THE ROLE OF BIOCHEMISTRY, MODEL ORGANISMS, AND SEPARATION SCIENCE TO IMPROVE METABOLOMICS OUTCOMES: A PATH TO INTEGRATION

by

GONÇALO JORGE PERES CARDOSO GOUVEIA

(Under the Direction of Arthur S. Edison)

ABSTRACT

Metabolomics is a multidisciplinary field with a wide range of applications from basic science to translational medicine. The small molecules it can detect are direct measurements of cellular function and can aptly describe physiological states and/or pathologies even in the absence of phenotypes. The larger the number of identified metabolites, the greater the scope and detail of the metabolic network uncovered. However, metabolomics is limited in the number of metabolites it identifies and their function in a metabolic network. Metabolite identification is technically challenging, with most metabolomics studies relying on chemical standards and metabolite spectral databases that represent only but a small fraction of the metabolites that can be detected (predominantly using mass spectrometry and nuclear magnetic resonance spectroscopy; MS and NMR). Thus, novel strategies are needed for *de novo* structural elucidation of unknown metabolites. Here, I describe two critical developments to overcome this challenge. First, a method capable of generating ample quantities of a reference material (RM) of any matrix type that is robust to changes over the course of time. Currently there are no easily available

metabolomics RMs for the model organism *Caenorhabditis elegans* therefore, we generated the first metabolomics *C. elegans* RM. Second, using this RM as a pivotal element, I developed an experimental design that uses semi-preparative fractionation to integrate liquid chromatography (LC)-MS and NMR, two analytical platforms challenging to integrate, and yet essential for the confident identification of metabolites. Metabolomics alone, is necessary but not sufficient to derive mechanistic insight into the interactions of the metabolites it measures with other biomolecules. This functional characterization has been traditionally achieved through biochemical and genetic approaches that strongly rely on model organisms. Using the bacterium *Salmonella enterica*, I demonstrate that metabolomics can provide a wider view the metabolic effects of 2-amino acrylate (2AA) stress, while carefully planned media supplementation experiments confirmed and expanded on the damage to serine hydroxymethyltransferase and the ensuing effects on the measured metabolites. The complementarity of these three approaches helps addressing two longstanding challenges in metabolomics: metabolite identification and determining their role in the organism; ultimately expanding the tools needed to tackle the complexity of metabolism.

INDEX WORDS: Metabolomics; *C. elegans*; *S. enterica*; Metabolite identification; LC-MS; NMR; Reference Material; IBAT; ridA; Fractionation

THE ROLE OF BIOCHEMISTRY, MODEL ORGANISMS, AND SEPARATION SCIENCE TO IMPROVE METABOLOMICS OUTCOMES: A PATH TO INTEGRATION

by

GONÇALO JORGE PERES CARDOSO GOUVEIA

BSC. (HONS) UNIVERSITY OF GLAMORGAN, UNITED KINGDOM, 2005

MSC. UNIVERSITY OF SOUTH LONDON, UNITED KINGDOM, 2006

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2022

THE ROLE OF BIOCHEMISTRY, MODEL ORGANISMS, AND SEPARATION SCIENCE TO IMPROVE METABOLOMICS OUTCOMES: A PATH TO INTEGRATION

by

GONÇALO JORGE PERES CARDOSO GOUVEIA

Major Professor:	Arthur S. Edison
Committee:	Diana M Downs
	Maria Belen Cassera
	Christine M Szimansky
	Ford Ballantyne IV

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2022

DEDICATION

To my wife Ana, and my two fur babies Lola and Margot drivers of most of the inspiration for this journey…

The remaining inspiration came from the community, the people that came to my posters, the ones that asked questions during my presentations, the ones that gave me 5 min of their time when they didn't have to, the ones that cared… This dissertation is also for all of you.

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Arthur Edison. He is and will always be, my mentor, my teacher and I hope one day my colleague. It is an understatement to thank him for his support and advice throughout my journey, he was essential and without him impossible to complete. His encouragement allowed me to take on challenges without fear of failure, but when at times I did fail, he was there with answers, support, encouragement and always a path forward, or sometimes sideways, but towards a goal, an inspiring vision and the big picture in mind. The skills and training I have received from him are priceless, he made me a better scientist and all I learned will be front-and-center for the rest of my career and for that I will always be thankful to Art.

A thank you to Prof. Downs, Prof. Szymanski, Prof. Wells, Dr. Cassera, Dr. Ballantyne and Prof Wood for sage advice and support as well as the CIDC group that are too many to name but have my gratitude.

I want to acknowledge “team worm”, Brianna Garcia, Max Colonna and Amanda Shaver, without them this adventure would have not been anywhere near as rewarding, with every challenge, frustration and success that came along the way, you shared the weight and made the worst less bad, and the good much sweeter. A big thank you to Michael Judge for endless brainstorming sessions, discussions, scientific, personal and philosophical rabbit holes often just for fun! To Bif, Tyler, Francesca, Nick and Jackie the original Edison lab at UGA, they made PhD life a little more amusing and easier that would have been a lot more bitter without you all. Thank you to Iris and Karen, who sometimes

seemingly magically make things happen, they rock! To John, Laura and Pam for making everything just work. The list is long of the many countless people who helped me along the way, they may not be named but they know I'm wholeheartedly grateful! I would like to thank my family for their unwavering support in everything I do. Special thanks to Graziella Li-Ship my first ever scientific mentor, for showing me what science looks like, that details matter, that rigor is not just a word, but rather the glue that makes science truth.

Finally, a thank you that pales all other before, to my two tail wagging ladies Margot and Lola for their unconditional love and for keeping me sane and motivated. And lastly to my wife Ana Bento for being amazing, she challenges me to be a better human being and a better scientist and for that she deserves a whole page ….

To Ana Bento.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

CHAPTER

1 INTRODUCTION 1

1.1 Metabolomics 5

1.2 Reference materials in untargeted metabolomics 26

1.3 Metabolite identification 29

1.4 Integrating model organisms, genetics and biochemistry with metabolomics 35

1.5 References 38

2 LONG-TERM METABOLOMICS REFERENCE MATERIAL 51

Foreword 52

Abstract 53

Introduction 54

Results 57

Discussion and Conclusion 63

Methods 65

References 69

3 A METABOLITE FRACTION LIBRARY APPROACH FOR IMPROVED ANNOTATION IN UNTARGETED METABOLOMICS ACROSS ANALYTICAL PLATFORMS....72

Foreword 73

Abstract 74

Introduction 75

Methods....81

Results....87

Discussion and Conclusion 100

References 110

4 PROTON NUCLEAR MAGNETIC RESONANCE METABOLOMICS CORROBORATES SERINE HYDROXYMETHYLTRANSFERASE AS THE PRIMARY TARGET OF 2-AMINOACRYLATE IN A RIDA MUTANT OF SALMONELLA ENTERICA....119

Foreword 120

Abstract 121

Importance 121

Introduction 122

Results and Discussion 125

Conclusion 135

Materials and Methods 138

References 143

5 CONCLUSIONS AND FUTURE DIRECTIONS 150
A perspective on IBAT applications.....150
Metabolomics fraction library future directions 152
Final remark on metabolomics and biochemical/genetic approaches.....154
References.....155
BIOGRAPHICAL SKETCH 158
APPENDICES 159
A SUPPLEMENTAL MATERIAL FOR CHAPTER 2.....160
B SUPPLEMENTAL MATERIAL FOR CHAPTER 3.....167
C SUPPLEMENTAL MATERIAL FOR CHAPTER 4.....180

CHAPTER 1

INTRODUCTION

'Omics' technologies are defined as data driven approaches that aim at the comprehensive measurement of all the respective elements in a given biological system. Genomics, transcriptomics and proteomics measure DNA, RNA and proteins, with each aiming to comparatively assess changes that can be attributed to disease states, external stressors, genetic mutations, among others.[1] Similarly, metabolomics aims to measure the totality of metabolites, small chemical compounds with a molecular weight lower than 1500 Dalton.[2] These small molecules, once described as "simple" substrates of biochemical reactions at the end of the central dogma cascade, have far-reaching biochemical effects contributing to regulation at all levels of metabolism (Fig. 1.1).[3]

An organism's metabolism consists of countless complex and interconnected biochemical reactions that result in physiological response(s) and/or subsequent behavioral traits(s) (phenotype).[4] These reactions are modulated by fluctuating environments and internal mechanisms that lead to metabolic changes structured to respond to both internal and external stimuli. The collection of all metabolites, that act as precursors, activators, intermediates, inhibitors, products and substrates of numerous cellular processes (from signaling to energy production) constitute the metabolome and depicts the physiological state of an organism.[3-5] For this reason, metabolomics has aptly linked differences in the metabolome to perturbed metabolic pathways,[6] disease states,[7, 8] physiological processes,[9] as well as a proxy for enzyme regulation and genetic variation.[4, 10, 11] The metabolome is a

dynamic and sensitive molecular phenotype quickly changing and adapting to internal and external stimuli, and as such, metabolomics applications have been steadily growing for the past two decades (Fig. 1.2), spanning numerous disciplines with direct applications to real world problems.[12] Despite the scientific community enthusiasm and successful applications, its full potential is still limited by technical challenges as well as a vast number of biochemical reactions and respective metabolites that remain unknown.

Herein, I describe the methodologic workflow of a metabolomics experiment step-by-step. The strengths of this framework, limitations and nuances are raised in light of the work detailed in the following chapters.

Chapter 2 describes a novel method to generate a RM for metabolomics from any biological source. These materials are critical to ensure standardized, reproducible, high-quality data and have been considered a critical challenge in the field, that ultimately increase the confidence of the biological interpretation and translational applications of metabolomics findings.

Chapter 3 builds up on the outputs of chapter two. The approach detailed in this chapter harnesses elements from several disciplines and combines them, using RMs as a pivotal element to develop an experimental design capable of generating essential data for metabolite identification. This challenge has been a long-standing bottleneck for metabolomics' outputs, stemming from the challenging integration of analytical instrumentation, essential to derive chemical structures of measured metabolites. This chapter sets a clear path towards addressing this gap in knowledge with exciting possibilities for downstream applications.

Chapter 4 details the application ^{1}H NMR metabolomics to uncover global metabolic shifts that occur in the bacterium *Salmonella enterica ridA* mutant and considers these findings to complement the current biochemical model describing the effect of 2-amino acrylate (2AA) stress in *S. enterica*. Metabolomics experiments often lack in the mechanistic and functional understanding of the measured metabolites. Thus, this approach illustrates the powerful integration of metabolomics with biochemistry/genetics methodologies to overcome this limitation, ultimately deepening our understanding of the metabolism and biochemical reactions.

Finally, Chapter 5 comprises of a summary of my work presented here providing a coherent connection between chapters and future perspectives of the methods developed.

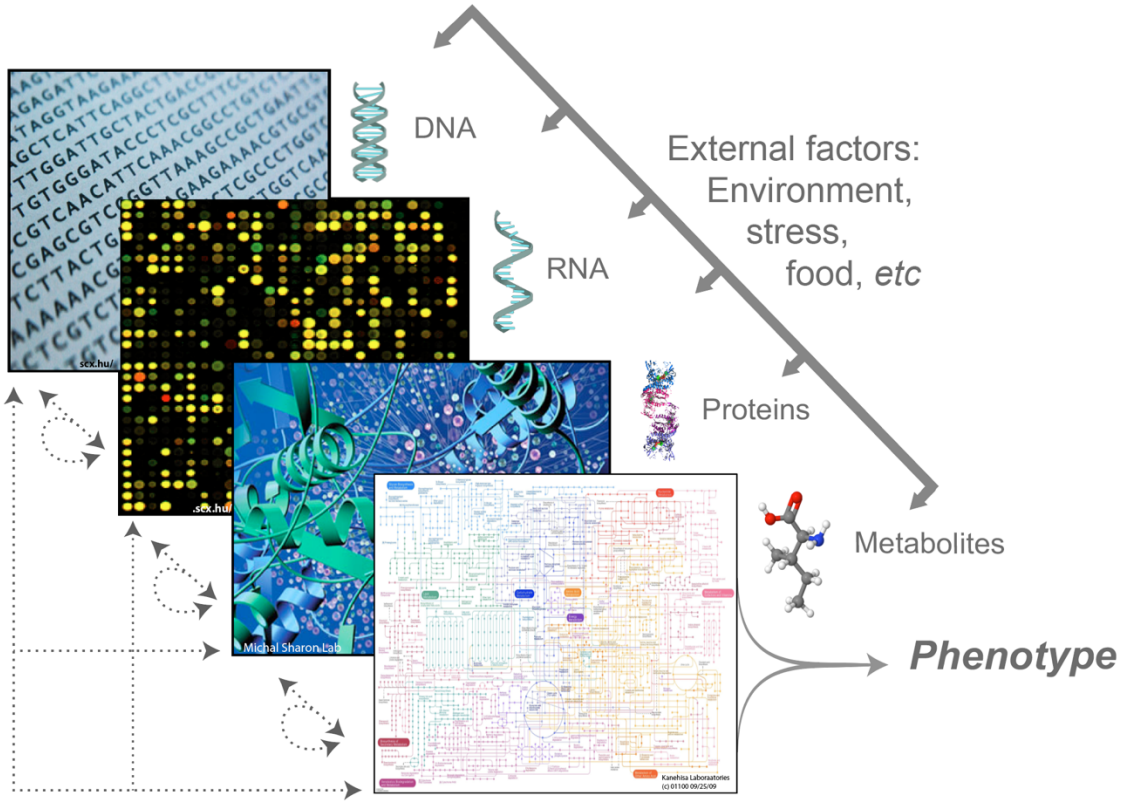


Figure 1.1. Central Dogma of Molecular Biology. The canonical flow of information from DNA to metabolism where the regulation and cascade of information flows from top to bottom. Both curved arrows and two-sided arrows indicate regulation from resulting elements of downstream processes. External factors affect all levels of metabolism. At the last level the vast number of biochemical reactions and their byproducts induce a phenotype.[13]

1.1 Metabolomics

Metabolomics is a transdisciplinary field. At its core relies on analytical chemistry technology to measure and identify small molecules. Developments to automate and improve analytical instrumentation, drove the most commonly used platforms to high-throughput status capable of generating datasets of unprecedented volume.[4, 14, 15] Mass Spectrometry (MS) in tandem with Liquid Chromatography (LC), Gas Chromatography-MS and Nuclear Magnetic Resonance (NMR) spectroscopy (Fig. 1.2) are the most prevalent analytical platforms and while LC-MS accounts for much of this growth, each technology has its own advantages and disadvantages.[16] In addition to the technological limitations, metabolites occupy a diverse chemical space and wide concentration ranges, creating technical challenges that make it impossible to characterize the metabolome in its entirety with a single analytical platform.[17]

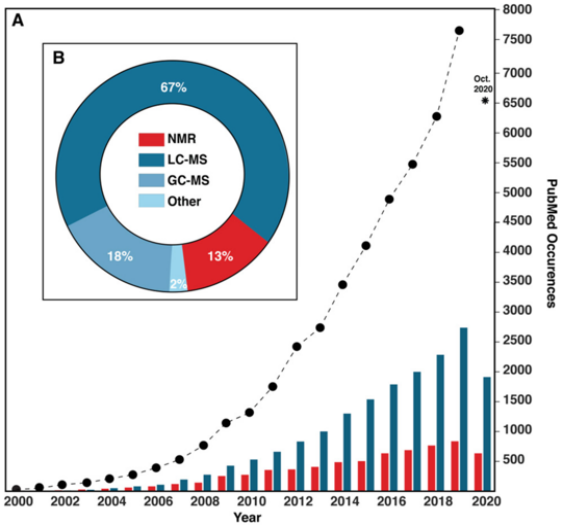


Figure 1.2. Analytical technology trends in metabolomics. A) PubMed search results for metabolomics articles are represented by black dots. Blue bars represent number of publications matching the search terms "metabolomics" and "mass spectrometry". Red bars represent search terms "metabolomics" and "NMR". B) Illustrates the distribution of analytical platforms as percentage of studies deposited to Metabolomics Workbench. All data were obtained Oct. 10, 2020. Reprinted with permission from Anal. Chem. 2021, 93, 1, 478–499. Copyright © 2020 American Chemical Society.[16]

Agnostic to the technology used, metabolomics studies follow the same standard structure and can generally be described in six sequential steps (Fig. 1.3): i) experimental design, ii) sample production/collection, iii) sample preparation, iv) data acquisition, v) data processing, and vi) data analysis.[18] The aim of the following section is not to comprehensively and exhaustively review each step, methods, techniques and materials, since the list would be too large for a single document, but to provide a background to the subsequent chapters and the challenges and limitations I have faced, and to certain degree overcome.

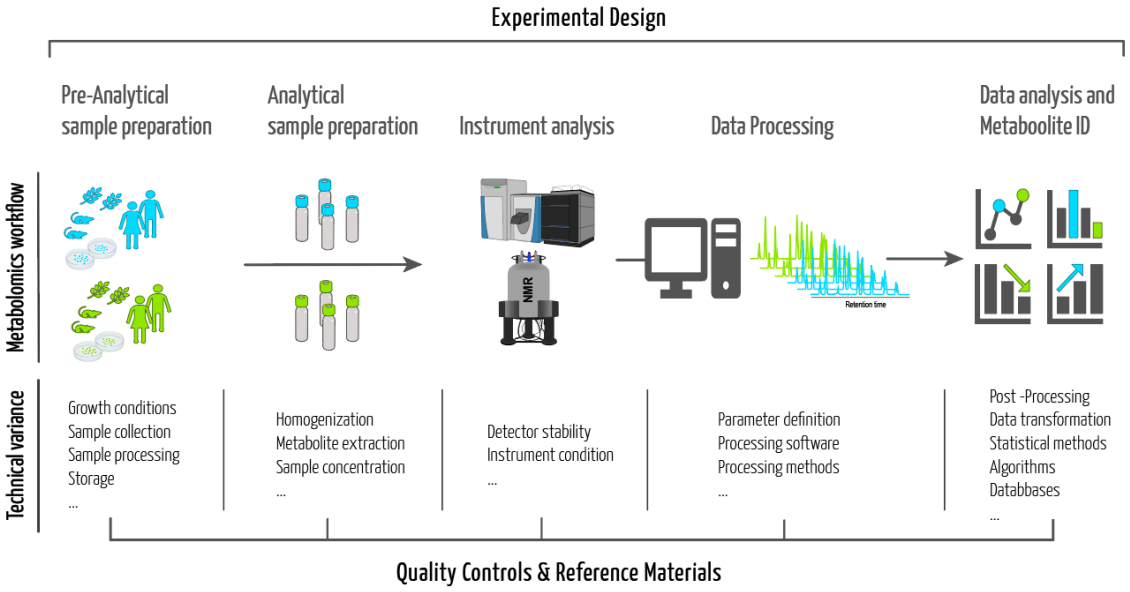


Figure 1.3. Generic metabolomics workflow steps. The process runs from left to right. At each step non-exhaustive examples of technical variance are listed. Experimental design overarches all the steps while quality controls and RMs can provide input at each step.

1.1.1 Experimental design

The design or planning stage is the critical step for a successful metabolomics experiment. This pre-experiment process ensures that the final analytical output is appropriate to answer the biological question at hand through statistical inference.[19] All the

possible parameters at each of the above-mentioned steps are considered; from the number and type of biological samples required and extraction protocols, to the most suitable analytical platform, and the controls used throughout. Pilot studies are often ideal to determine these parameters as well as sample numbers in light of the expected biological effect size and technical variance.[20] In addition, cost, sample logistics, throughput, and instrument type have considerable weight on designing a metabolomics experiment. [15, 21, 22]

The fast-paced-ever-changing nature of the metabolome makes it challenging to predict, plan and control for variance external to the biological system. A vast number of dynamic variables can impact the analytical readout and obscure or confound the true biological effect under investigation. For this reason, Quality Assurance and Quality Control (QA/QC) measures are critical for a successful experiment.[23] Quality management systems have deep historical roots in manufacturing and industrial production where it was developed and implemented to ensure the products met the required specifications.[24]

In metabolomics, QA/QC are critical to generate high quality data and reproducible results.[25] Succinctly, QA can be considered a framework of systematic processes put in place to ensure the quality of the analysis (*e.g.*, standardized protocols, training, reporting, instrument calibrations and maintenance) often arching beyond a study, laboratory and even institution. QCs are measurement specific; they are often materials that can provide a metric/criterion of performance and, can be used to identify and measure unwanted variance of sources external to a single study or means of comparison between studies, instrument and laboratories.[23, 25] Part of the experimental design is determining the most appropriate QC for the study at hand and ensuring the QA processes are complied with.

The diversity of the possible configurations a metabolomics study can take, makes the QC selection process challenging and difficult to standardize particularly in respect to the available instrumentation, sample type and importantly the aims or goals of the study. The latter allows to broadly classify metabolomics experiments as either targeted or untargeted.[23]

The aims of the experiment are deeply connected to the type of QCs used. Targeted studies work in similar ways to well-established analytical assays of testing laboratories (*e.g.*, forensic, drug testing, biomedical), where the study aims to investigate the effects of a panel or group of known metabolites. The metabolites of interest, or closely related compounds, are generally procured as pure chemical standards (or mixtures of pure chemicals) to generate calibration curves and derive the respective concentrations in the samples.[23, 26] In addition, these can be isotopically labeled and added pre-analysis to the study samples to define quality criteria for the analytical process and the unambiguous identification of the unlabeled metabolite.[27] In contrast, untargeted metabolomics is not limited by a select number of known compounds, measuring as many metabolites as possible and leveraging semi-quantitative comparisons between study samples to determine biologically relevant features that are then identified post- data acquisition. Because these metabolites are selected from the data analysis without *a priori* knowledge of their relevance, untargeted metabolomics is often useful to generate novel hypothesis or highlight previously unknown relationships between metabolites and the study conditions.[19, 22, 23, 28, 29]

1.1.2 Sample production/collection

The most common QA practice associated with generating samples for metabolomics is randomization. The concept of experimental randomization was made an essential component of experimental design by R.A. Fisher in the early 1920's, while working at the Rothamsted Experimental Station, a major center for ecological and agricultural research. Fisher posited that experimental design and statistical analysis go hand-in-hand, and that randomly assigning treatments to the agricultural plots was the only valid way to measure the true error (or variance) associated with the fields that had received the same treatment. Previous experiments fell short to rigorously measure (and to correct) this variance, introducing experimenter bias to the treatment layout and ultimately deriving misleading conclusions.[30, 31] Because of the metabolome fast and dynamic response, unrandomized experiments can incorrectly associate metabolic shifts from extraneous factors to study conditions. Randomization allows for these effects to be distributed over all the study samples and therefore allowing for between conditions and between samples variances to be aptly calculated.[19, 23, 25]

This approach still remains a key experimental consideration, especially in plant metabolomics, where field experiments require careful planning, time and considerable resources to ensure that samples of different experimental groups represent (as best as possible) the varying environmental conditions.[32] Similarly, this concept is transferable to any biological sample, from microbial cultures, where environmental factors can be carefully controlled, to human samples where controlled experiments are extremely hard to implement. Nevertheless, every experiment will have some form of external factor (*i.e.,* collection time/order, container type, temperature, processing speed, storage time, operator

capacity, *etc.*) of unknown magnitude. Appropriate QA and QC allow for these effects in the final measured metabolite concentration to be quantified and adjusted accordingly. [19]

1.1.3 Sample preparation

Following sample collection/production, the integrity of the sample must be maintained until analysis. Flash freezing after collection with liquid nitrogen (approximately -196 ^{O}C) is a popular and logistically feasible method, suitable for storage and transport.[18] These low temperatures, prevent metabolic changes by increasing the activation energy necessary for both enzymatic and non-enzymatic reactions.[33] However, the succeeding steps after collection are often incompatible with such low temperatures, and in such cases, samples can be kept cold with ice (0 ^{O}C) or dry-ice (-79 ^{O}C). Other alternatives such as the addition of organic solvents, can effectively quench metabolism, but because of the metabolome diversity solvent-metabolite reactions are possible and difficult to determine its effects.[34, 35]

Sample preparation or sample extraction refers to the process of separating the small molecules of interest from proteins, cellular structures and other larger biological molecules in the biological sample.[32, 34, 36-39] To effectively carry this process out and reduce extraction variability, it is important that cells and/or tissues (unnecessary for biofluids) are broken down into small particles. This usually requires a mechanical force (*i.e.*, bead-beating, pestle and mortar, *etc.*) at low temperatures, and often combined with organic solvents to denature and precipitate proteins and macro-molecules. For both biological fluids and homogenized materials, a sequential high-speed centrifugation separates the metabolites in solution from the remnants of the sample. [18, 34, 36, 37]

This simplified description of the sample extraction method misrepresents the complexity and vast number of possible parameters that have a dramatic effect on the number and intensity of the measured metabolites (*e.g.*, extraction time, solvent to sample amount ratio, temperature, centrifuging speed, analyte concentration method, *etc.*).[2, 40] Due to the metabolite's large physicochemical properties, no single extraction protocol is capable of extracting all chemical classes into a single extractant.[41] Despite this diversity, metabolomics extraction methods can be broadly classified as targeting polar or non-polar metabolites. Under these two categories, the methanol-water solvent system and the Bligh and Dyer methods have been used as templates to a wide range of modifications for specific applications and adjusted to different polarity ranges.[36, 38, 39, 41, 42] These modified methods, aside niche applications, generally favor simplicity (especially for non-polar extractions where a single-phase extraction is preferred), efficiency and reproducibility while maximizing the number of metabolites extracted.[39, 40, 43] A soon to be published manuscript detailing the optimization of some elements of this parameter space has been spearheaded by Brianna Garcia, *et al*. in our laboratory, of which I was a contributing author to its development.

Given the large effect of the extraction protocol on the analytical readout, and the large number of possible protocols, these optimal criteria pose a significant analytical challenge to standardize. Thus, the onus to determine quality criteria for individual metabolites is on the selection of appropriate QCs. For targeted metabolomics, the effort is placed pre-analysis, determining concentration ranges, extraction efficiencies, matrix effects and instrument performance so that QA criteria can then provide a high degree of precision and accuracy to the analytes that fall within the QA thresholds.[23, 25, 44]

For untargeted approaches, it is difficult (nearing impossible or monetarily prohibitive) to attain the same level of analytical scrutiny as there are no pre-determined metabolites, respective QA criteria, defined limits of quantification nor the identity of the yet-to-be-determined metabolites of interest. Thus, effective untargeted QC materials need to represent the entirety of the sample under investigation. This gave rise to the development of pooled QCs,[45] where a small aliquot from every sample in the study is combined, mixed and divided into identical replicates that could undergo the same processes as the study samples. Through multiple measurements of the identical material, it becomes possible to characterize the precision of the measurements and the magnitude of variance originating from the analytical process.[23] Despite being extremely popular, and useful, these materials have significant limitations in terms of metabolite coverage, logistics and sustainability.[46]

Metabolomics experiments invariably rely on statistical testing to determine the significance of metabolite changes between the study conditions. With a large dynamic range of metabolite concentrations, an equally vast range of effect sizes is expected. Thus, higher number of samples per condition results in better estimation of the population size and structure, and therefore better statistical inferences.[20] However, large sample sizes push the boundaries of the number of samples that can effectively processed and analyzed.[47] Typically, the rate limiting step is at the sample processing stage requiring the study samples to be divided into smaller groups that are then logistically feasible to process. Despite following detailed standard operating protocols (SOPs), these smaller batches of samples will always undergo slightly different and unique set of conditions that are outside of the operator control and generate unique batch effects. Pooled QCs in these instances

are difficult to prepare and often inadequate to account for the variability across batches.[46] This problem has been highlighted by the metabolomics community as a main challenge of the field[48] and is further discussed, together with novel alternatives in Chapter 2.

1.1.4 Data acquisition

The selection criteria for the analytical platform to be used is often dictated by the instrument available and respective expertise. However, with the growing availability of commercial metabolomics service providers and university core facilities that can offer multiple platforms; cost per sample, metabolite coverage, reproducibility and limits of quantification become important considerations.[49] Because the analytical platform to be used has repercussions throughout the entire metabolomics workflow, sample specific characteristics (*i.e.*, sample size, number of samples, collection and storage, *etc.*) can also dictate the most suitable platform, in addition to the metabolite class of interest to the biological question at hand (*i.e.,* polar, non-polar, aminoacids, nucleic acids, sugars, lipids, *etc.*). Samples with limited biomass or low metabolite extraction yield and/or the metabolites of interest are low level metabolites, sensitivity is a major decision factor; while epidemiologic studies with several thousands of samples favor analytical robustness and reproducibility. Amongst the wide selection of analytical chemistry detectors, NMR and MS have been the predominant technologies in metabolomics (Fig. 1.2). Throughout my thesis, I will focus specifically on NMR and LC-MS as two orthogonal and complementary analytical platforms, each with their own advantages and disadvantages.[2, 15-17, 29] The following background for these two platforms and considerations for metabolomics will highlight concepts further discussed in the coming chapters.

NMR:

Nuclear Magnetic Resonance spectroscopy, a non-destructive technique, relies on an external magnetic field to align magnetically active nuclei excited with a radiofrequency wavelength pulse that on returning to ground state, generate an NMR signal. This signal or resonance is indicative of the chemical environment of the nuclei in a molecule. The simplest NMR experiment of a single compound yields a spectrum that details (i) the distinct environments of each nucleus in a molecule as a chemical shift, (ii) the number of the same nucleus in the sample as a peak of proportional area, and (iii) the nucleus relationship with neighboring interacting nuclei that can split the signal into smaller patterns known as multiplets.[50] Thus, the collection of peaks (or features) in an NMR spectrum can provide not only atomic level detail about the chemical structure of a molecule but also quantification without the requirement of calibration curves or matched chemical standards. Despite the extremely useful information that can be drawn from other magnetically active nuclei (*e.g.*, ^{15}N, ^{13}C, ^{31}P, *etc.*), metabolomics predominantly measures hydrogen resonances (^{1}H) due to their high natural abundance (99.98%), prevalence in organic molecules. [51, 52]

After metabolite extraction, NMR samples are required to be solubilized in deuterium-containing solvents. Deuterium atoms (^{2}H) in solution serve as a reference to control the magnetic field in the sample (lock) and because the spectrometer is set to measure ^{1}H resonances, the large amplitude features that would arise from undeuterated solvents are reduced. Deuterated water (D_2O), methanol-d4 (MeOH-D4) and chloroform ($CDCl_3$) are usual solubilizing solutions depending on the extraction solvent system and the target metabolites polarity.[52, 53] In addition, metabolites that contain acidic or basic

groups can be affected by pH especially in aqueous solutions, resulting in chemical shift changes. Therefore, sodium or potassium phosphate salts are added as pH buffers to minimize chemical shift changes that are detrimental for metabolomics analyses [53]

To obtain an NMR spectrum, the spectrometer needs to receive a set of instructions and parameters of how the data is generated and collected. This set of instructions is termed pulse sequence. For metabolomics four main pulse sequences are routinely used to collect one-dimensional ^{1}H NMR profiles of biological samples: noesypr and PURGE,[54] are sequences that are intended to minimize the resultant large resonances from solvents, particularly useful in mixed D_2O/H_2O solutions and, CPMG and PROJECT,[6] which define parameters that allow to remove/minimize unwanted broad signals originating from co-solvated macromolecules. The detailed composition and design of these pulse sequences are beyond the remit of this document. However, these have been optimized and designed to generate high quality metabolite profiles of complex biological samples in less than 10 min per sample. The fast acquisition times paired with temperature-controlled autosamplers and fast sample exchange, allow for 1000s of samples to be analyzed without interruption, with very small (if any) reduction in instrument performance. These advantages make NMR spectroscopy high-throughput and highly reproducible, thus, particularly appealing for long-term, large-scale metabolomics studies. [2, 16, 49, 51, 52]

In respect to the quantification of metabolites, NMR limit of detection (LOD) is highly dependent on a number of factors: the strength of the magnetic field, the type of probe, the volume, diameter and type of tube, the type of sample/matrix as well as the spectrometer itself and acquisition parameters have a significant effect on the intensity of the observed peak.[55] As such it is not straightforward to assign a general lowest

concentration for NMR metabolomics. However, as long as the feature is detected and above the baseline it can be quantified as the dynamic range of NMR is unrivaled to any other analytical platform used in metabolomics even at its lowest intensities.[16] Nonetheless, it is customary in our lab to prepare pure chemical standards above 100 µM in a 600 MHz magnet with a 5mm NMR tube, to ensure complete characterization of all ^{1}H chemical shifts. One of the biggest challenges of NMR is spectral overlap in complex mixtures. The large number of features that originate from biological samples create overlapping regions that are difficult to quantify.[51, 52] Several pre-analysis methods have been developed to reduce this overlap which have been shown to work for specific applications but have not become standard NMR metabolomics practices.[43, 56] An alternative to experimental methods to simplify spectra, is the development of pulse sequences capable of removing peaks multiplicity (decoupling). Despite being an attractive approach there are still concerns over implementation, increase of acquisition time, reduction of peak intensities and peak broadening and the loss of structural information codded into the multiplet patterns.[57, 58]

All things considered, NMR has unique advantages over other platforms and remains a powerful analytical technique for metabolomics showing a steady increase in the number of publications over the past decade (Fig. 1.2). A more in-depth discussion of NMR applications, strengths and limitations are further discussed in the following sections and chapter 3.

LC-MS:

A high-resolution LC-MS profile of a metabolomics sample typically contains thousands of spectral features. This large swath of data is only possible due to the physical separation of the analytes by LC and the high detection sensitivity of the mass spectrometer.[2] The separation of analytes by LC relies on the equilibrium achieved between two immiscible phases, the liquid phase and the stationary phase. The transfer of the analytes between these two phases is driven by the chemical properties of the mobile phase, often a mixture of solvents with different physicochemical properties that changes in composition over the course of time (*i.e.*, gradient) and a stationary phase, colloquially termed LC column, that is typically commercially available with varying dimensions and physicochemical properties.[26, 28, 39] As the analytes interact with each of the phases, driven by their specific chemical properties (*i.e.,* size, polarity, charge, *etc.*) they get retained differently, start separating from each other and leave the system at different times aided by a constant flow of mobile phase and respective system pressure. This exit timing, or elution, can be measured in time and is normally referred to as retention time.[59]

The liquid output of the LC and the gaseous and ionized requirement for MS inputs was a major scientific challenge for several decades. These incompatibilities were finally overcome with the development of several LC to MS interfaces. Of note to metabolomics, the electrospray ionization interface (ESI) is considered a soft ionization method that favors the stability of the resulting ionized analytes by minimizing molecular fragmentation pre-mass analysis.[60] The ESI atomizes the LC liquid eluate into a fine spray that is then converted into gas phase ions by an electric field from a high voltage current. However, the charge applied to transform metabolites into positive or negative charged species and

atomize the liquid droplets also induces other ions in solution to bind to metabolites creating adducts (denoted as M+, molecular ion, and the corresponding adducts (*e.g.*, $[M+Na]^+$, $[M+K]^+$, $[M+NH_4]^+$, *etc.*)).[61] This well-documented consequence of the ionization process has implications in metabolite identification and quantification, which led to the development of several computational methods to correct or mitigate this effect, post- data collection. [62]

Succinctly, the general schematic of a mass spectrometer consists of an inlet, an ionization source, a mass/ion analyzer and a detector. In hyphenated systems, the inlet and ionization source are often one and the same, where the sample introduction and ionization are carried out by the type of interface. From here the ions travel into the mass analyzer and are sorted according to their mass and charge. As these ions reach the detector a signal is generated that encodes the mass to charge ratio (m/z, which can be converted into molecular weight) and respective intensity proportional to the number of ions detected at a specific chromatographic retention time.[2, 61, 63] The collection of these three elements over the course of the length of an LC run, constitute an LC-MS chromatogram where the retention time and m/z pair constitute a singular feature. It is often said that MS enables the detection of picomolar (10^{-12}) metabolite concentration as an example of its high sensitivity, even though in routine metabolomics studies, such low concentrations of a compound are difficult to measure due to background and matrix effects. The LOD for MS (including sample preparation) with most cases ranges from micromolar (10^{-6}) to nanomolar (10^{-9}) concentrations at its minimum. [64]

LC-MS systems are highly customizable to the analytes of interest. The extraction method and the LC conditions are intrinsically linked and drive the selectivity and

resolution of the chromatographic separation. Extracted non-polar metabolites are typically paired with Reverse Phase (RP) chromatographic conditions, whereas polar extractions are paired with HILIC chromatographic conditions (hydrophilic interaction liquid chromatography). These can vary both in terms of mobile phase composition, gradient, time and type of stationary phase chemistry that have a wide range of commercially available columns for different applications.[2, 15, 36, 60, 65] However, variations of these setups are common, particularly for intermediate polar/non-polar analytes where either stationary phase or solvent systems (or both) can be adapted to the separation goals.

With respect to the mass spectrometer, different configurations of the four generic elements described above give rise a wide variety of commercially available options. Of note to metabolomics, High-Resolution instruments (benchtop time-of-flight and orbitrap) have been increasingly popular providing accurate mass measurements that bring an additional layer of information for metabolite identification in the form of elemental formulas and (in some cases) isotopic distributions.[66] However, due to its high sensitivity, the LC only effectively separates a fraction of the number of metabolites detected, which means that simultaneous detection of ions at the same retention time can lead to the misrepresentation of the number of ions present (ion suppression).[67]

The large number of possible configurations of LC-MS platforms depending on the application is incredibly appealing, however it also makes standardization challenging. Both targeted and untargeted LC-MS metabolomics rely heavily on QCs. These aim to address contaminants originating from solvents and containers, chromatographic performance reduction (columns have limited number of separations due to matrix effects), and small changes in the mass accuracy inherent to the detector. These additional

considerations, often require that large-scale LC-MS studies to be run in batches adding complexity to data analysis and focus on appropriate QCs.[23] Nonetheless, LC-MS is a powerful platform, particularly for secondary metabolites which are often present at low-levels and would otherwise be difficult to detect. Additional discussion of LC-MS limitations and strengths is addressed in following sections and chapter 3.

1.1.5 Data processing

Data generated from either LC-MS or NMR instrumentation is generated as a signal. A considerable number of steps are needed to transform this signal into usable metabolomics data. For NMR there are a number of available software both commercial (*i.e.*, TopSpin, Mnova, ACDLabs, etc.) and, freely available (*i.e.*, NMRpipe[68], ccpNMR[69], *etc*.). NMR data after collection are transformed by a mathematical process (a Fourier Transform, FT) from the time domain (time the nuclei take to return to equilibrium) to a frequency domain which generates a spectrum. This process was first applied to NMR in the 1960s and was extremely laborious, but it is now automatic, parameter free and almost instantaneous requiring only the click of a button.[70] Additional data processing steps (phasing, apodization and baseline correction) unlike the FT operation require operator input and assessment to maximize the amount of information from each spectrum. [71]

The alignment of these spectra from multiple biological measurements in a metabolomics study is critical, and as such, two operations are routinely carried out: referencing, an operation relying on an internal standard (or known metabolite feature) common to all samples that is set at the same chemical shift axis coordinate (i.e., Sodium trimethylsilylpropane-sulfonate at 0 ^{1}H ppm),[71] and alignment, an operation that uses

specialized mathematical algorithms to detect and correct slight deviations of individual features across all spectra (*e.g.*, pH changes, *etc.*).[72] Following spectral and peak alignment, peak picking is a common process to extract peak heights or areas but has some limitations that can result in poor statistical outcomes. In collaboration with Dr. Michael Judge a former Edison lab colleague, we developed a computational workflow to improve and curate peak quantification, while providing a metric for alignment for every peak across all spectra in a study (available at https://github.com/artedison/Edison_Lab _Shared_ Metabolomics_UGA). The result of all these processing steps is a datamatrix where each row represents the intensities of a single spectrum for each column that represents the same chemical shift across all the spectra in a study.

Similar to NMR, LC-MS processing relies on software to carry out a considerable number of processing steps to reach a format where data can be used for statistical analysis and compound identification. Because of the larger number of LC-MS instrument vendors, there is a considerable number of available options, which has been thoroughly debated by the community in respect to the consequences to standardization (particularly of data formats and processing steps).[73] Nonetheless, open source, freely available and community supported options remain available and popular (*e.g.*, MZmine,[74] XCMS,[75] MS-DIAL,[76] *etc.*).

Succinctly, LC-MS metabolomics data processing consists of four distinct steps: (i) peak-picking/detection, (ii) alignment, (iii) adduct removal and (iv) isotope pattern extraction.[32, 74] These steps are a product of significant development, each having a wide range of different methods and algorithms, and as such, their inner workings are beyond the scope of this document. Nonetheless, these steps are critical and have dramatic effects

on metabolite detection and quantification, and therefore, the statistical inferences that can be drawn. Furthermore, the large number of parameters for each of the steps are highly dependent on the user expertise, the type of sample, the instrument used and the metabolites of interest, making data processing challenging to standardize.[77] QCs can be used to benchmark and define processing outcomes, but these need to be well characterized/quantified, which are generally matrix specific and relatively expensive (*e.g.*, SRM1950 NIST human plasma,[78] credentialed *E. coli* cell extracts,[62] QC TruQant from IROA technologies, *etc.*) and therefore difficult to apply to every study.

Independent from the analytical platform of choice, data processing has a significant effect of the final outcomes, and as such accurate and detailed records of parameters and individual steps is to be reported alongside data and respective analysis outputs.[44]

1.1.6 Data analysis

Metabolomics studies invariably measure more metabolites than number of samples. Because of this data structure with more variables than observations, traditional univariate methods alone are not suitable to identify differences between the study conditions in the numerous spectral features.[79] Thus, multivariate methods such as PCA (principal component analysis) and PLS-da (partially least squares discriminant analysis), are effective at handling these matrices, and therefore, widely adopted for metabolite profiling.[2, 16] These methods use linear decompositions to collapse all the measured variables in each spectrum into a single set of coordinates, while retaining much of the variance from the original data (Fig. 1.4). This output dramatically improves the ability to

represent the differences (or lack of) between samples and between the biological conditions under investigation. These mapped coordinates, or score plots, are most useful to visualize the variation between samples of the same group that form a single cluster, and between groups forming separate clusters, as well as to quickly assess the analysis performance from the various QC samples.[80]

Despite either method generating similar decompositions, PCA scores plots are agnostic to the study conditions and provide an unbiased representation of variance and dimensionality reduction which is useful for QA/QC. PLS-da tend to be less descriptive of the dataset variance as they require the class membership of each sample as an input (supervised), thus, introducing a bias in the dimensionality decomposition (Fig. 1.4b). However, the introduction of class membership to the PLS-da algorithm allows to better expose the separation between conditions and better estimate the contributing elements by loading plots and VIP scores.[81] These provide a measure of each variable's importance based on the percentage variation explained by the model and compared to the original data. The assumptions drawn from PLS-da need to be carefully weighted and cross-validated to reduce the potential to overestimate their importance. [80, 81] The Q^2 and R^2 values are typical metrics used to indicate the model fit and the of model consistency (Fig. 1.4c), however, to test the validity of the class separation of the model permutation tests are required.[79, 80]

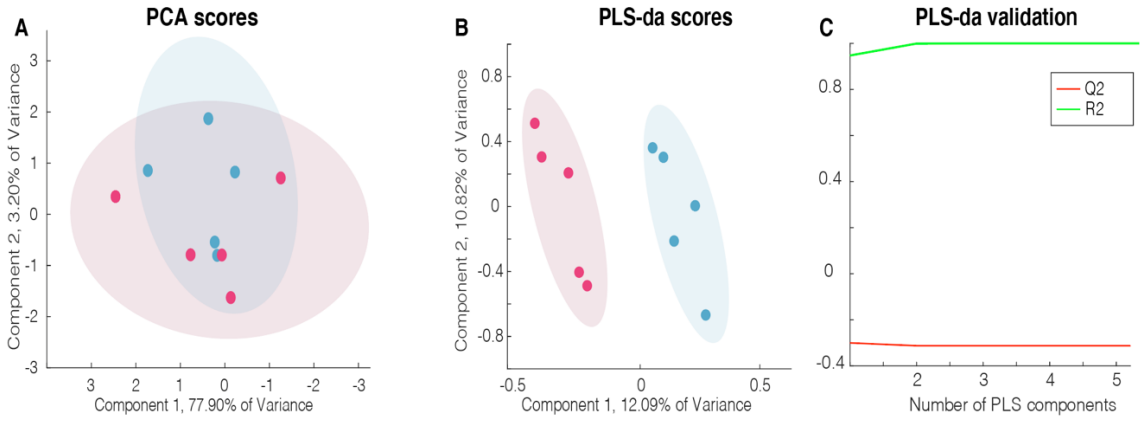


Figure 1.4. Multivariate analysis. The data used to generate the plots above was the output of a random number generator between 0 and 1. The datamatrix consisted of 10 rows (observations/samples) and 100 columns (variables/metabolites). The first five datapoints were assigned to red and the second five datapoints assigned to blue. Each colored point represents one row of the data matrix and 95% confidence interval represented by the shaded circle of the same color. Panels A and B correspond to the PCA and PLS-da scores plots respectively. Panel C illustrates the cross-validation outputs Q^2 and R^2 with the number of components chosen.

