STATE-OF-THE-ART METABOLOMICS: FROM METHOD DEVELOPMENT TO RAPID DIAGNOSTICS, DETECTION, AND PROFILING

by

Allyson L. Dailey
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Biochemistry

Committee:

_____	Dr. Robin Couch, Dissertation Director
_____	Dr. Barney Bishop, Committee Member
_____	Dr. Geraldine Grant, Committee Member
_____	Dr. John Schreifels, Committee Member
_____	Dr. Joseph Urban, Committee Member
_____	Dr. Gerald L.R. Weatherspoon, Department Chairperson
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: _____	Summer Semester 2017 George Mason University Fairfax, VA

State-of-the-Art Metabolomics: From Method Development to Rapid Diagnostics, Detection, and Profiling

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Allyson L. Dailey
Master of Science
George Mason University, 2013
Bachelor of Science
George Mason University, 2011

Director: Robin Couch, Professor
Department of Chemistry and Biochemistry

Summer Semester 2017
George Mason University
Fairfax, VA

DEDICATION

To my loving and supportive husband Jonathan, who has been by my side since we were kids. To my parents, for never letting me forget all the places I will go in life. To my brother Jamie and my sister Tissy, for making me tough and resilient. And finally to my right-hand man Jasper and my faithful "little wolf" Faolan for giving the best licks and cuddles.

ACKNOWLEDGEMENTS

This dissertation could not have been possible without the love and continuous support of my friends and family. To all of the past and present members of the Couch lab, especially Justin, Amanda, Haley, Fatima, and Hameed, thank you for your assistance, advice, and being the very best lab family. To Charlie Wells and Jason Calhoun, thank you for encouraging and inspiring me all those years ago to pursue a degree in biochemistry. And finally to my mentor Dr. Couch, thank you for your constant support, guidance, and being the best mentor a student could ask for.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Chemical Ionization CI
Collision Induced Dissociation ..... CID
Extracted Ion Chromatogram ..... EIC
Electron Impact ..... EI
Electrospray Ionization ..... ESI
Gas Chromatography ..... GC
Flame Ionization Detector ..... FID
Headspace Solid-Phase Microextraction ..... hSPME
Hierarchical Clustering Analysis ..... HCA
Human Metabolome Database ..... HMDB
Hydrophilic Interaction Chromatography ..... HILIC
Liquid Chromatography Mass Spectrometry ..... LCMS
Liquid Chromatography Quadrupole Time of Flight ..... LC-QToF
Microbial Volatile Organic Compound ..... mVOC
National Institute of Standards and Technology ..... NIST
Nuclear Magnetic Resonance ..... NMR
Partial Least Squares Discriminate Analysis ..... PLS-DA
Principal Component Analysis ..... PCA
Proximal Colon Mucosa ..... PCM
Receiver Operating Characteristics ..... ROC
Reversed Phase ..... RP
Simultaneous Solid-Phase Microextraction ..... SMULTI
Solid-Phase Microextraction ..... SPME
Tandem Mass Spectrometry ..... MS/MS
Total Ion Chromatogram ..... TIC
Ulcerative Colitis ..... UC
Volatile Organic Compound ..... VOC

ABSTRACT

STATE-OF-THE-ART METABOLOMICS: FROM METHOD DEVELOPMENT TO RAPID DIAGNOSTICS, DETECTION, AND PROFILING
Allyson L. Dailey, Ph.D.

George Mason University, 2017

Dissertation Director: Dr. Robin Couch

Metabolomics is a science concerned with the isolation and identification of small molecules (metabolites) from biological samples. A common goal of metabolomics is to investigate these molecules in order to better understand the intricate interactions between them and their relationship to human health. Metabolomic profiling has shown great potential in detecting and understanding biomarkers of a variety of diseases, of infectious or non-infectious origin. The goal of this dissertation research was to develop a state-of-the-art metabolomics approach to the investigation of metabolite alterations in juxtaposed biological cohorts (eg. healthy vs. disease; infected vs. uninfected). To establish a metabolomics pipeline of data acquisition and analysis, a series of research projects were pursued within the general context of rapid diagnostics, pathogen detection, and/or molecular profiling. From gastrointestinal health to identification of bacterial infection, this dissertation research aimed to develop the tools necessary for rapid-diagnostics of an array of diseases.

INTRODUCTION

Since the advent of medicine, physicians have called upon various tools for rapid and accurate disease diagnosis. From Aristotle's idea of the four humors to the examination of the color, smell, and consistency of excrement to determine a patient's disease state, there has always been this notion of a "biomarker" and its relation to human health[1,2]. As medicine and science have evolved, the desire for elucidating a biomarker has flooded the journals. Though some of these molecules have been validated and are now routinely tested as a true measurement of disease, many act as mere puzzle pieces in our understanding of the complexities of disease. By identifying and/or monitoring metabolite concentration and presence, conclusions can be drawn about the potential phenotypic changes occurring within the patient, which often complements information derived from genomics, transcriptomics, and/or proteomics[3,4].

The following overview will introduce metabolomics as a tool to understanding changes in health and disease, and will discuss the instrumentation often used in a metabolomics analysis.

Metabolomics: An Overview

The goal of metabolomics is to investigate small molecule metabolites and understand the intricate interactions among them. Metabolomic profiling has shown great potential in detecting and understanding biomarkers of a variety of diseases, which may

or may not have an infectious origin[5,6]. For example, through metabolomic profiling, it is possible to screen for various protozoan infections (Nuclear Magnetic Resonance (NMR)), the presence of candidiasis (Gas Chromatography-Mass Spectrometry (GC-MS)), and early stages of breast cancer (NMR) using human urine[7–9]. The elucidation of altered metabolites related to atherosclerosis has been studied using plasma or serum samples (GC-Flame Ionization Detector (GC-FID), NMR, Liquid Chromatography-Mass Spectrometry (LC-MS))[10]. In addition, the identification of changes to the porcine intestinal tract upon whipworm (*Trichuris suis*) infection is possible using collected fecal material[11]. Further, metabolomic profiling, whether achieved through a broad spectrum NMR study or lipidomics using LC-MS, has also been shown to be promising in the detection of schizophrenia. In this case, the profiles obtained can be linked to side effects of antipsychotic drugs as well as metabolite changes due to natural disease progression[12,13].

A metabolomics analysis is typically based on GC-MS, LC-MS, and/or NMR platforms, and is often targeted (i.e. focused on a group of specific analytes) or global in design (i.e. profiling of all detectable analytes (also known as non-targeted)). Regardless of the approach, the metabolites are generally extracted from the biological sample (e.g. tissue, feces, breath, bacterial culture, etc) and then analyzed by LC, GC, or NMR. The resulting chromatograms/spectra subsequently undergo data processing, chemometrics, and statistical analysis to identify key metabolites of interest (Figure 1).

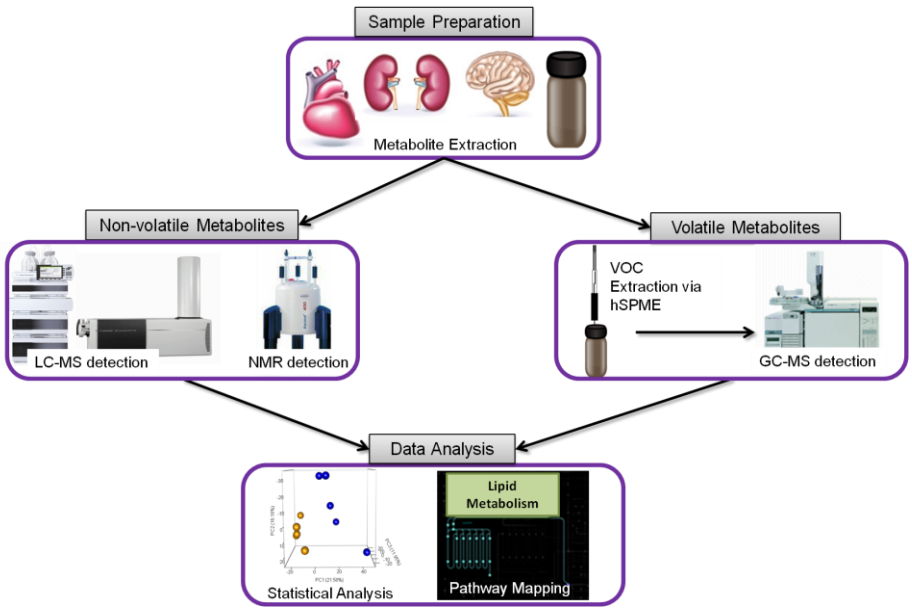


Figure 1. A Typical Approach to a Metabolomics Analysis.
Illustrated here is a flowchart detailing the components of a typical metabolomics investigation. The first stage of analysis is the preparation of the sample and metabolite extraction. Depending on the course of the study, there are two stages to this analysis: analysis of volatile metabolite or analysis of non-volatile metabolites. To extract the volatile metabolites, I employ the use of headspace solid-phase microextraction (hSPME) and subsequently use GC coupled to an MS detector or flame ionization detector (FID, not shown). To extract the non-volatile metabolites, I use an organic extraction and subsequently run the extractions on our LC-QToF (labeled here as the LC-MS). Following data acquisition, molecular features (metabolites) are extracted from the chromatograms and spreadsheets are generated for further analysis. The data analysis pipeline will be described in detail later, but encompasses multivariate and univariate statistical analysis, the results of which can be used to determine alterations to biochemical pathways in juxtaposed biological systems.

A metabolomics investigation typically focuses on both analyte presence and its relative abundance, and as such, the resulting data set is multidimensional. Accordingly, multivariate statistical techniques (e.g., Principal Component Analysis and Hierarchical Clustering Analysis) are employed to analyze the data[14]. These statistical techniques make it possible to explore what is occurring within the data and identify trends.

The final phase of a metabolomics investigation often involves biomarker identification or pathway analysis. These can be performed by looking at the area under a

Receiver Operating Characteristics (ROC) curve to evaluate a metabolite's potential for serving as a biomarker of disease, or through the use of Network Analysis, which allows the visualization of relationships and interactions amongst various metabolites to further understand the disease[15–19].

1. Metabolite Extraction

In the course of a metabolomics investigation, the metabolites are typically extracted from the sample matrix and are then analyzed. Biological molecules of interest fall into two broad categories: volatile metabolites (also known as volatile organic compounds (VOCs)) and non-volatile metabolites (Figure 1). For the volatile analyses, solid-phase microextraction (SPME) is used to extract and concentrate the VOCs present in the headspace above the sample. On the other hand, non-volatile metabolites are generally extracted using an organic solvent, following cryogenic homogenization of the sample via a bead mill. The specific approach to metabolite extraction can however be project specific, and will be elaborated on throughout this dissertation.

2. Metabolite Detection

Following metabolite extraction, the extracts are separated/resolved using gas (volatile metabolites) or liquid (non-volatile metabolites) chromatography[20–23]. The chromatographic separation of the analytes within the complex mixtures typically enables better detection and identification of the individual analytes.

2.1 GC-Based Analysis of Volatile Metabolites

Following extraction of the volatile metabolites, a Gas Chromatograph (GC) is commonly used to separate/resolve the complex mixture. To accomplish this, the VOCs are typically captured and concentrated using SPME, the SPME fibers are then inserted into the inlet of the GC, the VOCs desorb off of the fiber into the chromatography column and are then resolved by gas-liquid partition chromatography (commonly using a capillary column (5-100 m in length)). Partitioning of the volatile analytes is achieved via their interaction with the stationary phase of the capillary column. For example, a column containing (5%Phenyl)-methylpolysiloxane (DB-5) will retain nonpolar analytes. By starting the analysis at low temperatures (35°C), the analytes with a higher affinity to the stationary phase (low polarity) are retained while the others (high polarity) flow through. By gradually increasing the temperature, retained nonpolar analytes begin to elute of the stationary phase and through the column, typically in order of their relative volatility. Following separation, the analytes travel into the detector. A variety of different detectors are available for use in a GC based analysis of VOCs, two of the most common being the Flame Ionization Detector and the Mass Spectrometer.

2.1.1 Flame Ionization Detector (FID)

The FID employs a hydrogen fueled flame to burn the hydrocarbons eluting from GC column, producing detectable ions (Figure 2). More specifically, the analytes elute into a bias potential electrode, undergo pyrolysis, and become ionized[24]. The newly formed ions accumulate on a collecting electrode (conventionally the cathode) located adjacent to the tip of the flame. The resulting ion current is amplified and converted into

a chromatogram. Though it is a highly sensitive detector, an important caveat for the use of the FID is that the sample must contain hydrocarbons that are ionizable by the flame[25]. This is typically the case for biological samples.

Examples of the use of the FID in VOC analyses includes evaluation of the human fecal metabolome performed in the Couch lab[26], detection of coffee flavor ageing[27], and environmental applications including wastewater monitoring[28].

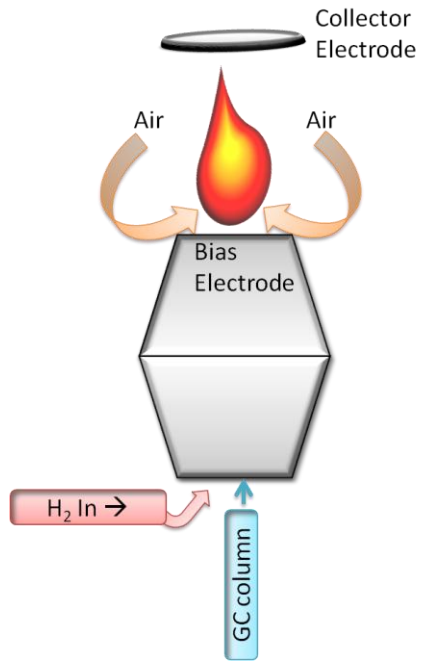


Figure 2. Schematic depicting the components of the flame ionization detector.
The analyte passes through the GC column and into the FID. Inside, the hydrogen flame burns the analyte, producing ions which collect on the collector electrode. The resulting current is then amplified and digitized by the computer.

2.1.2 Mass Spectrometer (MS)

Another common detector used in GC-based analysis of volatile metabolites is the mass spectrometer (MS). For example, GC-MS has been utilized in metabolomic profiling of breath composition in lung cancer[5,29] and cystic fibrosis patients[30,31], examination of the human fecal metabolome[26,32,33], alterations to intestinal mucosa in infected pigs[11], detection of candidiasis[7] and bacterial strains[6], and in the identification of VOCs found in wine[34]. The MS generates and separates gas-phase ions by their mass to charge (m/z) ratio, facilitating the identification of molecules of interest[35]. As the production of ions is the foundation to the analysis, an important caveat of MS analysis is that the analytes must be readily ionizable[36–38]. To achieve this, the MS has three main components, resulting in a three step process:

1. *Ion Source*

The first step is ionization, which occurs in the ion source. In GC-MS, there are two main approaches to ionization: Electron Impact (Electron Ionization, EI) and Chemical Ionization (CI). EI, characterized as a hard ionization, occurs when the VOCs enter the ion source following separation via the column and are bombarded by high energy electrons emitted from a heated filament (typically comprised of tungsten or rhenium). The collision of the analyte and electrons results in the expulsion of an analyte's electron causing the analyte to become a radical cation (Figure 3). This collision of electron and analyte is not really an impact but rather an interaction between a wave and chemical bond. Each electron is characterized as a wave with a specific wavelength at a particular kinetic energy (e.g., 1.4 Å for 70 eV). If the wavelength of the electron is

close to any of the bond lengths within the analyte, the wave is perturbed and an energy transfer can occur. If enough energy is transferred, the analyte's electron is expelled, resulting in a radical cation. As the energy transferred is between 10-20 eV, any excess energy leads to fragmentation of the parent molecule. This fragmentation results in a characteristic spectral "fingerprint" that is then used to identify molecules. However, due to the nature of the collision, the fragmentation patterns can vary with the energy of the electron. To circumvent this variability, the accepted standard of electron energy for populating molecular databases is 70 eV[36–38].

$$A_{(g)} + e^- \rightarrow A_{(g)}^{+\bullet} + 2e^-$$

Figure 3. Chemical equation depicting electron impact ionization.
Analyte (A) collides with the electron (e^-) to form a radical cation ($A^{+\bullet}$).

CI, a soft ionization technique, occurs when a known reagent gas, separate from the sample molecules, is introduced into the ion source. The gas (usually methane, iso-butane, or ammonia) is then bombarded by the electrons emitted by a heated filament and ionizes. When the sample analyte is introduced into the ion source, the analyte then reacts with the now ionized reagent gas and thus becomes ionized itself (Figure 4). As the energy used for ionization is lower than in EI, there is very little subsequent fragmentation that results. Rather, the resulting mass spectrum acquired typically

contains adducts of the intact analyte ion, formed as a result of proton transfer from the reagent gas and the analyte[36–38].

$$NH_3 + e^- \rightarrow {NH_3}^{+\bullet} + 2e^-$$
$${NH_3}^{+\bullet} + NH_3 \rightarrow {NH_4}^+ + {NH_2}^\bullet$$
$${NH_4}^+ + A \rightarrow NH_3 + [A+H]^+$$
$${NH_4}^+ + A \rightarrow [A+NH_4]^+$$

Figure 4. Chemical equation depicting chemical ionization.
Ammonia as the reagent gas collides with the electron to form an ammonia radical cation. The radical cation then reacts with additional ammonia gas to form ammonium and an amino radical. Ammonium reacts with the analyte (A) to form two different species.

2. Mass Analyzers

Following ionization, ions traverse the mass analyzer where they are separated/filtered according to their masses. While mass is predominately measured and subsequently separated or filtered via the m/z ratio of the ions (as seen in the quadrupole), mass analyzers such as those based on time of flight use principles such as velocity to separate ions (the time of flight mass analyzer is described in the LC-MS section of the thesis, below). Therefore, a fundamental difference between mass analyzers is the way in which a static and/or oscillating electromagnetic field is used to achieve proper separation of the ions.

Quadrupole mass analyzers are made up of four circular which are held parallel to each other but oriented to approximate a hyperbolic path (Figure 5). The mass analyzer

separates and filters ions according to the ion's m/z ratio by tracking the stability of the ion's trajectory as the ion passes through an oscillating electromagnetic field.

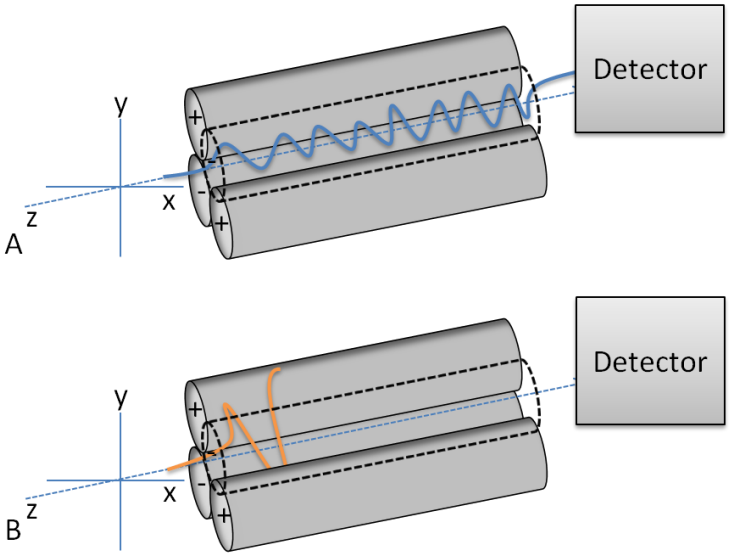


Figure 5. Schematic depicting the Quadrupole mass analyzer.
A. The blue line demonstrates an ion with a stable trajectory.
B. The orange line demonstrates an ion whose trajectory became unstable and will be filtered out. The instability is a byproduct of the acceleration of the ion along the x and y axis. Depending on the ion's mass, this rate of acceleration will be either too fast or too slow causing the ion to become dysregulated with respect to the rate of the rod's change in potential.

The ions do not travel along a straight line, but rather in a three-dimensional space (Figure 5). Following ionization from the source, the ion enters the space between the rods (z-axis) at a fixed velocity. As the ion travels through the space, it will be drawn towards an oppositely charged rod at either the x or y-axis in a circular fashion. As the ion approaches the charged rod, the potential of the rod changes polarity, resulting in the ion changing direction. However, if the ion collides with the rod prior to this occurring,

the ion will become discharged and therefore filtered out. Thus, along the z-axis, the ions maintain their velocity whereas, along the x and y-axis they are subjected to acceleration from the alterations of the electric fields[37–39]. As the rate of acceleration is proportional to the mass of the ion, only those ions with the stable trajectory pattern will be included in the analysis. The trajectory of those ions which are too light or too heavy will become destabilized, thus filtering them from the analysis[38].

3. Detectors

After the ions are separated by the mass analyzer, they are detected and the signal produced is transformed by an ion detector. As the quadrupole is a scanning instrument, it detects the ions of one mass individually. Therefore, the detector that is coupled to this mass analyzer must be capable of measuring the signal produced in this approach. Continuous-dynode electron multipliers is the ion detector used for our quadrupole (Figure 6). Following the separation of the ions, the ions enter the electron multiplier and collide into an electrode called the continuous dynode[38]. This electrode is set at a high potential with a charge opposite to the ion. As the ions collide into the electrode, secondary particles (electrons) are emitted. The secondary particles are then converted into electrons by colliding with the inner wall of the detector. As the electrons flow down the electron multiplier, they continuously collide with the wall, because there is a continuous voltage drop that attracts the secondary electrons to the outlet, thereby producing more electrons and amplifying the signal by as much as a factor of 10^8. Finally, at the very end of the electron multiplier tube, a metal anode collects the

incoming electrons. The resulting current is then measured and digitized by a computer via an analog-to-digital converter (ADC)[37,38].

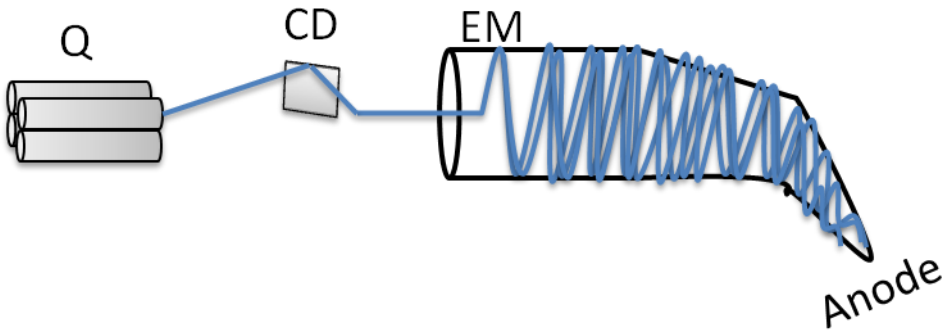


Figure 6. Schematic depicting the components of the continuous dynode electron multiplier. The blue line depicts the ion as it travels from the quadrupole (Q) and into the conversion dynode (CD) where it is converted into secondary electrons. From here, the electrons enter the electron multiplier (EM) where the electron collides into the wall of the electron multiplier, thereby, producing additional electrons. This process continues as the electrons travel down to the anode resistor where they are detected.

2.2 LC Based Analysis of Non-Volatile Metabolites

Following the extraction of the non-volatile analytes, a Liquid Chromatograph (LC) is often used to separate the complex extract. In the LC, a liquid mobile phase is continuously pumped into the system where it mixes and interacts with the liquid sample extract as it is injected into the instrument (either manually or via an autosampler)[40]. The mobile phase and sample then travel to a packed (stationary phase) column where separation occurs. In reversed phase liquid chromatography, the stationary phase is comprised of non-polar alkyl hydrocarbons (typically of C-18, C-8 or C-5 in carbon chain length) bound to an inert support (typically silica), while the mobile phase is comprised

of an aqueous and an organic component. Initially, the instrument begins with a higher concentration of the aqueous solvent resulting in a polar mobile phase. As the mixture of polar mobile phase and sample flows into the column, those analytes with greater affinity for the column (in this case higher hydrophobicity), will interact and bind to the stationary phase of the column while the polar (hydrophilic) analytes will remain interacting with the mobile phase and flow through. As the concentration the organic solvent increases within the mobile phase, the mobile phase decreases in polarity and those analytes whose polarity is similar to the mobile phase will elute off the column and into the detector. Therefore on a reverse phase chromatogram, the order of analyte elution is polar (shortest retention time), mixed polarity, and finally non polar (longest retention time).

The ability to properly separate complex mixtures is related to the resolution power of the column used. One way to increase the resolution, thereby increasing separation, is by decreasing the particle size of the column[40]. However, as particle size decreases, the resistance to flow increases. In conventional liquid chromatography, this process occurs at ambient pressure, requiring the force of gravity to separate the complex mixture. Therefore, the increase in resolution will decrease the flow rate. In an effort to increase resolution without sacrificing the flow rate, techniques such as high performance liquid chromatography (HPLC) were developed[40,41]. In HPLC, the instrument is capable of pumping the mobile phase through the stationary phase at higher pressures (500 bar), allowing the flow rate to remain the same. Further, as HPLC instruments are capable of running at much higher pressures, increasing the flow rate also allows an increase in

analysis speed. However, the increase in flow rate reduces the efficiency of column retention. Thus, in an effort to further decrease the particle size, increase resolution, and increase speed, ultra performance liquid chromatography (UPLC) was developed[42]. In UPLC, the instrument is capable of performing at even higher pressures (1200 bar) with a smaller particle size (1.7µm versus the 3.5µm or 4.6µm seen in HPLC), allowing the flow rate to increase without sacrificing the column efficiency. Another feature of UPLC instrumentation is the decrease in column size. A column's resolving power is related to its length and particle size[40], therefore by decreasing the length of the column, the resolving power can be comparable to HPLC. Overall, this decrease in particle size and column length allows the loading capacity to decrease, reducing the amount of sample needed for an analysis[42].

Following elution from the column, the analytes travel into the detector. Two of the most common detectors used in a LC based metabolomics analysis is the UV-Vis detector and the Mass Spectrometer.

2.2.1 UV-Vis Diode Array

Following elution from the chromatography column, the analytes flow into the UV-Vis diode array detector (DAD) which detects the absorption spectrum in the ultraviolet (100-400nm) and visible (400-700nm) regions of the electromagnetic spectrum. The DAD has two separate lamps emitting light in either the ultraviolet or visible regions. First, a tungsten (W) lamp emits light from the visible and near infrared range (370-900nm). The emitted light enters a deuterium discharge lamp (D_2). Here, the D_2 lamp couples UV light (180-370nm) to the visible light and the resulting beam passes

through the flow cell. In the flow cell, the beam of light can be absorbed by an appropriate chromophore and the resulting absorption profile is recorded by a detector[43]. As each molecule has a characteristic absorption spectrum, analytes can be identified, particularly when the absorption spectrum is compared to a reference library.

2.2.2 Mass Spectrometer (MS)

LC-MS based analyses permit the detection of a wide array of molecules found within complex biological samples. The instrumentation itself is an intricate coupling between HPLC (or UPLC) and a Mass Spectrometer[23]. However, this coupling presents some challenges. First, the elution solvent needs to be eliminated to ensure a proper vacuum in the MS. Second, all molecules need to be converted into gas-phase ions. This includes any buffers found within the sample/mobile phase, therefore analysts need to be aware of the volatility of the buffers or they risk introducing salts into the mass analyzer during the ionization process[22]. As in GC-MS, the mass spectrometer can be broken down into three main components:

1. *Ion Source*

Ionization in LC-MS occurs within a condensed or vapor phase under atmospheric pressure, a condition referred to as Atmospheric Pressure Ionization (API). There are 3 main API techniques used as an ion source within LC-MS; Electrospray Ionization (ESI), Atmospheric Pressure Chemical Ionization (APCI), and Atmospheric Pressure Photoionization (APPI). As with CI in GC-MS, these three techniques are

examples of soft ionization, as molecular fragmentation is not extensive relative to EI[44–46].

ESI works by ionizing the molecules directly from the liquid matrix that elutes from the column[37–39,44,47,48]. Depending on the pH of the solvents used, the analytes can favor the ionic state in the solution, travels through the column, and ends at the capillary tip. At this point, the charged media containing the analytes is sprayed into a chamber containing a strong electrostatic field. The droplet that is formed becomes desolvated using a heated drying gas, thereby leaving ionized gas phase analytes. The analytes are then ejected into the sampling cone and continue to the mass analyzer. Because of the desolvation process, ESI permits an analysis of a wide array of biomolecules including larger molecular weight proteins and thermally unstable analytes[44,47,48]. In addition, molecules such as proteins which contain multiple ionizable sites produce multiply charged ions. While the production of multiply charged species complicates the spectra, it allows researchers to examine the intact molecule and is advantageous when using a mass analyzer that has a narrow measurable mass range[38,48]. The one major drawback to ESI is ion suppression due to a high concentration of non-volatile analytes. Ionization will first occur at the surface of the liquid droplets. If any non-volatile analytes are present on the surface, they will become ionized, thereby inhibiting the formation of ionized gas phase analytes and leading to ion suppression[39,46,48,49].

APCI is another technique used to ionize non-volatile analytes of interest. In this ionization technique, the analytes elute from the column and into a nebulizer where they are converted into droplets[37,38]. Unlike in ESI where the droplets themselves are ionized,

the liquid droplets produced by APCI travel into a heated chamber where they vaporize resulting in gaseous analyte and mobile phase. The heated mixture then travels along a corona discharge electrode where the mobile phase becomes ionized (ionizing gas). The analytes and ionizing gas react resulting in the analytes becoming ionized, akin to CI in GC-MS (however, the electrons emitted from the corona discharge electrode do not come from a heated filament, like they do with CI). This technique in particular works well with small and thermally stable molecules which may not be ionized via ESI[38]. However, due to the high temperatures, not all analytes will be detected, and unlike ESI, multiple charging is not seen.

The third technique, APPI, ionizes the analytes using a photon of light to excite and ionize gas phase molecules[38]. Similar to APCI, the liquid analytes elute from the column and are desolvated and vaporized by a heated nebulizer. Once vaporized, the gaseous analytes then interact with photons emitted from a UV lamp, resulting in ionization. The UV lamp emits photons at a lower energy than the ionization potential of the solvent, thereby preventing the solvent from becoming ionized, consequently reducing the background noise[38]. Additionally, this technique is able to ionize compounds such as non-polar molecules that are not readily ionized by ESI or APCI[38]. However, this technique is the most sensitive to suboptimal experimental conditions. Hence, proper solvent selection is imperative, as acidic solvents lead to analyte ion suppression. As in APCI, this technique will result in only singly-charged ions[44,45,49,46,50,39,36].

2. Mass Analyzer

After becoming ionized, the ions traverse through a mass analyzer where they are separated/filtered according to their masses. Our LC-MS instrument (LC-QToF, Figure 7) contains both a quadrupole and a ToF mass analyzer.

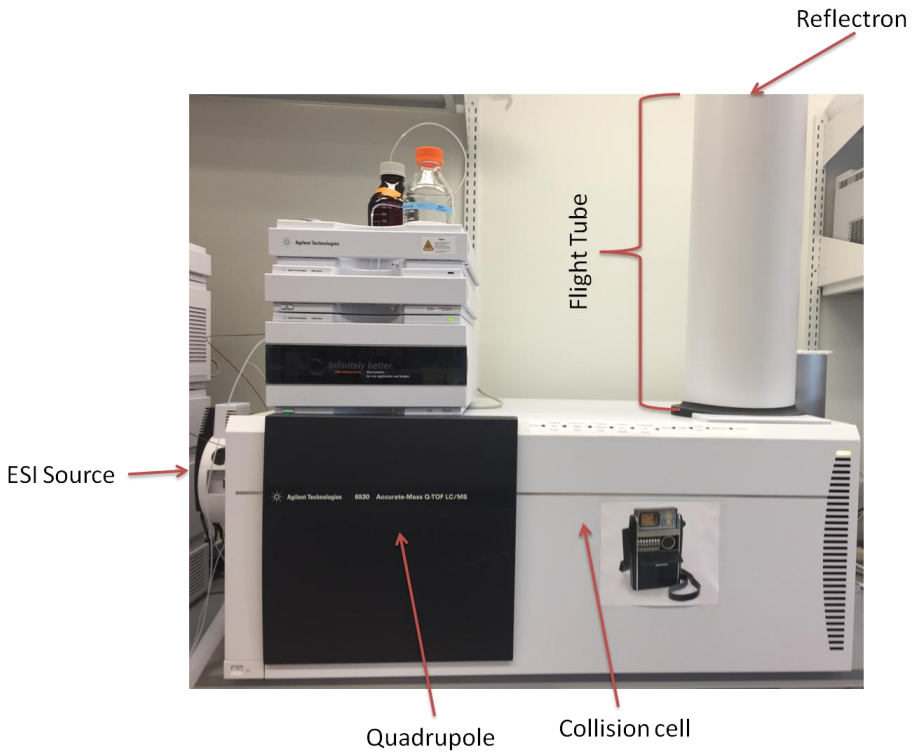


Figure 7. The Couch Lab's Agilent 6530 QToF.
The QToF is an instrument capable of MS and MS/MS analyses. The QToF has an ESI ionization source, a quadrupole as the first mass analyzer, a hexapole collision cell, and an orthogonal acceleration ToF. Indicated on the ToF is the flight tube and reflectron. We affectionately refer to the QToF as a tricorder, as it is capable of identifying metabolite composition in samples of interest. A photo of a tricorder from Star Trek is seen affixed to the QToF.

As stated previously (see the GC-MS section, above), quadrupole mass analyzers consist of four circular rods which are held parallel to each other. The ions that elute from the ion source are separated and filtered according to their m/z ratio by tracking the stability of the trajectory of the ions through an oscillating electromagnetic field[37–39].

Time of flight (ToF) mass analyzers separate ions via their velocity through a space devoid of an electromagnetic field (i.e. field-free region) known as the flight tube. Following ionization, the ions produced travel into an accelerated region where they acquire the same kinetic energy (E_k). Therefore, once the accelerated ions travel into the field-free region, their m/z ratio can then be related to the time (t) it takes to drift towards the detector using the equation $m/z = Kt^2$, where K is a fixed constant related to the energy applied, the length of the accelerated region, and the length of the flight tube[37]. Therefore, the mass of the analyte is defined by the charge (z) multiplied by the squared travel time (t) and the fixed constant K. Since all of the ions are produced in a short amount of time and are temporally separated, all of the formed ions will reach the detector at different times. Early ToF mass analyzers, known as linear ToF perform this in one direction. However, due to a kinetic energy spread among ions with the same m/z ratio, the mass resolution was poor. This has since been corrected by using either delayed pulse extraction or a reflectron[38].

In delayed pulse extraction, ions are initially allowed to travel and separate through the field-free region. After a specified amount of time, a voltage (extraction pulse) is applied which provides additional energy to the ions still lingering by the

source. This process then allows the less energetic ions to reach the detector at the same time as those more initially energetic ions of the same m/z ratio[38].

Another way to improve mass resolution is by using an electrostatic reflector, referred to as a reflectron. The reflectron contains an opposing electric field that is of greater magnitude than the acceleration region. Ions with higher kinetic energy will penetrate the deepest, thereby taking more time to reflect back to the detector and correcting the kinetic energy dispersion between ions of the same m/z ratio[38].

ToF mass analyzers were developed initially for use with pulsed ionization techniques (lasers, plasma) because they provide concise ionization regions at defined time intervals. To incorporate the use of continuous ionization techniques such as electrospray ionization, orthogonal acceleration (orthogonal injection) was developed[37–39]. In this technique, the sample is continuously ionized in the source and enters into the orthogonal accelerator as a parallel beam using ion optics as a guide. Inside the accelerator are a plate and two grids. In the first stage, the field-free space between the plate and the first grid become filled with the ions. Once filled, the plate is supplied a specified injection pulse voltage and the resulting change in the electric field thrusts the ions in an orthogonal direction where they are accelerated past the second grid and into the field-free flight tube where mass separation occurs. Once all the ions enter the flight tube, the plate's voltage is restored and new ions from the source begin to accumulate the space between the plate and the first grid again. The injection pulse will not be applied a second time until the ion with the highest m/z ratio currently in the flight tube reaches the detector[38].

Mass analyzers can be combined to form a technique known as Tandem Mass Spectrometry (MS/MS)[37–39]. In MS/MS, two or more mass analyzers are coupled to obtain better sensitivity and selectivity by fragmenting the ions isolated during the first MS experiment. The first mass analyzer selects and isolates a precursor ion, the ion then undergoes spontaneous or activated fragmentation, and the subsequent mass analyzer(s) separates and analyzes the resulting product ions. This technique can be achieved in two distinct ways, in time or in space. Instruments conducting this technique in time contain an ion storage area which will conduct the steps throughout a sequence of events. Those conducted in space contain separate instruments. With the in space instruments, fragmentation occurs within a collision cell which lies between the mass analyzers. In the collision cell, the ions can undergo fragmentation by colliding into a high pressure gas or other fragments in a process known as Collision Induced Dissociation (CID). The amount of energy used for the collision can vary resulting in different degrees of fragmentation.

The instrument I use for the LC-MS analyses is a Quadrupole Time of Flight (QToF, Figure 8) and is one example of an MS/MS instrument. In the QToF, the quadrupole acts as the first mass analyzer (Q1), a hexapole (q2) acts as the collision cell, and the ToF is an orthogonal acceleration-ToF and acts as the second mass analyzer. When acting in MS mode, the ToF is the sole mass analyzer as the two quadrupoles act purely as ion guides. The collision cell, q2 may contain a collision gas to improve resolution and the sensitivity of the ToF. In MS/MS mode, Q1 acts as the precursor ion filter. The selected ions are then accelerated into q2 where they undergo fragmentation by nitrogen at a specified collision induced dissociation. The resulting daughter ions and

parent ions continue to collide and fragment resulting in a decrease in their kinetic energy, and increasing the resolution prior to analysis via the orthogonal acceleratation-ToF[37,38].

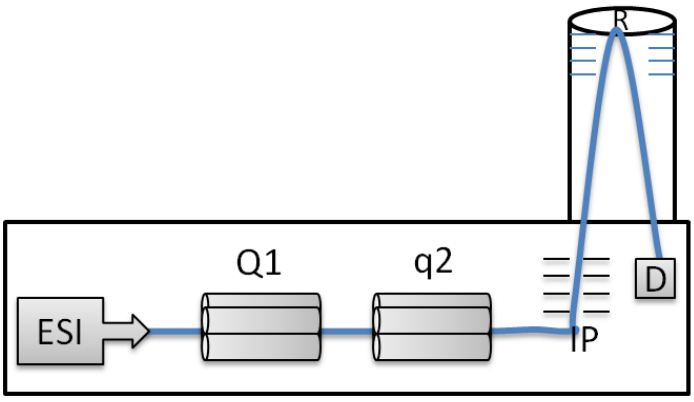


Figure 8. Schematic depicting the various components of the QToF.
Shown here is a depiction of the QToF currently housed in the Couch lab. The blue line depicts the ion's course as it is first generated by the ion source (ESI), guided into the first mass analyzer via an octapole (not shown), travels through the quadrupole mass analyzer (Q1) and into the hexapole collision cell (q2). The ion is then transferred into the ion pulser (IP) where through orthogonal acceleration it travels into the flight tube, deflects off the reflectron (R), and strikes the detector (D).

3. Detector

After the ions travel through the mass analyzer, they proceed to the detector. While the quadrupole scans the ions from one m/z value at a time, the ToF detects all of the ions of any given mass all at once[37,38]. Therefore, the LC-QToF contains an array detector which collects all of the ions of various masses isolated in the instrument. To detect the large collection of ions and ensure a precise arrival time, a microchannel plate containing an electron multiplier is used (Figure 9). The plate contains microscopic

parallel cylindrical channels and each channel acts as an electron multiplier[38]. When an ion with enough energy collides into the microchannel plate, secondary electrons are produced and travel through the channel becoming amplified. Because the signal produced is at approximately -6000V, the amplified electrons will then collide with a scintillator, emitting photons. The resulting photons are directed in a photomultiplier tube where they are further amplified and the current is measured. By incorporating the photomultiplier tube, the potential increases from -6000V to 0V (ground potential) allowing it to be measured by the anode resistor and subsequently digitized by the computer[38,51].

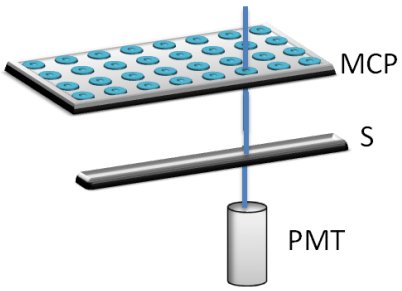


Figure 9. Schematic depicting the components of the microchannel plate and photomultiplier detector.
Shown here is a depiction of our QToF's detector. The blue line depicts the ion as it travels from the flight tube (not shown) and into a well in the microchannel plate (MCP). The resulting secondary electrons emerge and travel into the scintillator (S) where they are converted into photons. The photons enter the photomultiplier tube (PMT) where they are converted back into secondary electrons and strike the anode resistor at the bottom.

3. Metabolomics Data Analysis

Following metabolite detection, the resulting molecular features are extracted and subsequently analyzed using multivariate statistical techniques. The approach to data analysis was a central theme in this dissertation and the techniques will be described in detail throughout.

The overarching goal of this research was to develop a state-of-the-art metabolomics approach to the investigation of metabolite alterations in juxtaposed biological cohorts (eg. healthy vs. disease; infected vs. uninfected). To establish a metabolomics pipeline of data acquisition and analysis, a series of research projects were pursued within the general contexts of rapid diagnostics, pathogen detection, and/or molecular profiling.

SPECIFIC AIMS AND PROJECT OVERVIEW

Project Overview

The goal of this research project is to develop and employ a metabolomics pipeline to visualize and compare the metabolomes of biological systems. To achieve this goal, I simultaneously performed 4 research investigations, each of which offering insight into metabolomics methods of data acquisition and analysis. Collectively, these projects defined the overall metabolomics pipeline that I developed. For clarity in this dissertation document however, I present my research as two specific and separate aims, the first describing the metabolomics pipeline and the second listing the series of projects that were used to develop the metabolomics pipeline (each project organized into a sub-aim of Specific Aim 2).

Specific Aims

Specific Aim 1: Develop a metabolomics pipeline of data acquisition and analysis. Develop a method that utilizes GC and/or LC based platforms to perform a global (untargeted) assessment of the metabolic state of biological samples. The pipeline developed here will serve as the foundation for future metabolomics analyses.

Specific Aim 2: Application of the metabolomics pipeline. Develop and refine the metabolomics pipeline by exploring alterations to the metabolome as a consequence of a change in health.

Specific Aim 2.1: Metabolomics as a biosensor; developing an electronic nose. Use metabolomics as an electronic nose to determine if global metabolic profiling of microbial volatile organic compounds can differentiate select biological warfare agents.

Specific Aim 2.2: Alterations to the human fecal metabolome due to alcohol consumption. Use metabolomics as an electronic nose to explore the odor profile of fecal material collected at home or via an endoscopy procedure to potentially identify biomarkers of chronic alcohol consumption.

Specific Aim 2.3: Tissue-Related effects of a high fat diet and probiotic supplementation in pigs. Use metabolomics to evaluate if a high fat diet and/or probiotic supplementation have an effect on the pig tissue metabolome.

Specific Aim 2.4: Metabolic changes in pigs due to whipworm (Trichuris suis) infection. Use metabolomics to characterize the metabolic changes associated with various stages of *T. suis* infection in pigs and demonstrate how changes in certain metabolites relate to mammalian inflammatory pathways.

SPECIFIC AIM 1. DEVELOP A METABOLOMICS PIPELINE OF DATA ACQUISITION AND ANALYSIS.

Objective: Develop a method that utilizes GC and/or LC based platforms to perform a global (untargeted) assessment of the metabolic state of biological samples. The pipeline developed here will serve as the foundation for future metabolomics analyses.

Introduction

As a branch of the "-omics" sciences, metabolomics is concerned with the complete collection and analysis of small molecule metabolites (generally <600 Da) in a biological sample[3,4]. By identifying and/or monitoring the presence and concentration of these metabolites, hypotheses can be made regarding the origin and meaning of the phenotypic changes that are observed. The conclusions derived from a metabolomics investigation often complements information resulting from genomics, transcriptomics, and/or proteomics investigations. Due to the nature of the –omics sciences, it is not uncommon for these analyses to generate data matrices containing thousands of molecular features (i.e. the presence and relative abundance of metabolites, transcripts, or proteins from metabolomics, transcriptomics, and proteomics investigations, respectively). In order to analyze and interpret such large data sets, the –omics sciences rely on multivariate statistical analyses (i.e. statistics involving two or more variable quantities). Our goal was to develop a metabolomics investigative pipeline based upon

our analytical instrumentation (GC and LC), utilizing relevant univariate and multivariate statistical data analysis techniques. From sample processing to data acquisition and analysis, I evaluated and implemented techniques/approaches to establish a metabolomics pipeline, permitting both global (untargeted) and targeted metabolomics analyses.

As first mentioned in the Specific Aims section above, the development of this metabolomics pipeline was an iterative process developed concomitantly with performing the select research projects described in Specific Aim 2. Hence, the pipeline was refined and developed over the course of this entire thesis work. While metabolite extraction details specific to the biological samples are presented within Specific Aim 2, I present here in Specific Aim 1 a description of the core data analysis pipeline pertaining to GC and LC based analyses, a pipeline that is used regardless of the biological sample under investigation. Within this section I present a discussion of the theory and/or rationale behind each informatics step implemented in our current approach to performing a metabolomics investigation.

Metabolomics Data Analysis

Our current approach to metabolomics data analysis is summarized in Figure 10, with detailed Standard Operating Procedures (SOPs) presented in Appendix 1 and Appendix 2. This approach was developed specifically for use in a global metabolomics investigation, with LC and/or GC derived data, although it may also be readily used as-is for targeted metabolomics investigations. The data analysis pipeline employs univariate and multivariate statistical analyses, providing a comprehensive examination of the metabolomes derived from biological samples. As described in detail below, the data

analysis pipeline occurs in four stages: 1. Molecular Feature Identification, 2. Data Preparation, 3. Multivariate Statistical Analysis and Data Visualization, and 4. Determination and Identification of Metabolites of Interest. The first stage includes identification and extraction of the molecular features from the chromatogram/spectra. The second stage transforms the data set in preparation for multivariate analysis. The third stage employs Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), and Correlation Network Analysis to elucidate trends within the data. Finally, in the fourth stage, to identify the key metabolites of interest (i.e. those that differentiate cohorts from one another), I use a combination of Fold Change, P-value, Box and Whisker Plots, and Receiver Operating Characteristics curves (ROC).

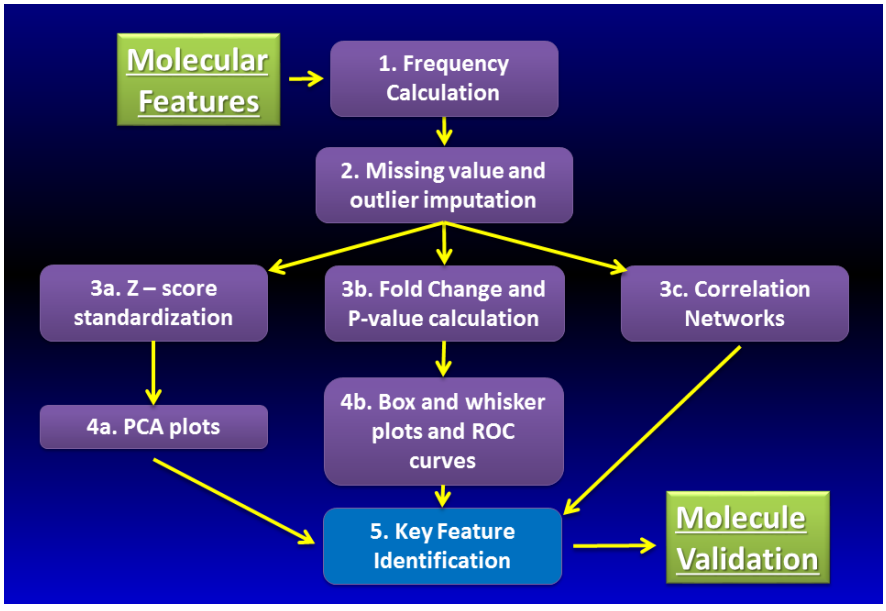


Figure 10. Metabolomics Pipeline.
The metabolomics pipeline allows us to examine the metabolome of biological systems. By examining key contributors to variations and their connectivity to other metabolites present in the sample, I can identify biomarkers and gain insight into the etiology of disease. See text for further discussion.

Stage 1 of the Data Analysis Pipeline: Molecular Feature Identification

The first stage of the metabolomics pipeline is to identify the molecular features extracted from the biological sample. Note that at this stage I refer to the extracted metabolites as molecular features, as their identity is either unknown or is only speculative. For non-MS derived data (e.g. chromatograms generated using the FID or UV-Vis detector), the molecular feature is simply referred to by its chromatographic retention time. For MS derived data, while retention time can also suffice for molecular feature identification, molecular features are typically identified by deconvoluting the chromatogram and matching the m/z value of a molecular feature to a molecular database. More specifically, feature identification for GC-MS acquired data is performed by using software known as the Automatic Mass Spectral Deconvolution and Identification System (AMDIS), whereas features in LC-MS acquired data is identified by using Agilent Technologies' MassHunter Qualitative Analysis, each of which is described in more detail below.

GC-MS Molecular Feature Identification

AMDIS is a computer program that evaluates GC-MS chromatograms and identifies molecular features (metabolites) within the chromatographic peaks by comparison of extracted mass spectra with a mass spectral database (the National Institute of Standards and Technology (NIST) Mass Spectral Library). This process is achieved in four successive steps: 1. Noise Analysis, 2. Component Perception, 3. Spectrum Deconvolution, and 4. Compound Identification[52].

Step 1. Noise Analysis

In this stage, AMDIS analyzes the chromatographic background and calculates a noise threshold for each data file (Figure 11). This is performed by first calculating the noise factor (N_f) which is the average random signal fluctuation divided by the square root of the signal intensity[52]. AMDIS generates a unique N_f for every data file by first examining regions of constant signal intensities within the ion chromatogram. Next, the average of all the signal intensities in this region is calculated. This information is then plotted and if less than one-half of the number of scans do not cross the mean, then the segment is rejected. For every accepted segment, the median value for the deviation is calculated and then divided the square root of the mean intensity to obtain the N_f value for that segment. Following the processing of the data file, the median of all the N_f values is calculated and presented as the representative N_f for that sample data file. The median is used in place of the average to prevent a high N_f resulting from normal chromatography.

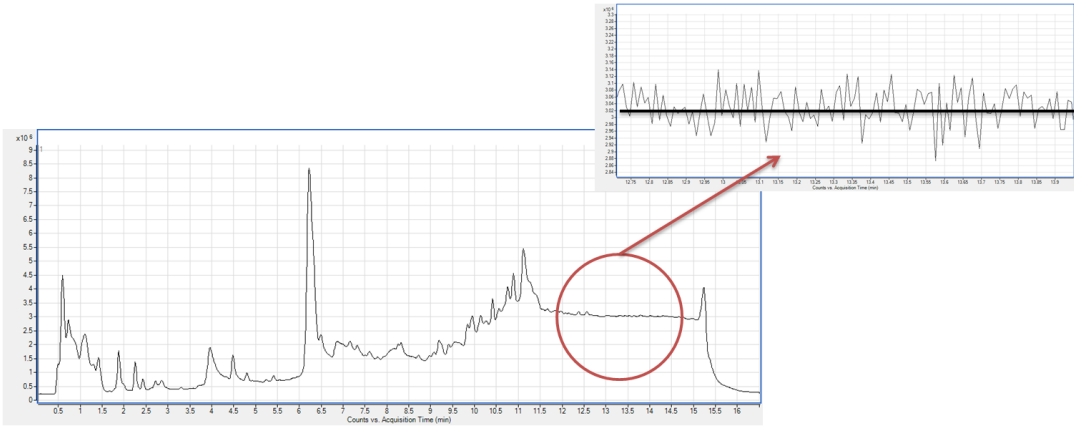


Figure 11. Determination of noise in a chromatogram.
The portion circled in red is a baseline region which will be used for the noise calculation. The average intensities of the scans found in this region is calculated and the information is plotted (see insert). If less than one half of the scans do not cross the average line (the black line) then this segment will be rejected. If the segment is accepted, the N_f value is calculated for that segment. See text for further discussion.

The second portion of the noise analysis is to identify the threshold transitions[52]. Typically when the instrument is tuned, a baseline abundance value (A_T) is established and retained. During a sample run, any signal intensity that is above that value is collected while those that fall below that threshold are not collected and given a value of 0. While this transition from a zero to a nonzero value is common in background signals, they can be misinterpreted as a chromatographic peak. Therefore, AMDIS establishes a threshold for each data file to eliminate such noise. This is calculated by first assuming the smallest nonzero abundance is equal to A_T. Next, the ion chromatogram is divided into equal length segments and for every m/z in the segments, the number of times a transition from zero to nonzero occurs is counted. For those m/z ratios which were given a value of 0, a new value is provided by multiplying A_T by the square root of the number of transitions.

The third and final portion of noise analysis is to establish m/z peak uniqueness[52]. This is performed during the identification of the threshold transition. Here, for every m/z value the number of nonzero abundance values is calculated. Then, the signal to noise threshold is multiplied by the square root of the number of nonzero m/z values. This process ensures that all m/z values are properly elucidated regardless of their signal levels.

