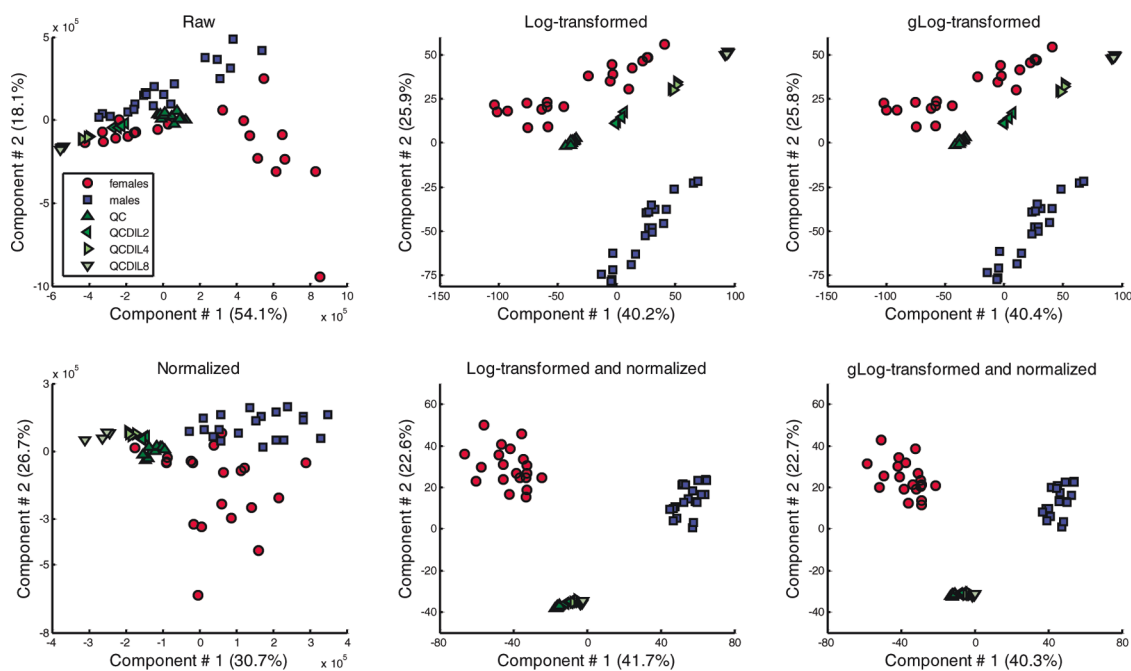


Figure 4. Impact of variance-stabilizing normalization on metabolic information recovery via PCA. QC, quality control urine samples of Wistar rats; QCDIL2, 2-fold diluted quality control urine samples; QCDIL4, 4-fold diluted quality control urine samples; QCDIL8, 8-fold diluted quality control urine samples. The data were normalized by median fold change normalization.

CONCLUSIONS

We have compared several variance-stabilizing normalization strategies based on a statistical error model in order to account for systematic sources of variation, such as that induced by variable sample dilution, and in order to stabilize the technical variance as a function of peak intensity observed in UPLC/MS metabolic profiles of urine.

It can be concluded that the UPLC/MS peak intensities of urine metabolic profiles are likely to respond linearly to variable sample dilution across the intensity range. In other words, peak-intensity-dependent biases between samples as a result of ion suppression or detector saturation that would cause highly abundant peaks to respond differentially to dilution were not observed. This indicates that the majority of urinary metabolites are within the detector dynamic range.

Second, all four studied normalization methods (LOESS, quantile, total intensity, and median fold change) performed comparatively well in reducing dilution-induced variation of urine samples in the absence of biological variation. However, the median fold change normalization is least compromised by biologically relevant changes in mixture components and is thus preferable.

Third, the higher peak intensities of urine profiles exhibit larger variability when repeatedly measured, mainly due to the presence of multiplicative noise. This violation of constant variance across the measurement range imposes a serious challenge when standard statistical techniques are applied, as demonstrated by principal component analysis. The UPLC/MS peak intensity of urine samples can be brought in line with this assumption by applying a log-based transformation, which successfully stabilizes the technical variance across the intensity range.

Finally, variance-stabilizing transformation and normalization are critical preprocessing steps that can greatly benefit the metabolic information recovery from urine UPLC/MS data sets via commonly used pattern recognition tools.

Although applied to urine studies, the proposed framework for assessing the performance of normalization methods and variance-stabilizing transformation is general in scope and thus can be applied to other biological sample types.