Despite protocols clearly defining the amounts of material per sample slight discrepancies are unavoidable and readily observed as differences in total peak intensities between samples. Normalization is a data transformation method carried out pre multivariate analysis to minimize/correct these small differences. A wide variety of normalization methods have been used in metabolomics studies (*e.g.*, PQN,[82] total area, integral, median, *etc.*) and their performance is highly dependent on the concentration/dilution differences between samples.[71, 83]

Often used immediately after normalization, scaling (also known as equalizing) is a data transformation method that aims at minimizing the magnitude differences between the spectral features of interest across the whole dataset.[71, 75] Because of the large dynamic range of the metabolome, different metabolites can have orders of magnitude difference between their concentrations. Multivariate models give more weight and higher rank to

high-concentration metabolites which doesn't necessarily mean they are more important for the biological question at hand. A variety of scaling transformations are commonly used (average based, dispersion based or non-linear) to adjust the magnitude of each peak across all samples (*i.e.*, a column of a datamatrix), to be as similar as possible to the neighboring peaks, while keeping their variance the same (which is not always the case).[84] Both scaling and normalization transform the data so that multivariate models can highlight candidate features of biological relevance that are then tested for significance using univariate methods (*e.g.*, ANOVA, Student's t test *etc.*).[85] Because one sample contributes to a large number of variables, multiple-comparison corrections (Bonferroni,[86] false discovery rate,[85] *etc.*) ensure that metabolites are not significantly different between groups just by random chance alone.[87] These data analysis methods (and other variations) allow to focus the efforts of metabolite identification and annotation to a manageable number of select metabolites that are relevant for the biological question at hand as it would be unrealistic to identify the hundreds to thousands of metabolites present in each spectrum in one single study.[85]

It is reasonable to infer that these approaches are far from perfect. The never-ending efforts to control and understand the sources of undesired variance are often insufficient. The way data are collected, data processing steps, multivariate methods and data transformations have limitations and assumptions which change from metabolite to metabolite and highlight the need of appropriate QC materials that can address these variations to ensure that the effort for the following metabolite identification steps are focused on relevant metabolites.

Finally, metabolite identification is difficult. Database matching facilitates fast spectral annotation, but it has limitations, while identification of unknown metabolites is costly, in time, effort and demands expertise.[52, 66, 88, 89] Once these metabolites are identified, strategies to better understand the function and role of metabolites in the biological system are needed. The detailed mechanistic understanding of the function and role of metabolites is not adequately addressed by metabolomics alone but can be derived by integrating metabolomics with well-established biochemistry/genetics experimental designs.[5, 9] The following three sections introduce these concepts in more detail and will inform the remaining chapters of this document.

1.2 Reference materials in untargeted metabolomics

The focus of untargeted metabolomics is to (i) detect the maximum number of features, (ii) determine patterns of features relevant to the biological question, (iii) minimize the sources of variance outside of the biological conditions and (iv) identify the biologically relevant features. As detailed above, these four goals are intrinsically connected with the QA/QC system, but more importantly rely on QCs to ensure that these findings are translated into real world applications and advance our understanding of biology.[23]

As previously introduced, QCs are materials or physical samples that measure the performance of various metabolomics steps. Extraction blanks and solvents blanks are effective QCs to determine contaminants but fail to address extraction variance, matrix effects and batch effects. For this purpose, materials that comprehensively represent the study samples such as pooled QCs are preferred.[23, 45] However, these are not always

possible to generate (especially in large-scale experiments), are difficult to reproduce and have limited applications beyond a single study failing to act as a link between different studies and analytical platforms (Fig. 1.5a).[46] Alternate materials that can address these limitations are RMs. The general concept of these materials can be broad, often getting denominations based on their characterization, contents or application (*i.e.*, certified, synthetic, biological, calibration, validation, *etc.*).

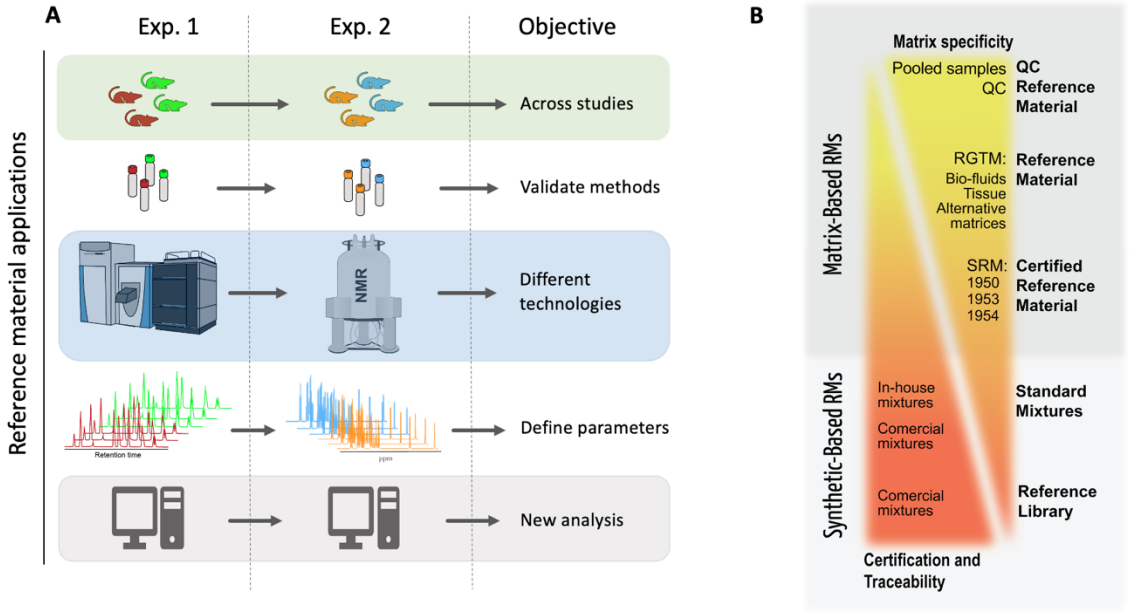


Figure 1.5. Reference Materials in metabolomics. A) represents a non-exhaustive list of different applications for RMs. Each horizontal panel represents comparisons beyond a single study or experiment. B) Classification of the different types of RMs, synthetic and matrix-base. Specific examples of each type of RM can be found inside the colored triangles. From top to bottom the different types of RMs are in order of their specificity to the study biological samples (matrix). In the opposite direction lists the same materials in respect to their ability to be traced to a certified chemical standard. Panel B) is a figure I created for the manuscript Lippa, K.A., […], Wilson, I., Ubhi, B.K., **2021**, *Reference Materials for MS-based Untargeted Metabolomics and Lipidomics: Review by mQACC, Metabolomics*, Metabolomics. I am a contributing author to this manuscript which has been accepted but not yet released.[90]

The official definition from the ISO 2016, 17034 Section 3.3 states that a RM is "a material, sufficiently homogeneous and stable with respect to one or more specified properties, which has been established to be fit for its intended use in a measurement process". Thus, Reference Material is a generic term. The term "properties" can be a quantitative or qualitative measurement (*e.g.*, presence or pre-defined amount of a substances). Its uses may include the calibration of an instrument or method, or an assessment of performance (*e.g.,* system suitability) or/and a quality control. In other words, RMs include CRMs (certified reference materials) that are highly characterized RMs supplied with a certificate of analysis,[78] synthetic-based reference standards, solutions and standard mixtures, and Reference Library (RL) products that are also comprised of high purity standards (Fig. 1.5b). It also includes pooled QC samples despite their short-lived use, which are further distinguished from QCRMs that are available for long-term reference uses and can be defined as adequate materials to measure technical variance, instrument harmonization and comparison, benchmark analyses and, to develop novel methods.[23, 48, 78, 90]

CRMs are the gold standard for metabolomics, where considerable effort is made to identify and quantify individual metabolites, as well as to define expiry dates and storage stability. This, however, is reflected in their cost and matrix specificity. Alternatively, matrix specific, lower-cost QCRMs can be produced in-house to initially address QA/QC needs, however production and sampling variability limit their applicability and long-term use. A significant challenge in design of RMs is a recognized need for contiguous supply of stable, matrix-specific materials which is addressed and further expanded by chapter 2.

1.3 Metabolite identification

Metabolomics has made significant contributions to science that have translated into real world applications.[12, 91, 92] Part of its appeal to the scientific community is the ability to identify novel compounds of biological significance. However, the pace at which the field can identify new metabolites is slow and a limiting factor on the impact of metabolomics as a field.[22, 93] Unlike its "omics" precedents, transcriptomics and proteomics, that have linear combinations of elements (four base pairs and 20 aminoacids respectively), can be sequenced and their structure can be predicted, metabolites have an incredible diversity and an even larger number of possible structural combinations of the same number of atoms which makes analytical measurements capable of capturing thiss information and predictions a significant challenge (*i.e.*, glucose molecular formula, $C_6H_{12}O_6$, generates 496 known chemical structures; chemspider.com).

Therefore, the determination of a metabolite structure (depending on its complexity) can require various analytical measurements that can determine the type and number of atoms present but also how thesy are connected to one another. This has long been a difficult problem, even pre- modern metabolomics, where chemists relied on a large number of measurements (from physical properties to analytical measurements) from numerous different sources (*i.e.*, infrared, ultraviolet, Raman, NMR spectroscopy, MS, *etc*.) to derive the chemical structure of unknown chemicals. With technological developments, the quality and detail of the structural information drawn from these analytical techniques has greatly improved, especially NMR and MS technologies which were once considered difficult, not very specific and of lower value have now become essential.[94] Moreover, the structural information now encoded by NMR and MS spectra

alone is sufficient for searching and retrieving matching spectra to spectral databases of known chemical standards. This provides a fast and automated tentative identification of a large number of metabolites present in biological samples. [95]

Metabolite database annotation by 1D NMR can be done by determining the features that correspond to a single metabolite and then using these features to query against one or several NMR spectral databases. This processes is a common practice in NMR spectroscopy but also challenging, particularly for complex metabolomics one-dimensional (1D) spectra.[96] A method of particular note to this process is the statistical method STOCSY[97] (statistical total correlation spectroscopy). This post-acquisition statistical method takes advantage of the collinearity between features of the same molecule. As the concentration of a metabolite of interest changes from spectra to spectra it is possible to measure the relationship between features as a correlation coefficient. Because there is an inherent variance structure to the dataset, features of the same compound will have high correlation values (>0.9) whereas other features not belonging to the same molecule will have lower correlation values (<0.75). Therefore, it is possible to use high correlation and high covariance values for each feature as indicators of belonging to the same molecule. Albeit useful, because STOCSY works from collected data, it is dependent on the variance of each peak, baseline, phasing and overlap leading to potential false relationships and missed features. [97, 98]

Spectral matching of 1D spectra is a common practice particularly since the launch of the commercial platform Chenomx (Chenomx Inc.). This software allows for 1D spectra to be matched, quantified and adjusted to a database of known reference standards collected at different pH and at different magnetic field strengths. However, with both these

approaches annotation of features is challenging because of spectral overlap. However, the NMR metabolomics overlap problem observed in 1D spectra is significantly reduced with 2D experiments while also providing additional structural information, thus, also stronger evidence of the identity of the metabolite of interest.[49, 99]

In metabolomics the usually collected experiments that provide this added level of structural information are HSQC (heteronuclear single quantum coherence) and TOCSY (total correlation spectroscopy). HSQC experiments are fundamental for metabolite identification giving insight into the chemical structure of metabolites and reduce spectral overlap. HSQC spectra measure correlations between directly bonded ^{1}H-^{13}C nuclei and introduce the carbon chemical shift as a second dimension providing remarkable resolution. This greatly increases the specificity of the spectra and modified versions of HSQC pulse sequence can further distinguish CH_2 from CH and CH_3 features, particularly helpful for *de novo* structural identification.[51, 99] However, this experiment relies on the low natural ^{13}C abundance and long acquisition times are needed to collect high-quality data. Thus, it is not practical to collect these experiments in every sample of a metabolomics study, as such, to minimize the number of samples that require 2D analysis, our lab customarily uses pooled QC samples providing a good representation of the metabolites present in a study.

However, an HSQC experiment on its own does not inform how (and if) carbons are connected to one another. For this, the TOCSY spectra provides important complementary information. ^{1}H-^{1}H TOCSY experiments measure ^{1}H in two dimensions, but also the correlation between ^{1}H in the same molecule. This correlation comes in the form of additional peaks along both dimensions of the spectra. The identification of these

connected ^{1}H nuclei can be directly mapped on to the HSQC spectra ^{1}H dimension and the chemical structure of an unknown compound can start being assembled. Furthermore, these two experiments are the minimum prerequisite for the online resource COLMAR[42, 88], an incredibly powerful tool that allows for the quick and semi-automated annotation of metabolomics samples.

COLMAR allows for spectra to be directly imported onto the webserver, which then generates a peak list that is matched against several different databases providing a matching score and fitting criteria. This annotation method removes the need to deconvolvve individual metabolites and searches peak lists to find patterns of peaks from known metabolites. However, user input is still required to validate the annotations as differences in acquisition parameters, magnet strength and sample pH as well as metabolites with few discerning features can lead to misannotations.[52]

These experiments are only but a subset of the number of NMR experiments that can provide unique complementary structural information. For example, HMBC (heteronuclear multiple bond correlation) experiment is a common experiment for structural elucidation that is rarely reported in NMR metabolomics studies. It generates further complementary data that help connect different spins-systems that are separated by two or three bonds apart. However, the biggest challenges for NMR metabolite identification is the low sensitivity of experiments that measure natural abundance ^{13}C, acquisition time, ^{1}H resonances overlap and interpretation expertise.[49]

LC-MS metabolite identification relies on retention time, accurate mass and fragmentation pattern. As introduced before, high resolution MS provide additional information to the identity of the metabolite of interest, but to achieve high degree of

confidence these values need to match a chemical standard under the same experimental conditions.[44] The accurate mass from an experiment can generate one or more molecular formulas scaling in number with the size of the metabolite but reducing with the accuracy of the mass measurement.[66, 100] Even with a known unique elemental formula, it is still not possible to narrow down to a single compound. Multiple chemical structures are possible form one elemental formula, that yet again, increase in number with the size of the molecule of interest.[100] Tandem MS (MS^2) data is generally collected to further reduce this number of possibilities. These experiments use high collision energies to break the parent ion into smaller fragments with respective *m/z* and intensities, that can then be used to match against spectral libraries.[93]

MS-based databases have grown over the past decade increasing in coverage. However, these are still limited and further complicated by the different instrumentation, collision energies and matching algorithms.[95] In fact, very few metabolomics studies in recent years have identified more than 20% of the detected features which has motivated the development of *in silico*/predicted fragmentation patterns as a means to increase that number.[101] However, there are still concerns over the accuracy of the predictions, but they are expected to improve with the development of novel methods.[102]

Matching experimental data to databases entries relies on similarity metrics to define the criteria and rank of the annotation. However, different database depositories as well as peak detection and deconvolution software adopt a wide variety of different algorithms that perform differently. As such, the same data searched against the same libraries can generate different results. Mass spectral database search is fast and outputs a large number of annotated features, which can be hard to verify or validate. Thus, there is

a need to develop novel approaches that outperform current methods to increase the confidence of the annotation and cater for metabolites that only fragment into a small number of ions or between spectra that have been collected in different instruments and/or with different parameters.

To definitively identify a compound of interest several sources of complementary data are historically required, and arguably, purification and determination of its covalent structure is mandatory.[89] However, this is highly dependent on the complexity of the metabolite of interest. As such, the Chemical Analysis Working Group (CAWG) of the Metabolomics Standards Initiative (MSI)[44] devised a minimum reporting requirement and respective classification for the identification of metabolites in metabolomics studies. After several adjustments since its inception in 2007,[103] current guidelines define metabolite identification confidence as five different levels form (0) strong evidence of the metabolite identification to (4) weak evidence: (0) full 3D structure (1) identified compounds, (2) putatively annotated compounds, (3) putatively characterized compounds, and (4) unknown compounds.[44]

To an extent, the level of confidence is paired with the level of difficulty for each category. Level zero requires additional experiments, different analytical platforms and specific methods to achieve stereochemistry determination and is rarely carried out in metabolomics. Slightly less effort is required for level 1 identifications, where two orthogonal analytical techniques of both the metabolite of interest and a chemical reference standard need to be carried out under identical analytical conditions within the same laboratory. This level is easily achieved by targeted metabolomics studies, where chemical standards dictate the measured metabolites and can provide orthogonal information as to

the identity of the metabolite. Retention time, accurate mass, fragmentation pattern, isotopic composition can be matched in the case of LC-MS studies, whereas for NMR, experiments that offer complementary and orthogonal ^{13}C and ^{1}H data are sufficient to validate the identity of a metabolite.

All database spectral matching annotations that use a single method fall within level two.[44] Thus, most untargeted metabolomics studies that do not rely on chemical standards, will ever only achieve level one annotations unless deemed critical to carry out further analysis for structural elucidation of an unknown metabolite or confirmation of the metabolite annotation. The latter is generally less likely, and often reliant on database metrics outputs to satisfy an annotation requirement to proceed with a biological interpretation. Despite the progress that has been made, there is a need not only to improve compound annotation confidence but also strategies that allow for faster and accurate identification of novel metabolites. Further discussion and details of methods that can address these needs can be found in chapter 4.

1.4 Integrating model organisms, genetics and biochemistry with metabolomics

Metabolomics experiments provide a valuable global view of the metabolic network. The larger the number of metabolites identified, the larger the scope and detail of metabolism that is possible to observe.[4] However, this global view lacks the understanding of how these metabolites interact with other biomolecules within a metabolic network.[3, 5] For this purpose, biochemistry and genetics research have a long history of testing and

defining detailed models of specific biochemical reactions. Naturally, the combination of these two disciplines facilitates detailed mechanistic insights into specific biochemical reactions, while also their relationship to neighboring pathways and metabolites that were not previously thought to be connected (Chapter 4). In addition, because the metabolome is the closest measurement to cellular function, it can define molecular phenotypes specially in studies where no other behavior is observed (*e.g.*, growth, motility, *etc*.).[104]

The natural complementarity between these approaches relies heavily on experimental design and validation of metabolomics findings by biochemical models ultimately maximizing the output of both approaches. Much of what is known about metabolism has been carried out by classical biochemistry and genetics methods that greatly rely on model organisms' short generation times, easy manipulation, and fast response to systematic perturbations. These traits become even more apparent in model microorganisms, of which the bacterium *Escherichia coli* is likely to be the most well-studied model organism. Surprisingly, out of 40,880 metabolomics publications (PubMed: "metabolomics" between 2012 and 2022) only 307 entries were returned with the similar search parameters for *E. coli* (PubMed: "E.coli" and "metabolomics" between 2012 and 2022). The similarly well studied organism *C. elegans* also returned a meek 197 entries.

This highlights the outstanding opportunities in model microorganism metabolomics research. The wealth of information known for these microorganisms, such as fully sequence genomes and repositories of annotated biochemical reactions and pathways (*i.e.*, EcoCyc, YeastCyc, *etc.*) as well as genetic stock centers for numerous organisms that maintain, catalogue and readily distribute different strains and mutants at reasonable costs, are appealing benefits not only for metabolomics but also for the

integration of biochemical experiments with metabolomics approaches.[16, 89] In addition, These resources and organism traits allow to generate replicates with very low genetic variability (often isogenic) and can be grown under well-defined and strictly controlled growth conditions. These qualities are particularly appealing for metabolomics leading to analytical measurements of replicates low in variance improving downstream statistical analyses, and growth inputs can be defined to help inform quantitative models of metabolism. [105, 106]

As introduced earlier, metabolomics relies on materials capable of detecting and quantifying technical variance. The motivation for Chapter 2 was the lack of a *C. elegans* RM for metabolomics studies. A RM of a model organism reference genotype opens the possibility to a systematic metabolome annotation focusing community efforts and validating findings. Because of the large diversity of the metabolome and the large combinatorial possibilities of the building blocks of a metabolite it is hard to predict the composition of a metabolome and the detected features that are yet to be identified.

Model organisms can again be an added value for metabolomics. Annotated genomes and defined biochemical reactions can be used to define and validate expected metabolites. While conserved metabolic pathways can be used to extrapolate findings to other organisms and/or species, classical biochemical methodologies can provide important new targets of investigation in pathways of interest homologous to human biology and biomedicine.[107]

The complementarity between model organisms, biochemistry/genetics methods and metabolomics can address longstanding problems in metabolomics from metabolite identification to determining their role in metabolism, but not without challenges, and

surely a long-term community driven effort is needed, but it is undoubtedly made harder without a common RM.

1.5 REFERENCES

1. IOM, *Evolution of Translational Omics Lessons Learned and the Path Forward*. The National Academies Press: 2012.

2. Dunn, W. B.; Broadhurst, D. I.; Atherton, H. J.; Goodacre, R.; Griffin, J. L., Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev* **2011,** *40* (1), 387-426.

3. Rinschen, M. M.; Ivanisevic, J.; Giera, M.; Siuzdak, G., Identification of bioactive metabolites using activity metabolomics. *Nat Rev Mol Cell Biol* **2019,** *20* (6), 353-367.

4. Downs, D. M.; Bazurto, J. V.; Gupta, A.; Fonseca, L. L.; Voit, E. O., The three-legged stool of understanding metabolism: integrating metabolomics with biochemical genetics and computational modeling. *AIMS Microbiol* **2018,** *4* (2), 289-303.

5. Johnson, C. H.; Ivanisevic, J.; Siuzdak, G., Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol* **2016,** *17* (7), 451-9.

6. Borchert, A. J.; Walejko, J. M.; Guennec, A. L.; Ernst, D. C.; Edison, A. S.; Downs, D. M., Integrated Metabolomics and Transcriptomics Suggest the Global Metabolic Response to 2-Aminoacrylate Stress in Salmonella enterica. *Metabolites* **2019,** *10* (1).

7. Mamas, M.; Dunn, W. B.; Neyses, L.; Goodacre, R., The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Arch Toxicol* **2011,** *85* (1), 5-17.

8. Brindle, J. T.; Antti, H.; Holmes, E.; Tranter, G.; Nicholson, J. K.; Bethell, H. W.; Clarke, S.; Schofield, P. M.; McKilligin, E.; Mosedale, D. E.; Grainger, D. J., Rapid

and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nat Med* **2002,** *8* (12), 1439-44.

9. Joshua, C. J., Metabolomics: A Microbial Physiology and Metabolism Perspective. *Methods Mol Biol* **2019,** *1859*, 71-94.

10. Oliver, S., Systematic functional analysis of the yeast genome. *Trends in Biotechnology* **1998,** *16* (9), 373-378.

11. Gieger, C.; Geistlinger, L.; Altmaier, E.; Hrabe de Angelis, M.; Kronenberg, F.; Meitinger, T.; Mewes, H. W.; Wichmann, H. E.; Weinberger, K. M.; Adamski, J.; Illig, T.; Suhre, K., Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet* **2008,** *4* (11), e1000282.

12. Pinu, F. R.; Goldansaz, S. A.; Jaine, J., Translational Metabolomics: Current Challenges and Future Opportunities. *Metabolites* **2019,** *9* (6).

13. Kanehisa, M.; Goto, S., KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **2000,** *28* (1), 27-30.

14. Bartel, J.; Krumsiek, J.; Theis, F. J., Statistical methods for the analysis of high-throughput metabolomics data. *Comput Struct Biotechnol J* **2013,** *4*, e201301009.

15. Alonso, A.; Marsal, S.; Julia, A., Analytical methods in untargeted metabolomics: state of the art in 2015. *Front Bioeng Biotechnol* **2015,** *3*, 23.

16. Edison, A. S.; Colonna, M.; Gouveia, G. J.; Holderman, N. R.; Judge, M. T.; Shen, X.; Zhang, S., NMR: Unique Strengths That Enhance Modern Metabolomics Research. *Anal Chem* **2021,** *93* (1), 478-499.

17. Bhinderwala, F.; Wase, N.; DiRusso, C.; Powers, R., Combining Mass Spectrometry and NMR Improves Metabolite Detection and Annotation. *J Proteome Res* **2018,** *17* (11), 4017-4022.

18. Higashi, T. W.-M. F. N. L. M., *The Handbook of Metabolomics*. 2012.

19. Dunn, W. B.; Wilson, I. D.; Nicholls, A. W.; Broadhurst, D., The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **2012,** *4* (18), 2249-64.

20. Nyamundanda, G.; Gormley, I. C.; Fan, Y.; Gallagher, W. M.; Brennan, L., MetSizeR: selecting the optimal sample size for metabolomic studies using an analysis based approach. *BMC Bioinformatics* **2013,** *14*, 338.

21. Aristizabal-Henao, J. J.; Lemas, D. J.; Griffin, E. K.; Costa, K. A.; Camacho, C.; Bowden, J. A., Metabolomic Profiling of Biological Reference Materials using a Multiplatform High-Resolution Mass Spectrometric Approach. *J Am Soc Mass Spectrom* **2021,** *32* (9), 2481-2489.

22. Johnson, C. H.; Gonzalez, F. J., Challenges and opportunities of metabolomics. *J Cell Physiol* **2012,** *227* (8), 2975-81.

23. Broadhurst, D.; Goodacre, R.; Reinke, S. N.; Kuligowski, J.; Wilson, I. D.; Lewis, M. R.; Dunn, W. B., Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* **2018,** *14* (6), 72.

24. Khosla, R.; Srivastava, V., Historical Overview of Quality Assurance in Biological Research. In *Quality Assurance Implementation in Research Labs*, Anand, A., Ed. Springer Singapore: Singapore, 2021; pp 1-14.

25. Evans, A. M.; O'Donovan, C.; Playdon, M.; Beecher, C.; Beger, R. D.; Bowden, J. A.; Broadhurst, D.; Clish, C. B.; Dasari, S.; Dunn, W. B.; Griffin, J. L.; Hartung, T.; Hsu, P. C.; Huan, T.; Jans, J.; Jones, C. M.; Kachman, M.; Kleensang, A.; Lewis, M. R.; Monge, M. E.; Mosley, J. D.; Taylor, E.; Tayyari, F.; Theodoridis, G.; Torta, F.; Ubhi, B. K.; Vuckovic, D.; Metabolomics Quality Assurance, Q. C. C., Dissemination and analysis of the quality assurance (QA) and quality control (QC) practices of LC-MS based untargeted metabolomics practitioners. *Metabolomics* **2020,** *16* (10), 113.

26. Ribbenstedt, A.; Ziarrusta, H.; Benskin, J. P., Development, characterization and comparisons of targeted and non-targeted metabolomics methods. *PLoS One* **2018,** *13* (11), e0207082.

27. Weindl, D.; Wegner, A.; Hiller, K., Metabolome-Wide Analysis of Stable Isotope Labeling-Is It Worth the Effort? *Front Physiol* **2015,** *6*, 344.

28. Whiley, L.; Chekmeneva, E.; Berry, D. J.; Jimenez, B.; Yuen, A. H. Y.; Salam, A.; Hussain, H.; Witt, M.; Takats, Z.; Nicholson, J.; Lewis, M. R., Systematic Isolation and Structure Elucidation of Urinary Metabolites Optimized for the Analytical-Scale Molecular Profiling Laboratory. *Anal Chem* **2019,** *91* (14), 8873-8882.

29. Schrimpe-Rutledge, A. C.; Codreanu, S. G.; Sherrod, S. D.; McLean, J. A., Untargeted Metabolomics Strategies-Challenges and Emerging Directions. *J Am Soc Mass Spectrom* **2016,** *27* (12), 1897-1905.

30. Armitage, P., Fisher, Bradford Hill, and randomization. *Int J Epidemiol* **2003,** *32* (6), 925-8; discussion 945-8.

31. Fisher, R. A., The Design of Experiments. **1935**.

32. Rodrigues, A. M.; Ribeiro-Barros, A. I.; Antonio, C., Experimental Design and Sample Preparation in Forest Tree Metabolomics. *Metabolites* **2019,** *9* (12).

33. Stevens, V. L.; Hoover, E.; Wang, Y.; Zanetti, K. A., Pre-Analytical Factors that Affect Metabolite Stability in Human Urine, Plasma, and Serum: A Review. *Metabolites* **2019,** *9* (8).

34. Beltran, A.; Suarez, M.; Rodriguez, M. A.; Vinaixa, M.; Samino, S.; Arola, L.; Correig, X.; Yanes, O., Assessment of compatibility between extraction methods for NMR- and LC/MS-based metabolomics. *Anal Chem* **2012,** *84* (14), 5838-44.

35. Sauerschnig, C.; Doppler, M.; Bueschl, C.; Schuhmacher, R., Methanol Generates Numerous Artifacts during Sample Extraction and Storage of Extracts in Metabolomics Research. *Metabolites* **2017,** *8* (1).

36. Gil, A.; Zhang, W.; Wolters, J. C.; Permentier, H.; Boer, T.; Horvatovich, P.; Heiner-Fokkema, M. R.; Reijngoud, D. J.; Bischoff, R., One- vs two-phase extraction: re-evaluation of sample preparation procedures for untargeted lipidomics in plasma samples. *Anal Bioanal Chem* **2018,** *410* (23), 5859-5870.

37. Lin, C. Y.; Wu, H.; Tjeerdema, R. S.; Viant, M. R., Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics. *Metabolomics* **2007,** *3* (1), 55-67.

38. Nagana Gowda, G. A.; Raftery, D., Quantitating metabolites in protein precipitated serum using NMR spectroscopy. *Anal Chem* **2014,** *86* (11), 5433-40.

39. Ulmer, C. Z.; Jones, C. M.; Yost, R. A.; Garrett, T. J.; Bowden, J. A., Optimization of Folch, Bligh-Dyer, and Matyash sample-to-extraction solvent ratios for human plasma-based lipidomics studies. *Anal Chim Acta* **2018,** *1037*, 351-357.

40. McHugh, C. E.; Flott, T. L.; Schooff, C. R.; Smiley, Z.; Puskarich, M. A.; Myers, D. D.; Younger, J. G.; Jones, A. E.; Stringer, K. A., Rapid, Reproducible, Quantifiable NMR Metabolomics: Methanol and Methanol: Chloroform Precipitation for Removal of Macromolecules in Serum and Whole Blood. *Metabolites* **2018,** *8* (4).

41. Nakayasu, E. S.; Nicora, C. D.; Sims, A. C.; Burnum-Johnson, K. E.; Kim, Y. M.; Kyle, J. E.; Matzke, M. M.; Shukla, A. K.; Chu, R. K.; Schepmoes, A. A.; Jacobs, J. M.; Baric, R. S.; Webb-Robertson, B. J.; Smith, R. D.; Metz, T. O., MPLEx: a Robust and Universal Protocol for Single-Sample Integrative Proteomic, Metabolomic, and Lipidomic Analyses. *mSystems* **2016,** *1* (3).

42. Wang, C.; Timari, I.; Zhang, B.; Li, D. W.; Leggett, A.; Amer, A. O.; Bruschweiler-Li, L.; Kopec, R. E.; Bruschweiler, R., COLMAR Lipids Web Server and Ultrahigh-Resolution Methods for Two-Dimensional Nuclear Magnetic Resonance- and Mass Spectrometry-Based Lipidomics. *J Proteome Res* **2020,** *19* (4), 1674-1683.

43. Ye, T.; Zheng, C.; Zhang, S.; Gowda, G. A.; Vitek, O.; Raftery, D., "Add to subtract": a simple method to remove complex background signals from the 1H nuclear magnetic resonance spectra of mixtures. *Anal Chem* **2012,** *84* (2), 994-1002.

44. Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R., Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **2007,** *3* (3), 211-221.

45. Sangster, T.; Major, H.; Plumb, R.; Wilson, A. J.; Wilson, I. D., A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabonomic analysis. *Analyst* **2006,** *131* (10), 1075-8.

46. Gouveia, G. J.; Shaver, A. O.; Garcia, B. M.; Morse, A. M.; Andersen, E. C.; Edison, A. S.; McIntyre, L. M., Long-Term Metabolomics Reference Material. *Anal Chem* **2021,** *93* (26), 9193-9199.

47. Du, X.; Smirnov, A.; Pluskal, T.; Jia, W.; Sumner, S., Metabolomics Data Preprocessing Using ADAP and MZmine 2. *Methods Mol Biol* **2020,** *2104*, 25-48.

48. Beger, R. D.; Dunn, W. B.; Bandukwala, A.; Bethan, B.; Broadhurst, D.; Clish, C. B.; Dasari, S.; Derr, L.; Evans, A.; Fischer, S.; Flynn, T.; Hartung, T.; Herrington, D.; Higashi, R.; Hsu, P. C.; Jones, C.; Kachman, M.; Karuso, H.; Kruppa, G.; Lippa, K.; Maruvada, P.; Mosley, J.; Ntai, I.; O'Donovan, C.; Playdon, M.; Raftery, D.; Shaughnessy, D.; Souza, A.; Spaeder, T.; Spalholz, B.; Tayyari, F.; Ubhi, B.; Verma, M.; Walk, T.; Wilson, I.; Witkin, K.; Bearden, D. W.; Zanetti, K. A., Towards quality assurance and quality control in untargeted metabolomics studies. *Metabolomics* **2019,** *15* (1), 4.

49. Emwas, A. H.; Roy, R.; McKay, R. T.; Tenori, L.; Saccenti, E.; Gowda, G. A. N.; Raftery, D.; Alahmari, F.; Jaremko, L.; Jaremko, M.; Wishart, D. S., NMR Spectroscopy for Metabolomics Research. *Metabolites* **2019,** *9* (7).

50. Kruk, J.; Doskocz, M.; Jodlowska, E.; Zacharzewska, A.; Lakomiec, J.; Czaja, K.; Kujawski, J., NMR Techniques in Metabolomic Studies: A Quick Overview on Examples of Utilization. *Appl Magn Reson* **2017,** *48* (1), 1-21.

51. Markley, J. L.; Bruschweiler, R.; Edison, A. S.; Eghbalnia, H. R.; Powers, R.; Raftery, D.; Wishart, D. S., The future of NMR-based metabolomics. *Curr Opin Biotechnol* **2017,** *43*, 34-40.

52. Markley, J. L.; Dashti, H.; Wedell, J. R.; Westler, W. M.; Eghbalnia, H. R., Tools for Enhanced NMR-Based Metabolomics Analysis. *Methods Mol Biol* **2019,** *2037*, 413-427.

53. Emwas, A. H.; Roy, R.; McKay, R. T.; Ryan, D.; Brennan, L.; Tenori, L.; Luchinat, C.; Gao, X.; Zeri, A. C.; Gowda, G. A.; Raftery, D.; Steinbeck, C.; Salek, R. M.; Wishart, D. S., Recommendations and Standardization of Biomarker Quantification Using NMR-Based Metabolomics with Particular Focus on Urinary Analysis. *J Proteome Res* **2016,** *15* (2), 360-73.

54. Le Guennec, A.; Tayyari, F.; Edison, A. S., Alternatives to Nuclear Overhauser Enhancement Spectroscopy Presat and Carr-Purcell-Meiboom-Gill Presat for NMR-Based Metabolomics. *Anal Chem* **2017,** *89* (17), 8582-8588.

55. Mompean, M.; Sanchez-Donoso, R. M.; de la Hoz, A.; Saggiomo, V.; Velders, A. H.; Gomez, M. V., Pushing nuclear magnetic resonance sensitivity limits with microfluidics and photo-chemically induced dynamic nuclear polarization. *Nat Commun* **2018,** *9* (1), 108.

56. Yuan, J.; Zhang, B.; Wang, C.; Bruschweiler, R., Carbohydrate Background Removal in Metabolomics Samples. *Anal Chem* **2018,** *90* (24), 14100-14104.

57. Zangger, K., Pure shift NMR. *Prog Nucl Magn Reson Spectrosc* **2015,** *86-87*, 1-20.

58. Ludwig, C.; Viant, M. R., Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochem Anal* **2010,** *21* (1), 22-32.

59. Giddings, C., Dynamics of Chromatography, part 1: Principles and Theory. **1965**.

60. Lu, W.; Bennett, B. D.; Rabinowitz, J. D., Analytical strategies for LC-MS-based targeted metabolomics. *J Chromatogr B Analyt Technol Biomed Life Sci* **2008,** *871* (2), 236-42.

61. Ho, C. S.; Lam, C. W.; Chan, M. H.; Cheung, R. C.; Law, L. K.; Lit, L. C.; Ng, K. F.; Suen, M. W.; Tai, H. L., Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin Biochem Rev* **2003,** *24* (1), 3-12.

62. Mahieu, N. G.; Patti, G. J., Systems-Level Annotation of a Metabolomics Data Set Reduces 25000 Features to Fewer than 1000 Unique Metabolites. *Anal Chem* **2017,** *89* (19), 10397-10406.

63. Zhou, B.; Xiao, J. F.; Tuli, L.; Ressom, H. W., LC-MS-based metabolomics. *Mol Biosyst* **2012,** *8* (2), 470-81.

64. Evard, H.; Kruve, A.; Leito, I., Tutorial on estimating the limit of detection using LC-MS analysis, part I: Theoretical review. *Anal Chim Acta* **2016,** *942*, 23-39.

65. Commisso, M.; Anesi, A.; Dal Santo, S.; Guzzo, F., Performance comparison of electrospray ionization and atmospheric pressure chemical ionization in untargeted and targeted liquid chromatography/mass spectrometry based metabolomics analysis of grapeberry metabolites. *Rapid Commun Mass Spectrom* **2017,** *31* (3), 292-300.

66. Rathahao-Paris, E.; Alves, S.; Junot, C.; Tabet, J.-C., High resolution mass spectrometry for structural identification of metabolites in metabolomics. *Metabolomics* **2015,** *12* (1).

67. Metz, T. O.; Page, J. S.; Baker, E. S.; Tang, K.; Ding, J.; Shen, Y.; Smith, R. D., High Resolution Separations and Improved Ion Production and Transmission in Metabolomics. *Trends Analyt Chem* **2008,** *27* (3), 205-214.

68. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **1995,** *6* (3), 277-93.

69. Skinner, S. P.; Fogh, R. H.; Boucher, W.; Ragan, T. J.; Mureddu, L. G.; Vuister, G. W., CcpNmr AnalysisAssign: a flexible platform for integrated NMR analysis. *J Biomol NMR* **2016,** *66* (2), 111-124.

70. Hosur, R. V.; Kakita, V. M. R., Fourier Transform NMR. In *A Graduate Course in NMR Spectroscopy*, Springer International Publishing: Cham, 2022; pp 89-142.

71. Emwas, A. H.; Saccenti, E.; Gao, X.; McKay, R. T.; Dos Santos, V.; Roy, R.; Wishart, D. S., Recommended strategies for spectral processing and post-processing of 1D (1)H-NMR data of biofluids with a particular focus on urine. *Metabolomics* **2018,** *14* (3), 31.

72. Vu, T. N.; Laukens, K., Getting your peaks in line: a review of alignment methods for NMR spectral data. *Metabolites* **2013,** *3* (2), 259-76.

73. Weber, R. J. M.; Lawson, T. N.; Salek, R. M.; Ebbels, T. M. D.; Glen, R. C.; Goodacre, R.; Griffin, J. L.; Haug, K.; Koulman, A.; Moreno, P.; Ralser, M.; Steinbeck, C.; Dunn, W. B.; Viant, M. R., Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy. *Metabolomics* **2017,** *13* (2), 12.

74. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M., MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010,** *11*, 395.

75. Gowda, H.; Ivanisevic, J.; Johnson, C. H.; Kurczy, M. E.; Benton, H. P.; Rinehart, D.; Nguyen, T.; Ray, J.; Kuehl, J.; Arevalo, B.; Westenskow, P. D.; Wang, J.; Arkin, A. P.; Deutschbauer, A. M.; Patti, G. J.; Siuzdak, G., Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. *Anal Chem* **2014,** *86* (14), 6931-9.

76. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M., MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* **2015,** *12* (6), 523-6.

77. Delabriere, A.; Warmer, P.; Brennsteiner, V.; Zamboni, N., SLAW: A Scalable and Self-Optimizing Processing Workflow for Untargeted LC-MS. *Anal Chem* **2021,** *93* (45), 15024-15032.

78. Simon-Manso, Y.; Lowenthal, M. S.; Kilpatrick, L. E.; Sampson, M. L.; Telu, K. H.; Rudnick, P. A.; Mallard, W. G.; Bearden, D. W.; Schock, T. B.; Tchekhovskoi, D. V.; Blonder, N.; Yan, X.; Liang, Y.; Zheng, Y.; Wallace, W. E.; Neta, P.; Phinney, K. W.; Remaley, A. T.; Stein, S. E., Metabolite profiling of a NIST Standard Reference Material for human plasma (SRM 1950): GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources. *Anal Chem* **2013,** *85* (24), 11725-31.

79. Vu, T.; Siemek, P.; Bhinderwala, F.; Xu, Y.; Powers, R., Evaluation of Multivariate Classification Models for Analyzing NMR Metabolomics Data. *J Proteome Res* **2019,** *18* (9), 3282-3294.

80. Worley, B.; Powers, R., Multivariate Analysis in Metabolomics. *Curr Metabolomics* **2013,** *1* (1), 92-107.

81. Mendez, K. M.; Reinke, S. N.; Broadhurst, D. I., A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics* **2019,** *15* (12), 150.

82. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H., Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem* **2006,** *78* (13), 4281-90.

83. Kohl, S. M.; Klein, M. S.; Hochrein, J.; Oefner, P. J.; Spang, R.; Gronwald, W., State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics* **2012,** *8* (Suppl 1), 146-160.

84. van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J., Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **2006,** *7*, 142.

85. Vinaixa, M.; Samino, S.; Saez, I.; Duran, J.; Guinovart, J. J.; Yanes, O., A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. *Metabolites* **2012,** *2* (4), 775-95.

86. Haynes, W., Bonferroni Correction. In *Encyclopedia of Systems Biology*, Dubitzky, W.; Wolkenhauer, O.; Cho, K.-H.; Yokota, H., Eds. Springer New York: New York, NY, 2013; pp 154-154.

87. Peluso, A.; Glen, R.; Ebbels, T. M. D., Multiple-testing correction in metabolome-wide association studies. *BMC Bioinformatics* **2021,** *22* (1), 67.

88. Bingol, K.; Li, D. W.; Zhang, B.; Bruschweiler, R., Comprehensive Metabolite Identification Strategy Using Multiple Two-Dimensional NMR Spectra of a Complex Mixture Implemented in the COLMARm Web Server. *Anal Chem* **2016,** *88* (24), 12411-12418.

89. Robinette, S. L.; Bruschweiler, R.; Schroeder, F. C.; Edison, A. S., NMR in metabolomics and natural products research: two sides of the same coin. *Acc Chem Res* **2012,** *45* (2), 288-97.

90. Lippa, K. A.; Aristizabal-Henao, J. J.; Beger, R. D.; Bowden, J. A.; Broeckling, C.; Beecher, C.; Davis, W. C.; Dunn, W. B.; Flores, R.; Goodacre, R.; Gouveia, G. J.; Harms, A. C.; Hartung, T.; Jones, C. M.; Lewis, M. R.; Ntai, I.; Percy, A. J.; Raftery, D.; Schock, T. B.; Sun, J.; Theodoridis, G.; Tayyari, F.; Torta, F.; Ulmer, C. Z.; Wilson, I.; Ubhi, B. K., Reference Materials for MS-based Untargeted Metabolomics and Lipidomics: A Review by the Metabolomics Quality Assurance and Quality Control Consortium (mQACC). **2021,** (soon to be released in Metabolomics), accepted

91. Mehta, K. Y.; Wu, H. J.; Menon, S. S.; Fallah, Y.; Zhong, X.; Rizk, N.; Unger, K.; Mapstone, M.; Fiandaca, M. S.; Federoff, H. J.; Cheema, A. K., Metabolomic

biomarkers of pancreatic cancer: a meta-analysis study. *Oncotarget* **2017,** *8* (40), 68899-68915.

92. Li, S.; Tian, Y.; Jiang, P.; Lin, Y.; Liu, X.; Yang, H., Recent advances in the application of metabolomics for food safety control and food quality analyses. *Crit Rev Food Sci Nutr* **2021,** *61* (9), 1448-1469.

93. Monge, M. E.; Dodds, J. N.; Baker, E. S.; Edison, A. S.; Fernandez, F. M., Challenges in Identifying the Dark Molecules of Life. *Annu Rev Anal Chem (Palo Alto Calif)* **2019,** *12* (1), 177-199.

94. R. G. Scholz, E. S. S. a. M. E. W., FEASIBILITY STUDY OF THE DEVELOPMENT OF A SPECIALIZED COMPUTER SYSTEM OF ORGANIC CHEMICAL SIGNATURES OF SPECTRAL DATA. **1968**.

95. Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; Berjanskii, M.; Mah, R.; Yamamoto, M.; Jovel, J.; Torres-Calzada, C.; Hiebert-Giesbrecht, M.; Lui, V. W.; Varshavi, D.; Varshavi, D.; Allen, D.; Arndt, D.; Khetarpal, N.; Sivakumaran, A.; Harford, K.; Sanford, S.; Yee, K.; Cao, X.; Budinski, Z.; Liigand, J.; Zhang, L.; Zheng, J.; Mandal, R.; Karu, N.; Dambrova, M.; Schioth, H. B.; Greiner, R.; Gautam, V., HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res* **2022,** *50* (D1), D622-D631.

96. Bingol, K.; Bruschweiler-Li, L.; Yu, C.; Somogyi, A.; Zhang, F.; Bruschweiler, R., Metabolomics beyond spectroscopic databases: a combined MS/NMR strategy for the rapid identification of new metabolites in complex mixtures. *Anal Chem* **2015,** *87* (7), 3864-70.

97. 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* (5), 1282-9.

98. 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 HgCl2 Toxicity. *Journal of Proteome Research* **2006,** *5* (6), 1313-1320.

99. Nagana Gowda, G. A.; Raftery, D., Can NMR solve some significant challenges in metabolomics? *J Magn Reson* **2015,** *260*, 144-60.

100. Stupp, G. S.; Clendinen, C. S.; Ajredini, R.; Szewc, M. A.; Garrett, T.; Menger, R. F.; Yost, R. A.; Beecher, C.; Edison, A. S., Isotopic ratio outlier analysis global metabolomics of Caenorhabditis elegans. *Anal Chem* **2013,** *85* (24), 11858-11865.

101. Ghosh, R.; Bu, G.; Nannenga, B. L.; Sumner, L. W., Recent Developments Toward Integrated Metabolomics Technologies (UHPLC-MS-SPE-NMR and MicroED) for Higher-Throughput Confident Metabolite Identifications. *Front Mol Biosci* **2021,** *8*, 720955.