Step 2. Component Perception

In the second step of the molecular feature identification process, AMDIS analyzes the chromatogram and tracks the increase and decrease in abundance for any given ion and develops a model of the peak. If any other ions are found within that same retention time and exhibit the same profile, then those ions are assumed to be a part of that component (molecular feature)[52]. As discussed previously, quadrupoles acquire data by scanning a m/z range sequentially as the analytes elute from the column[37,38]. Therefore, different mass spectral peaks are acquired at various times during the elution profile.

The components (molecular features) are identified when ions for that feature increase at the same time[52]. This is established by first constructing a scan window using the minimum intensity value on each side of the chromatographic peak and then drawing a baseline between the two signal intensity points. Next, a second baseline (least-squares) is calculated and drawn through the lower half of the points relative to the baseline. Finally, the signal height is calculated between the least squares line and the maximum signal intensity value. In order to be considered a peak, the height must exceed four times

the noise factor (N_f) multiplied by the square root of the maximum signal intensity. If the component passes, a parabola is fitted through this peak and will proceed onto step 3.

Step 3. Spectrum Deconvolution

The co-elution of metabolites during chromatography complicates the mass spectra of a molecular feature through the extraneous contributions of ions from the co-eluted metabolite[52]. The presence of the co-eluted ions may contribute peaks which may not be characteristic of the molecular feature (Figure 12).

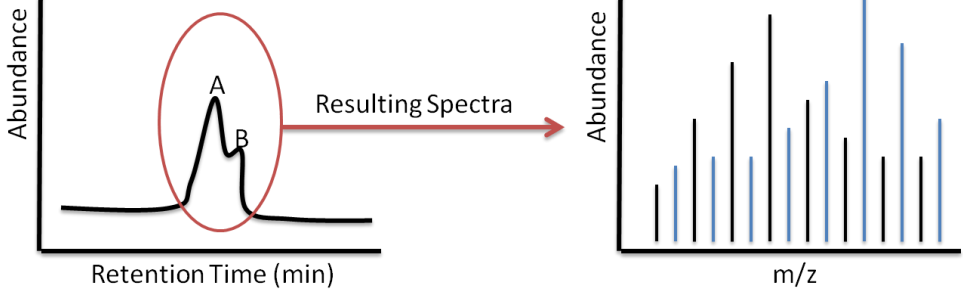


Figure 12. Mass Spectra of Co-eluting Metabolites.
As indicated on the chromatogram, metabolites A and B are co-eluting metabolites. The resulting mass spectra is then a combination of the mass spectra of these two metabolites. Shown in black is the mass spectra of metabolite A while in blue the mass spectra of metabolite B. AMDIS will resolve the spectra of these two metabolites.

This then provides a challenge when comparing the derived spectra to a database of purified molecular standards, devoid of the extraneous ions. Thus, in this deconvolution step, AMDIS corrects and resolves the mass spectra for a given molecular

feature by extracting the ions unique to that feature. This is achieved by using the model shape developed in step 2 and least squares method. Each ion chromatogram is fit into a unique model profile resulting in a linear baseline. The signals from nearby features are then subtracted from the first molecular feature thereby resolving the spectra[52,53].

Step 4. Compound Identification

The final step in AMDIS is the identification of the molecular feature via spectral comparison with the NIST Mass Spectral library. The resolved spectra are compared to the library spectra (containing 242,466 metabolites) and a score is given to those known standard spectra which closely resemble the unknown spectra (Figure 13). The higher the score value, the more closely related the spectra[52].

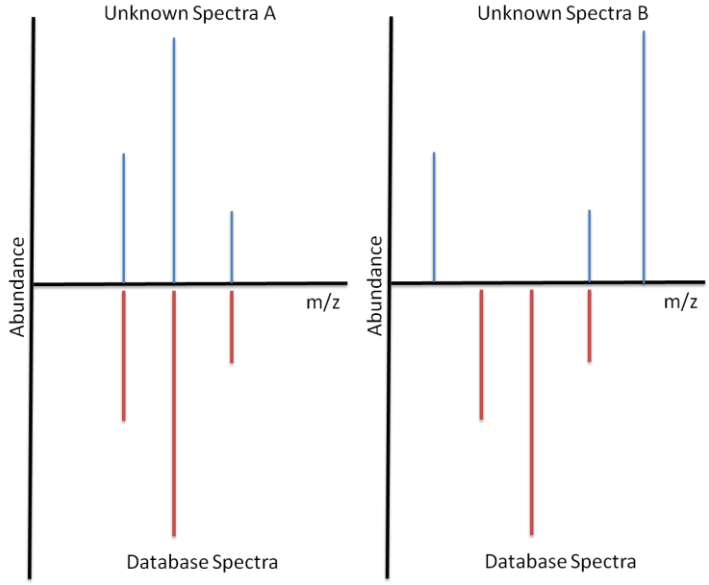


Figure 13. Molecular Spectral Library Matching.
The spectra for metabolite A matches the spectra found in the database while the spectra for metabolite B does not match the one found in the database.

LC-MS Molecular Feature Identification

Molecular feature identification with LC-MS based analysis utilizes the molecular feature extraction algorithm located in Agilent Technologies' MassHunter Qualitative Analysis software. This proprietary software resolves and locates the mass spectra of a molecular feature by elucidating the ions that are covariant (proportionally increase and decrease in abundance), similar to the method employed by AMDIS[54,55]. The ions are then grouped by their charge state, isotopic distribution, and/or the presence of adducts or dimers. Finally, the algorithm assigns these ions to a neutral molecule referred to as a molecular feature.

Stage 2 of the Data Analysis Pipeline: Data Matrix Preparation

Prior to analysis with statistical tools, the molecular features derived from the biological samples are compiled in a spreadsheet and are examined for missing data between samples, outliers are identified, and the measured abundance values are all scaled for comparative examination via multivariate analyses. It is notable that this initial step has a profound effect on the results obtained downstream and the importance of how the data is prepared should not be overlooked or understated (Figure 14).

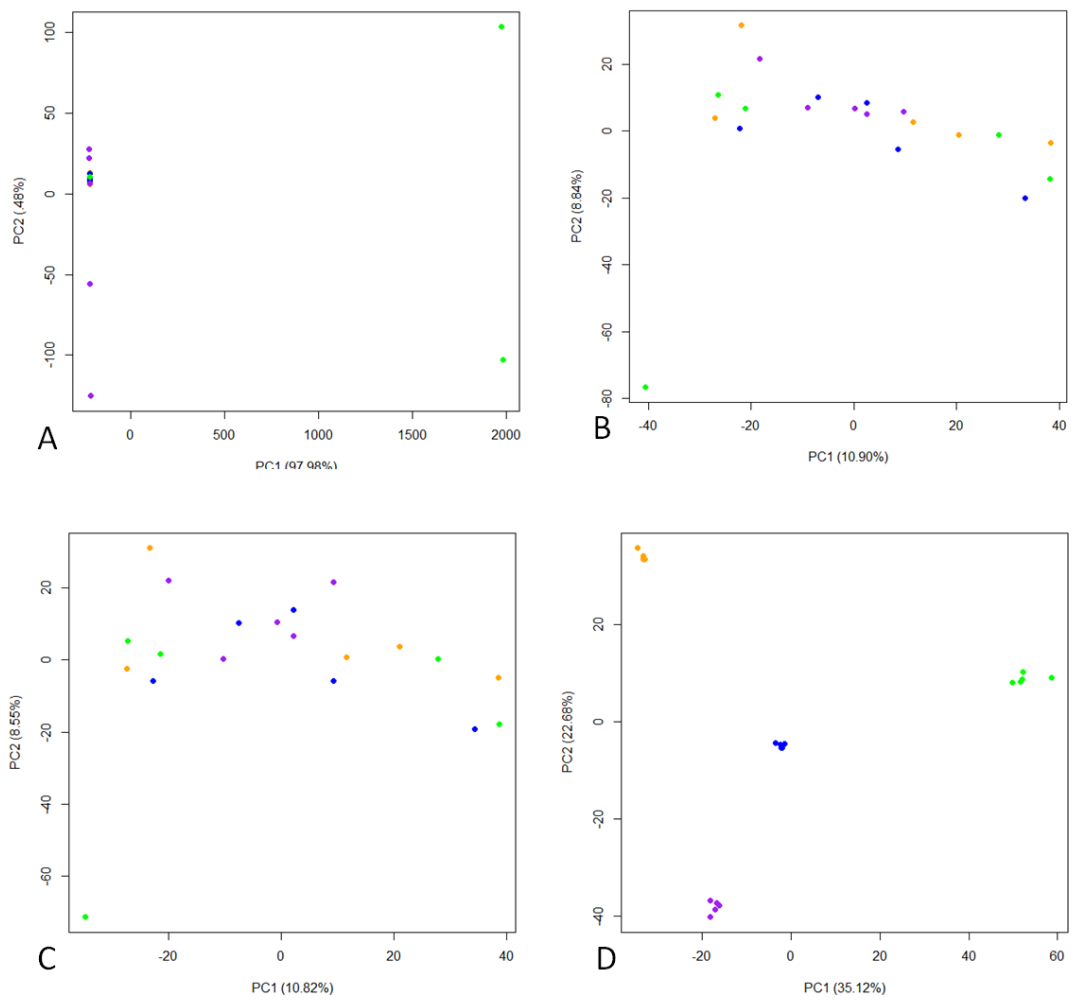


Figure 14. Effects of Data Preparation Prior To Principal Component Analysis.
In the PCA plots shown, each data point represents the metabolome from a biological sample, color coded by their associated cohort (5 samples per each of the 4 cohorts). The first plot (A) is the result of an analysis performed with metabolites identified in at least 2 of the 20 samples analyzed, without any additional data preparation. Contrast this with the second plot (B) resulting from an analysis of metabolites appearing in at least 4 of the 20 samples analyzed. The third plot (C) depicts the PCA from B but performed following outlier removal from the data set. Finally, the fourth plot (D) illustrates the PCA from B performed following median replacement, outlier removal, and z-score normalization. The approach to data preparation clearly has an effect on the outcome of the analysis, with the latter approach clearly differentiating each of the samples into their associated cohorts (as illustrated by the color coded clustering of the points within the plot). The principles of PCA are described in detail within the text.

As a first step in data matrix preparation, I eliminate what I refer to as “one-off” metabolites (that is, molecular features that appear sporadically in only a few number of samples in the analysis). To accomplish this, the frequency of appearance of each identified molecular feature is calculated across each cohort (cohort frequency), in addition to across the entire sample set (total frequency). A histogram is generated by plotting the number of metabolites as a function of total frequency of appearance (Figure 15). By examining the shape of the histogram and slope in the decline in total metabolites, I am able to determine a minimum total frequency value. For example, in Figure 15, there is significant decline in the number of metabolites present between 10% and 11% of samples. Then from 11% to 31% of samples, the number of metabolites present continues to decline but at a slower rate. However, from 31% to 100%, the number of metabolites present in each 10% interval remains similar. Given this, 31% would be set as the value for our minimum total frequency value.

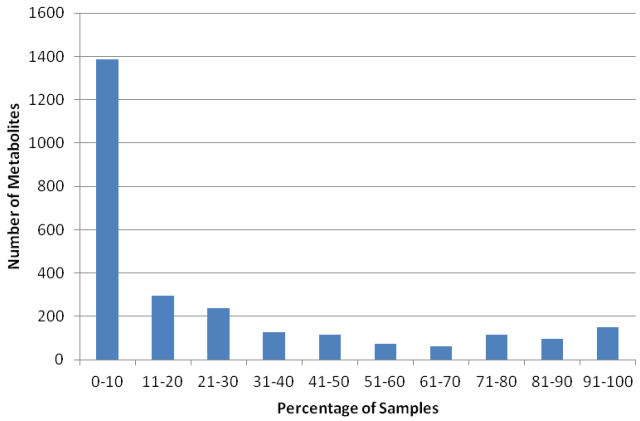


Figure 15. Histogram of Metabolite Appearance Among Numerous Biological Samples.
To visualize the distribution of metabolites across all of the biological samples, I generate histograms. The histogram shown here highlights numerous one-offs in the data matrix (nearly 1400 metabolites appear in 10% or fewer biological samples). Once identified, these one-offs are typically removed from the metabolite matrix, so as to not impact subsequent statistical analyses.

To further filter the data, I impose a second frequency cutoff that considers the frequency of a molecular feature within each of the cohorts (rather than across all samples in the analysis, as described for the first frequency cutoff step).

$$S = A^n + B^n + C^n + \cdots$$

Equation 1. Scoring criteria to determine low frequency metabolites.
S: score of metabolite
A: Frequency of metabolite in Cohort A
B: Frequency of metabolite in Cohort B
C: Frequency of metabolite in Cohort C
n: total number of cohorts

As indicated in Equation 1, I calculate a score for each metabolite then simply use a score value greater than or equal to the one acquired from the minimum total frequency value. For example, if the data set has 5 samples in 4 distinct cohorts (i.e. 20 samples in total) and the minimum total frequency value is set equal to 20%, as determined by the histogram (i.e. the metabolite is required to appear in at least 4 of the 20 samples), but I am specifically interested in a cohort frequency of 80% (i.e. 4 of 5 samples in any one of the four cohorts must contain the molecular feature), then the minimum score value would be $4^4+0^4+0^4+0^4$ which equates to 4^4 or 256. Therefore, I would first restrict our data set to metabolites with a total frequency of 20% or greater, then would retain only those metabolites with a score value greater than or equal to 256.

Once the one-off metabolites have been removed from the data set, missing values within the matrix of data are then addressed. Missing values arise for various reasons such as the stringency of the peak identification settings (stage 1 molecular feature identification), the complexity of the chromatographic peak (masking of a feature by co-eluting metabolites), metabolite abundance below the limit of detection, or a metabolite simply being absent from a sample. Due to the nature of the downstream statistical analysis techniques and their associated calculations, there cannot be any empty (valueless) cells within the metabolite matrix (data set). Therefore, either the molecular feature with the missing value must be entirely removed from the data set or a value must be inserted (imputed) in place of the missing value. Since the former is extreme and can needlessly eliminate important metabolites from an analysis, missing values are typically imputed.

There are a variety of different ways to impute a missing value, some approaches include populating the empty cells with a value of 1 or 0, using the average value of that metabolite detected across a cohort, or using the median value for that metabolite[56]. Since metabolite abundance is a direct reflection of the ion abundance value reported by the MS, and ion abundance is typically on the order of 10^4 counts or greater, imputing a value of 0 or 1 for missing data can have a profound and misleading effect on the distribution of metabolite abundance in that cohort. Therefore, imputation with the statistical mean or median metabolite value is preferred. The median metabolite value is used since it is least likely to have a significant effect on the Gaussian-like distribution of metabolite abundance across a cohort.

Following missing data imputation, metabolite abundance outliers are then palliated. Outliers may result from unintended variation in the extraction process among samples, from sample degradation during storage or processing, from chromatographic variations, or may simply reflect the metabolic state of that particular sample. Outliers are not uncommon in biological data sets and several mathematical methods of outlier identification have been developed[57–60]. Our approach is to determine the relative error (outlier value) in the distribution of our data set (Equation 2). Any instance where the outlier value is 1.5 or greater, the outlier is identified and replaced with the median value for that metabolite.

$$O = \frac{(mean - median)}{median}$$

Equation 2. Determination of an outlier present in a biological cohort.
O: outlier value
mean: average value of metabolite in that cohort
median: median value of metabolite in that cohort

The final step in data preparation involves the scaling of the relative metabolite abundance values across all samples to permit their comparison to one another. Data scaling specifically addresses differences in the ionizability of molecules in FID and MS and differences in absorbance at various wavelengths in UV-Vis. Some metabolites may be in low abundance yet ionize or absorb readily and produce a large peak intensity, whereas some metabolites may be in high abundance yet ionize or absorb poorly thereby resulting in small peak intensities. In fact, peak intensities in MS, FID, or UV-Vis can vary across several orders of magnitude among a data set of metabolites. A comparison of raw peak intensity values among these metabolites would then lead to erroneous conclusions.

Molecular feature standardization such as z-score standardization (Equation 3) converts all of the metabolite abundance values to a common scale, results in the production of a zero-mean, and a produces a standard deviation equal to one[61]. When applied, the z-score standardization approach allows the data to then be analyzed on the basis of correlation and allows each metabolite to be equally important, regardless of its ionizability or absorbance at a specific wavelength[62].

$$Z\ score = \frac{(x - mean)}{standard\ deviation}$$

Equation 3. Z-score standardization approach for molecular feature standardization.
x: value of metabolite in the sample
mean: average value of metabolite in all samples/cohorts
standard deviation: standard deviation value of metabolite in all samples/cohorts

Stage 3 of the Data Analysis Pipeline: Multivariate Statistical Analysis and Data Visualization

Multivariate statistics is often used in a metabolomic analysis and is a necessity during a non-targeted investigation[14]. While univariate statistical analyses only examine one variable at a time, multivariate statistics examines the large collection of variables (molecular features) found within the data. Here I describe two different techniques adopted for use in the multivariate statistical analysis portion our metabolomics analyses; dimension reduction and cluster analysis.

Dimension Reduction: Principal Component Analysis

As stated previously, the data set acquired during a metabolomics investigation is comprised of a large collection of variables (molecular features and their relative abundance) and associated observations (samples that were analyzed). In statistical analysis techniques, each variable is supplied its own dimension. Therefore for an analysis which identified over 1000 metabolites, there are over 1000 dimensions. Because of this, it is challenging to elucidate specific trends between the samples in the data set by plotting every possible metabolite combination. Dimension reduction addresses this challenge. Dimension reduction is the process of minimizing the number of dimensions while still conveying the same information, thereby allowing trends to be identified within the data set[63–65].

Principal component analysis (PCA, Figure 16) is a form of dimension reduction that uses linear transformations to elucidate the covariance structure of the metabolites identified in the biological sample, culminating in the description of the total variance identified in the data set. By using a covariance approach, not only are the differences (eg. abundance levels) between the metabolites examined, the influences those variations have on each of the metabolites are examined as well[63–65].

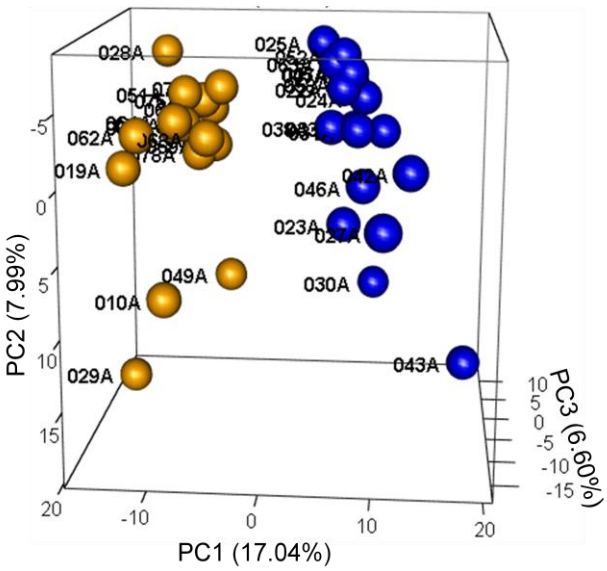


Figure 16. A 3D-PCA Plot demonstrating the variations between juxtaposed biological cohorts.
An illustrative three dimensional PCA plot depicting the differences in the metabolome composition of a healthy (blue spheres) and diseased (orange spheres) state.

The principal components are derived from the direction that illustrates the maximal variation within the data set. To illustrate the mathematical steps in principal component analysis, Figure 17 depicts a three dimensional diamond characterized by three variables: width, length, and height. This description of the three dimensional diamond can be simplified by redefining the shape on an alternate (two dimensional) coordinate system.

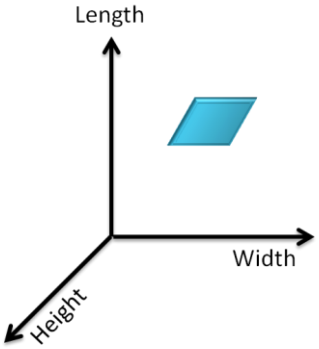


Figure 17. Diamond with a known width, length, and height depicted in a three dimensional graph. The diamond depicted here has three variables: width, length, and height. The values of each variable provides the coordinates needed to plot the diamond on the graph.

The alternate coordinate system is developed by first identifying the center of the diamond (i.e. the center of the X, Y, Z coordinate values) and then establishing the direction of a line that will provide the longest radius encompassing all of the data points (Figure 18A). This direction (new variable) is denoted as principal component 1 (PC1). Note that PC1 reflects the greatest variance in the data set. After defining PC1, the direction with the second longest radius is defined. This direction is defined by first drawing a line that is perpendicular to PC1 and then rotating the direction of the line to define its longest radius. This second direction (second variable) is labeled as principal component 2 (PC2), which reflects the second greatest variance within the data set. The three dimensional diamond can now be described using these two directions by re-plotting the diamond on a new coordinate system, defined by the two variables PC1 and PC2 (Figure 18B). Consequently, the dimensions (variables) can be reduced from three to

two while still retaining the same relative information about the diamond coordinate values (width, length, and height). Thus, PCA is method of simplifying the data set without losing any of the original information.

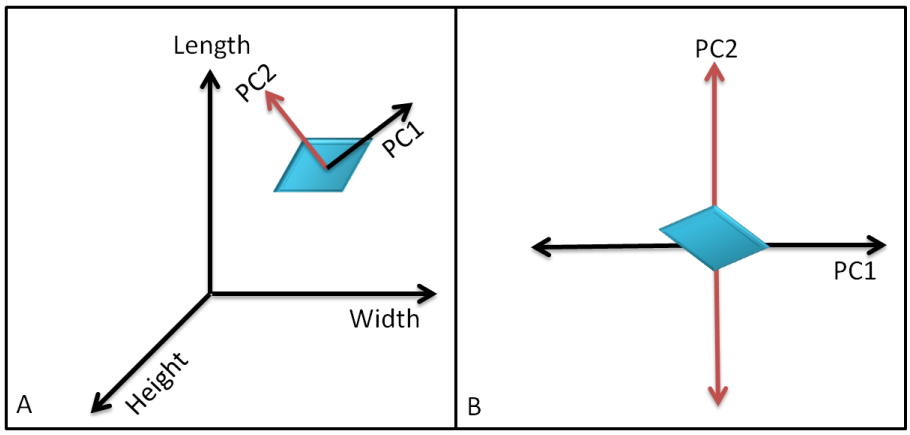


Figure 18. Determining an alternate coordinate system using principal component analysis.
In Figure A, the center of the diamond is first identified and a straight line is drawn (black line) through the center to determine the direction which provides the largest radius. This line is now referred to as PC1. Next, a line (red line) is drawn perpendicular to the direction of PC1. The direction of the red line which provides the second largest radius is labeled PC2. Using the largest and second largest radius and associate directions as new coordinates, the diamond can be re-plotted. Figure B depicts the diamond in a new coordinate system using the results obtained in Figure A.

A similar approach to dimension reduction can also be applied to metabolomics data. Suppose there are ‘n’ number of metabolites for ‘x’ number of samples. If the data provided in Figure 19 is plotted, it would appear as shown in Figure 19. Each point in the multidimensional plot is derived from an individual sample and its position in multidimensional space is a reflection of all of the metabolites pertaining to that sample.

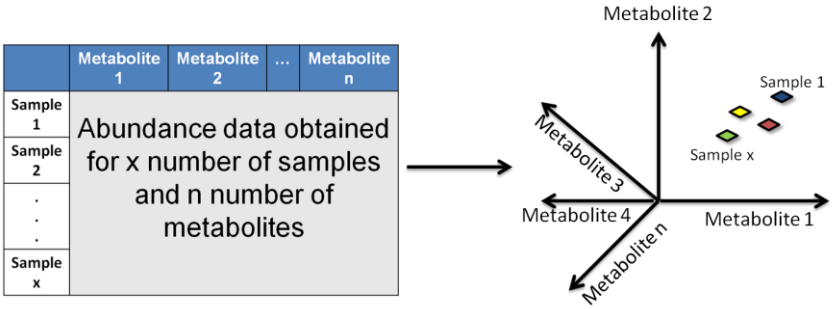


Figure 19. Acquired metabolomics data set visualized on a multidimensional graph.
In this figure, the table represents a metabolomics data set with associated abundance values acquired from GC or LC based analyses. In this data set, each metabolite is designated as a variable and provides its own dimension. Therefore in the graph on the right, the diamonds represent the samples and all of the metabolite information obtained for that sample. The positioning of the samples on the graph is a reflection of metabolite abundance.

To simplify the description of the samples, the data is transformed using the same approach as described with the diamond. First the center of all of the data is identified and the radius that provides the largest variance is used to define PC1 (Figure 20A). Next, the direction perpendicular to the first that describes the second largest variance is identified and labeled PC2. This process continues for m number of directions, so long as it results in a measurable (nonzero) variance (m < n (the number of metabolites)). After determining the direction of each component, the plot is redefined using the principal components as the new axes and the resulting variance measured in each component as the new coordinates (Figure 20B). As previously stated, each sample still contains the same amount of information (x number of samples with the abundance values for n

number of metabolites), but the dimensions are reduced from n number of metabolites to m number of components.

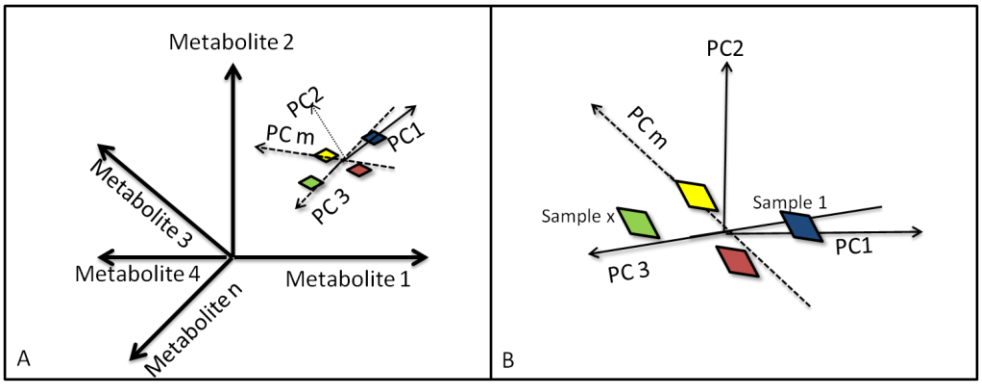


Figure 20. Using principal component analysis to reduce the dimensions in metabolomics data. In Figure A, using the data acquired in Figure 19, the center of all of the diamonds (samples) is determined first and the direction that provides the most variation (the radius of the diamond) between the samples. This direction is referred to as PC1. Next, a line is formed perpendicular to PC1, defining the direction that provides the second largest variation (PC2). This process continues for m number of directions (a value less than the n number of metabolites). The directions of the PC and their corresponding variance becomes the new coordinates for the samples. In Figure B, the samples are re-plotted using these new coordinates. By expressing the data in this manner, the number of dimensions is reduced while still retaining the original information.

To illustrate this analysis with actual metabolomics data, a PCA plot is shown in Figure 21. This analysis investigated the influence of diet on the pancreatic metabolome (fully described in Specific Aim 2.3). In this PCA plot, colored spheres represent the pancreatic samples, and their positioning in the plot is a reflection of their metabolite composition. The data has been transformed from a form similar to Figure 19 to the new PCA axes which describe the variance in the metabolic profile obtained from the samples. Therefore, spheres that are closely juxtaposed in the plot reflect similar

metabolome composition, whereas spheres which are segregated from one another exhibit significant variations in their acquired metabolomic profile. It is important to note, the relative positioning of the spheres is a reflection of the samples included in the analysis. Therefore, if another analysis was performed focused exclusively on the samples that appear closely juxtaposed in Figure 21, those samples will be resolved in a PCA plot, providing insight into the degree of variation amongst those particular samples.

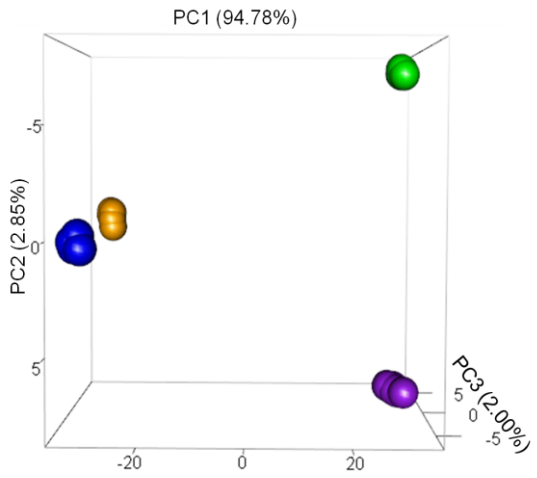


Figure 21. PCA plot depicting the variation in the metabolic profile obtained from the pancreas of pigs consuming different diets. Three dimensional PCA plots of metabolome composition derived from the pancreas differentiates the four pig dietary cohorts (basal, basal+probiotic, high fat, and high fat+probiotic) from each other, demonstrating that dietary consumption (including probiotic supplementation) can have a significant influence on the metabolite composition of the organ tissues. Each sphere in the plot reflects the metabolome of the pig tissue sample. Spheres are colored according to the diet (orange - basal, blue - basal+probiotic, green - high fat, purple - high fat+probiotic). Interestingly, grand variations between high fat and high+probiotic supplemented pigs are observed. Conversely, basal and basal+probiotic supplemented pigs display smaller variations, as they are closely juxtaposed in the plot.

Cluster Analysis: Hierarchical Clustering Analysis

Cluster analysis utilizes multidimensional data and categorizes the samples into groups known as clusters, based on their similarities[14,66,67]. These clusters, while capturing the data's overall structure, is used as an exploratory analysis to provide an assessment of the relatedness of the samples in regards to their metabolome composition. There are different varieties of hierarchical clustering analyses, but agglomerative hierarchical clustering is the only analysis which does not require the samples to classified into their respective cohorts. This is important for metabolomics because it reduces the amount of bias in your data (by not predefining cohorts, the samples are grouped solely on their degree of relatedness in metabolome composition). Agglomerative hierarchical clustering results are typically displayed as a dendrogram (Figure 22), which resembles a phylogenetic tree, depicting the similarities between each sample. In a metabolomics analysis, samples that are highly similar in metabolite composition will appear closely juxtaposed in the dendrogram[68].

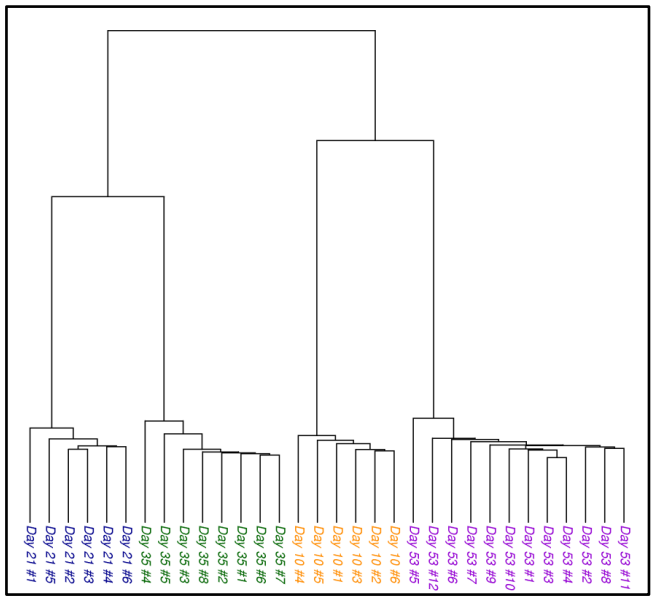


Figure 22. Dendrogram depicting the similarities in the metabolic profile obtained from samples at various stages of infection.
This dendrogram describes the similarities between different infection time points which are color coded at the bottom. While the plot depicts 4 distinct time points with all samples in those time point clustering together, the first major grouping between the time points denotes similarities in the metabolic profiles between the early (Day 10, orange) and late (Day 53, purple). Opposite to that is the clustering of the middle stages of infection (Day 21 and Day 35, blue and green respectively).

Agglomerative hierarchical clustering begins by comparing the metabolomes of each sample (Figure 23). Next, the samples that have the most similar metabolic profile are connected and referred to as a cluster (the mathematical approach to determining similarities among the samples is described below). Each cluster is then connected to another cluster based on the similarities of their metabolic profile. This process continues

until all of the samples have been connected. Therefore, the closer the distance between the samples, the more similar the metabolic profile[67].

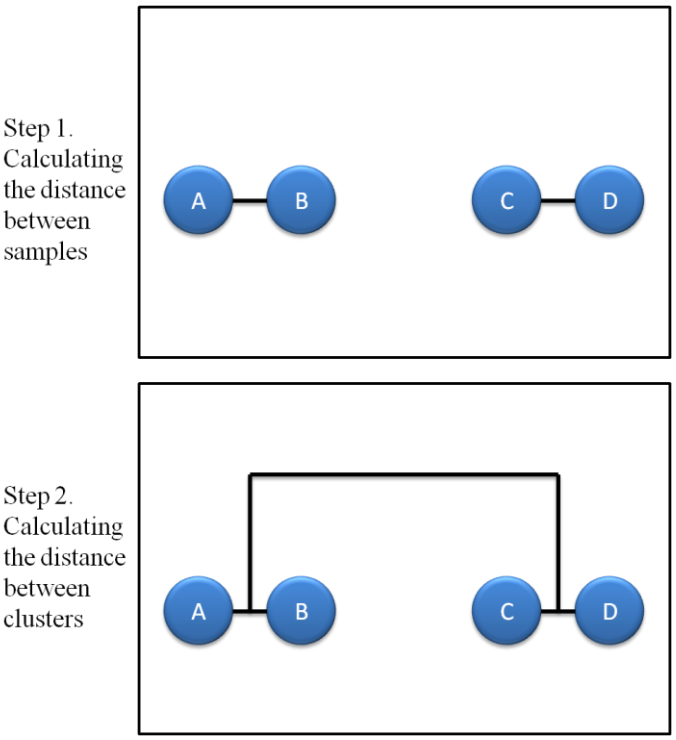


Figure 23. Steps involved in Agglomerative Hierarchical Clustering analysis to determine the similarities between samples.
Step 1 calculates the distance between the samples (A, B, C, and D). As in the PCA, each sphere represents all of the metabolite information found within the sample. If the metabolite profile (appearance and abundance) is similar, a connection will be made. Therefore, since the metabolic profile of A and B are similar, they are placed closer together, and a connection is made, likewise for C and D. After the first connection is made, the groups of samples are described as clusters. Step 2 calculates the distance between the clusters. If the cluster contains a similar metabolic profile, a connection will be made. Therefore, the closer the clusters and ultimately the samples are to each other, the more related their metabolic composition.

There are two steps to calculating the distances in the dendrogram. The first step is to calculate the distance between the samples. The second is to calculate the distance between the sample clusters.

Step 1. Calculating the distance between samples

The most common approach to calculating the distance between two samples is the Euclidean method[67,69]. This distance calculation uses the Pythagorean theorem applied to multidimensional data. In this procedure, the distance is described as a line segment connecting two distinct points (samples). The points exist in a multidimensional space and are provided coordinates which equal the number of dimensions. As stated in the PCA section of this dissertation, each variable (molecular feature) is supplied its own dimension. For example, for an analysis examining 3 metabolites, there will be 3 dimensions, and the coordinates for samples a and b are described as (a_1, a_2, a_3) and (b_1, b_2, b_3), respectively. Thus, for a metabolomics investigation one can assume that the coordinates for sample a would equal (a_1, a_2, ..., a_n) where n is equal to the number of metabolites. The distance between samples a and b would then be described as in Equation 4.

$$d(a,b) = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + \cdots + (b_n - a_n)^2}$$

Equation 4. Euclidean Distance calculation to determine the distance between two points.
Distance (d) between two points a and b is calculated by the square root of the sum of their squared differences where n is equal to the number of metabolites.

Step 2.Calculating the distance between sample clusters

The second step in calculating the hierarchical cluster is to calculate the distance between the clusters. The most common method, the Ward's method, aims to minimize the variation within clusters[67]. In this approach, the mean value of all the distances calculated between the samples in the cluster (the distances calculated in step 1) is used as the representative. The distance between clusters is then calculated by determining the sum of the squared distance between those clusters. This process continues until the distances between all clusters is calculated. The clusters with the shortest distance are then connected and the process begins again[70].

Correlation Network Analysis

Correlation Network Analysis (Figure 24) permits the visualization of the positive and/or negative relationships and interactions amongst various metabolites (and/or other biomolecules) in a sample[15,18,71]. Depending on the nature of the study, the networks can include data derived from the metabolome, proteome, and/or genome, bridging a gap which often occurs between the various "-omics" sciences, providing a more cohesive analysis.

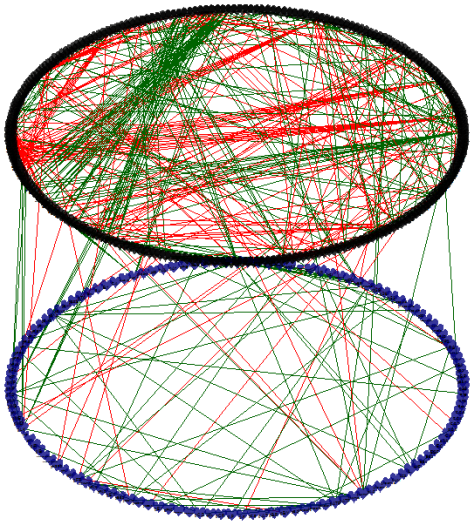


Figure 24. Correlation Network depicting the connections made within and between metabolites and proteins.
The correlation network presented depicts the correlations ($r > |0.95|$) identified within and between the metabolites and proteins acquired from a biological sample. The metabolites are shown in the top circle and are illustrated using black points. Within this circle, there are positive (green lines) and negative (red lines) correlations identified between the metabolites. The proteins are shown in the bottom circle and are illustrated using blue points. As with the metabolites, there are positive and negative correlations identified between the various proteins found within the biological sample. In addition, there are connections drawn between various metabolites and proteins. These connections illustrate the positive and negative interactions/regulations between the metabolites and proteins.

The networks pictorially represent the correlations between the metabolites, where green lines represent a positive correlation and red lines denote a negative correlation. When a unique network is prepared for each cohort, the connections made within the plots can provide insight into the alterations to metabolite interactions. This in

turn may lead to the identification of potential biomarkers or to mechanistic insight (Figure 25)[18,19,71].

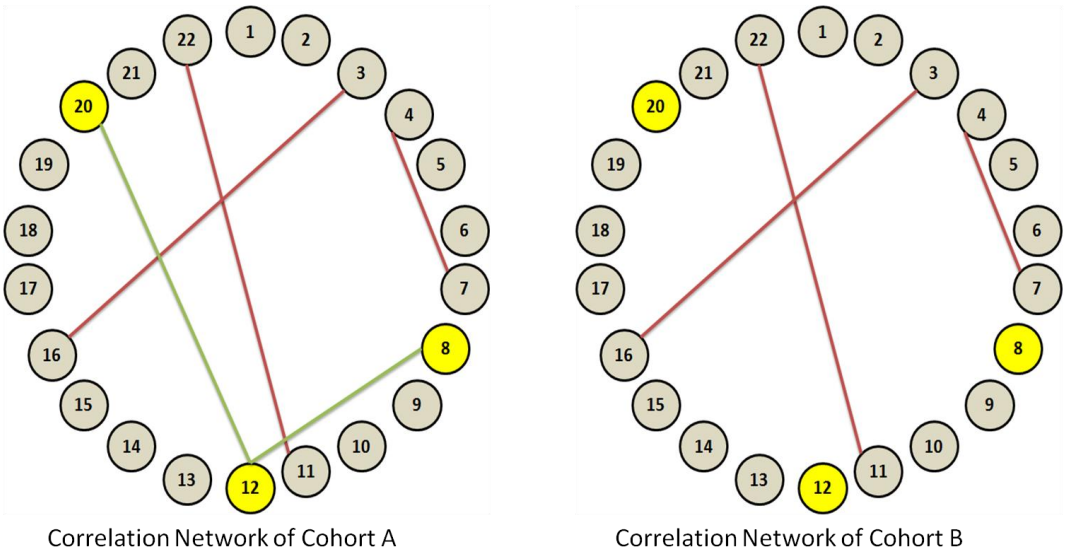


Figure 25. Correlation networks of two juxtaposed biological cohorts.
The correlation network presented is a reflection of the correlations identified between the metabolites in cohorts A and B. The plot depicts 22 different metabolites labeled numerically and are conserved between both plots. That is, the metabolite labeled 1in cohort A is the same metabolite labeled 1 in cohort B. While it appears that both cohorts exhibit similar correlations between the metabolites, the metabolites highlighted in yellow exhibit positive correlations in cohort A and no correlations in cohort B. This indicates an alteration in the relationship of metabolites 8, 12, and 20 between the two cohorts. By focusing my attention on those metabolites specifically, I can determine their identity, biological relevance, and elucidate other pathway specific information.

Stage 4 of the Data Analysis Pipeline: Determination and Identification of Metabolites of Interest

The final stages of a metabolomics investigation is to determine and identify key metabolites of interest. This can occur by identifying alterations to specific pathways as a consequence of disease or determining an appropriate biomarker for disease prediction.

To determine the top metabolites of interest, I filter the data set by using combination of univariate analyses such as fold change, p-value, and frequency values. I generally focus my attention on metabolites that exhibit a high level of fold change ($|\log_2(\text{fold change}) \geq 1.5|$), statistically significant p-value < 0.05, and at least a 50% appearance in one of the cohorts under investigation (the process for determining fold change and p-value is described in detail below). To identify the key metabolites, I use AMDIS in combination with the NIST database for the GC-MS acquired data set, while I compare the LC-MS/MS data to the Metlin and Human Metabolome Database (HMDB) to identify nonvolatile metabolites. Finally, potential biomarkers are assessed using the area under a Receiver Operating Characteristics (ROC) curve, as described below.

Fold Change

The relative fold change for a metabolite is calculated by comparing its median abundance value in each cohort under investigation (Equation 5).

$$log_2(Fold\ Change) = \log_2\left(\frac{median_A}{median_B}\right)$$

Equation 5. Calculation of $\log_2$ fold change for metabolite abundance in juxtaposed biological cohorts.
$median_A$: median value of metabolite in Cohort A
$median_B$: median value of metabolite in Cohort B

The $\log_2$ fold change value is used to determine whether the metabolite is up or down regulated with respect to cohort B. While this analysis is straightforward, one scenario arises where an additional consideration is required. That is, in the case where a metabolite is present in one cohort but completely absent in the other cohort (i.e. there is no median value for the latter cohort). In this case I impute a missing value of 1 for the cohort lacking the metabolite, to avoid a mathematical error reflecting division by zero (see Equation 5).

P-value

After calculating the fold change value for each metabolite in the data set, I calculate the p-value of that metabolite. Here I utilize a two tailed unpaired (two-sample) T-test and a statistically significant value of $p < 0.05$. As the selected T-test is unpaired and two tailed, I am able to compare the two independent cohorts and examine the data in both directions (the disease group can be either higher or lower than the healthy group)[72]. It is important to recognize that a p-value of 0.05 means that there is a 5% chance of obtaining the observed feature only if the null hypothesis is true (in this case, the feature's inclusion is the result of random sampling and not attributed to alterations in metabolites). Therefore, if a feature is retained because the p-value < 0.05, it may actually be a false positive (discovery)[73]. To reduce the false discovery rate, adjusted p-values are calculated using the Benjamini-Hochberg method. In this method, the p-values are listed in ascending order and assigned a rank relative to their position. Next, the critical value is

calculated (Equation 6) where the false discovery rate (Q) is a value chosen by the user. The higher the false discovery rate value, the more stringent the results[73]. After calculating the critical value, the p-values are compared to the critical value. The largest p-value that is still less than the critical value is then established as the new p-value cutoff. Once the new cutoff is determined, the new p-value cutoff is transformed to reflect a p-value < 0.05. This new adjusted p-value is calculated by multiplying the original p-value by the ratio of the total number of tests (m) to the rank of the individual (i).

$$critical\ value = (\frac{i}{m})Q$$

Equation 6. Benjamini-Hochberg Critical Value for the reduction of the false discovery rate.
i: the rank of the individual
m: total number of tests
Q: false discovery rate

Selection of Top Features

Following the calculation of fold change and adjusted p-values, I generate volcano plots (Figure 26) to visualize the number of biologically significant metabolites that are within our cut-off value ($|\log_2(\text{Fold Change}) \geq 1.5|$ and $\text{p-value}_{adj} < 0.05$). By selecting these cutoff values, I am able to focus my attention to only those highly varied but statistically significant metabolites.

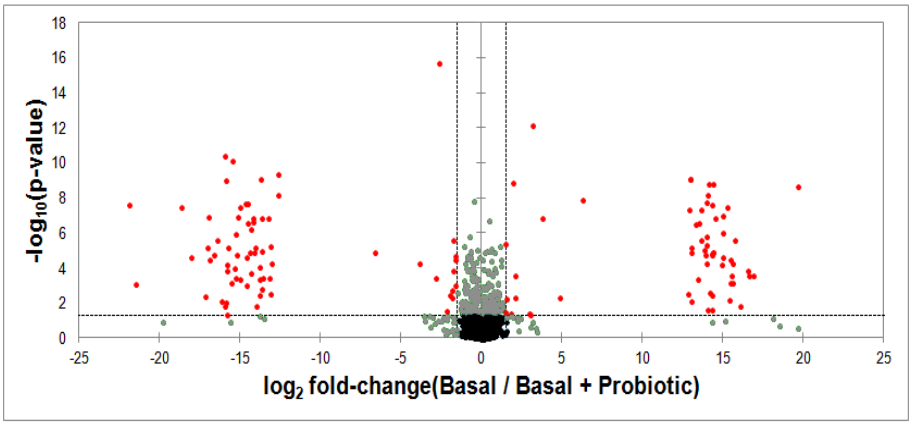


Figure 26. Volcano Plot depicting the fold change versus p-value acquired for differentially-compared metabolites.
The volcano plot depicts the $\log_2$ fold change values versus the $-\log_{10}$p-value. The metabolites that are within our cutoff are colored in red, those are within either p-value or fold change cutoff are in grey, and those that are not within either cutoff value are in black.

After determining which molecular features in the MS acquired data set are statistically different between the cohorts, they are then identified via a database. For GC-MS acquired data, the features have already been identified using the NIST database (Step 4 in AMDIS). However, for the LC-MS acquired data, there is a list of unknown molecular features. To identify the unknown features, I use the acquired MS/MS data obtained from each molecular feature and match the acquired molecular spectra to a library (Metlin and the Human Metabolome Database) molecular spectra using the CID energies of 10eV, 20eV, and 40eV. After metabolite identification, I determine biological relevance and if needed, whether the metabolite could serve as a good biomarker of the disease (via a ROC curve).

ROC Curves

The determination and identification of a biomarker for a particular disease or organism requires the use of many statistical analyses, testing, and reproducibility to ensure that marker is adequate for disease prediction. By using the statistical tools described above, it is possible to first elucidate the metabolites associated with the variations observed between healthy and diseased individuals. By identifying which metabolites cause the greater amount of variance, those metabolites can be extracted from the data set and a Receiver Operating Characteristics (ROC, Figure 27) curve is generated to evaluate the metabolite's ability to serve as a good biomarker[74].

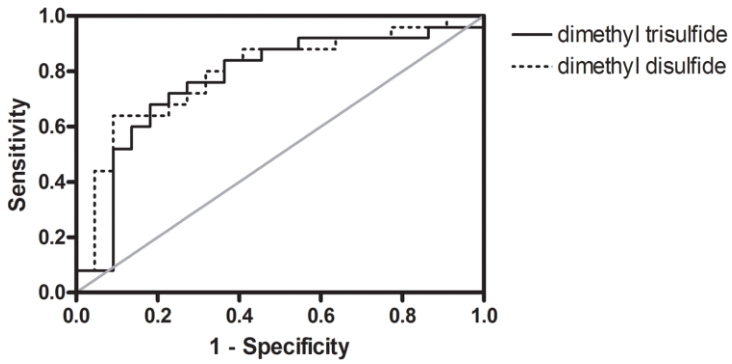


Figure 27. ROC Curve illustrating the potential of using dimethyl trisulfide or dimethyl disulfide as a biomarker.
This ROC curve is depicting relationship of the true positive rate (sensitivity) versus the false positive rate (1-specificity) of the metabolites dimethyl trisulfide and dimethyl disulfide[33]. The area under the ROC curve provides an indicator for the feasibility of a metabolite to be biomarker. An excellent biomarker is described as having an area under the curve (AUC) equal to 1, while an AUC of 0.5 means the analyte should not be used as a biomarker of disease. For these analytes, the AUCs are 0.77 and 0.80 respectively and are considered good biomarkers.

While the ROC curve itself serves as a visual tool, the value obtained from the area under the ROC curve (AUC, C-statistic), determines the metabolite's ability to be a good biomarker. The AUC is calculated using a non-parametric method known as the trapezoid rule. In this method, each adjacent point under a curve is depicted as a trapezoid. The area is then computed using the equation demonstrated in Equation 7.

$$AUC = \left(\Delta X \left(\frac{Y_1 + Y_2}{2}\right) + \Delta X \left(\frac{Y_2 + Y_3}{2}\right) + \cdots + \Delta X \left(\frac{Y_{n-1} + Y_n}{2}\right)\right)$$

Equation 7. Calculating the area under the curve using the trapezoid rule.
The area under the curve is calculated using the sum of the width of each trapezoid (Δx), multiplied by the sum of the length of the points from the baseline (y), divided by the number of points (2). This will continue for n number of points.

To be denoted as an excellent biomarker for a disease, the metabolite is described as having the maximum area under the curve (AUC also known as the C-statistic) possible and equates to a value of 1[16,17]. Conversely, those analytes whose curve lies along the diagonal line exhibit areas equaling 0.5. This value of 0.5 is attributed to random guessing and will exclude them from the analysis.

Summary

My goal in specific aim 1 was to develop an approach to a metabolomics based analysis (Figure 28).

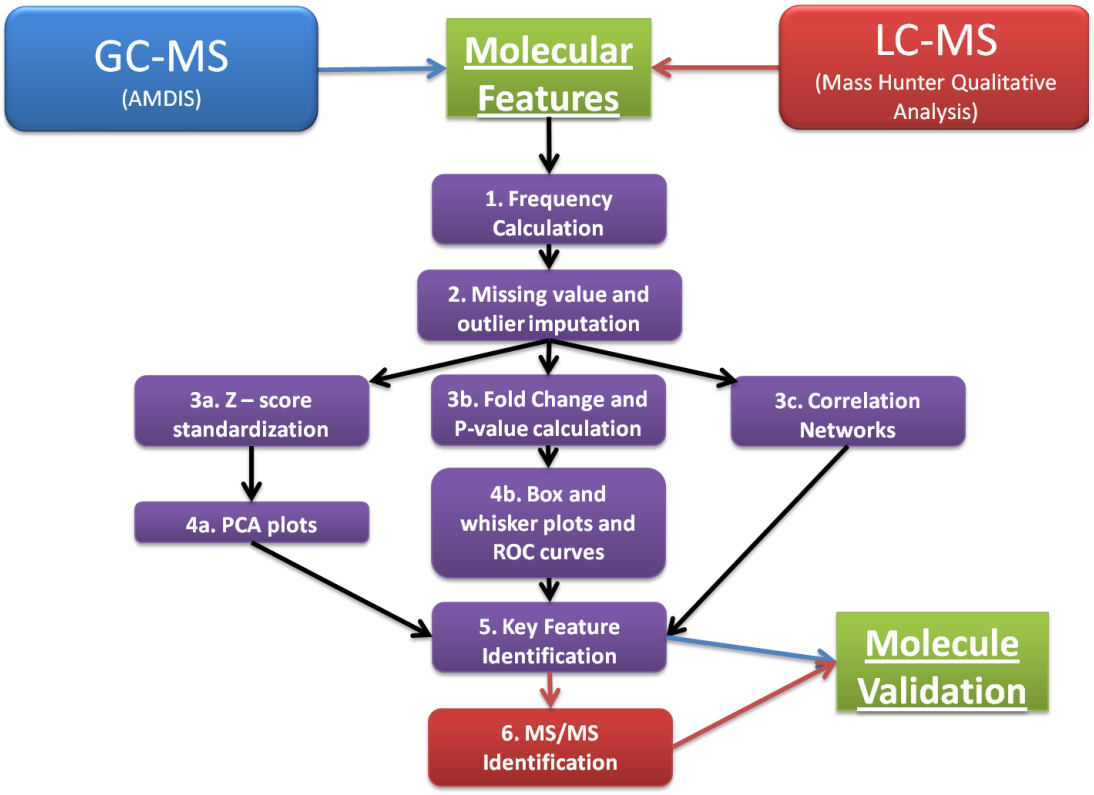


Figure 28. The Metabolomics Pipeline for Data Acquisition and Analysis.
The flowchart illustrated above is a modified version of the original metabolomics pipeline from Figure 10. In this modified version, I added the software used for determining the molecular features in GC-MS (AMDIS), and LC-MS (Agilent's MassHunter Qualitative Analysis) acquired data sets. I also illustrated the portion of the pipeline that differs between GC-MS and LC-MS data. This divergent step occurs after Step 5 in the original pipeline. For LC-MS data, I identify the molecular features using the acquired MS/MS data. For the GC-MS acquired data, AMDIS has already determined the identities of the molecular features in step 4 of its analysis.

As mentioned earlier, this pipeline was developed alongside the metabolomics analyses described in Specific Aim 2. The pipeline has been tailored primarily to a global metabolomics analysis and encompasses both volatile and non-volatile investigation routes, although is readily applicable to targeted analyses.