ASSOCIATED CONTENT

S Supporting Information. Two figures, showing box plots of log 2-fold change in peak intensities after various normalization techniques (LOESS, MFC, TI, and QN) and the impact of variance-stabilizing normalizations (LOESS, TI, and QN) on metabolic information recovery via PCA. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: j.nicholson@imperial.ac.uk.

ACKNOWLEDGMENT

K.A.V. and L.K.V. contributed equally to this work. K.A.V. gratefully acknowledges Technologie Servier for financial support.

REFERENCES

- (1) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155–171.
- (2) Lindon, J. C.; Nicholson, J. K.; Holmes, E.; Everett, J. R. *Concepts Magn. Reson.* **2000**, *12*, 289–320.
- (3) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *29*, 1181–1189.
- (4) Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. *Nat. Rev. Drug Discovery* **2002**, *1*, 153–161.

- (5) Want, E. J.; Wilson, I. D.; Gika, H.; Theodoridis, G.; Plumb, R. S.; Shockcor, J.; Holmes, E.; Nicholson, J. K. *Nat. Protocols* **2010**, *5*, 1005–1018.
- (6) Pan, Z.; Raftery, D. *Anal. Bioanal. Chem.* **2007**, *387*, 525–527.
- (7) Benton, H. P.; Wong, D. M.; Trauger, S. A.; Siuzdak, G. *Anal. Chem.* **2008**, *80*, 6382–6389.
- (8) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27*, 747–751.
- (9) Sangster, T.; Major, H.; Plumb, R.; Wilson, A. J.; Wilson, I. D. *Analyst* **2006**, *131*, 1075–1078.
- (10) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779–787.
- (11) Melamud, E.; Vastag, L.; Rabinowitz, J. D. *Anal. Chem.* **2010**, *82*, 9818–9826.
- (12) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinf.* **2010**, *11*, 395.
- (13) Katajamaa, M.; Miettinen, J.; Oresic, M. *Bioinformatics* **2006**, *22*, 634–636.
- (14) Sturm, M.; Bertsch, A.; Gropl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. *BMC Bioinf.* **2008**, *9*, 163.
- (15) Bertsch, A.; Gropl, C.; Reinert, K.; Kohlbacher, O. *Methods Mol. Biol.* **2010**, *696*, 353–367.
- (16) Zhang, X.; Asara, J. M.; Adamec, J.; Ouzzani, M.; Elmagarmid, A. K. *Bioinformatics* **2005**, *21*, 4054–4059.
- (17) Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Oresic, M. *BMC Bioinf.* **2007**, *8*, 93.
- (18) Anderle, M.; Roy, S.; Lin, H.; Becker, C.; Joho, K. *Bioinformatics* **2004**, *20*, 3575–3582.
- (19) Bollard, M. E.; Stanley, E. G.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *NMR Biomed.* **2005**, *18*, 143–162.
- (20) Kulthma, K.; Nilsson, A.; Scholz, B.; Rossbach, U. L.; Falth, M.; Andren, P. E. *Mol. Cell. Proteomics* **2009**, *8*, 2285–2295.
- (21) Tsuchiya, Y.; Takahashi, Y.; Jindo, T.; Furuhashi, K.; Suzuki, K. T. *Eur. J. Pharmacol.* **2003**, *475*, 119–128.
- (22) Bolstad, B. M.; Irizarry, R. A.; Astrand, M.; Speed, T. P. *Bioinformatics* **2003**, *19*, 185–193.
- (23) Ballman, K. V.; Grill, D. E.; Oberg, A. L.; Therneau, T. M. *Bioinformatics* **2004**, *20*, 2778–2786.
- (24) Warrack, B. M.; Hnatyshyn, S.; Ott, K. H.; Reily, M. D.; Sanders, M.; Zhang, H.; Drexler, D. M. *J. Chromatogr. B* **2009**, *877*, 547–552.
- (25) Zelena, E.; Dunn, W. B.; Broadhurst, D.; Francis-McIntyre, S.; Carroll, K. M.; Begley, P.; O'Hagan, S.; Knowles, J. D.; Halsall, A.; Wilson, I. D.; Kell, D. B. *Anal. Chem.* **2009**, *81*, 1357–1364.
- (26) Rocke, D. M.; Durbin, B. *Bioinformatics* **2003**, *19*, 966–972.
- (27) Huber, W.; von Heydebreck, A.; Sultmann, H.; Poustka, A.; Vingron, M. *Bioinformatics* **2002**, *18*, 96–104.
- (28) Karakach, T. K.; Wentzell, P. D.; Walter, J. A. *Anal. Chim. Acta* **2009**, *636*, 163–174.
- (29) Gika, H. G.; Theodoridis, G. A.; Wingate, J. E.; Wilson, I. D. *J. Proteome Res.* **2007**, *6*, 3291–3303.
- (30) Tautenhahn, R.; Bottcher, C.; Neumann, S. *BMC Bioinf.* **2008**, *9*, 504.
- (31) Durbin, B. P.; Hardin, J. S.; Hawkins, D. M.; Rocke, D. M. *Bioinformatics* **2002**, *18*, S105–S110.
- (32) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. *Anal. Chem.* **2006**, *78*, 4281–4290.
- (33) Dudoit, R.; Yang, Y. H.; Callow, M. J.; Speed, T. P. *Stat. Sin.* **2002**, *12*, 111–139.
- (34) Ritchie, M. E.; Silver, J.; Oshlack, A.; Holmes, M.; Diyagama, D.; Holloway, A.; Smyth, G. K. *Bioinformatics* **2007**, *23*, 2700–2707.
- (35) Wold, S.; Esbensen, K.; Geladi, P. *Chemometr. Intell. Lab. Syst.* **1987**, *2*, 37–52.