102. Blazenovic, I.; Kind, T.; Ji, J.; Fiehn, O., Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **2018,** *8* (2).

103. Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J., Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **2014,** *48* (4), 2097-8.

104. Borchert, A. J.; Gouveia, G. J.; Edison, A. S.; Downs, D. M., Proton Nuclear Magnetic Resonance Metabolomics Corroborates Serine Hydroxymethyltransferase as the Primary Target of 2-Aminoacrylate in a ridA Mutant of Salmonella enterica. *mSystems* **2020,** *5* (2).

105. Edison, A. S.; Hall, R. D.; Junot, C.; Karp, P. D.; Kurland, I. J.; Mistrik, R.; Reed, L. K.; Saito, K.; Salek, R. M.; Steinbeck, C.; Sumner, L. W.; Viant, M. R., The Time Is Right to Focus on Model Organism Metabolomes. *Metabolites* **2016,** *6* (1).

106. Viant, M. R.; Kurland, I. J.; Jones, M. R.; Dunn, W. B., How close are we to complete annotation of metabolomes? *Curr Opin Chem Biol* **2017,** *36*, 64-69.

107. Reed, L. K.; Baer, C. F.; Edison, A. S., Considerations when choosing a genetic model organism for metabolomics studies. *Curr Opin Chem Biol* **2017,** *36*, 7-14.

CHAPTER 2

LONG-TERM METABOLOMICS REFERENCE MATERIAL[*]

[*] Gouveia, G. J.; Shaver, A. O.; Garcia, B. M.; Morse, A. M.; Andersen, E. C.; Edison, A. S.; McIntyre, L. M., Long-Term Metabolomics Reference Material. *Anal Chem* **2021,** *93* (26), 9193-9199.

FOREWORD

This chapter is reprinted from Gouveia, G. J.; Shaver, A. O.; Garcia, B. M.; Morse, A. M.; Andersen, E. C.; Edison, A. S.; McIntyre, L. M., Long-Term Metabolomics Reference Material. Anal Chem 2021, 93 (26), 9193-9199, and is available at https://pubs.acs.org/doi/10.1021/acs.analchem.1c01294. The following work was carried out as part of an NIH metabolomics Common fund project titled "Genetics and quantum chemistry as tools for unknown metabolite identification" (1U2CES030167-01). The motivation for this work was driven by the senior authors Arthur S. Edison and Lauren M. McIntyre. My contribution to this work consisted of: (i) generate *E. coli* and *C. elegans* as well as develop and optimize protocols to run and maintain the bioreactors, (ii) storage, handling and preparation of the produced material, (iii) NMR experimental design, (iv) bacteria sample preparation for NMR analysis, (v) generation of the *C. elegans* reference material (vi) all NMR data processing, (vii) all data analysis (viii) results interpretation and figures and writing the draft versions, and finally (ix) manuscript submission, addressing reviewers comments and resubmission. The collaborators roles were as follows: Arthur S. Edison and Lauren M. McIntyre reviewed, edited, responded to reviewers and defined the direction and goals of the work. Amanda O. Shaver carried out *C. elegans* NMR sample preparation and *C. elegans* NMR analysis. Brianna M. Garcia aided in the experimental design and result interpretation, Alison M. Morse wrote the SECIM analysis steps, Erik C. Andersen provided expertise, advice and guidance on growing bacterial and *C. elegans* cultures as well as final draft feedback. The supplementary materials in the chapter were added to Appendix A.

ABSTRACT

The use of quality control samples in metabolomics ensures data quality, reproducibility and comparability between studies, analytical platforms and laboratories. Long-term, stable and sustainable reference materials (RMs) are a critical component of the QA/QC system, however, the limited selection of currently available matrix matched RMs reduce their applicability for widespread use.

To produce a RM in any context, for any matrix that is robust to changes over the course of time we developed IBAT (Iterative Batch Averaging meThod). To illustrate this method, we generated 11 independently grown *E. coli* batches and made a RM over the course of 10 IBAT iterations. We measured the variance of these materials by NMR and showed that IBAT produces a stable and sustainable RM over time. This *E. coli* RM was then used as food source to produce a *Caenorhabditis elegans* RM for a metabolomics experiment. The metabolite extraction of this material, alongside 41 independently grown individual *C. elegans* samples of the same genotype, allowed us to estimate the proportion of sample variation in pre-analytical steps. From the NMR data, we found that 40% of the metabolite variance is due to the metabolite extraction process and analysis and 60% is due to sample-to-sample variance.

The availability of RMs in untargeted metabolomics is one of the predominant needs of the metabolomics community that reach beyond quality control practices. IBAT addresses this need by facilitating the production of biologically relevant RMs and increasing their widespread use.

INTRODUCTION

Biological reference materials are needed to compare metabolomics data across multiple instruments, studies and batches. Whenever there are more samples collected than can be processed in a single 'run' there is added unwanted variation that if captured can be modeled and removed, leading to more powerful tests.[1]

Readily available long-term biologically relevant reference materials (RMs) represent a critical component to achieve reproducibility.[2, 3] Commercially available RMs and standard reference materials (SRMs) address some of these needs, but can be expensive to purchase, offer limited quantities, matrix diversity, and have an expiration date.[3] The National Institute of Standards and Technology (NIST) has a long history of producing biofluid-based SRMs to facilitate standardization and improve comparability and reproducibility of analytical measurements. These SRMs are trademarked Certified Reference Materials (CRMs) and specifically designed to provide certified metabolite levels that serve strict objectives (i.e., calibration, method validation, measurement accuracy).[4-6]

Pooled quality control (QC) samples produced from experimental samples are valuable as they capture instrument variation within the experiment, but have limited value in comparing across experiments, or in synthesizing results from large experiments.[7, 8] The individual variation intrinsic in subjecting biological material to extraction and quantification is not captured by pooled samples or by chemical standards made after extraction. There is a recognized need for matrix specific stable RMs that can be used to compare data across long-term studies with multiple batches or across different laboratories and instrumention.[9]

Homogeneous and stable materials that are fit for purpose are reference materials (as per the International Vocabulary of Metrology-VIM).[10] RM does not require a metrologically valid metabolite quantification (certification) and should be straightforward to produce and maintain. For untargeted metabolomics, additional criteria for a RM are important. Namely, it should (i) be made from the same biological matrix as the experimental samples, (ii) have a profile that is as complex as the experimental samples, (iii) be sustainably produced over time and (iv) facilitate the annotation of known and unknown compounds.

The proteomics community devoted substantial effort to the development and application of RMs, which greatly improved standardization and reproducibility in the field.[4, 9] The metabolomics community has highlighted the need for RMs as part of the development of resources and practices to measure, detect and prevent unwanted pre-analytical and instrumental variation.[2, 3, 5, 8, 11]

Here we introduce IBAT (Iterative Batch Averaging meThod) that can be used to create a stable RM produced over time in any context. The concept is straightforward: multiple small batches of starting material are produced and aliquoted, and then pooled to generate the RM. A stable and long-lasting RM can be generated by repeating the process over time, as illustrated in Fig 2.1.

IBAT results in a RM that (i) is robust to changes over time, (ii) minimizes variance between batches of RM, (iii) can be used over the course of large-scale experiments, (iv) can be made with a small amount of constant effort and smaller storage space, (v) can be applied to any organism or biological matrix of interest and (vi) can be used for evaluation of multiple sources of variation at multiple points in a metabolomics experiment. To

illustrate IBAT, we made and characterized a *Caenorhabditis elegans* reference material. *C. elegans* eats bacteria, which is also subject to variation over time, so to make a stable *C. elegans* RM, we first needed to make an *Escherichia coli* RM that can be fed to *C. elegans*. This two-step IBAT shows the flexibility of the approach, and in the Discussion section we outline strategies to apply IBAT to create other RMs of interest to metabolomics researchers.

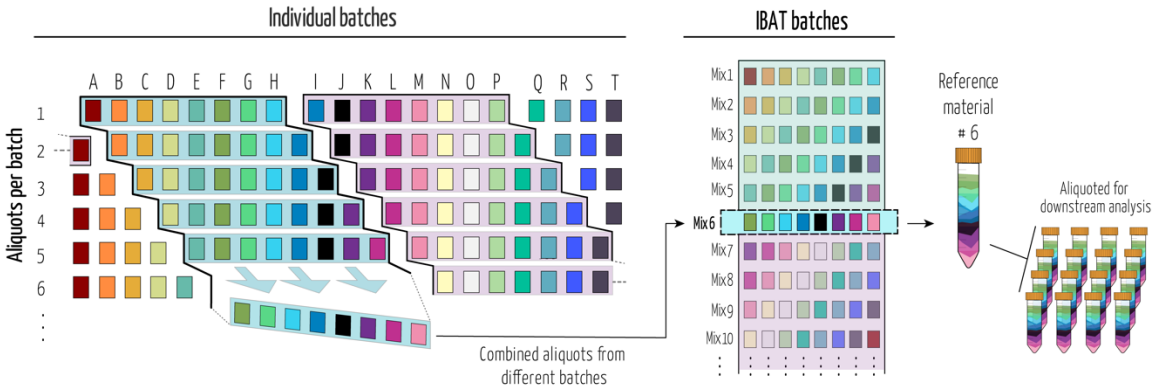


Figure 2.1: Iterative batch average method (IBAT). Batches of material are represented by columns (same-colored squares and letters). Rows represent homogeneous aliquots from each batch. Examples of sequential batch combinations are rows shaded from blue to purple. Right panel illustrates the IBAT generated pools from individual batches. IBAT is only limited by the number of individual batches produced and can be adjusted to the number of aliquots required and to any material.

RESULTS

Production and analysis of an IBAT E. coli as a food source for C. elegans

For this RM, we used a bioreactor to generate large quantities of bacteria in each batch, but the principle holds on a smaller scale with flasks and a shaker/incubator. We grew 11 different 2 L bioreactor batches (columns in Fig. 2.1) that each produced an average of 84 g of bacterial paste. Each batch was then aliquoted into 60-90 tubes (rows in Fig. 2.1) containing 1 g each, with mixing to maintain homogeneity of the aliquots.

We combined single aliquots from five different batches for this *E. coli* RM, such that each tube of IBAT RM contained the same amount of material. The first IBAT sample was made by combining batches A-E (columns in Fig. 2.1), the second IBAT sample combined batches B-F, etc. When we reached the end of the 11 batches (G-K with 11 batches), the next IBAT sample was made from H-K and an aliquot of A (Fig. 2.1). A similar IBAT process was applied to *C. elegans*, as described below. We compared the 10 different *E. coli* IBAT samples (Table 2.1) with individual replicates from all 11 batches. The 33 samples from 11 individual batches of *E. coli* and 30 IBAT samples were analyzed by nuclear magnetic resonance (NMR) spectroscopy.

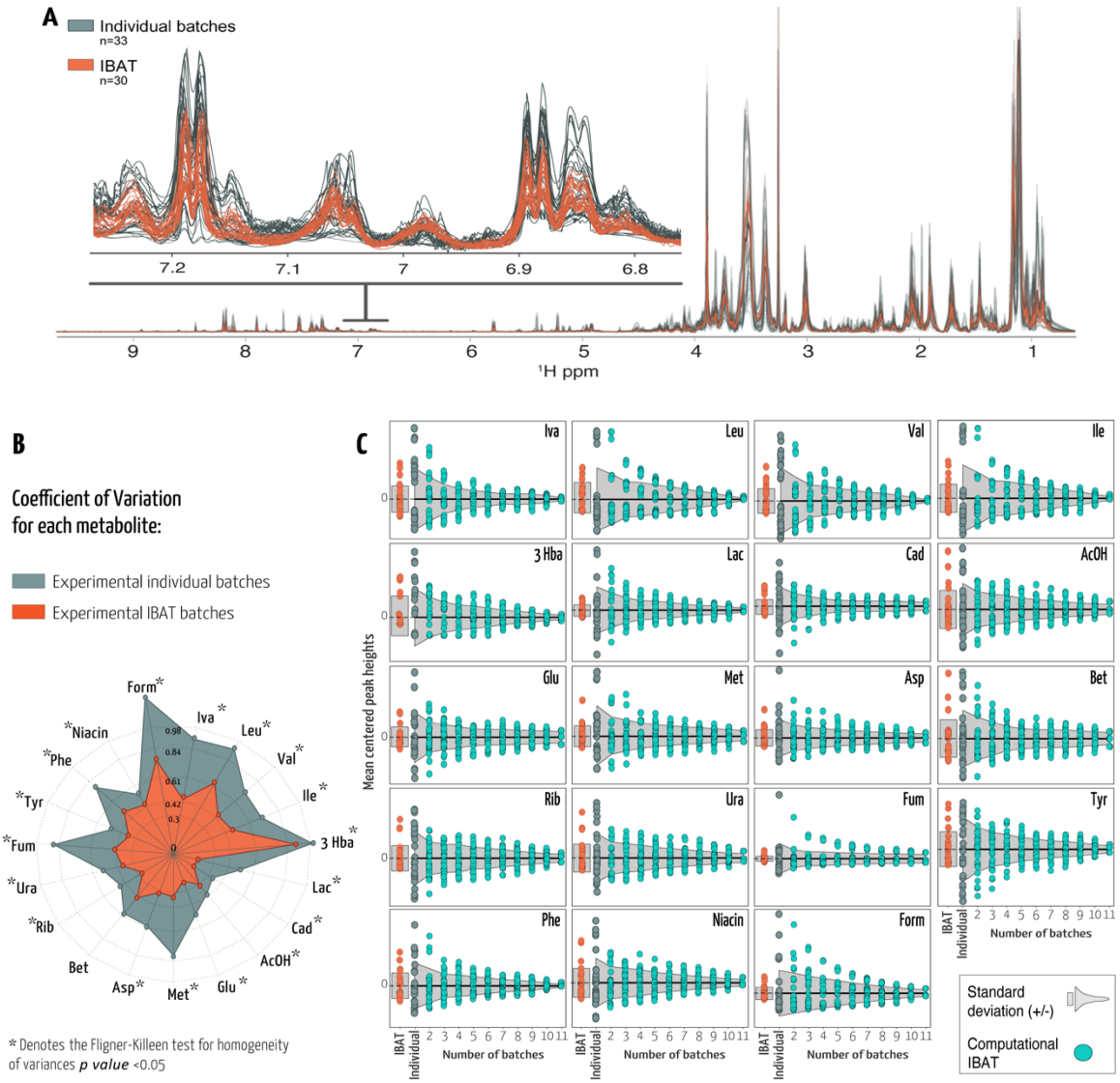


Figure 2.2: A) Untargeted full resolution ^{1}H NMR profile of *E. coli* and spectral expansion between 6.8 and 7.2 ppm. NMR spectra in grey or orange correspond to IBAT or individual batches, respectively. B) Radial plot representing the coefficient of variation (CV) for annotated metabolites using the same colors. The length of spokes corresponds to the CV of each metabolite. C) Each data point represents the mean centered peak height in each sample. Experimental IBAT samples are depicted in orange and individual batches in grey. Cyan data points represent simulated metabolite peak heights per number of averaged batches. Light grey shaded areas represent +/- one standard deviation from the mean. Iva – isovalerate, Leu – leucine, Val – valine, Ile– isoleucine, 3 Hba – 3-hydroxybutyrate, Lac – lactate, Cad – cadaverine, AcOH – acetate, Glu– glutamate, Met – methionine, Asp – aspartate, Bet – betaine, Rib – ribose, Ura – uracil, Fum – fumarate, Tyr – tyrosine, Phe – phenylalanine, Niacin – nicotinic acid and Form – formate.

Table 2.1: List of individual batches pooled together for each stable food source iteration. This process follows the same methodology described in fig 2.1.

IBAT iterations	1	2	3	4	5	6	7	8	9	10
Combined individual batches	A to E	B to F	C to G	D to H	E to I	F to J	G to K	H to A	I to B	J to C

The 33 samples from 11 individual batches of *E. coli* and 30 IBAT samples were analyzed by nuclear magnetic resonance (NMR) spectroscopy. The IBAT method reduces variance between different tubes of RM. The NMR spectra for these samples are nearly identical, with a very low variance (Fig. 2.2a). The variance here is due to extraction and quantification. In contrast, the variance between the 33 individual spectra is much larger, reflecting a combination of biological variance and technical variance.

To quantify variance, we selected 19 metabolites that we could identify, were present in all the samples, were consistent between replicate measurements, with clear individual peaks enabling accurate quantification of individual metabolites. The coefficient of variation (CV – standard deviation/mean) was calculated separately for each metabolite within each group (Fig. 2.2b, Supp. Table 2.1). Similar to the overlaid NMR spectra (Fig. 2.2a), the CV was lower for IBAT generated samples (between 0.19 and 0.91) than for individual samples (0.36 to 1.26). Using the Fligner-Killeen[12] test for homogeneity of variances for each of the selected metabolites showed significantly different variances between IBAT produced samples and individual batch samples (p value < 0.05) except for betaine (p value $= 0.21$). The IBAT process depends on pooling batches. We used the individual batch data to simulate the IBAT process. We generated 10 iterations for

combining 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11 individual batches to generate an IBAT compliant RM. We used the individual data to estimate the mean centered peak heights and respective standard deviations (Sd) for our 19 metabolites. The variance (10 iterations) decreases as the number of batches used increases (Fig. 2.2c). (Fig. 2.1 and Table. 2.1). This is consistent with the predictions of Spearman-Brown.[13, 14]

Production and application of a C. elegans PD1074 reference material:

To create an IBAT *C. elegans* RM, we used a 2 L bioreactor and fed the worms the *E. coli* RM. Each batch of the bioreactor produced between 40-60 million mixed-stage worms. These were harvested and aliquoted into 20-30 tubes so that every tube contained 2 million worms. These were then frozen at -80 ºC. After three bioreactor batches, we combined one aliquot from each batch for a total of 6 million batch-averaged worms. This was divided into 30 aliquots of *C. elegans* RM with 200 thousand worms each and refrozen until use (Fig. 2.3).

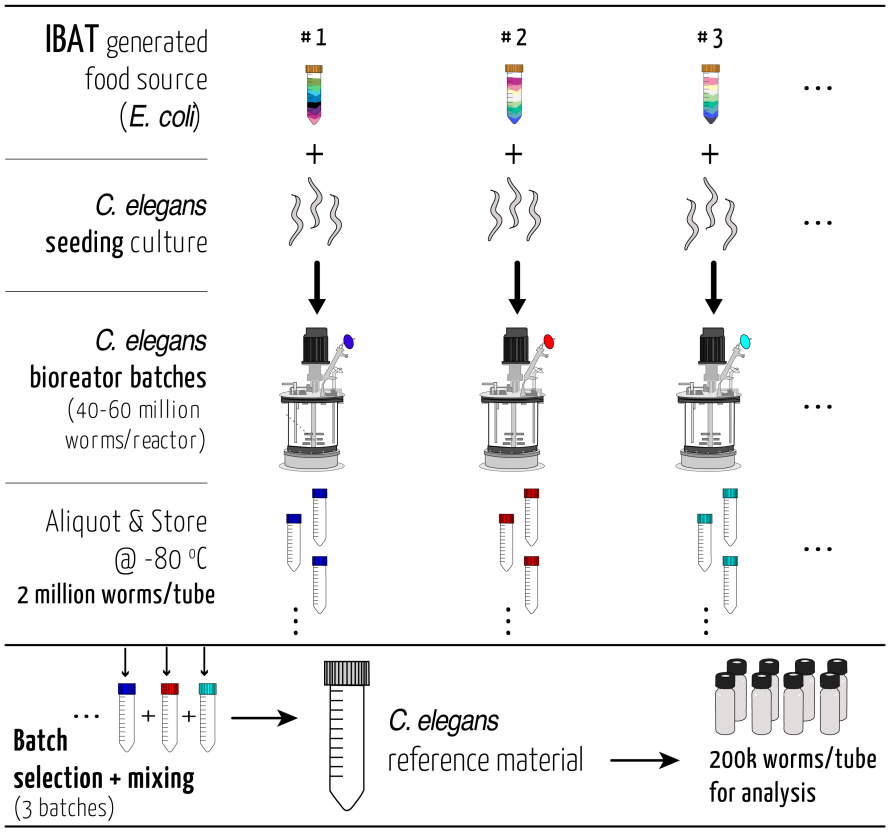


Figure 2.3: Schematic overview of the *C. elegans* reference material production. The reference strain PD1074 nematodes were seeded from cryo-preserved stocks and fed an *E. coli* RM (supplementary methods). Harvested material from each bioreactor was washed, aliquoted and stored. Aliquots from each reactor iteration were combined to produce a stable *C. elegans* reference material. This material can be divided into different sized aliquots according to the downstream application needs.

In a metabolomics experiment there are three main sources of variation: the sample material itself, the extraction, and data acquisition (supp. Fig 2.1). An experimental sample will encompass all three of those sources. The IBAT RM reduces the sample material variation, pooled samples average over both the sample variance and the extraction variation. We compared the *C. elegans* RM to 41 independent samples of the same strain (PD1074).[15] These individual samples were prepared in three sets of two extraction blocks. For each set an equimolar pool was formed from all individual samples, for three pools.

One *C. elegans* RM aliquot was included in each extraction block. In NMR data collection, one block was analyzed per each run. We selected 26 annotated features that were common to all samples and computed pairwise standardized Euclidean distances (SED) for each sample (Fig. 2.4). The distances between samples in the IBAT material reflect instrument variability (pools) and extraction variability. The distances between individual sample data include extraction and instrument variability but also sample variability.

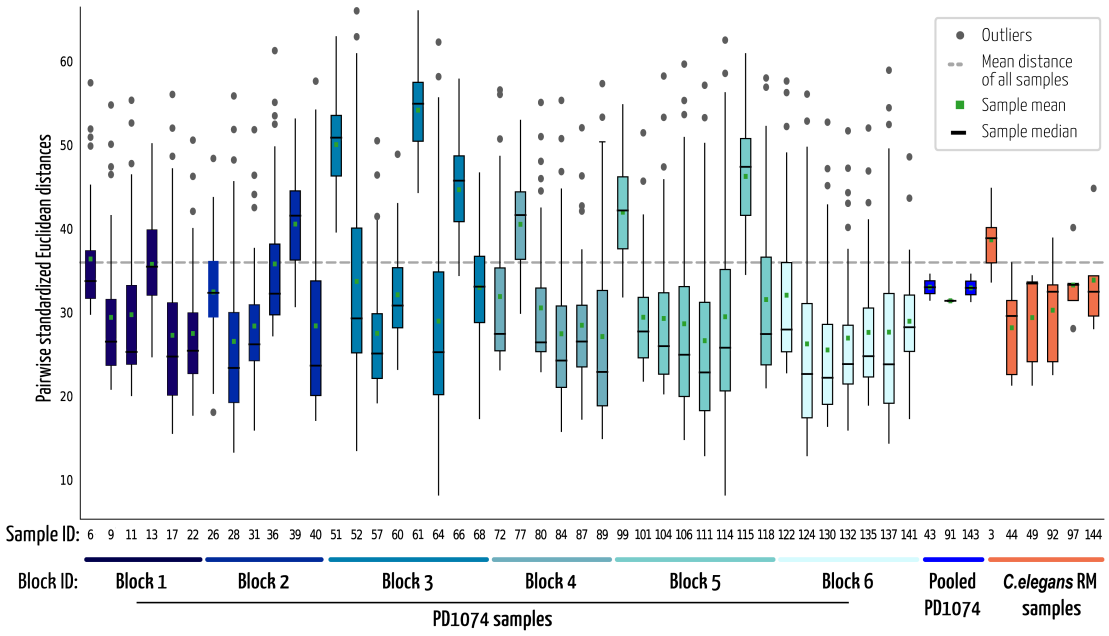


Figure 2.4: Boxplots of pair-wise standardized Euclidean distances. Each boxplot represents the distribution of distances from one sample to all the other samples of the same group. Mean and median distances for each sample are indicated by markers. Blue colored boxplots represent PD1074 samples that were processed in each block. The three pooled PD1074 samples were created from the samples in blocks 1+2, 3+4 and 5+6 respectively. *C. elegans* RM samples were generated using IBAT and processed alongside the PD1074 samples, one per each block.

The mean and median distances, minimum and maximum values and sample distribution for each of these groups allow us to estimate the variability from these different sources of variation. The individual PD1074 samples, which include all three sources of variation, have the largest variability with mean values from 25.5 to 54.1 and the min/max of 8.19 and 66 (blocks 1 through 6 in Fig. 2.4). The IBAT samples, representing the extraction and technical variance, have a smaller range of mean distances (28.2 to 38.7) and min/max values of 21.3 and 44.8. As expected, the pooled individual PD1074 samples representing differences in the manual preparation and instrumentation between sets, have the smallest range with the respective boxplot bounds between 31.4 and 33.

DISCUSSION AND CONCLUSION

The IBAT process reduces the growth and sampling contributions to variance by creating a common source of material from which homogeneous aliquots are produced. The advantage here is that instead of producing a single large batch, which will have its own challenges in achieving homogeneity, material is continuously generated over time, with each iteration using only small amounts of new material, thus capturing small changes over time while having minimal variance between experiments. This minimal variance can be theoretically predicted as a function of the number of distinct batches combined and the variance between the continuously produced material or, estimated from empirical data (Fig. 2.2c) to take into account the overlap between iterations. The IBAT process is flexible and can be adjusted to production throughput, the type of material, the quantities produced the degree of variance reduction, and the metabolomics technology.

We demonstrated this concept for two different types of matrices, *E. coli* and *C. elegans*. However, the method is general and can be applied to any biological matrix. In non-model systems studies it is common to use human plasma or urine or commercially available materials that are aliquoted from a single large batch and frozen. But when a batch runs out, shifting to a new external standard will often not be comparable to the prior standard. IBAT can be used by making pools from different batches of material as illustrated in Figure 2.1. New batches can be incorporated over time, and this will minimize the change in the RM over time. Similar strategies can be used with diverse applications such as plants or cultured mammalian cells for biotherapeutics. In these scenarios the main issue is minimizing the freeze thaw cycles and so, the size of the initial aliquots for future blending must be planned.

A RM of the same biological matrix as the study samples together with a carefully planned experimental design can be used to determine the magnitude and variance in the extraction, a major source of variation in metabolomics experiments.[3, 16] It can also facilitate comparison among separate experiments. The IBAT can then be used to separate the extraction variance from the sample-to-sample variance in the individually grown and processed samples, as demonstrated here. The individual *C. elegans* samples are genetically identical to the RM. Variance in metabolite intensities were larger as a result of sample variation during growth, handling, storage and sampling, added to the technical variation in extraction and data acquisition. The pooled individual *C. elegans* PD1074 samples minimizes the sample and extraction variance by averaging over both samples and extractions and reflects only the variation in the analytical measurement (which is low for NMR) and the pooling strategy. By processing the experimental replicates and RM

aliquots, one can independently estimate the contribution of the metabolite extraction step to individual metabolite variation. The IBAT *C. elegans* RM samples can be used to estimate variance due to extraction. We find that 40% of the total variance as estimated by the variation between individually grown, extracted and quantified samples is due to extraction variance and analysis and of that variance ~15% is due to technical variation.

IBAT increases the efficacy of QA/QC and is expected to improve the performance of biological reference materials by allowing estimation of process-derived variance including facilitating studies across multiple labs. Finally, the cost of using an IBAT process should be lower than acquiring a single large batch of reference material thus enabling labs to amortize the process over time while maintaining the stability of the material and facilitating comparison of experiments conducted months or years apart.

METHODS

E. coli individual batches production and storage:

In order to produce a stable and consistent *C. elegans* food source, batches of *E. coli* HT115 were grown in bioreactors (Biostat, Sartorius) using standardized protocols (Supplementary methods). A total of 11 batches were produced and each batch was divided into approximately 60-90 aliquots, flash frozen and stored at -80 ^{o}C. Each aliquot was comprised of 2 mL of bacterial suspension (1 g of wet bacterial paste and OD600 ranging from 17.5 to 24).

NMR sample preparation of E. coli IBAT and individual batches:

All 33 individual batch samples and 30 IBAT generated samples were prepared for NMR analysis. Approximately 200 µL of 0.7 mm silica beads (BioSpec products) were added to each of the 63 samples. These were homogenized at 1800 rpm for 300s (FastPrep 96 - MPBIO) and centrifuged at 20,000 x G for 15 minutes. From each sample, 450 µL of supernatant were transferred to a new tube and 150 µL of deuterated water were added (D_2O, D, 99.9%, Cambridge Isotope Laboratories). Each sample was vortex-mixed for 1 min before transferring into 5 mm SampleJet NMR tubes. Details of NMR acquisition and spectra processing can be found in the supplementary methods.

NMR sample preparation of C. elegans samples:

For the NMR analysis six IBAT RM aliquots were prepared for alongside 41 individual samples of the *C. elegans* strain PD1074 that were grown according to our previously published method.[15] Each of these samples contained approximately 200,000 nematodes. All samples were previously flash frozen and then lyophilized until dry. Approximately 200 µL of 1 mm Zirconia beads (BioSpec products) were added to each dry sample and homogenized at 1800 rpm for a total of 270 seconds (FastPrep 96 - MPBIO). The samples were then delipidated by adding 1 mL of cold (-20 ºC) isopropanol (Optima, LC/MS Grade, Fisher Scientific) and left overnight (12 hours) at -20 ºC after a 20 min resting period at room temperature. The supernatant was removed after being centrifuged for 30 min at 20,000 x G and 1 mL of cold (4 ºC) 80/20 methanol/water (Optima, LC/MS Grade, Fisher Scientific) was added to the remaining contents. The tubes were shaken for 30 min at 4 ºC and centrifuged at 20,000 x G for 30 minutes. The methanol/water

supernatant was transferred to new tubes and these were vacuum dried using a CentriVac benchtop vacuum concentrator (Labconco). The extracts were reconstituted in 45 µl of deuterated (D_2O, D, 99.9%, Cambridge Isotope Laboratories) 100 mM sodium phosphate buffer (mono- and dibasic; Fisher BioReagents) containing 0.11 mM of the internal standard DSS (sodium 2,2-dimethyl-2-silapentane-5-sulfonate, D6, 98%; Cambridge Isotope Laboratories) at pH 7.0 and vortex mixed for <1 min prior to transfer into 1.7 mm SampleJet NMR tubes. The three pooled PD1074 samples were created by adding together 6 µl from the samples in each NMR run (12, 14 and 15 samples respectively), after having been reconstituted in the internal standard containing NMR solvent. Details of NMR acquisition and spectra processing can be found in the supplementary methods.

Data analysis:

Following acquisition and processing, spectra were imported into Matlab programing software (MATLAB, MathWorks, R2019a). Using a toolbox developed in-house and available at (https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA) the following was carried out: plotting, referencing, baseline correction, alignment (CCOW[17]) and solvent peaks removal. Feature detection (peak picking) was automated using a combination of an in-house peak picking function and binning algorithm[18] to extract peak heights. Data were exported for Bland-Altman analysis, to select features that are in agreement between its replicates (cut-offs used: sample flag of 0.2, feature flag of 0.05 and residual of 3), and pairwise Standardized Euclidean Distances (SED) analysis using the SouthEast Center for Integrated Metabolomics Tools

(SECIMTools)[19]. Coefficient of variation calculations (CV), variance, %variance and Fligner-Killeen test were carried out in Matlab.

Data availability

All raw and processed data, along with detailed experimental NMR and data analysis methods, are available under project identifier PR001106 at the Metabolomics Workbench (www.metabolomicsworkbench.org). The data can be accessed directly via its project doi: 10.21228/M8R395. Metabolomics Workbench is supported by NIH grant U2C-DK119886.

Supporting information

Additional methods detailing the bioreactor production of both *C. elegans* and *E. coli*, making an *E. coli* reference material, NMR acquisition and processing, database matching procedures, metabolite summary table and a figure detailing the sources of variance.

Acknowledgement

Research reported in this publication was supported by the National Institutes of Health under Award Number 1U2CES030167-01. The authors would like to thank Dr. David Blum and Ron Garrison from the Bio-expression and Fermentation Facility at the University of Georgia for training and advice using the Bioreactors and Pamela Kirby at the Edison Lab for assistance with material handling and storage logistics.

REFERENCES

1. Cochran, W. G. a. C., G.M. , *Experimental Design*. 2nd edition ed.; New York, 1957.

2. Dunn, W. B.; Broadhurst, D. I.; Edison, A.; Guillou, C.; Viant, M. R.; Bearden, D. W.; Beger, R. D., Quality assurance and quality control processes: summary of a metabolomics community questionnaire. *Metabolomics* **2017,** *13* (5).

3. Broadhurst, D.; Goodacre, R.; Reinke, S. N.; Kuligowski, J.; Wilson, I. D.; Lewis, M. R.; Dunn, W. B., Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* **2018,** *14* (6), 72.

4. Paulovich, A. G.; Billheimer, D.; Ham, A. J.; Vega-Montoto, L.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Clauser, K. R.; Kinsinger, C. R.; Schilling, B.; Tegeler, T. J.; Variyath, A. M.; Wang, M.; Whiteaker, J. R.; Zimmerman, L. J.; Fenyo, D.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Mesri, M.; Neubert, T. A.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Stein, S. E.; Tempst, P.; Liebler, D. C., Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol Cell Proteomics* **2010,** *9* (2), 242-54.

5. Phinney, K. W.; Ballihaut, G.; Bedner, M.; Benford, B. S.; Camara, J. E.; Christopher, S. J.; Davis, W. C.; Dodder, N. G.; Eppe, G.; Lang, B. E.; Long, S. E.; Lowenthal, M. S.; McGaw, E. A.; Murphy, K. E.; Nelson, B. C.; Prendergast, J. L.; Reiner, J. L.; Rimmer, C. A.; Sander, L. C.; Schantz, M. M.; Sharpless, K. E.; Sniegoski, L. T.; Tai, S. S.; Thomas, J. B.; Vetter, T. W.; Welch, M. J.; Wise, S. A.; Wood, L. J.; Guthrie, W. F.; Hagwood, C. R.; Leigh, S. D.; Yen, J. H.; Zhang, N. F.; Chaudhary-Webb, M.; Chen, H.; Fazili, Z.; LaVoie, D. J.; McCoy, L. F.; Momin, S. S.; Paladugula, N.; Pendergrast, E. C.; Pfeiffer, C. M.; Powers, C. D.; Rabinowitz, D.; Rybak, M. E.; Schleicher, R. L.; Toombs, B. M.; Xu, M.; Zhang, M.; Castle, A. L., Development of a Standard Reference Material for metabolomics research. *Anal Chem* **2013,** *85* (24), 11732-8.

6. Simon-Manso, Y.; Lowenthal, M. S.; Kilpatrick, L. E.; Sampson, M. L.; Telu, K. H.; Rudnick, P. A.; Mallard, W. G.; Bearden, D. W.; Schock, T. B.; Tchekhovskoi, D. V.; Blonder, N.; Yan, X.; Liang, Y.; Zheng, Y.; Wallace, W. E.; Neta, P.; Phinney, K. W.; Remaley, A. T.; Stein, S. E., Metabolite profiling of a NIST Standard Reference Material for human plasma (SRM 1950): GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources. *Anal Chem* **2013,** *85* (24), 11725-31.

7. Han, W.; Li, L., Evaluating and minimizing batch effects in metabolomics. *Mass Spectrom Rev* **2020**.

8. Peng, J.; Chen, Y. T.; Chen, C. L.; Li, L., Development of a universal metabolome-standard method for long-term LC-MS metabolome profiling and its application for bladder cancer urine-metabolite-biomarker discovery. *Anal Chem* **2014,** *86* (13), 6540-7.

9. Bunk, D. M., Design considerations for proteomic reference materials. *Proteomics* **2010,** *10* (23), 4220-5.

10. Metrology, J. C. f. G. i., International vocabulary of metrology - Basic and general concepts and associated terms. *VIM* **2012,** (200).

11. Beger, R. D.; Dunn, W. B.; Bandukwala, A.; Bethan, B.; Broadhurst, D.; Clish, C. B.; Dasari, S.; Derr, L.; Evans, A.; Fischer, S.; Flynn, T.; Hartung, T.; Herrington, D.; Higashi, R.; Hsu, P. C.; Jones, C.; Kachman, M.; Karuso, H.; Kruppa, G.; Lippa, K.; Maruvada, P.; Mosley, J.; Ntai, I.; O'Donovan, C.; Playdon, M.; Raftery, D.; Shaughnessy, D.; Souza, A.; Spaeder, T.; Spalholz, B.; Tayyari, F.; Ubhi, B.; Verma, M.; Walk, T.; Wilson, I.; Witkin, K.; Bearden, D. W.; Zanetti, K. A., Towards quality assurance and quality control in untargeted metabolomics studies. *Metabolomics* **2019,** *15* (1), 4.

12. Conover, W. J.; Johnson, M. E.; Johnson, M. M., A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data. *Technometrics* **1981,** *23* (4), 351-361.

13. Brown, W., Some Experimental Results in the Correlation of Mental Abilities1. *British Journal of Psychology, 1904-1920* **1910,** *3* (3), 296-322.

14. Spearman, C., Correlation Calculated from Faulty Data. *British Journal of Psychology, 1904-1920* **1910,** *3* (3), 271-295.

15. Amanda O. Shaver, G. J. G., Pamela S. Kirby, Erik Andersen, Arthur S. Edison, Culture and assay of Large-Scale Mixed Stage Caenorhabditis elegans Population. *JOVE - J. Vis. Exp* **2020,** *e61453*.

16. Liu, Q.; Walker, D.; Uppal, K.; Liu, Z.; Ma, C.; Tran, V.; Li, S.; Jones, D. P.; Yu, T., Addressing the batch effect issue for LC/MS metabolomics data in data preprocessing. *Sci Rep* **2020,** *10* (1), 13856.

17. Tomasi, G.; van den Berg, F.; Andersson, C., Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics* **2004,** *18* (5), 231-241.

18. Sousa, S. A. A.; Magalhães, A.; Ferreira, M. M. C., Optimized bucketing for NMR spectra: Three case studies. *Chemometrics and Intelligent Laboratory Systems* **2013,** *122*, 93-102.

19. Kirpich, A. S.; Ibarra, M.; Moskalenko, O.; Fear, J. M.; Gerken, J.; Mi, X.; Ashrafi, A.; Morse, A. M.; McIntyre, L. M., SECIMTools: a suite of metabolomics data analysis tools. *BMC Bioinformatics* **2018,** *19* (1), 151.

CHAPTER 3

A METABOLITE FRACTION LIBRARY APPROACH FOR IMPROVED ANNOTATION IN UNTARGETED METABOLOMICS ACROSS ANALYTICAL PLATFORMS[*]

[*] Gouveia, G.J., Garcia, B.M., Asef, C.K., Shaver, A.O., Borges, R.M., Leach III, F.E., Fernández, F.M., Amster, I.J, Edison, A.S. Increasing confidence: Building a Fraction Library for Untargeted Metabolomics. To be submitted to *Analytical Chemistry*

FOREWORD

This chapter details the collaborative work of Goncalo J. Gouveia (GJG) and Brianna M. Garcia (BMG). The motivation for this work was driven by the senior author Arthur S. Edison. This work will be published, with Goncalo J. Gouveia and Brianna M. Garcia as joint co-first authors. All the non-overlapping steps detailed below that were carried out by either BMG or GJG, were all part of cross platform training and skill development. My contribution to this work as a co-first author consisted of: (i) experimental design aided, (ii) carry out metabolite extraction, (iii) design, test and validate the HPLC fractionation, (iv) prepare the fractions for further analysis, (v) analyze and process all the NMR data (vi) metabolite identification by NMR, (vii) NMR fraction data analysis and integration with LC-MS data (viii) generating figures and writing manuscript draft. BMG contribution as a co-first author consisted of: (i) validate and develop the sample extraction protocol, (ii) advise and define HPLC methods for fractionation, (iii) prepare the fractions for LC-MS analysis, (iv) analyze and process all LC-MS data, (v) metabolite identification by LC-MS, (vi) LC-MS fraction data analysis and integration with NMR data (vii) generating figures and writing manuscript draft. The senior authors roles were as follows: Arthur S. Edison, Franklin E. Leach III, I. Jonathan Amster reviewed, edited, and defined the direction and goals of the work. Carter K. Asef provided the metabolomics study data to relate the fractionated data to, Amanda O. Shaver generated the individual PD1074 samples that were used to make the RM, Facundo M. Fernández and Ricardo M. Borges provided advice and intellectual contributions to the analysis. Research reported in this manuscript was supported by the National Institutes of Health Award Number U2CES030167. Supplementary materials can be found in APPENDIX B.

ABSTRACT

Fractionation methods to purify and concentrate extracts containing a compound of interest have been widely used in natural products research. However, applying this approach to untargeted metabolomics is still a considerable challenge, in part, due to the number of metabolites under investigation and limited sample availability. Nonetheless, fractionation of metabolomics samples has the potential to improve compound annotation by reducing spectral overlap and concentrating metabolites. Further, it can bridge the sensitivity gap between nuclear magnetic resonance (NMR) spectroscopy and liquid chromatography-tandem mass spectrometry (LC-MS/MS), two analytical platforms often difficult to integrate, and yet essential for the annotation of metabolites. Herein, a semi-preparative fractionation approach is used on a Caenorhabditis elegans reference material (RM) to generate 100 concentrated 30-second fractions. Fractions were split approximately 100:1 between NMR and LC-MS/MS, respectively, and compared with spectral databases (SIRIUS, GNPS, and COLMARm). Putative annotations with the same InChIKey and fraction numbers were merged into a combined table containing their respective similarity and matching scores. Five fractions were analyzed as proof of concept. Seven annotations were matched (score >0.7) across four different database query methods (SIRIUS, GNPS, HSQC, and TOCSY). An additional 14 putative annotations were obtained when the search space was expanded to annotations with similarities scores < 0.7. A total of 31 (NMR) and 49 (LC-MS/MS) putative annotations were shown to be platform-dependent with no overlap between technologies. This approach demonstrates that spectral matching has limitations and is highly platform and metabolite dependent. However, the creation of a metabolite fraction library allows for multiple complementary analytical measurements on

a simplified fraction of a complex matrix, increasing key experimental outcomes such as S/N, database coverage, and annotation confidence.

INTRODUCTION

Metabolomics is the investigation of the structure and activity of biogenic small molecules, or metabolites, and attempts to relate changes in their concentrations to specific phenotypes or disease states. The measurement of metabolites provides fundamental insights into biochemical pathways; however, it requires the confident identification and quantification of metabolites. Mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy are the two most powerful analytical methods used in metabolomics research. Both platforms can detect individual molecular species within a complex biological matrix while providing unique and complementary chemical structure information. To date, metabolomics has primarily relied on the separate application of LC-MS/MS and NMR. The combination of these platforms has been shown to increase overall metabolome coverage, increase the accuracy of metabolite annotation, and provide validation of metabolite changes.[1-4] However, for either analytical technology used, detection alone does not always lead to the unambiguous identification of metabolites.[5]

Metabolite annotation and identification by LC-MS/MS has widely been accomplished by *de novo* identification or database searching, where fragments in a tandem mass spectrometry dataset (MS^2 spectra) are assigned to a structure or matched against standard spectral libraries. However, the large-scale annotation of metabolites presents additional challenges. Electrospray ionization (ESI), one of the most common MS interfaces employed for metabolomics, results in the generation of multiple signals that can

be attributed to a single molecule of interest. This ionization method generates protonated, deprotonated, and neutral loss ions, as well adducts (e.g., sodium, potassium, ammonium, etc.) and oligomeric or fragment species. Additionally, the presence of artifacts and contaminants further necessitates methods to distinguish biologically relevant and artifactual features.[6, 7] As a result, these mass spectra are complex and a considerable challenge to efficiently resolve and identify.

Accurate mass measurement of a compound permits the determination of its chemical formula, but the number of molecular structures with the same formula (isomers) grows with the molecular weight.[8] To distinguish between isomeric species, tandem MS data collection is required where the fragment masses are used as fingerprints for the identification of specific structures. However, metabolites similar in structure or molecules producing a low number of fragment ions can be difficult or impossible to distinguish by LC-MS/MS alone.[9]

These fingerprints are also used to compare against reference data hosted in mass spectral databases (e.g., HMDB,[10] NIST,[11] METLIN,[12] MassBank,[13] etc.). Additionally, both MS^1 and MS^2 data can be used to create molecular networks. These networks expand beyond database matching by connecting related molecules by their spectral similarity (GNPS-feature based molecular networking[14]) thereby increasing annotation confidence and providing new avenues to elucidate unknown structures. Database interfaces and software (GNPS,[15] SIRIUS,[16] etc.) use different metrics (cosine, Jaccard/Tanimoto,[17, 18] etc.) to quantify the similarity between experimental spectra and database reference spectra. This allows for a fast putative annotation of a large number of MS^2 spectra and their corresponding MS^1 features. However, both experimental and known compound

fragmentation patterns stored in databases can be collected using a wide range of different fragmentation methods, detector types, collision energies and acquisition parameters resulting in incorrect or poor metabolite annotations.

Metabolite annotation by NMR can be accomplished by pattern matching, similar to MS, or by determining the resonances (or groups of resonances) that belong to an individual compound. These resonances or spectral fingerprints can then be used to query against one or several NMR spectral databases (e.g., BMRB,[19] HMDB,[10] NMRShiftDB,[20] etc.). The latter approach heavily depends on the quality of the spectral deconvolution often requiring additional NMR experiments.[5] In contrast, spectral peak pattern matching, relies solely on peak-picked spectra. The resulting chemical shifts are used to search for peak patterns that match reference databases peak lists. [21] COLMARm, a free online resource uses HSQC (heteronuclear single quantum coherence) and TOCSY (total correlation spectroscopy) spectra to query several NMR spectral databases.[22]

Unlike MS database similarity matching, COLMAR uses three scoring criteria for metabolite identification: (i) a matching ratio from 0 to 1 represents the fraction of experimental features that match the available reference, (ii) the average difference between ^{1}H and ^{13}C chemical shifts of experimental and database, and (iii) a uniqueness score, represented as a fraction of the number of unique compound-specific HSQC features (that do not overlap with other database hits) over the number of features that are matched.