Pitfalls and Limitations to the Pipeline

As illustrated in this specific aim, metabolomic based analyses require a multitude of different components. A major limitation to this pipeline is the length of time for the analysis. To obtain a comprehensive assessment of the metabolite composition of the samples, I utilize multiple chromatography approaches (GC and LC based). While this increases the amount of information I obtain from the samples, this also increases the data acquisition analysis time. In addition, there is not one sole piece of software that can acquire and analyze the data set, therefore I had to incorporate a vast amount of different software platforms and algorithms, requiring me to develop technical expertise in each.

The ability to determine all metabolites of interest is limited for the LC-MS acquired data set. Following the acquisition of MS data, MS/MS data is acquired for the five most abundant ions using CID energies of 10eV, 20eV, and 40eV. After analyzing MS/MS data generated from another lab, I discovered that their technique of pooling all samples into one vial for MS/MS analysis resulted in very few of the top features acquiring MS/MS data, due to out-competing of metabolites with similar retention times. Because of this, I decided to collect MS/MS data on all samples. While I found that this new approach significantly increases the number of features with MS/MS acquired data (from 20% to at least 90%), this also significantly increases the analysis time as each sample will have 4 chromatographic runs (MS, 10eV MS/MS, 20eV MS/MS, and 40eV MS/MS).

Following the acquisition of the LC-MS acquired data, I match the acquired mass spectra of the molecular feature obtained during the analysis to a mass spectra of a known

compound in the Metlin and Human Metabolome Database (HMDB). While these databases are among the most commonly cited, there is a limited number of metabolites with MS/MS derived data within these databases. With the added stringency of matching the 10eV, 20eV, and 40eV mass spectra to the database, many of the top features remain unidentified.

SPECIFIC AIM 2.1: METABOLOMICS AS A BIOSENSOR; DEVELOPING AN ELECTRONIC NOSE.

Objective: Use metabolomics as an electronic nose to determine if global metabolic profiling of microbial volatile organic compounds can differentiate select biological warfare agents.

Introduction

Francisella tularensis (the causative agent of tularemia), *Burkholderia pseudomallei* (melioidosis), and *Brucella melitensis* (brucellosis) are zoonotic intracellular parasites considered by the US Centers for Disease Control and Prevention (CDC) as high risk biological warfare agents due to their high morbidity/mortality rate, ease of dissemination, emergency response procedures, and projected social impact[75]. Because these agents can be properly managed with early detection, rapid and accurate pathogen identification is essential.

Though sensitive and selective detection techniques involving Polymerase Chain Reaction (PCR), microbial culturing, and/or Enzyme-Linked Immunosorbent Assays (ELISA) are well established,[76] the techniques are typically time consuming, laborious, and costly. To circumvent these issues and to enhance the capabilities of the biodefense and public health sector, continued exploration of methods and approaches for the detection of biological warfare agents is vital.

Microbial volatile organic compounds (mVOCs) are a large class of structurally diverse, microbial-derived organic molecules, generally related by their volatility at room temperature. In our lab, we have demonstrated the diagnostic potential of VOCs derived from biological samples[77–80]. Similarly, other studies report on the differentiation of bacteria using the mVOC metabolome. For example, studies comparing the mVOCs produced by sepsis inducing bacteria have identified unique biomarkers indicative of *S. aureus*, *P. aeruginosa*, and *E. coli*[81–83]. Differentiation of antibiotic susceptible and resistant strains has also been achieved through mVOC profiling[84].

To perform our analysis of VOCs emanating from biological samples, we utilize a specialized sampling method known as headspace solid phase microextraction (hSPME). By using hSPME, we isolate and concentrate the analytes present in the sample, greatly facilitating their detection (Figure 29)[6,26,85–87].

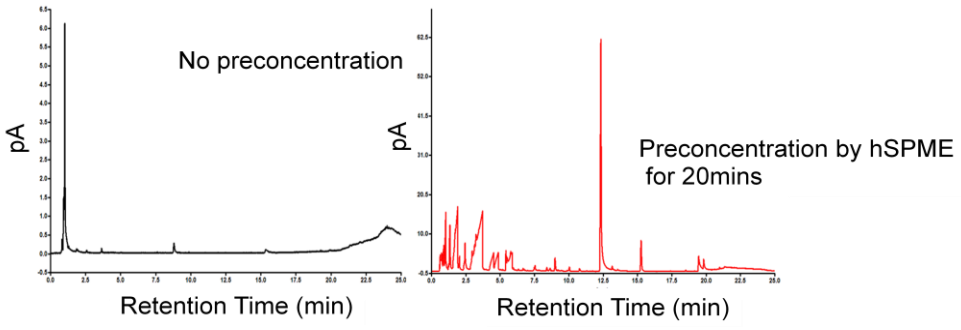


Figure 29. Enhanced analysis of VOCs using hSPME.
With direct injection, headspace VOCs are too dilute and difficult to detect by GC (left chromatogram). However, using hSPME, a sorbent (a polymer coating on a silica fiber) preconcentrates the headspace VOCs before analysis (see Figure 30), thereby permitting their detection (right chromatogram). Both chromatograms were derived from an identical biological sample and were analyzed using a GC equipped with a flame ionization detector (FID). Note the difference in the magnitude of the Y-axis (signal strength) and in the number of peaks (detected metabolites) when hSPME is used. A single hSPME fiber type was used for the right chromatogram. Also note that with a 20 min hSPME and 25 min GC run, VOC fingerprinting is complete in less than 1hr.

A typical hSPME analysis (Figure 30) involves the extraction of VOCs via partitioning into a polymeric coating adhered to a fused silica rod (fiber), subsequent desorption of the VOCs by heating the fiber in the injection port of a gas chromatograph, separation of the VOCs by gas-liquid partition chromatography, and detection of the VOCs via flame ionization or mass spectrometry. Comparison to a reference database enables VOC identification.

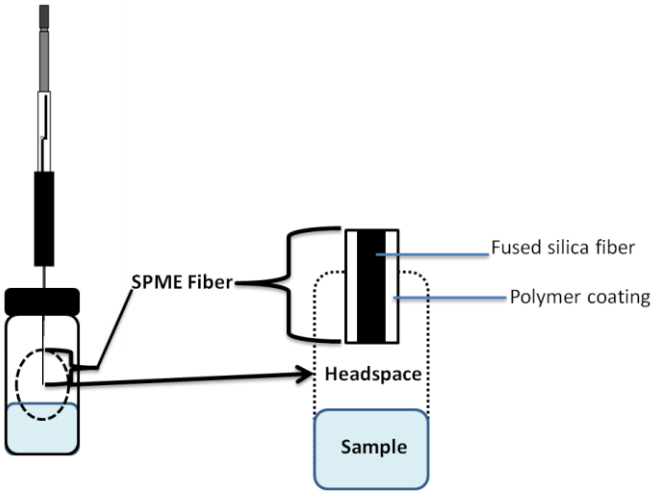


Figure 30. Solid-phase microextraction of VOCs in the headspace above a biological sample.
The SPME fiber is exposed to the headspace above the sample to facilitate the extraction of VOCs. The volatile analytes emanating from the sample enter the headspace where they interact and bind to the polymeric coating on the SPME fiber.

The choice of polymeric coating is an important consideration when performing a hSPME analysis, as it dictates the type of analytes that are extracted ("like dissolves like"). Several SPME fiber coatings are commercially available, including polyacrylate (PA), polydimethylsiloxane (PDMS), carbowax-polyethylene glycol (PEG), and mixed

phases of carboxen (CAR)-PDMS, divinylbenzene (DVB)-PDMS, and CAR-DVB-PDMS. While the polarity of the analyte of interest is typically used to guide the selection of a particular SPME fiber, the Couch lab has shown that the multifarious nature of biological sample composition dictates the use of multiple SPME fiber types for maximal metabolomic coverage of the total VOCs[80]. The greater the number of metabolites identified, the greater the probability of uniquely differentiating a biological cohort. Thus, we have developed and currently employ a multi-hSPME approach to VOC metabolomic investigations.

While multi-hSPME greatly increases the number of metabolites identified, multiple sample extractions are required (at least one per fiber type), thereby slowing the throughput of sample analysis. Our initial approach to addressing this issue was the use of a combinatorial analysis to identify core three-, four-, and five-fiber groupings that afford substantial (89-96%) coverage of the total VOC metabolome[80]. However, maximal (i.e. 100%) coverage of the VOC metabolome is preferred, so we developed and patented a novel extraction device that permits us to perform simultaneous multifiber extractions of a sample (simulti-hSPME; Figure 31) using all of the commercially available fiber types simultaneously, thereby ensuring maximal VOC metabolome coverage, while dramatically increasing the throughput of a VOC metabolomics investigation.

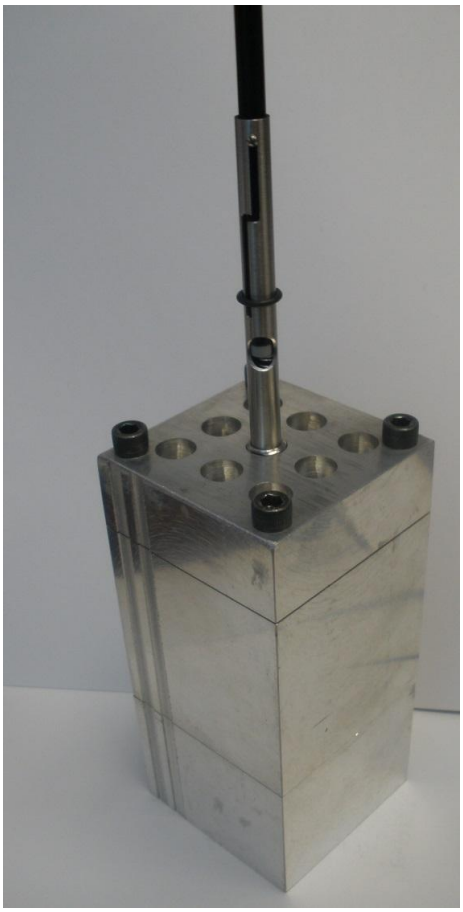


Figure 31. The simulti-hSPME device for multifiber extraction.
For clarity, the device is shown with only one hSPME syringe in place, but it can accommodate up to nine syringes simultaneously. To extract the VOCs, sample vials are placed within the device and then the SPME fibers are exposed to the VOCs in a headspace chamber located directly above the sample vial. Alternatively, the device is assembled with a base piece designed to connect directly in-line with a sample, for example to monitor ambient, inhaled, or exhaled air. Following the extraction, the fibers are then desorbed into a standard GC. We have used the simulti-hSPME device to analyze a diverse array of samples including blood, cell cultures, breath, feces, and ocular fluid.

Extraction duration is another important consideration when performing a simulti-hSPME analysis, as it can dictate the number of analytes that are extracted. To illustrate the effect of extraction duration, the Couch Lab used two different SPME fibers (a CAR-DVB-PDMS and PA fiber) in conjunction with GC-MS to identify and quantify the

VOCs extracted from a human fecal sample[26]. To perform the analysis, samples were placed in the simulti-hSPME device and then an individual SPME fiber was placed into the headspace above the sample for various time intervals (ranging from 1 min to 16 hr; the temperature was held constant throughout the extraction). Extracted analytes were then immediately desorbed into a GC-MS and spectral comparison with the NIST08 database facilitated analyte identification (only compounds with a 90% or greater probability of match to a molecule in the NIST08 library were scored). Figure 32 illustrates a plot of the number of analytes identified as a function of extraction duration.

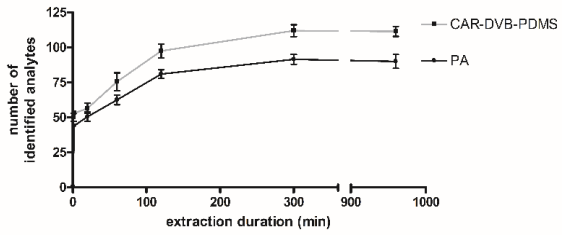


Figure 32. Extraction duration and headspace SPME of human feces[26].
A plot of the number of identified analytes as a function of extraction duration is presented. Nonlinear regression fitting the hyperbolic extraction curves yields a Y_{max} of 114 +/- 3 for the CAR-DVB-PDMS fiber (R^2= 0.9937) and 94 +/- 4 for the PA fiber (R^2= 0.9791).

As seen in Figure 32, the CAR-DVB-PDMS fiber isolated a greater number of identifiable analytes from the sample than did the PA fiber, regardless of the extraction duration. This result highlights the influence of fiber choice on VOC metabolomics. For both the PA and CAR-DVB-PDMS fibers, total analyte extraction appears hyperbolic,

with a near maximum value (Y_{max}) occurring with a 960 minute (16 hour) extraction duration (98% of Y_{max}). As illustrated with the CAR-DVB-PDMS fiber (Figure 33), individual analyte extraction rates are analyte specific, with some metabolites (such as indole and methyl indole) rapidly reaching equilibrium and others (such as acetic acid, propanoic acid, and caryophyllene) proceeding more slowly. In some cases (such as observed with methyl phenol and farnesene), metabolite titers plateau then subsequently wane with increased exposure duration, a phenomena attributed to higher affinity compounds displacing those with lower affinity for the fiber, thereby lowering the titer of the latter.

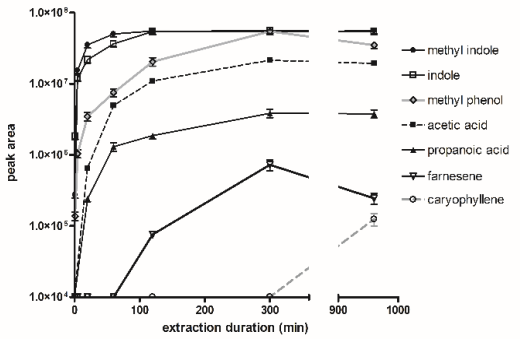


Figure 33. A plot of area under the chromatographic curve as a function of time for the indicated analytes obtained using the CAR-DVB-PDMS fiber.
This plot illustrates the differences in the extraction rates for the indicated metabolites. See text for further discussion.

The key to a successful comparative metabolomics analysis is the inclusion of a sufficient number of metabolites to permit differentiation of the biological cohorts. To reiterate, the greater the number of metabolites identified, the greater the probability of uniquely differentiating a biological cohort. As the extraction profile of a biological sample is hyperbolic, the extraction duration can be altered/optimized to increase or decrease the total number of metabolites included in the derived VOC metabolome (i.e. the number of metabolites in the resulting spreadsheet/data set is a reflection of the extraction duration). For clinical diagnostics, the goal is to define the shortest extraction duration permitting clear differentiation of the cohorts. The Couch Lab has illustrated the impact of extraction duration in a fecal VOC analysis, wherein 18 hour and 20 minute extractions (via simulti-hSPME) result in VOC metabolomes with a sufficient number of metabolites to differentiate alcoholics from healthy study participants, while a 2 minute extraction duration does not (Figure 34).

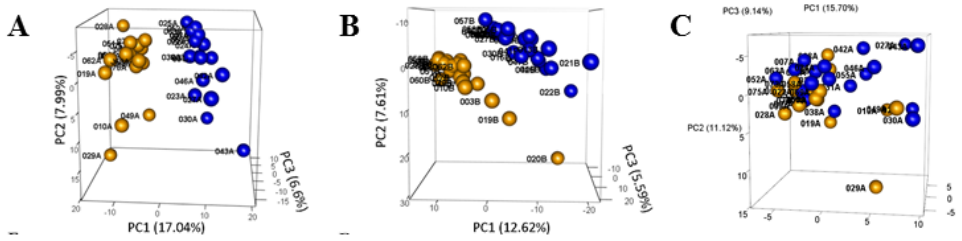


Figure 34. 3D-PCA illustrating the effects of extraction duration and cohort differentiation.
Extraction duration determines the number of metabolites included in a metabolomics data matrix, and thus effects the probability of successfully differentiating biological cohorts using statistical analyses. To illustrate this, we performed a comparative fecal VOC analysis of healthy study participants vs. clinically diagnosed alcoholics. Three dimensional PCA plots generated from the derived fecal VOC metabolomes are shown. Healthy participants are identified as blue spheres, while alcoholic patients are denoted as yellow spheres. While an 18 hr (A) and 20 min (B) simulti-hSPME analysis of the feces permits the clear differentiation of the healthy and alcoholic patients, a 2 min (C) simulti-hSMPE does not. Through optimization of the extraction duration used with simulti-hSPME, a rapid diagnostic test can be developed.

To address the objective of Specific Aim 2.1, I sought to determine if global metabolomic profiling of the mVOC metabolome could generate a microbial fingerprint to uniquely identify and differentiate select biological warfare agents (*Francisella tularensis*, *Burkholderia cenocepacia*, and *Brucella neotomae*). A multifiber hSPME analysis was performed, and to enable rapid visualization of the mVOC metabolomes, I converted the chromatograms into binary plots resembling retail Universal Product Codes (UPC bar codes). In these binary plots, the presence of a white line indicates the presence of a peak in the chromatogram at the specified retention time (Figure 35).

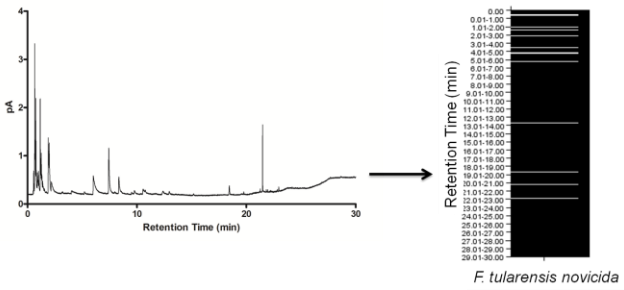


Figure 35. Conversion of GC-FID chromatogram to a binary plot.
On the left is a chromatogram obtained from a hSPME extraction of an *F. tularensis* liquid culture. The analytes acquired during the extraction were separated via gas-liquid partitioning and the resulting chromatographic peaks were converted to the binary plot. On the right, is the binary plot. This plot resembles a UPC code where the white bars are indicative of a peak at a specified retention time (y-axis).

This multifiber hSPME analysis was performed prior to the development of simulti-hSPME in the Couch lab, and as such, single hSPME fibers were individually used for the mVOC extraction. The corresponding binary plots were generated for each

of the fibers employed, then the plots were condensed/compiled into one single plot to create a mVOC fingerprint for the bacterium.

In addition, I also explored the influence of select environmental effectors on the mVOC fingerprints generated above, and went on to produce mVOC fingerprints that differentiate antibiotic sensitive and antibiotic resistant strains of the biothreat agents. All of this work is described in detail below.

Materials and Methods

1. Bacterial Strains and Culture Media

The following reagents were obtained through the NIH Biodefense and Emerging Infections Research Resources Repository, NIAID, NIH: *Francisella tularensis spp. novicida*, Strain CG62, NR-580; *Burkholderia cenocepacia*, Strain LMG 16656, NR-701; *Brucella neotomae*, Strain 5K33, NR-684, *Yersinia pestis*, Strain A1122, and *F. tularensis*, Strain NIH B38. Bacterial cultures were grown using rich media containing Tryptic Soy Broth (TSB) supplemented with 0.1% cysteine (TSBC), TSBC supplemented with 50 µg/mL kanamycin, modified Muller-Hinton media (mMH; 0.025% ferric pyrophosphate, 1 mM $CaCl_2$, 1 mM $MgCl_2$, 0.1% glucose, and 2% Isovitalex supplement), mMH + 10ug/mL of kanamycin, or a defined minimal media (referred to as Modified Chamberlin's Defined Media (MCDM)) containing 0.4 g/L L-Arginine, 0.4 g/L L-Aspartic Acid, 0.2 g/L L-Cysteine, 0.2 g/L L-Histidine, 0.4 g/L L-Isoleucine, 0.4 g/L L-Leucine, 0.4 g/L L-Lysine, 0.4 g/L L-Methionine, 2.0 g/L L-Proline, 0.4 g/L L-Serine, 2.0 g/L L-Threonine, 0.4 g/L L-Tyrosine, 0.4 g/L L-Valine, 0.04 g/L Spermine

disphosphate, 0.004 g/L Thiamine HCl, 0.002 g/L L-Calcium pantothenate, 4.0 g/L Glucose, 10 g/L NaCl, 0.135 g/L $MgSO_4 \cdot 7H_2O$, 1.0 g/L KH_2PO_4, 1.0 g/L K_2HPO_4, 1.92 g/L Sodium Citrate, 0.002 g/L $FeSO_4 \cdot 7H_2O$, and pH adjusted to 6.2. Agar was added at a concentration of 20 g/L to prepare solid media for petri plates. All work was conducted using aseptic technique in an approved BSL2 laboratory following approved protocols.

2. Bacterial Cultures

2.1. Multifiber Differentiation and Environmental Effectors

Individual bacterial colonies isolated from TSBC agar plates were used to inoculate 2 mL of TSBC in a 15mL Falcon conical tube. The tube was capped with a foam plug and the liquid cultures were incubated for either 8 or 18 hours at 37 °C, 250 rpm. A sterile serological pipette was used to dispense 250 µL aliquots of the culture into amber autosampler vials (VWR Screw Top Vial Amber Glass 15 x 45mm; 4 mL), and the vials were stored at -80 °C until the sample was analyzed.

Alternatively, cells from an overnight TSBC culture were harvested by centrifugation (4 °C, 3800 x g, 15 min), washed three times with 1 mL of MCDM (cells were collected by centrifugation after each wash), then resuspended to an OD_{600} = 1.0 in MCDM or MCDM supplemented with 25 mM $MgCl_2$, 12.5 mM $NiCl_2$, or 6.25 mM NaCl. Cultures were then incubated in a foam capped 125 mL Erlenmeyer flask at 37 °C, 250 rpm, for 8 hours, dispensed in 250 µL aliquots into amber autosampler vials, snap frozen in liquid nitrogen, and stored at -80 °C until analyzed.

2.2. Preparation of competent cells

Yersinia pestis competent cells were prepared as previously described with few modifications[88]. Overnight liquid cultures of *Yersinia pestis* A1122 were grown at 28°C in Tryptic Soy Broth with 0.1% cysteine (TSB-C) at 250 rpm. A healthy overnight culture was diluted 1:50 in 50 mL fresh TSB-C and allowed to grow to an OD_{600} of 0.5 (approximately 3 hours). Cells were harvested by centrifugation at 4000 x g for 10 min at 4°C, then washed once with sterile MilliQ water, and once with sterile *Yersinia* transformation buffer (15% glycerol, 272 mM sucrose). Cells were resuspended in 400 µL of transformation buffer, aliquoted, snap frozen, and stored at -80 °C until used.

Francisella tularensis competent cells were prepared as previously described with some modification[89]. Liquid cultures of *Francisella tularensis* subsp. *tularensis* NIH B-38 were grown in modified Muller-Hinton media (mMH; 0.025% ferric pyrophosphate, 1 mM $CaCl_2$, 1 mM $MgCl_2$, 0.1% glucose, and 2% Isovitalex supplement) at 200 rpm for 3 days. A healthy culture was diluted 1:10 in 80 mL fresh mMH and allowed to grow to an OD_{600} of 0.6 (approximately 3 days). Cells were harvested by centrifugation at 4000 x g for 10 min at 4°C, then washed twice with sterile *Francisella* transformation buffer (500 mM sucrose). Cells were resuspended in 400 µL of transformation buffer, and used fresh the same day for transformation.

2.3. Transformation of microbes

Transformation of *Yersinia pestis* was conducted essentially as described previously[88]. 40 µL of competent cells were incubated with 5 uL of pHSG298 (~100 ng/µL) over ice for 1 min then transferred an ice-cold 0.2 cm electroporation cuvette.

Cells were electroporated with a single electric pulse (25 µF, 200 Ω) at a field strength of 12.5 kV/cm. Transformed cells were added to 1 mL of SOC medium and incubated at 28°C, 250 rpm for 2 hours. Transformants were identified by plating on TSB-C agar plates with 50 µg/mL kanamycin; transformant colonies were visible within 24 hours.

Transformation of *Francisella tularensis* was conducted as described previously, with some modification[89]. 200 µL of competent cells were incubated at room temperature for 10 minutes with 2 uL of pFNLTP1 (~100 ng/µL) then transferred to an ice-cold 0.2 cm electroporation cuvette. Cells were electroporated with a single electric pulse (25 µF, 600 Ω) at a field strength of 12.5 kV/cm. Transformed cells were added to 1 mL of mMH broth and incubated at 37°C, 250 rpm for 6 hours. Transformants were identified by plating on mMH agar plates with 10 µg/mL kanamycin; transformant colonies were visible within 4 days.

2.4. Differentiation of Wild type and Kanamycin Resistance

Individual bacterial colonies isolated from TSBC agar plates for *Y. pestis* wild type or TSBC + 50 µg/mL kanamycin agar plates for *Y. pestis* kanamycin resistance were used to inoculate overnight seed cultures using 10 mL of TSBC (or 10 mL of TSBC + 50 µg/mL kanamycin) in a 25mL Falcon conical tube. For *F. tularensis*, individual bacterial colonies isolated from MMH agar plates for wild type or MMH + 10 µg/mL kanamycin agar plates for kanamycin resistance were used to inoculate overnight seed cultures using 10 mL of MMH (or 10 mL of MMH + 10 µg/mL kanamycin) in a 25mL Falcon conical tube. *The* tube was capped with a loose lid and tapped to prevent the lid from falling off. The liquid cultures were incubated for 18 hours at 37 °C, 250 rpm. The next morning,

OD_{600} was measured and production flasks were generated using 500 µL of an OD_{600} = 1 culture to 25 mL of fresh TSBC (or MMH) media in a 125 mL Erlenmeyer flasks. For the resistance strains, 50 µg/mL (or 10 µg/mL) of kanamycin was added to the culture. The liquid production flasks were incubated for 24 hours at 24 °C, 200 rpm. For the simulti-hSPME extraction, the simulti-hSPME device was placed over top the culture. For all other extractions, 250 µL aliquots of the production flask and corresponding media were dispensed into amber autosampler vials, snap frozen in liquid nitrogen, and stored at -80 °C until analyzed.

3. mVOC Profiling

3.1. Multifiber Differentiation and Environmental Effectors

All samples were analyzed by hSPME. Sample vials were preheated to 37 °C for 30 minutes then a SPME fiber assembly was manually positioned into the headspace above the culture and the fiber exposed to the volatiles for 60 minutes (the sample vial temperature was held at 37 ^{o}C for the duration of the exposure). The fiber assembly was then placed into the GC inlet for thermal desorption of the analytes. The following SPME fibers (Supelco, Bellefonte, PA) were used in the investigation: DVB-PDMS 65 µm, PA 85 µm, CAR-PDMS 75 µm, CAR-PDMS 85 µm with stableflex, PDMS 100 µm, PDMS 7 µm, PEG 60 µm, and CAR-DVB-PDMS 50/30 µm with stableflex. All fibers were preconditioned before use, as per the manufacturer's instructions (Table 1). After every sample analysis, the fiber was reanalyzed (without exposure to a sample) to ensure complete desorption of analytes. All analyses were performed in triplicate.

Table 1. SPME Fiber Operational Conditions.

Fiber	Inlet Temperature (°C)	Precondition Time (min)
CAR/PDMS 75 µm	300	60
PDMS/DVB 65 µm	250	30
PDMS 7 µm	320	60
DVB/CAR/PDMS 50/30µm Stableflex	270	60
CAR/PDMS 85 µm Stableflex	300	60
CW (PEG) 60 µm	240	30
PDMS 100 µm	250	30
PA 85 µm	280	60

3.2. Differentiation of Wild type and Kanamycin Resistance

All samples were analyzed by hSPME. For the 60 minutes extraction trials, sample vials were preheated to 37 °C for 30 minutes then a SPME fiber assembly was manually positioned into the headspace above the culture and the fiber exposed to the volatiles for 60 minutes (the sample vial temperature was held at 37 ^{o}C for the duration of the exposure). The fiber assembly was then placed into the GC inlet for thermal desorption of the analytes. The following SPME fibers (Supelco, Bellefonte, PA) were used in the investigation CAR/PDMS 85 µm with stableflex, 100 µm PDMS, PDMS/DVB 65 µm, PA 85 µm, DVB/CAR/PDMS 50/30 µm, PEG 60 µm. All fibers were preconditioned before use, as per the manufacturer's instructions but for 15 minutes each. Prior to the initial sample analysis, the fiber was analyzed (without exposure to a sample) to ensure complete desorption of analytes. The analysis was performed in triplicate.

For the 15 minutes extraction trials, sample vials were preheated to 80 °C for 15 minutes (to compensate for the shorter incubation and extraction duration) then a SPME fiber assembly was manually positioned into the headspace above the culture and the

fiber exposed to the volatiles for 15 minutes (the sample vial temperature was held at 80 ^{o}C for the duration of the exposure). The fiber assembly was then placed into the GC inlet for thermal desorption of the analytes. The following SPME fibers (Supelco, Bellefonte, PA) were used in the investigation CAR/PDMS 85 µm with stableflex, 100 µm PDMS, PDMS/DVB 65 µm, PA 85 µm, DVB/CAR/PDMS 50/30 µm, PEG 60 µm. All fibers were preconditioned before use, as per the manufacturer's instructions but for 15 minutes each. Prior to the initial sample analysis, the fiber was analyzed (without exposure to a sample) to ensure complete desorption of analytes.

For the simulti-hSPME trials, the sample flasks were preheated to 37 °C for 30 minutes then simulti-hSPME device was manually positioned on top of the flask. The SPME fibers were inserted into the device and exposed to the volatiles in the headspace above the culture 30 minutes (the sample vial temperature was held at 37 ^{o}C for the duration of the exposure). The fiber assembly was then placed into the GC inlet for thermal desorption of the analytes. The following SPME fibers (Supelco, Bellefonte, PA) were used in the investigation CAR/PDMS 85 µm with stableflex, 100 µm PDMS, PDMS/DVB 65 µm, PA 85 µm, DVB/CAR/PDMS 50/30 µm, PEG 60 µm. All fibers were preconditioned before use, as per the manufacturer's instructions but for 15 minutes each. Prior to the initial sample analysis, the fiber was analyzed (without exposure to a sample) to ensure complete desorption of analytes. The analysis was performed in triplicate.

4. Instruments

4.1. Multifiber Differentiation and Environmental Effectors

Samples were analyzed using an Agilent 6890 Plus GC-FID equipped with a DB5-MS capillary column (Agilent, Palo Alto, CA), 15 m in length, 0.25 mm ID, and 0.25 μm film thickness, and a 0.75 mm ID SPME injection port liner operated in splitless mode at varying inlet temperatures (Table 1). Helium carrier gas was used at a flow rate of 1.5 mL/min and the GC oven was held at an initial temperature of 35 °C for 1 min, ramped to 80 °C at 3 °C/min, then to 120 °C at 10 °C/min, to 260 °C at 40 °C/min and held for 2 min, and finally to 280 °C at 40 °C/min. The final temperature of 280 °C was held for 2.5 min. The total run time for each analysis was 30 min.

4.2. Differentiation of Wild type and Kanamycin Resistance

For the GC-FID trials using the 60 minutes extraction duration, the samples were analyzed using an Agilent 6890 Plus GC-FID equipped with a RXI-5Sil MS capillary column (Restek, Bellefonte, PA), 30 m in length, 0.25 mm ID, and 0.5 μm film thickness, and a 0.75 mm ID SPME injection port liner operated in splitless mode at varying inlet temperatures (Table 1). Helium carrier gas was used at a flow rate of 1.5 mL/min and the GC oven was held at an initial temperature of 35°C for 1 min, ramped to 80°C at 3°C/min, then to 120°C at 10°C/min, to 260 °C at 40°C/min and held for 2 min, and finally to 280°C at 40°C/min. The final temperature of 280°C was held for 2.5 min. The total run time for each analysis was 30 min.

For the GC-FID trials using the 15 minutes extraction duration, samples were analyzed using an Agilent 6890 Plus GC-FID equipped with a RXI-5Sil MS capillary column (Restek, Bellefonte, PA), 30 m in length, 0.25 mm ID, and 0.5 μm film thickness, and a 0.75 mm ID SPME injection port liner operated in splitless mode at varying inlet temperatures (Table 1). Helium carrier gas was used at a flow rate of 1.5 mL/min and the GC oven was held at an initial temperature of 35°C for 1 min, ramped to 50°C at 3°C/min, then to 300°C at 27.5°C/min. The total run time for each analysis was 15.09 min.

For the GC-MS trials using the 60 minutes extraction duration, the samples were analyzed using an Agilent 5977B MSD equipped with a Agilent HP-5ms ultra inert column, 30 m in length, 0.25 mm ID, and 0.25 μm film thickness, and a 35 μL ultra inert SPME injection port liner operated in splitless mode at varying inlet temperatures (Table 1). Helium carrier gas was used at a flow rate of 1.5 mL/min and the GC oven was held at an initial temperature of 35°C for 1 min, ramped to 80°C at 3°C/min, then to 120°C at 10°C/min, to 260 °C at 40°C/min and held for 2 min, and finally to 280°C at 40°C/min. The final temperature of 280°C was held for 2.5 min. The total run time for each analysis was 30 min.

For the GC-MS trials using the 15 minutes extraction duration, samples were analyzed using an Agilent 5977B MSD equipped with a Agilent HP-5ms ultra inert column, 30 m in length, 0.25 mm ID, and 0.25 μm film thickness, and a 35 μL ultra inert SPME injection port liner operated in splitless mode at varying inlet temperatures (Table 1). Helium carrier gas was used at a flow rate of 1.5 mL/min and the GC oven was held at

an initial temperature of 35°C for 1 min, ramped to 50°C at 3°C/min, then to 300°C at 27.5°C/min. The total run time for each analysis was 15.09 min.

For the simulti-hSPME trials, the samples were analyzed using an Agilent 5977B MSD equipped with a Agilent HP-5ms ultra inert column, 30 m in length, 0.25 mm ID, and 0.25 μm film thickness, and a 35 μL ultra inert SPME injection port liner operated in splitless mode at 240 °C. Helium carrier gas was used at a flow rate of 1.5 mL/min and the GC oven was held at an initial temperature of 35°C for 1 min, ramped to 80°C at 3°C/min, then to 120°C at 10°C/min, to 260 °C at 40°C/min and held for 2 min, and finally to 280°C at 40°C/min. The final temperature of 280°C was held for 2.5 min. The total run time for each analysis was 30 min.

5. Data processing

Chromatograms were converted into binary plots via SciLab utilizing a custom script (see Appendix 3 for the script) written in the MatLab technical computing language. To expedite the throughput of binary plot generation, I wrote a custom Perl script to automatically generate the plots once data was provided (see Appendix 4 for the script). Following a sample run, the peaks in the chromatogram were integrated using Agilent Technologies' ChemStation software. The information was exported into a .csv file and converted into a binary matrix (where 1 denotes the presence of a peak and 0 an absence at the specific retention time). The resulting matrix was graphed using white and black horizontal bars to denote a peak presence or absence, respectively. The binary plots only contain analytes attributed to the bacteria and disregard those found in the media alone (i.e. blank subtracted). In addition, PCA plots, generated in R statistical software,

were used to aid in the differentiation of mVOCs. For a more detailed description of PCA, see Specific Aim 1.

Results and Discussion

1. Single Fiber Differentiation of Biothreat Agents

Little is reported in the literature concerning the mVOCs produced by the microorganisms *F. tularensis*, *B. cenocepacia*, and *B. neotomae*. To expand the current knowledge in the field, I sought to ascertain whether the volatiles exuded by these microorganisms are different and thus could be used to detect and differentiate the bacteria. Liquid cultures of rich media (Tryptic Soy Broth (TSB) + 0.1% cysteine) containing *Francisella tularensis spp. novicida*, Strain CG62, *Burkholderia cenocepacia*, Strain LMG 16656, or *Brucella neotomae*, Strain 5K33 were incubated overnight. Aliquots from each culture were acquired, and the mVOCs were extracted via hSPME and subsequently desorbed onto a GC-FID. As depicted in Figure 36, a unique mVOC profile was identified and attributed to each of the different bacterial cultures.

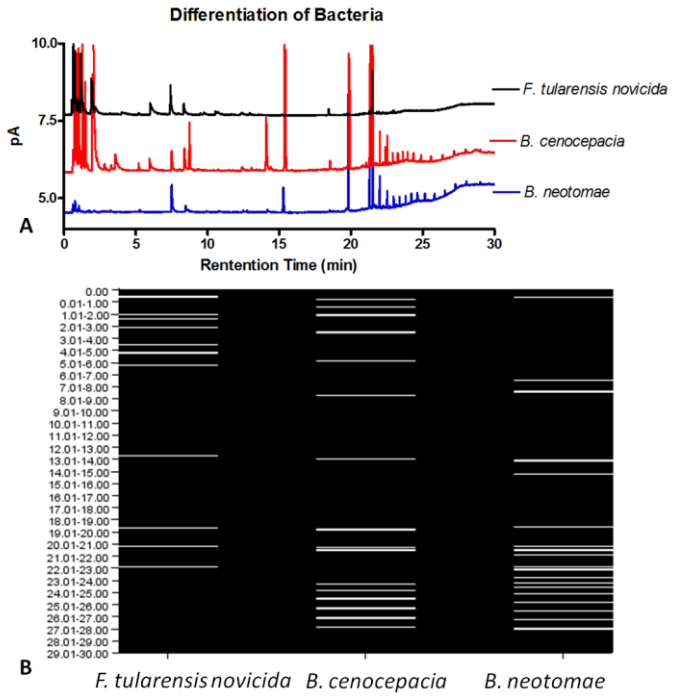


Figure 36. The mVOC fingerprints derived from liquid cultures of *F. tularensis novicida, B. cenocepacia,* and *B. neotomae*.
A single fiber analysis using the CAR/PDMS 75µm fiber was performed. Figure A depicts the resulting GC-FID chromatograms for the three bacteria (*F. tularensis*, *B. cenocepacia*, and *B. neotomae*). Figure B depicts the corresponding binary plots derived from the chromatograms presented in Figure A. The bacteria are identified along the x-axis while on the y-axis is the retention time in minutes. The white bars represent the peak that occurred at that particular retention time. Each fingerprint is blank subtracted (media alone). As illustrated in the binary plot, the mVOC fingerprint clearly differentiates the three bacteria.

To evaluate variability in this analysis, replicate extractions were performed and the resulting mVOC fingerprints were compared. As seen in Figure 37, replicate analyses performed with *F. tularensis novicida* liquid cultures generate identical mVOC fingerprints.

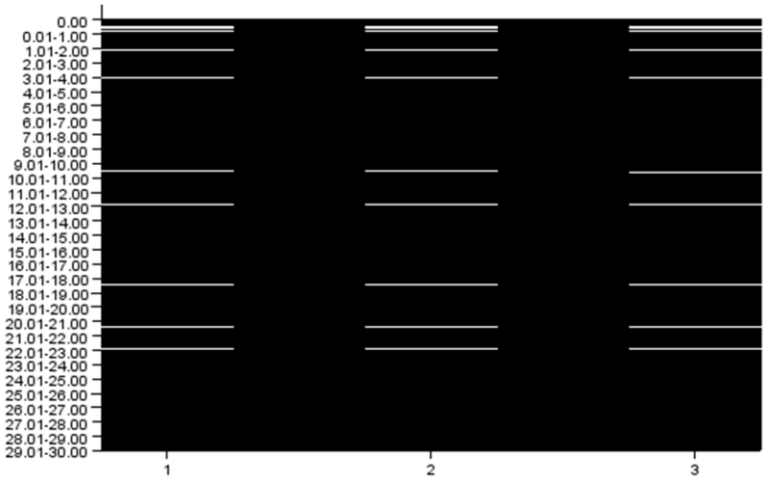


Figure 37. Reproducibility of mVOC analysis.
The three binary plots reflect three replicate extractions performed using three different cultures of *F. tularensis novicida*. The extractions were performed with a CAR/PDMS 85µm fiber. The replicate extractions (1, 2, and 3) are indicated along the x axis, while the y-axis indicates the retention time (minutes). The binary plots derived from each culture are identical.

2. Multifiber hSPME

Although the single CAR/PDMS fiber was successfully used to differentiate *F. tularensis*, *B. cenocepacia*, and *B. neotomae*, a multifiber analysis will extract a diverse array of mVOC metabolites (like dissolves like), further enabling the resolving power of an mVOC analysis. An 8-fiber analysis was next performed for each of the three microbes under investigation. Overnight liquid cultures of bacteria were dispensed into aliquots, and the aliquots were singly extracted via hSPME, each extraction performed using one of the 8 commercially available SPME fibers. The extracted analytes were analyzed via GC-FID, and the resulting chromatograms were converted into binary plots

then compiled into one composite fingerprint (Figure 38). As seen in the Figure, these composite mVOC fingerprints clearly differentiate these bacteria.

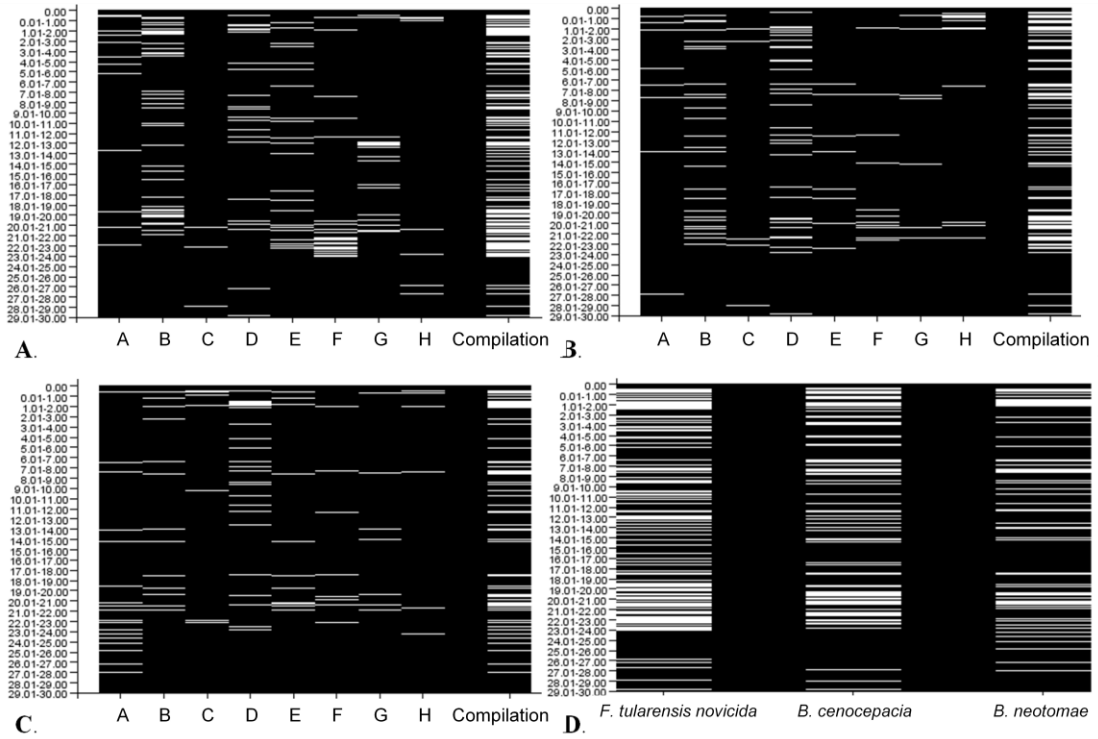


Figure 38. An 8 fiber hSPME analysis of the mVOC metabolome emanating from liquid cultures of *F. tularensis novicida, B. cenocepacia,* and *B. neotomae*.
An 8 fiber hSPME analysis of mVOCs from *F. tularensis novicida* (A), *B. cenocepacia* (B), and *B. neotomae* (C) liquid cultures (TSB + 0.1% cysteine). Binary plots were derived from chromatograms obtained using the following SPME fibers: A: CAR/PDMS 75 µm, B: PDMS/DVB 65 µm, C: PDMS 7 µm, D: DVB/CAR/PDMS 50/30µm Stableflex, E: CAR/PDMS 85 µm Stableflex, F: CW (PEG) 60 µm, G: PDMS 100 µm, H: PA 85 µm. A compilation plot is generated by combining all of the plots into a single binary plot. The compilation plot shown in A, B, and C are shown juxtaposed in panel D, clearly differentiating the three bacteria from one another. In the binary plots, the y-axis is the retention time (min) for the corresponding peaks and the white bars represent the chromatographic peaks.

3. Environmental Effectors

Environmental effectors such as metal salts are well known to influence the growth and metabolism of microbes, often through the regulation of gene expression[90,91]. Hence, I sought to ascertain if select metal salts might influence the derived mVOC fingerprints obtained from the three bacteria under investigation. As shown in Figure 39, these metal salts can have an impact on *F. tularensis novicida*, *B. cenocepacia,* and/or *B. neotomae* growth in liquid cultures. With *F. tularensis novicida,* the addition of either 25 mM $MgCl_2$ or 6.25 mM NaCl has little effect on the overall cell density of an overnight culture, compared to MCDM alone, whereas the addition of 12.5 mM $NiCl_2$ has a drastic influence, as it significantly inhibits the growth of the bacteria. On the other hand, the $MgCl_2$, NaCl, and $NiCl_2$ have little influence on *B. cenocepacia* overnight growth. For *B. neotomae*, the addition of $MgCl_2$ significantly inhibited growth, but not to the extent demonstrated by $NiCl_2$. Somewhat surprisingly, the addition of NaCl to the *B. neotomae* culture slightly increases the cell density of an overnight culture. Collectively, from these growth profiles, conditions were selected that permit the growth of the bacteria and will enable an evaluation of whether or not the added effector will have an influence on the derived mVOC fingerprint.

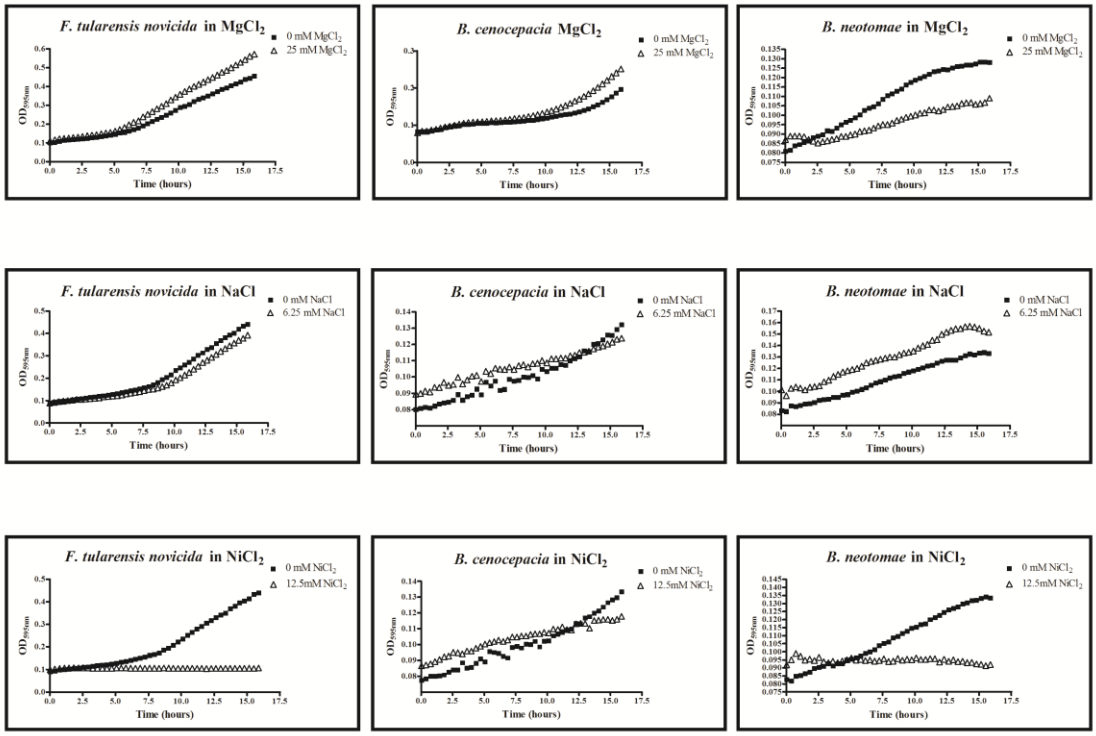


Figure 39. Effectors of Bacterial Growth in Minimal Media.
The indicated bacteria were cultured in minimal liquid media (MCDM) supplemented with various selected effectors (25mM $MgCl_2$, 6.25mM NaCl, and 12.5mM $NiCl_2$). Cultures were monitored for growth (OD_{600}) every 20 minutes for 16 hours at 37°C. See text for further discussion.

To evaluate the effects of the salt addition on the metabolic profile, overnight seed cultures were used to inoculate the amended minimal media, and the bacteria were then incubated in the media with shaking at 37°C for 8 hours. The mVOCs were then extracted from the cultures via multifiber hSPME and subsequently analyzed by GC-FID. The resulting fingerprints are presented in Figure 40.

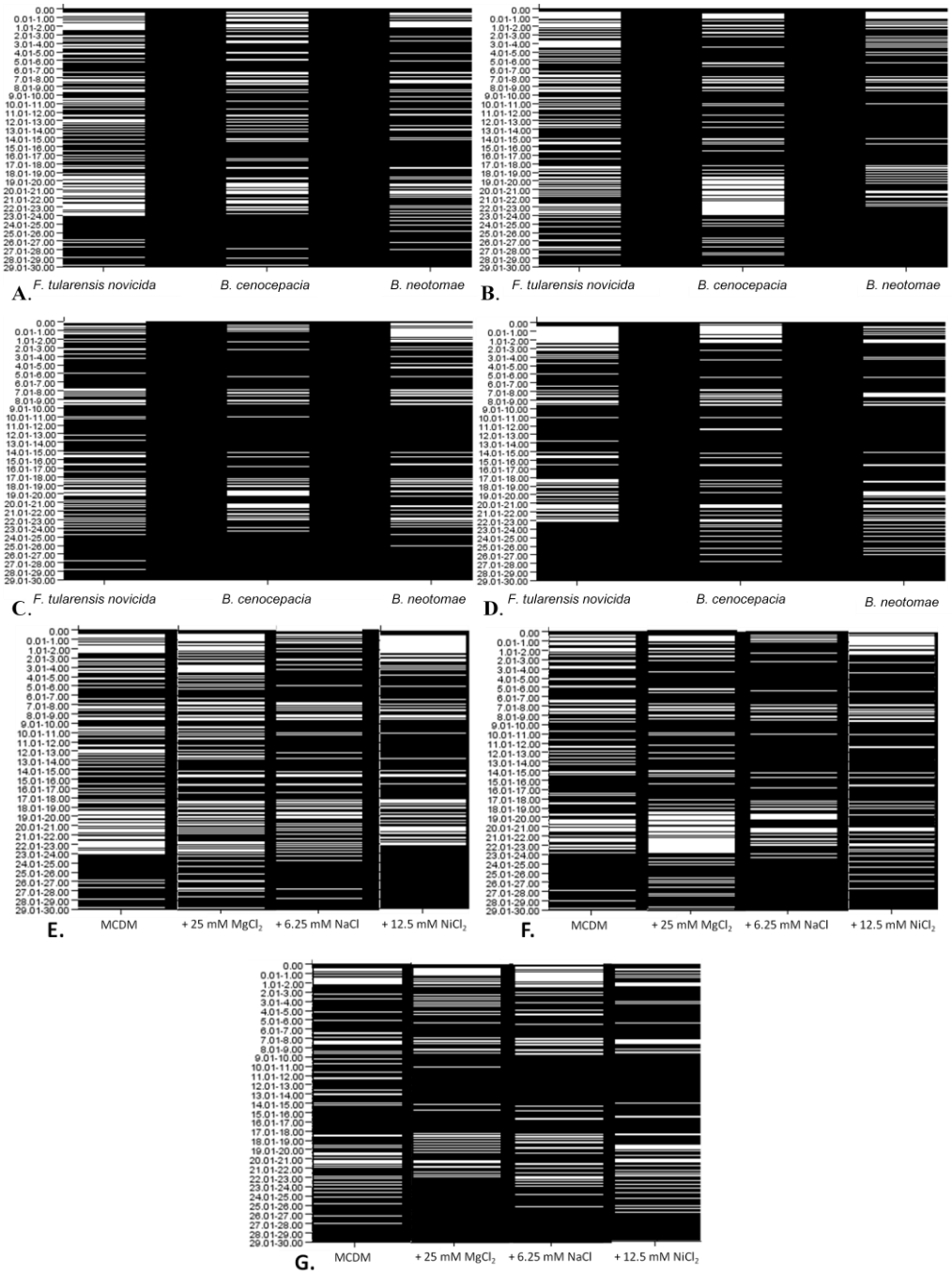


Figure 40. Binary plots exhibiting the mVOC profile obtained for *F. tularensis novicida, B. cenocepacia,* and *B. neotomae* cultured in MCDM with and without the addition of various salts.
For each of the three indicated bacteria under investigation, the corresponding mVOC fingerprint derived from minimal media (MCDM) is shown in panel A, from minimal media supplemented with 25 mM $MgCl_2$ is shown in panel B, from minimal media supplemented with 6.25 mM NaCl is shown in panel C, from minimal media supplemented with 12.5 mM $NiCl_2$ is shown in panel D. In all media compositions tested, the mVOC fingerprints uniquely differentiate the three bacteria. Panels E, F, and G illustrate how the media composition alters the derived mVOC fingerprint for *F. tularensis novicida*, *B. cenocepacia*, and *B. neotomae*, respectively.