Appendix F: Reproduction Rights



Title: 2D NMR-Based Metabolomics
Uncovers Interactions between
Conserved Biochemical Pathways
in the Model Organism
Caenorhabditis elegans

Logged in as:
Steven Robinette

[LOGOUT](#)

Author: Yevgeniy Izrayelit, Steven L.
Robinette, Neelanjana Bose,
Stephan H. von Reuss, and Frank
C. Schroeder

Publication: ACS Chemical Biology

Publisher: American Chemical Society

Date: Feb 1, 2013

Copyright © 2013, American Chemical Society

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

[BACK](#)[CLOSE WINDOW](#)



Title: 2D NMR-Based Metabolomics
Uncovers Interactions between
Conserved Biochemical Pathways
in the Model Organism
Caenorhabditis elegans

Logged in as:
Steven Robinette

[LOGOUT](#)

Author: Yevgeniy Izrayelit, Steven L.
Robinette, Neelanjana Bose,
Stephan H. von Reuss, and Frank
C. Schroeder

Publication: ACS Chemical Biology

Publisher: American Chemical Society

Date: Feb 1, 2013

Copyright © 2013, American Chemical Society

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

[BACK](#)[CLOSE WINDOW](#)



Title: 2D NMR-Based Metabolomics
Uncovers Interactions between
Conserved Biochemical Pathways
in the Model Organism
Caenorhabditis elegans

Logged in as:
Steven Robinette

[LOGOUT](#)

Author: Yevgeniy Izrayelit, Steven L.
Robinette, Neelanjana Bose,
Stephan H. von Reuss, and Frank
C. Schroeder

Publication: ACS Chemical Biology

Publisher: American Chemical Society

Date: Feb 1, 2013

Copyright © 2013, American Chemical Society

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

[BACK](#)[CLOSE WINDOW](#)

From: Rights and Permissions <permission@karger.com>
Subject: **AW: Reproduction permission for thesis**
Date: 10 May 2013 09:24:16 GMT+01:00
To: "Robinette, Steven" <steven.robinette08@imperial.ac.uk>
▶ 1 Attachment, 7.0 KB

Dear Dr. Robinette,

Thank you for your e-mail. As to your request, I am pleased to inform you that permission herewith is granted to use your article

Jaureguiberry, G. et al: Nephron Physiol 2012;122:1-6 (DOI:10.1159/0003-9989)

to be reproduced in your upcoming PhD thesis.

Please note that this is a non-exclusive permission, hence any further use, edition, translation or distribution, either in print or electronically, requires written permission again as this permission is valid for the above mentioned purpose only.

This permission applies only to copyrighted content that S. Karger AG owns, and not to copyrighted content from other sources. If any material in our work appears with credit to another source, you must also obtain permission from the original source cited in our work. All content reproduced from copyrighted material owned by S. Karger AG remains the sole and exclusive property of S. Karger AG. The right to grant permission to a third party is reserved solely by S. Karger AG.

Thank you for your understanding and cooperation.

Hopefully, I have been of assistance to you with the above.