The accuracy and reliability of the database matching relies on the peak picking method used[23] for the reference and experimental spectra as well as a variety of factors including: the correct chemical shift annotation of the reference compound, the presence

of impurities and artifacts, the pH of the sample, the signal-to-noise ratio of the experimental spectra, and its respective resolution and complexity.[21, 24]

Regardless of the analytical platform used, the confident annotation of metabolites continues to be a major challenge in the field. Nonetheless, the growth of publicly available databases has greatly improved metabolite annotation, despite low level metabolites, biologically modified xenobiotics, and chemically complex metabolites still being underrepresented and limited by the insufficient number of available chemical standards. Currently, the Human Metabolome Database (HMDB) contains over 250k metabolite entries from a wide variety of biospecimens, however only <5% of these metabolites have been experimentally collected.[10]

The expansion of these databases is time- and labor-intensive and the number of submitted reference spectra has been declining in recent years.[10] Thus, in silico MS (e.g., HMDB, CFM-ID,[25] CSI:FingerID,[26] MS-FINDER,[27] etc.) and NMR (HMDB, NMRShiftDB[20], etc.)[28, 29] computational methods have been developed from curated literature searches and the ~110 million chemical structures in public chemical databases (e.g., PubChem, ChemSpider, etc.) to predict ESI MS^2 spectra as well as ^1H and/or ^{13}C chemical shifts. Generating MS^2 and/or NMR in silico predictions for structures can come at a large computational cost. While the quality of in silico generated data is improving, these calculations follow a normal distribution resulting in a small number of low and high-quality predictions, and a large number of average accuracy predictions.[30,8] Additional research is still needed to improve accuracy and rearrangement reactions that remain underestimated. However, the addition of in silico predictions to a metabolite annotation workflow can significantly reduce the chemical space gap present in current experimental

databases such as enzymatic transformations, exogenous compounds, or chemical artifacts.[27,28]

Similar to metabolomics, natural products chemistry (NP) strives to identify molecules of interest from complex biological samples. Despite this commonality, NP routinely uses a biological activity of interest (e.g., antibacterial, anti-inflammatory, etc.) to narrow down the number of molecules of interest. This approach requires large amounts of starting material that are separated by chromatography and collected at different time intervals. Depending on their activity, the collected fractions are subsequently isolated and/or purified for analytical characterization by NMR, MS, X-ray crystallography, etc. [31] As such, this fractionation method is challenging for metabolomics owing to (i) the amount of effort necessary (time and resources) for the identification of a large number of features, and (ii) the overarching goal of untargeted metabolomic studies precludes the use of bioactivity guided or targeted separation. Furthermore, metabolomics studies heavily rely on statistical analysis favoring more replicates over larger quantities and are difficult to translate to preceding or succeeding studies, as such, additional purification and characterization for every study would be cost prohibitive.

Without large quantities of the starting material, these methods typically result in a decrease in metabolite concentration posing issues for metabolite identification and limited by the instrument's sensitivity. Methodologies to overcome the described challenges associated with the fractionation of metabolomics samples have the potential to greatly improve compound annotation.

Here, we detail an approach to create a metabolite fraction library (mFL) of complementary analytical spectra (LC-MS/MS and NMR) collected from simplified and

concentrated fractions. These fractions were generated from a semi-preparative scale HPLC separation of a metabolite rich *C. elegans* reference material (RM) that was also used as a QC sample in an ongoing metabolomics study. The method described herein showed reduced spectral overlap for NMR spectroscopy and increased metabolite concentrations. The latter reduced the sensitivity gap between NMR and LC-MS/MS, a common impediment to the integration of these technologies.[32] This library consists of both analytical data and archived aliquots of each fraction that will serve as a resource for future experiments containing the same RM. The analytical data collected facilitates fast and confident metabolite annotation, while archived aliquots can be used for additional analytical measurements (*i.e.,* ion mobility and Fourier transform ion cyclotron resonance mass spectrometry). This multistep approach accelerates the systematic annotation of the *C. elegans* metabolome and addresses a considerable and longstanding bottleneck in the field of metabolomics.

Metabolite Fraction Library (mFL) Workflow

The steps carried out to make and analyze the mFL are illustrated in Fig 3.1. An in-depth description of the *C. elegans* reference extraction, semi-preparative fractionation, fraction concentration, data acquisition, processing, and analysis, and database matching can be found in the Supplemental Methods.

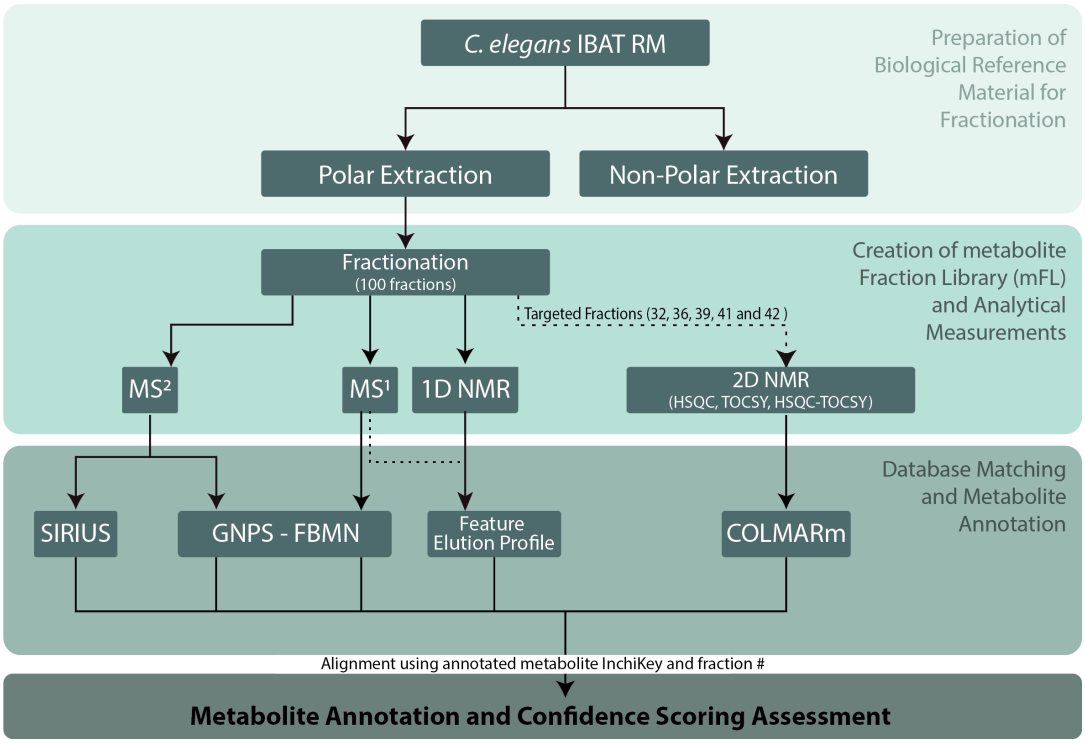


Figure 3.1: Workflow for the pipeline starting from the preparation of the *C. elegans* IBAT RM through to metabolite annotation by LC-MS/MS and NMR (1D and 2D) and ending with the assessment of annotations made through various database matching software effectively bridging LC-MS/MS and NMR for metabolite annotation with increased confidence.

C. elegans reference material preparation

To generate the RM used for fractionation 10 vials, each containing approximately 200,000 frozen and lyophilized nematodes (grown and stored using the LSPC method[33]), were selected from different batches according to the iterative batch averaging method (IBAT).[34]

Metabolite extraction

The RM was homogenized by bead beating under low temperatures. A fine homogenate was then extracted with 100% isopropanol (IPA) (chilled to -20°C) and left at room temperature for 30 minutes prior to a 12h extraction period at -20°C. The supernatant was transferred to a single tube, labelled "non-polar extract" and lyophilized in a Centrivap (Labconco) at room temperature until completely dry. The dry non-polar supernatant was stored at -80°C, and not included for analysis. The remaining pellet was mixed with 80:20 methanol:water (MeOH:H_2O) for polar analyte extraction. This polar fraction was vortex mixed for 30 minutes at low temperatures and the resulting supernatant was transferred to a single tube dried and stored at -80°C.

Semi-preparative HPLC fractionation

The polar extract was reconstituted in a total of 800 µL of 50/50 MeOH/H_2O. A total of six 90µL sample injections were separated using an Agilent 1260 infinity with an XBridge BEH Amide OBD Prep Column, 130Å, 5 µm, 10 mm X 250 mm HILIC column. The eluents were collected at 30 sec interval for a total of 100 fractions using a MeOH/H_2O solvent system. After fractionation, four 100 µL aliquots from each fraction were

transferred into LC-MS vials. These vials and the remaining fraction tubes containing 10 mL each were placed in a Centrivap at room temperature until dry and stored at -80°C. In addition, a standard AA mixture (Sigma AA-S-18) was fractionated and analyzed under the same conditions.

NMR data collection

HPLC polar eluents were reconstituted in 70 μL of deuterated water, containing 0.11 mM DSS (sodium 2,2-dimethyl-2-silapentane-5-sulfonate) as an internal standard and transferred into 1.7 mm NMR tubes (SampleJet, Bruker). These were then loaded onto a SampleJet automated sample changer and kept at 6°C. One-dimensional (1D) ^{1}H NMR spectra for each fraction and blanks were collected on a Bruker NEO 800MHz equipped with a 1.7mm TCI cryoprobe. Two-dimensional (2D) ^{1}H-^{1}H TOCSY, ^{1}H-^{13}C HSQC experiments were collected on select fractions (f32, f36, f39, f41, f42, f51, f55, f57) on a Bruker advance 600MHz with a TCI cryoprobe.

NMR database matching and 2D spectral annotation

HSQC and TOCSY experiments were separately matched to databases using the "HSQC query" and "TOCSY query" methods in COLMAR for the above listed fractions. HSQC, TOCSY and HSQC-TOCSY experiments from the unfractionated sample were used for spectral matching using COLMARm[22]. Matching chemical shift cutoffs of 0.04 and 0.3 ppm were used for ^{1}H and ^{13}C, respectively for all methods and queries. All data were processed using NMRPipe. [35] Manual spectral annotation and visualization was carried out using a combination of MNova (version 14.2.0) and NMR View J[36].

LC-MS/MS data collection

All 100 dry HPLC polar fraction aliquots for LC-MS/MS were reconstituted in 200 µL 80/20 ACN/H_2O and separated using a Vanquish UHPLC (ThermoFisher Scientific) equipped with a Waters Acquity UPLC BEH Amide column (2.1 x 150 mm, 1.7 µm particle size). A solvent system and gradient matching that used for the HPLC fractionation was used for separation. Gradient details and MS parameters and settings can be found in the Supplemental Methods.

LC-MS/MS database matching

LC-MS/MS data was processed using SLAW.[37] Data were blank filtered using a 10x sample-to-blank ratio using MATLAB where the intensity of each feature was retained if the feature had an intensity 10 times greater than the average intensity across all solvent blanks, reconstitution solvent blanks, and the first/last five fractions. MS^2 data were exported as an .mgf file. SIRIUS[16] + CSI:FingerID[26] was used to generate elemental formulas, *in silico* predicted fragmentation trees, and conduct database matching using the .mgf file. GNPS was used to create feature based molecular networking using the MS^1 data matrix and .mgf file outputs from SLAW.[14, 15] Similarity scores for GNPS (modified cosine) and SIRIUS (Jaccard[17] - Tanimoto[18]) were exported and used for downstream analyses.

Comparison of LC-MS/MS and NMR database annotations

The results from the independent LC-MS/MS (GNPS and SIRIUS) and NMR (COLMAR) database matching platforms were combined to investigate the overlap between the complementary methods. In order to compare the results, metabolite names

and/or database identifiers provided by each software were used to generate InChIKeys using the Chemical Translation Service.[38] The top ranked database matches (similarity score > 0.7) from GNPS, SIRIUS, and COLMAR (TOCSY and HSQC separate queries) were then compared using a subset of five fractions for preliminary analysis (f32, f36, f39, f41, and f42). Using an in-house MATLAB script, individual matrices were matched by planar InChIKeys and fraction number, generating a merged data matrix containing both LC-MS/MS and NMR data. A consensus chemical name list was obtained using the "Query Chemical Identifier Resolver" from the "Webchem" R package.[39] Putative annotations were transformed into a logical vector to determine the presence of each annotation within the different platforms. The UpsetR[40] package was used to generate an Upset plot of the putative annotations across platforms, and ClassyFire[41] was used to classify and organize annotated compounds into 11 different possible levels of categorization.

Relating metabolomics data to mFL data.

LC-MS/MS and NMR data was collected on 42 samples of the same strain (PD1074) and sample generation method (LSCP) as the RM each containing approximately 200,000 nematodes. These datasets were used as starting point to relate metabolomics features detected in differently collected datasets to the mFL. Of the 42 NMR spectra collected, one sample was excluded as an outlier. The remaining 41 processed 1D ^{1}H NMR spectra were imported into MATLAB. The statistical total correlation spectroscopy (STOCSY)[42] analysis was used to obtain highly correlated features from a peak of interest. These chemical shifts were then used to create a covariance matrix and determine the fractions with the highest covariance values. Similarly, LC-MS/MS spectra were collected

on the 42 individual PD1074 samples and 12 pooled samples under identical chromatographic conditions. The use of HPLC to generate the fractions resulted in a retention time (RT) shift between the two datasets, however the elution order of metabolites was maintained. In order to relate features from each study, data matrices for the PD1074 samples and fraction library were aligned using the metabCombiner[43] R package (Table 3.1).

Table 3.1. Combined table output from metabCombiner. Individual *C. elegans* strain PD1074 LC-MS/MS features (id_PD1074) from an unpublished study were aligned to features in the mFL (id_FL). The mass-to-charge (mz), retention time (rt), and abundance measure (Q) are shown for each feature. Alignment scores as generated and shown in the grey column (scores) where scores < 0.5 are highlighted in red text. The alignment rank for each feature is shown as an additional metric.

rowID	id_FL	id_PD1074	mz_FL	mz_PD1074	rt_FL	rt_PD1074	rtProj	Q_FL	Q_PD1074	group	score	rank_FL	rank_PD1074
136.1	36	met_1268	269.0874	269.0881	6.19	5.19	5.22	0.89	0.89	136	0.9018	1	1
17.5	287	met_1851	104.0706	104.0706	8.04	7.05	6.95	0.79	0.89	17	0.8506	1	1
18.7	290	met_793	104.1070	104.1002	5.49	4.28	4.58	0.93	0.53	18	0.373	1	1
18.5	290	met_789	104.1070	104.0943	5.49	4.28	4.58	0.93	0.40	18	0.2429	2	1
27.2	314	met_1740	118.0934	118.0862	7.55	6.43	6.44	0.31	1.00	27	0.5878	1	2
31.1	320	met_1690	120.0807	120.0808	7.37	6.21	6.26	0.46	0.99	31	0.9034	1	1
42.2	353	met_1707	132.1018	132.1019	7.36	6.22	6.25	0.83	1.00	42	0.9498	1	1
42.6	354	met_1744	132.1018	132.1019	8.22	6.45	7.14	0.61	0.99	42	0.3026	1	2
54.3	393	met_1366	146.0923	146.0924	7.22	5.39	6.11	0.82	0.98	54	0.2886	1	1
4.2	404	met_2240	86.0598	86.0601	7.96	7.88	6.86	0.55	0.76	4	0.1708	1	1
71.2	454	met_1793	162.1124	162.1124	8.01	6.91	6.91	0.70	0.98	71	0.9859	1	1
74.1	459	met_1689	166.0861	166.0862	7.35	6.21	6.24	0.52	1.00	74	0.9323	1	1
91.2	500	met_1792	184.0942	184.0944	8.03	6.91	6.93	0.12	0.94	91	0.9459	1	1
96.1	512	met_2170	189.1231	189.1234	8.09	7.70	7.00	0.12	1.00	96	0.2934	1	2
105.1	535	met_1730	205.0969	205.0971	7.41	6.31	6.30	0.12	0.99	105	0.9646	1	1

RESULTS

Fractionation concentrates metabolites and reduces NMR spectral overlap

The concentrated polar extract of the IBAT *C. elegans* RM was fractionated using a semi-preparative scale HPLC that generated 100 fractions collected at 30 seconds intervals. From each fraction approximately 10 mL were aliquoted for NMR analysis and dried down. Each fraction was reconstituted in D_2O and analyzed by 1D NMR, introducing

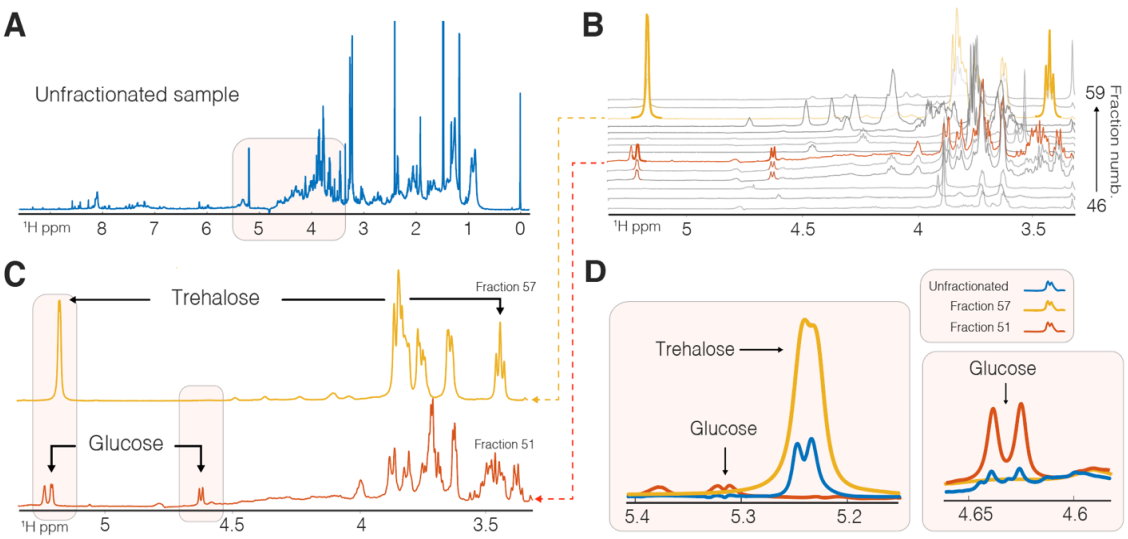


Figure 3.2. A) Unfractionated ^{1}H NMR spectra of the *C. elegans* reference strain PD1074. **B)** Stacked plot of fraction library spectra for fractions f46 to f59. Each spectrum (row) corresponds to a 1D NMR ^{1}H spectra of a single fraction. The chemical shift range was selected to highlight the sugar region, which is a highly overlapped region of the spectrum. Yellow and red colored spectra indicate the fractions where the respective intensities of trehalose and glucose are at their maximum. Colored features in the grey spectra indicate lower intensity glucose and trehalose features in other fractions. **C)** Fractions f51 and f57 in red and yellow illustrate each spectrum distinct profiles and the trehalose and glucose prominent features. Highlighted sections illustrate the regions in panel D) zoomed insets. **D)** Colored boxes illustrate the zoomed insets of glucose and trehalose anomeric protons. Signal to noise ratios (S/N = peak height/mean square root of the baseline) were calculated: trehalose peak 1007 and 4673 for the unfractionated and fraction f57 spectra; glucose peak 59.24 and 274 for the unfractionated and fraction f51 spectra.

a new dimension to the data, the fraction number (Fig 3.2b). Depending on the elution time of each metabolite, neighboring fractions can contain the same metabolite (Fig 3.2b, features shown in red at 5.187 and 4.639 ppm). The repeated injection, separation and collection of the chromatographic eluents allowed for the concentration of each fraction. Figure 3.2b shows the fractionated data for f46 to f59 focusing on the highly overlapped spectral region highlighted in Figure 3.2a (3.5-5.5 ppm). Chemical shifts related to carbohydrates trehalose and glucose are shown in yellow and red and are present in varied abundance across f49-f51 and f57-f58, respectively. Fractions f51 and f57 had the highest concentration of glucose and trehalose, respectively, and are shown in Figure 3.2c. Comparisons of the anomeric proton resonances at 5.226, 5.187, and 4.639 ppm in the unfractionated material and f51 and f57 are shown in Figure 3.2d.

The fractionation starting material contained approximately 2 million nematodes that were extracted to a final volume of 800 μL. A total of 540 μL were injected and separated which is equivalent to 1.35 million nematodes. The theoretical maximum for the increase in concentration in the fractions is 6.75 times more than the unfractionated spectra containing 200,000 nematodes. This theoretical value is consistent with the calculated signal-to-noise (S/N) ratios 4.6 times between the fractionated features compared to the unfractionated material.

We used COLMAR to putatively annotate the unfractionated material and f51 and f57 using HSQC and TOCSY 2D NMR experiments. The matching criteria for trehalose and glucose in the unfractionated material was 1 but with a uniqueness score of 1/7 for trehalose and a combined 1/14 for glucose indicating most features were highly overlapped. [44] In fraction f57, trehalose received a similarity score of 5/7 and in f51 glucose was

annotated with a score of 9/14 demonstrating an increase in annotation confidence through the separation of otherwise overlapping resonances.

NMR annotation of features in each fraction

Two-dimensional (2D) NMR experiments were collected for 8 fractions. As a proof of concept, ^{1}H-^{13}C HSQC and ^{1}H-^{1}H TOCSY of fractions f32, f36, f39, f41, and f42. were used for COLMAR's "TOCSY Query" and "HSQC Query" database matching methods. Results for f51, f57, and f58 were collected at a later date and used solely for identification of trehalose and glucose and not included in the below described database matching steps. For these five fractions we obtained metabolite names, matching ratios and uniqueness scores as well as BMRB or HMDB metabolite identifiers. A total of 365 unique metabolites were annotated with matching scores ranging from 0.01 to 1. The TOCSY query alone generated 284 metabolite annotations for all 5 fractions, while the HSQC query generated the remaining 81 annotations. The highest number of annotations came from fraction f42 queries (171), followed by fraction f41 (65 annotations), f36 (57), f32 (43) and finally f39 with 29 annotations.

LC-MS/MS annotation of features in each fraction

After fractionation, 100 μL of each fraction was aliquoted for LC-MS/MS analysis. HILIC LC-MS/MS data in positive mode was exported as an .mgf file and deconvoluted data matrix from SLAW.[37] A total of 2,312 MS^{1} features were present in the LC-MS/MS fraction dataset after blank filtering (10x). In line with the NMR fractions, a new dimension

of data was created by this process using the combination of mass-to-charge (m/z), RT, and fraction number for each feature.

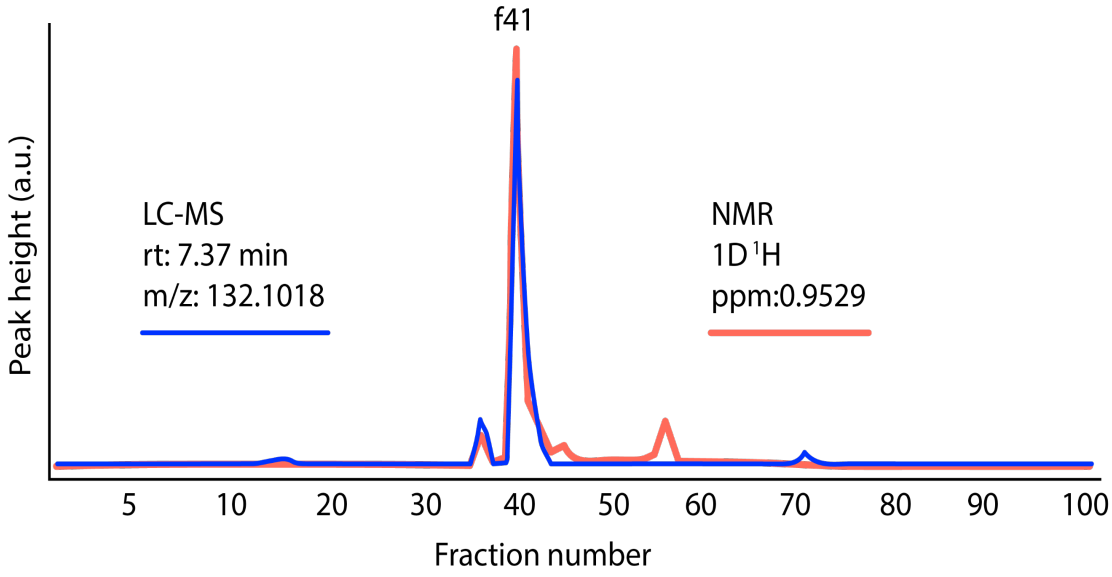


Figure 3.3: Overlaid traces of the peak height (a.u.) corresponding to chemical shift 0.9529 from 1D ^{1}H NMR (red) and m/z-RT pair 132.1019-7.37min from LC-MS/MS (blue) across all 100 fractions.

Elution profiles of features in adjacent fractions can be matched between LC-MS/MS and NMR where the abundance of a feature across fractions is mapped (Fig 3.3). MS^2 data collected from each fraction generated a total of 584 MS^2 scans that were used in GNPS FBMN[15]. These data generated a molecular similarity network where each MS^2 scan is color coded by the fraction where it is detected (Fig 3.4). In addition, we obtained putative annotations for each MS^2 scan from two independent platforms, GNPS and SIRIUS (SIRIUS[16]+CSI:FingerID[26]). SIRIUS also produced elemental formulas and *in silico* fragmentation trees as additional outputs. Both methods generate annotation lists by querying selected databases and calculate similarity scores using different metrics. GNPS provided 102 database matches to database reference standards. SIRIUS was able to

calculate 581/584 elemental formulas, and a total of 395 MS^2 scans were matched to putative annotations. Elemental formulas and therefore database matches were undetermined for features 207, 291, 313, and 314 .

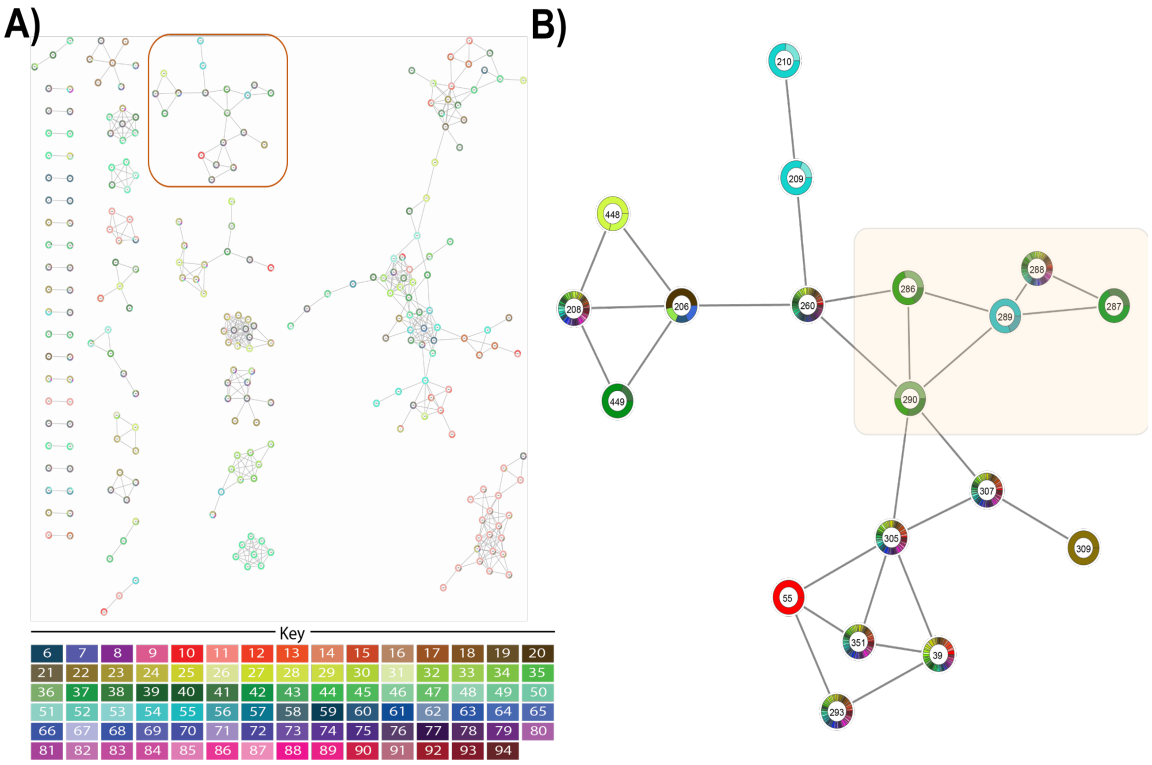


Figure 3.4: A) GNPS FBMN of HILIC LC-MS/MS data displaying nodes where the number of connections is ≥ 2. Each node is color coded using a donut plot where each color corresponds to a specific fraction (shown in the key below) and the portion of that color within the donut plot represents the intensity of that feature across all 100 fractions. B) Zoomed in region corresponding to the orange square in panel A. The yellow box displays a region where multiple features (represented as unique numbers within each donut plot) are matched to the same database entry. These features can serve as future targets for

Annotation coverage by NMR and LC-MS/MS

Individual matrices were matched by planar InChIKeys and fraction number, to generate a merged data matrix containing both LC-MS/MS and NMR data. An UpSet plot displaying the number of unique and overlapping putative annotations is shown in Figure 3.5. A total of 30 annotations were matched between at least two independent methods (*i.e.,* HSQC and TOCSY queries, GNPS and SIRIUS, HSQC and GNPS, etc.). LC-MS/MS and NMR had a total of 49 and 31 uniquely annotated metabolites, respectively. Seven annotations were matched between LC-MS/MS and NMR, with four annotations having matched to HSQC, TOCSY, and GNPS and three annotations matching all four platforms. A total of 87 putative annotations were made for these 5 fractions using either HILIC positive mode MS spectra, NMR spectra or both.

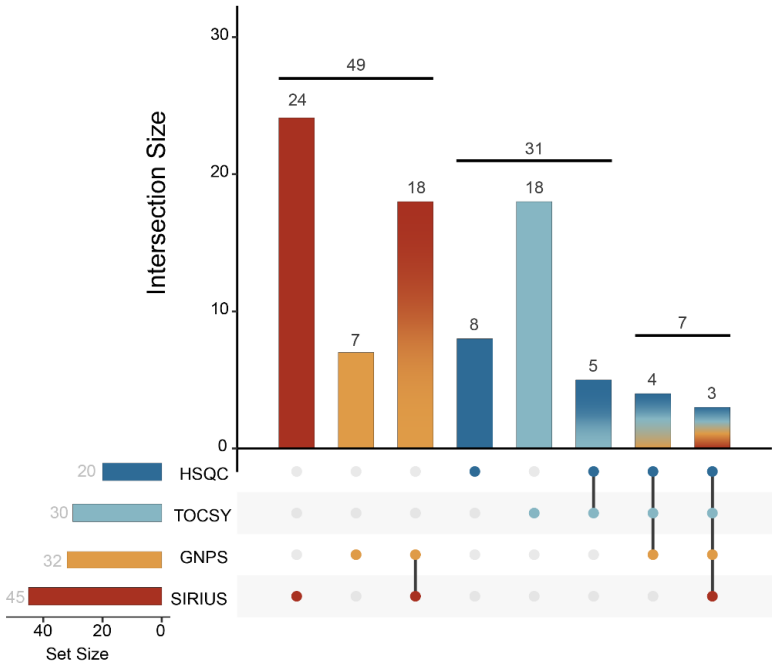


Figure 3.5. UpSet plot representing the number of unique putatively annotated metabolites in fractions f32, f36, f39, f41 and f42 across two different analytical platforms (NMR and LC-MS/MS) and different annotation methods (SIRIUS, GNPS, COLMAR HSQC query, and TOCSY query). Top ranked database candidates and a cutoff score of 0.7 for matching/similarity criteria were used for each method. The height of each bar illustrates the number of unique and overlapping metabolites annotated for each method indicated by the colored dots along the same column. The horizontal black line and number above adjacent columns detail the sum of those columns. The bottom left bar chart represents the horizontal sum of the number of metabolites identified by each individual method.

ClassyFire[41] was used to classify and organize the compounds annotated in Figure 3.5. Annotations within the five investigated fractions were organized into seven SuperClasses. The percentage of the 49 and 31 annotations individually determined by LC-MS/MS and NMR, respectively, and the seven overlapping annotations that fall into each of the SuperClasses determined by ClassyFire were calculated and shown in Figure 3.6. Organic acids and derivatives had the highest cumulative percentage (32.2%) with 4.6% of these annotations overlapping with LC-MS/MS and NMR. Lipid and lipid-like molecules

showed the second highest cumulative percentage (20.7%); however, the majority (18.4%) were only annotated by LC-MS/MS. Interestingly, benzenoids showed a very small percent coverage (10.3%), however, the majority of these annotations were from NMR alone (6.9%) highlighting the benefits of NMR spectroscopy for the annotation of these stable ring structures. The structures for all 87 annotations were compiled into a spreadsheet and examined.

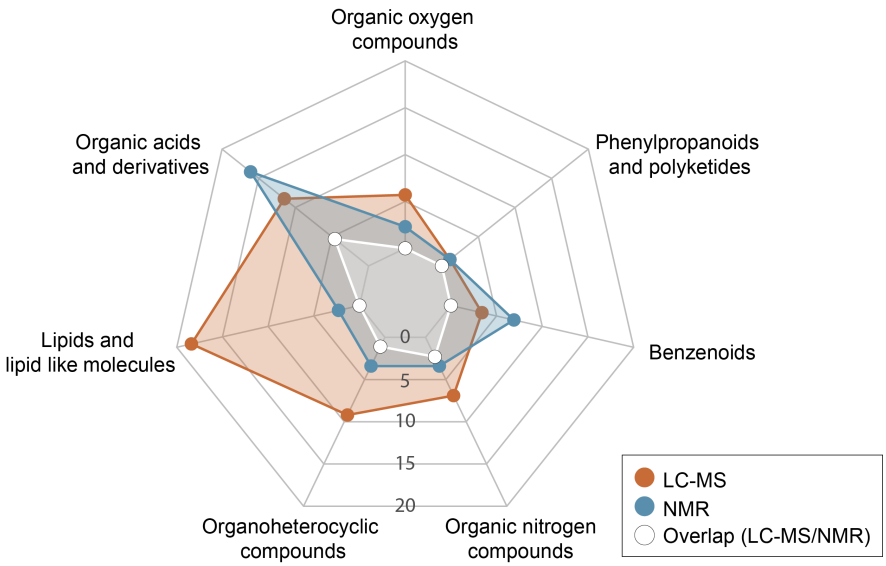


Figure 3.6. Radar plot of annotated metabolites. Each vertices represents a compound superclass determined by ClassyFire. The percentage of annotations that fall within each superclass by LC-MS/MS and NMR individually are shown in orange and blue, respectively. The seven overlapping LC-MS/MS and NMR annotations (Figure 3.3) are displayed in white. Percentage increases from 0% at the inner web to a maximum of 20% at the outer edge.

Bridging NMR and LC-MS/MS for database matching

To expand the search space beyond only top ranked database hits, a similar approach was used on the collective sum of putative annotations from each platform regardless of similarity score. A total of 21 matches between LC-MS/MS and NMR were

identified within the five investigated fractions (Table 3.2). The seven top ranked database matches shown in Figure 3.5 were also identified when using the expanded search space.

For duplicate matches (same MS scan number and annotation) the highest similarity and matching scores were used. The intensity of a feature (m/z and RT pair for LC-MS/MS and a single chemical shift for NMR) for each annotation was extracted across all fractions. The maximum intensity fraction was recorded as an additional layer of complementarity between the two platforms. NMR features that could not be matched between the database spectra and the respective fraction spectra determined by the maximum MS intensity were defined as “undetermined”.

The list of fractions for each MS^2 scan was obtained from GNPS-FBMN based on the corresponding MS^1 feature intensities. From the 21 annotations only 16 correspond to unique MS^2 scans and 17 to unique compounds. 3-aminoisobutyric acid, gamma-aminobutyric acid and 2-aminobutyric acid all have the same scan number. The isomers Ile and Leu share the same scan number with high confidence scores for both GNPS and NMR, however a second scan for Ile has been matched in both GNPS and SIRIUS but with a relatively low score and maximum intensity at fraction f39. Phe and Trp have high NMR matching ratios in both TOCSY and HSQC queries as well as consistently high similarity scores for SIRIUS and GNPS. Val and betaine NMR annotations were made from spectra f42, but their maximum intensities were determined at fractions f44 and f48 respectively and a 1D overlay of the database spectra were visually matched. Choline has high similarity scores for MS but lower NMR matching ratios. Ala-Ala has consistent low scores, and the database NMR spectra was generated by HMDB spectral prediction. All other annotations were matched to experimentally collected database spectra.

Table 3.2. List of putative annotations that are matched between NMR and LC-MS/MS and are detected in the same fraction. TOCSY and HSQC searches were done separately using different methods within the COLMAR web server. The cross-platform matches were carried out using InChIKeys. The maximum height for a particular m/z-RT combination was calculated from MS^1 data. For NMR, a feature is selected according to the overlap between the 1D database standard and the fraction spectrum specified by the MS maximum. The intensities at that chemical shift in each fraction were then used to calculate the maximum intensity point. The similarity scores and matching ratios were obtained from MS^2 and respective 2D NMR experiments. Asterisks represent the seven features in Figure 3.5 that were annotated by both LC-MS/MS and NMR.

Scan (MS^2) identifier	Putative Annotation	SIRIUS similarity score	GNPS similarity score	TOCSY query matching ratio	HSQC query matching ratio	Fraction Number(s) (MS)	Fraction(s) @ max intensity (MS)	Fraction(s) @ max intensity (NMR)
36	Inosine	n/a	0.959	n/a	0.63	42,41	42	undetermined
287	3-aminoisobutyric acid*	0.500	0.729	0.33	1	41, 42	42	undetermined
287	gamma-aminobutyric acid	0.477	n/a	1	1	42, 41	42	42
287	2-aminobutyric acid	0.489	0.540	0.33	n/a	42	42	undetermined
290	Choline	0.942	0.989	n/a	0.67	36, 38, 37	36	undetermined
314	Betaine*	n/a	0.967	n/a	1	44, 48, 46, 45, 41, 47, 42	48	48-1D annot. confirmation
314	Valine	n/a	0.111	0.83	1	44, 48, 46, 45, 41, 47, 42	44	44-1D annot. confirmation
320	2-amino-1-phenylethanol	n/a	0.521	0.33	n/a	42, 41	42	undetermined
353	Isoleucine*	n/a	0.910	n/a	1	42, 39, 41	41	42
353	Leucine	n/a	0.910	1	1	42, 39, 41	41	41
354	Isoleucine*	0.290	0.468	1	1	42, 39, 41	39	42
393	4-guanidinobutyric acid	0.895	0.780	0.33	0.67	39	39	undetermined
404	gamma-aminobutyric acid	n/a	0.155	1	1	42, 46, 41	42	42
426	Valine	0.289	n/a	0.83	1	43, 44, 42, 41, 87, 45, 47	44	44-1D annot. confirmation
426	5-aminopentanoic acid	0.234	n/a	0.2	n/a	43, 44, 42, 41, 87, 45, 47	44	undetermined
452	Alanine-alanine	0.410	n/a	0.25	n/a	42, 32, 39	42	undetermined
454	Carnitine*	0.939	0.959	0.33	0.75	41, 36, 39	40	undetermined
459	Phenylalanine*	0.951	0.986	1	1	42, 32, 36, 41, 42, 41	42	42
500	Carnitine	0.621	0.911	0.33	0.75	40, 39, 41, 47, 36	40	undetermined
512	Glycil-leucine	0.591	0.303	0.2	n/a	39, 41, 42	41	undetermined
535	Tryptophan*	0.970	0.795	0.67	1	39, 41, 42	41	41

Experimental datasets can be related back to the mFL to increase throughput for future annotations

A driver peak at 0.9529 ppm was selected for STOCSY analysis using 41 unfractionated individual PD1074 samples from a metabolomics study (Figure 3.7a). A total of 81 datapoints had a correlation to the driver peak that was larger than 0.95 (Figure 3.7b). These chemical shifts were extracted and used to query the fraction library. The covariance of these features in the fraction library was calculated and represented as a matrix in Figure 3.7C. Fractions f41-f42 showed large covariance for those chemical shifts. NMR spectra from the corresponding fractions (Fig 3.7d) were then compared to spectral databases for metabolite annotation and validated with 2D NMR experiments (Fig 3.8).[19]

Similarly, LC-MS/MS spectra were collected on 42 individual and 12 pooled PD1074 samples from the same metabolomics study described above. R package metabCombiner was used to align the dataset (Table 3.1) with features identified in the mFL based on their RT, m/z, and intensity (Q). All but two of the unique MS^2 identifiers shown in Table 3.2 were matched to MS^1 features in the PD1074 dataset showcasing how this approach can link disparately collected LC-MS/MS datasets when chromatography and/or elution order is maintained. Of the 14 aligned MS^2 features shown in Table 3.1, six received alignment scores less than 0.5 and require additional confirmation.

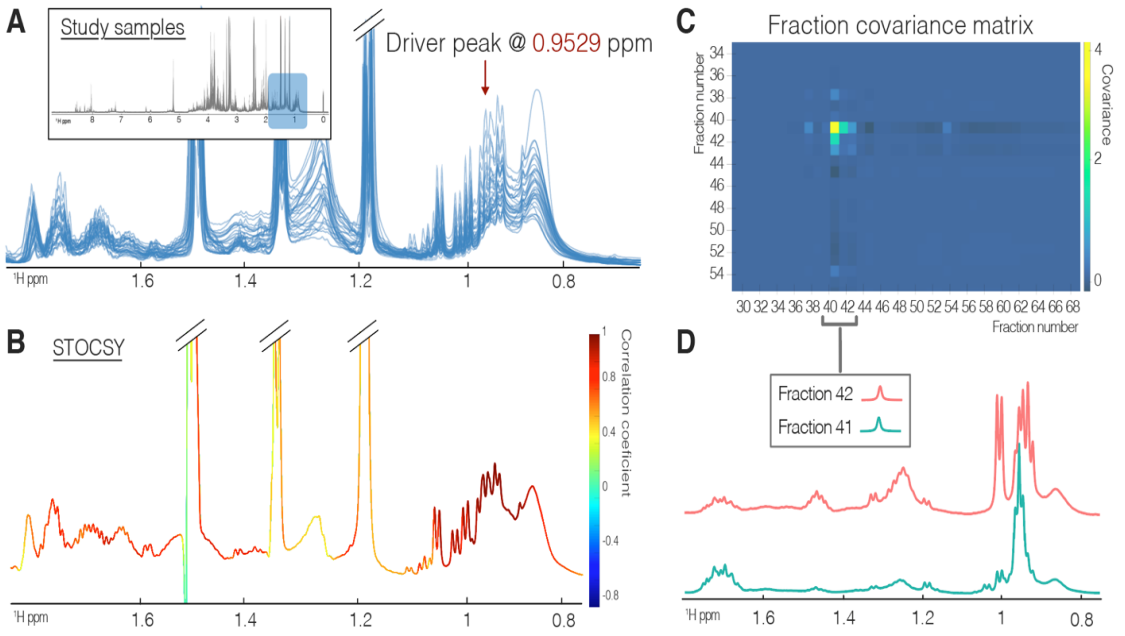


Figure 3.7. A) Full resolution spectra of the *C. elegans* reference strain PD1074 from a metabolomics study are represented as an inset. The blue square highlights the region of interest displayed by the spectra in blue. The feature of interest corresponds to an overlapped Leu feature at 0.953 ppm. **B)** STOCSY output highlighting the correlated features to the driver peak in panel A calculated from the PD1074 spectra in a study. **C)** Covariance matrix calculated using the fractionated spectra only. Features with a correlation above 0.95 from the STOCSY output were used to obtain fraction peak heights and calculate the fractions covariance. Yellow and green squares indicate the fractions where those features are prominent. **D)** Individual spectra from each fraction identified by the covariance matrix.