As seen in Figure 40, regardless of the media composition, the mVOC fingerprints generated all uniquely differentiate the three bacteria. Also seen in the Figure is the manner in which the mVOC fingerprint changes with media composition. When compared to the MCDM binary plot (Figure 40A), the mVOC metabolic profile obtained for *F. tularensis novicida* exhibited different patterns for each of the salt conditions. This alteration in the microbial fingerprint was also true for *B. cenocepacia,* and *B. neotomae.* In addition, while the growth of *F. tularensis novicida* and *B. neotomae* was significantly inhibited in $NiCl_2$, the bacteria were still alive and exuding volatile metabolites, producing a unique mVOC fingerprint. Overall, while the addition of various salts had a profound effect on the mVOC metabolome, it is still possible to differentiate the microbes in each environmental condition.

4. Differentiation of Wildtype and Antibiotic Resistant Bacteria

Following the success of differentiating the three bacteria (*F. tularensis novicida, B. cenocepacia*, and *B. neotomae*), I next sought to determine if multifiber hSPME can differentiate a kanamycin sensitive (wildtype) and engineered kanamycin resistant strain of *Yersinia pestis*. A 500 µL aliquot of a 10 mL overnight liquid seed culture of each strain (cultured in TSB + 0.1% Cys media, supplemented with 50 µg/mL kanamycin where appropriate) was adjusted with sterile media to an OD_{600} = 1.0 and used to inoculate 25 mL media in a 125 mL shake flask (the Kan^R strain was cultured in media containing 50 µg/mL kanamycin). The flasks were incubated at 250 rpm, 30 ^oC, for 24 hours, then 250 µL aliquots from each culture were dispensed into vials, stored at -80 °C, then analyzed by multifiber hSPME. Each extraction proceeded for 60 minutes, using one

of 6 commercially available SPME fibers (CAR/PDMS 85 µm with stableflex, 100 µm PDMS, PDMS/DVB 65 µm, PA 85 µm, DVB/CAR/PDMS 50/30 µm, PEG 60 µm). The extracted analytes were analyzed via GC-FID, and the resulting chromatograms were converted into binary plots and compiled into a composite fingerprint (Figure 41). As seen in the Figure, the composite mVOC fingerprints and the companion PCA plot clearly differentiate the bacterial strains.

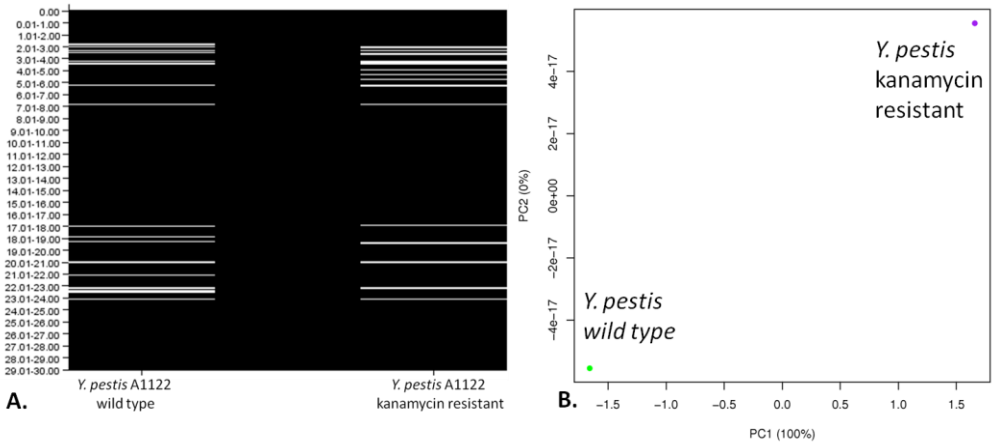


Figure 41. Differentiation of the mVOC metabolome of *Y. pestis* wild type and kanamycin resistant strains.
For each bacterial strain, a 6 fiber analysis was performed using an extraction duration of 60 minutes. Figure A depicts the composite binary plot results from all 6 fibers and demonstrates the differentiation in the extracted mVOC metabolome of the two strains. The samples are indicated along the x-axis, while retention time (min) is indicated along the y-axis. Figure B is a companion PCA plot illustrating the metabolic variations that exist within the mVOC metabolome. Each sphere in the plot reflects the metabolome of the bacterial sample, with *Y. pestis* wild type colored green, and the kanamycin resistant strain of *Y. pestis* colored purple.

To determine if differentiation could still be achieved using a shorter extraction duration, I repeated the analysis using an extraction time of 15 minutes (Figure 42).

Again, the composite mVOC fingerprints and companion PCA plot clearly differentiate the *Y. pestis* wild type and kanamycin resistant strains. While the binary plot for the shorter extraction exhibits a greater number of peaks, this is a reflection of the increase in the pre-incubation temperature from 37°C to 80°C to ensure enough volatiles were present.

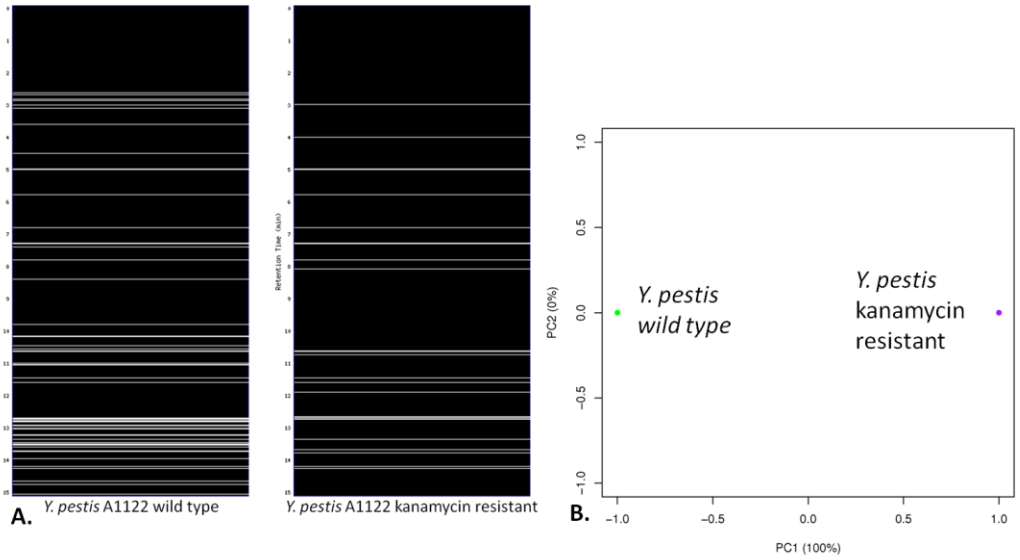


Figure 42. Differentiation of the mVOC metabolome of *Y. pestis* wild type and kanamycin resistant strains using an extraction duration of 15 minutes.
For each bacterial strain, a 6 fiber analysis was performed using an extraction duration of 15 minutes. Figure A depicts the composite binary plot from all 6 fibers and demonstrates the differentiation in the extracted mVOC metabolome of the two strains. The samples are indicated along the x-axis, while retention time (min) is indicated along the y-axis. Figure B is a companion PCA plot illustrating the metabolic variations that exist within the mVOC metabolome. Each sphere in the plot reflects the metabolome of the bacterial sample, with *Y. pestis* wild type colored green and the kanamycin resistant *Y. pestis* strain colored purple. Clearly an extraction duration of 15 minutes is sufficient for differentiating the two bacterial strains.

Since the GC-FID based analysis successfully differentiates the two strains of bacteria, I elected to determine if the GC-MS could do so as well. To this end, I repeated

the procedure using a 60 minute then 15 minute extraction duration. As shown in Figure 43, with either extraction duration, the composite mVOC fingerprints are able to distinguish the two bacterial strains. In addition, the PCA plots demonstrate not only how the two bacterial strains are differentiated, but how they relate to the VOCs associated with the media alone. Therefore, like the GC-FID, mVOC fingerprints generated via the GC-MS can also be used to differentiate these strains of bacteria.

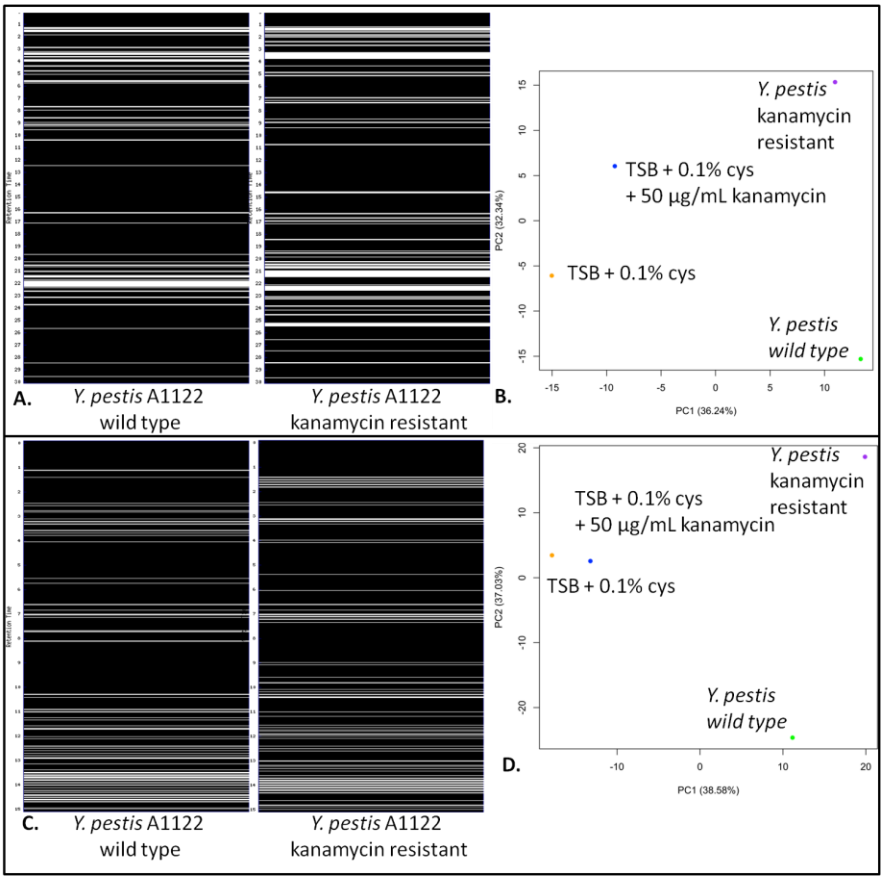


Figure 43. Differentiation of the mVOC metabolome of *Y. pestis* wild type and kanamycin resistant strains using a GC-MS and an extraction duration of either 60 minutes or 15 minutes.
For each bacterial strain, a 6 fiber analysis was performed using an extraction duration of 60 minutes (A and B) or 15 minutes (C or D). Figure A and C depicts the composite binary plots from all 6 fibers and illustrates the differences in the extracted mVOC metabolome of the two strains using extraction durations of either 60 minutes (A) or 15 minutes (C). The samples are indicated along the x-axis, while retention time (min) is indicated along the y-axis. Figure B and D are corresponding PCA plots illustrating the metabolic variations that exist within the mVOC metabolome using an extraction duration of 60 minutes or 15 minutes, respectively. Each sphere in the plot reflects the metabolome of the bacteria or blank media sample, with TSB + 0.1% cys colored orange, TSB + 0.1% cys + 50 µg/mL kanamycin colored blue, *Y. pestis* wild type colored green, and kanamycin resistant *Y. pestis* colored purple. A multifiber hSPME analysis of mVOCs using the GC-MS and an extraction duration of either 60 or 15 minutes readily differentiates the two bacterial strains.

In light of the success with *Y. pestis*, I next performed the analysis with *F. tularensis*. The wildtype strain (NIH B38) and engineered kanamycin resistant strain were each cultured and analyzed as described above for *Y. pestis*. A multifiber hSPME

analysis was performed with the 6 commercially available fibers, a 15 minute extraction duration, and either the GC-FID or GC-MS for analysis (Figure 44).

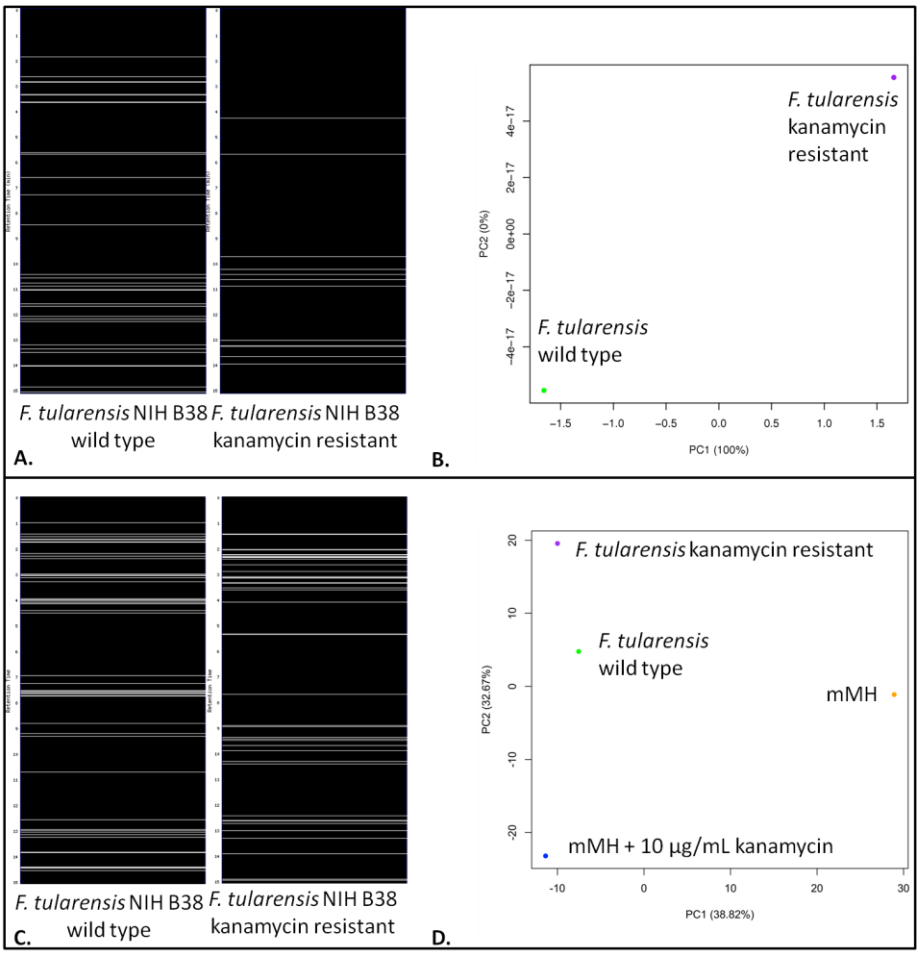


Figure 44. Differentiation of the mVOC metabolome of *F. tularensis* wild type and kanamycin resistant strains using a GC-FID or GC-MS with an extraction duration of 15 minutes.
For each bacterial strain, a 6 fiber analysis was performed using an extraction duration of 15 minutes. Figures A and B depicts the composite binary plots from all 6 fibers and the companion PCA plot using the GC-FID. *F. tularensis* with and without kanamycin resistance are clearly differentiated in each plot. Figures C and D illustrate the composite binary plot from all 6 fibers and the companion PCA plot using the GC-MS. The GC-MS can also differentiate *F. tularensis* with and without kanamycin resistance. For Figures A and C, the samples are indicated along the x-axis, while retention time (min) is indicated along the y-axis. Each sphere in the PCA plot reflects the metabolome of the bacteria or a blank media sample, with mMH colored orange, mMH+ 10 µg/mL kanamycin colored blue, *F. tularensis* wild type colored green, and kanamycin resistant *F. tularensis* colored purple.

Concomitant with performing the above multifiber analysis, the Couch lab developed the simulti-hSPME device (Figure 31). To determine if the device could be used to successfully differentiate the wildtype and kanamycin resistant strains of *Y. pestis*, a 6 fiber simulti-hSPME was performed with the GC-MS, using a 30 min mVOC extraction duration. As shown in Figure 45, the resulting PCA illustrates that the two strains are well resolved from each other, based upon their mVOC composition.

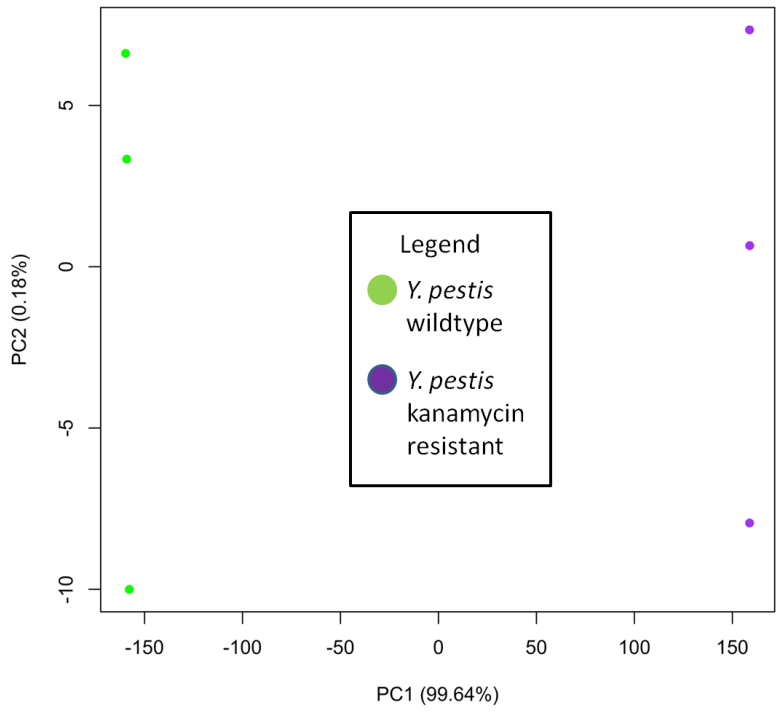


Figure 45. PCA plot demonstrating the difference in mVOC metabolome derived from liquid cultures of wild type and kanamycin resistant strains of *Y. pestis*.
For each bacterial strain, a 6 fiber analysis was performed in the simulti-hSPME device using an extraction duration of 30 minutes. The segregation of the samples along PC1 reflects the metabolic variation between the mVOC metabolomes for the two strains. The analysis was performed with three different aliquots of each bacterial culture. Each sphere in the plot reflects the metabolome of the bacterial sample, with *Y. pestis* wild type colored green and kanamycin resistant *Y. pestis* colored purple. The contribution of PC2 to variation among the samples is small (0.18%), yet accounts for the difference among the replicate runs for each strain.

Conclusions and Future Applications

By utilizing hSPME, a VOC metabolomic fingerprint was developed for liquid cultures of the bacteria *F. tularensis novicida*, *B. cenocepacia*, and *B. neotomae*. While I was able to successfully differentiate between the three different bacteria, I also determined that the environment/media in which the bacteria are cultured has a significant impact on the associated mVOC metabolome. In addition, I was able to successfully differentiate wild type and kanamycin resistant strains of both *Y. pestis* and *F. tularensis*. By incorporating the simulti-hSPME device, the overall analysis time was reduced (due to the ability to perform 6 extractions simultaneously) while the bacterial strains were still readily differentiated.

One future application of this work is the detection of infectious disease in exhaled breath. There are several reports in the literature on performing breath analysis[29,30,92–95]. In general, there are two distinct ways to use SPME to preconcentrate and extract breath-derived VOCs: namely, active and passive approaches[96]. In an active approach, the extraction is performed as the sample is being collected (i.e. in-line); while in a passive approach, the sample is collected in a bag or a vial, then the exhaled metabolites are extracted (Figure 46). It is noteworthy that the Couch lab has begun preliminary trials using both of these approaches.

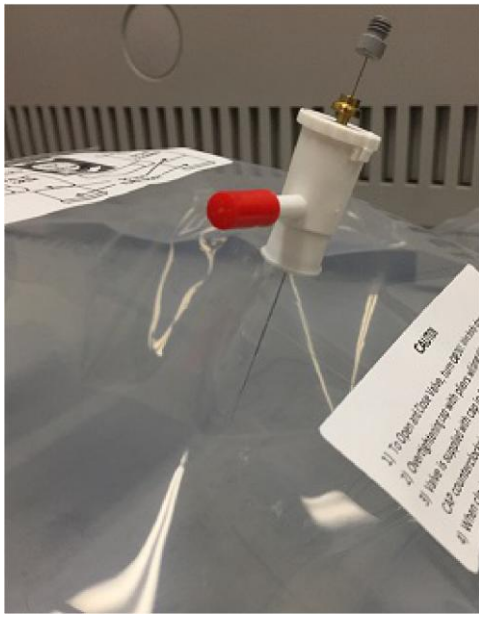


Figure 46. Passive approach to SPME using breath collection bags.
The patient provides a breath sample into the bag. Following collection, the bag is sealed and extraction via a SPME fiber can be performed.

An Active Approach to Breath Analysis

Along these lines, I took an initial foray into an active approach to breath analysis. While not a defined objective of Specific Aim 2.1, I report here my preliminary results.

To address the feasibility of performing a simulti-hSPME fiber extraction during breath collection, I first used a nose-only exposure chamber to extract the breath-derived VOCs produced by healthy and influenza H3N2 infected mice (Figure 47).

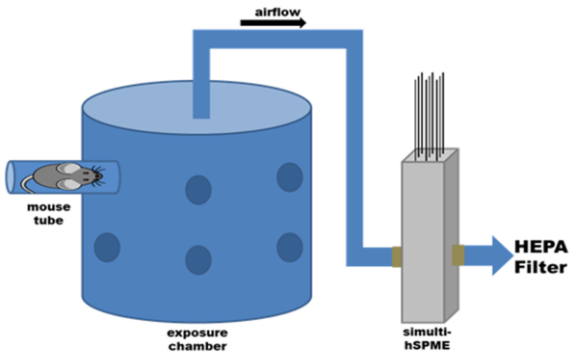


Figure 47. Schematic illustrating the active extraction process of breath VOCs.
In this approach, the mice were restrained in a funnel which was connected to the exposure chamber. Air is supplied to the mice and the mixture of supplied air and exhaled breath travels through the chamber and into the simulti-hSPME device for metabolite extraction via the SPME fibers. The air then travel out through a HEPA filter and into the environment.

In this chamber, the mice are restrained in a plastic funnel with their nose exposed to the chamber. Air is supplied through the chamber to the mice as they breathe normally. The supplied air and breath travel through the chamber and are directed into the simulti-hSPME extraction device containing the SPME fibers for extraction. Due to the use of live animals and the breath collection procedure, I used an extraction duration of 30 minutes (and no longer). While one trial demonstrated differentiation amongst the two cohorts (Figure 48), the second attempt did not. While the purpose of my analysis was to provide proof of principle (which it did), further investigation is clearly required to elucidate the cause of this variation.

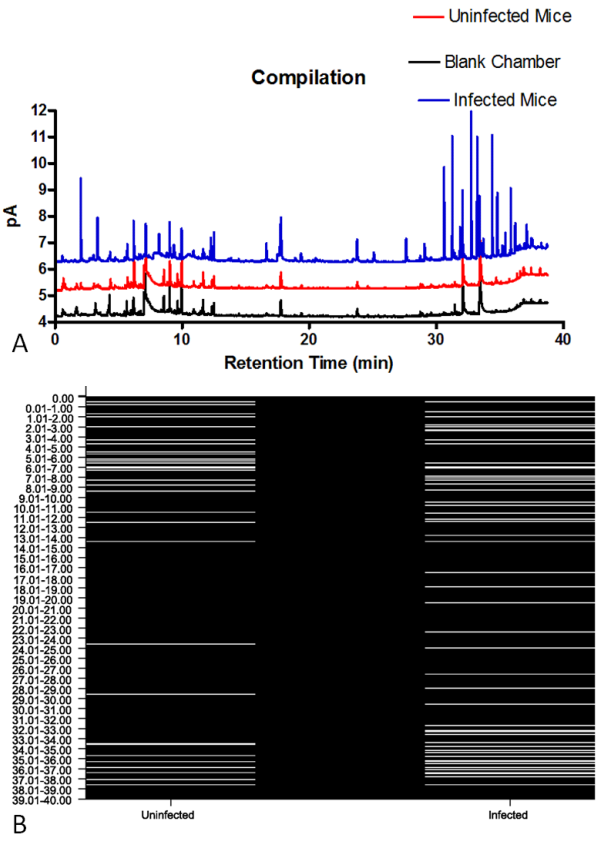


Figure 48. Breath Analysis of Mice Infected with the Flu Virus.
The chromatogram obtained in Figure A demonstrates the differences obtained between infected mice, uninfected mice, and the blank device. The binary plot in B indicates how the infected mice are differentiated from the healthy mice.

SPECIFIC AIM 2.2: ALTERATIONS TO THE HUMAN FECAL METABOLOME DUE TO ALCOHOL CONSUMPTION.

Objective Use metabolomics as an electronic nose to explore the VOC profile of fecal material, collected at home or via an endoscopy procedure, to potentially identify biomarkers of chronic alcohol consumption.

This research performed in this specific aim resulted in a co-authorship in PLOS One[32].

Paper I: Alcohol Induced Alterations to the Human Fecal VOC Metabolome
Couch RD, **Dailey A**, Zaidi F, Navarro K, Forsyth CB, et al. PLoS One. 2015; 10 (3):e0119362. doi: 10.1371/journal.pone.0119362

Synopsis

A comparative analysis of healthy versus alcoholic participants demonstrated the profound affect alcohol has on the fecal VOC metabolome. Irrespective of the collection method (home or endoscopy) we identified numerous metabolic alterations within the alcoholic cohort. Some of the most notable alterations include an increase in oxidative stress, decrease in short chain fatty acids responsible for maintaining epithelial health, and decrease in the attenuation of hepatic steatosis.

Attributions and Contributions

Reproduced with permission from the Public Library of Science and per the Creative Commons License agreement[97]. I was responsible for performing the experiments, analyzing the data, and generating Figures 3 through 10 and Table 2 in the

manuscript. The published manuscript and all supplementary information is presented in Appendix 5 and 6 (per Mason guidelines, embedded papers must be placed in the Appendix of the thesis).

SPECIFIC AIM 2.3: TISSUE-RELATED EFFECTS OF A HIGH FAT DIET AND PROBIOTIC SUPPLEMENTATION IN PIGS

Objective: Use metabolomics to evaluate if a high fat diet and/or probiotic supplementation have an effect on the pig tissue metabolome.

Introduction

Probiotics are live microorganisms which resemble/reflect those found in the natural gut flora[98,99]. Probiotics, particularly from the genera *Bifidobacterium* and *Lactobacillus*, have been associated with a number of health effects including competitive exclusion of food pathogens, stimulation of immune function, lowering gas distension, aiding in food digestion and adsorption, synthesis of vitamins, lowering of cholesterol levels,[100–106] and may aid in the curing of gastrointestinal diseases and autoimmune disorders[98,107–110]. While few studies have yet addressed the potential of probiotics for the management of obesity, a high-fat diet is known to induce diabetes and endotoxemia in mice, and negatively correlates with the level of *Bifidobacterium spp*. in the mouse intestine[111]. By supplementing these mice with the prebiotic oligofructose, enhanced growth of intestinal *Bifidobacterium* occurs with accompanying normalization of the inflammatory state (decreased endotoxemia, decreased plasma and adipose tissue pro-inflammatory cytokines), suggesting that *Bifidobacterium* in the gut microbiota may prevent the deleterious effects of high-fat diet-induced metabolic disease. Furthermore, a

selected strain of *L. rhamnossus* has also been reported to protect mice from diet-induced obesity, likely due to the production of conjugated linoleic acid by the bacteria[112]. Even with the increasing accumulation of clinical data on the benefits of pre- and probiotics, the mechanism of modulation still remains largely uncharacterized, especially with respect to the host metabolome. Hence, to further explore the effect of a high fat diet and the consequences of probiotic supplementation, in collaboration with Drs. Gloria Solano-Aguilar and Joe Urban at the United States Department of Agriculture (USDA) and Dr. Kati Hanhineva at the University of Eastern Finland, Specific Aim 2.3 examines pig tissue metabolomes through non-targeted metabolomic profiling.

Given their anatomical and metabolic similarities to humans, Ossabaw pigs (Figure 49) were selected for use as a model organism in this study.

Figure 49. Ossabaw pigs are the model organism used in this probiotic supplementation study.
Ossabaw pigs serve as models of juvenile obesity[106].

In a previous metabolomics investigation, it was shown that a high fat diet alters the metabolomic composition of select Ossabaw pig tissues, relative to tissues in pigs fed a nutritionally balanced diet[113]. In light of this, my investigation sought to determine if dietary supplementation with probiotics also has a measurable effect on the host tissue metabolism. To accomplish this, I specifically compared metabolomic profiles from select organ tissues obtained from pigs fed either a nutritionally balanced (basal) or obesogenic (high fat) diet, with or without probiotic supplementation (Figure 50).

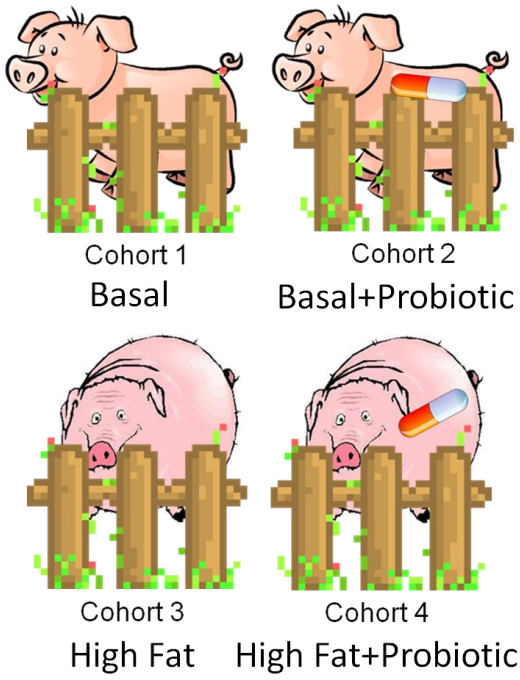


Figure 50. Schematic depicting the four distinct dietary cohorts of pigs used in this investigation.
For this study, twenty pigs were equally distributed among the four indicated cohorts (5 pigs in each cohort). The first two cohorts contain pigs fed a basal diet, with pigs in cohort 2 also receiving a probiotic supplement (the supplement is depicted as an orange and white colored capsule in the illustration). The second two cohorts (3 & 4) contain pigs fed a high fat diet, with pigs in cohort 4 receiving a probiotic supplement.

Twenty pigs were equally distributed among four cohorts (basal, basal+probiotic, high fat, and high fat+probiotic), and after 38 weeks of a cohort-specific diet, the pigs were euthanized, tissue samples were collected, snap frozen and stored at -80 ^{o}C, and a non-targeted LC−QToF based metabolite profiling of cortex, heart, kidney, liver, muscle, and pancreas tissue was performed on every member of each cohort.

Tissue Sample Processing

To extract the metabolites from the frozen tissue samples, the samples are homogenized while frozen (cryo-ground using a bead beater) then extracted using a 1:1 methanol:water solution (Figure 51). The tissue extracts are clarified by centrifugation and subsequently filtered to remove any residual particulates. The filtered samples are then analyzed by LC-QToF using both ESI positive and ESI negative modes, with a reversed phase (C18) or HILIC (Si) column (i.e. 4 MS chromatograms per sample are obtained (ESI pos+C18, ESI neg+C18, ESI pos+HILIC, ESI neg+HILIC), for a total of 480 chromatograms generated for the six tissue types (not including blanks, quality control samples, and MS/MS chromatograms)). By using the two ionization modes (ESI positive and negative) in conjunction with two types of chromatography columns (reversed phase and HILIC), the number of metabolites identified within the extract increases (relative to a single ionization mode with a single column). The specific details used to perform this investigation are described in the **Materials and Methods** section, below.

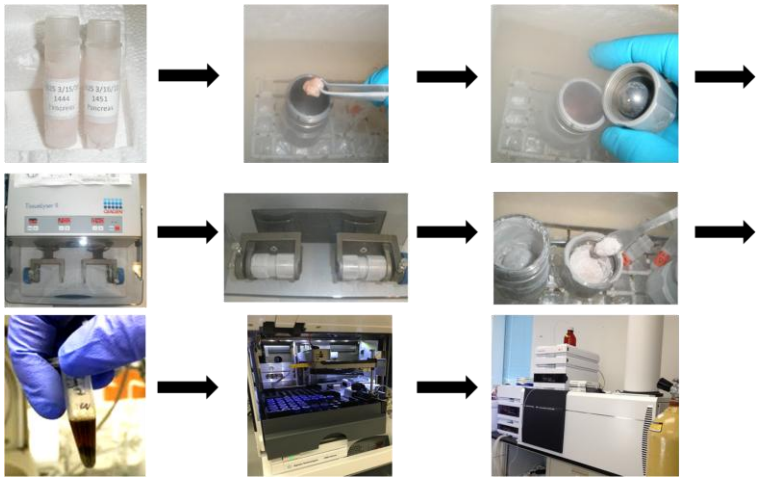


Figure 51. Tissue Sample Processing for Metabolome Determination
The frozen tissues are homogenized via a TissueLyser (bead mill). The stainless steel cylinder and ball bearing used to grind the samples are housed in liquid nitrogen to prevent thawing of the sample while grinding. Once powderized, the sample is dispensed into a microcentrifuge tube (100 mg sample), and extracted with a 1:1 methanol:water solution. The extract is clarified by centrifugation, filtered, then placed into an autosampler vial at 4 $^{\circ}$C and analyzed by the LC-QToF.

Materials and Methods

For Specific Aim 2.3, the pigs were housed and the tissue samples collected at the USDA. The tissues were processed through the combined effort of Dr. Robin Couch and the laboratory of Dr. Kati Hanhineva (University of Eastern Finland). I assisted with performing select MS and MS/MS evaluations using the LC-QToF, and I exclusively performed the data analysis as described in ***Data Processing, Chemometrics and Statistical Analysis***, below**.**

1. Maintenance of Animals and Experimentation

Ossabaw pigs, born at the Indiana University Ossabaw production Unit, were transported in kennels and delivered overnight to the USDA-Beltsville Maryland animal

facility according to standardized procedures of quarantine and under the approval of Beltsville Area Animal Care and Use Committee (protocol number 12-02). After arrival, the pigs were housed in an isolated building with individual pens with free access to water and fed a standard mini pig grower diet (5L80 Purina TestDiet, Inc, Richmond, IN) specifically designed for miniature swine (18.5% kcal from protein, 71.0% carbohydrates and 10.5% from fat)[114]. After acclimatization of pigs, all eight week old pigs were randomized by weight and split into four treatment groups. Pigs in cohorts I (n=5) and II (n=5) continued eating the mini pig grower diet 5L80 (CONV) with gradual bi-weekly step increases from 750 to 3600 kcal/day to adjust for nutrient requirements for growth during the 24 weeks of the study. A standard high fat, obesogenic pig diet was given to the other two treatment groups (cohort III (n=5) and IV (n=5)). This diet was prepared at the Beltsville feed mill by mixing a commercial diet (5KA6 Purina TestDiet, Inc, Richmond, IN; composed of ground corn, soybean meal, wheat, and alfalfa) with 17% hydrogenated soybean oil containing 56% trans fatty acids (#170, Columbus Foods, Chicago, IL), 2.4% corn oil, 1% cholesterol, 0.7% cholic acid, and recommended levels of minerals and vitamins for swine. This mixture yields 13.0% of total kcal from protein, 57% kcal from carbohydrates, and 30% kcal from fat. The pigs on the high fat diet received bi-weekly step increases from 2000 to 4500 kcal/day for the duration of the study. Daily food rations were pre-weighed and consumption was individually monitored and recorded daily. Starting at week 8, all dietary groups were supplemented daily with an oral gavage of either the probiotic bacteria (*Lactobacillus paracasei* (W8); 1 x 10^{10} cfu/day) (Ch. Hansen, Denmark) (cohorts II and IV) or an equivalent volume of a

probiotic-free vehicle (placebo) dissolved in 5ml of phosphate buffered saline (PBS) solution (cohorts I and III). Aliquots of lyophilized probiotic provided by the manufacturer (Ch. Hansen) were tested weekly to verify dose and viability. At the completion of the study, each pig was euthanized using the standard approach (500 mg ketamine (Ketaset, Fort Dodge Animal Health, Iowa), 80 mg tiletamine (Telazol, Fort Dodge Animal Health, Iowa), 80 mg zolazepam (Telazol) and 333 mg xylazine (Xyla-Ject, Phoenix Pharmaceutical, St Joseph, MO) per 100 kg body weight) and the tissue samples were collected, immediately snap frozen in liquid nitrogen, and stored at -80 ^{o}C. All animal experiments and procedures were conducted in accordance with guidelines established and approved by the Beltsville Area Animal Care and Use Committee.

2. Sample Preparation for Metabolite Profiling

While using liquid nitrogen to keep the tissue samples frozen, the samples were ground into fine powder (TissueLyser II, Qiagen, Germantown, MD) then weighed into pre-chilled microcentrifuge tubes (~100 mg aliquots). The powdered tissues were then extracted using a methanol:water solution (1:1), added in a ratio of 3 μL of solvent/mg frozen tissue. The sample was vortexed, sonicated at room temperature for 5 min, clarified by centrifugation (15 min at 16.1 rcf), and the supernatant was collected and stored on ice until analyzed.

3. LC−QToF Analysis

Samples were analyzed via an Agilent 1290 Infinity UPLC coupled to an Agilent 6450 Accurate Mass QToF running in ESI positive and negative modes. Analyte

separation was performed using two different chromatographic systems in order to gain wide coverage of metabolites (Amide HILIC column (BEH, Waters) and a C18 reversed phase column). A quality control (QC) sample was prepared by pooling a small aliquot of each sample to represent the total metabolite composition present in the analysis. This QC sample was injected after every 12 randomized analytical samples to serve as to monitor for potential problems in the chromatography and/or in ion response.

4. Data Processing, Chemometrics and Statistical Analysis

Molecular features were identified in the raw chromatograms using Agilent Technologies' MassHunter Qualitative Analysis software (ver B.06.00). The molecular features and their relative abundance (peak height) were then tabulated using Agilent Technologies' Mass Profiler Professional software (ver 12.6), resulting in a metabolomics data set containing the molecular features identified by HILIC ESI positive, HILIC ESI negative, reverse phase ESI positive, and reverse phase ESI negative modes. Metabolites present in ≤19% of the total number of samples processed were treated as one-offs and were removed from the matrix. The score value for each metabolite was calculated, and the data was filtered to include only those which were present in at least 4 of 5 pigs for any 1 cohort (Equation 1). The samples in the metabolite matrix were organized by their appropriate cohort (a nutritionally balanced basal cohort, a basal cohort supplemented with probiotics, a high fat obesogenic cohort, and a high fat cohort supplemented with probiotics) and the outlier peak area values were identified in each cohort using an analysis of (mean-median)/median for each analyte and a cutoff value ≥1.5 (Equation 2). Outliers were replaced with the median value for that metabolite within the cohort.

Metabolite peak height values were then standardized across the two cohorts by conversion to Z-scores (Equation 3). A principal component analysis was then performed using the standardized metabolite matrices and the R statistical package. XLSTAT was used to perform two-sample T tests between cohorts for each metabolite and Benjamini-Hochberg critical values were calculated (Equation 6) to adjust for the false discovery rate. Pearson's correlation coefficients were calculated using Microsoft Excel and a correlation network was created using the R statistical package. Fold change calculations were performed using Microsoft Excel. Finally, for the molecular features of interest, Extracted Ion Chromatograms (EIC) were generated using Agilent's Mass Hunter Qualitative Analysis software. Features were deemed significant when they appear as a single symmetrical peak at the expected retention time (see Appendix 7). Only significant molecular features were considered in the assessment of the effects of probiotic supplementation.

Results and Discussion

1. Data Preparation

A total of 480 chromatograms were generated (via LC-QToF) from the isolated pig tissue samples. Molecular features were identified in the chromatograms using Agilent Technologies' MassHunter Qualitative Analysis software (see ***Specific Aim 1: Stage 1 of the Metabolomics Pipeline*** for more information). Following the identification of molecular features, an all-inclusive molecular feature data set was then compiled. This data set comprises all of the tissue metabolomes from each member of the four cohorts

and contains a grand total of 70,957 molecular features. To filter out low frequency one-off metabolites, I restricted the analysis to include only those features detected in ≥20% of the tissue-specific samples profiled (i.e. for each tissue type, metabolites appearing in ≥4 of 20 pigs were retained in the data set, regardless of their distribution among the four dietary cohorts, while metabolites detected in ≤3 of the 20 pigs were removed from the data set). Subsequently, I further refined the data set by retaining and analyzing only those tissue-specific molecular features that appeared in a minimum of 3 of 5 pigs in any one of the four dietary cohorts, as determined by my score value calculation (see Equation 1 in ***Specific Aim 1: Stage 2 of the metabolomics pipeline***). As a consequence of this data filtration process, the data set was constrained to include between 2600-6000 molecular features per tissue type, with a combined total of 9788 molecular features.

2. Multivariate Statistical Analysis and Data Visualization

Principal component analysis (PCA) of the 9788 molecular features clearly differentiates each of the pig organs from one another, underscoring the tissue-specific makeup of the metabolome composition (Figure 52).

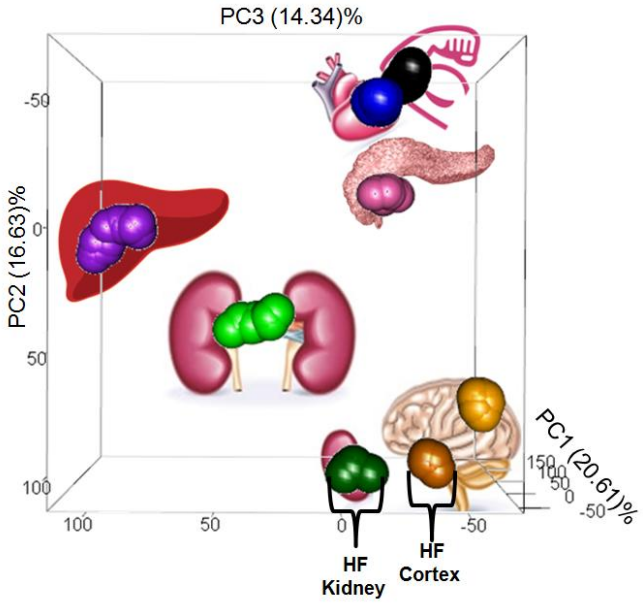


Figure 52. PCA plot of all derived pig tissues metabolomes.
A total of 20 pigs were distributed equally into 4 distinct dietary cohorts (a nutritionally balanced basal cohort, a high fat obesogenic cohort, a basal cohort supplemented with probiotics, and a high fat cohort supplemented with probiotics). After maintaining their diet for 38 weeks, the pigs were euthanized and the metabolomic makeup of the indicated organ tissues were derived and compared. The 3D PCA plot distinguishes the pig tissue metabolomes from each other (that is, the samples distinctly aggregate and segregate by tissue type in the plot). Each sphere in the plot reflects the metabolome of the pig tissue sample, with liver samples colored purple, kidney samples green, heart muscle samples blue, skeletal muscle black, pancreas pink, and cortex orange. For emphasis, the kidney samples obtained from the high fat pig cohort are colored dark green and the high fat cortex samples are colored dark orange (they are also denoted HF in the plot). This PCA plot demonstrates that a high fat diet particularly alters the composition of the kidney and cortex tissue metabolomes such that these samples are distinct from the other kidney and cortex dietary cohorts (i.e. for these tissues, the high fat samples are uniquely clustered away from the basal, basal+probiotic, and high fat+probiotic cohort samples).

Not surprisingly, the heart and skeletal muscles appear closely juxtaposed in the PCA plot, a reflection of the greater similarity in their overall metabolome composition in comparison to the other tissues. This is further apparent in Figure 53, where the heart and muscle tissues organize as sister clades in a dendrogram produced from the tissue metabolomes.

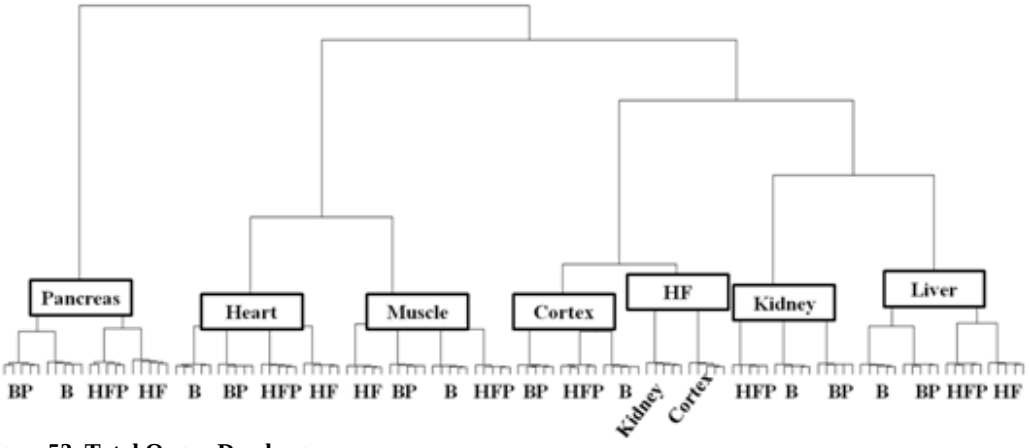


Figure 53. Total Organ Dendrogram.
The tissue samples are identified in the figure, labeled to reflect the basal (B), high fat (HF), basal+probiotic (BP), and high fat+probiotic (HFP) cohorts. Note that without exception, the 5 samples from each dietary cohort cluster together by tissue type. While the pancreas uniquely separates into its own clade in the dendrogram, differentiating it from the other tissue types, the heart and muscle tissues organize into sister clades, based upon the similarities in their metabolomes. It is also noteworthy that the high fat cortex and high fat kidney samples cluster together, whereas the remainder of the cortex and kidney samples are within their own corresponding clades.

Interestingly, both the PCA plot and dendrogram suggest that the metabolomic composition of the kidney and cortex tissues are particularly sensitive to a high fat diet. These tissues obtained from the high fat pig cohort appear distinctly clustered away from those same tissues obtained from the cohorts of pigs fed either a basal, basal+probiotic, or high fat+probiotic diet. It is also noteworthy that probiotic supplementation to the high fat diet appears to alter the high fat kidney and cortex tissue metabolomes. Here, the high fat+probiotic samples cluster tightly with the basal and basal+probiotic samples and differentiate from high fat in both the dendrogram and PCA plot.

To gain higher resolution in comparing the metabolomes, focused tissue-specific PCAs were performed. The resulting plots illustrate how the metabolomes derived from

each of the four pig dietary cohorts differ from one another, within the context of each tissue type (Figure 54).

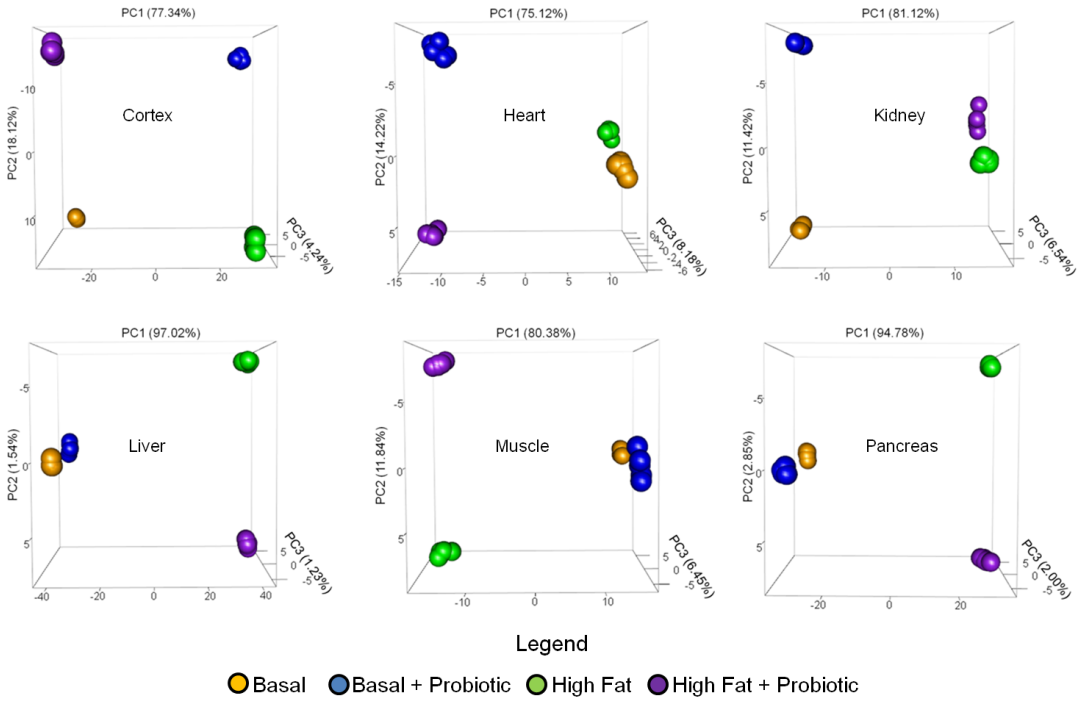


Figure 54.3D PCA plots demonstrating the diet derived metabolome variations in individual tissues. Three dimensional PCA plots of metabolome composition derived from each of the indicated tissue types clearly differentiate the four pig dietary cohorts from each other, demonstrating that dietary consumption (including probiotic supplementation) can have a significant influence on the metabolite composition of the organ tissues. Each sphere in the plot reflects the metabolome of a pig tissue sample. Spheres are colored by cohort, as indicated in the figure legend. The analysis was confined to molecular features appearing in ≥20% of the total number of pigs and a minimum of 3 of 5 pigs in any one of the four pig cohorts, as per Figure 52.

Regardless of tissue type, the samples distinctly cluster and segregate into basal, basal+probiotic, high fat, and high fat+probiotic groups in the PCA plots, a reflection of their unique diet-associated metabolome composition. Thus, dietary consumption has a

significant effect on the metabolite composition of the organ tissues. For example, the cortex metabolome is altered when pigs are fed an obesogenic high fat diet rather than a nutritionally balanced basal diet (i.e. the metabolomes derived from the 5 pigs fed a basal diet are clustered in the cortex PCA plot and clearly segregate from the clustered metabolomes of the 5 pigs in the high fat cohort; Figure 54). Similarly, high fat diet associated alterations in the metabolome are also observed among the other pig tissues analyzed, emphasizing the effect of a high fat diet on the organs of the body. In agreement with previous findings[113], metabolites such as phosphatidylcholines and lysophosphatidylcholines contribute to the difference between the cohorts.

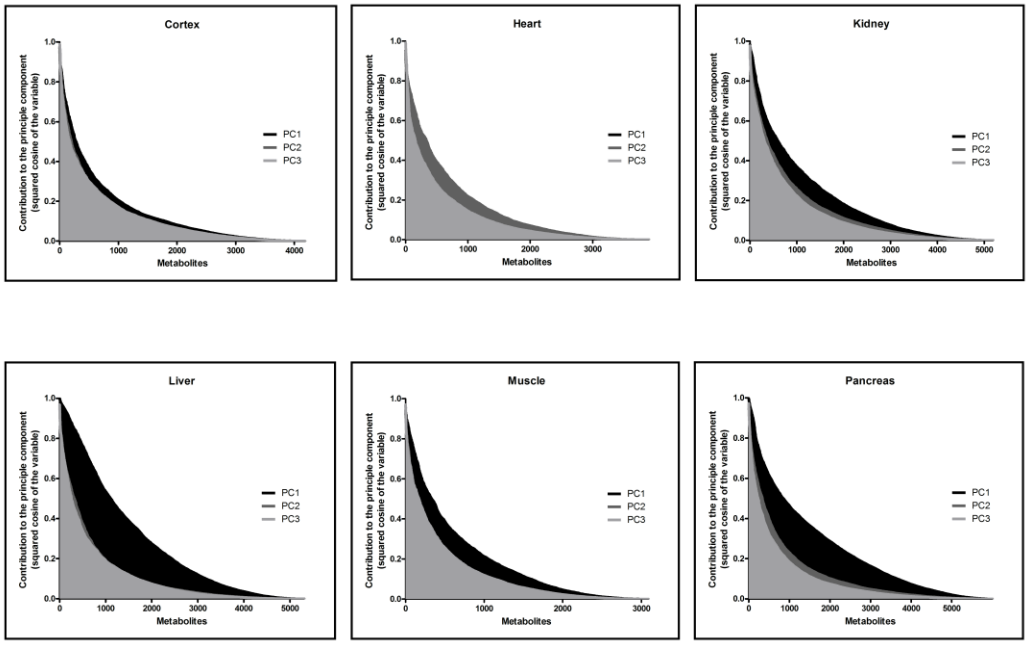


Figure 55. Number of metabolites contributing to the variation among the four tissue cohorts.
Each graph is restricted to the first three principal components and metabolites are arranged by descending contribution to the principal component (squared cosine of the variable value). The plots indicate that for each organ analyzed, numerous metabolites collectively contribute to cohort segregation. It is noteworthy that for PC1 more metabolites contribute to the variation observed in the liver and pancreas, relative to the other organs.

Although numerous metabolites collectively contribute to cohort segregation (Figure 55), the top 30 contributors are sufficient to isolate the four cohorts in a PCA plot of each tissue type (Figure 56).