Yours sincerely,

Silvia Meier
Rights and Permissions
t +- 1 61 306 1- 75
permission@karger.com
S. Karger AG, Medical and Scientific Publishers, Allschwilerstrasse 10, -009 Basel, Switzerland
t +- 1 61 306 11 11, f +- 1 61 306 12 3- , www.karger.com



Von: Robinette, Steven [mailto:steven.robinette08@imperial.ac.uk]

Gesendet: Mittwoch, 8. Mai 2013 18:22

An: Rights and Permissions

Betreff: Reproduction permission for thesis

Dear Karger Permissions,

I am an author on the paper below and would like reproduction permission for my PhD thesis, which I will submit at the end of this month. There is no link to the RightsLink service on this paper and I would appreciate the relevant documents or instructions to obtain reproduction rights.

Thanks,
-Steven

<http://www.karger.com/Article/FullText/349989>

**Title:** Statistical Spectroscopic Tools for
Biomarker Discovery and
Systems Medicine**Author:** Steven L. Robinette, John C.
Lindon, and Jeremy K. Nicholson**Publication:** Analytical Chemistry**Publisher:** American Chemical Society**Date:** May 1, 2013

Copyright © 2013, American Chemical Society

Logged in as:
Steven Robinette
Account #:
3000654864

LOGOUT

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK

CLOSE WINDOW

From: "Robinette, Steven" <steven.robinette08@imperial.ac.uk>
Subject: **Reproduction permission for thesis**
Date: 28 May 2013 14:54:36 GMT+01:00
To: reprints@biomedcentral.com

Dear BioMed Central,

I am an author on the review below and would like reproduction permission for my PhD thesis, which I will submit at the end of this month. I would appreciate the relevant documents or instructions to obtain them.

<http://genomemedicine.com/content/4/4/30>

Best,
-Steven

This publication was originally printed in *Genome Medicine* by BioMed Central.

Open access

All Research, Case reports, Databases, Methods, Open debates and Software articles published by *Genome Medicine* are made freely and permanently accessible online immediately upon publication, without subscription charges or registration barriers. Further information about open access can be found [here](#). In addition, the journal also publishes reviews and other article types that are only available by subscription.

Authors of Research, Case reports, Databases, Methods, Open debates and Software articles published in *Genome Medicine* are the copyright holders of their articles and have granted to any third party, in advance and in perpetuity, the right to use, reproduce or disseminate the article, according to the [BioMed Central copyright and license agreement](#).

BioMed Central Open Access Charter

Every peer-reviewed research article appearing in any journal published by BioMed Central is 'open access', meaning that:

1. The article is universally and freely accessible via the Internet, in an easily readable format and deposited immediately upon publication, without embargo, in an agreed format - current preference is XML with a declared DTD - in at least one widely and internationally recognized open access repository (such as PubMed Central).
2. The author(s) or copyright owner(s) irrevocably grant(s) to any third party, in advance and in perpetuity, the right to use, reproduce or disseminate the research article in its entirety or in part, in any format or medium, provided that no substantive errors are introduced in the process, proper attribution of authorship and correct citation details are given, and that the bibliographic details are not changed. If the article is reproduced or disseminated in part, this must be clearly and unequivocally indicated.

BioMed Central is committed permanently to maintaining this open access publishing policy, retrospectively and prospectively, in all eventualities, including any future changes in ownership.



Title: NMR in Metabolomics and Natural
Products Research: Two Sides of
the Same Coin

Logged in as:
Steven Robinette

[LOGOUT](#)

Author: Steven L. Robinette, Rafael
Brüschweiler, Frank C.
Schroeder, and Arthur S. Edison

Publication: Accounts of Chemical Research

Publisher: American Chemical Society

Date: Feb 1, 2012

Copyright © 2012, American Chemical Society

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

[BACK](#)[CLOSE WINDOW](#)



Title: Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery

Logged in as:
Steven Robinette

LOGOUT

Author: Kirill A. Veselkov, Lisa K. Vingara, Perrine Masson, Steven L. Robinette, Elizabeth Want, Jia V. Li, Richard H. Barton, Claire Boursier-Neyret, Bernard Walther, Timothy M. Ebbels, István Pelczer, Elaine Holmes, John C. Lindon, and Jeremy K. Nicholson

Publication: Analytical Chemistry

Publisher: American Chemical Society

Date: Aug 1, 2011

Copyright © 2011, American Chemical Society

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK

CLOSE WINDOW