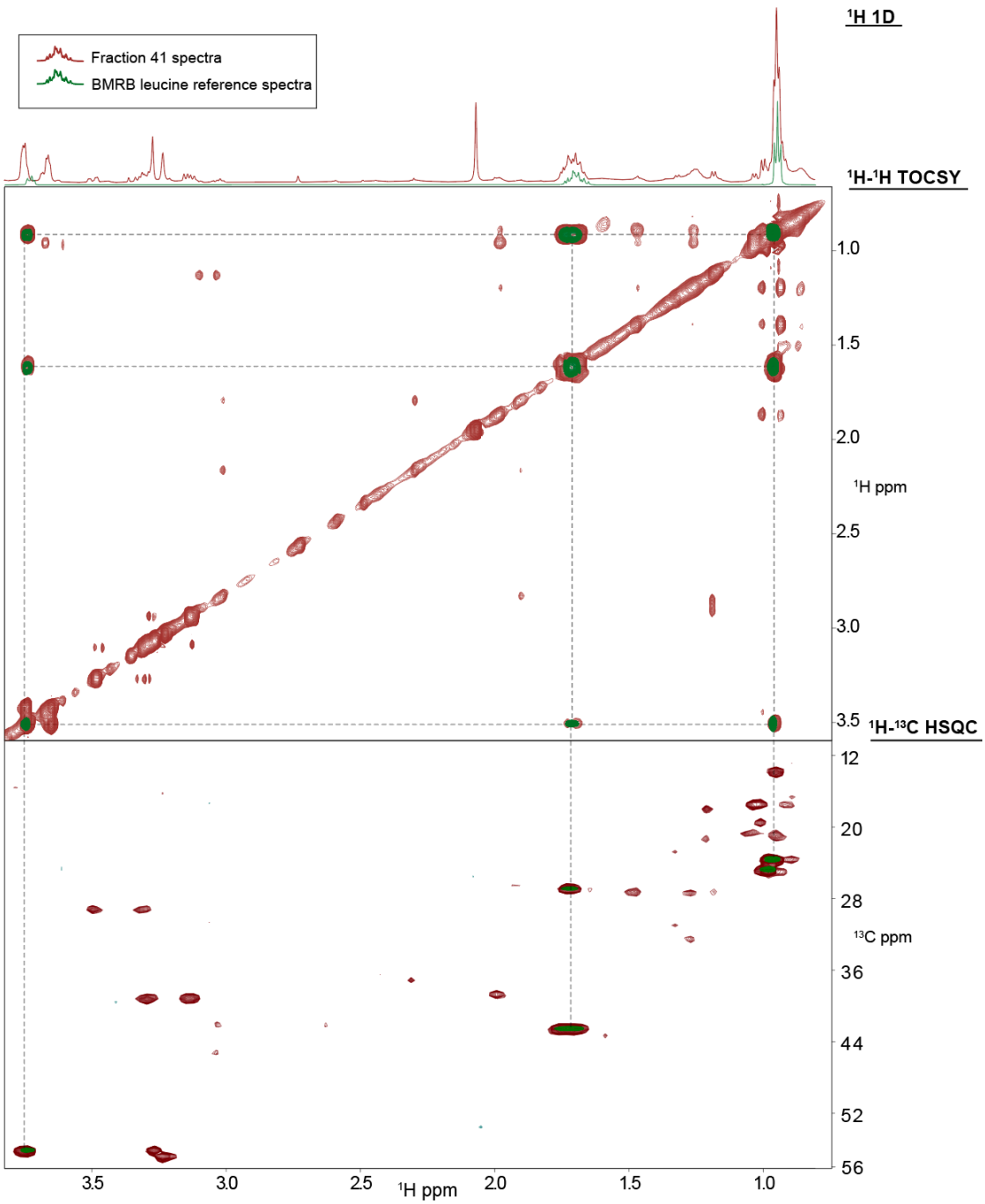


Figure 3.8. Leu NMR annotation confirmation. The spectra in green were downloaded from BMRB from a Leu chemical standard at concentration of 100mM and pH of 7.4 on a 600MHz magnet. The spectra in red correspond to fraction number 41 and were collected on a 600Mhz magnet in our laboratory. The intensities of the features were adjusted for visualization.

DISCUSSION AND CONCLUSION

Fractionation increases concentration and decreases spectral overlap of NMR spectra while retaining features

Due to their narrow line widths and ubiquitous presence in organic molecules, ^{1}H resonances generate information rich NMR spectra, but are also hard to interpret because of the overlap between features. Methods to reduce overlap at the sample preparation stage have been developed but generate artifacts, require previous knowledge of the metabolites present within the sample, or considerably reduce metabolite coverage.[45-47] Two-dimensional NMR experiments markedly reduce this overlap but require longer acquisition times. ^{1}H-^{1}H homonuclear experiments are particularly attractive, as they retain much of the same information as a 1D ^{1}H spectra but also retain considerable overlap. ^{1}H-^{13}C heteronuclear experiments provide unrivaled separation of signals because of the added carbon dimension. However, while capable of dramatically increasing annotation confidence, the low sensitivity of these experiments is a limiting factor even with long acquisition times to compensate for the low natural abundance of ^{13}C. Furthermore, some overlap still exists, especially for compounds that share similar structures, often making the structural elucidation of a molecule of interest difficult if not impossible. Thus, there is a need for alternate approaches that can reduce spectral overlap and simultaneously increase concentrations for metabolite annotation and identification.

Preparative scale separations have been predominant in other fields to address these limitations but are still not routine practice in untargeted metabolomics experiments. The lack of large quantities of material and connecting elements between studies are the biggest challenges. However, matrix matched RMs provide the necessary conditions to overcome

these limitations. The diversity of commercially available RMs is limited, and in-house materials often lack stability and longevity. An alternative to commercially available RMs is IBAT, a logistically simple method developed by the authors that generates, a long-term, stable RM of any matrix.[34]

Here, we have demonstrated how a RM and fixed-time interval fractionation can improve the S/N of NMR spectra and improve metabolite annotation. Peak overlap is particularly problematic for resonances of similar structures. The ^{1}H resonances of sugar rings are a good example of a problematic region (Fig 3.2). These structures generate ^{1}H resonances between 3.5-6 ppm and are common to all monosaccharides, oligosaccharides, polysaccharides and other carbohydrate derivatives. Glucose and trehalose are typical sugars found in *C. elegans* and their anomeric resonances are readily noticeable indications of their presence in complex mixtures. Database matching relies on the matching ratio between experimental and database spectra, and the uniqueness of a spectral match. As mentioned above NMR relies on multiple features to define a compound and because of overlap, the matching ratios for sugars are typically high, and their uniqueness scores low.

This is illustrated by the unfractionated COLMARm scores for glucose and trehalose. However, the separation highlighted in Figure 3.2 increased those uniqueness scores and therefore the confidence in their previously ambiguous annotation. Furthermore, the intensity of each feature is dependent not only on the metabolite concentration but also on the number of ^{1}H resonances that generate the signal. This translates into some features that may not be detected out of a whole molecule, and therefore, fail to provide sufficient detail to annotate a compound of interest confidently. Low intensity features and overlap

are detrimentally additive and a common occurrence in NMR database matching. As such, fractionation methods can significantly improve NMR annotation and database matching.

Bridging NMR and LC-MS/MS with database matching to improve annotation

Database matching is fast and allows for a large number of features to become key elements of the biological question under investigation. As such, incorrect annotations can lead to misinterpretation of biological findings that are reported and propagated throughout the scientific literature until proven wrong. To ensure higher metabolite annotation confidence and transparent reporting, the Metabolomics Standards Initiative implemented guidelines for quantifying annotation confidence into four different levels, from best (1) to worst (4): (1) identified compound, (2) putatively annotated compound (3) putatively characterized compound classes and (4) unknown compounds. As defined, and later modified to address developing technologies, [48] level 1 requires a minimum of two independent and orthogonal data confirmations relative to an authentic compound analyzed under identical experimental conditions. Thus, all database spectral matching annotations that rely on a single analytical platform fall within level 2. [49] Thus, most untargeted metabolomics studies that do not rely on chemical standards, will only ever achieve level 1 annotations if they're deemed critical to carry out further analysis for structural elucidation of an unknown metabolite or confirmation of the metabolite annotation. The latter is generally less likely, and often reliant on database metric outputs to satisfy an annotation requirement to proceed with a biological interpretation.

The level 1 requirement for orthogonality ensures validation. Current practices for LC-MS/MS untargeted metabolomics entail the use of several spectral matching services

to provide an additional level of confidence on the annotation. However, the problem becomes more complex. With different matching algorithms and different metabolite coverage, what criteria is deemed acceptable for a metabolite annotation? Here, we have used both GNPS and SIRIUS similarity score cut-offs of 0.7 to determine the overlap between the two methods. As described previously, each of these methods uses a different scoring metric to quantify similarity (from 0 to 1, where 1 represents the highest level of similarity between experimental and library spectra), however both methods use the same data, and so do not meet the orthogonality criteria. A total of 37% of the features annotated in five individual fractions were in common between the two methods (Fig 3.5) effectively expanding the number of annotations by querying against different databases. However, it becomes increasingly hard to determine whether one-sided annotations are due to coverage or inherent to the matching method (*i.e.,* matching algorithm, quality of the spectra, *etc.*) and even harder when the matching ratios disagree (>0.7 and <0.7). Nonetheless, high concordant matching ratios across two analytical instruments, satisfy the orthogonality requirements, and provide stronger indication of the correct identification of the feature of interest to the extent of the limitations of both methods and the metabolite chemical properties.

This complementarity is well illustrated by Phe (Table 3.2), which had scores >0.95 in all categories. Using NMR data, it was possible to completely match all the reference standard spectral features to the fraction library data, however this is time consuming and not always possible or feasible for a large number of molecules. This annotation was also confirmed by LC-MS/MS using an amino acid standard which was fractionated and analyzed in the same manner as the complex mixture. The same validation method was

used to confirm the annotation of gamma-aminoisobutyric acid (GABA), although its MS based similarity scores were bellow passing grade (<0.4). GABA was matched as a possible annotation by LC-MS/MS to two individual MS^2 scan numbers (287 and 404), of which scan 287 showed matches to 3-aminoisobutyric acid (BAIB), GABA, and 2-aminobutyric acid (AABA) further showcasing the complexity of MS-based annotation, especially when dealing with isomers with very similar fragmentation patterns that differ mainly in intensity. Based on the NMR similarity score and maximum fraction shown in Table 3.2, GABA looks like a promising annotation for MS^2 scan 287, however spike-in experiments with a commercially available standard are needed to definitively confirm this to a level 1 annotation and identify scan 404 as a misannotation.

Conversely, the annotation of choline shows high confidence scores in LC-MS/MS with a score >0.9 for both GNPS and SIRIUS but is met with a lacking NMR score resulting from a single feature in the experimental HSQC spectrum. The choline HSQC reference spectrum consists of a strong singlet at 3.195 ppm, corresponding to the nine equivalent ^1H of the three methyl groups, and two low intensity peaks at 3.514 and 4.053 ppm of the CH_2 groups. The low-level peaks in the fraction spectrum are too low level to be detected even after concentration which leaves the annotation of choline solely reliant on a single peak which is not specific enough to confirm the annotation by NMR alone. Further increase in concentration would be required with additional semi- preparative HPLC injections and subsequent combination of eluting fractions for metabolites with low intensity features or low concentration.

Cross-platform analysis of the same fixed-time interval fractionation provides an unexpected layer of confidence. Because eluting metabolites can be collected over more

than one fraction, the resulting change in intensity, if matched, adds further evidence of concordant annotations. This is illustrated by several examples in Table 3.2 (*i.e.,* GABA, betaine, Phe, Trp) but furthermore, can also help discern between isomers. Leu and Ile represented by two entries of the same MS^2 scan number (353) with a maximum intensity at f41 and a second MS^2 scan (354) with a maximum intensity at fraction f39. MS^2 fragmentation does not readily distinguish these isomers; however, manual inspection of the chromatograms revealed baseline separation using HILIC separation methods. These isomers are easily differentiated by NMR as shown in Fig 3.7. Their corresponding maximum intensities demonstrate the correct Leu annotation for scan 353 and a misannotation of Ile for scan number 354. The missing Ile LC-MS/MS annotation was manually investigated and determined to stem from the dynamic exclusion settings used for DDA, showcasing the variability introduced through instrument specific parameter settings. An amino acid standard mixture was fractionated and confirmed the RT, ,m/z, and fraction number of both features in the mFL , confirming their identity and elution order. Interestingly the discrepancy between NMR and MS platforms and their complementarity is well illustrated in Figure 3.5.

A total of 31 matches were unique to NMR that were not annotated by LC-MS/MS. However, these MS data are highly restricted by the chromatography and polarity of the measurements (HILIC positive mode) which is not the case for NMR. Nonetheless, there is a clear consistent trend in the class of analytes measured by both platforms (Fig 3.6). Lipid and lipid like molecules are mostly detected by LC-MS/MS. This is consistent with lipids being historically hard species to analyze by NMR in metabolomics due to their repeating structures and overlapping resonances, but to certain extent misconstrued as

recently demonstrated,[50] and can be further improved with fractionation strategies. Additionally, molecules with intermittent oxygen and nitrogen atoms also showed higher coverage by LC-MS/MS which can leverage fine isotopic structure and fragmentation for identification. As expected, small organic acids and derivatives are the most overlapped class of molecules between the two platforms while also the most abundant class detected by NMR. As additional 2D NMR spectra are collected on fractions the true complementarity and uniqueness of each platform will be better elucidated. However, a benefit of the fraction library is the ability to query these fractions and collect data on an as-needed basis as discoveries and hypotheses are made.

Experimental datasets can be related back to the mFL for rapid annotation

To effectively relate study features of interest to the mFL it is necessary to devise methods that can circumvent the manual search and comparison between mFL and a metabolomics study. STOCSY is a well-established method to extract resonances that belong to the same molecule,[51] however, it is limited by the variance of the dataset and peak overlap that can lead to ambiguous results with high rate of missed or falsely selected features. Unlike LC-MS/MS, that utilizes the paired combination of m/z and RT to create a unique identifier for a compound of interest, NMR relies on multiple chemical shifts to define a compound. A single NMR feature can easily be overlapped with other features in a complex matrix making it extremely difficult to associate to a single compound. However, high correlation thresholds from STOCSY (>0.85) help reduce the number of possible candidate features from the whole spectra. We demonstrated that this filtered list of chemical shifts can then be leveraged to identify fractions where the majority of features

belonging to that molecule covary. This approach enables NMR resonances that belong to the same molecule to be identified, not solely on statistical correlations, but now as well on the added dimension of 'fraction number', creating a unique identifier similar to m/z and RT for LC-MS/MS studies. This approach is scalable and allows for 100's of spectra in the FL to be queried semi-automatically without prior knowledge of the compound's identification. Similarly, peak lists from reference standards or computational predictions can be used in this pipeline to identify fractions where that compound is present and validate the match using 2D experiments or previously collected data.

Herein, LC-MS/MS data was collected using a similar separation method as that used for the fractionation process. While this approach does not further separate metabolites in the chromatographic dimension, it does allow for metabolites to be matched and queried between metabolomics experiments and the fraction library. A shift in RT was observed for metabolites (Table 3.1), but the elution order of metabolites remained consistent. This allows for computational methods, such as metabCombiner described by Habra *et al.,* to align features between disparately acquired LC-MS/MS metabolomics data using m/z, RT, and relative abundance similarities (Table 3.1). The use of orthogonal separation methods can be an added benefit; however, this necessitates the use of retention indexing between divergent chromatographic methods or the use of RT prediction, which requires large training datasets and therefore significant additional instrument time and benchwork.[52-54]

Data-dependent acquisition (DDA) LC-MS/MS was used for the collection of MS^2 data and database matching. However, the increased concentration provided through the fractionation process coupled with low volume requirements for MS provides additional

avenues to be explored. For example, coupling a matched LC-MS/MS separation followed by ion mobility separation (IMS) would allow individual datasets to be RT-matched to the fraction library while simultaneously improving the separation of the LC-MS/MS data and decreasing complexity. More importantly for the structural elucidation process, this approach would provide highly complementary structural information in the form of collisional cross section measurements (CCS) that can then be compared to machine learning-based approaches for CCS prediction.[55, 56]

The ability to relate features identified in the fraction library to separately collected metabolomics datasets was showcased here using an IBAT RM. One clear limitation of this approach is that a metabolite of interest may not be present in the RM. In this event, a feature of interest identified in the LC-MS/MS dataset can undergo the typical targeted isolation and/or purification with the added guidance of the adjoining features in the fraction library, reducing the usual effort needed to select the optimal fraction. Furthermore, the advantage of a RM is that it reflects the matrix effects of the study samples, as such, high overlap between the RM and the study samples metabolome is expected. While this approach can be used on any material that is regularly included in metabolomics studies, the inclusion of a RM presents many additional benefits. Creating a fraction library of a RM ensures the variation in the measured features is kept low and consistently detected in every study. Over the course time and multiple experiments, using different methods and instrumentation, it is possible to achieve a systematic annotation of a RM. In addition, features of non-biological origin (*i.e.,* plasticizers, impurities, artifacts, etc.) that are often laboratory specific can be appropriately annotated and removed prior to statistical analysis. Finally, fractions can be analyzed by multiple analytical methods (*i.e.,*

ion mobility, direct infusion MS, high resolution MS (HRMS) and/or different chromatographic conditions) as well as further concentrated by repeating the fractionation process or targeting a specific metabolite for isolation/purification. Therefore, the effort required to elucidate an unknown metabolite present in the mFL is only carried out once and used throughout future studies.

While an IBAT RM is critical for the fractionation to be cost effective for metabolomics studies, alternative materials can be used. Human biofluid studies, for example, often rely on large quantities of procured (commercially or sourced) urine or plasma samples that are regularly included in multiple studies over time and therefore are fit for purpose to create a mFL. These materials can be used to generate an IBAT RM, thereby prolonging the stability of the material, and improving management logistics. The approach detailed here is well suited as an annotation resource and continuously evolving library, as demonstrated for laboratories that lack semi-scale equipment.[32]

We have demonstrated that spectral matching has limitations and that these are highly platform dependent and even more so metabolite dependent. The creation of a mFL allows for multiple analytical measurements to be made on a simplified fraction of a complex matrix, increasing both S/N and database coverage through the use of complementary technologies. The effort required to generate a mFL is outweighed by the benefits of being able to leverage the information rich spectra and confirmed annotations in downstream experiments through the inclusion of the same sample type (*i.e.,* IBAT RM or commercially purchased biofluid). Additionally, the generation of archived concentrated aliquots of each fraction facilitates additional analytical measurements that can further overcome these limitations by orthogonalizing database matching, and therefore increasing

the confidence of the annotations. This library has the potential to serve as a long-standing resource that can increase the turnaround time from analytical measurement to confident annotation, as well as improve unknown metabolite identification through the inclusion of molecular networking and computational predictions (concepts to be demonstrated in an upcoming publication).

REFERENCES

1. Bhinderwala, F.; Wase, N.; DiRusso, C.; Powers, R., Combining Mass Spectrometry and NMR Improves Metabolite Detection and Annotation. *Journal of Proteome Research* **2018,** *17* (11), 4017-4022.

2. Wang, C.; He, L.; Li, D.-W.; Bruschweiler-Li, L.; Marshall, A. G.; Brüschweiler, R., Accurate Identification of Unknown and Known Metabolic Mixture Components by Combining 3D NMR with Fourier Transform Ion Cyclotron Resonance Tandem Mass Spectrometry. *Journal of Proteome Research* **2017,** *16* (10), 3774-3786.

3. Bingol, K.; Brüschweiler, R., Knowns and unknowns in metabolomics identified by multidimensional NMR and hybrid MS/NMR methods. *Current Opinion in Biotechnology* **2017,** *43*, 17-24.

4. Bingol, K.; Brüschweiler, R., Two elephants in the room: new hybrid nuclear magnetic resonance and mass spectrometry approaches for metabolomics. *Current opinion in clinical nutrition and metabolic care* **2015,** *18* (5), 471.

5. Bingol, K.; Bruschweiler-Li, L.; Yu, C.; Somogyi, A.; Zhang, F.; Brüschweiler, R., Metabolomics Beyond Spectroscopic Databases: A Combined MS/NMR Strategy for the Rapid Identification of New Metabolites in Complex Mixtures. *Analytical Chemistry* **2015,** *87* (7), 3864-3870.

6. Mahieu, N. G.; Huang, X.; Chen, Y., Jr.; Patti, G. J., Credentialing Features: A Platform to Benchmark and Optimize Untargeted Metabolomic Methods. *Analytical Chemistry* **2014,** *86* (19), 9583-9589.

7. de Jong, F. A.; Beecher, C., Addressing the current bottlenecks of metabolomics: Isotopic Ratio Outlier Analysis, an isotopic-labeling technique for accurate biochemical profiling. *Bioanalysis* **2012,** *4* (18), 2303-14.

8. Stupp, G. S.; Clendinen, C. S.; Ajredini, R.; Szewc, M. A.; Garrett, T.; Menger, R. F.; Yost, R. A.; Beecher, C.; Edison, A. S., Isotopic ratio outlier analysis global metabolomics of Caenorhabditis elegans. *Anal Chem* **2013,** *85* (24), 11858-11865.

9. Guan, S.; Armbruster, M. R.; Huang, T.; Edwards, J. L.; Bythell, B. J., Isomeric Differentiation and Acidic Metabolite Identification by Piperidine-Based Tagging, LC-MS/MS, and Understanding of the Dissociation Chemistries. *Anal Chem* **2020,** *92* (13), 9305-9311.

10. Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, Brian L.; Berjanskii, M.; Mah, R.; Yamamoto, M.; Jovel, J.; Torres-Calzada, C.; Hiebert-Giesbrecht, M.; Lui, Vicki W.; Varshavi, D.; Varshavi, D.; Allen, D.; Arndt, D.; Khetarpal, N.; Sivakumaran, A.; Harford, K.; Sanford, S.; Yee, K.; Cao, X.; Budinski, Z.; Liigand, J.; Zhang, L.; Zheng, J.; Mandal, R.; Karu, N.; Dambrova, M.; Schiöth, Helgi B.; Greiner, R.; Gautam, V., HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Research* **2022,** *50* (D1), D622-D631.

11. Stein, S. E.; Scott, D. R., Optimization and testing of mass spectral library search algorithms for compound identification. (1044-0305 (Print)).

12. Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G., METLIN: a metabolite mass spectral database. *Therapeutic drug monitoring* **2005,** *27* (6), 747-751.

13. Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K., MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry* **2010,** *45* (7), 703-714.

14. Nothias, L.-F.; Petras, D.; Schmid, R.; Dührkop, K.; Rainer, J.; Sarvepalli, A.; Protsyuk, I.; Ernst, M.; Tsugawa, H.; Fleischauer, M.; Aicheler, F.; Aksenov, A. A.; Alka, O.; Allard, P.-M.; Barsch, A.; Cachet, X.; Caraballo-Rodriguez, A. M.; Da Silva, R. R.; Dang, T.; Garg, N.; Gauglitz, J. M.; Gurevich, A.; Isaac, G.; Jarmusch, A. K.; Kameník, Z.; Kang, K. B.; Kessler, N.; Koester, I.; Korf, A.; Le Gouellec, A.; Ludwig, M.; Martin H, C.; McCall, L.-I.; McSayles, J.; Meyer, S. W.; Mohimani, H.; Morsy, M.; Moyne, O.; Neumann, S.; Neuweger, H.; Nguyen, N. H.; Nothias-Esposito, M.; Paolini, J.; Phelan, V. V.; Pluskal, T.; Quinn, R. A.; Rogers, S.; Shrestha, B.; Tripathi, A.; van der Hooft, J. J. J.; Vargas, F.; Weldon, K. C.; Witting, M.; Yang, H.; Zhang, Z.; Zubeil, F.; Kohlbacher, O.; Böcker, S.; Alexandrov, T.; Bandeira, N.; Wang, M.; Dorrestein, P. C., Feature-based molecular networking in the GNPS analysis environment. *Nature Methods* **2020,** *17* (9), 905-908.

15. Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W.-T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.;

Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. Ø.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016,** *34* (8), 828-837.

16. Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S., SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* **2019,** *16* (4), 299-302.

17. Jaccard, P., THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytologist* **1912,** *11* (2), 37-50.

18. Tanimoto, T. T., Elementary mathematical theory of classification and prediction. **1958**.

19. Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z., BioMagResBank. *Nucleic acids research* **2007,** *36* (suppl_1), D402-D408.

20. Steinbeck, C.; Kuhn, S., NMRShiftDB -- compound identification and structure elucidation support through a free community-built web database. (0031-9422 (Print)).

21. Markley, J. L.; Dashti, H.; Wedell, J. R.; Westler, W. M.; Eghbalnia, H. R., Tools for Enhanced NMR-Based Metabolomics Analysis. *Methods Mol Biol* **2019,** *2037*, 413-427.

22. Bingol, K.; Li, D. W.; Zhang, B.; Bruschweiler, R., Comprehensive Metabolite Identification Strategy Using Multiple Two-Dimensional NMR Spectra of a Complex

Mixture Implemented in the COLMARm Web Server. *Anal Chem* **2016,** *88* (24), 12411-12418.

23. Li, D.-W.; Hansen, A. L.; Yuan, C.; Bruschweiler-Li, L.; Brüschweiler, R., DEEP picker is a deep neural network for accurate deconvolution of complex two-dimensional NMR spectra. *Nature Communications* **2021,** *12* (1), 5229.

24. Ravanbakhsh, S.; Liu, P.; Bjorndahl Tc Fau - Mandal, R.; Mandal, R.; Grant, J. R.; Wilson, M.; Eisner, R.; Sinelnikov, I.; Hu, X.; Luchinat, C.; Greiner, R.; Wishart, D. S., Accurate, fully-automated NMR spectral profiling for metabolomics. (1932-6203 (Electronic)).

25. Wang, F. A.-O.; Liigand, J. A.-O.; Tian, S. A.-O.; Arndt, D.; Greiner, R.; Wishart, D. A.-O., CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. (1520-6882 (Electronic)).

26. Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S., Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences* **2015,** *112* (41), 12580.

27. Lai, Z.; Tsugawa, H.; Wohlgemuth, G.; Mehta, S.; Mueller, M.; Zheng, Y.; Ogiwara, A.; Meissen, J.; Showalter, M.; Takeuchi, K.; Kind, T.; Beal, P.; Arita, M.; Fiehn, O., Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nature Methods* **2018,** *15* (1), 53-56.

28. Das, S.; Edison, A. S.; Merz, K. M., Metabolite Structure Assignment Using In Silico NMR Techniques. *Analytical Chemistry* **2020,** *92* (15), 10412-10419.

29. Grimme, S.; Bannwarth, C.; Dohm, S.; Hansen, A.; Pisarek, J.; Pracht, P.; Seibert, J.; Neese, F., Fully Automated Quantum-Chemistry-Based Computation of Spin–Spin-Coupled Nuclear Magnetic Resonance Spectra. *Angewandte Chemie International Edition* **2017,** *56* (46), 14763-14769.

30. Blaženović, I.; Kind, T.; Ji, J.; Fiehn, O., Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **2018,** *8* (2).

31. Robinette, S. L.; Bruschweiler, R.; Schroeder, F. C.; Edison, A. S., NMR in metabolomics and natural products research: two sides of the same coin. *Acc Chem Res* **2012,** *45* (2), 288-97.

32. Whiley, L.; Chekmeneva, E.; Berry, D. J.; Jimenez, B.; Yuen, A. H. Y.; Salam, A.; Hussain, H.; Witt, M.; Takats, Z.; Nicholson, J.; Lewis, M. R., Systematic Isolation and Structure Elucidation of Urinary Metabolites Optimized for the Analytical-Scale Molecular Profiling Laboratory. *Anal Chem* **2019,** *91* (14), 8873-8882.

33. Au - Shaver, A. O.; Au - Gouveia, G. J.; Au - Kirby, P. S.; Au - Andersen, E. C.; Au - Edison, A. S., Culture and Assay of Large-Scale Mixed-Stage Caenorhabditis elegans Populations. *JoVE* **2021,** (171), e61453.

34. Gouveia, G. J.; Shaver, A. O.; Garcia, B. M.; Morse, A. M.; Andersen, E. C.; Edison, A. S.; McIntyre, L. M., Long-Term Metabolomics Reference Material. *Anal Chem* **2021,** *93* (26), 9193-9199.

35. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **1995,** *6* (3), 277-93.

36. Johnson, B. A.; Blevins, R. A., NMR View: A computer program for the visualization and analysis of NMR data. *J Biomol NMR* **1994,** *4* (5), 603-14.

37. Delabriere, A.; Warmer, P.; Brennsteiner, V.; Zamboni, N., SLAW: A Scalable and Self-Optimizing Processing Workflow for Untargeted LC-MS. *Analytical Chemistry* **2021,** *93* (45), 15024-15032.

38. Wohlgemuth, G.; Haldiya, P. K.; Willighagen, E.; Kind, T.; Fiehn, O., The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports. *Bioinformatics* **2010,** *26* (20), 2647-2648.

39. Szöcs, E.; Stirling, T.; Scott, E. R.; Scharmüller, A.; Schäfer, R. B., webchem: An R Package to Retrieve Chemical Information from the Web. *Journal of Statistical Software* **2020,** *93* (13), 1 - 17.

40. Conway, J. R.; Lex, A.; Gehlenborg, N., UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **2017,** *33* (18), 2938-2940.

41. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S., ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **2016,** *8* (1), 61.

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* (5), 1282-9.

43. Habra, H.; Kachman, M.; Bullock, K.; Clish, C.; Evans, C. R.; Karnovsky, A., metabCombiner: Paired Untargeted LC-HRMS Metabolomics Feature Matching and Concatenation of Disparately Acquired Data Sets. *Analytical Chemistry* **2021,** *93* (12), 5028-5036.

44. Bingol, K.; Li, D. W.; Bruschweiler-Li, L.; Cabrera, O. A.; Megraw, T.; Zhang, F.; Bruschweiler, R., Unified and isomer-specific NMR metabolomics database for the accurate analysis of (13)C-(1)H HSQC spectra. *ACS Chem Biol* **2015,** *10* (2), 452-9.

45. Ye, T.; Zheng, C.; Zhang, S.; Gowda, G. A.; Vitek, O.; Raftery, D., "Add to subtract": a simple method to remove complex background signals from the 1H nuclear magnetic resonance spectra of mixtures. *Anal Chem* **2012,** *84* (2), 994-1002.

46. Yuan, J.; Zhang, B.; Wang, C.; Bruschweiler, R., Carbohydrate Background Removal in Metabolomics Samples. *Anal Chem* **2018,** *90* (24), 14100-14104.

47. Zangger, K., Pure shift NMR. *Prog Nucl Magn Reson Spectrosc* **2015,** *86-87*, 1-20.

48. Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J., Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **2014,** *48* (4), 2097-8.

49. Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R., Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **2007,** *3* (3), 211-221.

50. Wang, C.; Timari, I.; Zhang, B.; Li, D. W.; Leggett, A.; Amer, A. O.; Bruschweiler-Li, L.; Kopec, R. E.; Bruschweiler, R., COLMAR Lipids Web Server and Ultrahigh-Resolution Methods for Two-Dimensional Nuclear Magnetic Resonance- and Mass Spectrometry-Based Lipidomics. *J Proteome Res* **2020,** *19* (4), 1674-1683.

51. 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 HgCl2 Toxicity. *Journal of Proteome Research* **2006,** *5* (6), 1313-1320.

52. Bonini, P.; Kind, T.; Tsugawa, H.; Barupal, D. K.; Fiehn, O., Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Analytical Chemistry* **2020,** *92* (11), 7515-7522.

53. Bouwmeester, R.; Martens, L.; Degroeve, S., Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction. *Analytical Chemistry* **2019,** *91* (5), 3694-3703.

54. Samaraweera, M. A.; Hall, L. M.; Hill, D. W.; Grant, D. F., Evaluation of an Artificial Neural Network Retention Index Model for Chemical Structure Identification in Nontargeted Metabolomics. *Analytical Chemistry* **2018,** *90* (21), 12752-12760.

55. Soper-Hopper, M. T.; Vandegrift, J.; Baker, E. S.; Fernández, F. M., Metabolite collision cross section prediction without energy-minimized structures. *The Analyst* **2020,** *145* (16), 5414-5418.

56. Borges, R. M.; Colby, S. M.; Das, S.; Edison, A. S.; Fiehn, O.; Kind, T.; Lee, J.; Merrill, A. T.; Merz, K. M.; Metz, T. O.; Nunez, J. R.; Tantillo, D. J.; Wang, L.-P.; Wang, S.; Renslow, R. S., Quantum Chemistry Calculations for Metabolomics. *Chemical Reviews* **2021,** *121* (10), 5633-5670.

CHAPTER 4

^{1}H NMR METABOLOMICS CORROBORATES SERINE HYDROXYMETHYLTRANSFERASE AS THE PRIMARY TARGET OF 2-AMINOACRYLATE IN A RIDA MUTANT OF *SALMONELLA ENTERICA*.[*]

[*] Gouveia GJ*, Borchert AJ*, Edison AS, Downs DM. Proton Nuclear Magnetic Resonance Metabolomics Corroborates Serine Hydroxymethyl-transferase as the Primary Target of 2-Aminoacrylate in a *ridA* Mutant of *Salmonella enterica*. mSystems. 2020 Mar 10;5(2):e00843-19. doi: 10.1128/mSystems.00843-19. PMID: 32156800; PMCID: PMC7065518. *co-first authors

FOREWORD

This chapter is reprinted from Borchert AJ, Gouveia GJ, Edison AS, Downs DM. Proton Nuclear Magnetic Resonance Metabolomics Corroborates Serine Hydroxymethyl-transferase as the Primary Target of 2-Aminoacrylate in a *ridA* Mutant of *Salmonella enterica*. mSystems. 2020 Mar 10;5(2) : e00843-19 and is available at https://journals asm.org/doi/10.1128/mSystems.00843-19. The motivation for this work was driven by the senior authors Diana M. Downs and Arthur S. Edison. My contribution to this work as a co-first author consisted of: (i) develop the experimental design in conjunction with Andrew J Borchert, (ii) carry out metabolite extraction for the samples and media, (iii) design the NMR run order, (iv) setup and run the NMR analysis, (v) process the NMR data (vi) carryout data post-processing steps, (vii) metabolomics data analysis (viii) metabolite identification and generating metabolomics figures and writing the metabolomics sections of the draft, and finally (ix) addressing reviewers comments upon request. The steps carried above were carried out as a part of Dr. Borchert training in metabolomics, as such, most of those steps were carried out with his help. In addition, Dr. Borchert contribution was as follows: (i) generation of bacterial cells and separation and quenching of media and bacterial pellets, (ii) spearhead the design of the experiment, (iii) biochemical interpretation of the metabolomics data (iv) construct the biochemical model figures, growth curves plots and figures, (v) wrote most of the draft and finally manuscript submission and addressed reviewers' comments. The senior authors roles were as follows: Arthur S. Edison and Diana M. Downs reviewed, edited, responded to reviewers and defined the direction and goals of the work. Supplemental materials can be found in APPENDIX C.

ABSTRACT

The reactive intermediate deaminase RidA (EC: 3.5.99.10) is conserved across all domains of life and deaminates reactive enamine species. When *S. enterica ridA* mutants are grown in minimal medium, 2-aminoacrylate (2AA) accumulates, damages several pyridoxal 5'-phosphate (PLP)- dependent enzymes, and elicits an observable growth defect. Genetic studies suggested that damage to serine hydroxymethyltransferase (GlyA), and the resultant depletion of 5,10-methelenetetrahydrofolate (5,10-mTHF), was responsible for the observed growth defect. However, the downstream metabolic consequence from GlyA damage by 2AA remains relatively unexplored. This study sought to use untargeted ^{1}H NMR metabolomics to determine whether the metabolic state of a *S. enterica ridA* mutant was accurately reflected by characterizing growth phenotypes. The data supported the conclusion that metabolic changes in a *ridA* mutant were due to the IlvA-dependent generation of 2AA, and that the majority of these changes were a consequence of damage to GlyA. While many of the shifts in the metabolome of a *ridA* mutant could be explained, changes in some metabolites were not easily modeled, suggesting that additional levels of metabolic complexity remain to be unraveled.

IMPORTANCE

Accumulation of the reactive enamine intermediate, 2-aminoacrylate (2AA), elicits global metabolic stress in many prokaryotes and eukaryotes by simultaneously damaging multiple pyridoxal 5'-phosphate(PLP)-dependent enzymes. This work employed ^{1}H NMR to expand our understanding of the consequence(s) of 2AA stress on metabolite pools and effectively identify the metabolic changes stemming from one damaged target: GlyA. This

study shows that nutrient supplementation during ^{1}H NMR metabolomics experiments can disentangle complex metabolic outcomes stemming from a general metabolic stress. Metabolomics shows great potential to complement classical reductionist approaches to cost-effectively accelerate the rate of progress in expanding our global understanding of metabolic network structure and physiology. To that end, this work demonstrates the utility in implementing nutrient supplementation and genetic perturbation into metabolomics workflows as a means to connect metabolic outputs to physiological phenomena and establish causal relationships.

INTRODUCTION

The metabolic state of the cell at a given time reflects the cumulative result of inputs to the system of cellular metabolism that include but are not limited to, transcription, translation, enzyme activity, metabolic flux. Deconvoluting the role of a specific cellular process in this complex network requires both global and local knowledge, acquired by the integration of multidisciplinary approaches [1, 2]. Metabolomics approaches have been successful in accelerating the elucidation of complex metabolic and physiological states of an organism and help complement biochemical and genetic approaches that can require significant resources and time [1, 3-7]. Metabolomics provides the benefit of providing a snapshot of all metabolic changes in a system without requiring that these shifts produce an observable (growth) phenotype. Integration of large metabolomics datasets with reductionist biochemical genetic analyses is advantageous, since the former allows detection of underlying metabolic shifts caused by genetic or environmental perturbation, while data from the latter provides biological relevancy to frame conclusions. The RidA

paradigm of endogenous metabolic stress provides an opportunity to explore the utility of integrating metabolomics analysis with the biochemical genetic approaches that have defined the framework of this stress.

Pyridoxal 5'-phosphate (PLP)-dependent α,β-eliminases generate reactive enamine species as important reaction intermediates from amino acid substrates. A subset of these α,β-eliminases release reactive enamine intermediates into the cellular milieu. The **r**eactive **i**ntermediate **d**eaminase, RidA, catalyzes the deamination of free enamine species. 2-aminoacrylate (2AA) is a reactive enamine species generated from L-serine by the biosynthetic serine/threonine dehydratase (IlvA; EC: 4.3.1.19) [8]. The absence of RidA in *S. enterica* allows the 2AA produced by IlvA to accumulate and damage multiple other PLP-dependent enzymes, generating a number of detectable mutant phenotypes [8-11]. Relevant to this study, serine hydroxymethyltransferase (GlyA; EC: 2.1.2.1) is the most physiologically significant target for 2AA damage in *S. enterica,* since its damage causes a growth-limiting reduction in 5,10-methylenetetrahydrofolate (5,10-mTHF) [12]. Importantly, exogenous glycine can bypass the 5,10-mTHF limitation and restore growth by allowing 5,10-mTHF production via the glycine cleavage complex (GCV). IlvA is subject to allosteric control by L-isoleucine, thus exogenous L-isoleucine restores growth to a *ridA* mutant by preventing 2AA generation [13-15]. Therefore, isoleucine and glycine supplements provide mechanistically distinct means to restore full growth to an *S. enterica ridA* mutant. With the former, the 2AA stress is eliminated, and with the latter, one impact from the stress is circumvented. A summary of the RidA paradigm for *S. enterica* is provided in Figure 4.1.

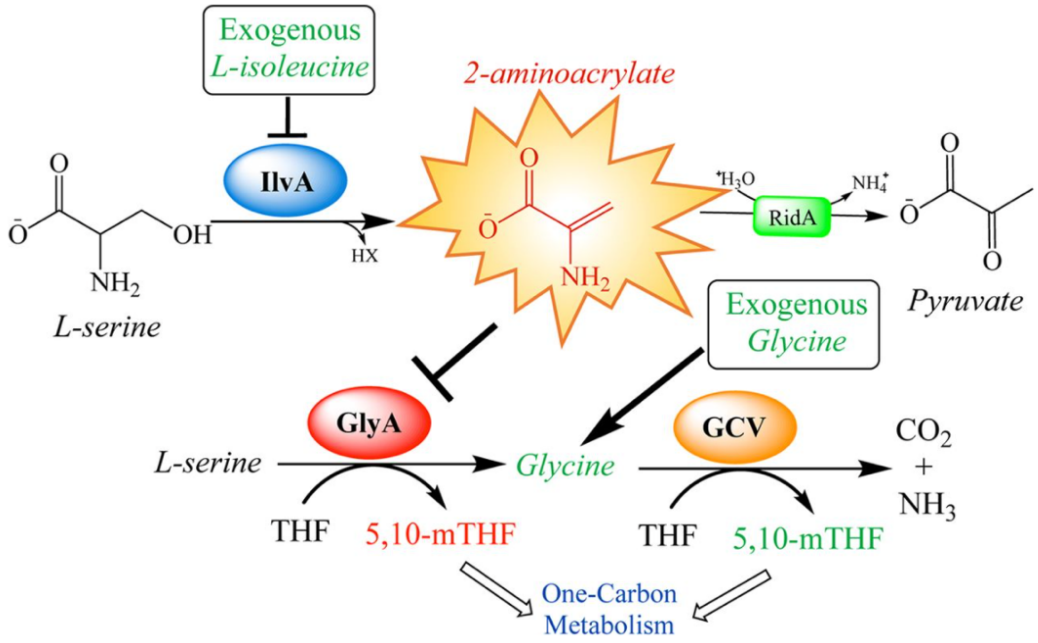


Figure. 4.1 RidA paradigm of 2-aminoacrylate stress in *S. enterica*. Biosynthetic serine/threonine dehydratase (IlvA) catalyzes the β-elimination of l-serine to generate the reactive enamine 2-aminoacrylate (2AA). The activity of IlvA is prevented via allosteric inhibition by l-isoleucine. 2AA is hydrolyzed to pyruvate by the reactive intermediate deaminase A (RidA). In the absence of RidA, 2AA accumulates and can damage a number of PLP-dependent enzymes. The most physiologically sensitive target of 2AA damage in *S. enterica* grown in minimal glucose medium is serine hydroxymethyltransferase (GlyA), as judged by nutrient supplementation (9, 12). GlyA is responsible for the reversible transfer of the hydroxymethyl from serine to tetrahydrofolate (THF), generating glycine and 5,10-methylenetetrahydrofolate (5,10-mTHF). The glycine cleavage complex (GCV) can further catabolize glycine, generating additional 5,10-mTHF.

Damage to GlyA by 2AA perturbs glycine and 5,10-methylenetoetrahydrofolate (5,10-mTHF) synthesis, but the extent of the changes to the global metabolic network caused by this perturbation is less clear. In this study, untargeted proton nuclear magnetic resonance (^{1}H NMR) metabolomics and nutrient supplementation was used to dissect the global metabolic consequences associated with the accumulation of 2AA, extending those

deduced from past growth studies. ^{1}H NMR, was used to measure the endogenous and exogenous (e.g. spent culture media) metabolomes of *S. enterica* wild-type and *ridA* mutant strains in various media. The strengths of NMR-based metabolomics, including simple sample preparation, broad chemical coverage, confident chemical assignments, and straightforward quantification of metabolites, benefited this study [7, 16-18]. The data showed a clear metabolic 'fingerprint' associated with an *S. enterica ridA* mutant grown in minimal glucose medium. Significantly, addition of isoleucine to the growth medium restored the 'fingerprint' to that of the wild-type strain. Further, addition of glycine to the growth medium almost completely moved the *ridA* 'fingerprint' back to that of wildtype, suggesting that the primary impact of 2AA stress is via damaged GlyA. Importantly, this conclusion could not be reached from biochemical genetic data alone. Overall, this work demonstrates the potential for appropriate metabolomics experiments, in combination with biochemical genetic insights, to dissect perturbations to the metabolic network and isolate systems and subsystems impacted by these perturbations.

RESULTS AND DISCUSSION

Metabolic shifts in a ridA mutant are a consequence of IlvA-dependent 2AA generation.

In a *S. enterica ridA* mutant, 2AA accumulates and damages a number of PLP-dependent enzymes, causing a mild growth defect in minimal glucose medium (Fig. 4.2A,B). Both the metabolome and transcriptome of a *ridA* mutant differ from a wild-type strain grown in minimal medium [19, 20]. The growth phenotypes associated with a *S. enterica ridA* mutant result from the accumulation of 2AA, and all RidA orthologs described to date share enamine/imine deaminase activity [8, 11, 21, 22]. Other functions for RidA have been

proposed, including a role as a translational inhibitor/ribonuclease [23-26] or a molecular chaperone [27-29]. Because of the potential for multiple functions for this protein, it was important to determine whether the metabolic restructuring of a *ridA* mutant was caused by 2AA accumulation. The endogenous and exogenous metabolomes of a *ridA* mutant (DM3480) were compared to those obtained for the isogenic parental strain (DM9404, wild-type) in multiple growth conditions using untargeted ^{1}H NMR metabolomics. Principle component analysis (PCA) showed clear visual separation of the metabolomes (both exogenous and endogenous) from the *ridA* mutant and wild-type strains grown in minimal glucose medium, consistent with the growth difference between the two strains (Fig. 4.2C-F).

In a *S. enterica ridA* mutant grown in minimal glucose medium IlvA acts as the dominant, if not sole, generator of 2AA [8]. L-isoleucine allosterically inhibits IlvA, lowering activity, reducing production of 2AA, and restoring wild-type growth to the *ridA* mutant [8] (Fig 4.2A). Presence of L-isoleucine in the medium will therefore eliminate any metabolic effects that are the consequence of 2AA accumulation. The endogenous and exogenous metabolomes of *ridA* and wild-type strains grown in the presence of 1 mM exogenous L-isoleucine were not distinguishable based on PCA analyses (Fig 4.2C, E). These data supported the conclusion that differences found in the metabolomes of the strains in minimal medium were the result of 2AA accumulation. This result was the first to demonstrate that the elimination of detectable *ridA* mutant phenotypes was mirrored by the restoration of metabolite balance. Significantly, this conclusion reinforced that a valid interpretation of the system had been obtained by the biochemical genetic analyses reported previously.