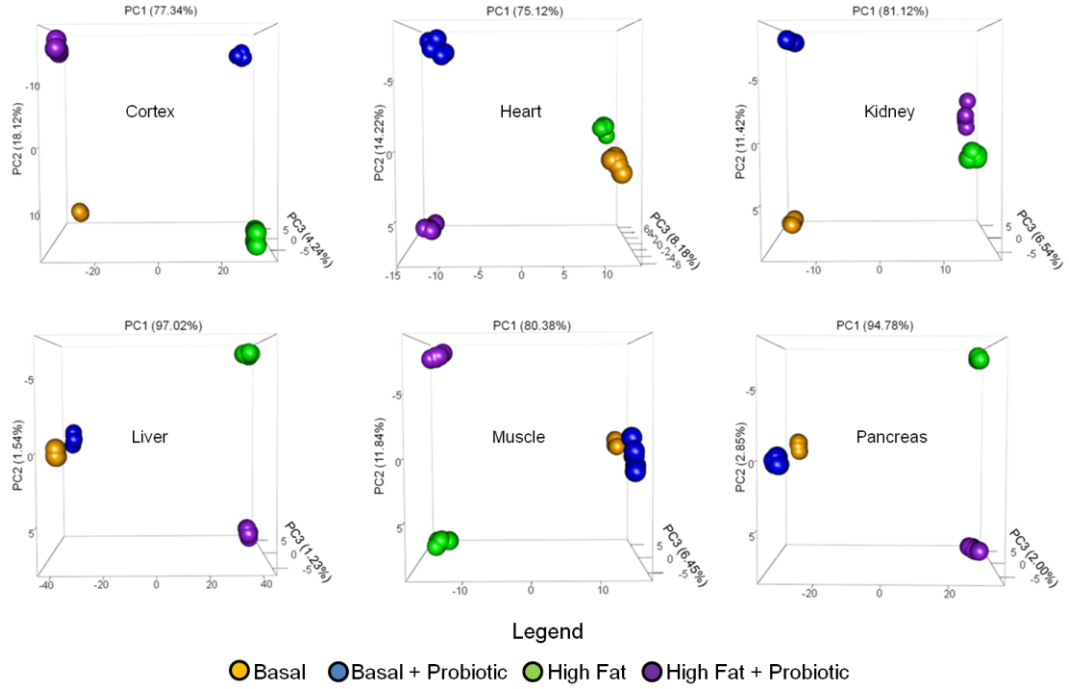


Figure 56. Tissue-specific PCA plots based on the top scoring metabolites.
The top ten metabolites were determined by the weight loadings (squared cosines of the variable) for each of PC1, PC2, and PC3 (derived from Figure 54), which represent the three largest contributions to variation among the samples, to identify the top 30 metabolites differentiating the cohorts within each tissue type. The PCA plot shows that by using these top 30 metabolites, the four cohorts are still distinguishable. The analysis was confined to molecular features appearing in ≥20% of the total number of pigs and a minimum of 3 of 5 pigs in any one of the four pig cohorts, as per

The top 30 metabolites are listed in Table 2, organized by principle component and tissue type.

Table 2. Top 30 metabolites for each tissue type contributing the most to cohort segregation in the PCA plot.

Top Contributors to PC1				
Cortex				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
N-lactoyl-phenylalanine	0	1	1	3
6-Succinoaminopurine	1	3	0	1
3-(4-Hydroxy-3-methoxyphenyl)-2-methyllactic acid	0	1	3	1
Malyngamide I	0	0	1	3
Prostaglandin A1-biotin	0	0	3	1
(R)-2-Hydroxybutane-1,2,4-tricarboxylate	2	3	1	1
Dihydroferuloylglycine	0	0	1	4
Molecular feature_110.01	3	1	0	1
Molecular feature_390.30	0	0	3	1
Tumonoic Acid F	0	0	2	4
Heart				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
Quinapril hydrochloride	1	1	3	0
Molecular feature_365.14	1	1	3	0
Molecular feature_584.36	4	3	1	0
Molecular feature_549.17	1	3	1	0
Malonyl CoA	5	1	2	0
(R)-Byakangelicin 2'-glucoside	1	0	0	3
Molecular feature_158.03	1	3	1	0
PC(2:0/18:1(9Z))	2	4	2	0
PG(19:0/22:2(13Z,16Z))	3	1	1	4
Palmitoylglycerone phosphate	4	4	3	0
Kidney				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
Molecular feature_564.33	1	1	4	1
3,3-Difluoro-5alpha-androstan-17beta-yl acetate	3	2	0	0
Molecular feature_564.35	3	3	1	1
Molecular feature_870.26	3	1	0	0
Succinic acid	3	2	1	0
Cyclocommunin	0	0	2	3
6-Hydroxymelatonin glucuronide	0	0	2	3
Phenylamil	0	0	2	4
Molecular feature_1151.66	0	0	4	3
Molecular feature_432.42	3	2	0	0
Liver				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**

8-Epidiosbulbin E acetate	1	3	0	0
Triglochinin	1	3	0	0
Indole-3-acetic-acid-O-glucuronide	0	0	3	1
IPSP	3	1	0	0
FMNH2	0	0	3	1
(±)Fenfluramine	0	0	1	3
Molecular feature_192.93	3	1	0	0
Molecular feature_702.31	0	0	1	3
Deserpidine	0	0	5	5
Molecular feature_1035.67	0	0	1	4
Muscle				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
Molecular feature_2470.57	0	0	2	4
GDP-glucose	0	0	2	3
Molecular feature_360.13	0	0	1	3
Molecular feature_617.15	4	1	0	1
Molecular feature_315.03	0	0	3	1
Luteolin 3'-methyl ether 7-glucuronosyl-(1->2)-glucuronide	0	0	1	3
Molecular feature_1054.18	0	0	1	3
(S)-10,16-Dihydroxyhexadecanoic acid	0	0	2	4
Notoginsenoside T2	1	1	3	2
Cholesterol glucuronide	0	0	3	4
Pancreas				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
Molecular feature_379.97	1	3	0	0
Molecular feature_642.11	0	0	1	3
Molecular feature_322.99	1	3	0	0
3-Hydroxy-2-nitroestra-1,3,5(10)-trien-17-one	0	0	2	4
PA(6:0/6:0)	0	0	1	3
6-(2-Carboxyethyl)-7-hydroxy-2,2-dimethyl-4-chromanone glucoside	2	3	0	0
Molecular feature_1317.18	0	0	2	3
Hydroxychloroquine	3	2	0	0
7alpha-1(10->19)-Abeo-7-acetoxyobacun-9(11)-ene	0	0	3	4
PC(16:0/5:0(COOH))	3	2	0	0
Top Contributors to PC2				
Cortex				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
Molecular feature_676.93	1	0	0	3
Umbelliprenin	0	3	2	0
Afrormosin 7-O-(6''-malonylglucoside)	1	0	0	3

GW 7647	3	1	1	0
Molecular feature_712.06	0	3	1	0
3-hexanoyl-NBD Cholesterol	1	1	3	2
Quinidinone	4	0	0	1
5,7,3'-Trihydroxy-6,4',5'-trimethoxyflavanone	3	0	0	1
O-Acetylserine	1	0	0	3
Molecular feature_279.14	1	0	0	4
Heart				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
N1-(2-Methoxy-4-methylbenzyl)-n2-(2-(pyridin-2-yl) ethyl)oxalamide	1	3	2	2
4-Hydroxyazobenzene	3	4	0	1
Malvidin 3-(6-acetylglucoside)	2	1	1	3
Acepromazine	1	1	0	4
Ganoderenic acid C	0	1	2	3
Molecular feature_116.93	1	2	0	3
Molecular feature_966.01	0	1	1	3
24-Nor-5β-cholane-3α,12α,23-triol	1	3	0	0
Spergualin	1	1	1	3
PS(P-18:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z))	0	1	4	2
Kidney				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
Molecular feature_778.432	2	4	0	1
Neorauteen	1	1	0	3
Ziprasidone	3	1	1	1
Molecular feature_1713.73	4	3	1	0
Benzethonium chloride	2	3	1	0
PC(16:1(9E)/0:0)	4	2	3	3
Vilazodone	3	1	1	0
Aminoethoxyacetic acid	4	4	1	0
Ganoderenic acid E	2	3	1	3
2,4-Dihydroxyacetophenone 5-sulfate	1	5	1	0
Liver				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
Isogingerenone B	1	4	0	2
Arachidonoyl-CoA	3	2	0	1
Molecular feature_972.67	1	3	0	1
Cyclic adenosine diphosphate ribose	3	2	0	1
Bromodiphenhydramine	3	2	1	1
Norstictic Acid Pentaacetate	3	3	0	2
PE(13:0/20:5(5Z,8Z,11Z,14Z,17Z))	1	3	1	0
2R-hydroxy-stearic acid	3	3	1	0
Molecular feature_712.28	3	2	1	2
Molecular feature_590.33	4	4	0	1

Muscle				
Metabolite Name	Frequency: Basal (n = 5)	Frequency: Basal+Probiotic (n = 5)	Frequency: High Fat (n = 5)	Frequency: High Fat+Probiotic (n = 5)
Glucoliquiritin apioside	2	3	2	0
Neobifurcose	2	4	1	0
Molecular feature_1790.69	1	3	1	0
Molecular feature_176.14	3	1	2	0
2-Hydroxy-4,7-dimethoxy-2H-1,4-benzoxazin-3(4H)-one	4	3	0	1
Hydroxytetrabenazine glucuronide	1	2	3	1
Molecular feature_326.34	2	3	1	0
Molecular feature_205.09	4	2	2	1
Theasapogenol A	2	3	1	0
Molecular feature_149.02	3	0	0	1
Pancreas				
Metabolite Name	Frequency: Basal (n = 5)	Frequency: Basal+Probiotic (n = 5)	Frequency: High Fat (n = 5)	Frequency: High Fat+Probiotic (n = 5)
Molecular feature_554.22	1	1	3	2
LysoPC(0:0/18:0)	3	1	0	1
Malvidin 3-O-(6-O-(4-O-malonyl-alpha-rhamnopyranosyl)-beta-glucopyranoside)-5-O-(6-O-malonyl-beta-glucopyranoside)	1	3	0	2
Minoxidil-O-glucuronide	4	1	1	0
Molecular feature_1286.87	4	1	1	0
PE-Cer(d14:1(4E)/20:0(2OH))	2	4	0	1
Molecular feature_749.32	4	2	0	2
Molecular feature_339.29	3	2	1	0
PS(16:1(9Z)/22:2(13Z,16Z))	1	3	1	0
PPA(18:1(9Z)/18:1(9Z))	3	1	2	0
Top Contributors to PC3				
Cortex				
Metabolite Name	Frequency: Basal (n = 5)	Frequency: Basal+Probiotic (n = 5)	Frequency: High Fat (n = 5)	Frequency: High Fat+Probiotic (n = 5)
Molecular feature_557.15	1	0	3	0
Molecular feature_3098.70	0	3	0	4
UDP-N-acetyl-alpha-D-galactosamine	0	1	0	3
Avadharidine	4	0	1	0
Molecular feature_552.86	3	0	2	0
S-Seven	2	0	4	0
PS(14:1(9Z)/16:1(9Z))	4	0	2	0
7-Methylinosine	0	4	0	2
Oleoyl glycine	4	0	2	0
Molecular feature_164.98	0	3	2	2
Heart				
Metabolite Name	Frequency: Basal (n = 5)	Frequency: Basal+Probiotic (n = 5)	Frequency: High Fat (n = 5)	Frequency: High Fat+Probiotic

				(n = 5)
Molecular feature_3028.44	0	3	0	1
Molecular feature_499.99	0	3	0	3
Molecular feature_554.12	3	0	1	0
Normetanephrine glucuronide	1	0	3	0
Epinephrine glucuronide	0	2	0	3
3,4-Hexahydroxydiphenoylarabinose	0	3	0	2
Guanosine 3'-phosphate	0	3	0	2
Molecular feature_697.44	2	0	3	0
Molecular feature_635.14	2	0	3	0
Molecular feature_361.03	1	0	3	0
Kidney				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
Molecular feature_887.11	0	1	1	3
Molecular feature_398.09	3	0	1	2
Molecular feature_358.18	3	1	2	1
2-Ethyl-1-hexanethiol	5	0	1	2
b-D-Xylopyranosyl-(1->4)-a-L-rhamnopyranosyl-(1->2)-D-fucose	0	1	1	4
Molecular feature_892.13	0	1	1	3
Molecular feature_302.93	4	0	2	1
11'-Carboxy-alpha-tocotrienol	0	1	2	4
9-oxo-2E-decenoic acid	0	1	2	3
Molecular feature_922.80	0	1	1	3
Liver				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
PIP(16:0/22:2(13Z,16Z))	0	1	1	3
Molecular feature_560.14	0	1	1	3
Molecular feature_290.0343	2	1	3	1
26,27-diethyl-1α,25-dihydroxy-20,21-methano-23-oxavitamin D3 / 26,27-diethyl-1α,25-dihydroxy-20,21-methano-23-oxacholecalciferol	3	0	2	1
Molecular feature_365.25	0	1	1	4
Molecular feature_450.84	1	1	3	1
Molecular feature_998.29	3	1	1	2
Molecular feature_464.41	1	0	2	4
Molecular feature_561.37	0	1	3	4
Molecular feature_1635.51	0	2	1	4
Muscle				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
Xanthine	3	0	1	1
Dolichyl b-D-glucosyl phosphate	1	0	1	3
Diguanosine triphosphate	1	3	1	1

Moricizine sulfone	4	2	1	1
Cyanidin 3,5,3'-triglucoside	0	3	1	0
Oleamide	1	1	1	3
Mono-N-desisopropyldisopyramide	0	1	2	4
Butoctamide hydrogen succinate	2	0	3	3
(5-heptyl-6-methyloctahydroindolizin-8-yl)methanol	1	2	1	3
4-Bromophenol	1	1	2	3
Pancreas				
Metabolite Name	**Frequency: Basal (n = 5)**	**Frequency: Basal+Probiotic (n = 5)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+Probiotic (n = 5)**
PS(22:4(7Z,10Z,13Z,16Z)/0:0)	1	0	1	5
4'-O-Demethylrebeccamycin	1	0	1	3
Molecular feature_730.09	2	1	3	2
PI(22:1(11Z)/22:4(7Z,10Z,13Z,16Z))	3	0	1	1
Succinic acid	1	0	3	2
Molecular feature_690.00	1	3	2	2
Molecular feature_506.24	0	1	3	2
Molecular feature_1033.77	1	0	1	4
Molecular feature_932.55	1	0	3	4
6-bromo-tetracosa-5E,9Z-dienoic acid	1	0	3	4
n is equal to the number of samples in the cohort				
Features with associated names were putatively identified via MS. Metabolites which are not currently included in the Metlin or HMDB databases are provided the name Molecular Feature coupled to their respective m/z value.				

When examining all three principal components, the metabolites which contribute the most to cohort segregation in all tissues studied include phosphatidylcholines, phosphatidylserines, phosphatidylethanolamines, and phosphatidylinositols (all MS confirmed) which are all membrane glycerophospholipids, present in either the outer or inner sheet of membranes. Interestingly, phosphatidylcholines were among the top contributors to variance in principal component 1 and 2, the greatest and second greatest contribution of overall variation among the samples. Phosphatidylcholines are located primarily in the outermost sheet of membranes, interacting with extracellular environment. This is in contrast to phosphatidylserines, phosphatidylethanolamines, and

phosphatidylinositols, (contributors to principal components 2 and 3) which are located primarily in the inner sheet, interacting with the intracellular environment. Research has shown that dietary oils do indeed have an effect on membrane lipid composition, leading to metabolic syndrome[115]. Thus, it can be proposed that while a high fat diet has a profound effect on the membrane composition, probiotic supplementation, irrespective of diet, results in alterations to membrane composition with the greatest amount of variation occurring in the outermost sheet.

3. Metabolites Altered Due to Probiotic Supplementation

Supplementation of the diet with probiotics leads to detectable alterations in the tissue metabolomes, as evidenced in all six of the organs analyzed (Figure 54). That is, the basal and basal+probiotic cohorts are clearly segregated and distinct in the PCA plot, as are the high fat and high fat+probiotic cohorts, regardless of tissue type. Figure 57 presents tissue-specific volcano plots comparing the basal vs. basal+probiotic and the high fat vs. high fat+probiotic cohorts. Within every tissue examined, irrespective of the diet, numerous metabolites were found to be either unaffected, significantly increased in abundance, or significantly decreased by probiotic supplementation.

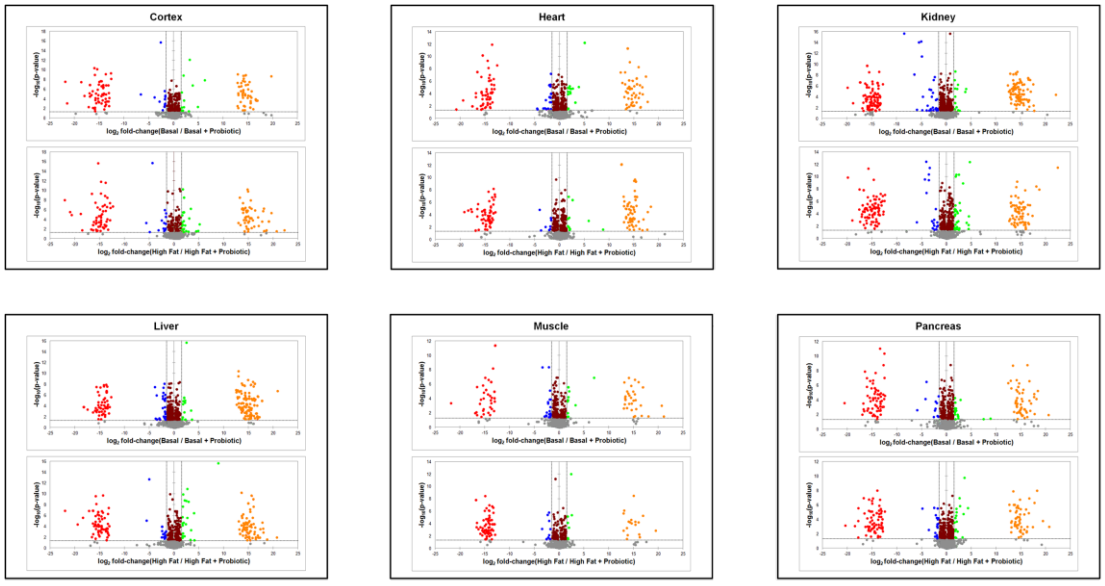


Figure 57. Tissue Specific Volcano Plots comparing the Basal versus Basal+Probiotic and High Fat vs High Fat+Probiotic Cohorts of Pigs.
Within each identified tissue type, the fold change in metabolite abundance is calculated as the ratio of the median chromatographic peak height in the basal cohort relative to the corresponding median peak height in the basal+probiotic cohort, or the median peak height in the high fat cohort relative to the corresponding median peak height in the high fat+probiotic cohort, as indicated. In the volcano plots, the $-\log_{10}$ of the Benjamini-Hochberg corrected *p*-value for each metabolite is presented as a function of the $\log_2$ of the relative abundance fold change, using a Benjamini-Hochberg critical cutoff value of 0.05 to perform the *p*-value corrections. The horizontal dashed line in the plots identifies a corrected *p*-value of 0.05 (i.e. $-\log_{10}(p\text{-value})=1.3$), while the black vertical dashed lines denote a 1.5 $\log_2$ fold change in metabolite abundance (increase and decrease). Each metabolite is represented as a circle in the plot, with circles colored grey having a statistically insignificant alteration in fold change (i.e. a corrected $-\log_{10}(p\text{-value})$ <1.3). Metabolites found to be generally unaffected by probiotic supplementation are colored brown (corrected $-\log_{10}(p\text{-value})$ >1.3 and a $\log_2$ fold change value between -1.5 and 1.5). Metabolites significantly increased by probiotic supplementation are colored blue (corrected $-\log_{10}(p\text{-value})$ >1.3 and a $\log_2$ fold change value between -1.5 and -10). Metabolites significantly decreased by probiotic supplementation are colored green (corrected $-\log_{10}(p\text{-value})$ >1.3 and a $\log_2$ fold change value between 1.5 and 10). Metabolites drastically reduced with probiotic supplementation are colored orange (corrected $-\log_{10}(p\text{-value})$ >1.3 and a $\log_2$ fold change value >10). Metabolites drastically increased with probiotic supplementation are colored red (corrected $-\log_{10}(p\text{-value})$ >1.3 and a $\log_2$ fold change value <-10). Within each tissue type, this analysis was restricted to only the molecular features appearing in a minimum of 3 of 5 pigs (60%) in any one of the two pig cohorts.

To identify the top 10 metabolites that are greatly influenced by probiotic supplementation (identified by comparison of basal versus basal+probiotic tissue samples along with comparison of high fat versus high fat+probiotic tissue samples), I focused my attention on tissue features present in at least 80% of any one cohort (i.e. the affected

tissue metabolite was identified in at least 4 of the 5 pigs in any one of the two compared cohorts, and thus appeared in at least 4 of the 10 pigs in the two compared cohorts), and ranked these metabolites by fold change, *p*-value, and quality of EIC, resulting in a list of 10 statistically significant features (Table 3; see Appendix 8 for the associated EICs and TICs). Of these 10 metabolites, 3 were subsequently identified via MS/MS fragmentation (10 eV, 20 eV, and 40 eV for each metabolite) and corresponding fragment pattern matching to fragmentation patterns of known compounds curated in the Metlin and/or Human Metabolome Database (HMDB). The MS/MS fragmentation pattern matching for these 3 metabolites is shown in Appendix 9. Conversely, the remaining 7 metabolites did not match the MS/MS fragmentation pattern of any molecules in Metlin or HMDB, and thus are not discussed further (these 7 metabolites are simply identified as a 'molecular feature' in Table 3, with the corresponding m/z value listed), the 3 identified metabolites are discussed in detail below.

Table 3. Top 10 metabolites exhibiting the greatest degree of change in all tissues when comparing basal versus basal+probiotic and high fat versus high fat+probiotic.

Cortex				
Basal versus Basal+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: Basal (n = 5)**	**Frequency: Basal+probiotic (n = 5)**
Molecular Feature_273.12	-17.94	4.67	0	4
High Fat versus High Fat+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+probiotic (n = 5)**
Molecular Feature_355.26	17.09	4.30	4	0
Heart				
Basal versus Basal+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: Basal (n = 5)**	**Frequency: Basal+probiotic (n = 5)**

Based on the imposed inclusion criteria, no metabolites were highlighted in this comparison.				
High Fat versus High Fat+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+probiotic (n = 5)**
Molecular Feature_348.15	-19.07	4.67	0	4
Kidney				
Basal versus Basal+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: Basal (n = 5)**	**Frequency: Basal+probiotic (n = 5)**
Uridine diphosphate-N-acetylglucosamine	22.10	4.62	4	0
High Fat versus High Fat+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+probiotic (n = 5)**
Molecular Feature_498.19	-17.38	2.34	0	4
Liver				
Basal versus Basal+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: Basal (n = 5)**	**Frequency: Basal+probiotic (n = 5)**
Saccharopine	16.94	5.80	4	0
High Fat versus High Fat+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+probiotic (n = 5)**
Molecular Feature_279.16	-22.02	6.80	0	4
Muscle				
Basal versus Basal+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: Basal (n = 5)**	**Frequency: Basal+probiotic (n = 5)**
Based on the imposed inclusion criteria, no metabolites were highlighted in this comparison.				
High Fat versus High Fat+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+probiotic (n = 5)**
Molecular Feature_291.07	19.55	2.70	4	0
Pancreas				
Basal versus Basal+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: Basal (n = 5)**	**Frequency: Basal+probiotic (n = 5)**
Molecular Feature_322.05	8.88	1.37	5	5
Proline	-6.04	2.53	5	5
High Fat versus High Fat+probiotic				
Metabolite Name	**log_2(Fold Change)**	**$-log_{10}$(p-value)**	**Frequency: High Fat (n = 5)**	**Frequency: High Fat+probiotic (n = 5)**
Based on the imposed inclusion criteria, no metabolites were highlighted in this comparison.				
Criteria for Selection: Frequency $\geq$ 80%; $-log_{10}$(p-value) $\geq$ 1.3; $\lvert log_2$(Fold Change) $\geq$ 1.5$\rvert$; Median abundance values $\geq$ 25000				

Note: A positive fold change when comparing basal diets versus basal+probiotic indicates an increase in abundance for the basal cohort. A positive fold change when comparing high fat versus high fat+probiotic diets indicates an increase in abundance for the high fat cohort. n equals the number of total samples in the cohort. Features with associated names are putatively identified via MS/MS using the Metlin and HMDB databases Metabolites which are not currently included in the Metlin or HMDB databases are provided the name Molecular Feature coupled to their respective m/z value.

1. Uridine Diphosphate-N-Acetylglucosamine

The tissue metabolomics investigation revealed that probiotic supplementation of the basal diet results in significantly decreased levels of Uridine Diphosphate-N-Acetylglucosamine (commonly referred to as UDP-GlcNAc) within the kidney tissue of Ossabaw pigs (Table 3). In fact, UDP-GlcNAc was undetected in all 5 of the kidney samples obtained from the basal+probiotic cohort pigs, while it was abundant in 4 of the 5 kidney samples obtained from the basal diet cohort of pigs. Synthesized *de novo* from fructose 6-phosphate and glucosamine via the hexosamine biosynthetic pathway[116,117], UDP-GlcNAc is utilized by the Golgi apparatus for O-linked protein glycosylation[118–120]. Metabolic flux through the hexosamine biosynthetic pathway is a reflection of nutrient availability, particularly glucose, glucosamine, UDP, and acetyl-CoA[118]. Since 3-5% of glucose is typically destined for UDP-GlcNAc production[116,117], it may be that probiotic supplementation results in alterations in the availability of these nutrients, thereby affecting the intracellular UDP-GlcNAc concentration. This probiotics-associated drop in UDP-GlcNAc level is unique to the kidney tissue, as it was not observed in any of the other tissues we analyzed (the UDP-GlcNAc level was consistent among the cohorts in these other tissues). However, the drop in UDP-GlcNAc level was observed in the high fat+probiotics cohort relative to the high fat cohort, with probiotic supplementation

linked to a complete absence of UDP-GlcNAc in all 5 of the pig kidney tissue samples (data not shown). While the cause of the probiotics-associated decrease in kidney UDP-GlcNAc levels remains uncertain, studies have shown that O-linked glycosylation is positively regulated by the intracellular concentration of UDP-GlcNAc[120], suggesting that a consequence of low levels of UDP-GlcNAc would be altered glycoprotein production in the basal+probiotic and high fat+probiotic pig cohort kidney cells. Interestingly, increased O-linked glycosylation of proteins has been implicated in insulin resistance and obesity[121]. It remains unknown if probiotic supplementation might serve to reduce the risk of metabolic syndrome by reducing intracellular kidney UDP-GlcNAc levels.

2. Saccharopine

The metabolomics investigation also discovered that probiotic supplementation to the basal diet leads to significantly decreased saccharopine concentration within the liver. While 4 of the 5 liver tissues obtained from the basal diet pig cohort have abundant saccharopine within, none of the basal+probiotic pig liver tissues had detectable levels of saccharopine. Saccharopine is an intermediate within a lysine catabolic pathway in the liver[122–125]. Unlike other amino acids, lysine does not undergo direct transamination. Rather, lysine and α-ketoglutarate condense to form saccharopine, which is subsequently oxidized to form α-aminoadipic δ-semialdehyde and glutamate via the bifunctional enzyme aminoadipic semialdehyde synthase[122,123,126]. Downstream metabolic pathways can route α-aminoadipic δ-semialdehyde and glutamate for nitrogenous biomolecule biosynthesis and energy production in the liver, so it is tempting to suppose that with the basal diet the pigs are utilizing dietary lysine as an energy source, but with probiotic

supplementation the added probiotic bacteria utilize the lysine pool in the GI tract, thereby lowering the saccharopine concentration within the liver. Notably, the liver saccharopine concentration is essentially the same in the high fat and high fat+probiotics cohorts, suggesting the caloric abundance of the high fat diet does not necessitate amino acid oxidation by the probiotic bacteria (or the liver). While the cause of the reduced liver saccharopine concentration in the basal+probiotic pigs remains speculative, perhaps more importantly the significance/effect of the reduced saccharopine concentration in the liver of the pigs requires further investigation.

3. Proline

The third top 10 metabolite identified by MS/MS as significantly influenced by probiotic supplementation is the amino acid proline. With a frequency of appearance of 100% in all 4 pig pancreas cohorts (5 of 5 pigs in the basal, basal+probiotic, high fat, and high fat+probiotic cohorts), probiotic supplementation results in a significant increase in the abundance of proline in the pancreas of the basal+probiotic cohort, relative to basal (~100 fold increase in median abundance). Notably, probiotic supplementation of a high fat diet increases pancreas proline levels approximately 2 fold relative to the high fat diet alone. More significantly, relative proline levels are approximately 100 fold higher in the basal+probiotic, high fat, and high fat+probiotic cohorts compared to the basal diet. A high activity of mitochondrial proline oxidase is known to reside in the intestinal absorptive cells in pigs, and as a consequence up to 90% of dietary proline becomes metabolized to ornithine, citrulline, and arginine, thereby lowering the amount of dietary proline available to the extraintestinal tissues[127,128]. Given that probiotic supplementation

of the basal diet increases the pancreatic level of proline 100 fold, probiotic supplementation of the high fat diet increases the pancreatic level of proline 2 fold, and that a high fat diet yields 100 fold more pancreatic proline than does a basal diet, it appears that both a high fat diet and probiotic supplementation influence intestinal mucosal proline metabolism, resulting in increased dietary uptake of proline.

4. Probiotic Induced Metabolic Reversions

While probiotic induced changes to the tissue metabolomes are of interest in this investigation, a more important goal was to identify if probiotic supplementation of a high fat diet can 'revert' metabolomic changes associated with the high fat diet (relative to a healthy basal diet) back to metabolite levels on par with the basal diet. Intriguingly, it was discovered that for many tissue metabolites, probiotic supplementation of a high fat diet does indeed lead to alterations in metabolite abundance such that they assume a level associated with the basal diet. That is, probiotic supplementation effectively reverts these high fat-associated metabolite levels to a level linked with a nutritionally balanced basal diet, even though the pigs continued to consume an obesogenic diet.

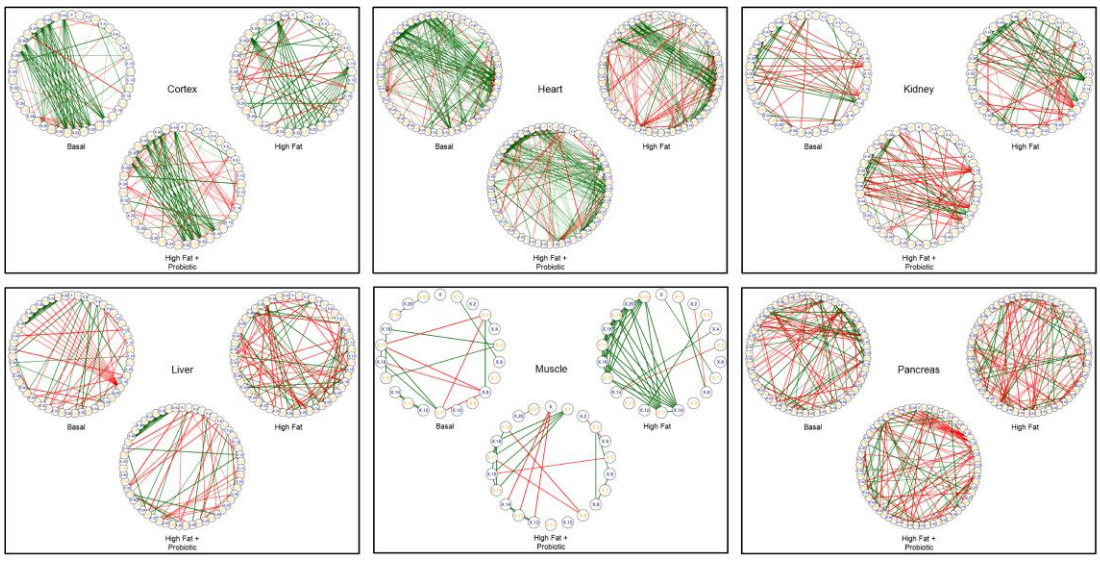


Figure 58. Tissue-specific metabolite correlation networks for the basal, high fat, and high fat+probiotic metabolomes

For each tissue type, Pearson's correlation coefficients were calculated for metabolites present in a minimum of 3 of 5 pigs in any one of the three pig cohorts, and only metabolites common to all three cohorts were included in the network. Positive Pearson correlation values, $r \geq 0.90$, are depicted as a green line between metabolites, while negative correlations, $r \leq -0.90$, are depicted as a red line between metabolites. To facilitate comparison of the networks, metabolites are represented as a small alpha numerically labeled circle and their placement around the circumference of each network is fixed among the plots for each tissue. For all tissues, comparison of the metabolite correlation networks indicates that probiotic supplementation of the high fat diet alters the metabolome relative to the high fat diet alone, and for the cortex, heart, and kidney tissues in particular, probiotic supplementation changes the high fat metabolome in such a manner so as to make it appear more similar to the correlation networks associated with the basal diet.

To demonstrate this reversion on a global metabolomic scale, I generated a series of network correlation maps depicting the relationship among metabolites within the tissues (Figure 58). Within the cortex tissue, for example, numerous metabolites demonstrate positive and/or negative correlations with each other in the basal diet cohort. However, for many of these metabolites, the network correlations are significantly altered within the high fat cohort, as evidence in the corresponding correlation map (compare the appearance of basal vs. high fat in Figure 58). Intriguingly, the supplementation of probiotics to the high fat diet results in a reversion of the interrelationship between many

of these metabolites to a network correlation pattern more reflective of that which occurs with the basal diet (note how high fat+probiotics is similar in appearance to basal in Figure 58). This reversion phenomenon is also apparent in the heart and kidney tissue. Hence, probiotic supplementation confers measurable changes to the composition of the tissue metabolome, and for some of the tissues, reverts a high fat associated metabolome to one that is more akin to the basal cohort (Table 4).

Table 4. Metabolites associated with probiotic induced reversion of the high fat metabolome.

;Cortex					
Metabolite Name	**log_2(Fold Change) High Fat vs Basal**	**log_2(Fold Change) High Fat+probiotic vs Basal**	**Frequency: High Fat (n = 5)**	**Frequency: Basal (n = 5)**	**Frequency: High Fat+probiotic (n = 5)**
S-Adenosylhomocysteine*	19.00	0.00	3	0	0
Heart					
Metabolite Name	**log_2(Fold Change) High Fat vs Basal**	**log_2(Fold Change) High Fat+probiotic vs Basal**	**Frequency: High Fat (n = 5)**	**Frequency: Basal (n = 5)**	**Frequency: High Fat+probiotic (n = 5)**
CDP-glycerol	16.77	0.00	3	0	0
Kidney					
Metabolite Name	**log_2(Fold Change) High Fat vs Basal**	**log_2(Fold Change) High Fat+probiotic vs Basal**	**Frequency: High Fat (n = 5)**	**Frequency: Basal (n = 5)**	**Frequency: High Fat+probiotic (n = 5)**
Indoxyl sulfate *	-17.57	1.61	0	3	3
Liver					
Metabolite Name	**log_2(Fold Change) High Fat vs Basal**	**log_2(Fold Change) High Fat+probiotic vs Basal**	**Frequency: High Fat (n = 5)**	**Frequency: Basal (n = 5)**	**Frequency: High Fat+probiotic (n = 5)**
Based on the imposed inclusion criteria, no metabolites were highlighted in this comparison.					
Muscle					
Metabolite Name	**log_2(Fold Change) High Fat vs Basal**	**log_2(Fold Change) High Fat+probiotic vs Basal**	**Frequency: High Fat (n = 5)**	**Frequency: Basal (n = 5)**	**Frequency: High Fat+probiotic (n = 5)**
D-Sedoheptulose 7-phosphate	19.55	0.00	4	0	0
Pancreas					

Metabolite Name	log_2(Fold Change) High Fat vs Basal	log_2(Fold Change) High Fat+probiotic vs Basal	Frequency: High Fat (n = 5)	Frequency: Basal (n = 5)	Frequency: High Fat+probiotic (n = 5)
Estradiol disulfate	-16.30	0.43	0	5	3
Criteria for Selection: Frequency ≥ 60% in High Fat or Basal AND High Fat+Probiotic; p value < 0.05; $\|log_2$(Fold Change) ≥ 10$\|$ in High Fat vs Basal AND $\|log_2$(Fold Change) ≤ 10$\|$ in High Fat+Probiotic vs Basal; Median abundance values ≥ 25000. n is equal to the number of samples in the cohort.					
Note: A positive fold change when comparing high fat versus basal diets indicates an increase in abundance for high fat pigs. A positive fold change when comparing high fat+probiotic versus basal diets indicates an increase in abundance for the high fat+probiotic cohort.					
* Indicates metabolite was putatively identified via MS/MS; all others were putatively identified via MS.					

To identify the top metabolites that are responsible for this reversion, I elected to include only those metabolites that are present in at least 80% of any one of the three cohorts (high fat, high fat+probiotic, and basal) with the caveat that if the metabolite was present in the high fat cohort, it must be undetected in the high fat+probiotic and basal cohorts. This resulted in a list which contained only 4 metabolites. Therefore, I lowered the frequency to 60% with the same restrictions resulting in a data set which included a total of 17 metabolites. Next, I further refined the data using the EIC for each metabolite, as described previously (see Appendix 9, for the EICs and companion TICs). Table 4 presents the top 5 metabolites that undergo reversion. Of these 5 metabolites, 2 could be successfully identified via MS/MS by fragmentation pattern matching known compounds curated in the Metlin library and HMDB. These two compounds are discussed in further detail below. The remaining 3 of the top 5 reversion metabolites could not be identified by MS/MS (the compounds are not in the reference database). However, putative identification is possible by MS, and a discussion of them is also included below.

1. S-adenosylhomocysteine

Within the cortex, probiotic supplementation of the high fat diet results in a suppression of the elevated intracellular concentration of S-adenosylhomocysteine (SAH) observed in 3 of the 5 pigs in the high fat cohort, thus reverting the metabolite to a phenotype more akin to that found within the healthy basal diet cohort of pigs. SAH is a member of the methyl cycle, and is an immediate product derived from S-adenosylmethionine (SAM), a powerful methylating agent in numerous biosynthetic reactions (catalyzed by a variety of methyl transferases)[129–131]. SAH is known to act as a negative regulator of several of these methyltransferases, thereby inhibiting the rate of methylation using SAM[132]. Studies have shown that increased intracellular concentration of SAH is correlated to DNA hypomethylation, leading to an increase in reactive oxygen species (ROS) levels related to inhibition of the methylation of tRNAs responsible for translating the selenoproteins which reduce H_2O_2 levels[132]. Additionally, SAH has been linked to an increase in cognitive impairment and disease progression in Alzheimer's patients[133,134]. Thus, while the mechanism underlying the phenomenon remains unknown, the effect of suppressing the intracellular SAH abundance via probiotic supplementation is highly desirable.

2. Indoxyl sulfate

Indoxyl sulfate (also referred to as indican) is a product of tryptophan degradation and protein putrefaction[135,136]. Within the intestines, tryptophan is converted by bacteria into indole, which can be absorbed into the blood stream and then further metabolized in the liver. Indole accumulation is toxic (due to its ability to disrupt membrane integrity[137]),

so indole is converted by the liver into indoxyl sulfate, which migrates to the kidneys and is excreted in the urine[136]. While a basal healthy diet leads to measurable levels of indoxyl sulfate in kidney tissue, kidneys obtained from pigs fed a high fat diet were found to be devoid of this metabolite. While the reason for this is unclear (perhaps the microbes preferentially utilize the abundance of fat for fuel, rather than protein degradation), probiotic supplementation of pigs fed the high fat diet reverts the level of indoxyl sulfate to that observed with the basal diet (possibly due to the abundance of probiotic bacteria in the GI tract consuming the additional fat and reestablishing protein putrefaction).

3. CDP-Glycerol

Probiotic supplementation of the high fat diet results in a suppression of the intracellular heart concentration of cytidine diphosphate-glycerol (CDP-glycerol; MS identified). Not detected in the hearts of pigs in the basal diet cohort, an elevated level of CDP-glycerol was found in 3 of 5 pigs in the high fat cohort. CDP-glycerol is an intermediate in glycerophospholipid metabolism[138]. As an increase in intermediates associated with glycerophospholipid metabolism have been linked to an increase in coronary heart disease (CHD)[139], it is interesting to postulate that probiotic supplementation to a high fat diet may result in an alteration to the pathways associated with glycerophospholipids, thereby reducing the risk of CHD.

4. Sedoheptulose 7-phosphate

Probiotic supplementation of the high fat diet also resulted in an alteration in the muscle sedoheptulose 7-phosphate concentration, an intermediate of the reversible nonoxidative phase of the pentose phosphate pathway[140–142]. Sedoheptulose 7-phosphate

was detected in 4 of the 5 pigs in the high fat cohort, with probiotic supplementation reducing the level below the limits of detection in 5 of 5 pigs, akin to that found in the basal diet cohort (0 of 5 pigs had detectable amounts of sedoheptulose 7-phosphate in the basal cohort). Increased sedoheptulose 7-phosphate levels may reflect enhanced degradation of xylulose 5-phosphate (a precursor of sedoheptulose 7-phosphate), an intermediate of the pentose phosphate pathway that directly upregulates the process of fatty acid biosynthesis (thus the high fat diet associated degradation of xylulose 5-phosphate to sedoheptulose 7-phosphate would decrease metabolic flux to fatty acid biosynthesis). However, since the expression of the fatty acid synthase gene is very low or nonexistent in skeletal muscle[143], the alteration of the sedoheptulose 7-phosphate levels detected here may in fact originate within muscle associated adipose cells (ie. marbling of the muscle) isolated when the muscle samples were obtained from the pigs. This, as well as elucidating the mechanism by which probiotic supplementation reduces sedoheptulose 7-phosphate levels, will require further investigation.

5. *Estradiol disulfate*

Within the pancreas, probiotic supplementation of the high fat diet resulted in an elevation in the intracellular concentration of the steroid conjugate, estradiol disulfate (MS identified; undetected in the high fat cohort, yet found in all of the basal diet pigs and 3 of the 5 high fat+probiotic diet pigs). Estradiol is a cholesterol derived steroid hormone, synthesized within the ovaries and adipose tissue. As with other steroid hormones, estradiol acts as a signaling molecule for a variety of different metabolic processes such as the regulation of sodium intake, inhibition of testosterone synthesis,

and energy homeostasis[144–146]. To control the intracellular concentration of estradiol, processes such as sulfation are used for inactivation, transport, and subsequent excretion from the body[144,147]. A reduction in estradiol has been shown to reduce leptin and increase appetite, thereby increasing a risk for obesity[146,148]. In addition, when compared to a healthy weight, a reduction in the levels of steroid sulfates, such as estradiol disulfate, has been shown in overweight and obese girls[147]. Thus, as probiotic supplementation of the high fat diet resulted in an increase in estradiol disulfate levels, one can propose that probiotic supplementation may alter steroid biosynthesis and decrease the risk for obesity.

Summary

Based on the tissue metabolomics results presented here, I conclude that a change in diet alone has a profound effect on the tissue metabolome, agreeing with the old adage that you are what you eat. As depicted in Figures 52 and 54, a high fat diet has an impact on all of the tissues examined in this investigation, from heart and skeletal muscle, to pancreas, liver, kidney, and cortex, with the greatest amount of variation resulting from an alteration in membrane lipid composition. The kidney and cortex appear particularly sensitive to intake of a high fat diet (Figures 52 and 53). Further, this study demonstrates how probiotic supplementation can also have an effect on the metabolome composition, creating metabolic phenotypes that are differentiated from those obtained with a basal and high fat diet (Figure 54). Notably, the metabolic differences identified by probiotic supplementation include alterations to metabolic flux through the hexosamine biosynthetic pathway resulting in a reduction of intracellular UDP-GlcNAc levels within

the kidney; suggesting an altered abundance in glycoprotein production and (when coupled to an obseogenic diet) a possible prevention of metabolic syndrome. Additionally, probiotic supplementation on the nutritionally balanced basal diet resulted in an alteration in lysine catabolism within the liver, presumably a reflection of preferential lysine utilization by the microbes within the GI tract. Finally, probiotic supplementation to the basal diet resulted in an increase in intracellular proline concentration within the kidney suggesting an alteration in its utilization comparable to what was measured within the high fat cohorts (with and without probiotic supplementation).

Through an in-depth analysis of the basal, high fat, and high fat+probiotic cohorts, I discovered that with probiotics supplementation, some potentially aberrant metabolites identified in the high fat cohort tissues were reverted to levels associated with those found in pigs fed a healthy, nutritionally balanced diet. Some of these metabolites include those that are associated with restoring processes associated with DNA methylation and the removal of toxic by-products of protein putrefaction. In addition, I identified possible alterations to glycerophospholipid metabolism and utilization of available fatty acids, a reduction in intermediates of the pentose phosphate pathway, and potential restoration of steroid biosynthesis. Coupled together, the reversion of the intracellular concentration of these metabolites to a level akin to a healthy diet depicts the beneficial nature of probiotic supplementation.

SPECIFIC AIM 2.4: METABOLIC CHANGES IN PIGS DUE TO WHIPWORM (*TRICHURIS SUIS)* INFECTION.

Objective: Use metabolomics to characterize the metabolic changes associated with various stages of *T. suis* infection in pigs and demonstrate how changes in certain metabolites relate to mammalian inflammatory pathways.

Introduction

The autoimmune disorder Inflammatory Bowel Disease (IBD), characterized by chronic inflammation of the intestinal tract, affects over 1.3 million people in the United States alone[149]. Two of the most widely discussed and common variants of this disease are Crohn's and Ulcerative Colitis (UC)[149–151]. Crohn's disease is characterized by the continual inflammation of the gastrointestinal tract (from mouth to rectum), predominately occurring throughout the intestinal lining of the ileum, resulting in severe abdominal pain and diarrhea during symptomatic periods. Alternatively, Ulcerative Colitis is characterized by chronic inflammation and ulcers found exclusively within the innermost lining of the colon and rectum. As IBD is a chronic disorder, treatments range from drug therapies such as anti-inflammatory drugs, immunosuppressants, and/or antibiotics (in the case of UC)[151], to surgical intervention to excise the afflicted portions. As the inflammation in UC often occurs throughout the entire colon or rectum, many UC patients have to undergo surgery resulting in partial or complete removal of the affected

organ. Conversely, inflammation associated with Crohn's disease occurs in patches; as a temporary solution, approximately 50% of Crohn's patients have the damaged portions of their intestines removed. While this may abate symptoms, lack of treatment of the underlying causes of inflammation result in possible reoccurrence, resulting in more surgery to excise the newly afflicted portions[152,153]. Because current therapies are so invasive and do not treat underlying biochemical causes of inflammation, researchers are exploring alternative approaches to ease patient pain and address IBD etiology.

One proposed therapy, helminthic therapy, is a treatment that stems from the hygiene hypothesis. The theory states that lack of exposure to parasites can increase the chance of development of an autoimmune disorder such as IBD[154,155]. Therefore, by administering helminths, in particular *Trichuris suis* ova which cannot reproduce within the human host, those who are afflicted with autoimmune disorders will see their symptoms subside. There have been several clinical trials involving the use of *T. suis* in IBD treatment[156–160]; current research has established that a continuous dose of 2500 *T. suis* ova is safe and effective in UC and Crohn's patients[161].

The helminth used in these trials, *T. suis*, is a parasitic nematode infectious to pigs. Though the natural host is the pig, *T. suis* has been shown to have immune modulating effects on people with IBD[155]. The ova has a characteristic oval shape with two polar plugs while the mature worms are long and "whip-like." The life cycle of *T. suis* is typical to most nematodes and begins once the embryonated egg (ova) is ingested. Once ingested, the first infective stage larvae (L1), hatch in the ileum, cecum, and proximal colon mucosa (PCM). The larva buries itself into the mucosal lining where it

undergoes four subsequent molts. The first molt (L1 to L2) begins at approximately day 10 post infection, the second, a transition to L3, occurs between days 16-21. At this point the worm's posterior emerges from the mucosa. The third and fourth molts, L4 and L5 (young adult), occur around days 32 and 37. By day 41, the worms are mature adults and begin to produce eggs[162].

In collaboration with Dr. Joe Urban at the United States Department of Agriculture (USDA), the goal of Specific Aim 2.4 was to characterize the metabolic changes associated with various stages of *T. suis* infection in pigs and determine if these metabolic changes are related to any inflammatory pathways implicated in the development of IBD. To achieve this, I separated the larval lifecycle into three specific stages. The first stage corresponds to the first and second molt (L1 to L2, Days 1-15). The second stage corresponds to the transition into the young adults (L3-L4, Days 16-37). The final stage corresponds to when the adult worm begins to produce and lay eggs (L5, Days 38-53). I then tracked the appearance and levels of the non-volatile metabolites identified from gastrointestinal tissue (Jejunum, Ileum, and Proximal Colon Mucosa (PCM)) and fecal material (Jejunum, Ileum, and PCM) (Figure 59). The data was acquired using an Agilent 6530 LC-QToF, and further processed using principal component analysis, hierarchical clustering analysis, correlation networks, and fold change analysis.

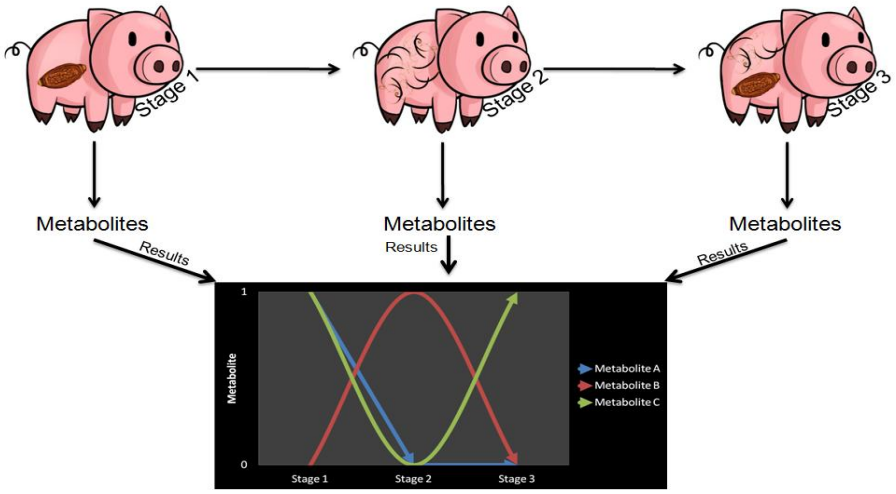


Figure 59. Schematic examining the metabolic alterations as a consequence of *T. suis* infected stage.
In this investigation, I examined the alterations to the metabolic composition for the various time points during the infection. For clarity, I have broken the investigation down into three stages, comprising of four different time points, 32 samples in total. For stage 1, I have Day 10 containing intestinal tissue and content from pigs infected with L2 larvae. For stage 2, I have two different time points. Day 21 illustrates the early portion of stage 2 with intestinal tissue and content from pigs infected with L3 larvae. Day 35, late portion of stage 2, contains intestinal tissue and content from pigs infected with L4 larvae. Finally, stage 3, I have Day 53, containing intestinal tissue and content from pigs infected with adult worms.

Materials and Methods

For Specific Aim 2.4, the pigs were housed and infected using the *T. suis* infection protocol at the USDA. In addition, all intestinal tissue and content samples were collected at the USDA. I exclusively processed the samples and performed the data analysis as described in methods below.

1. T. suis Infection Protocol.

Adult female worms were cultured in fresh complete Dulbecco's Modified Eagle's Medium (DMEM (pH 7.2) supplemented with 250 U/mL penicillin, 250 µg/mL streptomycin, 0.625 µg/mL amphotericin B, 400 µg/mL chloramphenicol at

approximately 20 worms/mL. Plates were stored in a 37 °C incubator with 10% CO_2 in air and left undisturbed for 24 hours. The adult worms release eggs into the media which were then harvested by centrifugation, washed in PBS with antibiotics, and maintained at room temperature for 45-60 days to allow development of infected larva inside the egg. The pigs were then inoculated with 2,000 - 20,000 eggs/per pig depending on the age of the pig and amount of larvae or adult worms that were needed. Following the specified infection duration, the pigs were euthanized and their organs were collected. The worms were removed from the intestines and their physiology was recorded. The tissue samples were stored at -80 °C until they are ready to be analyzed.

2. Sample Preparation.

Using liquid nitrogen to keep the tissue samples frozen, the samples were ground into fine powder using a TissueLyser II (Qiagen, Germantown, MD) then weighed into pre-chilled microcentrifuge tubes (~100 mg aliquots). The powdered tissues were then extracted using a methanol:water solution (1:1), added in a ratio of 3 µL of solvent/mg frozen tissue. The sample was vortexed, sonicated at room temperature for 5 min, clarified by centrifugation (15 min at 16.1 rcf), and the supernatant collected, filtered using Supelco (54145-U) Iso-disc, N-4-2 nylon, 4 mm x 0.2 µm filters (Sigma-Aldrich), and transferred to high-recovery amber vials (Agilent Technologies, Inc., Santa Clara, CA).