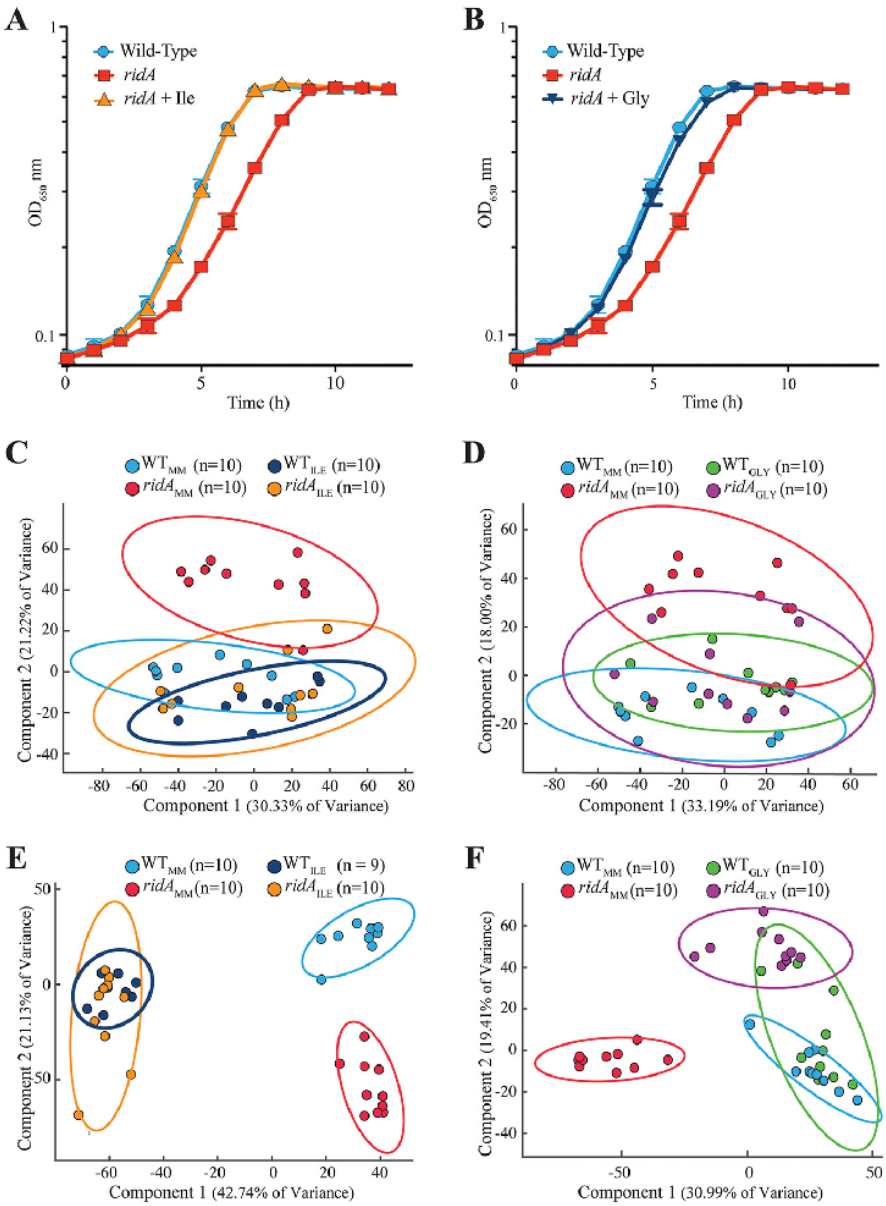


Figure. 4.2 Isoleucine and glycine restore growth and metabolic stability to a ridA mutant. *S. enterica* wild-type and ridA mutant strains were grown at 37°C in minimal glucose (11 mM) medium. Supplementation of the medium with 1mM L-isoleucine (A) or 1mM glycine (B) restored wild-type growth to a ridA mutant. Data are means from three biological replicates, where error bars represent the 95% confidence intervals. OD650, optical density at 650 nm. Principle component analysis (PCA) score plots show separation of endogenous (C and D) and exogenous (E and F) metabolite profiles for *S. enterica* wild-type and ridA mutant strains following 16 h of growth in minimal glucose medium at 37°C. Metabolomes obtained from growth with supplementation with isoleucine (C and E) or glycine (D and F) are shown. Colored ellipses represent the 95% confidence intervals for each group.

Bypassing the one-carbon starvation in a ridA mutant eliminates most, but not all, metabolic changes.

When 2AA accumulates in an *S. enterica ridA* mutant, it damages multiple PLP-dependent enzymes by covalently modifying PLP in the active site. One of the target enzymes is GlyA, whose activity is reduced to ~20% of wild-type in minimal glucose medium [9, 12]. GlyA catalyzes the transfer of the hydroxymethyl group from L-serine substrate to THF, forming glycine and 5,10-mTHF (Fig. 4.1). Damage of GlyA by 2AA causes the growth defect of a *S. enterica ridA* mutant in minimal medium, resulting in its designation as the most physiologically sensitive target in this organism [9, 12]. Addition of glycine to the medium restores growth of a *ridA* mutant since 5,10-mTHF can be generated from glycine by the glycine cleavage system (GCV) and bypasses the need for GlyA [12] (Fig 4.1; Fig 4.2B). Growth of a *ridA* mutant in the presence of glycine was expected to restore a subset of the *ridA* metabolome back to that of wildtype. Further, the metabolites in this subset would define the metabolic subsystem that was perturbed by the reduction or lack of GlyA-dependent formation of glycine/5,10-mTHF. With this logic, the metabolomes of strains grown with glycine were used to distinguish between the metabolic effects resulting from 2AA-dependent damage of GlyA and those resulting from other 2AA-dependent perturbations. PCA analysis showed that, when the cells were grown in the presence of 1 mM glycine, the endogenous metabolome of the *ridA* mutant strain was no longer be distinguishable from that of the wild-type (Fig 4.2D). Surprisingly, these data indicated that the majority of the metabolic perturbations, at least those detected by ^{1}H NMR, in the endogenous metabolome of a *ridA* mutant strain were downstream effects of the damage to GlyA. Consistently, PCA analysis of the external metabolome of a *ridA*

mutant grown with glycine and the corresponding metabolome from wild-type showed an obvious decrease in the separation of metabolic signatures, although the separation was not eliminated, as it was with isoleucine.

In total, the data in Figure 4.2 supported the conclusions that (i) ^{1}H NMR metabolomics detected a molecular signature unique to an *S. enterica ridA* mutant strain, (ii) the deviation in molecular signatures between *ridA* and wild-type strains was dependent upon the generation of 2AA by IlvA, and (iii), the metabolic consequences of 2AA-dependent GlyA damage dominated the differences detected between the *ridA* and wild-type strains. The latter conclusion suggests metabolomic data did not detect all metabolic changes present, possibly due to (i) lack of spectral resolution and/or sensitivity in the data set, or (ii) a limited impact other targets (i.e., IlvE) had on the overall metabolic profile.

2AA stress influences amino acid metabolism and mixed acid fermentation.

In total, sixteen endogenous and ten exogenous metabolites were identified from the NMR spectra and their patterns were considered in the context of *S. enterica* physiology (Supplementary Material File S1). Partial least squares discriminant analysis (PLS-DA) scores plots were used to identify the specific metabolic pathways perturbed by 2AA stress and further understand the metabolic differences between *ridA* mutant and wild-type strains when grown in minimal medium (Fig. 4.3A,C). Variable importance in projections (VIP) scores were determined for all NMR features in the endogenous and exogenous datasets that contributed most to the PLS-DA separation in the first component (Supplementary Material File S2) [30]. From these data, VIP plots were created with the identified metabolites that contributed to separation of endogenous (Fig. 4.3B) and exogenous (Fig. 4.3D)

metabolomes. Importantly, the VIP score reported in these plots represent the average of VIP scores taken from all the features comprising a given metabolite. VIP scores > 1 indicated that elevated threonine, valine, N-acetyl putrescine, glutamine, phenylalanine, alanine, glutamate, and pyruvate, and diminished ethanolamine, formate, putrescine, and coenzyme A (CoA) drove PLS-DA separation of *ridA* endogenous metabolomes (Fig. 4.3B). Similarly, VIP analysis revealed that elevated lactate, valine, putrescine, 2-isopropylmalate, and acetyl-phosphate, and diminished formate, uracil, and 2-aminobutyrate, drove PLS-DA separation of *ridA* exogenous metabolomes (Fig. 4.3D). In total, these data indicated that metabolites in amino acid metabolism and mixed acid fermentation were largely responsible for the separation of the metabolomes, as determined by PLS-DA analysis. Integration of peaks corresponding to all identifiable metabolites and comparison by Student's unpaired two-samples *t*-test showed that 12 of 16 endogenous metabolites and 8 of 10 exogenous metabolites were significantly altered in a *ridA* mutant strain (q-value < 0.1, Supplementary Material File S4.3, Fig. 4.4).

Integration of the peaks described above from the samples grown in minimal glucose medium supplemented with isoleucine showed that concentrations of the altered metabolites were restored to wild-type levels. (Supplementary Material File S4.3, Fig. 4.4). The only exception was of the abundance of exogenous uracil, which was 3.1-fold lower in a *ridA* mutant compared to wild-type following growth in minimal glucose medium, but appeared 1.3-fold elevated when grown in minimal glucose medium containing isoleucine. This discrepancy might be a consequence of the fact that the doublet integrated to define uracil concentration was poorly resolved in three of the *ridA* samples from growth with isoleucine (data not shown). Nonetheless, isoleucine clearly reversed the metabolic shifts

observed for a *ridA* mutant grown in minimal glucose medium. These results further supported the conclusion from the PCA plots that the metabolic perturbations detected with ^{1}H NMR were the consequence of IlvA-dependent 2AA stress.

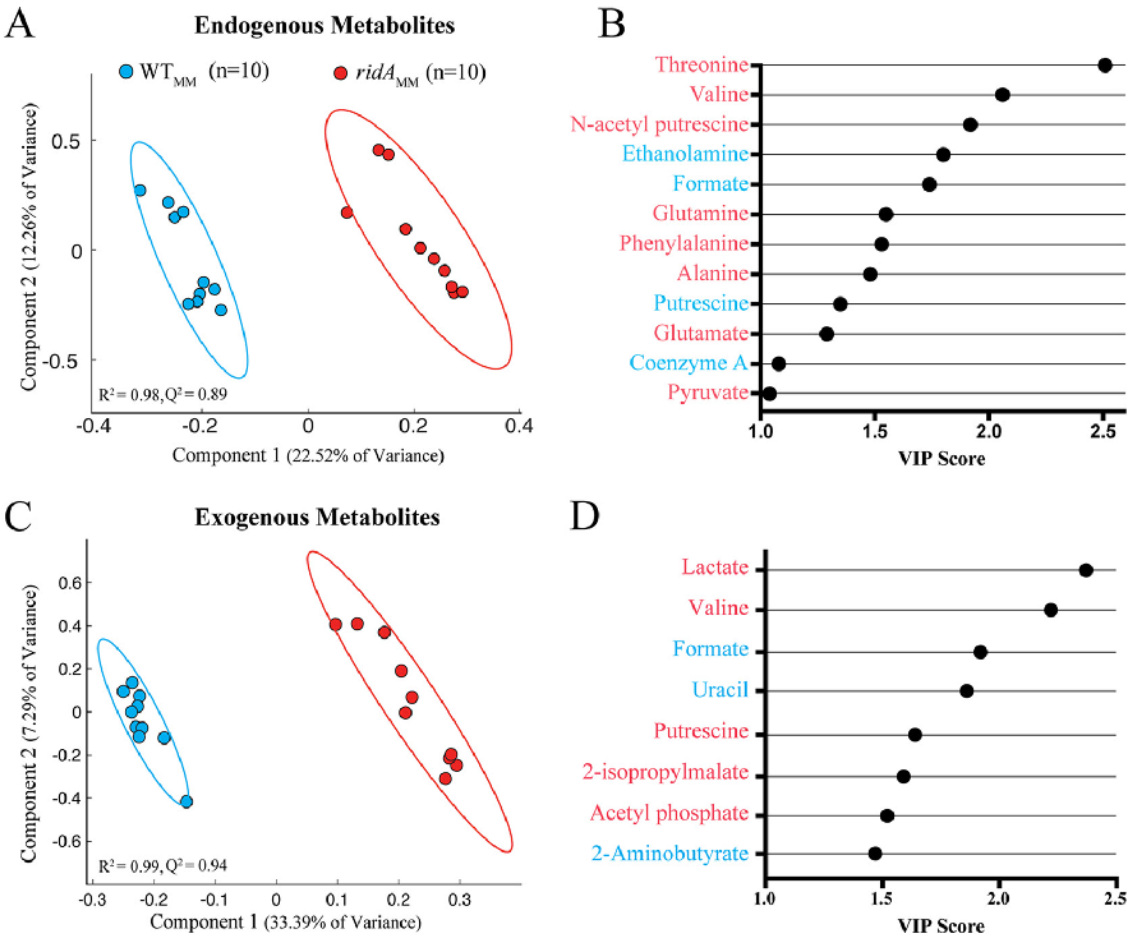


Figure.4.3 Partial least-squares discriminant analysis (PLS-DA) highlights separation of metabolomic profiles. PLS-DA score plots show clear separation by PLS component 1 of wild-type (n=10, blue) and ridA mutant (n=10, red) endogenous (A) and exogenous (B) metabolite samples following growth in minimal glucose medium. Variable importance of projection (VIP)scores were plotted for the metabolites that contributed significantly (VIP of <1) to separation by PLS-DA component 1 for endogenous (C) and exogenous (D) samples. VIP scores were determined as the average from all peak VIP scores belonging to the given metabolite. Metabolites colored blue were elevated in wild-type samples, while those colored red were elevated in ridA mutant samples.

CoA limitation in a ridA mutant is captured by untargeted ^{1}H NMR metabolomics.

During growth on minimal glucose medium, *S. enterica* derives most of its one-carbon units from serine via generation of 5,10-mTHF by the PLP-dependent enzyme, GlyA [31]. GlyA activity is decreased more than five-fold in a *ridA* mutant, when compared to wild-type, as a result of damage by 2AA [9]. The constraint on GlyA activity leads to significantly decreased CoA (3-fold) in a *ridA* mutant, since the biosynthesis of CoA involves the 5,10-mTHF-dependent enzyme 3-methyl-2-oxobutanoate hydroxymethyltransferase (PanB; EC: 2.1.2.11) [9]. Gratifyingly, the untargeted metabolomic experiments herein captured the lowered CoA levels in a *ridA* mutant (Fig. 4.4). Furthermore, addition of glycine to the growth medium eliminated the difference in CoA levels between the *ridA* mutant and wild-type (Fig. 4.4).

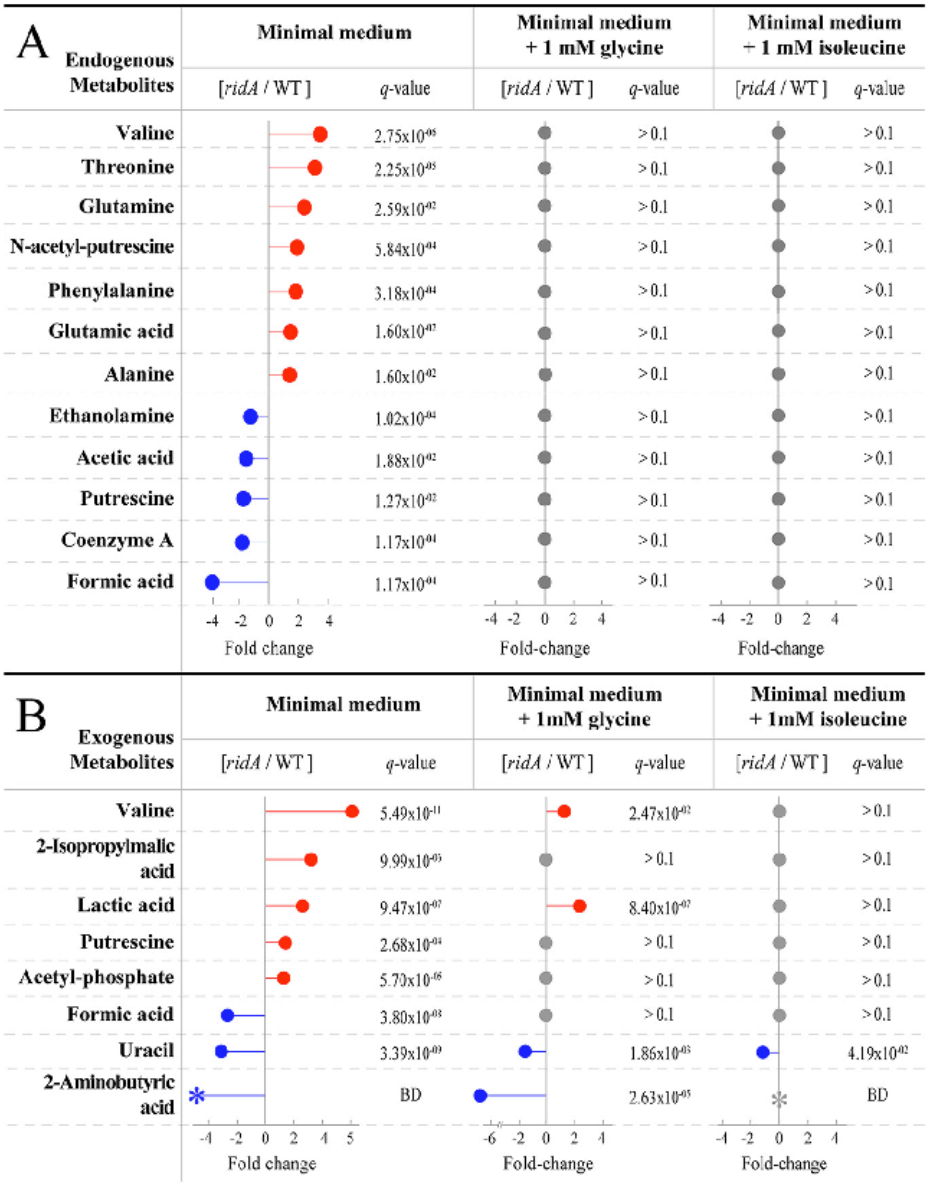


Figure. 4.4 Significantly altered metabolites under different conditions. Fold change differences between ridA mutant and wild-type strains were calculated for both intracellular metabolites (A) and external metabolites (B). Red circles indicate higher abundance in ridA mutants, blue circles indicate high abundance in the wild type, and gray circles indicate no significant (q value of <0.1) difference in metabolite abundance between ridA mutants and wild-type samples. q values represent false discovery ratecorrected P values according to the Benjamini-Hochberg method (51). Colored asterisks, with corresponding "BD" q value designation, specify that a fold change was not determined, since the feature in ridA (blue) or both (gray) groups was below the detection (BD) limit for multiple samples.

2AA-dependent decrease in 5,10-mTHF generates additional metabolic effects.

Glycine supplementation eliminated most of the metabolic shifts observed for a *ridA* mutant, supporting the conclusion that these changes were a consequence of the constrained function of GlyA. The metabolites restored to wild-type levels included all of the endogenous metabolites of known identity: valine, threonine, glutamine, N-acetyl putrescine, phenylalanine, glutamic acid, alanine, ethanolamine, acetic acid, CoA, putrescine and formic acid (Fig. 4.4A), as well as exogenous 2-isopropylmalic acid, putrescine, acetyl-phosphate, and formic acid (Fig. 4.4B). While the majority of metabolic pools were restored to balance, a few exogenous metabolites were not, notably valine and uracil. The discrepancy between *ridA* mutant and wild-type valine and uracil content was partially reduced (p-value <0.01) following the addition of exogenous glycine (5.0-fold higher to 1.4-fold higher and 3.1-fold lower to 1.5-fold lower in the *ridA* background for valine and uracil, respectively), suggesting that the state of 5,10-mTHF was partially responsible for the concentration shift (Supplementary Material File S3, Fig. 4.4).

A *ridA* mutant accumulated less endogenous formic acid and acetic acid than the wild-type strain. The finding that these trends were eliminated by the presence of glycine in the growth medium suggested a model in which CoA limitation triggered a shift in flux through mixed acid fermentation (Fig. 4.5). During mixed acid fermentation, pyruvate-formate lyase (PflB: EC: 2.3.1.54) uses CoA and pyruvate as substrates for the production of formate and acetyl-CoA, which is further processed to acetate [32]. A bottleneck in CoA biosynthesis would reduce PflB-dependent formation of formate and downstream production of acetate. The accumulation of pyruvate during late exponential phase growth in a *ridA* mutant is consistent with this model [9]. An increase in endogenous pyruvate in a

ridA mutant was suggested by the PLS-DA model; however, the increase did not meet the threshold for statistical significance (*q*-value = 0.12) and thus was not discussed (Supplementary Material File S3). Two molecules of pyruvate are used during valine synthesis, and in *Klebsiella aerogenes* pyruvate accumulation increases valine production [33, 34]. The increase in endogenous and exogenous valine and exogenous 2-isopropylmalic acid, which is formed from an intermediate in valine biosynthesis, in a *ridA* mutant is eliminated by growth in glycine (Fig. 4.4B). Therefore, the increase in valine and 2-isopropylmalic acid content may indicate that overflow pyruvate in a *ridA* mutant is diverted toward valine synthesis (Fig. 4.4, Fig. 4.5).

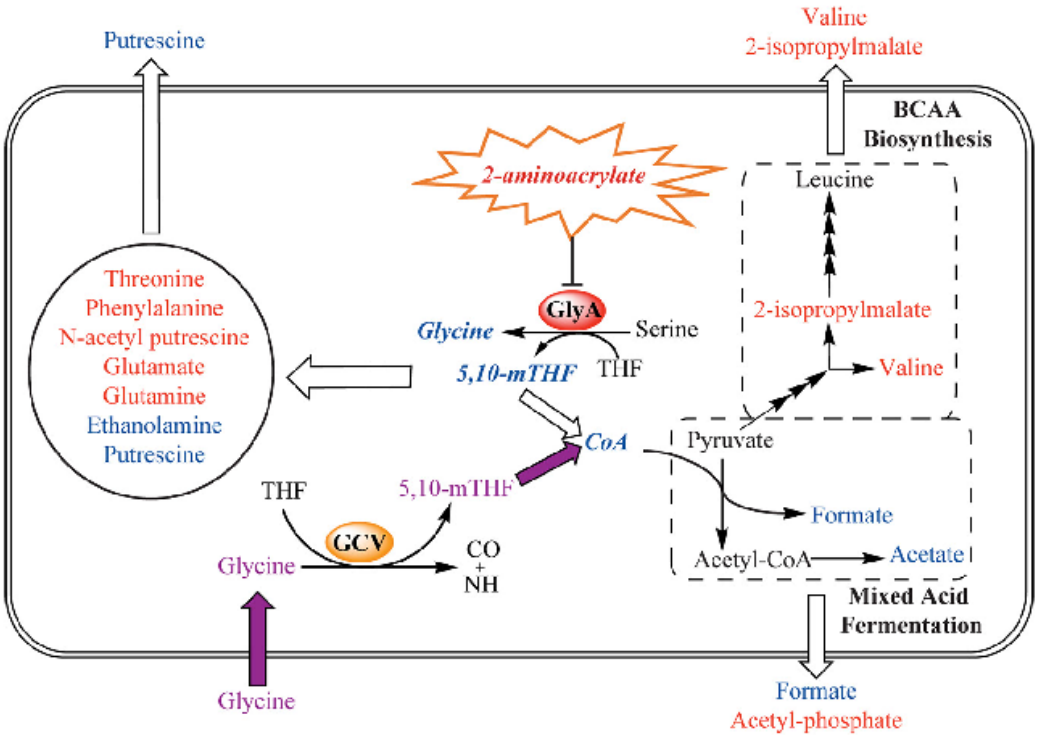


Figure. 4.5 Working model for metabolic outcomes resulting from 5,10-mTHF limitation. 2AA-dependent damage of GlyA in a ridA mutant of *S. enterica* causes a glycine/5,10-mTHF limitation. The 5,10-mTHF limitation leads to a downstream decrease in CoA, altering flux through CoA-dependent mixed acid fermentation. In this working model, accumulated pyruvate is rerouted toward branched-chain amino acid (BCAA) biosynthesis. The metabolites on the left side of the cell schematic are altered by an unknown, but glycine/5,10-mTHF-dependent, mechanism. Metabolites colored blue are elevated in the wild type, and those colored red are elevated in the *ridA* mutant. The purple pathway represents the effect of supplemented glycine in restoring endogenous glycine/5,10-mTHF levels. Metabolites that are boldfaced and italicized were not detected by ^{1}H NMR. Compounds listed outside the rounded rectangle represent metabolites detected by ^{1}H NMR in the growth medium.

CONCLUSION

The untargeted ^{1}H NMR metabolomics approach used here exposed the global metabolic consequences of eliminating RidA from *S. enterica*. Both endogenous and exogenous metabolomes were assessed and among the multitude of features visualized, 16

endogenous and 10 exogenous metabolites were confidently identified. Multivariate analysis by PCA and PLS-DA showed a clear difference between the metabolomes of a *ridA* and wild-type strain grown in minimal medium. Importantly, the PLS-DA models revealed a 'fingerprint' for a *ridA* metabolome and the high Q^2 scores from cross-validation showed that these models effectively captured the metabolic separation between the two groups. A VIP score cutoff >1 was used to identify metabolites contributing to separation of the respective 'fingerprints'. Findings from VIP analysis agreed well with the findings from univariate analysis of the ^{1}H NMR dataset, as only endogenous pyruvate had a VIP score >1 but did not differ significantly by univariate analysis (q-value > 0.1) and only endogenous acetate differed significantly by univariate analysis but failed to meet a VIP score > 1. Altogether, these data indicated the two PLS-DA models, without the need for orthogonal signal correction [35], accounted for most identifiable and significantly altered metabolites and effectively separated wild-type metabolomes from those associated with *ridA* mutants.

The ^{1}H NMR analyses of strains grown in medium containing isoleucine demonstrated that concentration shifts of metabolites between the *ridA* and wild-type strains are due to the IlvA-dependent generation of 2AA. This result significantly extended our understanding of the influence of 2AA on the metabolic network of a *ridA* mutant by showing the restoration of all metabolic feature discrepancies, including those that were not associated with a detectable growth phenotype. The growth defect of a *ridA* mutant is a consequence of damage to GlyA by 2AA [9, 12]. The ^{1}H NMR data showed decreased PCA separation between wild-type and *ridA* mutant metabolomes, when the damage to GlyA was bypassed by glycine addition to the medium. The glycine-dependent restoration of

wild-type levels for 12/12 endogenous and 4/8 exogenous identifiable metabolites supported a metabolic model connecting GlyA damage by 2AA to shifts in mixed-acid fermentation and BCAA metabolism, as depicted in Figure 4.5. The ^{1}H NMR data found metabolic differences that were not easily modeled as a consequence of lowered 5,10-mTHF or CoA. Further dissection of the network generating these changes could incorporate the supplementation of growth medium with pantothenate, which restores CoA levels but not 5,10-mTHF levels [9].

Since fewer metabolites are present in spent media samples, the spectra associated with the exogenous metabolome had less spectral overlap. Media samples also did not require homogenization or methanol extraction, making their processing more expedient and straightforward. Therefore, analysis of exogenous samples may offer a high-throughput and simplified way to continue characterization of the RidA paradigm. Such simplified and expedited analysis would be particularly useful during time-course experiments or studies containing more genetic backgrounds and/or media conditions, where dozens to hundreds of samples may be required.

Overall, the combination of ^{1}H NMR metabolomics and relevant nutrient supplementation was successful in expanding the RidA/2-aminoacrylate paradigm in *S. enterica* and in making first steps toward delineating downstream consequences of GlyA damage from metabolic effects independent of glycine/5,10-mTHF perturbation. Historically, metabolomics approaches have been valuable in identifying correlations to generate hypotheses/models; however, the design of metabolomics experiments to act as a high-throughput means of testing these models and identifying mechanistic/causal relationships is a nascent field [36, 37]. The RidA system provides an interesting case-study in

refining the experimental approach of metabolomics studies to include genetic manipulations and nutrient supplementations as a means to probe the underlying factors contributing to the metabolic shifts observed following a perturbation. The study herein highlighted, i) the benefit of using spent growth medium as an initial proxy for metabolome differences, and (ii) the value of nutritional supplementation as a way to help define metabolic sub-networks. If applied to the study of complex metabolic systems, these approaches have the potential to contribute to understanding how a network responds to perturbation, and drive our understanding of gene function and the physiological impact of various cellular components.

MATERIALS AND METHODS

Bacterial strains, chemicals, and media.

Strains used in this work are derivatives of *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2. Construction of the *ridA* null mutant (DM3480, *ridA3*::MudJ1734) is described elsewhere [38, 39]. The wild-type strain used in this work (DM9404) is an LT2 derivative isogenic to the *ridA* mutant. Rich medium was Difco nutrient broth (NB) (8 g/L) supplemented with 5 g/L NaCl. No-carbon E medium (NCE) containing 1 mM MgSO4 [40], trace metals [41], and 11 mM D-glucose was the designated minimal medium. Minimal medium was supplemented with 1 mM L-isoleucine or 1 mM glycine, as indicated. All chemicals were purchased from the Sigma-Aldrich Chemical Company (St. Louis, MO).

Generation of cell pellets and spent media.

Ten biologically independent wild-type and *ridA* mutant cultures were grown overnight in NB medium shaking at 37 °C and used to inoculate (1% inoculum) 250 mL non-baffled flasks holding 125 mL of medium. Each culture inoculated one each of the three media types (minimal medium, minimal medium with 1 mM L-isoleucine, and minimal medium with 1 mM glycine), for a total of 60 flasks. Flasks were randomly arranged in an Innova®44 incubator and cultures allowed to grow 16 h at 37 °C, shaking at 180 RPM. Cultures were chilled 5 min on ice and then harvested by centrifugation at 7,000 x G for 10 min at 4 °C. The supernatant was decanted, with 10 mL transferred to sterile 15 mL conical tubes and flash-frozen using liquid nitrogen for downstream analysis of external metabolites. Cell pellets were transferred to sterile 15 mL conical tubes after resuspension in 10 mL ddH_2O, prior to a second pelleting at 7,000 x G for 10 min at 4 °C. The supernatant was decanted and pellets were flash-frozen using liquid nitrogen and stored at -80 °C.

Preparation of growth medium samples.

Spent media from each bacterial culture was lyophilized (VirTis Benchtop K) for 48 h. Once dry, each lyophilized sample was reconstituted in 150 μL of 100 mM sodium phosphate buffer (Cambridge Isotope Laboratories), pH 7.0, containing 1/3 mM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid, Cambridge Isotope Laboratories) as an internal standard. Each sample was centrifuged at 20,000 X G for 30 min and 50 μL of supernatant was transferred by a Bruker SamplePro liquid handler into 1.7 mm SampleJet NMR tubes (Bruker Biospin).

Metabolite extraction from bacterial pellets.

Each frozen bacterial pellet was thawed on ice and 1 mL of ice cold 80/20 methanol/water together with approximately 200 mL of 0.7mm silica beads (BioSpec products). Homogenization was carried out using a FastPrep 96 (MPBIO). The samples and extraction blanks went through three cycles of homogenization at 1800 rpm for 300 s each. At the end of each cycle samples and controls were centrifuged at 20000 x G for 30 min. Each supernatant was transferred to a new tube and 1 mL of ice-cold methanol/water added to the original tubes before each new cycle. The combined supernatants from each cycle were pooled and concentrated overnight using a CentriVap Benchtop Vacuum Concentrator (Labconco) down to 0.1 mL. The samples were then diluted with 0.5 mL of methanol/water and transferred into 0.6 mL centrifuge tube and concentrated to dryness. The extracts were reconstituted in 150 μL of deuterated 100 mM sodium phosphate buffer containing 1/3 mM of the internal standard DSS (d6 4,4-dimethyl-4-silapentane-1-sulfonic acid) at pH 7.0 and vortex mixed for 5 min. Each sample was centrifuged at 20000 x G for 30 min and transferred by a Bruker SamplePro liquid handler into 1.7 mm SampleJet NMR tubes. Extraction blanks were prepared following the same procedure except the biological material was replaced with an equal volume of water. Solvent blanks consisted of the reconstituting NMR buffer (deuterated sodium phosphate buffer with DSS).

Acquisition and processing of NMR data.

One-dimensional ^{1}H NMR spectra for each sample and blanks were acquired using an optimized PROJECT (periodic refocusing of *J* evolution by coherence transfer) pulse sequence [42] on an Avance III HD 600 MHz Bruker NMR spectrometer equipped with a

TCI cryoprobe and a Bruker SampleJet autosampler cooled to 5.6 °C. During acquisition, 32,768 complex datapoints were collected for the FID, using 64 scans with 16 additional dummy scans. The spectral width was 20 ppm. A Fourier transform (FT), a polynomial baseline correction of order 3, a 2 Hz line broadening and phase correction were applied to each spectrum.

Two-dimensional ^{1}H-^{1}H total correlation spectroscopy (TOCSY), ^{1}H-^{13}C heteronuclear single quantum correlation (HSQC) and ^{1}H-^{13}C HSQC–total correlation spectroscopy (HSQC–TOCSY) experiments were collected on pooled samples, composed from a small aliquot of each study sample, for metabolite identification. During acquisition, all three experiments were collected for 32 scans and an additional 16 dummy scans, with 512 and 1,024 datapoints recorded on the direct and indirect dimensions respectively and, a spectral width of 200 ppm for ^{13}C and 12 ppm for ^{1}H. A 90 ms mixing time was used for both HSQC-TOCSY and TOCSY experiments. All spectral processing was carried out using NMRpipe [43]. Spectra referencing, baseline correction, and statistical analysis were carried out using in-house Matlab (Mathworks, R2019a) scripts which are publicly available. (https://github.com/artedison/Edison _Lab_ Shared_ Metabolomics_UGA).

Compound identification/database matching.

All three two-dimensional experiments were used for spectral matching against the BBiorefcode library using COLMARm [44] using a chemical shift cutoff of 0.03 and 0.3 ppm for ^{1}H and ^{13}C respectively. A total of 9 exogenous and 16 endogenous metabolites that could be integrated without overlapping features in their respective 1D ^{1}HNMR spectra were identified. A confidence level ranging from 1 to 5, was assigned to each metabolite

(Supplementary Material File S1) as described elsewhere [45]. Briefly this scale is defined as: (1) putatively characterized compound, (2) matched reported 1D spectra, (3) matched reported HSQC spectra, (4) matched reported HSQC and HSQC-TOCSY spectra, and (5) validated by spiking putative compound into sample.

Statistical analysis.

The data were normalized using probabilistic quotient normalization (PQN) and range-scaled before multivariate statistical analysis [46, 47]. The principal component analysis (PCA) scores were calculated using the NIPALS algorithm [48]. The partial least squares discriminant analysis (PLS-DA) using the SIMPLS algorithm was conducted with a 5-fold cross-validation and 30 permutations [49]. Goodness of prediction (Q^2) for the PLS-DA model was obtained and the model was used to identify features that differed between the wild-type and *ridA* mutant for both endogenous and exogenous datasets [50]. Univariate statistics were performed using PQN-normalized 1D ^{1}H NMR data for metabolites whose features could be integrated without the presence of overlapping features. Student's *t*-test with a Benjamini-Hochberg false discovery rate (FDR)-correction [51] was used to determine metabolites that differed significantly (q-value < 0.1) between wild-type and *ridA* mutant samples. All raw and processed data are available on the Metabolomics Workbench (www.metabolomicsworkbench.org), along with detailed experimental NMR and statistical analysis methods.

Acknowledgements

This work was supported by National Institutes of Health (GM095837 to D.M.D.) and the National Science Foundation (MCB1615373 to D.M.D.). The Georgia Research Alliance partially supported ASE.

REFERENCES

1. Downs, D.; V. Bazurto, J.; Gupta, A.; L. Fonseca, L.; O. Voit, E., The three-legged stool of understanding metabolism: integrating metabolomics with biochemical genetics and computational modeling. *AIMS Microbiology* **2018,** *4* (2), 289-303.

2. Johnson, C. H.; Ivanisevic, J.; Siuzdak, G., Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol* **2016,** *17* (7), 451-9.

3. Walejko, J. M.; Antolic, A.; Koelmel, J. P.; Garrett, T. J.; Edison, A. S.; Keller-Wood, M., Chronic maternal cortisol excess during late gestation leads to metabolic alterations in the newborn heart. *Am. J. Physiol. Endocrinol. Metab.* **2019,** *316* (3), E546-E556.

4. Lee-McMullen, B.; Chrzanowski, S. M.; Vohra, R.; Forbes, S. C.; Vandenborne, K.; Edison, A. S.; Walter, G. A., Age-dependent changes in metabolite profile and lipid saturation in dystrophic mice. *NMR Biomed.* **2019**, e4075.

5. Hattori, A.; Tsunoda, M.; Konuma, T.; Kobayashi, M.; Nagy, T.; Glushka, J.; Tayyari, F.; McSkimming, D.; Kannan, N.; Tojo, A.; Edison, A. S.; Ito, T., Cancer progression by reprogrammed BCAA metabolism in myeloid leukaemia. *Nature* **2017,** *545* (7655), 500-504.

6. Benjamin, D. I.; Cravatt, B. F.; Nomura, D. K., Global profiling strategies for mapping dysregulated metabolic pathways in cancer. *Cell Metab.* **2012,** *16* (5), 565-77.

7. Markley, J. L.; Bruschweiler, R.; Edison, A. S.; Eghbalnia, H. R.; Powers, R.; Raftery, D.; Wishart, D. S., The future of NMR-based metabolomics. *Curr Opin Biotechnol* **2017,** *43*, 34-40.

8. Lambrecht, J. A.; Schmitz, G. E.; Downs, D. M., RidA proteins prevent metabolic damage inflicted by PLP-dependent dehydratases in all domains of life. *mBio* **2013,** *4* (1), e00033-13-e00033-13.

9. Flynn, J. M.; Christopherson, M. R.; Downs, D. M., Decreased coenzyme A levels in *ridA* mutant strains of *Salmonella enterica* result from inactivated serine hydroxymethyltransferase: *ridA* mutants are deficient in one carbon metabolism. *Mol. Microbiol.* **2013,** *89* (4), 751-759.

10. Flynn, J. M.; Downs, D. M., In the absence of RidA, endogenous 2-aminoacrylate inactivates alanine racemases by modifying the pyridoxal 5'-phosphate cofactor. *J. Bacteriol.* **2013,** *195* (16), 3603-3609.

11. Borchert, A. J.; Downs, D. M., The response to 2-aminoacrylate differs in *Escherichia coli* and *Salmonella enterica*, despite shared metabolic components. *J. Bacteriol.* **2017,** *199* (14), e00140-17.

12. Ernst, D. C.; Downs, D. M., 2-aminoacrylate stress induces a context-dependent glycine requirement in *ridA* strains of *Salmonella enterica*. *J. Bacteriol.* **2016,** *198* (3), 536-543.

13. LaRossa, R. A.; Van Dyk, T. K., Metabolic mayhem caused by 2-ketoacid imbalances. *Bioessays* **1987,** *7*, 125-130.

14. Umbarger, H. E.; Brown, B., Threonine deamination in *Escherichia coli* ii.: Evidence for two l-threonine deaminases. *J. Bacteriol.* **1957,** *73* (1), 105.

15. Enos-Berlage, J. L.; Langendorf, M. J.; Downs, D. M., Complex metabolic phenotypes caused by a mutation in *yjgF*, encoding a member of the highly conserved YER057c/YjgF family of proteins. *J. Bacteriol.* **1998,** *180* (24), 6519-6528.

16. Aretz, I.; Meierhofer, D., Advantages and pitfalls of mass spectrometry based metabolome profiling in systems biology. *Int. J. Mol. Sci.* **2016,** *17* (5).

17. Nagana Gowda, G. A.; Gowda, Y. N.; Raftery, D., Expanding the limits of human blood metabolite quantitation using NMR spectroscopy. *Anal. Chem.* **2015,** *87* (1), 706-15.

18. Nagana Gowda, G. A.; Raftery, D., Can NMR solve some significant challenges in metabolomics? *J. Magn. Reson.* **2015,** *260*, 144-60.

19. Borchert, A. J.; Walejko, J. M.; Le Guennec, A.; Ernst, D. C.; Edison, A. S.; Downs, D. M., Integrated metabolomics and transcriptomics suggest the global metabolic response to 2-aminoacrylate stress in *Salmonella enterica*. *Submitted* **2019**.

20. Borchert, A. J.; Downs, D. M., Endogenously generated 2-aminoacrylate inhibits motility in *Salmonella enterica*. *Sci. Rep.* **2017,** *7* (1).

21. Niehaus, T. D.; Nguyen, T. N. D.; Gidda, S. K.; ElBadawi-Sidhu, M.; Lambrecht, J. A.; McCarty, D. R.; Downs, D. M.; Cooper, A. J. L.; Fiehn, O.; Mullen, R. T.; Hanson, A. D., *Arabidopsis* and maize RidA proteins preempt reactive enamine/imine damage to branched-chain amino acid biosynthesis in plastids. *Plant Cell* **2014,** *26* (7), 3010-3022.

22. ElRamlawy, K. G.; Fujimura, T.; Baba, K.; Kim, J. W.; Kawamoto, C.; Isobe, T.; Abe, T.; Hodge-Hanson, K.; Downs, D. M.; Refaat, I. H.; Beshr Al-Azhary, D.; Aki, T.; Asaoku, Y.; Hayashi, T.; Katsutani, T.; Tsuboi, S.; Ono, K.; Kawamoto, S., Der f 34, a novel major house dust mite allergen belonging to a highly conserved Rid/YjgF/YER057c/UK114 family of imine deaminases. *J. Biol. Chem.* **2016,** *291* (41), 21607-21615.

23. Oka, T.; Tsuji, H.; Noda, C.; Sakai, K.; Hong, Y. M.; Suzuki, I.; Munoz, S.; Natori, Y., Isolation and characterization of a novel perchloric acid-soluble protein inhibiting cell-free protein synthesis. *J. Biol. Chem.* **1995,** *270* (50), 30060-7.

24. Morishita, R.; Kawagoshi, A.; Sawasaki, T.; Madin, K.; Ogasawara, T.; Oka, T.; Endo, Y., Ribonuclease activity of rat liver perchloric acid-soluble protein, a potent inhibitor of protein synthesis. *J. Biol. Chem.* **1999,** *274* (29), 20688-92.

25. Schmiedeknecht, G.; Kerkhoff, C.; Orso, E.; Stohr, J.; Aslanidis, C.; Nagy, G. M.; Knuechel, R.; Schmitz, G., Isolation and characterization of a 14.5-kDa trichloroacetic-acid-soluble translational inhibitor protein from human monocytes that is upregulated upon cellular differentiation. *Eur. J. Biochem.* **1996,** *242* (2), 339-51.

26. Walejko, J. M.; Kim, S.; Goel, R.; Handberg, E. M.; Richards, E. M.; Pepine, C. J.; Raizada, M. K., Gut microbiota and serum metabolite differences in African Americans and White Americans with high blood pressure. *Int. J. Cardiol.* **2018,** *271*, 336-339.

27. Samuel, S. J.; Tzung, S. P.; Cohen, S. A., Hrp12, a novel heat-responsive, tissue-specific, phosphorylated protein isolated from mouse liver. *Hepatology* **1997,** *25* (5), 1213-22.

28. Farkas, A.; Nardai, G.; Csermely, P.; Tompa, P.; Friedrich, P., DUK114, the *Drosophila* orthologue of bovine brain calpain activator protein, is a molecular chaperone. *Biochem. J.* **2004,** *383* (Pt 1), 165-70.

29. Müller, A.; Langklotz, S.; Lupilova, N.; Kuhlmann, K.; Bandow, J. E.; Leichert, L. I. O., Activation of RidA chaperone function by N-chlorination. *Nat. Commun.* **2014,** *5*, 5804.

30. Chong, I. G.; Jun, C. H., Performance of some variable selection methods when multicollinearity is present. *Chemometrics Intellig. Lab. Syst.* **2005,** *78* (1-2), 103-112.

31. Green, J. M.; Nichols, B. P.; Matthews, R. G., Folate biosynthesis, reduction, and polyglutamylation. In *Escherichia coli and Salmonella Cellular and Molecular Biology*, Neidhart, F. C., Ed. ASM Press: Washington DC, 1996; Vol. 1, pp 665-673.

32. Trotter, E. W.; Rolfe, M. D.; Hounslow, A. M.; Craven, C. J.; Williamson, M. P.; Sanguinetti, G.; Poole, R. K.; Green, J., Reprogramming of *Escherichia coli* K-12 metabolism during the initial phase of transition from an anaerobic to a micro-aerobic environment. *PLoS ONE* **2011,** *6* (9), e25501.

33. Webb, M., Aminopterin inhibition in *Aerobacter aerogenes*: alanine and valine accumulation during the inhibition and their utilization on recovery. *Biochem. J.* **1958,** *70* (3), 472-489.

34. Webb, M., Pyruvate accumulation in growth-inhibited cultures of *Aerobacter aerogenes*. *Biochem. J.* **1968,** *106* (2), 375-380.

35. Vu, T.; Siemek, P.; Bhinderwala, F.; Xu, Y. H.; Powers, R., Evaluation of Multivariate Classification Models for Analyzing NMR Metabolomics Data. *J. Proteome Res.* **2019,** *18* (9), 3282-3294.

36. Bizzarri, M.; Brash, D. E.; Briscoe, J.; Grieneisen, V. A.; Stern, C. D.; Levin, M., A call for a better understanding of causation in cell biology. *Nat. Rev. Mol. Cell Biol.* **2019,** *20* (5), 261-262.

37. Stern, C. D., The 'Omics Revolution: How an Obsession with Compiling Lists Is Threatening the Ancient Art of Experimental Design. *Bioessays* **2019**.

38. Schmitz, G.; Downs, D. M., Reduced transaminase B (IlvE) activity caused by the lack of *yjgF* is dependent on the status of threonine deaminase (IlvA) in *Salmonella enterica* Serovar Typhimurium. *J. Bacteriol.* **2004,** *186* (3), 803-810.

39. Castilho, B. A.; Olfson, P.; Casadaban, M. J., Plasmid insertion mutagenesis and *lac* gene fusion with mini-mu bacteriophage transposons. *J. Bacteriol.* **1984,** *158* (2), 488-495.

40. Vogel, H. J.; Bonner, D. M., Acetylornithinase of *Escherichia coli*: partial purification and some properties. *J. Biol. Chem.* **1956,** *218* (1), 97-106.

41. Balch, W. E.; Wolfe, R. S., New approach to the cultivation of methanogenic bacteria: 2-mercaptoethanesulfonic acid (HS-CoM)-dependent growth of *Methanobacterium ruminantium* in a pressureized atmosphere. *Appl. Environ. Microbiol.* **1976,** *32* (6), 781-791.

42. Le Guennec, A.; Tayyari, F.; Edison, A. S., Alternatives to nuclear Overhauser enhancement spectroscopy presat and Carr-Purcell-Meiboom-Gill presat for NMR-based metabolomics. *Anal. Chem.* **2017,** *89* (17), 8582-8588.

43. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **1995,** *6* (3), 277-293.

44. Bingol, K.; Li, D. W.; Zhang, B.; Bruschweiler, R., Comprehensive metabolite identification strategy using multiple two-dimensional NMR spectra of a complex mixture implemented in the COLMARm web server. *Anal. Chem.* **2016,** *88* (24), 12411-12418.

45. Walejko, J. M.; Chelliah, A.; Keller-Wood, M.; Gregg, A.; Edison, A. S., Global metabolomics of the placenta reveals distinct metabolic profiles between maternal and fetal placental tissues following delivery in non-labored women. *Metabolites* **2018,** *8* (1).

46. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H., Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. *Anal. Chem.* **2006,** *78* (13), 4281-4290.

47. Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van-der Vat, B. J. C.; Jellema, R. H., Fusion of mass spectrometry-based metabolomics data. *Anal. Chem.* **2005,** *77* (20), 6729-6736.

48. Wold, H., Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability* **1975,** *12* (S1), 117-142.

49. de Jong, S., SIMPLS: An alternative approach to partial least squares regression. *Chemometrics Intellig. Lab. Syst.* **1993,** *18* (3), 251-263.

50. Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A., Assessment of PLSDA cross validation. *Metabolomics* **2008,** *4* (1), 81-89.

51. Benjamini, Y.; Hochberg, Y., Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc.* **1995,** *57* (1), 289-300.

CHAPTER 5
CONCLUSION AND FUTURE DIRECTIONS

Metabolite identification and QA/QC in untargeted metabolomics are critical aspects to generate reproducible and confident outputs as the unknown chemical space yet to be determined can significantly impact in our understanding of metabolism and/or translational applications to human health. Current challenges derive from a combination of instrumentation and methodological limitations. Thus, the presented work increments the tools necessary to address these challenges and improve metabolomics outputs: (i) the development of materials capable of integrating data collected from multiple studies, analytical instruments, and laboratories, (ii) novel experimental designs that facilitate metabolite identification and (iii) the integration of metabolomics with biochemistry and genetic approaches that increase our understanding of metabolism.

A PERSPECTIVE ON IBAT APPLICATIONS

Any material can be made into a RM with IBAT.[1] The method is simple but requires some logistic organization. A simple spreadsheet can be used to track and record the IBAT process, but this requires direct user input and becomes difficult to manage or query and has no integration with collected analytical data. A software or automated system that can calculate and inform which aliquots and how many are to be combined for the next iteration based on the variance of each individual batch would be a significant development. Furthermore, historical individual batches analytical data can be used to define quality

criteria for acceptance into the IBAT process. This system together with a record of metadata associated with each batch can create a multivariate quality control system, similar to that used in pharmaceutical industry. These integrated data become a rich resource to answer questions of metabolite stability, process deviations, non-conformant batches, IBAT RM variance trends over the course of time and allows to better characterize the IBAT generated RM and its respective metabolites.

The community has noted the utility of RMs through numerous publications,[1-7] think-tanks[8] and workshops, and is a prominent feature of this dissertation. However, a reference material for a model organism provides additional opportunities beyond the commonly discussed large-scale studies, reproducibility or inter-laboratory comparisons.

Current model organism repositories house and distribute genetic references strains, natural strains and mutants that are extremely well-characterized with defined phenotypic traits and growth conditions, population dynamics and fully sequenced genome sequences (specially for the genetic references). Yet, despite the well-established infrastructures they do not offer metabolomics RMs which are much in need. The variance reduction properties of IBAT rely on regularly spaced, small batches and does not need specialized equipment or large-scale storage, which would be ideal for genetic repositories and in line with their routine daily operations. A common source of material has the potential to consolidate metabolite identifications efforts. Because of the incredibly diverse composition of an organism's metabolome, it is impossible to capture its entirety with a single extraction method or a single instrument. Thus, various methods have been developed to measure subsets of the metabolome. These efforts are often specific to a set method and a particular sample, grown under specific conditions which are difficult to

generalize and consolidate. However, a RM designed for the community would consolidate the metabolite identification efforts over the course of time. Additionally, open access data repositories can house these data they are reused, easily validated and incremented by different analytical methodologies, reducing redundant community efforts and accelerate the systematic metabolome annotation of a single organism.

METABOLOMICS FRACTION LIBRARY FUTURE DIRECTIONS

Metabolite identification in complex biological matrices is difficult due to a multitude of reasons, particularly for untargeted metabolomics that strongly rely on databases of chemical reference standards. The quality of the collected spectra is a product of numerous variables, from the type of matrix, pH and the physicochemical properties of the metabolite to the type of instrument and parameters used, which then influence the accuracy of the database matching. The work in chapter 3 aims to improve and assess this process.

The fractionation reduces spectral overlap and concentrates metabolites which helps to generate better quality data but most importantly it creates a common sample that can be measured using different complementary analytical techniques, specifically LC-MS and NMR. These provide orthogonal information that allows to annotate features from metabolomics studies confidently, but other analytical techniques are also possible. Further chromatographic separation can additionally separate or even isolate metabolites of interest. Fourier transform ion cyclotron resonance mass spectrometry (FTICR) is also a particularly appealing complementary measurement due to its high mass accuracy. Direct infusion FTICR of simplified fractions alleviates the need for hyphenated chromatography

to simplify spectra, often an issue in these experiments. The high mass accuracy reduces the number of possible chemical structures from a single formula. This becomes particularly beneficial when integrating NMR and MS together for the structural elucidation of unknown molecules.

The SUMMIT[9] approach developed by the Bruschweiler lab takes advantage of this synergy and predicts NMR spectra from molecular formulas that are matched against chemical standards. The application of this approach to the metabolite fraction library has the potential to create a semi-automated pipeline for unknown metabolite identification. Chemical formulas derived from the fraction LC-MS data can generate candidate chemical structures, but the number of candidate structures can become considerably high. Additional methods like, FTICR and collisional cross section measurements from ion mobility experiments, can further reduce this number so that it is computationally feasible to predict NMR spectra which are then matched against the NMR fraction data.

Computational methods to predict analytical measurements are increasingly improving their accuracy and reducing the computational burden needed for these complex calculations. Our collaborators have recently demonstrated a reduction in the overall computational time by 2 orders of magnitude for NMR chemical shifts of a set of molecules, while still producing good agreement with experimental observations.[10] These computational chemistry tools are undoubtedly an asset for metabolomics workflows, especially for metabolite identification and increasing the scope of current databases. This approach was the motivation for NIH funded project “Genetics and quantum tools for unknown metabolite identification” that funded the work in chapters 2 and 3. The work

demonstrated in these chapters are the building blocks for this unknown metabolite identification pipeline which is still ongoing and soon to be published.

FINAL REMARKS OF METABOLOMIC AND BIOCHEMICAL/GENETIC APPROACHES

The large repertoire of mutant model organisms that can be purchased and/or created facilitates the design of experiments that inform targeted nodes along a pathway of interest. Together with carefully planned biochemical experiments and metabolomics measurements, relevant metabolites previously not thought to be connected can be associated with a specific gene, particular pathway and/or enzyme of interest.[11]

A metabolite fraction library is useful for identifying metabolites, but it is also a powerful resource with applications beyond metabolite identification. Because NMR is a nondestructive technique, it means that concentrated and simplified extracts of a complex biological matrix can be reused and reanalyzed under different circumstances. One of the exciting applications discussed in our lab is the potential for this material to be used as screening panels to determine protein-metabolite targets. Pharmaceutical drug discovery methods routinely use a number of available NMR methods that can effectively determine these interactions and respective dissociation constants,[12, 13] providing exciting insight into metabolite roles in biochemical reactions.

Similar to natural products chemistry activity assays, these extracts can also be used to carryout additional experiments to derive loss of function mechanisms and/or phenotype rescue experiments, similar to the work illustrated in chapter 4. The metabolomics data provide a global perspective into the metabolic network. This network can adapt to both

internal and external disruptions generating potential new targets for additional experiments. The metabolite rich fractions can then help confirm these targets and validate mechanisms, but also provide novel ways to identify metabolic modulators that can circumvent disruptions.[14]

These three projects have metabolomics as a common link and more importantly inform and improve each method. These have been paradigm shifting approaches embraced by the Edison lab and will surely continue to improve and evolve generating valuable insights into metabolism.

REFERENCES

1. Gouveia, G. J.; Shaver, A. O.; Garcia, B. M.; Morse, A. M.; Andersen, E. C.; Edison, A. S.; McIntyre, L. M., Long-Term Metabolomics Reference Material. *Anal Chem* **2021,** *93* (26), 9193-9199.

2. Aristizabal-Henao, J. J.; Lemas, D. J.; Griffin, E. K.; Costa, K. A.; Camacho, C.; Bowden, J. A., Metabolomic Profiling of Biological Reference Materials using a Multiplatform High-Resolution Mass Spectrometric Approach. *J Am Soc Mass Spectrom* **2021,** *32* (9), 2481-2489.

3. Broadhurst, D.; Goodacre, R.; Reinke, S. N.; Kuligowski, J.; Wilson, I. D.; Lewis, M. R.; Dunn, W. B., Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* **2018,** *14* (6), 72.

4. Dunn, W. B.; Wilson, I. D.; Nicholls, A. W.; Broadhurst, D., The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **2012,** *4* (18), 2249-64.

5. Mahieu, N. G.; Patti, G. J., Systems-Level Annotation of a Metabolomics Data Set Reduces 25000 Features to Fewer than 1000 Unique Metabolites. *Anal Chem* **2017,** *89* (19), 10397-10406.

6. Phinney, K. W.; Ballihaut, G.; Bedner, M.; Benford, B. S.; Camara, J. E.; Christopher, S. J.; Davis, W. C.; Dodder, N. G.; Eppe, G.; Lang, B. E.; Long, S. E.; Lowenthal, M. S.; McGaw, E. A.; Murphy, K. E.; Nelson, B. C.; Prendergast, J. L.; Reiner, J. L.; Rimmer, C. A.; Sander, L. C.; Schantz, M. M.; Sharpless, K. E.; Sniegoski, L. T.; Tai, S. S.; Thomas, J. B.; Vetter, T. W.; Welch, M. J.; Wise, S. A.; Wood, L. J.; Guthrie, W. F.; Hagwood, C. R.; Leigh, S. D.; Yen, J. H.; Zhang, N. F.; Chaudhary-Webb, M.; Chen, H.; Fazili, Z.; LaVoie, D. J.; McCoy, L. F.; Momin, S. S.; Paladugula, N.; Pendergrast, E. C.; Pfeiffer, C. M.; Powers, C. D.; Rabinowitz, D.; Rybak, M. E.; Schleicher, R. L.; Toombs, B. M.; Xu, M.; Zhang, M.; Castle, A. L., Development of a Standard Reference Material for metabolomics research. *Anal Chem* **2013,** *85* (24), 11732-8.

7. Simon-Manso, Y.; Lowenthal, M. S.; Kilpatrick, L. E.; Sampson, M. L.; Telu, K. H.; Rudnick, P. A.; Mallard, W. G.; Bearden, D. W.; Schock, T. B.; Tchekhovskoi, D. V.; Blonder, N.; Yan, X.; Liang, Y.; Zheng, Y.; Wallace, W. E.; Neta, P.; Phinney, K. W.; Remaley, A. T.; Stein, S. E., Metabolite profiling of a NIST Standard Reference Material for human plasma (SRM 1950): GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources. *Anal Chem* **2013,** *85* (24), 11725-31.

8. Beger, R. D.; Dunn, W. B.; Bandukwala, A.; Bethan, B.; Broadhurst, D.; Clish, C. B.; Dasari, S.; Derr, L.; Evans, A.; Fischer, S.; Flynn, T.; Hartung, T.; Herrington, D.; Higashi, R.; Hsu, P. C.; Jones, C.; Kachman, M.; Karuso, H.; Kruppa, G.; Lippa, K.; Maruvada, P.; Mosley, J.; Ntai, I.; O'Donovan, C.; Playdon, M.; Raftery, D.; Shaughnessy, D.; Souza, A.; Spaeder, T.; Spalholz, B.; Tayyari, F.; Ubhi, B.; Verma, M.; Walk, T.; Wilson, I.; Witkin, K.; Bearden, D. W.; Zanetti, K. A., Towards quality assurance and quality control in untargeted metabolomics studies. *Metabolomics* **2019,** *15* (1), 4.

9. Bingol, K.; Bruschweiler-Li, L.; Yu, C.; Somogyi, A.; Zhang, F.; Bruschweiler, R., Metabolomics beyond spectroscopic databases: a combined MS/NMR strategy for the rapid identification of new metabolites in complex mixtures. *Anal Chem* **2015,** *87* (7), 3864-70.

10. Das, S.; Edison, A. S.; Merz, K. M., Jr., Metabolite Structure Assignment Using In Silico NMR Techniques. *Anal Chem* **2020,** *92* (15), 10412-10419.

11. Reed, L. K.; Baer, C. F.; Edison, A. S., Considerations when choosing a genetic model organism for metabolomics studies. *Curr Opin Chem Biol* **2017,** *36*, 7-14.

12. Nikolaev, Y. V.; Kochanowski, K.; Link, H.; Sauer, U.; Allain, F. H., Systematic Identification of Protein-Metabolite Interactions in Complex Metabolite Mixtures by Ligand-Detected Nuclear Magnetic Resonance Spectroscopy. *Biochemistry* **2016,** *55* (18), 2590-600.

13. Edison, A. S.; Colonna, M.; Gouveia, G. J.; Holderman, N. R.; Judge, M. T.; Shen, X.; Zhang, S., NMR: Unique Strengths That Enhance Modern Metabolomics Research. *Anal Chem* **2021,** *93* (1), 478-499.

14. Peng, B.; Li, H.; Peng, X. X., Functional metabolomics: from biomarker discovery to metabolome reprogramming. *Protein Cell* **2015,** *6* (9), 628-37.

BIOGRAPHICAL SKETCH

Goncalo Jorge Gouveia started his scientific career after graduating with honors from the University of Glamorgan (now University of Wales) in the United Kingdom in Forensic Science with focus on analytical chemistry. After graduating he started a Master of Science degree at the University of South London in Forensic Science where his research project was to determine the time since death using NMR spectroscopy. In 2009 he joined the Forensic Science Services Drugs Department and started working as a drugs analyst. He then became a Drug reporting officer at LGC Forensics in 2010, where he additionally had the roles of Internal Auditor, Heroin Profiling trainer and Complex Case Scientist until he moved to the Toxicology department to become a Toxicology Reporting Officer. In 2018, he joined the PhD. Program at the University of Georgia under the direction and advisement of Dr. Arthur Edison where he applied his previous knowledge to develop methods to improve metabolomics outcomes, by creating a novel method to make reference materials and approaches to improve metabolite identification using booth NMR and LC-MS platforms. In the future, he hopes to continue improving his skills in metabolomics and its direct applications to real world problems.

APPENDICES

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

SUPPLEMENTARY METHODS

Bioreactor production of Escherichia coli:

Cultures of *E. coli* were started from frozen stocks kept at -80 oC. A streak plate with LB media and agar (LB broth, Miller – Novagen, agar, Bacto BD) was made under standard aseptic conditions and incubated for 32 hours at 37 ^{o}C. A single colony was transferred to 20 mL of Terrific Broth (TB - Fisher BioReagents) and placed in a shaker incubator for 24 hours at 37 ^{o}C and 250 rpm. A contamination control LB plate was then streaked under aseptic conditions and incubated overnight. This liquid culture was the starting inoculum for a bioreactor (Biostat A, Sartorius) containing 2 L of TB with 30 mL of glycerol (EMD millipore). The bioreactor was under automated control of temperature, dissolved oxygen, pH (37 oC, 30% and 7.5 respectively) and constant mixing (500 rpm). After 42 hours growth an OD_{600} measurement was taken, and the bioreactor harvested into 500 mL centrifuge bottles and centrifuged for 30 min at 10,000 x G. The supernatant was discarded and the pellet process repeated two more times with deionized water and finally the centrifuged pellets were combined, weighed and reconstituted in M9 minimal media1 to a concentration of 0.5 g/mL. Aliquots from this material were then made so that each tube contains 1 g (wet weight) of material by an automated pipetting robot (Andrew Robot – Andrew Alliance), with continuous mixing, flash frozen in liquid Nitrogen and stored at -

80 °C. Making a stable food source for C. *elegans growth*: Individual batches of *E. coli* were produced as described above. Six aliquots from 10 different individual bacterial batches were thawed on ice and pooled together as substrate for one *C. elegans* batch. For optimal *C. elegans* growth, a total of 60 bacterial aliquots were required to achieve a ratio of 3% (w/v)2 of substrate to volume of media in a 2 L bioreactor (Biostat, Sartorius). IBAT was used to create two additional batches of food, each containing 60 aliquots from 10 batches where, for each iteration, aliquots from one individual batch were removed, and new individual batch aliquots added. Growing *C. elegans* in bioreactors: Similar to the *E. coli* bioreactor process a starting inoculum of *C. elegans* was first made. This was a population of worms collected from a large scale culture plate as described previously.[3] Approximately 2 million worms were washed with M9 media and added to the bioreactor (Biostat A, Sartorius) containing 2 L of K-media1 and the stable food source created above. The Bioreactor was under automated control of temperature, dissolved oxygen, pH (20 °C, 10% and 7 respectively) and constant mixing (150 rpm). Two daily OD_{600} measurements were taken to monitor the amount of available food and the nematodes counted under the microscope to account for overcrowding. The bioreactor was harvested when food was below 0.5% w/v (calculated from OD_{600} measurements) and/or nematode density was above 30,000 individuals/mL. The harvested culture was then divided into 500 mL centrifuge bottles and centrifuged for 20 min at 5,000 x G and 4 °C. The supernatant discarded, and the wash process repeated two more times with M9 media and a final reconstitution with deionized water. The contents of each bottle were combined, and three 1 mL aliquots taken to count the number of nematodes[3]. The material was then aliquoted into 15 mL centrifuge tubes by an automated pipetting robot (Andrew Robot – Andrew

Alliance), with continuous mixing, each containing approximately 2,000,000 nematodes, flash frozen in liquid Nitrogen and stored at -80 ^{o}C.

NMR data acquisition and processing:

One-dimensional ^{1}H NMR spectra were acquired using moesypr1d with pre-saturation during relaxation delay and mixing time on an Avance III HD 600 MHz Bruker NMR spectrometer equipped with a TCI cryoprobe and a Bruker SampleJet autosampler cooled to 5.6 °C. During acquisition, 32,768 complex datapoints were collected for the FID, using 64 scans with 4 additional dummy scans. The spectral width was 20 ppm. A Fourier transform (FT), a polynomial baseline correction of order 2, a 0.5 Hz line broadening and phase correction were applied to each spectrum using NMRPipe processing software.[4] Two-dimensional ^{1}H-^{1}H total correlation spectroscopy (TOCSY- dipsi2esfbgpph), ^{1}H-^{13}C heteronuclear single quantum correlation (HSQC - hsqcedetgpsisp2.3) and ^{1}H-^{13}C HSQC–total correlation spectroscopy (HSQC–TOCSY - hsqcdietgpsisp.2) experiments were collected on both *C. elegans* and *E. coli* samples for metabolite identification. During acquisition, all three experiments were collected for 32 scans and an additional 16 dummy scans, with 512 and 1,024 datapoints recorded on the direct and indirect dimensions respectively and, a spectral width of 200 ppm for ^{13}C and 12 ppm for ^{1}H. A 90ms mixing time was used for both HSQC-TOCSY and TOCSY experiments. All spectral processing was carried out using NMRPipe [5]. Compound identification/database matching: All two-dimensional experiments were used for spectral matching against the BBiorefcode library using COLMARm6 and a chemical shift cutoff of 0.03 and 0.3 ppm for ^{1}H and ^{13}C respectively. Metabolites that could be quantified without overlap and were consistent

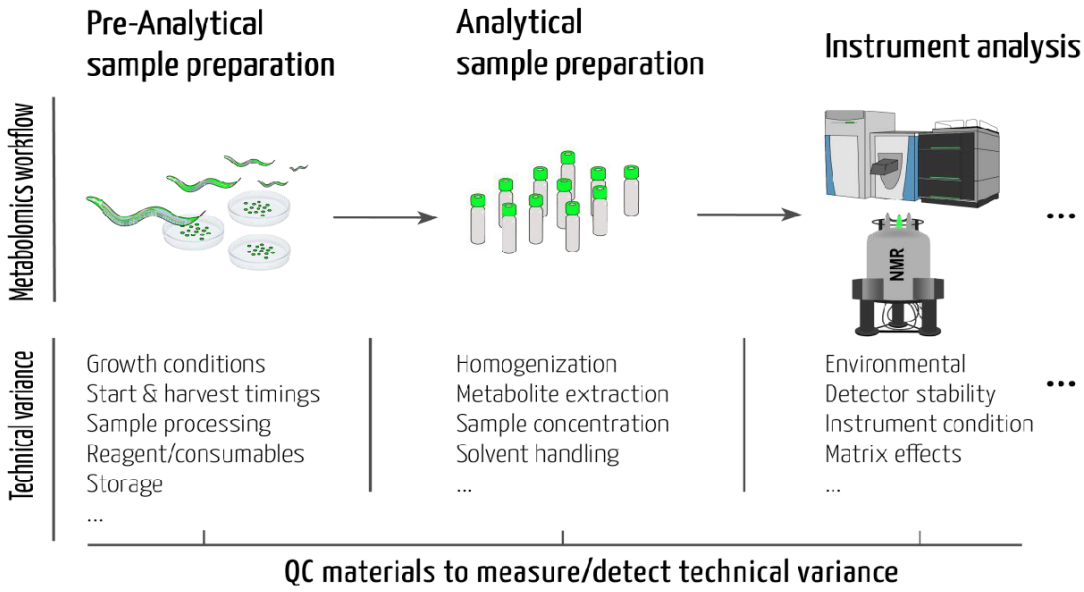
between replicates in their respective 1D 1HNMR spectra were selected to be identified. From the *E. coli* samples 19 features were annotated to metabolites and 26 in the *C. elegans* samples. A confidence level ranging from 1 to 5 (Supplementary Table 2.1), was assigned to each metabolite as described elsewhere [7]. Briefly this scale is defined as: (1) putatively characterized compound, (2) matched to reported 1D spectra, (3) matched to reported HSQC spectra, (4) matched to reported HSQC and HSQC- spectra, and (5) validated by spiking putative compound into sample.

Supplementary Table 2.1a: Table of metabolites isolated features that were common to all for *E. coli* samples.

			Individual batches			IBAT batches		
Compound name (COLMARm)	Identification level	ppm_1D	Mean	Standard deviation	CV (mean/std)	Mean	Standard deviation	CV (mean/std)
'Isovaleric_acid_1	4	0.8903	6.68E+07	6.20E+07	0.9281	5.83E+07	2.64E+07	0.4528
'Leucine_1	4	0.9375	8.25E+07	7.83E+07	0.9501	5.27E+07	3.42E+07	0.6494
'L_Valine_1	4	0.9695	5.69E+07	4.14E+07	0.7278	3.37E+07	1.56E+07	0.4622
'L_Isoleucine_1	4	0.9929	3.58E+07	2.57E+07	0.7185	2.25E+07	1.09E+07	0.4865
'3_Hydroxybutyrate_1	3	1.1741	3.91E+07	4.03E+07	1.0322	2.34E+07	2.14E+07	0.9116
'Lactic_acid_1	4	1.2943	3.10E+07	1.60E+07	0.5143	2.04E+07	4.04E+06	0.1978
'Cadaverine_1	3	1.4481	9.72E+07	3.49E+07	0.3594	7.48E+07	1.45E+07	0.1943
'Acetic_acid_1	3	1.8941	1.04E+08	4.31E+07	0.4136	8.88E+07	2.86E+07	0.3220
'L_Glutamic_acid_1	3	2.3281	4.45E+07	2.31E+07	0.5194	3.61E+07	8.71E+06	0.2413
'L_Methionine_1	4	2.6190	8.95E+06	7.27E+06	0.8119	7.33E+06	2.58E+06	0.3524
'D_Aspartate_1	3	2.6564	7.67E+06	4.70E+06	0.6119	6.16E+06	2.07E+06	0.3360
'Betaine_1	4	3.2377	8.80E+08	5.33E+08	0.6061	8.56E+08	3.54E+08	0.4138
'D_Ribose_2	4	4.8944	8.61E+06	4.07E+06	0.4726	6.62E+06	1.87E+06	0.2830
'Uracil_1	4	5.7637	8.69E+06	4.68E+06	0.5389	6.75E+06	2.69E+06	0.3981
'Fumaric_acid_1	3	6.4871	4.88E+05	4.34E+05	0.8898	1.69E+05	7.51E+04	0.4453
'L_Tyrosine_1	3	7.1577	6.45E+06	3.25E+06	0.5043	4.67E+06	1.74E+06	0.3716
L_Phenylalanine_1	4	7.4040	8.04E+06	6.34E+06	0.7890	7.04E+06	3.56E+06	0.5057
'Nicotinic_acid_1	4	8.2322	1.24E+06	6.80E+05	0.5498	8.79E+05	4.00E+05	0.4553
'Formate_1	3	8.4308	3.37E+06	4.24E+06	1.2577	1.50E+06	1.15E+06	0.7647

Supplementary Table 2.1b: Table of metabolites isolated features that were common to all for *C. elegans* samples.

Compound name (COLMARm)	Identification level	ppm_1D	PD1074 Mean	PD1074 SD	IBAT Mean	IBAT SD	PoolQC Mean	PoolQC SD
'AMP_sulfate_1'	3	8.2697	1.05E+10	6.43E+09	6.52E+09	6.18E+08	1.18E+10	2.73E+09
'Benzoate_1'	3	7.4853	2.41E+09	1.41E+09	3.11E+09	1.59E+09	2.50E+09	9.87E+08
'L_Phenylalanine_1'	4	7.4387	2.31E+09	1.73E+09	4.06E+09	1.70E+09	2.18E+09	2.89E+08
'L_Tyrosine_1'	4	6.8948	3.04E+09	2.04E+09	6.30E+09	1.34E+09	3.34E+09	5.80E+08
'UDP_GlcNAc_1'	3	5.5317	1.63E+09	8.80E+08	1.20E+09	1.47E+08	1.60E+09	1.93E+08
'Allantoin_1'	3	5.3865	4.11E+09	2.86E+09	3.54E+09	6.51E+08	4.40E+09	9.41E+08
'D_Trehalose_1'	4	5.188	6.04E+10	3.54E+10	1.47E+11	1.23E+10	6.25E+10	1.80E+09
'D_Glucose_1'	4	4.6492	5.13E+09	2.73E+09	8.45E+09	1.31E+09	5.28E+09	8.57E+08
'Betaine_1'	4	3.2672	3.15E+11	2.75E+11	1.08E+12	1.29E+11	3.82E+11	1.63E+11
'Lysine_1'	4	3.054	2.67E+10	1.53E+10	1.50E+10	2.18E+09	2.99E+10	3.78E+09
'L_Asparagine_1'	3	2.9527	4.40E+09	2.71E+09	1.62E+09	1.27E+09	5.37E+09	4.41E+08
'D_Aspartate_1'	3	2.7968	7.35E+09	4.32E+09	3.63E+09	1.71E+09	7.31E+09	1.16E+09
'Allocystathionine_1'	3	2.741	1.47E+10	8.99E+09	2.32E+09	2.54E+08	1.48E+10	1.07E+09
'Succinic_acid_1'	3	2.4119	2.78E+11	2.13E+11	2.44E+11	2.86E+10	3.13E+11	6.05E+10
'L_Glutamic_acid_1'	3	2.3585	5.98E+10	4.13E+10	7.65E+10	1.06E+10	6.42E+10	1.16E+10
'2_Aminoadipic_acid_1'	3	2.2476	1.36E+10	1.17E+10	1.21E+10	1.87E+09	1.18E+10	2.40E+08
'N_acetyl_putrescine_1'	4	1.9906	7.40E+10	7.02E+10	8.24E+10	9.18E+09	7.60E+10	1.19E+10
'Acetic_acid_1'	3	1.9208	6.15E+10	4.26E+10	1.59E+11	2.29E+10	6.57E+10	1.61E+10
'Putrescine_1'	4	1.7771	2.67E+10	1.51E+10	1.08E+10	4.98E+09	2.67E+10	2.57E+09
'Alanine_1'	4	1.4761	3.46E+11	2.29E+11	5.59E+11	1.09E+11	3.65E+11	5.96E+10
'Lactic_acid_1'	4	1.3238	1.02E+11	8.79E+10	1.81E+11	2.13E+10	1.11E+11	4.34E+09
'Propionic_acid_1'	3	1.0672	6.16E+09	4.94E+09	3.63E+10	4.89E+09	9.85E+09	5.60E+09
'L_Isoleucine_1'	4	1.0192	1.07E+10	5.97E+09	3.51E+10	1.82E+10	1.30E+10	2.67E+09
'L_Valine_1'	4	0.98857	1.82E+10	9.89E+09	6.06E+10	2.87E+10	2.01E+10	1.80E+09
'Leucine_1'	4	0.97397	1.95E+10	1.10E+10	5.78E+10	2.60E+10	2.17E+10	1.86E+09
'L_Isoleucine_1'	4	0.94007	3.37E+10	1.35E+10	6.01E+10	2.13E+10	3.55E+10	3.36E+09
'Pantothenate_1'	3	0.89793	2.42E+10	1.28E+10	4.53E+10	8.29E+09	2.86E+10	3.00E+09



Supplementary Figure 2.1: General metabolomics workflow from sample generation to instrument analysis. Non-exhaustive examples of pre-analytical technical variance at each step of the metabolomics process.

Supplementary References:

1. Steiernagle, T. Maintenance of C. elegans - Wormbook http://www.wormbook.org/.

2. Kaplan, F.; Srinivasan, J.; Mahanti, P.; Ajredini, R.; Durak, O.; Nimalendran, R.; Sternberg, P. W.; Teal, P. E.; Schroeder, F. C.; Edison, A. S.; Alborn, H. T., Ascaroside expression in Caenorhabditis elegans is strongly dependent on diet and developmental stage. PLoS One 2011, 6 (3), e17804.

3. Amanda O. Shaver, G. J. G., Pamela S. Kirby, Erik Andersen, Arthur S. Edison, Culture and assay of Large-Scale Mixed Stage Caenorhabditis elegans Population. JOVE - J. Vis. Exp 2020, e61453.

4. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., NMRPipe: a multidimensional spectral
processing system based on UNIX pipes. J Biomol NMR 1995, 6 (3), 277-93.

5. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., NMRPipe: A multidimensional spectral processing system based on UNIX pipes. J. Biomol. NMR 1995, 6 (3), 277-293.

6. Bingol, K.; Li, D. W.; Zhang, B.; Bruschweiler, R., Comprehensive metabolite identification strategy using multiple two-dimensional NMR spectra of a complex mixture implemented in the COLMARm web server. Anal. Chem. 2016, 88 (24), 12411-12418.

7. Walejko, J. M.; Chelliah, A.; Keller-Wood, M.; Gregg, A.; Edison, A. S., Global metabolomics of the placenta reveals distinct metabolic profiles between maternal and fetal placental tissues following delivery in non-labored women. Metabolites 2018, 8 (1).

APPENDIX B

SUPLEMENTAL MATERIAL FOR CHAPTER 3

SUPPLEMENTAL METHODS

Chemical and reagents

All reagents were LC/MS grade. Water (H_2O), methanol (MeOH), isopropanol (IPA), and acetonitrile (ACN) were purchased from Fisher Scientific (Optima™). Formic acid was purchased from Honeywell Fluka Chemicals. Ammonium formate and ammonium acetate (Optima® LC/MS) were purchased from Fisher Chemical. Reagents were stored in 5% hydrochloric acid (HCl) washed and solvent rinsed glassware. Deuterated water D_2O, D, 99.9%; and DSS sodium 2,2-dimethyl-2-silapentane-5-sulfonate, D6, 98% were purchased from Cambridge Isotope Laboratories).

C. elegans reference material preparation

To generate the reference material used for fractionation 10 vials each containing approximately 200,000 frozen and lyophilized nematodes (grown and stored as described elsewhere[1]), were selected according to the IBAT method.[2] The material from these vials was combined, mixed with a spatula and aliquoted into five separate 15 mL glass vials.

Metabolite extraction

To each 15mL glass vial 400 µL of 1 mm zirconia beads were added and homogenized at 1800 oscillations/min for 90 seconds in a FastPrep-96 (MPbio) homogenizer. This step was repeated three times with a 1min resting period in dry ice to prevent overheating. After homogenized, 3 mL of 100% IPA chilled to -20°C was added to each vial in three increments of 1000 µL and vortex mixed for approximately 30s after each addition. The vials were left at room temperature for 30 min prior to a 12h extraction period at -20°C. The vials were then centrifuged at 20,800 x G and 4°C for 30 minutes. The supernatant was transferred to a single tube, labelled "non-polar extract" and placed in a Centrivap (Labconco) at room temperature until completely dry together with the five pellet-containing vials. The dry non-polar supernatant was stored at -80°C and 3 mL of cold 80:20 MeOH:H_2O were added to each of the remaining five vials for polar analyte extraction. This polar fraction was vortex mixed for 30 minutes at 4°C. The tubes were then centrifuged at 20,800 x G and 4°C for 30 minutes and the supernatant was transferred to a single tube. This polar fraction was placed in a Centrivap at room temperature until dry and stored at -80°C.

Semi-preparative HILIC HPLC fractionation

The polar extract was reconstituted in a total of 800 µL of 50/50 MeOH/H_2O. The vial was vortex mixed for 10 min and centrifuged at 20,800 x G and 4°C for 30 minutes. The supernatant was transferred into a high recovery LC-MS vial which was then used for a total of six injections into an Agilent 1260 infinity for the fractionation process. Three 50 µL injections were carried out to equilibrate the column pre fractionation. An XBridge

BEH Amide OBD Prep Column, 130Å, 5 µm, 10 mm X 250 mm, was used with a flow rate of 3.5mL/min to separate the analytes and collect fractions at 30 second intervals for a total of 100 fractions after DAD detection using a Foxy Jr. fraction collector. The mobile phase composition consisted of Solvent A, 95% ACN and 5% H_2O and Solvent B 80% ACN and 20% H_2O using the following linear gradient program: 0.0-2 min 100% A; 2-5 min 96% A; 5-10 min 95% A; 10-15 min 85% A; 15-18 75% A; 18-21 min 65% A; 21-26 min 55% A; 26-32 min 53% A and 32-38 min 35% A for a total run oof 45 min plus five minutes of solvent switch to starting conditions. No fractions were collected during an additional equilibrium of 5 min at 100% A between injections.

The non-polar fractionation was also carried out, but the data was not used in this manuscript and not described further. After fractionation, four 100 µL aliquots from each fraction and each chromatography were transferred into LC-MS vials. These vials and the remaining fraction tubes containing approximately 10 mL for NMR analysis were placed in a Centrivap at room temperature until dry and stored at -80°C.

One-dimensional NMR analysis

The HPLC polar elutants were reconstituted in 70 µL of D_2O, containing 0.11 mM DSS as an internal standard and transferred into 1.7 mm NMR tubes (SampleJet, Bruker). These were then loaded onto a SampleJet automated sample changer and kept at 6°C. One-dimensional (1D) ^1H NMR spectra for each fraction and blanks were collected using the pulse sequence noesypr1d on a Bruker NEO 800MHz equipped with a 1.7mm TCI cryoprobe. During acquisition, 32,768 complex data points were collected using 128 scans with four additional dummy scans. The spectral width was 20 ppm. In addition,

immediately after each 1D acquisition, a two-dimensional (2D) J-resolved spectrum was collected using the Bruker pulse program jresgpprqf. A total of 8,192 and 40 points were collected using four scans, 16 dummy scans, and spectral widths of 20 and 0.13 ppm. Following data acquisition, the 1D data were Fourier transformed using NMRPipe[3]. Further processing consisted of an exponential line broadening of 2 Hz, automatic zero fill baseline and manual phase correction.

Two-dimensional NMR analysis

Two-dimensional ^{1}H-^{1}H total correlation spectroscopy (TOCSY; dipsi2esfbgpph), ^{1}H-^{13}C heteronuclear single quantum correlation (HSQC; hsqcedetgpsisp2.3) and ^{1}H-^{13}C HSQC-total correlation spectroscopy (HSQC-TOCSY; hsqcdietgpsisp.2), experiments were collected on select fractions. During acquisition, all three experiments were collected for 128 scans and an additional 16 dummy scans, with 512 and 1,024 data points recorded on the direct and indirect dimensions, respectively, and a spectral width of 200 ppm for ^{13}C and 12 ppm for ^{1}H. A 90-ms mixing time was used for both HSQC-TOCSY and TOCSY experiments. All spectral processing was carried out using NMRpipe[3].

Database matching and 2D spectral annotation

The above three 2D data were used for spectral matching against the COLMARm[4] database using matching chemical shift cutoffs of 0.04 and 0.3 ppm for ^{1}H and ^{13}C, respectively. In addition, HSQC only and TOCSY only queries were carried out using the same threshold criteria. Spectral annotation and visualization was carried out using a combination of MNova (version 14.2.0) and NMR View J[5].

Hydrophilic Interactive Liquid Chromatography (HILIC) methods

Dried polar extracts were resuspended in 200 µL of 80/20 ACN/H2O for LC-MS/MS. Polar extracts were separated using a Vanquish (ThermoFisher Scientific), fitted with a Waters Acquity UPLC BEH Amide column (2.1 x 150 mm, 1.7 µm particle size). The compounds were eluted with the following gradient: 80:20 H2O:ACN with 10 mM ammonium formate and 0.1% formic acid (mobile phase A) and 100% ACN with 0.1% formic acid (mobile phase B) using the following gradient program: -5.0 min 95% B; 0.0-0.5 min 95% B; 8.0-9.4 min 40% B; 9.5-11.0 min 95% B. A curve 5 value was set for -5.0 and 0.0 minutes, a curve 6 at 0.5 min, curve 7 at 8.0 min, and a curve 6 for the remainder of the gradient. The flow rate was set at 0.400 mL min^{-1}. The column temperature was set to 40°C, and the injection volume was 2 µL.

HILIC Mass spectrometer settings and methods

A Q Exactive HF (ThermoFisher Scientific) equipped with a HESI ion source was used for all mass spectrometry data collection. The mass spectrometer was run in full MS mode at a resolution of 240,000 (at m/z 200) for the duration of the chromatographic gradient. An automatic gain control (AGC) target of 1e5 was set with a maximum injection time of 150 ms. A tune file with the following source conditions was used for positive and negative mode: spray voltage (+) 3000, spray volage (-) 2800, capillary temperature: 275°C, sheath gas: 50, aux gas: 13, spare gas: 4.0, max spray current: 100, probe heater temperature: 425.0°C, and S-Lens RF level: 50. A m/z scan 70-1050 was used. Calibration was conducted using ThermoFisher Pierce™ Negative Ion Calibration Solution and Pierce™ LTQ Velos ESI Positive Ion Calibration Solution prior to the collection of negative and

positive mode data, respectively. A stepped normalized collision energy [(N)CE] was used for MS^2 data collection. MS^2 resolution was set to 30,000 at m/z 200. An AGC target of 5e4 was set with a maximum injection time of 50 ms and a minimum AGC target of 8.00e3. A loop count of 5 was set with an isolation window of 1.0 m/z. (N)CE was set to 10, 30, and 50. Dynamic exclusion was set to 10 s.

SLAW

LC-MS/MS data was processed using SLAW (https://github.com/zamboni-lab/SLAW).[6] SLAW is a scalable and self-optimizing processing workflow for untargeted LC-MS. Meta-analysis data was processed using 12 pooled samples of the *C. elegans* laboratory reference strain PD1074 for optimization of the ADAP[7] peak picking algorithm and alignment. A filtering threshold of 1.0 was set for the fraction of detection in QC samples for a feature to be kept (frac_qc). This ensured features were stable and present in all pooled PD1074 QCs. The optimized parameter file was then used to process the fraction library dataset. Gap filling (output_format ms1) was turned off (value: data matrix) in the fraction library due to the nature of the fractionation process and expectation for missing values for features across many individual fractions. Data was blank filtered using a 10x sample-to-blank ratio using MATLAB where the intensity of each feature was retained if the feature had an intensity 10 times greater than the average intensity across all solvent and reconstitution blanks and the first/last five fractions in at least one sample.

metabCombiner

Feature tables from the meta-analysis and *C. elegans* fractionation were combined using metabCombiner.[8]

(https://bioconductor.org/packages/release/bioc/html/metabCombiner.html)

Each blank filtered and deconvoluted data matrix containing a column with a unique identifier served as inputs. Scoring metrics parameters are used to determine the weight of m/z (A), RT (B), and peak intensity (C) between the two data sets. Scoring parameters were set to A = 70, B = 15, and C = 0 for analysis.

GNPS

A molecular network was created with the Feature-Based Molecular Networking (FBMN) workflow[9] on GNPS (https://gnps.ucsd.edu).[10] The mass spectrometry data were first processed with SLAW[6] and the results were exported for FBMN analysis. An .mgf file, MS^2 quantification table containing mass-to-charge (m/z), retention time (RT), and peak height (PH) across fractions, and a meta-data table relating file names to fraction numbers were used as inputs. MATLAB was used to filter the quantification table. Feature PHs were set to zero if the intensity of that feature was less than $1/5^{th}$ the maximum peak intensity. This was used to remove noise across fractions and to allow for visualization of the presence of features across fractions in Cytoscape.[11] The data was filtered by removing all MS^2 fragment ions within +/- 17 Da of the precursor m/z. MS^2 spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50 Da window throughout the spectrum. The precursor ion mass tolerance was set to 0.005 Da and the MS^2 fragment ion tolerance to 0.005 Da. A molecular network was then created where edges were filtered to

have a cosine score above 0.7 and more than 1 matched peaks. Further, edges between two nodes were kept in the network if and only if each of the nodes appeared in each others respective top 10 most similar nodes. Finally, the maximum size of a molecular family was set to 100, and the lowest scoring edges were removed from molecular families until the molecular family size was below this threshold. The spectra in the network were then searched against GNPS spectral libraries.[10, 12] The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have a score above 0.7 and at least 1 matched peak. The DEREPLICATOR was used to annotate MS^2 spectra.[13] The molecular networks were visualized using Cytoscape software.[11] Pie charts were used for visualization, using a modulated color scheme with fraction numbers chosen as the displayed attribute.

SIRIUS

SIRIUS 4.0 was used for LC-MS/MS molecular formula identification (SIRIUS[14]), network-based improvement of SIRIUS molecular formula rankings (ZODIAC), metabolite annotation (CSI:FingerID[15]), and compound class prediction (CANOPUS[16, 17]). The .mgf output from SLAW was used to input all MS^1 and MS^2 data. Within the graphical user interface (GUI) the Orbitrap was selected as the instrument type with a MS2 MassDev (ppm) set to five. A total of 10 candidates with one candidate per ion was used for SIRIUS, with [M+H]+, [M+K]+, and [M+Na]+ as possible ionizations. The default elements allowed in the molecular formula were used. Default values for ZODIAC[18] were used for analysis, and CSI:FingerID was set to search all included databases using [M+H]+ as the fallback adduct. No input parameters are required for CANOPUS.

Fraction analysis

Using a toolbox developed in-house (MATLAB MathWorks, R2019a), the polar fractions spectra were referenced at 0.00 ppm using DSS. Spectral correlation analysis STOCSY [19] was used to determine features highly correlated to the same driver peak and native Matlab covariance function was used to calculate the covariance matrix. Data visualization and data handling tools and functions are available from the Edison Lab toolbox at https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA.

Colmar unique HMDB and BMRB identifiers were converted into InChIKeys using the Chemical Translation Service.[20]

Using MATLAB, individual matrices were matched by planar InChIKeys and fraction number for SIRIUS, GNPS and COLMAR outputs. Planar InChIKeys consist of the first 14 characters of an InChIKey. These solely encode the molecular connectivity of each InChIKey. A consensus chemical name list was obtained using the "Query Chemical Identifier Resolver" from the "Webchem" R package.

REFERENCES

1. Shaver, A. O.; Gouveia, G. J.; Kirby, P. S.; Andersen, E. C.; Edison, A. S., Culture and Assay of Large-Scale Mixed-Stage Caenorhabditis elegans Populations. *J Vis Exp* **2021,** (171).

2. Gouveia, G. J.; Shaver, A. O.; Garcia, B. M.; Morse, A. M.; Andersen, E. C.; Edison, A. S.; McIntyre, L. M., Long-Term Metabolomics Reference Material. *Anal Chem* **2021,** *93* (26), 9193-9199.

3. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **1995,** *6* (3), 277-93.

4. Bingol, K.; Li, D. W.; Zhang, B.; Bruschweiler, R., Comprehensive Metabolite Identification Strategy Using Multiple Two-Dimensional NMR Spectra of a Complex Mixture Implemented in the COLMARm Web Server. *Anal Chem* **2016,** *88* (24), 12411-12418.

5. Johnson, B. A.; Blevins, R. A., NMR View: A computer program for the visualization and analysis of NMR data. *J Biomol NMR* **1994,** *4* (5), 603-14.

6. Delabriere, A.; Warmer, P.; Brennsteiner, V.; Zamboni, N., SLAW: A Scalable and Self-Optimizing Processing Workflow for Untargeted LC-MS. *Analytical Chemistry* **2021,** *93* (45), 15024-15032.

7. Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. A.-O., One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. (1520-6882 (Electronic)).

8. Habra, H.; Kachman, M.; Bullock, K.; Clish, C.; Evans, C. R.; Karnovsky, A., metabCombiner: Paired Untargeted LC-HRMS Metabolomics Feature Matching and Concatenation of Disparately Acquired Data Sets. *Analytical Chemistry* **2021,** *93* (12), 5028-5036.