3. UPLC-QToF Analysis.

Samples were analyzed using an Agilent 1290 Infinity UPLC equipped with a C18 column (Zorbax Eclipse XDB-C18, 2.1 x 100 mm, 1.8 µm) and coupled to an Agilent 6530 QToF and data was acquired in ESI positive and negative modes. The mobile phase was dispensed by a binary pump at a flow rate of 0.4 mL/min. In ESI positive mode, solvent A was composed of LCMS Grade Water + 0.1% v/v formic acid (Proteochem, Loves Park, IL) and solvent B was composed of HPLC Grade Methanol + 0.1% v/v formic acid (Fisher Scientific). In ESI negative mode, solvent A was composed of Milli-Q Water + 0.01% v/v formic acid and solvent B was composed of HPLC Grade Methanol + 0.01% v/v formic acid (Fisher Scientific). Regardless of the ESI mode used, the solvents were dispensed over a gradient: 0 min 2% solvent B, 10-14.5 min 100% B, 14.51-16.50 min 2% B. The flush port was set to clean the injection needle for 2 seconds. The column was maintained at an isothermal temperature of 50 °C. Mass spectrometric analysis was performed by the QToF and the ESI source was set for a detection mass range from mass-to-charge ratio (m/z) 100-600 and a scan rate of 1.67 spectra/sec. To achieve accurate mass correction, a dedicated isocratic pump continuously introduced reference standards of purine (Agilent Technologies, Inc.) and hexakis-H, 1H, 3H-tetrafluoropropoxy-phosphazine, or HP-921 (Agilent Technologies, Inc.) at a flow rate of 0.5 mL/min. The nebulizer pressure was set at 45 psi with a sheath gas temperature of 350 °C and a gas flow rate of 11.0 L/min. The drying gas temperature was set at 325 °C with a flow rate of 10.0 L/min. For tandem MS analysis, the collision energy varied between 10, 20, and 40 eV. The data was acquired using Agilent Technologies'

MassHunter Acquisition SW Version, 6200 series ToF/6500 series Q-ToF B.05.01 (B5125.1).

4. Data Processing, Chemometrics and Statistical Analysis

Molecular features were identified in the raw chromatograms using Agilent Technologies' MassHunter Qualitative Analysis software (ver B.06.00). Molecular features and their relative abundance (peak height) were tabulated using Agilent Technologies' Mass Profiler Professional software (ver 12.6), resulting in a metabolomics data set containing the features identified by reversed phase ESI positive and reversed phase ESI negative modes. Metabolites present in ≤9% of the total number of samples processed were treated as one-offs and were removed from the matrix[33]. The score values (Equation 1) were calculated for each of the metabolites and filtered to include only metabolites appearing in at least 3 of 6 samples in any 1 cohort. The samples in the data set were organized by their appropriate infection duration, and the outlier peak area values were identified in each cohort using an analysis of (mean-median)/median for each analyte and a cutoff value ≥1.5. Outliers were replaced with the median value for that metabolite within the cohort. Metabolite peak height values were then standardized across the two cohorts by conversion to Z-scores (Equation 3). A Pearson (n) principal component analysis was then performed using the standardized metabolite matrices and the R statistical package. XLSTAT was used to perform two-sample T tests between cohorts for each metabolite. Benjamini-Hochberg critical values were calculated. A correlation network was created using the R statistical package. Fold change calculations were performed using Microsoft Excel.

Results and Discussion

1. Data Preparation

Following data acquisition by the LC-QToF, molecular features were acquired via Agilent Technologies' MassHunter Qualitative Analysis software as illustrated in ***Specific Aim 1: Step 1 of the metabolomics pipeline*** for every MS acquired data. Following the extraction of molecular features, an all-inclusive metabolite data set was then compiled for each tissue and content, from each member of the four cohorts, containing a grand total of 331,407 molecular features. The data consisted of four infection time points along the three stages of infection as illustrated in Figure 60, with a total number of 32 samples per tissue and content type. For Stage 1, the time point was Day 10 post infection (6 samples) and consisted of pigs with L2 larvae. For Stage 2, there are two different time points examining the early and late portions: Day 21 post infection (6 samples) constituted early stage 2 with pigs infected with L3 larvae while Day 35 post infection (8 samples) constituted late stage 2 with pigs infected with L4 larvae. Finally, for stage 3, the time point was Day 53 (12 samples) which consisted of pigs infected with adult worms.

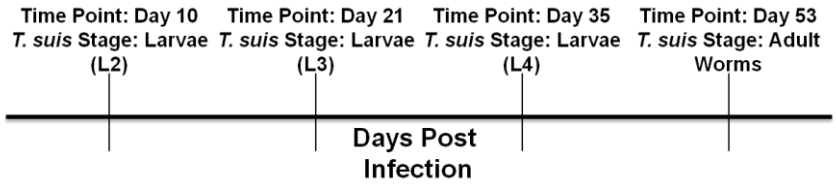


Figure 60. Timeline indicating the four infection time points used when examining *T. suis* disease progress.
The timeline shown above illustrates the four infection time points samples used for this analysis. Each time point contains a unique *T. suis* and belongs to one of the three infection stages used for this analysis as previously shown in Figure 59. See text for additional information.

As each time point did not include the same number of samples (minimum number of samples: 6; maximum number of samples: 12; total number of samples: 32); I restricted the analysis to include only those features detected in 3 of 6 pigs within one cohort, 3 of 32 overall. Due to the filtration process, the data set was constrained to include between 2520-3507 (grand total of 18036) molecular features within the six different intestinal tissue and content types analyzed, a significant reduction from the initial total.

2. Multivariate Statistical Analysis and Data Visualization

Due to the large number of molecular features acquired, I began my analysis by generating high resolution, focused PCAs of the intestinal tissue and content to examine the disease progression (Figure 61). The PCA plots depict significant alterations to the porcine intestinal tract as the disease progresses. During the initial stage of infection, L1 larvae hatch from the eggs and adhere to the porcine intestinal mucosa in the ileum, cecum, and proximal colon where they undergo their first molt into L2. In the subsequent

stages of infection, larvae molt from L2 into L3, at which point the posterior ends of the larvae begin to extend out of the mucosa, and into L4, as they transition into adult worms.

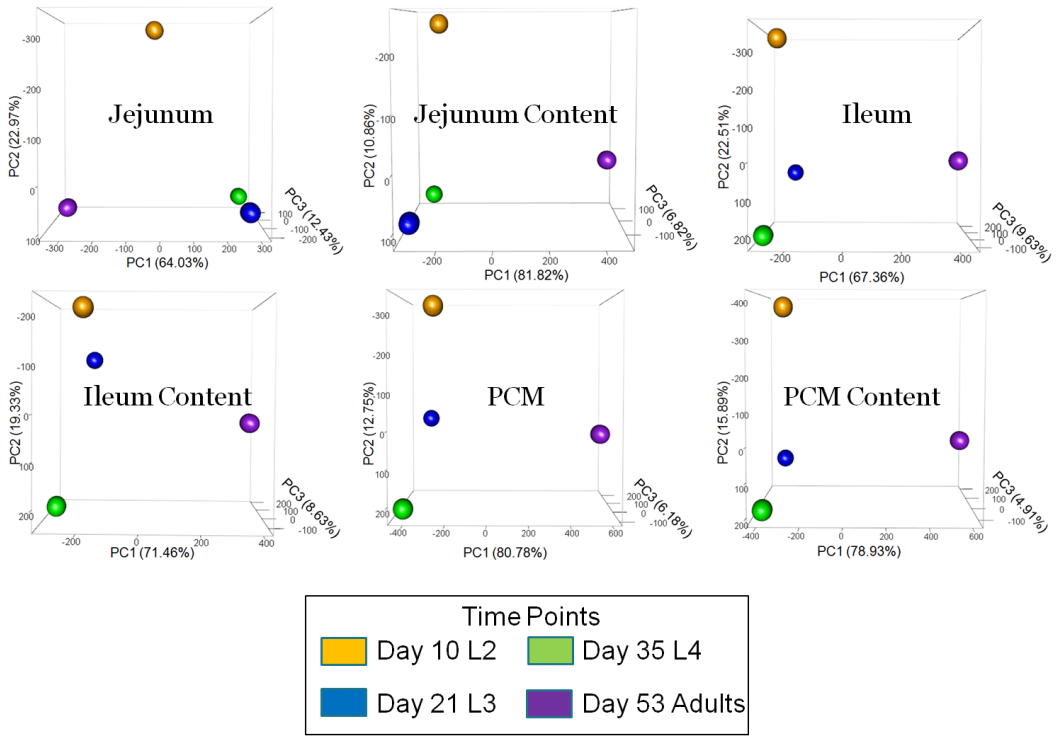


Figure 61. Disease Progression and the resulting intestinal tissue and content metabolome.
A total of 32 pigs were distributed among the four distinct time points (Day 10 post infection, Day 21 post infection, Day 35 post infection, and Day 53 post infection). The three dimensional PCA plot depicts the alterations in the metabolic profile obtained from each of the four time points. Each sphere on the plot reflects the metabolome of the pig intestinal tissue or content sample. Spheres are colored as indicated in the figure. The analysis was confined to molecular features appearing in a minimum of 3 of 6 pigs in any one of the four pig cohorts.

While all individual time points are distinct from the others, the two second stage cohorts are more tightly clustered in the PCA plots, demonstrating a greater similarity in their overall metabolome composition relative to the other time points (Days 10 and 53). The only exception to this trend is found in the ileum content where Day 10 and Day 21

are most tightly clustered. This is expected as the L2 larvae in Day 10 will molt into the L3 larvae of Day 21 within the intestinal mucosa of the ileum. Notably, within the tissue and content of the ileum and proximal colon mucosa, the relative positioning of the spheres are in counter-clockwise order as the infection progresses from stage 1 through stage 3. Finally, while it appears there is only one sphere for each time point, the PCA plot is a collection of all samples, illustrating minute metabolomic variations occurring within each time point.

While the PCA plots focused on the *differences* elucidated in the metabolic profile, dendrograms focusing on the intestinal tissue and content illustrate the *similarities* between the time points (Figure 62). Interestingly (with the exception of the jejunum and PCM tissue), the dendrograms cluster in a similar fashion as the PCA ordering from the distant to closest relative, Day 53, Day 10, Day 21 and Day 35. The dendrograms again illustrate the relative similarities of the metabolic profile obtained from Days 21 and 35 and the differentiation from the Day 53 (adult) profile. This clustering pattern is present in all sample types except for the PCM tissue where there are two distinct clusters, Days 21 and Days 53 versus Days 10 and Days 35.

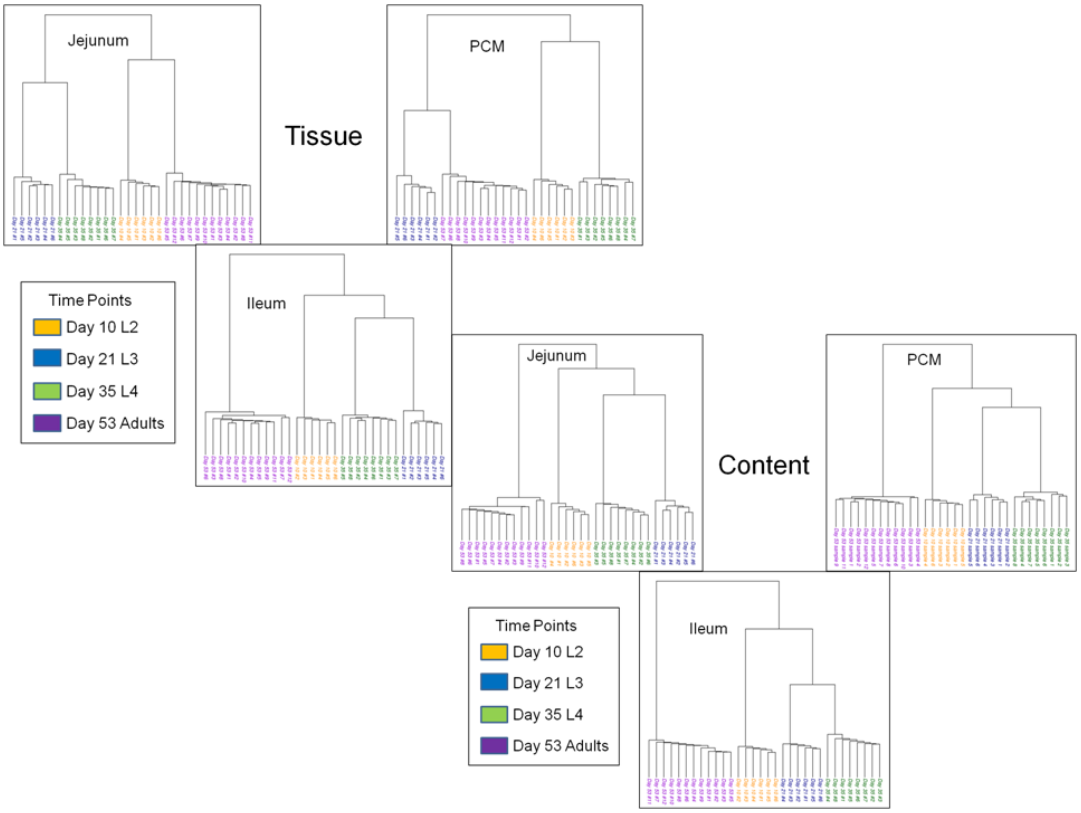


Figure 62. Dendrogram generated by unsupervised hierarchical clustering of the intestinal tissue and content samples according to the similarity of the metabolome composition.
The intestinal tissue and content samples are labeled in the figure and color coded based on their respective time points. Note that without exception, the samples from each time point, cluster together with their respective cohort. With the exception of the PCM tissue, the Day 21 and Day 35 cohorts organize into sister clades. It is also noteworthy that, with the exception of the jejunum and PCM tissue, the time points all follow the same clustering pattern: featuring sister clades Day 21 and 35, followed by segregation of Day 10 and higher Day 53. This illustrates the similar transition from adult worm to ova and subsequent molts from L1 to L4.

Due to the vast amount of positive and/or negative interrelationships amongst the metabolites, it was difficult to discern any potential trends in the data set using correlation networks ($r \geq |0.99|$, Figure 63). When I increased the stringency of the correlation value to $r \geq |0.999|$, I was still unable to elucidate any potential trends in the data. A correlation value of $r \geq |1|$ was deemed too stringent as there were no more than 7 metabolites that would appear in any of the networks. Based on this assessment, I

determined that correlation networks of $r \geq |0.99|$ indicated that further analysis beyond correlation mapping would be required to elucidate potential trends. While distinct trends were difficult to elucidate from the correlation mapping itself, it is clear that the relationships between metabolite abundance over time is drastically altered as the infection progresses, given the distinctly different maps generated at each time point.

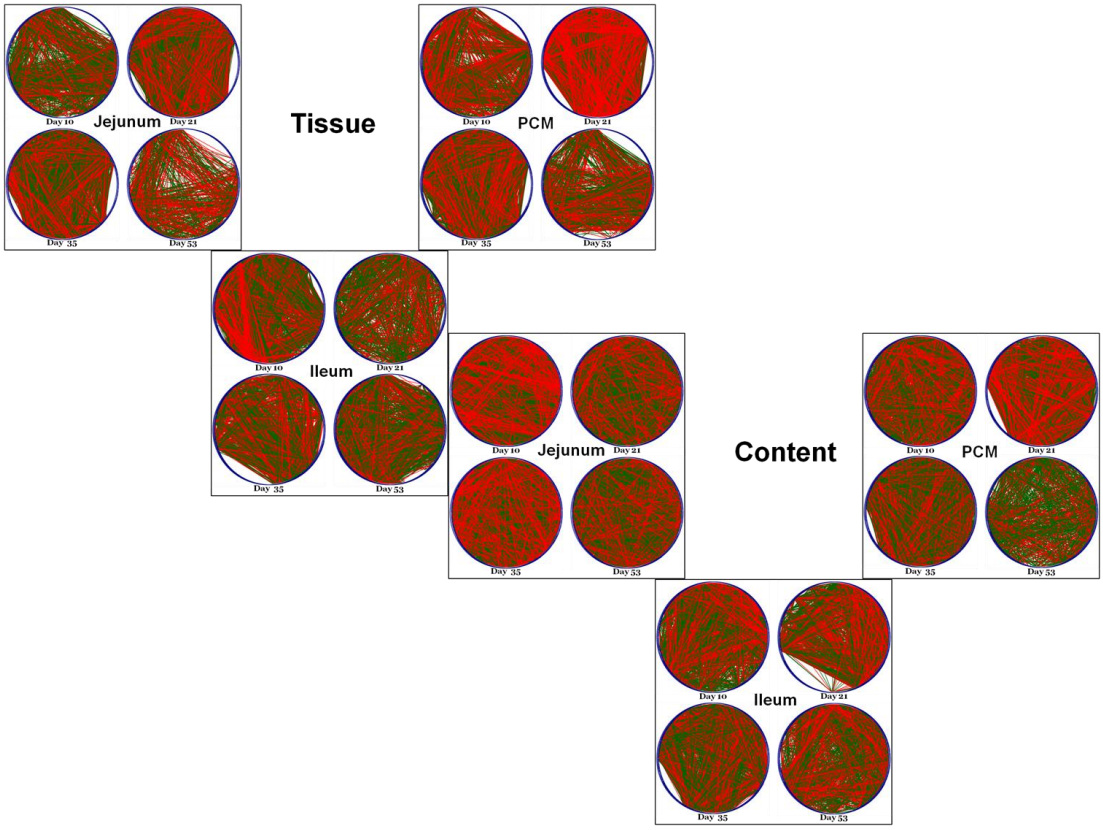


Figure 63. Intestinal tissue and content specific metabolite correlation networks for the time points: Day 10, Day 21, Day 35, and Day 53.
For each tissue and content type, Pearson's correlation coefficients were calculated for metabolites. Positive Pearson correlation values, $r \geq 0.99$, are depicted as a green line between metabolites, while negative correlations, $r \leq -0.99$, are depicted as a red line between metabolites. To facilitate comparison of the networks, metabolites are represented as a small, alpha numerically labeled circle. Their placement around the circumference of each network is fixed among the plots for each intestinal tissue and content. Each individual correlation network is representative of the following time points: Day 10 (top left), Day 21 (top right), Day 35 (bottom right), and Day 53 (bottom left). For all intestinal tissue and content, there is an enormous amount of positive and negative interrelationships occurring amongst the metabolites in each cohort. Increasing the stringency stepwise to $r \geq |1|$ significantly reduced the number of correlated metabolites in each cohort, but did not provide further insight. Based on the results presented, I can conclude that the metabolome is significantly altered as the disease progresses.

3. Metabolites Associated with Disease Progression

Worm maturation through the stages of infection leads to detectable alterations in the intestinal tissue and content metabolomes, as evidenced in all six of the intestinal

tissue and content samples. That is, stages 1, 2, and 3 are all clearly segregated and distinct in the PCA plot. Within every intestinal tissue and content examined, irrespective of the infection stage, numerous metabolites were found to be either unaffected, significantly increased, or significantly decreased by worm maturation. For each intestinal tissue and content sample, I compared the metabolome of all four time points culminating in 6 different comparisons (36 total comparisons for all four cohorts): Day 10 versus Day 21, Day 10 versus Day 35, Day 10 versus Day 53, Day 21 versus Day 35, Day 21 versus Day 53, and Day 35 versus Day 53. To focus my attention, I used the following cutoff values: $|\log_2 (\text{fold change}) \geq 1.5|$, $\text{p-value}_{adj} < 0.05$, and abundance value $> 2.4 \text{x} 10^4$, ensuring all significant features were above the baseline. This resulted in a grand total of 1477 significant metabolites, with the proximal colon mucosa and ileum tissue exhibiting the highest (611 molecular features) and second highest (334 molecular features) number of significant features, respectively. Following MS/MS identification, I identified a list of 7 metabolites (see Appendix 12 for the associated EICs and TICs) associated with intestinal motility and/or inflammatory pathways. Interestingly, all metabolites tabulated in Table 5 were exclusive to the stage 2 samples. The metabolites presented were all putatively identified via MS/MS and were undetected when compared to a concurrent analysis using the collected intestinal tissue and content from control (uninfected) pigs at the same time points. The top metabolites of interest are categorized into two distinct groups, serotonin metabolism and eicosanoid production, and are described in detail below.

Table 5. Metabolites exhibiting the greatest degree of change in all tissues and content and associated with pathways implicated in intestinal motility and/or inflammation.

Metabolite	Frequency: Day 10 (n = 6)	Frequency: Day 21 (n = 6)	Frequency: Day 35 (n = 8)	Frequency: Day 53 (n = 12)	Associated with:
Tryptophan	0	5	2	0	Serotonin Metabolism[134]
12-HETE	0	5	2	0	Eicosanoid Production[163,164]
6-Keto-prostaglandin F1a	0	5	3	0	Eicosanoid Production[163,164]
8-Isoprostane	0	5	5	0	Eicosanoid Production[163,164]
Niacin	0	6	8	0	Serotonin Metabolism[134]
5-Hydroxyindole acetic acid	0	3	8	0	Serotonin Metabolism[134]
15-deoxy-d-12,14-PGJ_2	0	5	7	0	Eicosanoid Production[163,164]
Criteria for Selection: Frequency ≥ 60%; $-\log_{10}$(p-value) ≥ 8; $\lvert\log_2$(Fold Change)$\rvert$ ≥ 1.5; Median abundance values ≥ 24000. All metabolites were identified via MS/MS (CID = 10 eV, 20 eV, 40 eV) using Metlin and the Human Metabolome Database (HMDB).					
n is equal to the number of samples in each cohort.					

1. Serotonin Metabolism

The analysis identified an increase in metabolites associated with serotonin metabolism (tryptophan, niacin, and 5-hydroxyindoleacetic acid). Approximately 95% of serotonin is produced in the intestines[165–167], where it aids in intestinal secretion and motility. Dysregulation of serotonin in the intestines can lead to diarrhea or constipation depending on its relative concentration in the body. In addition, serotonin can act as a pro-inflammatory molecule, where an increase in its bioavailability results in an increase in intestinal inflammation[165–167]. In the first step of the biosynthetic pathway, tryptophan, is converted into 5-hydroxy-L-tryptophan[134,166]. The results indicate a significant increase in the levels of tryptophan in the Day 21 samples (5 of 6 pigs) when compared to Days 10, 35, and 53. Conversely, the degradation of serotonin results in the production of 5-

hydroxyindoleacetic acid, a metabolite that is significantly increased in the Day 35 samples (8 of 8 pigs) when compared to Days 10, 21, and 53. Finally, there was a significant increase in niacin in the Day 21 (6 of 6) and Day 35 samples (8 of 8) when compared to Days 10 and 53. As niacin is generated via an alternative tryptophan degradation pathway, it is interesting to postulate that its appearance may indicate an alteration to potential serotonin levels[134,168]. Further investigation would be required to substantiate this claim.

2. Eicosanoid production

Eicosanoids (such as 12-HETE, 6-Keto-prostaglandin F1a, 8-isoprostane, and 15-deoxy-d-12,14 - PGJ2), another important class of intestinal health modulators[163,164], were significantly increased in Days 21 and 35 with respect to Days 10 and 53. While eicosanoids are commonly implicated in inflammation, they are also important for maintaining intestinal health by modulating epithelial cell proliferation/differentiation and inducing contractions and/or relaxation of the intestinal muscle[163,164]. Intriguingly, one of the metabolites identified, the prostaglandin 15-deoxy-d-12,14-PGJ_2 ($15dPGJ_2$), has been shown to decrease in individuals with Crohn's disease[169]. $15dPGJ_2$ has recently been described as an important inhibitor of intestinal epithelial cell proliferation and promoter of intestinal epithelial cell differentiation[169]. In addition, this prostaglandin has demonstrated anti-inflammatory properties by activating peroxisome proliferator-activated receptors (PPARs), transcription factors responsible for inhibiting activated T cell proliferation[170,171]. Given the potential of using *T. suis* as a treatment for patients

afflicted with Crohn's disease, this metabolite should undergo further study to determine if the increase of 15-deoxy-d-12,14-PGJ_2 can be attributed to symptom suppression.

Summary and Additional Examination

Worm maturation and disease progression has a profound effect on the intestinal tissue and content metabolome as depicted in Figure 62-64. According to the fold change analysis, the proximal colon mucosa and ileum are particularly sensitive to the transitions occurring between larval stages. Interestingly, metabolites associated with mammalian inflammation, including serotonin metabolism and eicosanoid production, were identified solely within the second stage of infection. Of those identified, the most notable was an increase in 15dPGJ_2, a prostaglandin which is known to be suppressed in individuals with Crohn's disease[169]. As this prostaglandin was present solely during the mid-stages of infection, its appearance and disappearance may potentially provide insight into the efficacy and longevity of this treatment. This metabolite, 15dPGJ_2, warrants further study as a potential therapeutic.

Investigation of the resistant versus susceptible phenotypic metabolome.

During the maturation to adult worms and late stages of infection, there is a development of two distinct phenotypes: those pigs who exhibit high worm burden (susceptible) and those who exhibit low worm burden (resistant)[172]. A similar occurrence was noted while collecting the samples. While many of the pigs experienced a high worm burden (approximately 1000 worms) during the late stages of infection, some were either devoid of worms or exhibited less than 60 worms. Throughout the course of infection,

these low worm burden pigs, however, still experienced symptoms indicating they were infected. Thus, it can be concluded that these pigs were able to clear the infection prior the worms reaching adulthood. To investigate this phenomenon, I performed a nonvolatile LC-QToF based analysis on the proximal colon mucosa and ileum tissue acquired from uninfected pigs and pigs exhibiting a high worm burden (susceptible) or a low worm burden (resistant).

Following the acquisition of the LC-QToF data, I extracted the molecular features using Agilent Technologies' MassHunter Qualitative Analysis software, complied the data set, and generated an all-inclusive spreadsheet for each tissue studied, containing a total of 99,873 molecular features. To filter out the low frequency metabolites, I restricted the analysis to include only those tissue-specific molecular features that appeared in a minimum of 3 of 5 pigs in any one of the three cohorts, significantly reducing the data to a total of 4562 molecular features (2496 in the PCM and 2066 in the ileum).

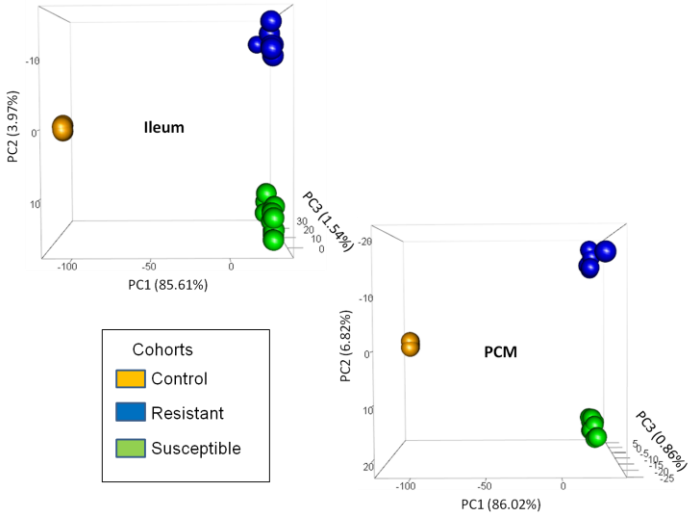


Figure 64. Focused intestinal tissue three-dimensional PCA examining the metabolomic alterations that occur as a consequence of worm burden.
A total of 24 pigs were distributed among three distinct cohorts (uninfected control, low worm burden (resistant), and high worm burden (susceptible)). The three dimensional PCA plot depicts the alterations in the metabolic profile obtained from each of the three different cohorts. Each sphere on the plot reflects the metabolome of the pig intestinal tissue sample. Spheres are colored as indicated in the figure legend. The analysis was confined to molecular features appearing in ≥25% and a minimum of 3 of 5 pigs in any one of the three pig cohorts.

Tissue-focused PCA plots (Figure 64) clearly differentiates the metabolome amongst the susceptible (high worm burden), resistant, and control (uninfected) samples. The first degree of separation occurs along principal component 1, separating the control from the infected (high and low worm burden) cohorts; explaining approximately 85% and 86% of the variance in the ileum and PCM tissues, respectively. Differentiation between the two infection phenotypes (high and low worm burden), occurs along principal component 2, however, cohort segregation is achieve by a substantially smaller degree of variation (approximately 4% and 7% in the ileum and PCM, respectively).

To further explore the metabolic variation, I then examined the relationships of the metabolites associated within each cohort using correlation networks ($r \geq |0.99|$,

Figure 65). The resulting correlation networks demonstrate the drastic alteration to the interrelationships between metabolites when comparing uninfected pigs with high worm burden (susceptible) and resistant pigs.

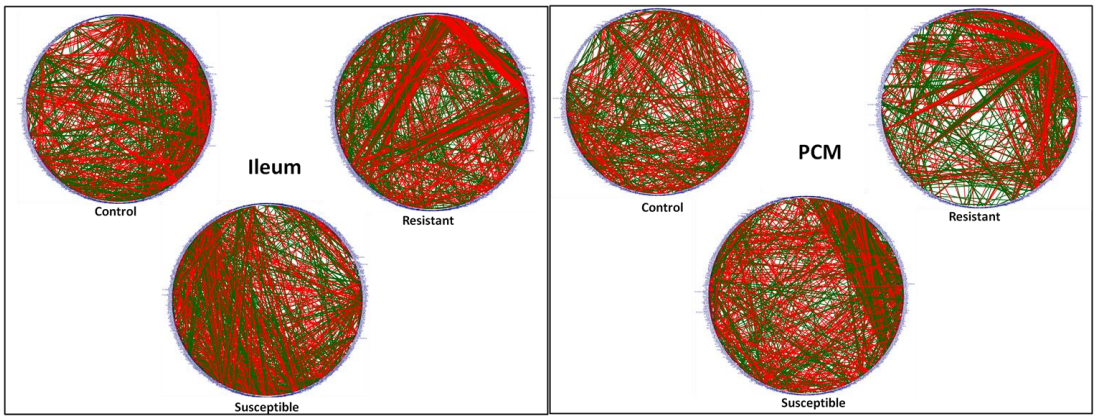


Figure 65. Intestinal tissue specific metabolite correlation networks for the two different worm burden phenotypes and uninfected pigs.
For each tissue and content type, Pearson's correlation coefficients were calculated for metabolites. Positive Pearson correlation values, $r \geq 0.99$, are depicted as a green line between metabolites, while negative correlations, $r \leq -0.99$, are depicted as a red line between metabolites. To facilitate comparison of the networks, metabolites are represented as a small alpha numerically labeled circle and their placement around the circumference of each network is fixed among the plots for each intestinal tissue. Each individual correlation network is a representative of the following phenotypes: uninfected (control, top left), low worm burden (resistant, top right), and high worm burden (susceptible, bottom middle). Based on the results presented, I can conclude that the metabolome is significantly altered due to worm burden levels.

To elucidate the potential metabolites responsible for these variations, fold change analysis was employed with the following cutoffs for inclusion: $|\log_2 (\text{fold change}) \geq 1.5|$, p-value cutoff of $\text{p-value}_{adj} < 0.05$, and metabolite abundance $> 2.4 \times 10^4$. The resulting filtration steps generated a list of 108 molecular features for the ileum (41 molecular features) and PCM (67 molecular features). However, upon close examination, it appears that many of these molecular features were only detected within a sub-population of each

worm burden phenotype. Those molecular features which were deemed cohort specific are not currently included in the Metlin or HMDB databases and thus, unidentified.

Based on the results of this tissue specific analysis, I can conclude metabolic variations exist between the resistant and susceptible phenotypes as evidenced in Figure 65 and 66. While the fold change analysis was inconclusive in determining the source of these variations, this analysis showed that there is a sub-population of sample within each phenotype, a phenomenon not readily observed within the PCA plot. While it appears that the alteration in metabolite abundance may not be factor in phenotype development, the correlation networks depicted grandiose alterations to the metabolic interactions occurring within each cohort. Thus, indicating that metabolomic variations between the two phenotypes may be a reflection of altered flux through metabolic pathways. Though this is speculative, further research should be performed to determine if any metabolic markers exist in other tissues, or if any genomic or proteomic information can be derived from the two phenotypes.

CONCLUSIONS

The goal of this research project was to develop and employ a metabolomics pipeline to visualize and compare the metabolomes of biological systems. Through the use of four different research investigations, I was able to refine the pipeline to incorporate the needs presented in each project. In Specific Aim 2.1, I used binary plots as an electronic noise to successfully differentiate mVOCs from biothreat agents, including those with kanamycin resistance, using hSPME as an extraction method. In Specific Aim 2.2, a GC-MS analysis was used to differentiate the VOCs exuded from the human fecal metabolome of healthy and alcoholic patients. This investigation revealed the analytical power of fecal metabolomics to identify alterations to the gastrointestinal tract as a result of chronic alcohol consumption. In Specific Aim 2.3, a LC-MS analysis was used to determine the cumulative effects of probiotic supplementation and dietary consumption on the Ossabaw pig. The analysis demonstrated the grandiose alterations to the global tissue metabolome and demonstrated the need for additional filtrations to properly analyze a multi-cohort and multi-tissue investigation. Finally, in Specific Aim 2.4, a LC-MS analysis was used to characterize the metabolic alterations induced by *T. suis* infection on the intestinal tract of pigs, as a function of the whipworm life cycle and the metabolic variations that can lead to a resistant phenotype. This study was the first introduction into a predominately targeted metabolomics analysis and demonstrated that

this current pipeline can be employed to both global and targeted metabolomics investigations.

Ultimately, this iterative approach to method development provided a metabolomics pipeline that incorporates many of the commonly used statistical tools while remaining adaptable to the multitude of different investigative techniques employed by the field. Although it is critical to state that each statistical analysis has its own limitations, a suitable combination of chromatographic and spectral analysis with the proper statistical tools can be employed for most metabolomic investigations.

APPENDIX 1 - GENERATING AN EXCEL SPREADSHEET

Generating a spreadsheet from LCMS Data

1. Raw data is opened in the Qualitative Analysis software.
2. Under the Method Explorer tab, click on the section called "Find Compounds"
   i. Select Find by Molecular Feature
   ii. Under the extraction tab, the target data type should state "Small molecules (chromatographic)"
   iii. Click on the play button (Find Compounds by Molecular Feature)
3. Highlight all the compounds identified
4. Return to the Method Explorer tab, click on the section called "Identify Compounds"
   i. Select Generate Formulas
      a. Under the charge state tab, the isotope model should state "Common organic molecules"
      b. Click on the play button (Generate Formulas from Compound)
5. Now you will need to export your file as a cef file.
   i. Click on File -> Export -> as CEF
6. Start Mass Profiler Professional
   i. i. Select 1 of the 3 choices: Create new project, Open existing project, Open recent project

ii. Create a new experiment
    a. Name the experiment
    b. Experiment type: Combined (Identified + Unidentified)
    c. Workflow type: Analysis: Significance Testing and Fold Change
iii. Select the Data files you would like to import (.cef files you just created)
iv. You may choose to reorder your samples
v. Add a parameter
    a. Name the parameter
    b. Parameter type: Non-Numeric
    d. Under Parameter values, double click and type in the name of the cohort for that sample -- continue this until all samples are in their respective cohorts
    e. click ok
vi. Under abundance filtering - select the minimum absolute abundance (default is 5000 counts)
vii. Under Normalization Algorithm - select none
viii. Under Baselining options - select none
ix. Click finish
x. Under Experiment Grouping highlight the parameter you created.
xi. Under Filter by Frequency
xii. Filtering Conditions - Click on re-run filter.
    a. Retain entities that appear in at least 80.0% of samples in at least 1 condition
xiii. Select IDBrowser Identification

a. Compound Selection -> Identify all compounds

b. Compound identification methods

1. Database Search

2. Molecular formula generator (MFG) for only unidentified compounds

c. Once completed, click on Save and Return

xiv. Click on Finish

7. Under the analysis folder, highlight "filtered by frequency" --> right click and select export list

i. Under interpretation select "All Samples"

ii. Select Raw Signal Values and Entitylist Data

iii. Columns to include: Compound Name, Frequency, Ionization Mode, Metlin ID, Retention Time, Mass, Alignment Value, Annotations, CAS number, ChEBI ID, Composite Spectrum.

iv. Click on OK and save your file

Generating a spreadsheet from AMDIS using GCMS Data

1. Open AMDIS

2. Click “File” and then “Open.” Load appropriate file.

    a. Tip: If the file looks like a folder, it is corrupted and AMDIS will not be able to analyze it.

3. Go under “Analyze” and click “Analyze GC/MS Data.” You will see a window popup that looks like the following (Figure 66):

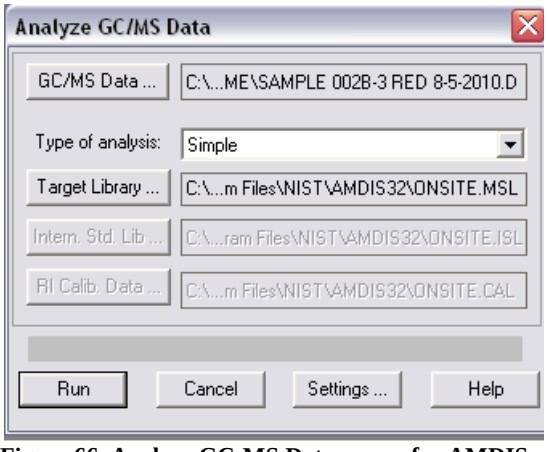


Figure 66. Analyze GC-MS Data screen for AMDIS.

4. Click “Settings” and check to make sure the settings match the following and click save:

    i. Identification Tab

a. Minimum Match Factor: 60

b. Only check off: "Multiple Identifications Per Compound"

c. Type of analysis: Simple

ii. Instrument Tab

a. Check off the Auto boxes for both Low M/Z and High M/Z

b. Threshold: Off

c. Scan Detection: High to Low

d. Instrument Type: Quadrupole

e. Data File Format: Agilent Files

iii. Deconvolution Tab

a. Component width: 12

b. Uncheck: omit m/z

c. Adjacent Peak Subtraction: One

d. Resolution: Medium

e. Sensitivity: Medium

f. Shape Requirements: Medium

iv. Library Tab

    a. MS libraries/RI data: Target Compounds Library

v. QA/QC Tab

    a. Uncheck Solvent Tailing

    b. Uncheck: Column Bleed

vi. Scan Sets Tab

    a. Number of sets: 0

5. Click “Run”

6. Go under “Analyze” and click “Search NIST Library.” Check to make sure the settings match the following (AMDIS will automatically substitute your file name for the one pictured here (Figure 67)):

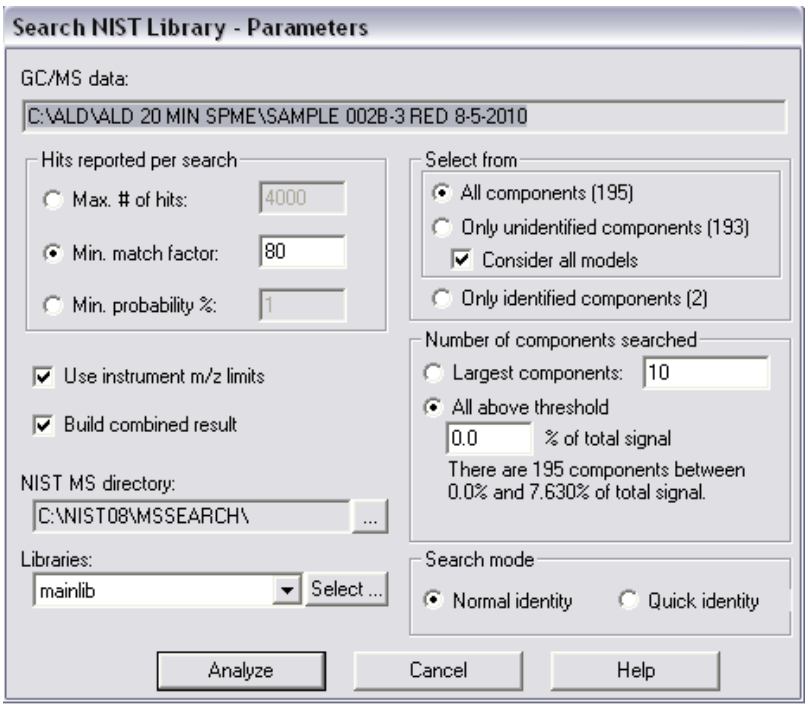


Figure 67. Search NIST Library - Parameters screen on AMDIS.

7. Click "Analyze"'

8. Go under "File" and click "Generate Report."

   i. **"Append to report file" will need to be unchecked every time you generate a report.**

   ii. Next to "Report File" there should be a file name. Click the […] next to the name, name your file appropriately, and click "Open."

9. Click "Generate"

APPENDIX 2 - METABOLOMICS ANALYSIS PIPELINE

What you will need:

1. Excel with an XLSTAT plug-in
2. R (can use either R or Rstudio)
    i. Libraries in R
        a. rgl - 3d PCAs
        b. pheatmap - heatmaps (dendrogram), correlation heatmap
        c. qgraph - visualize correlation network

Key Shortcuts

1. To select all of the numbers in a given area:
    i. CTRL+Shift +Directional Arrow
2. Jump back and forth on different sides of the data sheet
    i. CTRL+directional arrow
3. To keep the value in a cell fixed in an equation, use a dollar sign to keep the value for that cell fixed.
    i. =IF(C2<1,C$18,C2)

Data Preparation

2. Open the raw data and copy and paste the information into a second sheet.
    i. Label the first sheet as "infile"

ii. Label the second sheet as "frequency"

2. Create a distribution plot to visualize the metabolite's frequency

i. Go to the “infile", copy the data and paste it into the "frequency" sheet.

ii. Calculate the frequency for each cohort and the total frequency

Equation: =COUNTIF(____:___,">0")

the “___” is the cell values that need to be specified for the range; ex:

B2:B32 → tells Excel to refer to cells from B2 to B32

iii. Calculate the score value for each metabolite (Equation 1)

iv. Create histogram

a. Select data

b. Select “Column”

c. Right click on any bar in the graph to edit the data set under: “Select Data”

v. Determine your frequency cutoff and score value cutoff.

vi. Filter your data to reflect those cutoff values.

vii. Copy and paste the data into a new sheet called “Frequency Filtered”

3. Populate any missing values with the median value of that analyte for that cohort (Figure 68).

i. Calculate the median

ii. Open a new sheet and rename it "Median Replacement"

iii. In the sheet, use the equation:

=IF(ISBLANK(SHEET1cell), SHEET1medianvalue, SHEET1cell)

iv. Simply drag down to impute the values and drag across to complete the whole table

v. Repeat for any other cohort

SHEET1	A	B
1	Sample1	52701
2	Sample2	39045
3	Sample3	
4	Sample4	43600
5	Sample5	
6	Sample6	
7		
8	Median	43600

Insert Equation

SHEET2	A	B
1	Sample1	=IF(ISBLANK('SHEET1'!B1),'SHEET1'!B$8,'SHEET1'!B1)
2	Sample2	=IF(ISBLANK('SHEET1'!B2),'SHEET1'!B$8,'SHEET1'!B2)
3	Sample3	=IF(ISBLANK('SHEET1'!B3),'SHEET1'!B$8,'SHEET1'!B3)
4	Sample4	=IF(ISBLANK('SHEET1'!B4),'SHEET1'!B$8,'SHEET1'!B4)
5	Sample5	=IF(ISBLANK('SHEET1'!B5),'SHEET1'!B$8,'SHEET1'!B5)
6	Sample6	=IF(ISBLANK('SHEET1'!B6),'SHEET1'!B$8,'SHEET1'!B6)
7		
8	median	=MEDIAN(B1:B6)

Finished Product

SHEET2	A	B
1	Sample1	52701
2	Sample2	39045
3	Sample3	43600
4	Sample4	43600
5	Sample5	43600
6	Sample6	43600
7		
8	Median	43600

Figure 68. Example Median value.

4. Remove Outliers

ii. Calculate Mean for each cohort

iii. Calculate Median

iv. Calculate the outlier equation value (Equation 2)

a. Highlight the cells that don't meet this criteria (<1.5)

v. Replace the outlying value with the median and highlight that cell!

5. Highlight and Remove Duplicates - common in the AMDIS generated spreadsheets

i. Select one row of values at a time

ii. Go to the "Home" tab > "Conditional Formatting" tab in the top panel

iii. Select "Highlight Cells" > "Duplicate Values"

iv. Repeat for the next 2 rows

v. Select the whole cohort (including the metabolite labels) > sort data again, using the top row

vi. Select **all of the data for both cohorts**

vii. Sort the data by clicking on the icon "AZ" > "Custom Sort"

viii. Click on "Options" button on the bottom of the dialog box

ix. Select "sort right to left"

x. Sort according to the first sample's data (1st row)

xi. Whenever you find a replicate, cut the whole column of cells, and "**insert cut cells**" into a new sheet entitled: "removed replicates"

xii. Repeat until all replicates have been removed

xiii. Copy the completed table, and rename the copy "with replicates removed"

xiv. Make another copy of this sheet and rename: "Standard Values"

6. Standardize the data and save it as a .csv (Equation 3)(Figure 69)

i. Calculate the mean for that metabolite

ii. Calculate the standard deviation for that metabolite

iii. Standardize the value

iv. Copy all of the newly calculated standard values with their ID and metabolite names into a new sheet entitled: “Standard Values”

	A	B	C	D	E
1	Sample	Label	Metabolite name 1	Metabolite name 2	Metabolite name 3
2	A	Cohort 1	3418090	1805186	406468
3	B	Cohort 1	2006374	8247583	576579
4	C	Cohort 1	867217	3245076	746690
5	D	Cohort 2	928189	2159001	990743
6	E	Cohort 2	17199488	2798675	144010
7	F	Cohort 2	769733	3438348	1837476
8					
9	Average		=AVERAGE(C2:C7)	=AVERAGE(D2:D7)	=AVERAGE(E2:E7)
10	Standard Deviation		=STDEV(C2:C7)	=STDEV(D2:D7)	=STDEV(E2:E7)
11					
12	Standard Values		Metabolite name 1	Metabolite name 2	Metabolite name 3
13	A	Cohort 1	1.838218437	-0.76932049	-0.637739
14	B	Cohort 1	0.494883712	1.968255287	-0.350124
15	C	Cohort 1	-0.589094336	-0.15746624	-0.062509
16	D	Cohort 2	-0.531075723	-0.61897344	0.3501238
17	E	Cohort 2	-0.531075723	-0.34715619	-1.08149
18	F	Cohort 2	-0.681856367	-0.07533893	1.7817374

Figure 69. Z-score standardization.

XLSTAT Analysis

Can use XLSTAT plugin for Excel or R

PCA

1. Turn on the XLSTAT plugin

2. In the first toolbar select “Analyzing Data”
3. In the second toolbar select “Principal Component Analysis (PCA)”
4. Select the “Observation/Variable Tables”
   i. Select all of the Metabolite names and values
5. Select “Observation Labels”
   i. Select from “Sample” to the last patient ID
6. Make sure that:
   i. “Sheet” is selected
   ii. PCA type is: “Pearson (n)"
   iii. Under the “Outputs” tab > Significance level is “5”
   iv. Under the “Charts” tab > under the “Biplots” tab > make sure all are selected
7. The new PCA should appear in the “PCA” tab

Squared Cosines Table

*Note: you may want to look at F1, F2, and F3 values separately.

1. Refer to the newly generated “PCA” sheet and look at the “Squared Cosines of Variables”
2. Copy the F1-F3 values and paste them into a new sheet entitled: “Squared Cosines”
3. Next to this new column of data, calculate the sum of the F1-3 values

   Equation: =SUM(____:_____)
4. Copy the entire data table (metabolite names, F1, F2, F3 values, and the SUM), and paste it in the next column
5. Then select all of the data in this new table, and Sort the data by SUM

i. Go to the “Sort” button (AZ) in the top menu bar

ii. Select “Custom Sort”

iii. Sort according to the column “Sum”, and set the values from: “Largest to Smallest”

6. Then select the top 5 metabolites found in the table, and simply paste them on the side along with their Sum (F1-3) values

R Analysis

After opening R, set your working directory. You may do this by the simple command:

setwd("directorymap")

example:

setwd("C://Users/Lab/Rfiles")

You may also do this by clicking on File -> change dir... Then click on the file you want and hit ok!

Coding PCA plots in R

1. Open your file.

data <- read.csv ("filename.csv", header = T, row.names = 1)

*note: you will want to put in your file name in the " " (*don't forget the .csv*), header will be the first row of your excel file denoting sample and metabolite names, row.names is when the first column of your spreadsheet has sample names

2. To examine your data:

head(data)

3. Calculate the principal components.

example: **pca <- prcomp (na.omit(data[1:20,], scale = T))**

*note: you will not be able to set scale to true if you have any metabolite columns all labeled as 1. [1:20,] is the row numbers (*in this example I have 20 rows*). If you want column numbers it would be [,1:20] -- note position of the comma. na.omit removes any blank cells. Princomp CANNOT be used in this case since there are 20 rows and 3823 columns. Princomp can ONLY be used if columns are smaller than rows.*

4. Pull up the PCA output.

summary(pca)

5. Plot 2d plot of pc's 1 and 2.

plot(pca$x, xlab = "PC1 (%)", ylab= "PC2 (%)")

*note: do not forget to put the actual percentages in the labels

6. Color the points -- color coding the cohorts.

points(pca$x[1:5,], col = "red")

*note: repeat for all rows changing the color for each cohort

7. f desired, add a legend.

legend(locator (1), title ="cohorts", c("cohort1","cohort2"), fill = c("red", "blue")

*note: locator(1) allows you to click on where the legend goes. c("cohort1") refers to the variables that are colored, and fill allows you to colors the boxes for those variables

3D PCA plot

Make sure to install your package - you will want rgl - this one allows you to rotate your plot!

1. Open rgl.

library(rgl)

2. Input your PC data.

ttt = textConnection("

sample,x,y,z

1, value, value, value

2, value, value, value

3, value, value, value

4, value, value, value

5, value, value, value

6, value, value, value

")

*note: To get the numbers (here labeled as value) listed below, you will first want to calculate the pca values like you would create the 2d plot. Then type head (pca) or what you saved your pca values as -- x y z refers to pc1, pc2, and pc3 respectively. Sample refers to the sample used - you will need this if you want to label your points later on! You will need to actually put a number where it says value - do not leave it saying value!

3. Set the colors.

colors = c('orange','orange','orange','blue','blue', 'blue')

4. Read the table line by line, column by column.

pca<-read.table(ttt, header = TRUE, sep = ",")

5. Plot the graph - you will need to put the actual percentages in the labels.

p3d<-plot3d(pcax, pcay, pca$z, xlab="PC1 (%)", ylab="PC2 (%)", zlab="PC3 (%)", type="s", size=3, col = colors)

6. If desired, label the points.

p3d<-text3d(pcax,pcay,pcaz,pcasample, adj=1.5, font = 2)

*note: adj relates to the position of the labels where 0.5 is center, font refers to the style. font = 1 (standard); font = 2 (bold); font = 3 (italics); font = 4 (bold and italics).

7. Save your graph.

rgl.snapshot ('filename.png', fmt = "png", top ="TRUE")

*note: You can do this multiple times at multiple angles just change the filename. rgl will only recognizes png.

Pairs plot

1. Color based on samples (assuming the sample names are going down a column).

cols <-character(nrow(data))

*note: Here data represents what you saved your table as it was read into R

2. Set up colors (you can use whatever colors you want, just keep it consistent) based on the row number.
   i. Set up a case where the point(s) do not fall into the described row numbers.

      cols[] <- "blue"

   ii. Describe the row numbers and setting them to equal a color.

      cols[2:16] <-"red"

      cols[17:35]<-"green"

*note: No "," is needed after the row numbers since at the beginning R was told that only rows will be used (eg. normally [2:17,] is used to describe a row in the matrix)

3. Plot the pairs plot.

pairs(pca$x[,1:5], pch = 16, col=cols)

*note: This assumes that the pca calculation was performed in R and resulting value had a scalar variable of pca. In addition, [,1:5] states that pcs 1 -> 5 will be plotted. Finally, pch describes the point's shape and cex describes the points size. Please see http://www.statmethods.net/advgraphs/parameters.html for more information on which number goes with which shape!

Saving data to a .csv file

example: correlation matrix

1. Read in the csv file.

data <- read.csv("filename.csv", header = TRUE, row.names = 1)

*note: Make sure you have set your working directory so R knows where to look!

2. Calculate the correlation.

q <- cor(data)

q <- cor(data, method = "pearson")

*note: This defaults to pearson method but you may choose the method (spearman, kendal or pearson).

3. To see the output in the R command line:

head(q)

4. Write the data out as a csv file.

write.csv(q, file = "correlated.csv")

Generating a Dendrogram

1. Read in the csv file.

data <- read.csv("filename.csv", header = TRUE, row.names = 1)

*note: Make sure you have set your working directory so R knows where to look!

2. Open pheatmap.

library(pheatmap)

3. Perform an unsupervised hierarchical clustering analysis. This will create a heatmap using the data.

pheatmap(as.matrix(data), color = colorRampPalette(rev(c("#FF0000", "#000000", "#00FF00")))(100), cluster_rows = TRUE, cluster_cols = TRUE, clustering_distance_rows = "correlation", clustering_distance_cols ="correlation", clustering_method = "ward", width = "48", height = "48", fontsize_col="2", fontsize_row="6")

*note: There are other clustering methods to use. Different examples of the methods are shown below and should be tried. The changes are highlighted.