9. Nothias, L.-F.; Petras, D.; Schmid, R.; Dührkop, K.; Rainer, J.; Sarvepalli, A.; Protsyuk, I.; Ernst, M.; Tsugawa, H.; Fleischauer, M.; Aicheler, F.; Aksenov, A. A.; Alka, O.; Allard, P.-M.; Barsch, A.; Cachet, X.; Caraballo-Rodriguez, A. M.; Da Silva, R. R.; Dang, T.; Garg, N.; Gauglitz, J. M.; Gurevich, A.; Isaac, G.; Jarmusch, A. K.; Kameník, Z.; Kang, K. B.; Kessler, N.; Koester, I.; Korf, A.; Le Gouellec, A.; Ludwig, M.; Martin H, C.; McCall, L.-I.; McSayles, J.; Meyer, S. W.; Mohimani, H.; Morsy, M.; Moyne, O.; Neumann, S.; Neuweger, H.; Nguyen, N. H.; Nothias-Esposito, M.; Paolini, J.; Phelan, V. V.; Pluskal, T.; Quinn, R. A.; Rogers, S.; Shrestha, B.; Tripathi,

A.; van der Hooft, J. J. J.; Vargas, F.; Weldon, K. C.; Witting, M.; Yang, H.; Zhang, Z.; Zubeil, F.; Kohlbacher, O.; Böcker, S.; Alexandrov, T.; Bandeira, N.; Wang, M.; Dorrestein, P. C., Feature-based molecular networking in the GNPS analysis environment. *Nature Methods* **2020,** *17* (9), 905-908.

10. Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W.-T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. Ø.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016,** *34* (8), 828-837.

11. Shannon, P.; Markiel A Fau - Ozier, O.; Ozier O Fau - Baliga, N. S.; Baliga Ns Fau - Wang, J. T.; Wang Jt Fau - Ramage, D.; Ramage D Fau - Amin, N.; Amin N Fau -

Schwikowski, B.; Schwikowski B Fau - Ideker, T.; Ideker, T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. (1088-9051 (Print)).

12. Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K., MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry* **2010,** *45* (7), 703-714.

13. Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.; Shcherbin, E.; Nothias, L.-F.; Dorrestein, P. C.; Pevzner, P. A., Dereplication of microbial metabolites through database search of mass spectra. *Nature Communications* **2018,** *9* (1), 4035.

14. Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S., SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* **2019,** *16* (4), 299-302.

15. Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S., Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences* **2015,** *112* (41), 12580.

16. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S., ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **2016,** *8* (1), 61.

17. Dührkop, K.; Nothias, L.-F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M. A.; Petras, D.; Gerwick, W. H.; Rousu, J.; Dorrestein, P. C.; Böcker, S., Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology* **2021,** *39* (4), 462-471.

18. Ludwig, M.; Nothias, L.-F.; Dührkop, K.; Koester, I.; Fleischauer, M.; Hoffmann, M. A.; Petras, D.; Vargas, F.; Morsy, M.; Aluwihare, L.; Dorrestein, P. C.;

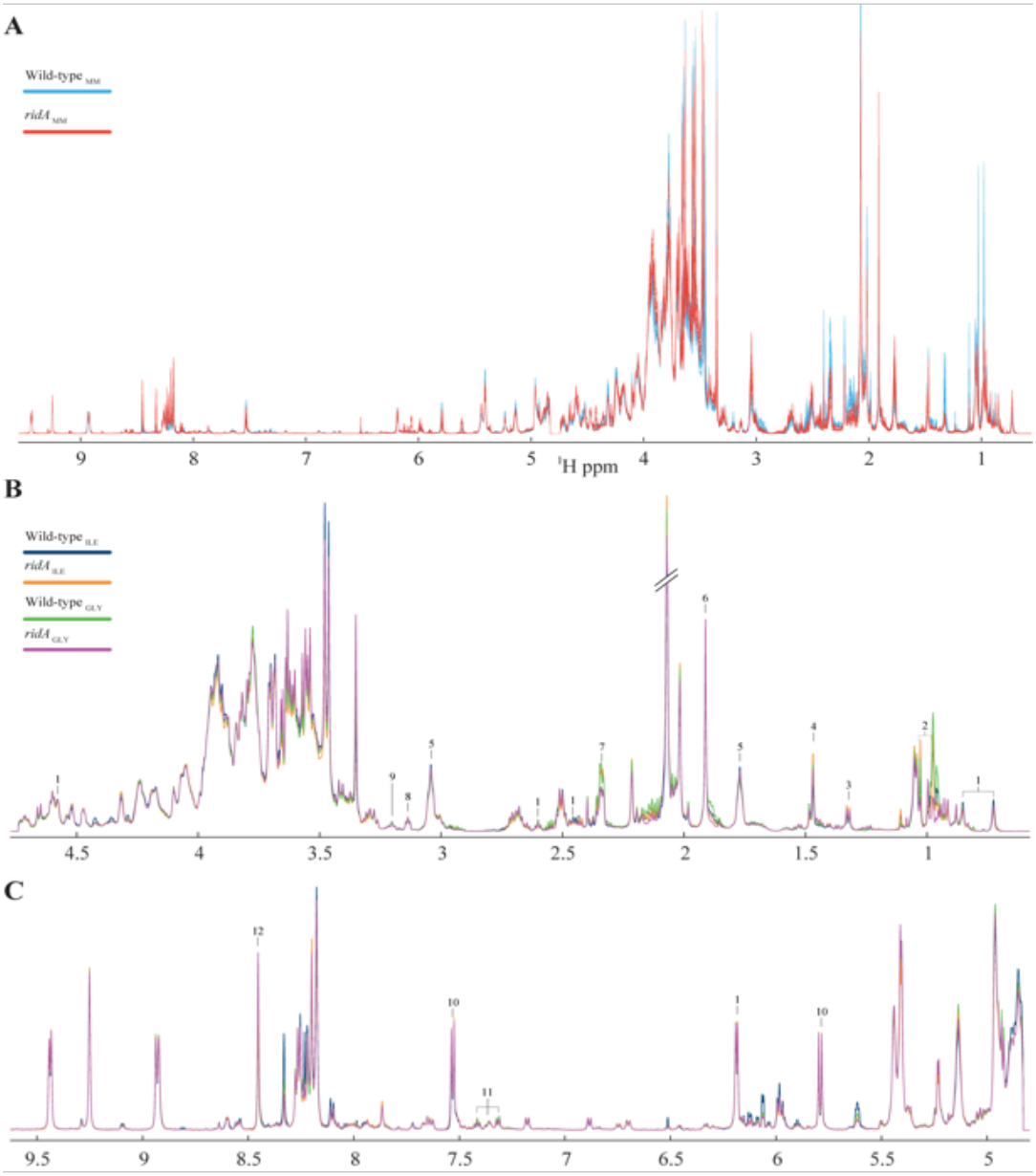
Böcker, S., ZODIAC: database-independent molecular formula annotation using Gibbs sampling reveals unknown small molecules. *bioRxiv* **2019**, 842740.

19. 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* (5), 1282-9.

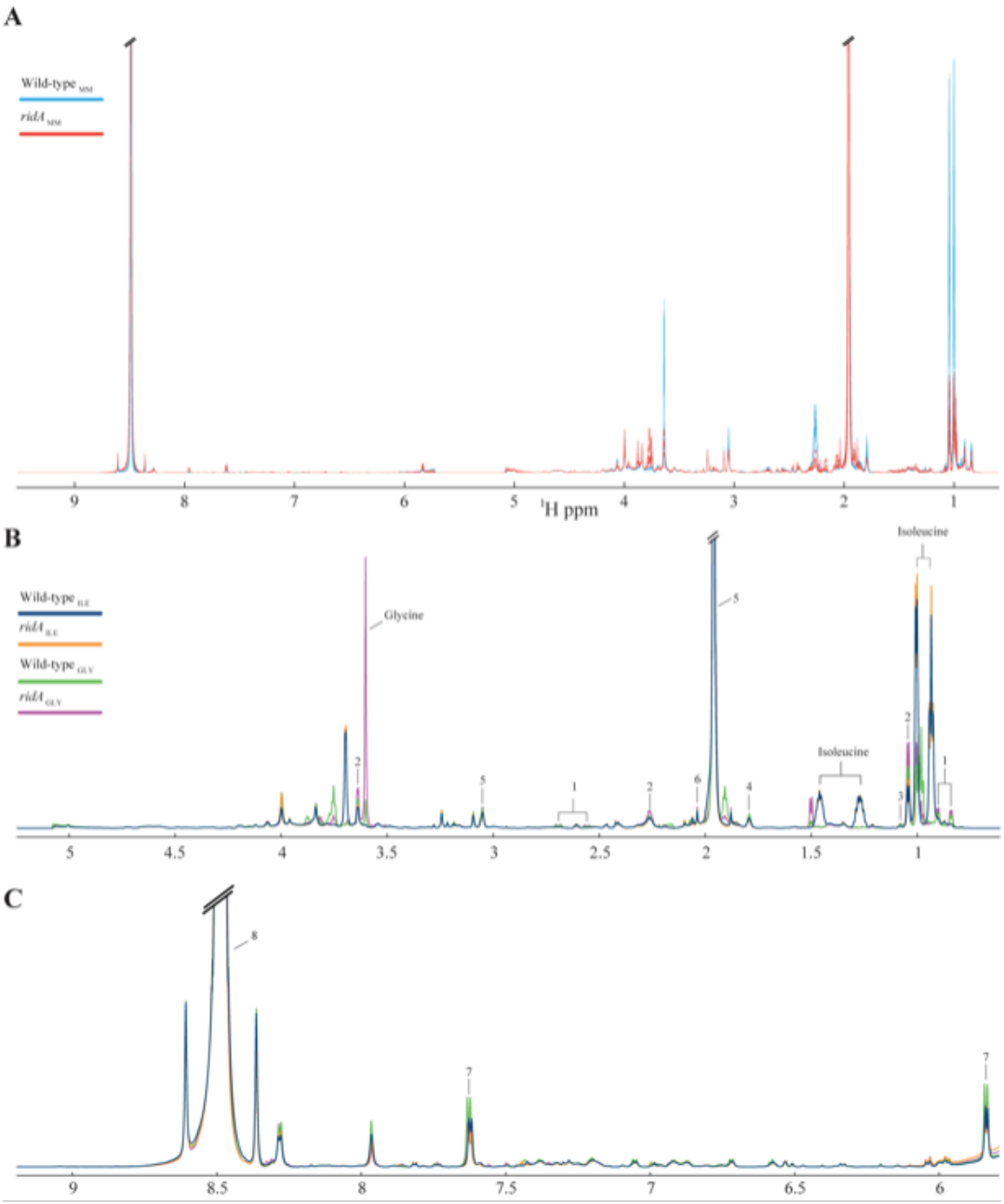
20. Wohlgemuth, G.; Haldiya, P. K.; Willighagen, E.; Kind, T.; Fiehn, O., The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports. *Bioinformatics* **2010,** *26* (20), 2647-2648.

APPENDIX C

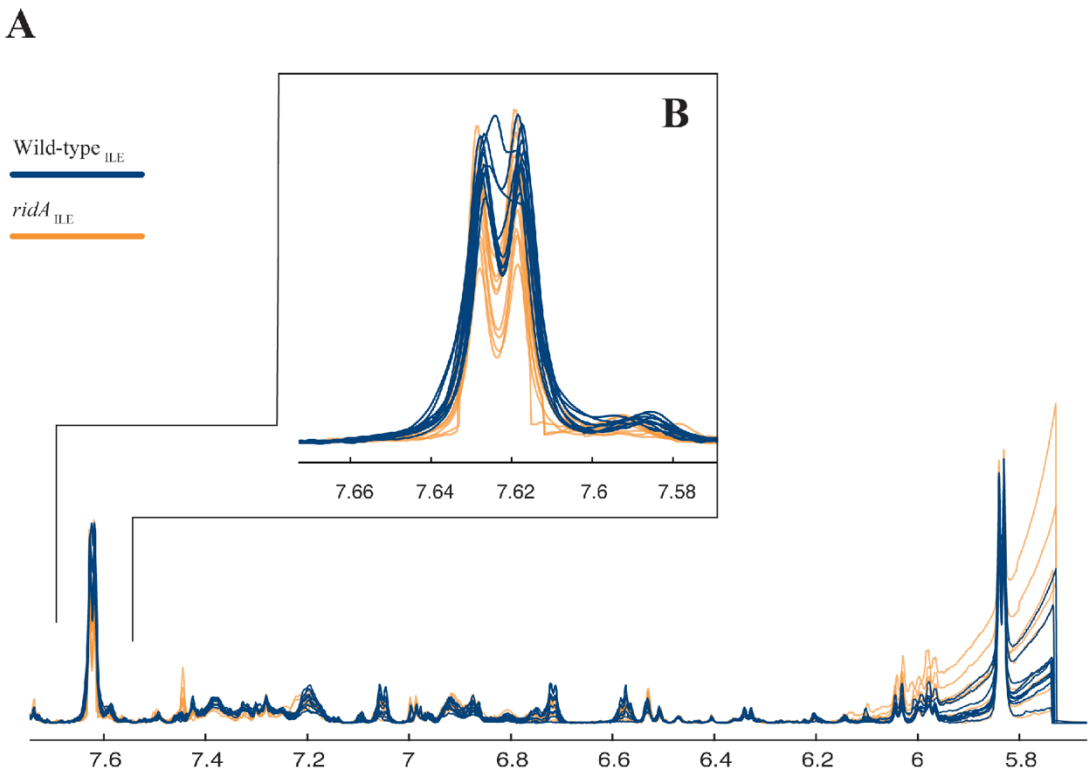
SUPLEMENTAL MATERIAL FOR CHAPTER 4



Supplemental Figure S4.1.[1]H NMR spectra for endogenous samples. (A) Endometabolome overlaid spectra of wild-type samples (n = 10, blue) and *ridA* mutant samples (n = 10, red) after growth in minimal medium. Expansions from 0 to 4.5 ppm (B) and 5 to 9.5 ppm (C) of the mean spectra obtained from wild-type and ridA endogenous samples following growth in minimal medium containing isoleucine and minimal medium containing glycine are also provided. The water region was removed and not displayed (4.7 to 4.8 ppm). Annotations: 1, CoA; 2, valine; 3, threonine; 4, alanine; 5, putrescine; 6, acetic acid; 7, glutamic acid; 8, ethanolamine; 9, N-acetyl-putrescine; 10, uracil; 11, phenylalanine; 12, formic acid.



Supplemental Figure. S4.2. ^{1}H NMR spectra for exogenous samples. (A) Exometabolome overlaid spectra of wild-type samples (n = 10, blue) and ridA mutant samples (n = 10, red), grown in minimal medium. Expansions from 0 to 4.5 ppm (B) and 5 to 9.5 ppm (C) of the mean spectra obtained from wild-type and ridA exogenous samples following growth in minimal medium containing isoleucine and minimal medium containing glycine are also provided. The residual water resonance between 5.1 and 5.7 ppm (due to proton exchange with large amounts of formic acid) was removed during processing. Annotations: 1, isopropylmalic acid; 2, valine; 3, lactic acid; 4, putrescine; 5, acetic acid; 6, acetyl phosphate; 7, uracil; 8, formic acid.



Supplemental figure S4.3. Spectral distortions for samples grown with isoleucine at the uracil region. (A) Exometabolome overlaid spectra expansion from 5.76 ppm to 7.7 ppm of wild-type (n = 10, blue) and ridA mutant (n = 10, orange) samples following growth in minimal medium containing isoleucine. (B) Expansion from 7.58 to 7.66 ppm illustrates the uracil peak used for integration. Distortions of the doublet peak shape are noted for a portion of the wild-type samples. In addition, alignment artifacts create peak shape changes for a portion of the ridA samples. These two factors contribute to the area under the curve calculation.

Supplemental Table. S4.1. Endogenous metabolites identified by ^{1}H-NMR in pellet samples with confidence levels.

Endogenous Metabolites				***Endogenous Metabolites***			
Metabolite	Assignment	^{1}H Chemical shift peaks (ppm) [Multiplicity]	Confidence Level	Metabolite	Assignment	^{1}H Chemical shift peaks (ppm) [Multiplicity]	Confidence Level
Acetate	CH_3	1.912 [s]	4	Nicotinate		8.596 [s]	4
Alanine	CH_3	1.469, 1.484 [d]	4			8.603 [s]	
Coenzyme A	CH_3	0.726 [s]	4			9.289 [s]	
	CH_3	0.853 [s]		Phenylalanine	CH	7.311, 7.323 [d]	4
	CH_2	2.445, 2.456, 2.470 [t]			CH	7.349-7.374 [m]	
	CH_2	3.308 [s]			CH	7.403-7.428 [m]	
	CH	6.182, 6.191 [d]		Putrescine	CH_2	1.770 [m]	4
Ethanolamine	CH_2NH_2	3.128, 3.136, 3.145 [t]	4		CH_2NH_2	3.042 [t]	
Formate	CH	8.453 [s]	4	Pyruvate	CH_3	2.359 [s]	4
Glutamate	αCH_2	2.017 [m]	4	Succinate	CH_2	2.398 [s]	4
	βCH_2	2.334-2.346 [m]		Threonine	CH_3	1.319, 1.330 [d]	4
	CH	3.774 [m]			CH	4.238-4.245 [m]	
Glutamine		2.170 [s]	4	Uracil	CH	5.784, 5.798 [d]	4
N-acetylputrescine	CH_2	1.565-1.602 [m]	4		CHNH	7.525, 7.537 [d]	
	CH_2	1.667-1.695 [m]		Valine	γCH_3	0.976 [d]	4
	CH_3	1.983 [s]			γCH_3	1.028 [d]	
	CH_2NH_2	2.992, 3.006, 3.019 [t]			αCH	3.653, 3.66 [d]	
	CH_2	3.192, 3.203, 3.214 [t]			βCH	2.251-2.291 [m]	

Supplemental Table. S4.2. Endogenous metabolites identified by ^{1}H-NMR in pellet samples with confidence levels.

Exogenous Metabolites				***Exogenous Metabolites***			
Metabolite	Assignment	^{1}H Chemical shift peaks (ppm) [Multiplicity]	Confidence Level	Metabolite	Assignment	^{1}H Chemical shift peaks (ppm) [Multiplicity]	Confidence Level
Lactic acid	CH_3	1.259, 1.266 [d]	4	2-isopropylmalic acid	CH_2	2.547, 2.567, 2.681, 2.700 [d of d]	4
	CH	4.062 [q]			CH_3	0.838, 0.845 [d]	
Valine	αCH	3.636, 3.641 [d]	4		CH_3	0.899, 0.907 [d]	
	βCH	2.253-2.302 [m]		Putrescine	CH_2	1.793 [m]	4
	γCH_3	0.997, 1.006 [d]			CH_2NH_2	3.050 [t]	
	γCH_3	1.039, 1.048 [d]		2-aminobutyric acid	CH_3	0.974, 0.984 [t]	4
Isoleucine	αCH	3.63, 3.64 [d]	4		CH_2	1.90, 1.91, 1.92 [m]	
	CH_2	1.45-1.48 [m]			CH	3.75, 3.75, 3.76 [t]	
	CH_2	1.25-1.30 [m]		Acetate	CH_3	1.96 [s]	4
	CH_3	.926, 0.935, 0.944 [t]		acetyl-phosphate	CH_3	2.095 [s]	4
	CH_3	1.00, 1.01 [d]		Uracil	CH	5.830, 5.839 [d]	4
Formate	CH	8.488 [s]	4		CHNH	7.617, 7.625 [d]	

Supplemental Table S4.3. VIP scores for endogenous PLS-da component 1.

Peak ppm	VIP Score	Peak Identity	Peak ppm	VIP Score	Peak Identity
1.409	2.92		1.565	1.84	N-acetyl putrescine
1.990	2.76		3.192	1.83	N-acetyl putrescine
1.398	2.67		5.010	1.81	
2.500	2.55		7.458	1.81	
1.319	2.52	Threonine	2.489	1.81	
3.462	2.51		4.238	1.80	Threonine
1.330	2.51	Threonine	3.392	1.80	
5.441	2.51		7.720	1.78	
1.516	2.48		8.176	1.77	
2.679	2.43		3.774	1.76	Glutamate
3.479	2.36		5.032	1.75	
2.215	2.33		8.453	1.74	Formate
2.690	2.31		2.303	1.74	
2.667	2.31		3.708	1.73	
4.245	2.28	Threonine	7.428	1.73	Phenylalanine
3.902	2.27		5.017	1.71	
2.512	2.27		1.999	1.71	
2.251	2.27	Valine	3.948	1.71	
5.310	2.24		7.846	1.70	
1.430	2.24		7.415	1.69	Phenylalanine
3.918	2.23		1.288	1.68	
1.028	2.21	Valine	1.591	1.68	N-acetyl putrescine
1.527	2.19		3.622	1.66	
4.539	2.17		3.883	1.66	
3.145	2.13	Ethanolamine	7.311	1.65	Phenylalanine
0.889	2.11		1.454	1.63	
2.266	2.07	Valine	7.444	1.63	
2.717	2.06		6.849	1.62	
3.386	2.05		1.199	1.62	
2.291	2.04	Valine	1.667	1.62	N-acetyl putrescine
1.442	2.03		2.194	1.62	
3.136	2.01	Ethanolamine	2.557	1.59	
2.131	2.01		1.961	1.59	
1.983	1.99	N-acetyl putrescine	7.323	1.58	Phenylalanine
3.875	1.98		4.995	1.57	
3.612	1.97		2.545	1.56	
5.048	1.96		2.170	1.55	Glutamine
1.577	1.95	N-acetyl putrescine	3.006	1.55	N-acetyl putrescine
3.214	1.95	N-acetyl putrescine	6.877	1.54	
2.706	1.95		0.726	1.54	Coenzyme A
1.188	1.93		4.714	1.54	
3.602	1.92		1.680	1.54	N-acetyl putrescine
2.144	1.92		8.253	1.51	
0.976	1.91	Valine	7.895	1.50	
3.203	1.91	N-acetyl putrescine	6.153	1.48	
3.498	1.90		8.421	1.48	
5.319	1.88		1.469	1.48	Alanine
0.928	1.86		1.042	1.47	
2.622	1.85		8.638	1.47	
4.318	1.84		7.911	1.45	

Cont. Supplemental Table S4.3. VIP scores for endogenous PLS-da component 1.

Peak ppm	VIP Score	Peak Identity		Peak ppm	VIP Score	Peak Identity
2.747	1.44			2.316	1.20	
8.076	1.44			0.767	1.20	
1.052	1.44			2.789	1.20	
2.570	1.43			5.235	1.20	
2.964	1.43			2.342	1.18	Glutamate
3.733	1.39			2.346	1.18	Glutamate
7.403	1.39	Phenylalanine		8.088	1.18	
8.234	1.39			5.136	1.18	
2.757	1.39			0.880	1.18	
4.868	1.38			6.066	1.17	
1.770	1.38	Putrescine		4.922	1.17	
3.525	1.36			2.120	1.16	
7.666	1.36			8.043	1.16	
3.683	1.36			4.932	1.15	
2.992	1.36	N-acetyl putrescine		3.232	1.14	
3.375	1.36			4.209	1.14	
				1.695	1.13	N-acetyl putrescine
1.484	1.35	Alanine		2.390	1.13	
6.326	1.35			1.173	1.13	
8.109	1.35			8.031	1.13	
2.930	1.35			1.137	1.11	
6.836	1.34			3.261	1.11	
3.042	1.33	Putrescine		8.381	1.09	
7.374	1.33	Phenylalanine		1.148	1.09	
7.359	1.33	Phenylalanine		2.334	1.08	Glutamate
3.702	1.32			8.364	1.06	
2.071	1.32			2.918	1.06	
8.096	1.31			3.174	1.05	
2.048	1.30			1.855	1.05	
2.940	1.28			3.308	1.04	Coenzyme A
1.602	1.27	N-acetyl putrescine		7.673	1.04	
3.128	1.26	Ethanolamine		2.360	1.04	Pyruvate
2.955	1.26			8.937	1.03	
6.337	1.25			0.810	1.03	
4.478	1.25			3.228	1.03	
2.456	1.25	Coenzyme A		5.296	1.03	
2.157	1.24			2.636	1.03	
8.135	1.24			4.578	1.02	
1.246	1.24			4.424	1.01	
3.594	1.24			6.891	1.00	
2.017	1.24	Glutamate		7.651	1.00	
7.958	1.23			5.984	1.00	
8.330	1.22			7.185	1.00	
4.065	1.22			4.394	0.99	
6.059	1.22			7.172	0.99	
2.798	1.22			0.853	0.98	Coenzyme A
8.221	1.22			5.411	0.98	
3.980	1.22			4.608	0.98	
2.035	1.21			9.433	0.98	
5.229	1.21			2.979	0.97	

Cont. Supplemental Table S4.3. VIP scores for endogenous PLS-da component 1.

Peak ppm	VIP Score	Peak Identity	Peak ppm	VIP Score	Peak Identity
6.191	0.97	Coenzyme A	2.375	0.74	
4.883	0.96		3.277	0.73	
4.473	0.96		1.268	0.73	
4.185	0.96		5.899	0.72	
1.122	0.96		6.182	0.72	Coenzyme A
5.068	0.95		5.247	0.71	
8.923	0.95		1.912	0.71	Acetate
1.946	0.95		1.545	0.70	
3.450	0.94		6.693	0.70	
2.533	0.94		2.108	0.70	
3.350	0.94		4.594	0.70	
2.906	0.93		4.414	0.69	
5.784	0.93	Uracil	3.106	0.69	
1.281	0.92		0.916	0.69	
8.603	0.92	Nicotinate	4.683	0.68	
3.019	0.92	N-acetyl putrescine	6.287	0.67	
5.268	0.91		1.887	0.67	
1.374	0.91		4.292	0.66	
7.502	0.91		4.288	0.66	
5.969	0.90		2.810	0.66	
1.554	0.90		4.174	0.64	
4.600	0.90		3.827	0.63	
2.781	0.90		6.739	0.62	
1.161	0.90		8.596	0.62	Nicotinate
5.494	0.89		1.881	0.62	
6.753	0.89		9.442	0.61	
7.537	0.88	Uracil	4.131	0.60	
5.798	0.87	Uracil	1.624	0.60	
2.445	0.87	Coenzyme A	7.986	0.59	
8.199	0.87		4.351	0.59	
5.406	0.87		3.244	0.59	
8.265	0.87		7.936	0.58	
2.657	0.86		1.361	0.58	
1.108	0.86		5.607	0.58	
6.707	0.85		4.342	0.57	
1.216	0.84		0.956	0.57	
8.142	0.82		2.398	0.57	Succinate
6.167	0.81		2.601	0.57	
4.730	0.79		0.944	0.56	
1.892	0.79		3.292	0.56	
2.770	0.78		3.588	0.55	
5.911	0.78		8.404	0.55	
7.525	0.77	Uracil	7.489	0.55	
1.803	0.77		2.093	0.55	
4.380	0.76		7.998	0.54	
1.234	0.76		4.359	0.54	
4.050	0.76		5.623	0.54	
0.832	0.76		5.616	0.53	
9.251	0.74		7.624	0.53	
7.475	0.74		5.614	0.52	

Supplemental Table S4.3. VIP scores for endogenous PLS-da component 1.

Peak ppm	VIP Score	Peak Identity	Peak ppm	VIP Score	Peak Identity
3.317	0.52		7.950	0.28	
4.407	0.52		0.959	0.28	
5.851	0.51		4.661	0.28	
6.140	0.50		4.100	0.27	
5.504	0.49		3.571	0.27	
2.522	0.48		5.960	0.27	
5.888	0.48		3.559	0.26	
3.436	0.48		3.993	0.25	
1.707	0.48		6.083	0.24	
4.962	0.47		2.590	0.24	
5.363	0.47		3.162	0.22	
6.103	0.45		6.029	0.21	
8.559	0.45		2.422	0.19	
3.405	0.45		8.538	0.19	
1.636	0.45		8.276	0.18	
6.455	0.44		1.725	0.17	
2.470	0.44	Coenzyme A	6.296	0.15	
3.819	0.43		6.038	0.15	
1.501	0.43		6.302	0.14	
0.996	0.43		3.094	0.14	
8.119	0.42		8.551	0.14	
3.659	0.42	Valine	3.842	0.14	
4.853	0.42		5.944	0.13	
5.376	0.41		9.100	0.13	
7.866	0.41		3.540	0.13	
5.169	0.41		3.798	0.13	
2.612	0.40		3.550	0.13	
8.013	0.39		4.466	0.13	
5.089	0.38		6.130	0.13	
1.823	0.38		5.501	0.12	
1.740	0.38		7.832	0.12	
5.842	0.37		3.361	0.12	
3.653	0.36	Valine	9.090	0.08	
1.257	0.35		2.433	0.08	
4.739	0.35		1.084	0.07	
7.788	0.34		3.667	0.06	
4.520	0.34		9.289	0.06	Nicotinate
1.869	0.31		3.632	0.06	
4.839	0.31		3.420	0.05	
6.092	0.31		3.810	0.05	
5.183	0.31		5.873	0.04	
6.512	0.30		6.463	0.04	
6.275	0.29		2.646	0.04	
3.640	0.29		6.118	0.03	
7.639	0.29		4.648	0.02	
5.994	0.28		1.715	0.01	
3.790	0.28				

Supplemental Table S4.4. VIP scores for exogenous PLS-DA plot component 1

Peak ppm	VIP Score	Peak Identity		Peak ppm	VIP Score	Peak Identity
1.259	2.66	Lactate		0.838	1.75	2-isopropylmalate
7.000	2.64			6.988	1.74	
1.266	2.55	Lactate		7.351	1.74	
1.116	2.46			1.915	1.73	2-aminobutyrate
1.124	2.46			6.876	1.71	
1.108	2.43			3.050	1.71	Putrescine
0.963	2.39			0.845	1.71	2-isopropylmalate
7.494	2.38			3.350	1.69	
7.960	2.36			1.906	1.65	2-aminobutyrate
3.641	2.30	Valine		2.320	1.65	
6.938	2.28			7.424	1.65	
2.283	2.27	Valine		0.899	1.64	2-isopropylmalate
1.039	2.26	Valine		3.120	1.63	
1.074	2.26			3.341	1.63	
2.261	2.26	Valine		2.681	1.62	2-isopropylmalate
1.048	2.25	Valine		7.556	1.60	
5.967	2.25			6.868	1.60	
1.006	2.25	Valine		3.144	1.59	
2.275	2.24	Valine		2.700	1.59	2-isopropylmalate
0.997	2.22	Valine		1.793	1.58	Putrescine
0.925	2.21			2.472	1.57	
3.636	2.21	Valine		2.517	1.56	
2.253	2.19	Valine		0.907	1.56	2-isopropylmalate
2.917	2.18			6.201	1.56	
2.267	2.17	Valine		1.409	1.53	
2.922	2.14			4.155	1.53	
7.726	2.12			2.998	1.53	
5.977	2.09			2.095	1.52	Acetyl-phosphate
2.302	2.07	Valine		1.856	1.51	
1.084	2.04			3.441	1.51	
6.031	2.01			7.301	1.50	
0.955	2.00			3.747	1.50	2-aminobutyrate
6.044	1.99			4.574	1.50	
2.464	1.95			0.787	1.49	
8.362	1.94			3.753	1.49	2-aminobutyrate
2.904	1.93			2.650	1.48	
8.607	1.93			1.415	1.48	
8.488	1.92	Formate		0.936	1.47	
7.625	1.91	Uracil		2.435	1.47	
7.736	1.91			8.629	1.47	
4.062	1.90	Lactate		1.870	1.46	
5.830	1.86	Uracil		3.193	1.45	
7.820	1.85			2.567	1.44	2-isopropylmalate
0.871	1.84			4.197	1.43	
5.839	1.83	Uracil		4.148	1.43	
7.809	1.83			2.547	1.43	2-isopropylmalate
7.617	1.82	Uracil		0.795	1.41	
6.926	1.82			0.974	1.39	2-aminobutyrate
7.440	1.79			4.597	1.37	
3.364	1.78			2.606	1.36	

Cont. Supplemental Table S4.4. VIP scores for exogenous PLS-DA plot component 1

Peak ppm	VIP Score	Peak Identity		Peak ppm	VIP Score	Peak Identity
3.185	1.33			1.180	0.96	
2.133	1.32			2.597	0.96	
2.527	1.32			2.065	0.95	
7.747	1.32			3.876	0.94	
3.760	1.31	2-aminobutyrate		2.225	0.91	
8.287	1.30			5.993	0.91	
8.279	1.30			4.324	0.91	
1.400	1.27			3.919	0.90	
2.444	1.26			2.056	0.90	
3.696	1.25			8.383	0.89	
0.984	1.25	2-aminobutyrate		1.388	0.88	
6.508	1.23			5.727	0.88	
2.616	1.22			7.281	0.87	
4.250	1.21			3.901	0.87	
4.275	1.18			0.944	0.86	
3.429	1.18			3.510	0.86	
8.678	1.15			1.731	0.85	
3.687	1.14			1.863	0.84	
2.730	1.14			2.171	0.84	
4.143	1.13			4.092	0.84	
4.590	1.12			5.740	0.82	
7.344	1.11			3.016	0.82	
8.304	1.10			5.942	0.81	
6.585	1.10			2.012	0.81	
4.513	1.09			2.836	0.81	
2.421	1.08			3.542	0.80	
2.313	1.08			1.722	0.79	
4.370	1.07			3.414	0.79	
4.345	1.07			2.400	0.78	
6.722	1.06			3.869	0.77	
3.784	1.06			4.861	0.77	
7.057	1.05			1.379	0.77	
6.566	1.05			2.183	0.76	
1.623	1.05			4.204	0.76	
7.046	1.05			5.939	0.76	
2.501	1.05			6.893	0.75	
3.700	1.05			2.345	0.75	
1.367	1.04			2.722	0.75	
4.193	1.04			3.726	0.73	
6.575	1.03			4.485	0.73	
3.769	1.03			4.622	0.73	
6.714	1.02			7.590	0.70	
2.163	1.02			5.947	0.70	
3.777	1.02			3.930	0.70	
3.176	1.01			5.735	0.69	
3.883	1.00			3.331	0.69	
1.217	0.99			8.327	0.68	
6.916	0.98			5.067	0.67	
1.209	0.98			4.456	0.66	
3.798	0.97			4.580	0.64	

Cont. Supplemental Table S4.4. VIP scores for exogenous PLS-DA plot component 1

Peak ppm	VIP Score	Peak Identity		Peak ppm	VIP Score	Peak Identity
6.001	0.62			1.234	0.30	
7.598	0.59			5.082	0.30	
3.155	0.56			2.371	0.30	
4.358	0.56			3.004	0.27	
3.589	0.54			4.887	0.27	
5.069	0.53			3.216	0.27	
1.533	0.51			1.188	0.25	
3.672	0.51			3.475	0.25	
3.579	0.51			3.243	0.25	
1.578	0.50			4.878	0.24	
3.499	0.49			7.249	0.23	
1.575	0.48			2.073	0.22	
4.040	0.47			4.185	0.22	
1.960	0.46	Acetate		3.960	0.21	
2.207	0.46			2.751	0.18	
4.565	0.46			4.916	0.17	
1.245	0.45			4.381	0.16	
2.195	0.45			2.938	0.16	
4.445	0.44			1.138	0.16	
7.377	0.43			3.837	0.15	
1.445	0.42			6.860	0.14	
3.944	0.42			6.341	0.14	
3.459	0.42			3.280	0.13	
4.715	0.41			7.416	0.13	
2.821	0.41			2.352	0.11	
3.093	0.41			3.025	0.11	
1.353	0.41			1.281	0.10	
2.789	0.39			1.465	0.09	
2.490	0.39			2.410	0.09	
2.805	0.39			6.328	0.08	
2.217	0.39			4.724	0.08	
3.168	0.38			1.692	0.07	
3.161	0.38			5.006	0.07	
1.552	0.38			0.000	0.06	
1.879	0.38			1.608	0.05	
7.384	0.37			1.290	0.05	
7.290	0.37			1.095	0.05	
4.422	0.36			4.735	0.05	
1.500	0.35			7.324	0.04	
3.487	0.34			1.489	0.04	
3.997	0.34			3.398	0.04	
7.203	0.34			3.302	0.03	
1.543	0.34			1.474	0.03	
6.531	0.34			3.308	0.03	
7.380	0.31			4.119	0.03	
3.601	0.30			3.711	0.02	
2.038	0.30			1.685	0.00	

Supplemental Table S4.5. Endogenous metabolites integration values and descriptive statistics.

Minimal

	Coenzyme A	Valine	Threonine	Alanine	Putrescine	Acetate	Glutamine	Glutamate	Pyruvate	Succinate	Ethanolamine	N-acetylputrescine	Uracil	Phenylalanine	Formate	Nicotinate
Peak_ppm	0.726	1.028	1.331	1.470	1.770	1.912	2.170	2.342	2.360	2.399	3.137	3.203	5.798	7.311	8.453	8.596
WT_mean	0.305	0.149	0.155	0.214	0.921	1.026	0.079	0.507	0.028	0.083	0.144	0.072	0.169	0.021	0.216	0.028
WT_Stdev	0.057	0.057	0.021	0.082	0.270	0.331	0.027	0.167	0.008	0.061	0.014	0.018	0.046	0.006	0.085	0.009
Mut_mean	0.189	0.518	0.486	0.304	0.593	0.691	0.190	0.747	0.037	0.121	0.110	0.138	0.216	0.039	0.057	0.034
Mut_Stdev	0.037	0.130	0.155	0.051	0.201	0.243	0.134	0.205	0.014	0.114	0.013	0.042	0.077	0.010	0.034	0.008
Fold_Change	0.619	3.466	3.127	1.418	0.644	0.674	2.402	1.474	1.310	1.444	0.763	1.918	1.280	1.830	0.263	1.228
p_value	3.67E-05	1.72E-07	2.82E-06	9.00E-03	6.34E-03	1.88E-02	1.94E-02	1.00E-02	1.01E-01	3.76E-01	1.92E-05	2.56E-04	1.12E-01	1.19E-04	3.43E-05	1.12E-01
FDR_value	1.17E-04	2.75E-06	2.25E-05	1.60E-02	1.27E-02	2.59E-02	2.59E-02	1.60E-02	1.20E-01	3.76E-01	1.02E-04	5.84E-04	1.20E-01	3.18E-04	1.17E-04	1.20E-01

Minimal_Gly

	Coenzyme A	Valine	Threonine	Alanine	Putrescine	Acetate	Glutamine	Glutamate	Pyruvate	Succinate	Ethanolamine	N-acetylputrescine	Uracil	Phenylalanine	Formate	Nicotinate
Peak_ppm	0.726	1.028	1.331	1.470	1.770	1.912	2.171	2.342	2.360	2.398	3.137	3.203	5.798	7.311	8.453	8.596
WT_mean	0.249	0.202	0.185	0.317	0.773	0.910	0.100	0.722	0.046	0.077	0.137	0.087	0.194	0.026	0.184	0.028
WT_Stdev	0.057	0.066	0.054	0.111	0.187	0.251	0.067	0.138	0.013	0.064	0.024	0.019	0.046	0.006	0.100	0.009
Mut_mean	0.190	0.259	0.229	0.418	0.784	1.018	0.116	0.736	0.037	0.089	0.142	0.083	0.189	0.033	0.156	0.029
Mut_Stdev	0.042	0.116	0.085	0.055	0.294	0.292	0.090	0.232	0.020	0.053	0.011	0.044	0.053	0.016	0.071	0.008
Fold_Change	0.761	1.282	1.240	1.318	1.014	1.118	1.164	1.019	0.807	1.157	1.037	0.958	0.975	1.292	0.846	1.011
p_value	1.56E-02	1.95E-01	1.80E-01	1.89E-02	9.23E-01	3.87E-01	6.52E-01	8.71E-01	2.51E-01	6.54E-01	5.50E-01	8.13E-01	8.28E-01	1.74E-01	4.77E-01	9.31E-01
FDR_value	1.51E-01	6.25E-01	6.25E-01	1.51E-01	9.31E-01	8.85E-01	9.31E-01	9.31E-01	6.69E-01	9.31E-01	9.31E-01	9.31E-01	9.31E-01	6.25E-01	9.31E-01	9.31E-01

Minimal_Ile

	Coenzyme A	Valine	Threonine	Alanine	Putrescine	Acetate	Glutamine	Glutamate	Pyruvate	Succinate	Ethanolamine	N-acetylputrescine	Uracil	Phenylalanine	Formate	Nicotinate
Peak_ppm	0.726	1.028	1.331	1.470	1.770	1.912	2.172	2.342	2.360	2.398	3.136	3.203	5.798	7.311	8.453	8.603
WT_mean	0.199	0.082	0.227	0.323	0.751	0.874	0.065	0.521	0.028	0.059	0.145	0.088	0.215	0.035	0.174	0.027
WT_Stdev	0.066	0.036	0.072	0.106	0.292	0.383	0.017	0.141	0.007	0.033	0.020	0.054	0.062	0.018	0.130	0.010
Mut_mean	0.248	0.081	0.211	0.286	0.901	1.146	0.068	0.483	0.029	0.079	0.152	0.065	0.202	0.028	0.248	0.025
Mut_Stdev	0.122	0.035	0.048	0.072	0.306	0.319	0.026	0.149	0.010	0.050	0.011	0.027	0.053	0.010	0.078	0.009
Fold_Change	1.242	0.988	0.931	0.885	1.199	1.311	1.046	0.926	1.013	1.344	1.048	0.742	0.940	0.786	1.430	0.940
p_value	2.84E-01	9.50E-01	5.78E-01	3.70E-01	2.77E-01	1.02E-01	7.68E-01	5.60E-01	9.26E-01	3.01E-01	3.47E-01	2.49E-01	6.21E-01	2.71E-01	1.37E-01	7.04E-01
FDR_value	6.57E-01	9.50E-01	8.28E-01	6.57E-01	6.57E-01	6.57E-01	8.77E-01	8.28E-01	9.50E-01	6.57E-01	6.57E-01	6.57E-01	8.28E-01	6.57E-01	6.57E-01	8.67E-01

Fold_change is expressed as (*ridA* /WT)

Supplemental Table S4.6. Exogenous metabolites integration values and descriptive statistics.

Minimal

	2-isopropylmalate	Valine	2-aminobutyrate	Acetate	Acetyl-phosphate	Putrescine	Lactate	Uracil	Formate
Peak_ppm	0.838	1.040	1.907	1.960	2.096	3.051	4.063	7.626	8.488
WT_mean	0.215	1.231	0.294	42.503	0.040	0.544	0.134	0.145	38.811
WT_Stdev	0.082	0.330	0.184	15.001	0.005	0.057	0.052	0.025	5.409
Mut_mean	0.510	6.118	N/A	42.479	0.054	0.789	0.274	0.046	14.080
Mut_Stdev	0.122	0.928	N/A	9.322	0.007	0.159	0.024	0.008	5.894
Fold_Change	2.368	4.971	N/A	0.999	1.356	1.449	2.044	0.319	0.363
p_value	5.55E-06	6.10E-12	Undetermined	9.97E-01	3.80E-05	2.38E-04	4.21E-07	7.52E-10	1.27E-08
FDR_value	9.99E-06	5.49E-11	Undetermined	9.97E-01	5.70E-05	2.68E-04	9.47E-07	3.39E-09	3.80E-08

Minimal_Gly

	2-isopropylmalate	Valine	2-aminobutyrate	Acetate	Acetyl-phosphate	Putrescine	Lactate	Uracil	Formate
Peak_ppm	0.839	1.040	1.908	1.961	2.097	3.051	4.064	7.633	8.489
WT_mean	0.270	1.364	0.985	32.242	0.041	0.504	0.072	0.139	34.911
WT_Stdev	0.095	0.505	0.394	17.838	0.008	0.083	0.016	0.020	5.383
Mut_mean	0.413	1.969	0.152	41.363	0.045	0.538	0.149	0.095	34.432
Mut_Stdev	0.209	0.447	0.135	13.936	0.004	0.160	0.024	0.026	7.504
Fold_Change	1.528	1.443	0.154	1.283	1.085	1.067	2.073	0.687	0.986
p_value	6.52E-02	1.10E-02	5.85E-06	2.19E-01	2.58E-01	5.62E-01	9.33E-08	6.21E-04	8.72E-01
FDR_value	1.17E-01	2.47E-02	2.63E-05	3.28E-01	3.31E-01	6.33E-01	8.40E-07	1.86E-03	8.72E-01

Minimal_Ile

	2-isopropylmalate	Valine	2-aminobutyrate	Acetate	Acetyl-phosphate	Putrescine	Lactate	Uracil	Formate	Isoleucine
Peak_ppm	0.846	1.040	1.921	1.962	2.099	3.051	4.065	7.619	8.489	1.271
WT_mean	0.048	0.824	N/A	30.452	0.044	0.387	0.112	0.103	33.419	1.858
WT_Stdev	0.039	0.307	N/A	15.542	0.004	0.043	0.034	0.014	3.208	0.193
Mut_mean	0.049	0.928	N/A	27.515	0.080	0.368	0.108	0.078	30.654	1.841
Mut_Stdev	0.036	0.357	N/A	15.970	0.066	0.088	0.021	0.018	11.684	0.165
Fold_Change	1.016	1.127	N/A	0.904	1.817	0.949	0.964	0.757	0.917	0.990
p_value	9.66E-01	5.07E-01	Undetermined	6.90E-01	1.20E-01	5.51E-01	7.56E-01	4.19E-03	5.02E-01	8.32E-01
FDR_value	9.66E-01	9.18E-01	Undetermined	9.25E-01	6.01E-01	9.18E-01	9.25E-01	4.19E-02	9.18E-01	9.25E-01

N/A- Peak for one or more samples was not detectable above baseline signal
Undetermined - Fold-change could not be calculated since one or both conditions did not show peak instensity above the baseline;
Fold_change is expressed as (*ridA* /WT)