Clustering Distance Method: Euclidean

pheatmap(as.matrix(data), color = colorRampPalette(rev(c("#FF0000", "#000000", "#00FF00")))(100), cluster_rows = TRUE, cluster_cols = TRUE, clustering_distance_rows = "euclidean", clustering_distance_cols ="euclidean",

clustering_method = "ward", width = "48", height = "48", fontsize_col="2", fontsize_row="6")

Clustering Distance Method: Maximum

pheatmap(as.matrix(data), color = colorRampPalette(rev(c("#FF0000", "#000000", "#00FF00")))(100), cluster_rows = TRUE, cluster_cols = TRUE, clustering_distance_rows = "maximum", clustering_distance_cols ="maximum", clustering_method = "ward", width = "48", height = "48", fontsize_col="2", fontsize_row="6")

Clustering Distance Method: Manhattan

pheatmap(as.matrix(data), color = colorRampPalette(rev(c("#FF0000", "#000000", "#00FF00")))(100), cluster_rows = TRUE, cluster_cols = TRUE, clustering_distance_rows = "manhattan", clustering_distance_cols ="manhattan", clustering_method = "ward", width = "48", height = "48", fontsize_col="2", fontsize_row="6")

Clustering Distance Method: Canberra

pheatmap(as.matrix(data), color = colorRampPalette(rev(c("#FF0000", "#000000", "#00FF00")))(100), cluster_rows = TRUE, cluster_cols = TRUE, clustering_distance_rows = "canberra", clustering_distance_cols ="canberra",

clustering_method = "ward", width = "48", height = "48", fontsize_col="2", fontsize_row="6")

Clustering Distance Method: Binary

pheatmap(as.matrix(data), color = colorRampPalette(rev(c("#FF0000", "#000000", "#00FF00")))(100), cluster_rows = TRUE, cluster_cols = TRUE, clustering_distance_rows = "binary", clustering_distance_cols ="binary", clustering_method = "ward", width = "48", height = "48", fontsize_col="2", fontsize_row="6")

Clustering Distance Method: Minkowski

pheatmap(as.matrix(data), color = colorRampPalette(rev(c("#FF0000", "#000000", "#00FF00")))(100), cluster_rows = TRUE, cluster_cols = TRUE, clustering_distance_rows = "minkowski", clustering_distance_cols ="minkowski", clustering_method = "ward", width = "48", height = "48", fontsize_col="2", fontsize_row="6")

Creating a Correlation Network

1. Read in the csv file.

data <- read.csv("filename.csv", header = TRUE, row.names = 1)

*note: Make sure you have set your working directory so R knows where to look!

2. Open qgraph to generate the correlation network.

library(qgraph)

3. Plot your correlation network.

   b<-qgraph(cor(data), minimum = 0.9, layout = "spring", label.cex = 0.75, label.scale = FALSE, esize = 3, borders = FALSE, label.color = "navy")

*note: minimum = the min correlation value you want plotted; layout - can be either circular which will place the nodes in a circle and link each node with the same distance or spring which plots the nodes and the distance between the nodes is determined by the weight of the edges, defaults to circular; label.cex - this is the size of your node's label; label.scale - determines whether or not your label is scaled to the node's size - defaults to true; esize - the size of the edges (lines); borders - whether or not you want borders around your nodes - defaults to true; label.color - color of your label. defaults to black

Fold Change

1. Copy the data from the table: "with replicates removed", into a new sheet entitled: "Fold Change"
2. If a cohort has a metabolite that is completely missing from all the samples of the cohort, replace the missing values with a value of 1.
3. Then calculate:
   i. Median of Cohort 1

Equation: =MEDIAN(__:__)

ii. Median of Cohort 2

Equation: =MEDIAN(__:__)

iii. Fold Change

Equation: =(Median Infected values/Median Healthy values)

iv. Fold Change of log base 2

Equation: =LOG(Fold Change value,2)

P Values

1. Create a new sheet entitled: “P-Values”.
2. Copy the data from the file: "with replicates removed" and pass it into the new sheet
3. Beneath each column enter the following formula:

   Equation: =XLSTAT_TTest(C3:C13,C14:C28,,0,TRUE,TRUE)
4. Then, to be sure the values remain the same when XLSTAT is off, copy the values generated.
5. Use the function: "Paste Special" to insert the values into this new line.

LC-MS Identification

1. Retrieve the Retention time, and Composite Spectrum data for each of your top metabolites. (This data should be found within your original raw spreadsheet that you obtained from MPP).
2. Open the corresponding Raw Chromatogram files for 10 eV, 20 eV, and 40 eV as it relates to your samples.
3. Scan the files for your top metabolites using the retention time and their m/z values.
   i. Once you have located your compound, copy the m/z data which includes the mass and their intensities. If you cannot find your compound, denote that on the spreadsheet as the sample may need to run again.
6. Repeat this process for 10, 20, and 40 eVs for each metabolite.

7. Once you have your m/z and intensity table set up, go to Metlin (https://metlin.scripps.edu/index.php) and hover over Metlin, and click on MS/MS spectrum match. You may also want to use the Human Metabolome Database (HMDB) (http://www.hmdb.ca/).
    i. In Metlin, where it says peaks, copy and paste your m/z, intensity data there. Include all data above 5%. Your precursor ion tolerance = 100ppm, and your m/z tolerance = 0.1Da. Set to collision energy to the correct one that corresponds to your data and make sure to select the appropriate mode. Under precursor m/z use the value that you found while scanning the chromatogram. Finally, select Find Metabolites.

*note: The Metlin Scoring is an X-Rank Algorithm[173]. This algorithm first sorts the peak intensity of the spectrum and then establishes a correlation between the two spectrums. It then computes the probability that a rank from an experimental spectrum matches a rank from a reference library spectrum.

Roc curve

1. Use GraphPad Prism to make the Roc curve.

2. When you first open graph pad, under New table & graph, click on Column, and pick the scatter plot option.

    i. Then click create.

3. In your data table separate the data of your metabolite of interest into different columns: Ex Treatment and Control.

Make sure to label your metabolite on the left hand side of the page.

4. At the top of the page, under the analysis, click on “= Analyze". An Analyze Data page will open up.

5. Under Column analyses, click on ROC Curve.

6. On the right, make sure all boxes next to your cohorts are checked.

7. Click OK.

8. The next page that comes up is Parameters: ROC Curve.

    i. Make sure to double check that your patient and control values are set right, to match your data.

    ii. Set the confidence interval to 95%.

    iii. Report result as: Fraction.

    iv. New graph: check both boxes for line of identity and create a new graph of results.

    v. Click OK.

9. Your graph should have been created now.

    i. On the left, under Results you can look at the area under the curve.

    ii. On the left, under Graphs, you can see the ROC curve of your data.

10. To get rid of the data points right click on a data point > format entire data set>symbol size>0.

11. Double click on the “Y title” to change it to “Sensitivity.

*Instructions to superimpose 2 or more metabolites of interest.

12. Make a copy of the ROC curve graph one of your top metabolites of interest : Right click on the name on the left of the page and click duplicate, make sure to rename it.

13. At the top of the page, under the “Change” tab click on “Format axes (range, custom ticks, gridlines etc).

    i. Click on the “Left Y axis” tab.

    ii. Under “range”, make sure the maximum is 1.0, then click ok.

14. At the top of the page under “Change” click on “Add or remove data sets, and change their front to back or left to right order .”

    i. Go to the “Data Sets on Graph” tab.

    ii. On the right hand side click on “Add”.

    iii. Scroll down to the “Roc of [metabolite] and click on the sensitivity of the [metabolite] to add it to the curve. Do this until you have added all the metabolites you need.

    iv. Click OK.

15. Make a key to your graph:

    i. On the right hand side, replace the “Sensitivity” labels with the names of your metabolites.

    ii. To change the color of each one, right click on a data point > format entire data set> line/curve color> pick a color.

    iii. To get rid of the shapes on the line right click on a data point> format entire data set> shape size> 0.

16. Make a new column in your Excel Spreadsheet titled "Total Data Analysis" called ROC Curve Values and place the values into your spreadsheet.

APPENDIX 3 - SCILAB SCRIPT FOR BINARY PLOTS

```
M = [ insert your binary matrix here];
M=M*(color("white")-color("black"))+color("black");
 Matplot(M)
```

Example:

```
M = [0 0 0 0 0;
0 0 0 0 0;
1 0 1 0 1;
1 0 1 0 1;
0 0 0 0 1;
1 0 1 0 1;
0 0 0 0 0;
0 0 0 0 0;
1 0 1 0 1;
1 0 1 0 1;
1 0 1 0 1;
0 0 0 0 0;
0 0 0 0 0;
0 0 0 0 0;
0 0 0 0 0;
1 0 1 0 1;
0 0 0 0 0;
0 0 0 0 0;
0 0 0 0 0;
1 0 1 0 1;
0 0 0 0 0;
0 0 0 0 0;
0 0 0 0 0;
1 0 1 0 1;
];

M=M*(color("white")-color("black"))+color("black");
Matplot(M)
```

APPENDIX 4 - PERL SCRIPT FOR AUTOMATED BINARY PLOTS

The script below will take in binary results for multiple fibers, combine the results, compare those results it to a blank subtracting out peaks attributed to a blank, and generate an image. The script can be changed to take in any filename, increase or decrease the number of fibers, and increase or decrease the retention time windows.

```
#!/usr/bin/perl

use 5.12.3;
use strict;
use warnings;
use diagnostics;

# Open csv file, compare to blank, set window for retention time,
# subtract the blank --
# If the blank peak was there then set = 0 if not then =1.
# 1 means a bar and # 0 means no bar

# parsing csv

# Fiber 1

# Sample
my $file1 = 'filename_for_fiber1_sample.csv';
my @file1 = ();
open (my $fh1, '<', $file1) or die "Can't read sample file!: $!\n";
while (my $line = <$fh1>) {
  chomp $line;
  my @fields = split ",", $line;
  push @file1, $fields[1];
}

#Blank
my $file2 = 'filename_for_fiber1_blank.csv';
my @file2 = ();
open (my $fh2, '<', $file2) or die "Can't read blank file!: $!\n";
while (my $line = <$fh2>) {
  chomp $line;
  my @fields = split ",", $line;
```

```
  push @file2, $fields[1];
}

# Fiber 2

# Sample
my $file3 = 'filename_for_fiber2_sample.csv';
my @file3 = ();
open (my $fh3, '<', $file3) or die "Can't read sample file!: $!\n";
while (my $line = <$fh3>) {
  chomp $line;
  my @fields = split ",", $line;
  push @file3, $fields[1];
}

# Blank
my $file4 = 'filename_for_fiber2_blank.csv';
my @file4 = ();
open (my $fh4, '<', $file4) or die "Can't read blank file!: $!\n";
while (my $line = <$fh4>) {
  chomp $line;
  my @fields = split ",", $line;
  push @file4, $fields[1];
}

# Fiber 3

# Sample
my $file5 = 'filename_for_fiber3_sample.csv';
my @file5 = ();
open (my $fh5, '<', $file5) or die "Can't read sample file!: $!\n";
while (my $line = <$fh5>) {
  chomp $line;
  my @fields = split ",", $line;
  push @file5, $fields[1];
}

# Blank
my $file6 = 'filename_for_fiber3_blank.csv';
my @file6 = ();
open (my $fh6, '<', $file6) or die "Can't read blank file!: $!\n";
while (my $line = <$fh6>) {
  chomp $line;
  my @fields = split ",", $line;
  push @file6, $fields[1];
}

# Fiber 4

# Sample
my $file7 = 'filename_for_fiber4_sample.csv';
my @file7 = ();
open (my $fh7, '<', $file7) or die "Can't read sample file!: $!\n";
while (my $line = <$fh7>) {
```

```
  chomp $line;
  my @fields = split ",", $line;
  push @file7, $fields[1];
}

# Blank
my $file8 = 'filename_for_fiber4_blank.csv';
my @file8 = ();
open (my $fh8, '<', $file8) or die "Can't read blank file!: $!\n";
while (my $line = <$fh8>) {
  chomp $line;
  my @fields = split ",", $line;
  push @file8, $fields[1];
}

# Printing results
print "array size = ", @file1. "\n\n";
print "Sample Fiber 1: \n";
print "@file1 \n\n";

print "array size = ", @file2. "\n\n";
print "Blank Fiber 1: \n";
print "@file2 \n\n";

print "array size = ", @file3. "\n\n";
print "Sample Fiber 2: \n";
print "@file3 \n\n";

print "array size = ", @file4. "\n\n";
print "Blank Fiber 2: \n";
print "@file4 \n\n";

print "array size = ", @file5. "\n\n";
print "Sample Fiber 3: \n";
print "@file5 \n\n";

print "array size = ", @file6. "\n\n";
print "Blank Fiber 3: \n";
print "@file6 \n\n";

print "array size = ", @file7. "\n\n";
print "Sample Fiber 4: \n";
print "@file7 \n\n";

print "array size = ", @file8. "\n\n";
print "Blank Fiber 4: \n";
print "@file8 \n\n";

# Initialize empty x and y arrays.
my @x_array = ();
my @y_array = ();
my @y_array2 = ();
my @y_array3 = ();
my @y_array4 = ();
```

```
# Go over each time interval
for (my $x = 0.00001; $x <= 30.1; $x += .1) {
  my $diff = 0;
  # Set the diff based on the time interval.
  # Adjust to what I want as parameters.
  if ($x < 5) {
    $diff = 0.1;
  } elsif ($x < 10) {
    $diff = 0.1;
  } elsif ($x < 15) {
    $diff = 0.1;
  } elsif ($x < 20) {
    $diff = 0.1;
  } elsif ($x < 25) {
    $diff = 0.1;
  } else {
    $diff = 0.1;
  }

  my $y = 0;
  # Make sure there are data points to compare between each file.
  if (scalar(@file1) > 0) {
    # Find the closest match between the files.
    # Look only at the first item in file1;
       # Check to see if the time is within the interval.
    if ($file1 [0] < $x + 0.1) {
      # Check for a comparable time
      # Go over each item in other 2 files
      my $min = $diff;
      $y = 1;
      # For loop for each item in file 2
      for (my $index_in_file2 = 0; $index_in_file2 <
             scalar(@file2); $index_in_file2 += 1) {
        my $dt = abs ($file1 [0] - $file2 [$index_in_file2]);
        # Want to compare to current min diff
        if ($dt < $min) {
          $y = 0;
          }
      }
      # The following lines can be used if your File 3 should be
      # subtracted of your File 1. These lines can be
      # included in any fiber data.
            # Looking through file 3 to find comparable times
      #for (my $index_in_file3 = 0; $index_in_file3 <
              scalar(@file3); $index_in_file3 += 1) {
      #  my $dt = abs ($file3 [0] - $file4 [$index_in_file3]);
        # Want to compare to current min diff
      #  if ($dt < $min) {
      #    $y = 0;
      #    }
      #}
      # If comparable = 0 not comparable = 1
      #remove time in this interval from file1
```

```
      shift (@file1);
    }
  }

  my $y2 = 0;
  # Make sure there are data points to compare between each file.
  if (scalar(@file3) > 0) {
    # Find the closest match between the files.
    # Look at only the first item in file3;
       # Check if the time is within the interval.
    if ($file3 [0] < $x + 0.1) {
      # Check for comparable time
      # Go over each item in other 2 files
      my $min = $diff;
      $y2 = 1;
      # For loop for each item in file 4
      for (my $index_in_file4 = 0; $index_in_file4 <
             scalar(@file4); $index_in_file4 += 1) {
        my $dt = abs ($file3 [0] - $file4 [$index_in_file4]);
        # Want to compare to current min diff
        if ($dt < $min) {
          $y2 = 0;
        }
      }

      # If comparable = 0 not comparable = 1
      # Remove time in this interval from file1
      shift (@file3);
    }
  }

  my $y3 = 0;
  # Make sure there are data points to compare between each file.
  if (scalar(@file5) > 0) {
    # Find the closest match between the files.
    # Look at only the first item in file5; check if the time is
# within the interval.
    if ($file5 [0] < $x + 0.1) {
      # Check for comparable time
      # Go over each item in other 2 files
      my $min = $diff;
      $y3 = 1;
      # For loop for each item in file 6
      for (my $index_in_file6 = 0; $index_in_file6 <
             scalar(@file6); $index_in_file6 += 1) {
        my $dt = abs ($file5 [0] - $file6 [$index_in_file6]);
        # Want to compare to current min diff
        if ($dt < $min) {
          $y3 = 0;
        }
      }

      # If comparable = 0 not comparable = 1
      # Remove time in this interval from file1
```

```
      shift (@file5);
    }
  }

  my $y4 = 0;
  # Make sure there are data points to compare between each file.
  if (scalar(@file7) > 0) {
    # Find the closest match between the files.
    # Look at only the first item in file7; check if the time is
# within the interval.
    if ($file7 [0] < $x + 0.1) {
      # Check for comparable time
      # Go over each item in other 2 files
      my $min = $diff;
      $y4 = 1;
      # for loop for each item in file 8
      for (my $index_in_file8 = 0; $index_in_file8 <
             scalar(@file8); $index_in_file8 += 1) {
        my $dt = abs ($file7 [0] - $file8 [$index_in_file8]);
        #   Want to compare to current min diff
        if ($dt < $min) {
          $y4 = 0;
        }
      }

      # If comparable = 0 not comparable = 1
      # Remove time in this interval from file1
      shift (@file7);
    }
  }
  push (@x_array, int($x));
# Placing int in front of $x removes floating points
  push (@y_array, $y);
  push (@y_array2, $y2);
  push (@y_array3, $y3);
  push (@y_array4, $y4);
}

my @data = (\@x_array, \@y_array, \@y_array2, \@y_array3, \@y_array4);

use GD::Graph::hbars;
use GD::Text;

#create new image
my $graph = GD::Graph::hbars->new(400,800);

#setting parameters of graph
$graph->set(
  title => 'Sample's Binary Plot',
  x_label => 'Retention Time (min)',
  x_label_position => 1/2,
  x_label_skip => 10,
  y_tick_number => 0,
  dclrs => [ qw (white) ],
```

```
  boxclr => [qw (black)],
  labelclr => [qw (black)],
  axislabelclr => [qw (black)],
  textclr => [qw (black)],
) or die $graph->error;

my $font_spec = "./Dustismo_Sans";

$graph->set_title_font($font_spec, 5);

$graph->set_x_label_font($font_spec, 18);

my $gd = $graph->plot(\@data) or die $graph->error;

# Saving image
open(IMG, '>sample_binary.png') or die $!;
binmode IMG; # Tell the computer it is receiving binary data
print IMG $gd->png; # Generating a png image from the data and
# printing it to the png file handle
close IMG;

exit;
```

APPENDIX 5 - PUBLISHED WORK

Specific Aim 2.1 - Alterations to the human fecal metabolome due to alcohol consumption

Paper I: Alcohol Induced Alterations to the Human Fecal VOC Metabolome Couch RD, **Dailey A**, Zaidi F, Navarro K, Forsyth CB, et al. PLoS One. 2015; 10 (3):e0119362. doi: 10.1371/journal.pone.0119362

PLOS | ONE

RESEARCH ARTICLE

Alcohol Induced Alterations to the Human Fecal VOC Metabolome

Robin D. Couch[1]*, Allyson Dailey[1], Fatima Zaidi[1], Karl Navarro[1], Christopher B. Forsyth[2,3], Ece Mutlu[2], Phillip A. Engen[2], Ali Keshavarzian[2,4,5,6]

1 Department of Chemistry and Biochemistry, George Mason University, Manassas, Virginia, United States of America, **2** Department of Medicine, Division of Digestive Diseases and Nutrition, Rush University Medical Center, Chicago, Illinois, United States of America, **3** Department of Biochemistry, Rush University Medical Center, Chicago, Illinois, United States of America, **4** Department of Pharmacology, Rush University Medical Center, Chicago, Illinois, United States of America, **5** Department of Molecular Biophysics and Physiology, Rush University Medical Center, Chicago, Illinois, United States of America, **6** Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Utrecht University, Utrecht, The Netherlands

* rcouch@gmu.edu

Citation: Couch RD, Dailey A, Zaidi F, Navarro K, Forsyth CB, Mutlu E, et al. (2015) Alcohol Induced Alterations to the Human Fecal VOC Metabolome. PLoS ONE 10(3): e0119362. doi:10.1371/journal.pone.0119362

Academic Editor: Markus M. Heimesaat, Charité, Campus Benjamin Franklin, GERMANY

Received: September 3, 2014

Accepted: January 13, 2015

Published: March 9, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was supported by National Institutes of Health grant 1RC2AA019405. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Studies have shown that excessive alcohol consumption impacts the intestinal microbiota composition, causing disruption of homeostasis (dysbiosis). However, this observed change is not indicative of the dysbiotic intestinal microbiota function that could result in the production of injurious and toxic products. Thus, knowledge of the effects of alcohol on the intestinal microbiota function and their metabolites is warranted, in order to better understand the role of the intestinal microbiota in alcohol associated organ failure. Here, we report the results of a differential metabolomic analysis comparing volatile organic compounds (VOC) detected in the stool of alcoholics and non-alcoholic healthy controls. We performed the analysis with fecal samples collected after passage as well as with samples collected directly from the sigmoid lumen. Regardless of the approach to fecal collection, we found a stool VOC metabolomic signature in alcoholics that is different from healthy controls. The most notable metabolite alterations in the alcoholic samples include: (1) an elevation in the oxidative stress biomarker tetradecane; (2) a decrease in five fatty alcohols with anti-oxidant property; (3) a decrease in the short chain fatty acids propionate and isobutyrate, important in maintaining intestinal epithelial cell health and barrier integrity; (4) a decrease in alcohol consumption natural suppressant caryophyllene; (5) a decrease in natural product and hepatic steatosis attenuator camphene; and (6) decreased dimethyl disulfide and dimethyl trisulfide, microbial products of decomposition. Our results showed that intestinal microbiota function is altered in alcoholics which might promote alcohol associated pathologies.

Introduction

Clinical and experimental data have demonstrated that the intestinal microbiota plays a major role in maintaining a healthy state, while an abnormal bacterial community can contribute to

the development/progression of various pathological diseases [1]. It is also well established that diet impacts the intestinal microbiota composition and diversity [2]. Alcohol is a major component of diet in Western societies, which could potentially impact the intestinal microbiota community. Several studies, including our own, have shown that excessive alcohol consumption impacts the intestinal microbiota composition in both rodent models and humans, causing disruption of intestinal microbiota homeostasis (dysbiosis) [3–6]. The changes in the intestinal microbiota community may be a potential co-factor for the development of tissue injury and organ pathologies associated with excessive alcohol consumption, such as alcoholic steatohepatitis and cirrhosis (alcoholic liver disease (ALD)). Several epidemiologic and observational studies show that only a subset of alcoholics develop organ damage such as ALD, indicating that while chronic alcohol consumption is necessary, it is not sufficient to cause organ dysfunction [7,8]. Additional experimental studies indicate that proinflammatory, gut derived bacterial products like endotoxins (lipopolysaccharide; LPS) are required co-factors for alcohol-induced organ pathologies like ALD [9–11]. Further, human and experimental studies show that gut leakiness to LPS is one of the primary mechanisms of endotoxemia [12] and abnormal intestinal bacterial community composition (dysbiosis) that has been shown to occur in the subset of alcoholics and alcohol fed rodents [3,5] that can play a major role in oxidative stress, gut leakiness and endotoxemia and thus could potentially cause the development of alcohol-induced pathologies like ALD [12–17].

However, the observed change in the microbiota composition in alcoholics is not indicative of the dysbiotic intestinal microbiota function that could result in the production of injurious and toxic products. Thus, knowledge of the effects of alcohol on the intestinal microbiota function and their metabolites is warranted to complement the results of alcohol-induced changes to the intestinal microbiota composition, in order to better understand the role of the intestinal microbiota in alcohol associated organ pathologies. This knowledge is essential for identifying the potential intestinal microbiota directed therapeutic target(s) to prevent and treat alcoholic organ damage like ALD. However, to the best of our knowledge, there has not been a comprehensive report of the impact of alcohol consumption on the intestinal microbial metabolites.

Recent advancements in the field of metabolomics provide the opportunity to interrogate the impact of alcohol consumption on bacterial metabolites such as volatile organic compounds (VOC) in the stool of alcoholics. Related by their volatility at ambient temperatures, the VOCs comprise a large and structurally diverse family of carbon-based molecules, of both natural and man-made origin. Specialized sampling methods, such as headspace solid-phase microextraction (hSPME), greatly enable the isolation of VOCs from a wide array of biological samples [18–21], including feces [22–27]. hSPME typically involves the partitioning of the VOCs from the headspace above a sample into a polymeric sorbent adhered to a fused silica rod (fiber), subsequent desorption of the VOCs into the heated inlet of a gas chromatograph, separation of the VOC mixture by gas-liquid partition chromatography, and detection by mass spectrometry. Spectral comparison to a reference database enables VOC identification.

One of the challenges to interrogating microbiota metabolites is the selection of the appropriate samples and the method of sample collection in order to avoid potential confounding factors, such as the continual bacterial metabolic events *ex-vivo* after samples, like stool, are voided and exposed to ambient environment before freezing. Indeed, we recently reported that the VOC metabolome derived from stool collected at home was different than that obtained from stool collected during endoscopy and immediately frozen avoiding any *ex-vivo* metabolic events [28].

Here, we report the results of a differential metabolomic analysis comparing VOC metabolomes derived from the stool of alcoholics and non-alcoholic healthy controls. We performed the analysis with fecal samples collected after passage (patient's home) and then frozen after a

period of time, as well as with fecal samples collected directly from the sigmoid lumen (via unprep sigmoidoscopy) then immediately frozen to prevent metabolic events from occurring after stool collection. Regardless of the approach to fecal collection, we found a stool VOC metabolomic signature in alcoholics that is different from healthy controls.

Materials and Methods

Fecal samples

The Institutional Review Boards at George Mason University and Rush University Medical Center approved this investigation. An informed written research consent was signed by all study participants. Fecal samples were endoscopically collected from 18 healthy and 16 alcoholic subjects (the 'endoscopy collected samples') or were collected *ex vivo* after passage from 25 healthy and 22 alcoholic subjects (the 'home collected samples'), in the manner described below. (Table 1) depicts the demographic characteristics of the study subjects. Each subject completed a detailed health questionnaire that showed that healthy participants did not have any chronic GI or systemic disease or symptoms, none were taking any regular medication except for blood pressure and cholesterol, and none used supplements including probiotics or prebiotics. No subject took antibiotics, for at least three months, and none of the healthy participants were excessive drinkers of alcohol (women consumed less than 2 drinks per sitting per day or no more than 7 drinks per week and men consumed no more than 4 drinks per sitting per day or no more than 14 drinks per week.). Women were considered alcoholics if they consumed 4 or more drinks per day or 8 or more drinks per week, while men were considered alcoholics if they consumed 5 or more drinks per day or 15 or more drinks per week. All study participants were instructed not to change their usual dietary consumption and, as verified by a dietary questionnaire, all participants demonstrated no change in their typical diet or health status during and 7 days prior to stool collection. We compared the dietary consumption of the healthy and alcoholic cohorts and found no substantial differences between cohorts.

Study participants in the endoscopy collected group had their stool collected *in vivo* via unsedated sigmoidoscopy, after providing an informed, written consent. There was no colon preparation prior to sigmoidoscopy. The stool in the lumen of the distal sigmoid was obtained using a Roth Net (US Endoscopy, Mentor, OH), removed with the sigmoidoscope, and then placed in a cryovial and immediately snap frozen in liquid nitrogen. Upon removal from the liquid nitrogen, the cryovial was immediately stored in a -80°C freezer until analysis. For the home collected group, study participants were instructed on how to place their stool into a BD Gaspak EZ Anaerobe Gas Generating Pouch System with Indicator (Becton, Dickinson and Company, Sparks, MD), to minimize the exposure of stool to high oxygen ambient atmosphere. Study subjects were asked to keep the sealed anaerobic stool bag in a cold environment until bringing the anaerobic stool bag to the hospital. Upon receipt, the stool was immediately stored in a -80°C freezer. The interval between passage of stool and storage at -80°C was within 12 to 24 hours.

hSPME procedure

The frozen fecal samples were dispensed in 0.2 g aliquots into 4 mL WISP style screw thread amber glass vials, sealed with Black Top Hat PTFE/Silicone caps (J.G. Finneran, Vineland, NJ), and stored at -80°C until analyzed. Three different SPME fibers (Supelco, Bellefonte, PA) were used in our investigation; 75 μm carboxen-polydimethylsiloxane (CAR-PDMS), 85 μm polyacrylate (PA), and 50/30 μm divinylbenzene (DVB)-CAR-PDMS. Each study subject's fecal sample was extracted with each of the three SPME fibers, using a new fecal aliquot for each hSPME. All fibers were preconditioned before use, following the manufacturer's instructions.

Table 1. Characteristics of the study participants.

	Alcoholics Endo Collection (N = 16)	Healthy Controls Endo Collection (N = 18)	Alcoholics Home Collection (N = 22)	Healthy Controls Home Collection (N = 25)
Gender: male, M; female, F	15 M; 1 F	8 M; 10 F	18 M; 4 F	11 M; 14 F
Race: Caucasian, C; African American, AA;Asian, A	8 C; 8 AA	9 C; 8 AA; 1 A	10 C; 12 AA	12 C; 12 AA; 1 A
Age Range	30–64	20–63	30–64	20–63
Age Mean	49.9	39	48.4	37.7
BMI Range	15.9–43.9	19.6–45.4	15.9–43.9	19.6–45.4
BMI Mean	25.3	31.6	27.6	29.7
Alcohol Consumption History (Years) Mean	28.9	12.9	27.4	12
Smoking During Time of Study (1–2 packs per day)	8 out of 16	5 out of 18	11 out of 22	6 out of 25
NSAID Usage During Time of Study (Daily)	3 out of 16	0 out of 18	4 out of 22	0 out of 25

All analyses were performed in duplicate. The sample vials were heated to 60°C for 30 minutes prior to positioning the hSPME fiber into the headspace above the feces. The extraction was performed until equilibrium (18 hours; [26,28]), with the sample vial temperature held at 60°C for the duration of the extraction. The fiber assembly was then placed into the GC inlet for thermal desorption of the analytes.

GC-MS Instrument

Samples were analyzed using an Agilent 7890A GC equipped with a DB5-MS capillary column (Agilent, Palo Alta, CA; 30 m length, 0.25 mm ID, and 0.25 μm film thickness), a 0.75 mm ID SPME injection port liner, and a 5975 inert XL mass selective detector (MSD) with triple axis detector. The GC injection port was operated in splitless mode at select inlet temperatures, dependent upon the SPME fiber used (300°C, CAR-PDMS; 280°C PA; 270°C DVB-CAR-PDMS). Helium carrier gas was set to a flow rate of 1.17 mL/min. The GC oven was held at an initial temperature of 35°C for 1 min, ramped at 3°C/min to 80°C, then 10°C/min to 120°C, and finally 40°C/min to 260°C, where the temperature was held for 1.5 min. The total run time for the analysis was 25.0 min. The MSD was scanned from 30 to 550 amu at a rate of 2.81 scans/sec.

Data processing and analysis

The VOCs were identified in the GC-MS chromatograms using the National Institute of Standards and Technology (NIST, Washington, DC) Automated Mass Spectral Deconvolution and Identification System (AMDIS, ver. 2.69) software and mass spectral library (NIST08). Compounds with 85% or greater probability of match to a molecule in the NIST08 library were only considered. Each AMDIS outfile, containing a list of identified metabolites and their corresponding peak height values, was filtered using custom Perl scripts designed to remove background analytes (e.g. siloxanes) and eliminate metabolite redundancies (retaining the replicate with the highest peak value). Duplicate sample data sets were combined using Perl scripts created to merge AMDIS outfiles and average the corresponding peak height values. A comprehensive, three-fiber metabolite dataset was prepared for each sample by pooling the metabolites obtained using the CAR-PDMS, PA, and DVB-CAR-PDMS fibers and summing the corresponding peak height values (a peak height of zero was imputed for missing metabolites). A Perl script was then used to assemble two complete metabolite matrices; one containing all of the endoscopy collected healthy and alcoholic patient samples and their accompanying metabolites, and another containing all of the home collected healthy and

alcoholic patient samples and their accompanying metabolites. Metabolites present in ≤20% of the samples were treated as one-offs and were removed [28]. Each metabolite matrix was arranged into two cohorts (healthy and alcoholic) and the outlier peak height values were identified in each cohort using a plot of (mean-median)/median for each analyte and a cutoff value ≥1.5. Outliers were replaced with the median value for that metabolite within the cohort. Metabolite peak height values were then standardized across the two cohorts by conversion to Z-scores (peak height-mean/standard deviation). A Pearson (n) principal component analysis was then performed using the standardized metabolite matrices and the statistical package XLSTAT 2012.6.02. XLSTAT was also used to perform two sample T tests between cohorts for each metabolite. Benjamini-Hochberg critical values were calculated as (i/m)Q, where i is the rank in an ascending list of *p* values, m is the total number of tests, and Q is a false discovery rate of 0.15. Pearson's correlation coefficients were calculated using Microsoft Excel. A correlation network was created using the R statistical package. Unsupervised hierarchical clustering and heatmap generation was accomplished using R, with the Manhattan method and Pearson correlation for the distance measure. Fold change calculations were performed using Microsoft Excel. Custom Perl scripts were used to combine and compare the cohort metabolites to identify the common and unique metabolites and to group the metabolites and their relative abundance into defined chemical classes. Bar graphs and ROC curves were prepared using GraphPad Prism ver. 4.0.

Results and Discussion

To determine if the fecal VOC metabolome composition is altered by excessive alcohol consumption, we obtained a combined total of 81 stool samples from healthy and alcoholic volunteers. As we illustrated previously [28], the approach to collecting a fecal sample has an impact on the derived VOC metabolome, so we elected to acquire the fecal samples in each of two ways; *in vivo* by endoscopy and *ex vivo* by home collection after passage, as detailed in Materials and Methods. The VOCs from the collected samples were extracted by hSPME and identified by GC-MS. To ensure greater metabolome coverage while still accommodating reasonable sample throughput, three different hSPME fiber chemistries were used (CAR-PDMS, PA, and DVB-CAR-PDMS). All of the extractions were performed in duplicate (using different fecal aliquots) and the replicates combined by averaging the chromatographic peak height values. Hence, a total of 486 chromatograms were generated from the 81 participant fecal samples, resulting in both endoscopy collected (containing 16 alcoholic samples and 18 healthy samples) and home collected (containing 22 alcoholic and 25 healthy samples) VOC metabolome datasets.

When constraining metabolite identification to a minimum 85% molecular library match, a grand total of 2,659 different VOCs are identified in the endoscopy collected fecal samples. In contrast, the home collected samples collectively contain 2,883 total analytes, an additional 224 analytes relative to the endoscopy group. Fig. 1 presents a comparison of the alcoholic and healthy cohort composition in terms of the number of identified analytes and the relative abundance in each of the indicated chemical classes. Of greatest significance, in both the endoscopy collected and home collected VOC metabolomes, the overall chemical distribution appears similar among the healthy and alcoholic cohorts (Figs. 1A and 1B), with a slight bias towards alcohols, alkanes, and alkenes in the home collected healthy group. Further, regardless of the means by which the feces was isolated, very little difference in each of the chemical classes is apparent when comparing relative analyte abundance between the healthy and alcoholic cohorts (Figs. 1C and 1D). Additionally, while the metabolome composition as a whole is asymmetrically distributed across the various chemical classes, the relative distribution remains

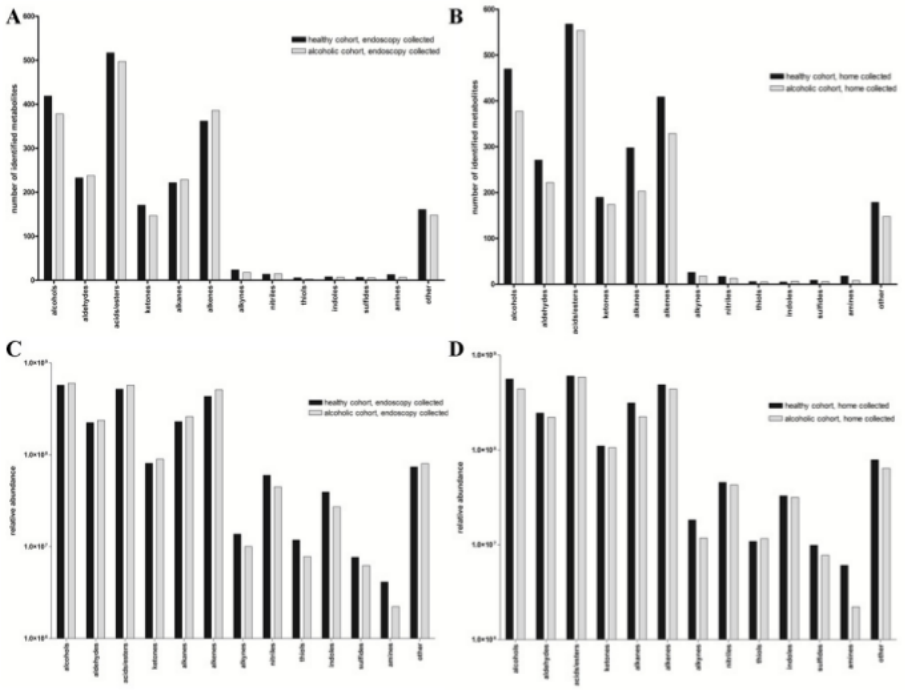


Fig 1. Metabolite composition and abundance. The pooled analytes present in the alcoholic and healthy cohorts were distributed among the listed chemical classes and then tallied. A and B) The bar graphs indicate the total number of analytes in each chemical class for the endoscopy (A) or home collected (B) fecal VOC metabolomes. C and D) The relative abundance (peak height) of the metabolites present in each cohort were distributed among the indicated chemical classes and then summed. The bar graphs indicate the relative abundance of each class for the endoscopy (C) or the home collected (D) fecal VOC metabolomes.

doi:10.1371/journal.pone.0119362.g001

consistent regardless of the cohort or means by which the feces were collected (e.g. the acids/esters group always has the greatest number of metabolites, followed by the alcohols and alkanes, and so on). While this latter observation may simply be a reflection of the three fiber hSPME technique (fiber chemistry dictates the nature of the isolated analytes and while a three fiber analysis expedites sample processing, it results in an incomplete metabolome relative to a study using five or more different fibers [26]), this distribution is also suggestive of a global homeostatic relationship among the chemical classes within the feces. Additional work is required to further explore this possibility.

Although Fig. 1 suggests that the overall chemical composition is very similar between the healthy and alcoholic cohorts, noteworthy differences become apparent when performing a higher resolution comparison of the specific analytes identified within each of the chemical classes. Fig. 2A presents the similarities and differences within the endoscopy collected VOC metabolomes. While a significant number of metabolites are common to both of the cohorts,

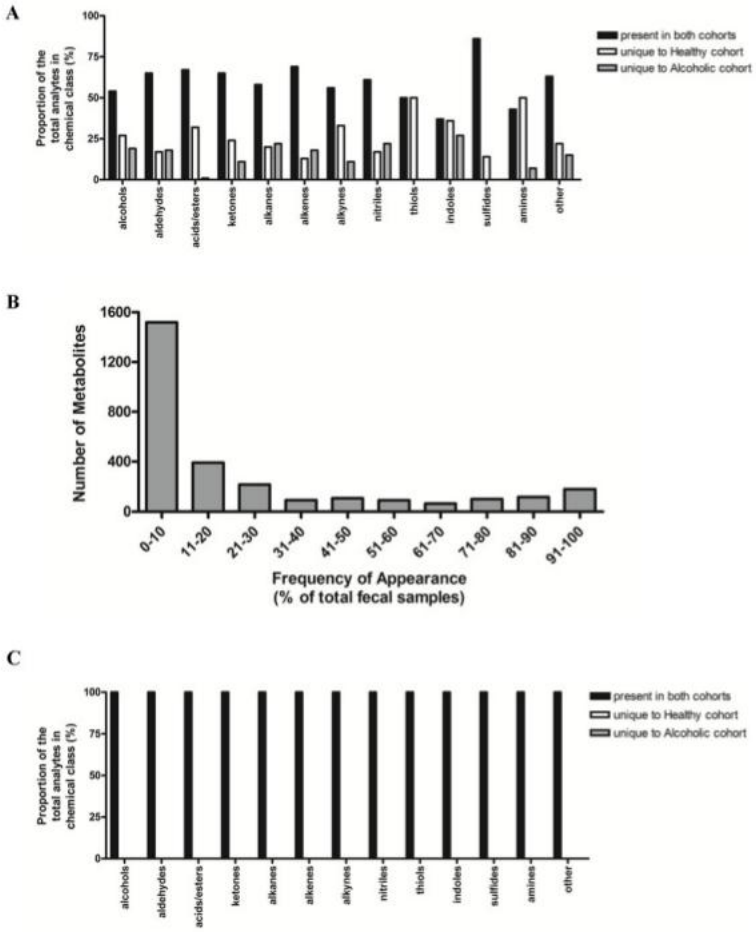


Fig 2. Metabolite distribution within the endoscopy collected fecal samples. A) The VOCs identified in the fecal samples were sorted according to the indicated chemical classes and then further arranged by their unique association with either the healthy or alcoholic cohorts, or their appearance in both cohorts. The percent distribution relates to the total number of metabolites within the chemical class. B) The number of identified VOCs as a function of frequency of appearance among the total number of fecal samples analyzed. A large number of analytes appear in a small number of fecal samples, likely a reflection of dietary variation among the study participants. C) The plot was prepared as described in A), but with the exclusion of the low frequency metabolites (≤20%) identified in B). Consequently, there are no longer any metabolites exclusive to either the healthy or alcoholic cohorts.

doi:10.1371/journal.pone.0119362.g002

in most of the chemical classes a substantial number are uniquely associated with either the healthy or alcoholic samples (equivalent results are also obtained when comparing the home collected metabolomes (data not shown)). However, all of these unique analytes appear in only a small proportion (20% or fewer) of the total number of stool samples analyzed (Fig. 2B). Hence, these 'cohort-unique' metabolites are most likely attributed to variations in dietary intake [28], and when these low frequency metabolites are excluded, the combined metabolome composition appears identical among the cohorts (Fig. 2C). Alternatively, since only a subset of alcoholics develop organ damage such as ALD [7,8], it is also possible that these low frequency metabolites comprise a unique VOC signature associated with eventual organ dysfunction. However, since our investigation is cross sectional by design, we cannot determine if the subset of alcoholics with the unique VOC metabolites will go on to develop organ damage. An additional longitudinal study is required to address this possibility.

We have indicated previously how the colonic microbiome is altered in alcoholism [5]. To ascertain how the metabolite composition and abundance relates among the healthy and alcoholic cohorts, a principal component analysis (PCA) was performed (restricted to analytes appearing in >20% of the samples). As seen in Figs. 3A through 3D, the PCA clearly segregates the healthy and alcoholic samples based upon their VOC metabolome composition, regardless of the approach to fecal sample acquisition. With the endoscopy collected samples (Figs. 3A and 3C), the first principal component clearly discriminates between the two cohorts (as evidenced by the samples segregating into separate groups along the PC1 axis of the PCA plots), whereas the second and third components reveal variation within each of the two segregated cohorts (particularly evident with the alcoholic samples 010A, 029A, 049A and healthy samples 023A, 027A, 030A, 042A, 043A, and 046A (Fig. 3C)). Cohort differentiation is also apparent in the home collected fecal VOC dataset, with healthy and alcoholic segregation readily apparent along the PC1, PC2, and PC3 axis (Figs. 3B and 3D). Numerous metabolites collectively contribute to the segregation of the healthy and alcoholic cohorts (as ranked by the squared cosine of the variable, Figs. 3E and 3F), the top ten of which alone cause segregation of the healthy and alcoholic samples in a PCA (S1 Fig.). A dendrogram and accompanying heat map further depict the clear differentiation of the healthy and alcoholic fecal VOC metabolomes (Fig. 4). Additionally, metabolite correlation networks derived from the healthy and alcoholic fecal VOC metabolomes also illustrate extensive alcohol related changes to the relationships among the metabolites (Fig. 5 and S2 Fig.). Further, a fold change analysis of the endoscopy and home collected fecal VOC metabolomes highlights several metabolites that undergo a significant abundance change associated with the excessive consumption of alcohol (Fig. 6).

As listed in Table 2, when restricting the comparison to only those metabolites found in ≥80% of the total samples present in either the alcoholic or healthy fecal cohort, of a total of 152 metabolites, 18 demonstrate a statistically significant difference in abundance between the healthy and alcoholic cohorts (p value <0.05 and fold change greater than 2), 9 of which are associated with the endoscopy collected samples and 9 are associated with the home collected samples. A PCA with these analytes alone clearly differentiates the healthy and alcoholic cohorts from one another, with either of the fecal collection techniques (not shown). Fig. 7 presents box plots comparing the abundance of these metabolites in the alcoholic and healthy control fecal samples. Of significance, tetradecane demonstrates a fourfold increase in its median abundance level in the endoscopy collected alcoholic fecal VOC metabolome, relative to the healthy controls (see Fig. 7A and Table 2). On the other hand, tetradecane has only a 1.1 fold increase in median abundance in the home collected alcoholic fecal VOC metabolome (data not shown). This discrepancy between the endoscopy and home collected samples corroborates our previous report that the approach to fecal sample acquisition can have a significant impact on the resulting derived VOC metabolome [28]. As the endoscopy collected samples are

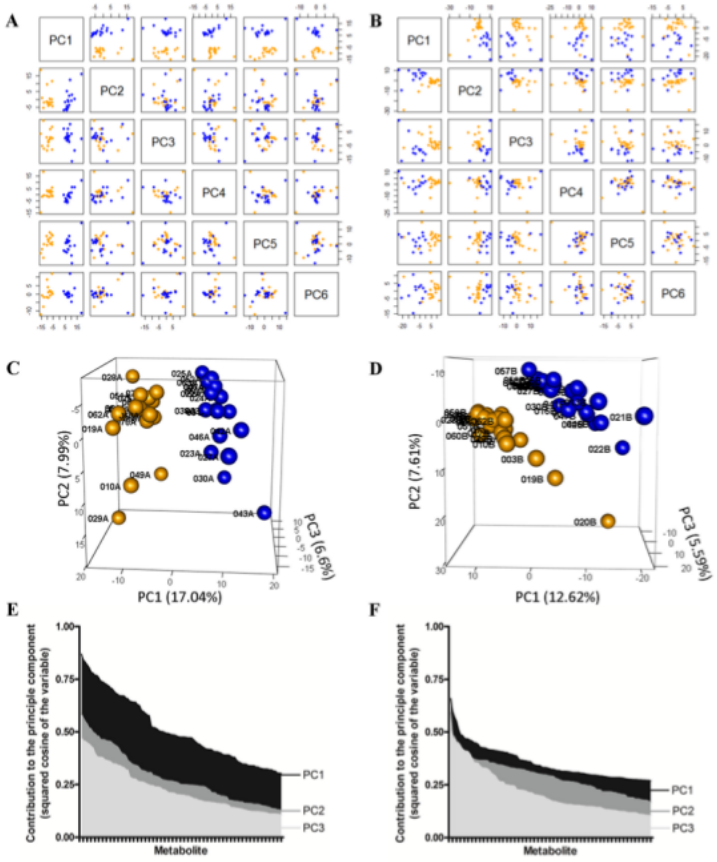


Fig 3. Principal component analysis (PCA) of the VOC metabolomes derived from the healthy and alcoholic human fecal samples. Infrequent metabolites were disregarded by restricting the analysis to analytes that appear in a minimum of 21% of all samples in each cohort (see Fig. 2). Hence, the endoscopy collected dataset contains 525 metabolites while the home collected dataset contains 641. A) PCA plots of the endoscopy collected fecal VOC metabolome. The first (PC1) through sixth (PC6) principal components are shown, relative to one another. Healthy samples are identified as blue spheres, while alcoholic samples are denoted as yellow spheres. The two cohorts clearly segregate in all plots involving PC1, while little to no difference is observed among the cohorts in the combinatorial plots with PC2 through PC6. B) PCA plots of the home collected fecal VOC metabolomes. The samples are colored as described in A). The two cohorts segregate in all plots involving PC1, PC2, and PC3 (to differing degrees), while segregation is not observed in the combinatorial plots of PC4, PC5, and PC6. C) A three dimensional PCA plot of the endoscopy collected fecal VOC metabolome (depicting the first, second and third principal components) clearly illustrates the distinctiveness of the healthy and alcoholic cohorts. Samples are colored as in A). Unique sample identifiers are shown adjacent to each data point. D) A three dimensional PCA plot of the home collected fecal VOC metabolome clearly differentiates the healthy and alcoholic cohorts. E and F) Metabolite contribution to the principal components. For clarity, each graph is restricted to the first three principal components and the top 104 contributing metabolites. Metabolites were arranged by descending contributions to each principal component, and the values plotted in the bar graph. The plots indicate that numerous metabolites collectively contribute to cohort segregation in the endoscopy collected (E) and home collected (F) fecal VOC metabolomes. See text for further discussion.

doi:10.1371/journal.pone.0119362.g003

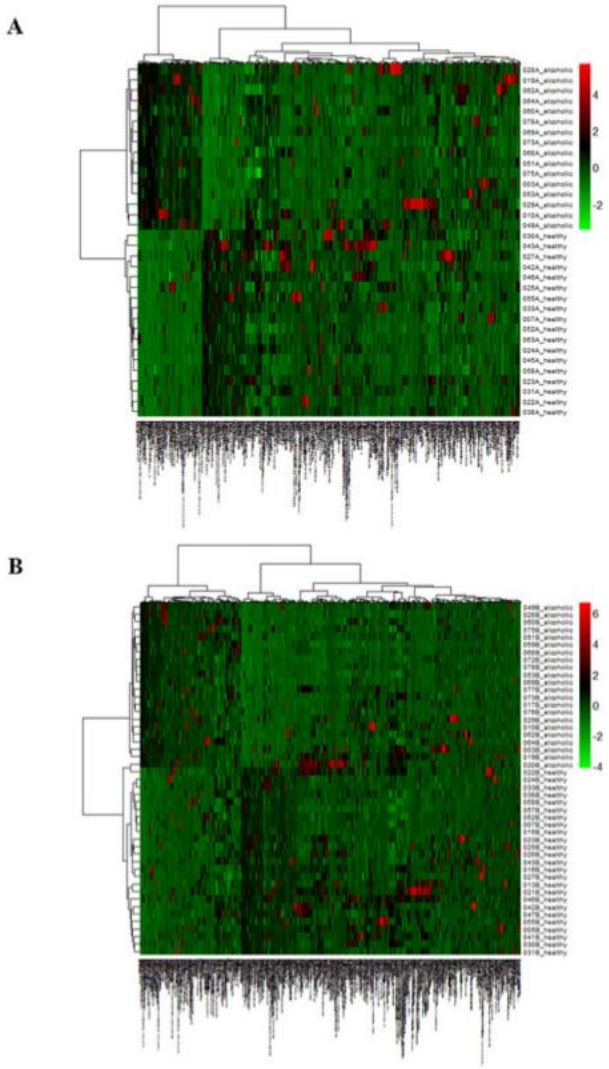


Fig 4. Heat map showing the unsupervised hierarchical clustering of the fecal samples according to the similarity of metabolome composition. The endoscopy collected fecal metabolomes are compared in (A) while the home collected fecal metabolomes are compared in (B). The samples are arranged in rows, the metabolites in columns, and shades of red represent elevation of a metabolite while shades of green represent decrease of a metabolite, relative to the median metabolite levels (see color scale). In the dendrograms, the clustering clearly differentiates the alcoholic and healthy fecal samples.

doi:10.1371/journal.pone.0119362.g004

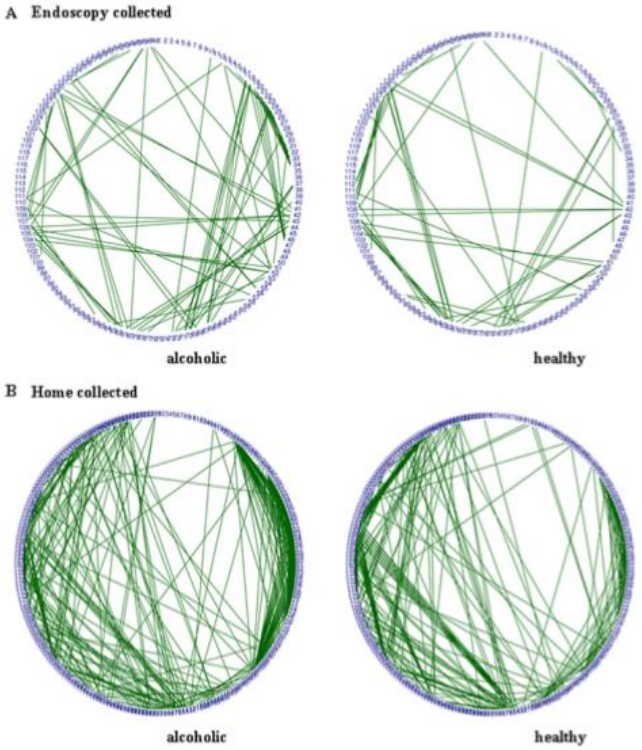


Fig 5. Metabolite correlation network of the endoscopy collected (A) and home collected (B) fecal VOC metabolomes. Pearson's correlation coefficients were calculated for metabolites present in 80% or greater of the total fecal samples. A Pearson correlation value greater than 0.95 is depicted as a green line between metabolites (negative correlations are not shown, as correlation values less than-0.95 were not obtained). To facilitate comparison of the networks, metabolites are numerically represented and their placement around the circumference of each network is fixed among the paired plots. Regardless of the fecal collection method used, the fecal samples from the alcoholic participants have a notably different correlation network than that seen in the fecal samples from non-alcoholics. This difference is even more apparent in correlation networks derived using metabolites present in ≥21% of all fecal samples (S2 Fig.).

doi:10.1371/journal.pone.0119362.g005

immediately snap frozen in liquid nitrogen after their *in vivo* isolation, these samples best reflect the *in situ* fecal VOC metabolome and avoid possible *ex vivo* fermentation/evaporation effects that may influence/alter the composition of the home collected fecal samples. Hence, the endoscopy collected samples are preferred. However, in light of the significantly increased cost associated with the endoscopy collection, the relative ease by which the home collected samples are obtained, and the fact that the home collected healthy and alcoholic cohorts are clearly segregated by principal component analysis of their corresponding VOC metabolomes (Fig. 3), we elected to continue our examination of the home collected samples, as they still offer insight

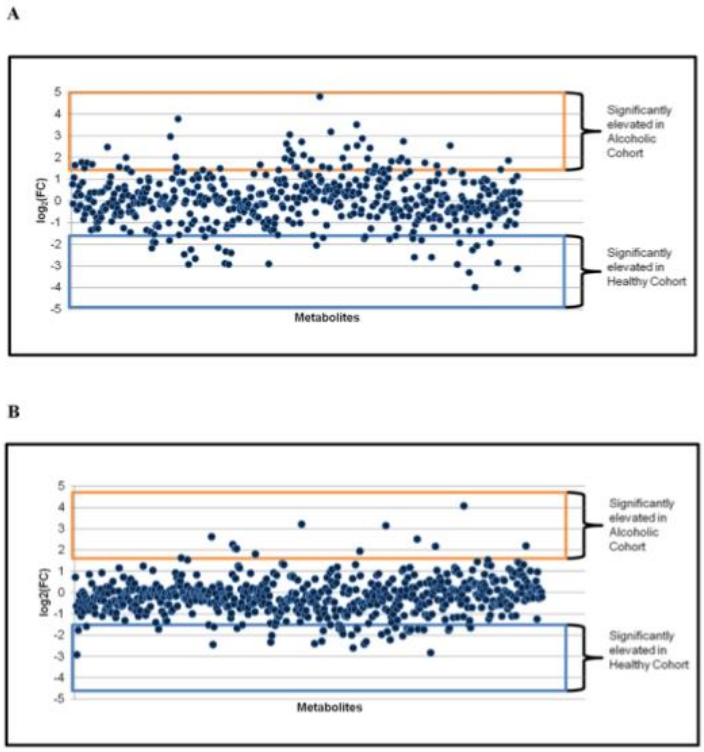


Fig 6. Fold change analysis of the metabolite abundance between the healthy and alcoholic fecal samples. The fold change (FC) is calculated as the log transformation of the ratio between the mean metabolite abundance in the alcoholic cohort relative to the healthy cohort. The analysis was performed with both the endoscopy (A) and home passage collected (B) metabolomes. A $\log_2$(FC) greater than 1.5 is deemed significant (equivalent to a threefold or greater change in metabolite abundance). A fold change analysis comparing median analyte values produces similar results (not shown).

doi:10.1371/journal.pone.0119362.g006

into alcohol induced changes to the fecal VOC metabolome, and implicate biomarkers indicative of alcoholism.

As mentioned, tetradecane is increased in abundance in the endoscopy collected alcoholic fecal samples, relative to the healthy controls. Additionally, the major contribution of tetradecane to cohort segregation is within the first principal component of the PCA plot (Fig. 3C), as indicated by the squared cosine of the variable (Fig. 3E). A member of the acyclic alkane family of molecules, tetradecane has been identified as a breath biomarker of oxidative stress [29]. Hence, in the context of fecal VOCs, the increased abundance of tetradecane in the alcoholic cohort might be indicative of a state of oxidative stress within the gut. Accordingly, alcohol-induced oxidative stress is associated with the development of gut hyperpermeability, endotoxemia, and subsequent alcoholic steatohepatitis [30].

Table 2. Metabolites with a statistically significant difference in abundance between the healthy and alcoholic cohorts ($p<0.05$; $p<$the Benjamini-Hochberg critical value at a false discovery rate of 0.15).

Endoscopy Collected Fecal Samples increased in alcoholics	p value	fold change	frequency in Healthy (of 18)	frequency in Alcoholic (of 16)	area under ROC curve
Tetradecane	0.013	4.07	15	13	0.68
decreased in alcoholics					
2-Tetradecen-1-ol	0.000	2.52	17	14	0.83
1-Undecanol	0.025	2.14	15	12	0.67
Propanoic acid	0.029	2.32	18	16	0.67
Cyclopropane, nonyl-	0.012	2.34	14	12	0.70
6-Pentadecen-1-ol	0.002	8.48	14	11	0.85
8-Tetradecen-1-yl acetate	0.008	2.33	14	7	0.77
1,15-Pentadecanediol	0.016	3.09	16	9	0.72
Eicosen-1-ol	0.035	3.98	14	10	0.74

Home Collected Fecal Samples increased in alcoholics	p value	fold change	frequency in Healthy (of 25)	frequency in Alcoholic (of 22)	area under ROC curve
none					
decreased in alcoholics					
caryophyllene	0.044	2.12	24	20	0.68
1-Naphthalenol	0.023	2.12	25	19	0.64
Phellandrene	0.022	3.12	21	17	0.70
Dimethyl disulfide	0.016	4.63	24	22	0.80
Dimethyl trisulfide	0.002	4.27	25	22	0.77
Camphene	0.012	5.56	20	16	0.80
2,5-Pyrrolidinedione, 1-(benzoyloxy)-	0.040	2.23	20	12	0.71
5-Hepten-2-one, 6-methyl-	0.033	3.86	21	14	0.75
(2-Aziridinylethyl)amine	0.030	2.55	21	16	0.66

doi:10.1371/journal.pone.0119362.t002

On the contrary, the observed change in relative abundance of tetradecane in the home collected alcoholic cohort samples is only minimal (with a 1.1 fold increase in alcoholics). In fact, no metabolites appearing in ≥80% of the home collected samples demonstrate a 2 fold or larger increase in abundance in the alcoholic samples (rather, the metabolites in the home collected samples demonstrating greater than 2 fold change in abundance were all decreased in alcoholics, as described below). As a point of interest, ascorbic acid 2,6-dihexadecanoate, a known phytochemical [31–33], was detected in 20 of the 25 healthy and 18 of the 22 alcoholic home collected fecal samples and demonstrates a 1.72 fold increase in the alcoholic cohort (p value = 0.02; data not shown). Ascorbic acid 2,6-dihexadecanoate is a known constituent of wine [34].

Of the eight metabolites identified as decreased in the endoscopy collected alcoholic fecal samples (Table 2), five are members of the fatty alcohol family of molecules (2-Tetradecen-1-ol, 1-undecanol, 6-pentadecen-1-ol, 1,15-pentadecanediol, and eicosen-1-ol). The first four of these fatty alcohols contribute maximally to segregation of the healthy and alcoholic cohorts along principal component 1 in the PCA (Fig. 3C), while eicosen-1-ol maximally contributes to principal component 3. As depicted in Fig. 7, of these five fatty alcohols, the abundance of 6-

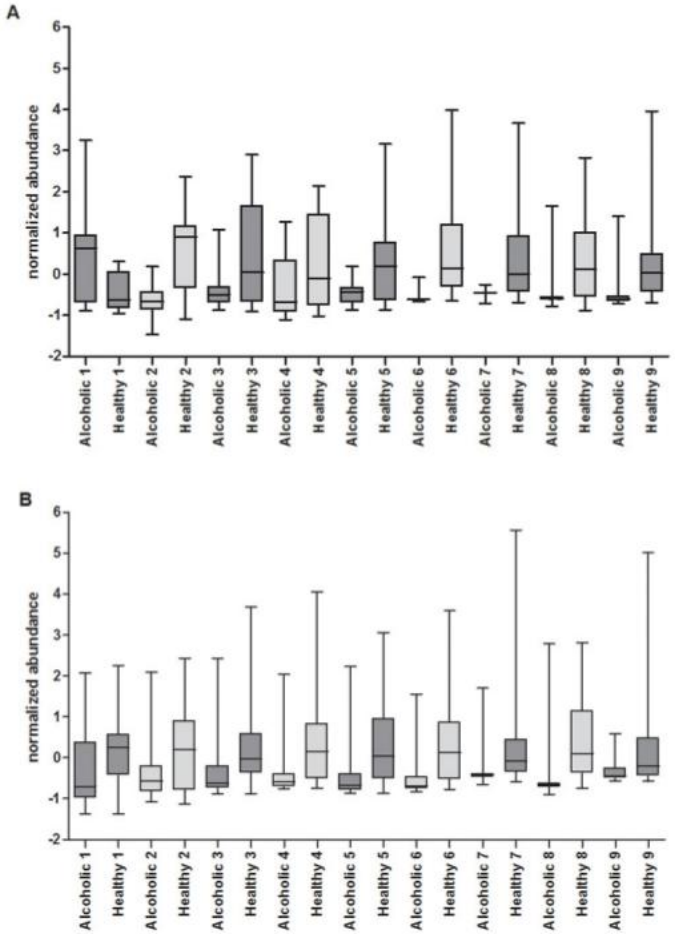


Fig 7. Metabolites with a statistically significant difference in median abundance levels. The metabolites listed in Table 2 were compared among the healthy and alcoholic cohorts. Box plots are shown, depicting the interquartile range of Z-score normalized abundance values, with whiskers extending from minimum to maximum. The median value is identified by a horizontal line within the box. Metabolites from the endoscopy collected samples are shown in A), and are numerically coded as follows; 1) tetradecane, 2) 2-tetradecen-1-ol, 3) 1-undecanol, 4) propanoic acid, 5) cyclopropane, nonyl-, 6) 6-pentadecen-1-ol, 7) 8-Tetradecen-1-yl acetate, 8) 1,15-Pentadecanediol, and 9) Eicosen-1-ol. Metabolites from the home collected samples are shown in B), and are numerically coded as follows; 1) caryophyllene, 2) 1-naphthalenol, 3) phellandrene, 4) dimethyl disulfide, 5) dimethyl trisulfide, 6) camphene, 7) 2,5-pyrrolidinedione, 1-(benzoyloxy)-, 8) 5-hepten-2-one, 6-methyl-, and 9) (2-aziridinylethyl)amine. See text for further discussion.

doi:10.1371/journal.pone.0119362.g007

pentadecen-1-ol, 1,15-pentadecanediol, and eicosen-1-ol are dramatically suppressed in the alcoholic fecal samples. In fact, these three metabolites are below their detection limits in several of the alcoholic fecal samples (ranging from five samples for 6-pentadecen-1-ol to seven for 1,15-pentadecanediol; see Table 2). *In vivo*, fatty alcohols are obtained from the diet or may be derived from fatty acids and fatty aldehydes via the fatty alcohol cycle [35]. Fatty alcohols are precursors for the biosynthesis of wax esters and ether glycerolipids such as the plasmalogens [35,36]. Of particular relevance, the plasmalogens play an important role in cell-cell interactions and gap junctions, and are known to protect against reactive oxygen species [37]. Further, rodents fed a diet rich in fatty alcohols show elevated plasmalogens in the liver [38]. Hence, it is interesting to speculate that the decreased abundance of these fatty alcohols in the alcoholic cohort might be a response to the alcohol-induced oxidative stress, causing an increase in the biosynthesis of plasmalogens (thereby a decrease in these particular fecal fatty alcohols). On the other hand, low levels of these fatty alcohols in alcoholics might represent low capacity of alcoholics to synthesize plasmalogens and thus render the intestine and liver more susceptible to alcohol-induced oxidative stress. While these fatty alcohols appear to be potential biomarkers of excessive alcohol consumption (particularly 6-pentadecen-1-ol), further investigation is needed to elucidate the details underlying the relationship between alcohol consumption, fecal fatty alcohol abundance, and the plasmalogens.

The level of 8-tetradecen-1-yl acetate in the endoscopy collected samples also appears to be significantly depleted due to excessive alcohol consumption (p value = 0.008; Table 2), although the biological meaning remains unclear. Common to most of the healthy samples, its median abundance is decreased 2.33 fold in the alcoholic fecal cohort, even falling below detection limits in over half of the cohort samples (Table 2 and Fig. 7). Nonylcyclopropane, a known phytochemical [39] and VOC associated with meat [40], is also significantly suppressed in the endoscopy collected alcoholic fecal cohort (p value = 0.012; 2.34 fold decrease in abundance in the alcoholic cohort relative to healthy). However, as with 8-tetradecen-1-yl acetate, the biological relevance of this decline in abundance is not readily apparent and requires further investigation.

As indicated in Table 2 and Fig. 7A, the median abundance of propanoic acid is significantly lower in the endoscopy collected alcoholic cohort, relative to the healthy controls (p value = 0.029; fold change = 2.32). Short chain fatty acids (SCFAs) such as acetate, propionate, butyrate, isobutyrate, pentanoate, and isopentanoate are products of microbial fermentation and are often considered to be essential to maintain and promote normal colonic epithelial cell barrier integrity [41,42]. Fig. 8 illustrates box plots depicting the abundance of these SCFAs, in both the healthy and alcoholic cohorts. With the endoscopy collected fecal samples (Fig. 8A), among all these SCFAs, only propanoic acid shows a statistically significant reduction in median abundance in the alcoholic cohort relative to the healthy cohort (acetate follows with a 1.49 fold reduction in median abundance and a p value of 0.34). In contrast, propanoic acid levels in the home collected alcoholic fecal samples are only reduced by 1.2 fold (p = 0.40; Fig. 8B). While the median abundance of isobutyrate appears comparable in the endoscopically collected healthy and alcoholic cohorts, it's notable that this SCFA was only detected in 8 of the 16 alcoholic fecal samples (whereas it was present in ~80% of the healthy samples), illustrating the alcohol related loss of this analyte. A decrease in isobutyrate was also observed in the home collected alcoholic fecal samples (p = 0.006; fold change = 1.72). Overall, the alcohol related decrease of propionate and isobutyrate may be a reflection of the alcohol induced changes to the microbiome composition and could provide a mechanism through which alcohol-induced changes to the microbiota composition contribute to alcohol-induced gut leakiness. However, it is not clear why the other SCFAs remain relatively unaffected.

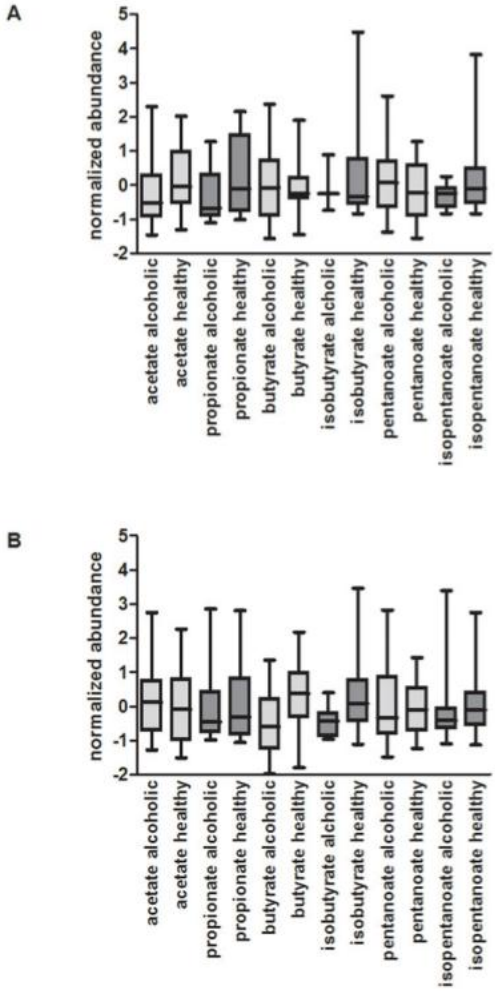


Fig 8. Comparison of SCFA abundance among the cohorts. Box plots are presented, depicted as described in Fig.8. A) Among the endoscopy collected samples, while most of the SCFAs are comparable in abundance within the healthy and alcoholic cohorts, propanoic acid demonstrates a statistically significant reduction in median abundance in the alcoholic cohort relative to the healthy cohort ($p = 0.03$, fold change = 2.32). B) In the home collected samples, butyrate has a 1.6 fold reduction in median abundance, while the remaining SCFAs are very similar between the healthy and alcoholic fecal samples. See text for further discussion.

doi:10.1371/journal.pone.0119362.g008

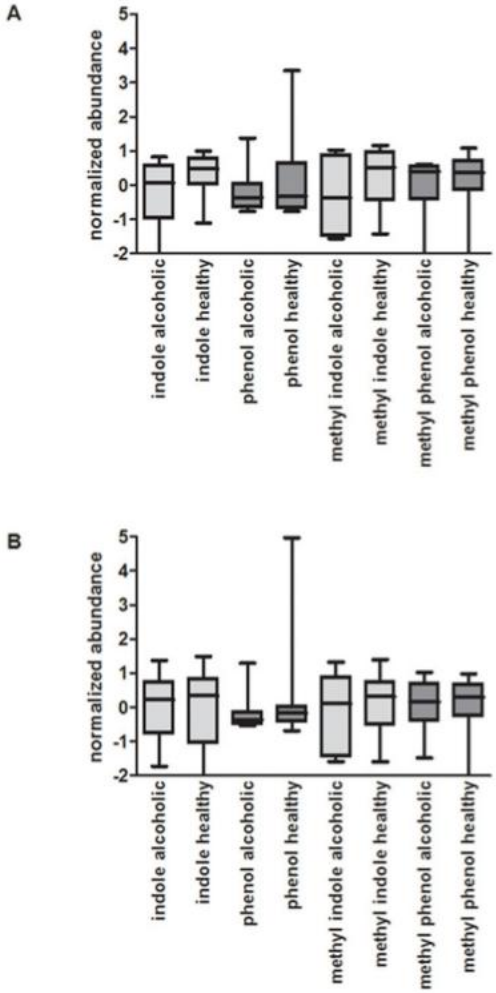


Fig 9. Comparison of protein putrefaction products among the cohorts. Box plots depicting relative metabolite abundance are shown, as described in Fig. 8. A) Among the endoscopy collected samples, with the exception of methyl indole, the metabolites are comparable in abundance within the healthy and alcoholic cohorts. Note however, the difference observed with methyl indole is not statistically significant ($p = 0.519$). B) In the home collected samples, metabolite abundance is very similar between the healthy and alcoholic fecal samples.

doi:10.1371/journal.pone.0119362.g009

In contrast to the SCFAs, protein putrefaction products such as indole, methyl indole, phenol, and methyl phenol have been shown to be injurious to intestinal epithelial cells and could disrupt intestinal barrier integrity [43]. Fig. 9 illustrates box plots depicting the abundance of these metabolites, in both the endoscopy and home collected healthy and alcoholic cohorts. While there is a 1.7 fold decrease in the median abundance of methyl indole in the endoscopy collected alcoholic cohort, relative to the healthy cohort, the difference is not statistically significant ($p = 0.519$). Hence, excessive alcohol consumption appears to have little influence on the abundance of these indicators of protein putrefaction.

As listed in Table 2, nine metabolites were found to be significantly decreased in the home collected alcoholic fecal samples. Of these, caryophyllene, a widely dispersed phytochemical [44–46], is a well characterized agonist of the cannabinoid receptor 2. Interestingly, the association of caryophyllene with the cannabinoid receptor 2 has been shown to reduce voluntary alcohol intake in mice and as such the receptor has been suggested as a target for the pharmaceutical intervention of alcoholism [47]. Within this context, it is noteworthy that the median level of caryophyllene abundance in the alcoholic cohort was found to be twofold lower than that found in the healthy cohort ($p = 0.044$; see Fig. 7B and Table 2).

The relative abundance of phellandrene and 1-naphthalenol are also decreased in the home collected alcoholic fecal cohort, relative to the healthy group (Table 2 and Fig. 7B). A monoterpenoid, phellandrene is a constituent of many plant extracts [48–50], whereas the naphthalene derivative 1-naphthalenol is implicated as a biomarker for exposure to polycyclic aromatic hydrocarbons [51–53]. While the median abundance of both phellandrene and 1-naphthalenol are significantly decreased in the alcoholic samples relative to healthy ($p < 0.05$, fold change >2), the biological significance and implications of this abundance change remain elusive.

In the home collected fecal samples, the most significant decrease in relative metabolite abundance in the alcoholic cohort is observed with dimethyl disulfide, dimethyl trisulfide, and camphene (all with a fold change >4; Table 2 and Fig. 7B). Microbial products of decomposition [54], dimethyl disulfide and dimethyl trisulfide were detected in all 22 of the tested alcoholic and 24 of the 25 healthy cohort fecal samples. With over a fourfold decrease in median abundance level relative to healthy controls ($p < 0.016$; Table 2), dimethyl disulfide and dimethyl trisulfide are potential biomarkers of chronic alcohol consumption in the home collected fecal samples. It is noteworthy though, these two analytes do not demonstrate a statistically significant change in the endoscopy collected fecal samples (dimethyl disulfide—$p = 0.93$, fold change = 1.1, dimethyl trisulfide—$p = 0.28$, fold change = 1.3). This may reflect the means by which they are formed (via the oxidation of methanethiol [55,56]), more likely to occur in passaged fecal samples exposed to air. In fact, in the home collected samples, a threefold reduction in median abundance of methanethiol is also apparent in the alcoholic cohort relative to the healthy cohort, whereas in the endoscopy collected samples this fold change is only 1.1. Also of note, there is a 3.5 fold greater median abundance of methanethiol in the home collected healthy samples as there is in the endoscopy collected healthy fecal cohort.

The phytochemical camphene, a bicyclic monoterpene shown to attenuate hepatic steatosis in mice [57], demonstrates a 5.5 fold reduction in median abundance in the alcoholic fecal samples relative to healthy (Table 2 and Fig. 7B). As hepatic steatosis occurs in the early stages of alcohol liver disease, it is interesting to reflect on the significant decrease in camphene levels in the alcoholic samples. There is no statistically significant change in camphene in the endoscopy collected samples.

The metabolite 1-benzoyloxy-2,5-pyrrolidinedione also demonstrates a significant reduction in median abundance in the home collected alcoholic fecal samples (Table 2 and Fig. 7B). Indeed, while 1-benzoyloxy-2,5-pyrrolidinedione was detected in 20 of the 25 healthy fecal samples, the analyte could only be detected in 55% of the alcoholic fecal samples, and with a

2.2 fold reduction in median abundance in the alcoholic cohort ($p = 0.04$). Similarly, (2-aziridinylethyl)amine, detected in 84% of the healthy samples and 73% of the alcoholic samples, is decreased in the alcoholic cohort 2.5 fold relative to the median abundance in the healthy cohort ($p = 0.03$; Table 2 and Fig. 7B). The biological relevance of 1-benzoyloxy-2,5-pyrrolidinedione and (2-aziridinylethyl)amine remains unclear. 6-Methyl-5-hepten-2-one, on the other hand, is a well-known mammalian VOC detected in skin, breath, and fecal samples [58–62]. Identified in 21 of 25 home collected healthy fecal samples, but only 14 of 22 alcoholic samples, 6-methyl-5-hepten-2-one is decreased 3.86 fold in median abundance in the alcoholic cohort ($p = 0.033$; Table 2 and Fig. 7B). However, 6-methyl-5-hepten-2-one levels have been linked to the estrus cycle in several female mammals [61,63,64]. It's noteworthy, as indicated in Table 1, the home collected healthy fecal cohort contains samples obtained from 14 female participants, whereas the alcoholic cohort has samples obtained from 4.

Biomarkers of Alcoholism

As we have demonstrated previously [28], the approach to fecal sample collection has a profound impact on the derived VOC metabolome. Hence, a strong argument can be made in favor of the endoscopy collected samples, as they are immediately snap frozen upon collection and thus best reflect the colonic metabolism occurring *in vivo*. Indeed, in studies seeking to understand the pathology of disease, the value of the endoscopy collected samples is clear. However, our comparative analysis of home collected fecal samples also demonstrates clear differentiation of healthy and alcoholic VOC metabolomes (Fig. 3). Thus, as the home collected samples are drastically cheaper to obtain, they appear well suited to investigations seeking potential biomarkers of disease.

Ideally, a biomarker of chronic alcohol consumption will appear in all of the healthy and/or alcoholic cohort samples and demonstrate a significant change in abundance between the two cohorts. Accordingly, dimethyl disulfide and dimethyl trisulfide are nearly ubiquitous among all of the home collected fecal samples, yet their relative abundance in the alcoholic feces is significantly suppressed (Table 2). Indeed, their area under the ROC curve is 0.80 and 0.77 respectively, indicative of a marker affording a reasonable balance between sensitivity and specificity (Fig. 10). Additionally, despite its absence in some of the home collected healthy and alcoholic fecal samples, camphene also has an area under the ROC curve indicative of a good biomarker (Table 2). On the other hand, with the endoscopy collected samples, although propanoic acid is ubiquitous among all of the tested samples, the area under the ROC curve is only 0.67, which affords poor accuracy in the determination of alcoholism. Instead, the metabolites 2-tetradecen-1-ol and 6-pentadecen-1-ol are much better biomarkers of alcoholism in the endoscopy collected samples (Table 2). Further studies are now needed to qualify/validate these potential biomarkers.

It is important to consider the impact of diet on the stool VOC profile as dietary products are substrates for bacterial metabolism and thus the type of diet can impact bacterial metabolism. We used a food frequency questionnaire to assess the global dietary habit of our subjects and there was no significant difference between alcoholic and control cohorts in regards to their dietary habit. Thus, the differences in stool VOC profile between alcoholic and control groups cannot be explained by a difference in dietary habit. However, we acknowledge that a recall dietary questionnaire has its limitation and a future longitudinal study with prospective collection of detailed information on diet is required to fully assess the impact of diet on fecal VOCs in alcoholics.

Summary

In summary, regardless of whether the samples were collected *in vivo* or *ex vivo* after passage, it is clear that excessive alcohol consumption has a significant effect on the composition of the

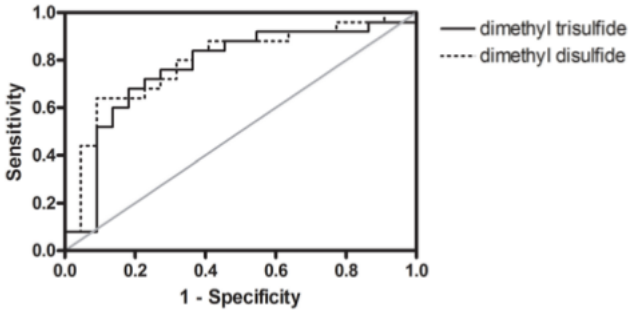


Fig 10. ROC curve of dimethyl disulfide and dimethyl trisulfide. The area under the curve is 0.80 and 0.77, respectively, which are indicative of a fairly good diagnostic test (albeit not excellent).

doi:10.1371/journal.pone.0119362.g010

VOC metabolome, as evidenced by the clear cohort separation in the PCA plots, the distinct healthy and alcoholic clades formed with unsupervised hierarchal clustering analysis, and the drastically distinct metabolite correlation networks within the healthy and alcoholic VOC metabolomes. Numerous metabolites undergo a significant fold change in abundance with excessive alcohol consumption, with many found to increase while others decrease in abundance in the alcoholic feces, relative to healthy controls. The most notable metabolite alterations in the alcoholic samples include: (1) an elevation in the oxidative stress biomarker tetradecane; (2) a decrease in five fatty alcohols with anti-oxidant property and a relationship to the abundance of plasmalogens, known to be linked to cell-cell interactions and gap junctions; (3) a decrease in the short chain fatty acids propionate and isobutyrate, important in maintaining intestinal epithelial cell health and barrier integrity; (4) a decrease in alcohol consumption natural suppressant caryophyllene; (5) a decrease in natural product and hepatic steatosis attenuator camphene; and (6) decreased dimethyl disulfide and dimethyl trisulfide, microbial products of decomposition. With this initial insight into alcohol associated VOC metabolomic change, the stage is set for additional studies associating these metabolites with the progression of alcohol associated pathologies and interventional studies directed to correct these abnormalities to determine whether alcohol associated pathologies such as ALD can be prevented.

Supporting Information

S1 Fig. PCA based only on the ten top scoring metabolites, determined by the weight loadings (squared cosines of the variable) in Fig. 3. The resulting three dimensional plot from the endoscopy collected fecal dataset is shown in A) and the home collected fecal dataset is shown in B).
(TIF)

S2 Fig. Correlation networks of the endoscopy collected (A) and home collected (B) fecal VOC metabolomes. Pearson's correlation coefficients were calculated for all metabolites present in at least 21% of the total fecal samples. A Pearson correlation value greater than 0.95 is depicted as a green line between metabolites, while a Pearson correlation value less than -0.95 is depicted as a red line. Metabolites are numerically represented in the network and their placement around the circumference of the network is fixed among the paired plots. Regardless of the approach to fecal collection, the fecal samples from the alcoholic participants have a

significantly different correlation network than that seen in the fecal samples from non-alcoholics.
(TIF)

Acknowledgments

Special thanks to the following for their contributions to this study: Cynthia Clubb, Trish Ike, Sara Pittman, Quyen Deroule, Alison Hight, Emma Dixon, Andrew Pham, Tiffani Willis, David Derby, Colleen Cook, Phung Vo, Purvang Mirani, and Taelim Kim.

Author Contributions

Conceived and designed the experiments: RDC AK CBF EM PAE. Performed the experiments: AD KN. Analyzed the data: RDC AD FZ KN. Contributed reagents/materials/analysis tools: RDC CBF EM PAE AK. Wrote the paper: RDC CBF PAE AK.

References

1. Quigley EMM. Gut bacteria in health and disease. Gastroenterol Hepatol. 2013; 9: 560–569. PMID: 24729765
2. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014; 505: 559–563. doi: 10.1038/nature12820 PMID: 24336217
3. Mutlu E, Keshavarzian A, Engen P, Forsyth CB, Sikaroodi M, Gillevet P. Intestinal dysbiosis: a possible mechanism of alcohol-induced endotoxemia and alcoholic steatohepatitis in rats. Alcohol Clin Exp Res. 2009; 33: 1836–1846. doi: 10.1111/j.1530-0277.2009.01022.x PMID: 19645728
4. Yan AW, Fouts DE, Brandl J, Stärkel P, Torralba M, Schott E, et al. Enteric dysbiosis associated with a mouse model of alcoholic liver disease. Hepatol Baltim Md. 2011; 53: 96–105. doi: 10.1002/hep.24018
5. Mutlu EA, Gillevet PM, Rangwala H, Sikaroodi M, Naqvi A, Engen PA, et al. Colonic microbiome is altered in alcoholism. Am J Physiol Gastrointest Liver Physiol. 2012; 302: G966–978. doi: 10.1152/ajpgi.00380.2011 PMID: 22241860
6. Bull-Otterson L, Feng W, Kirpich I, Wang Y, Qin X, Liu Y, et al. Metagenomic analyses of alcohol induced pathogenic alterations in the intestinal microbiome and the effect of Lactobacillus rhamnosus GG treatment. PloS One. 2013; 8: e53028. doi: 10.1371/journal.pone.0053028 PMID: 23326376
7. Grant BF, Dufour MC, Harford TC. Epidemiology of alcoholic liver disease. Semin Liver Dis. 1988; 8: 12–25. doi: 10.1055/s-2008-1040525 PMID: 3283941
8. Purohit V, Bode JC, Bode C, Brenner DA, Choudhry MA, Hamilton F, et al. Alcohol, intestinal bacterial growth, intestinal permeability to endotoxin, and medical consequences: summary of a symposium. Alcohol Fayettev N. 2008; 42: 349–361. doi: 10.1016/j.alcohol.2008.03.131 PMID: 18504085
9. Bigatello LM, Broitman SA, Fattori L, Di Paoli M, Pontello M, Bevilacqua G, et al. Endotoxemia, encephalopathy, and mortality in cirrhotic patients. Am J Gastroenterol. 1987; 82: 11–15. PMID: 3799574
10. Keshavarzian A, Farhadi A, Forsyth CB, Rangan J, Jakate S, Shaikh M, et al. Evidence that chronic alcohol exposure promotes intestinal oxidative stress, intestinal hyperpermeability and endotoxemia prior to development of alcoholic steatohepatitis in rats. J Hepatol. 2009; 50: 538–547. doi: 10.1016/j.jhep.2008.10.028 PMID: 19155080
11. Voigt RM, Forsyth CB, Keshavarzian A. Circadian disruption: potential implications in inflammatory and metabolic diseases associated with alcohol. Alcohol Res Curr Rev. 2013; 35: 87–96. PMID: 24313168
12. Nanji AA, Khettry U, Sadrzadeh SM, Yamanaka T. Severity of liver injury in experimental alcoholic liver disease. Correlation with plasma endotoxin, prostaglandin E2, leukotriene B4, and thromboxane B2. Am J Pathol. 1993; 142: 367–373. PMID: 8382006
13. Rao RK, Baker RD, Baker SS, Gupta A, Holycross M. Oxidant-induced disruption of intestinal epithelial barrier function: role of protein tyrosine phosphorylation. Am J Physiol. 1997; 273: G812–823. PMID: 9357822
14. Banan A, Choudhary S, Zhang Y, Fields JZ, Keshavarzian A. Ethanol-induced barrier dysfunction and its prevention by growth factors in human intestinal monolayers: evidence for oxidative and cytoskeletal mechanisms. J Pharmacol Exp Ther. 1999; 291: 1075–1085. PMID: 10565827

15. Keshavarzian A, Holmes EW, Patel M, Iber F, Fields JZ, Pethkar S. Leaky gut in alcoholic cirrhosis: a possible mechanism for alcohol-induced liver damage. Am J Gastroenterol. 1999; 94: 200–207. doi: 10.1111/j.1572-0241.1999.00797.x PMID: 9934756

16. Wang HJ, Gao B, Zakhari S, Nagy LE. Inflammation in alcoholic liver disease. Annu Rev Nutr. 2012; 32: 343–368. doi: 10.1146/annurev-nutr-072610-145138 PMID: 22524187

17. Frazier TH, DiBaise JK, McClain CJ. Gut microbiota, intestinal permeability, obesity-induced inflammation, and liver injury. JPEN J Parenter Enteral Nutr. 2011; 35: 14S–20S. doi: 10.1177/0148607111413772 PMID: 21807932

18. Boland W. Analysis of Volatiles. 1984.

19. Schulz S, Fuhlendorff J, Reichenbach H. Identification and synthesis of volatiles released by the myxobacterium Chondromyces crocatus. Tetrahedron. 2004; 60: 3863–3872.

20. Arthur CL, Pawliszyn J. Solid-Phase Microextraction with Thermal-Desorption Using Fused-Silica Optical Fibers. Anal Chem. 1990; 62: 2145–2148.

21. Alpendurada MD. Solid-phase microextraction: a promising technique for sample preparation in environmental analysis. J Chromatogr A. 2000; 889: 3–14. PMID: 10985530

22. Garner CE, Smith S, de Lacy Costello B, White P, Spencer R, Probert CS, et al. Volatile organic compounds from feces and their potential for diagnosis of gastrointestinal disease. Faseb J. 2007; 21: 1675–88. PMID: 17314143

23. De Lacy Costello B. An analysis of volatiles in the headspace of the faeces of neonates. J Breath Res. 2008; 2: 037023. doi: 10.1088/1752-7155/2/3/037023 PMID: 21386183

24. Garner CE, Smith S, Bardhan PK, Ratcliffe NM, Probert CS. A pilot study of faecal volatile organic compounds in faeces from cholera patients in Bangladesh to determine their utility in disease diagnosis. Trans R Soc Trop Med Hyg. 2009; 103: 1171–3. doi: 10.1016/j.trstmh.2009.02.004 PMID: 19268999

25. Garner CE, Ewer AK, Elasouad K, Power F, Greenwood R, Ratcliffe NM, et al. Analysis of faecal volatile organic compounds in preterm infants who develop necrotising enterocolitis: a pilot study. J Pediatr Gastroenterol Nutr. 2009; 49: 559–65. doi: 10.1097/MPG.0b013e3181a3bfbc PMID: 19668005

26. Dixon E, Clubb C, Pittman S, Ammann L, Rasheed Z, Kazmi N, et al. Solid-phase microextraction and the human fecal VOC metabolome. PloS One. 2011; 6: e18471. doi: 10.1371/journal.pone.0018471 PMID: 21494609

27. Li RW, Wu S, Li W, Navarro K, Couch RD, Hill D, et al. Alterations in the Porcine Colon Microbiota Induced by the Gastrointestinal Nematode Trichuris suis. Infect Immun. 2012; 80: 2150–2157. doi: 10.1128/IAI.00141-12 PMID: 22493085

28. Couch RD, Navarro K, Sikaroodi M, Gillevet P, Forsyth CB, Mutlu E, et al. The approach to sample acquisition and its impact on the derived human fecal microbiome and VOC metabolome. PloS One. 2013; 8: e81163. doi: 10.1371/journal.pone.0081163 PMID: 24260553

29. Phillips M, Cataneo RN, Cheema T, Greenberg J. Increased breath biomarkers of oxidative stress in diabetes mellitus. Clin Chim Acta. 2004; 344: 189–94. PMID: 15149888

30. Forsyth CB, Farhadi A, Jakate SM, Tang Y, Shaikh M, Keshavarzian A. Lactobacillus GG treatment ameliorates alcohol-induced intestinal oxidative stress, gut leakiness, and liver injury in a rat model of alcoholic steatohepatitis. Alcohol Fayettev N. 2009; 43: 163–172. doi: 10.1016/j.alcohol.2008.12.009 PMID: 19251117

31. Huang B, Ban X, He J, Tong J, Tian J, Wang Y. Comparative analysis of essential oil components and antioxidant activity of extracts of Nelumbo nucifera from various areas of China. J Agric Food Chem. 2010; 58: 441–448. doi: 10.1021/jf902643e PMID: 19919095

32. Aja PM, Nwachukwu N, Ibiam UA, Igwenyi IO, Offor CE, Orji UO. Chemical Constituents of Moringa oleifera Leaves and Seeds from Abakaliki, Nigeria. Am J Phytomedicine Clin Ther. 2014; 2: 310–321.

33. Okwu D, Ighodaro B. GC-MS Evaluation of Bioactive Compounds and Antibacterial Activity of the Oil Fraction from the Leaves of Alstonia boonei De Wild. Pharma Chem. 2010; 2: 261–272.

34. Chao W, Na Z. Analysis of Volatile Flavor Compounds in Wines by Headspace Solid-phase Microextraction Coupled with Gas Chromatography. Shimadzu.

35. Rizzo WB. Fatty aldehyde and fatty alcohol metabolism: review and importance for epidermal structure and function. Biochim Biophys Acta. 2014; 1841: 377–389. doi: 10.1016/j.bbalip.2013.09.001 PMID: 24036493

36. Blank ML, Cress EA, Smith ZL, Snyder F. Dietary supplementation with ether-linked lipids and tissue lipid composition. Lipids. 1991; 26: 166–169. PMID: 2051899

37. Gorgas K, Teigler A, Komljenovic D, Just WW. The ether lipid-deficient mouse: tracking down plasmalogen functions. Biochim Biophys Acta. 2006; 1763: 1511–1526. doi: 10.1016/j.bbamcr.2006.08.038 PMID: 17027098

38. Gelman RA, Gilbertson JR. Permeability of the blood-brain barrier to long-chain alcohols from plasma. Nutr Metab. 1975; 18: 169–175. PMID: 1196546

39. Adeosun CB, Olaseinde S, Opeifa AO, Atolani O. Essential oil from the stem bark of Cordia sebestena scavenges free radicals. J Acute Med. 2013; 3: 138–141. doi: 10.1016/j.jacme.2013.07.002

40. Ramarathnam N, Rubin LJ, Diosady LL. Studies on meat flavor. 4. Fractionation, characterization, and quantitation of volatiles from uncured and cured beed and chicken. J Agric Food Chem USA. 1993; Available: http://agris.fao.org/agris-search/search.do?recordID=US9439856

41. Wang H-B, Wang P-Y, Wang X, Wan Y-L, Liu Y-C. Butyrate enhances intestinal epithelial barrier function via up-regulation of tight junction protein Claudin-1 transcription. Dig Dis Sci. 2012; 57: 3126–3135. doi: 10.1007/s10620-012-2259-4 PMID: 22684624

42. Elamin EE, Masclee AA, Dekker J, Pieters H-J, Jonkers DM. Short-chain fatty acids activate AMP-activated protein kinase and ameliorate ethanol-induced intestinal barrier dysfunction in Caco-2 cell monolayers. J Nutr. 2013; 143: 1872–1881. doi: 10.3945/jn.113.179549 PMID: 24132573

43. Macfarlane GT, Macfarlane S. Bacteria, colonic fermentation, and gastrointestinal health. J AOAC Int. 2012; 95: 50–60. PMID: 22468341

44. Nikbin M, Kazemipour N, Maghsoodlou MT, Valizadeh J, Sepehrimanesh M, Davarimanesh A. Mineral elements and essential oil contents of Scutellaria luteo-caerulea Bomm. & Snit. Avicenna J Phytomedicine. 2014; 4: 182–190. PMID: 25050316

45. Seo S-M, Kim J, Kang J, Koh S-H, Ahn Y-J, Kang K-S, et al. Fumigant toxicity and acetylcholinesterase inhibitory activity of 4 Asteraceae plant essential oils and their constituents against Japanese termite (Reticulitermes speratus Kolbe). Pestic Biochem Physiol. 2014; 113: 55–61. doi: 10.1016/j.pestbp.2014.06.001 PMID: 25052527

46. Nurzyńska-Wierdak R, Bogucka-Kocka A, Szymczak G. Volatile constituents of Melissa officinalis leaves determined by plant age. Nat Prod Commun. 2014; 9: 703–706. PMID: 25026727

47. Al Mansouri S, Ojha S, Al Maamari E, Al Ameri M, Nurulain SM, Bahi A. The cannabinoid receptor 2 agonist, β-caryophyllene, reduced voluntary alcohol intake and attenuated ethanol-induced place preference and sensitivity in mice. Pharmacol Biochem Behav. 2014; 124C: 260–268. doi: 10.1016/j.pbb.2014.06.025

48. Das AK, Swamy PS. Comparison of the volatile oil composition of three Atalantia species. J Environ Biol Acad Environ Biol India. 2013; 34: 569–571.

49. Bukvicki D, Gottardi D, Veljic M, Marin PD, Vannini L, Guerzoni ME. Identification of volatile components of liverwort (Porella cordaeana) extracts using GC/MS-SPME and their antimicrobial activity. Mol Basel Switz. 2012; 17: 6982–6995. doi: 10.3390/molecules17066982

50. Bos R, Koulman A, Woerdenbag HJ, Quax WJ, Pras N. Volatile components from Anthriscus sylvestris (L.) Hoffm. J Chromatogr A. 2002; 966: 233–238. PMID: 12214699

51. Sreekanth R, Prasanthkumar KP, Sunil Paul MM, Aravind UK, Aravindakumar CT. Oxidation Reactions of 1- and 2-Naphthols: An Experimental and Theoretical Study. J Phys Chem A. 2013; 117: 11261–11270. doi: 10.1021/jp4081355 PMID: 24093754

52. Preuss R, Angerer J, Drexler H. Naphthalene—an environmental and occupational toxicant. Int Arch Occup Environ Health. 2003; 76: 556–576. doi: 10.1007/s00420-003-0458-1 PMID: 12920524

53. Smith CJ, Walcott CJ, Huang W, Maggio V, Grainger J, Patterson DG Jr.. Determination of selected monohydroxy metabolites of 2-, 3- and 4-ring polycyclic aromatic hydrocarbons in urine by solid-phase microextraction and isotope dilution gas chromatography–mass spectrometry. J Chromatogr B. 2002; 778: 157–164. doi: 10.1016/S0378-4347(01)00456-X PMID: 12376123

54. Segal W, Starkey RL. Microbial decomposition of methionine and identity of the resulting sulfur products. J Bacteriol. 1969; 98: 908–913. PMID: 5788717

55. Parliment TH, Kolor MG, Rizzo DJ. Volatile components of Limburger cheese. J Agric Food Chem. 1982; 30: 1006–1008. doi: 10.1021/jf00114a001

56. Chin H-W, Lindsay R c. Volatile Sulfur Compounds Formed in Disrupted Tissues of Different Cabbage Cultivars. J Food Sci. 1993; 58: 835–839. doi: 10.1111/j.1365-2621.1993.tb09370.x

57. Kim S, Choi Y, Choi S, Choi Y, Park T. Dietary camphene attenuates hepatic steatosis and insulin resistance in mice. Obes Silver Spring Md. 2014; 22: 408–417. doi: 10.1002/oby.20554 PMID: 23818423

58. Mochalski P, Krapf K, Ager C, Wiesenhofer H, Agapiou A, Statheropoulos M, et al. Temporal profiling of human urine VOCs and its potential role under the ruins of collapsed buildings. Toxicol Mech Methods. 2012; 22: 502–511. doi: 10.3109/15376516.2012.682664 PMID: 22482743

59. Mochalski P, King J, Unterkofler K, Hinterhuber H, Amann A. Emission rates of selected volatile organic compounds from skin of healthy volunteers. J Chromatogr B Analyt Technol Biomed Life Sci. 2014; 959: 62–70. doi: 10.1016/j.jchromb.2014.04.006 PMID: 24768920

60. Amal H, Leja M, Broza YY, Tisch U, Funka K, Liepniece-Karele I, et al. Geographical variation in the exhaled volatile organic compounds. J Breath Res. 2013; 7: 047102. doi: 10.1088/1752-7155/7/4/047102 PMID: 24184568

61. Nielsen BL, Jerôme N, Saint-Albin A, Rampin O, Maurin Y. Behavioural response of sexually naïve and experienced male rats to the smell of 6-methyl-5-hepten-2-one and female rat faeces. Physiol Behav. 2013; 120: 150–155. doi: 10.1016/j.physbeh.2013.07.012 PMID: 23911690

62. Curran AM, Prada PA, Furton KG. The differentiation of the volatile organic signatures of individuals through SPME-GC/MS of characteristic human scent compounds. J Forensic Sci. 2010; 55: 50–57. doi: 10.1111/j.1556-4029.2009.01236.x PMID: 20002268

63. Nielsen BL, Jerôme N, Saint-Albin A, Thonat C, Briant C, Boué F, et al. A mixture of odorant molecules potentially indicating oestrus in mammals elicits penile erections in male rats. Behav Brain Res. 2011; 225: 584–589. doi: 10.1016/j.bbr.2011.08.026 PMID: 21884731

64. Rampin O, Jérôme N, Briant C, Boué F, Maurin Y. Are oestrus odours species specific? Behav Brain Res. 2006; 172: 169–172. doi: 10.1016/j.bbr.2006.04.005 PMID: 16740322

APPENDIX 6 - SUPPLEMENTARY INFORMATION

Paper I: Alcohol Induced Alterations to the Human Fecal VOC Metabolome

The following supplementary figures are reproduced from Paper I[32].

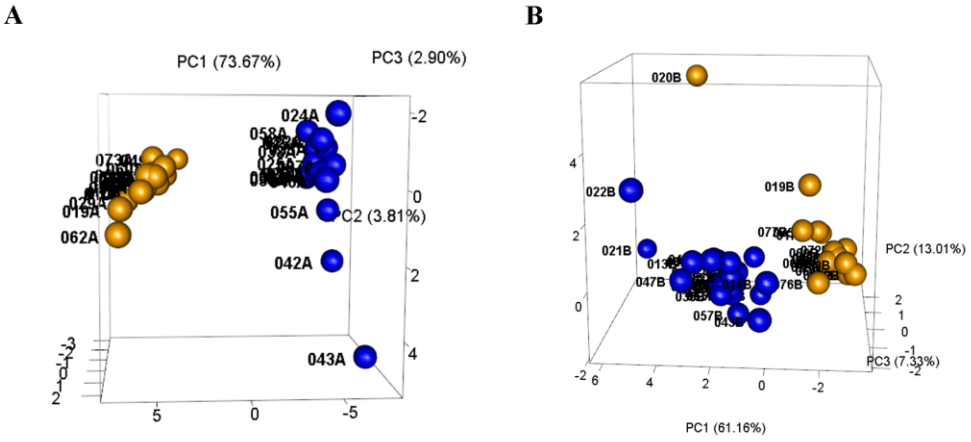


Figure 70. Supplementary Figure 1: PCA.
The resulting three dimensional plot from the endoscopy collected fecal data set is shown in A) and the home collected fecal data set is shown in B).

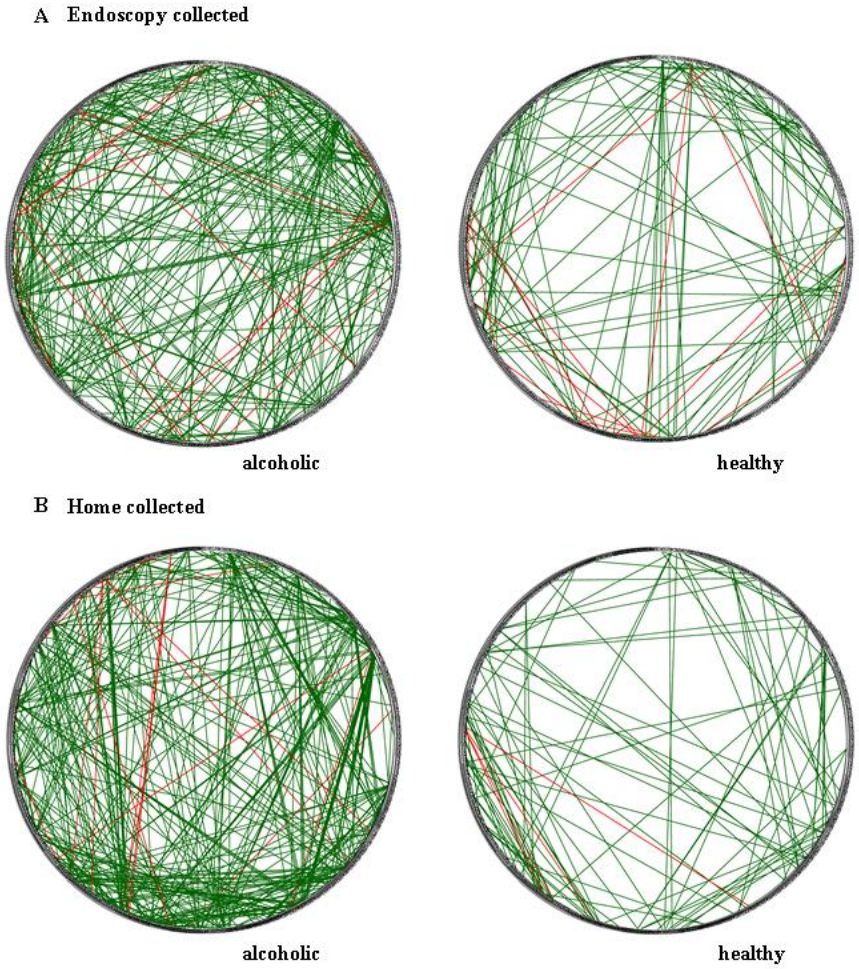


Figure 71. Supplementary Figure 2: Correlation Networks.
Pearson's correlation coefficients were calculated for all metabolites present in at least 21% of the total fecal samples. A Pearson correlation value greater than 0.95 is depicted as a green line between metabolites, while a Pearson correlation value less than-0.95 is depicted as a red line. Metabolites are numerically represented in the network and their placement around the circumference of the network is fixed among the paired plots. Regardless of the approach to fecal collection, the fecal samples from the alcoholic participants have a significantly different correlation network than that seen in the fecal samples from non-alcoholics.

APPENDIX 7 - USING THE EXTRACTED ION CHROMATOGRAM, TOTAL ION CHROMATOGRAM, AND ION ABUNDANCE RANK TO REFINE THE LIST OF STATISTICALLY SIGNIFICANT TOP MOLECULAR FEATURES.

1. Discard all molecular features eluting within the first 1 min or last min of the run, as these regions are most susceptible to chromatographic variation.
2. Generate an EIC (extracted ion chromatogram) for each molecular feature (parameters: symmetrical within 10ppm).
    i. Remove molecular features with a shoulder or a twin peaks (two peaks appearing within 0.2 min of each other). See Figure 72 for an example.

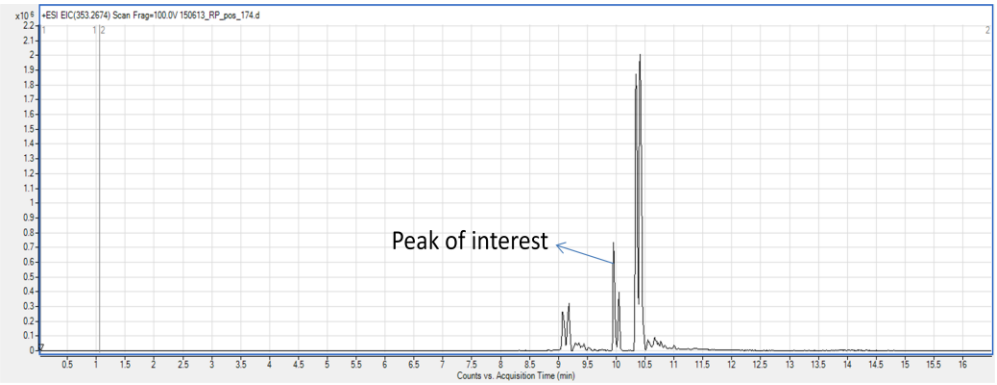


Figure 72. EIC of a molecular feature with twin peaks.
Depicted in the chromatogram is the signal produced by the molecular ion attributed to the molecular feature. The peak of interest is indicated on the plot (this peak was statistically identified as significantly different when comparing cohorts). However, the EIC shows two distinct peaks of differing height in very close proximity to one another, raising concerns that differential comparison may have highlighted the peak of interest by erroneously comparing the wrong peaks. Hence, removal of this peak of interest from subsequent analyses eliminates this risk.

ii. Remove molecular features with an EIC topology resembling noise. See Figure 73 for an example.

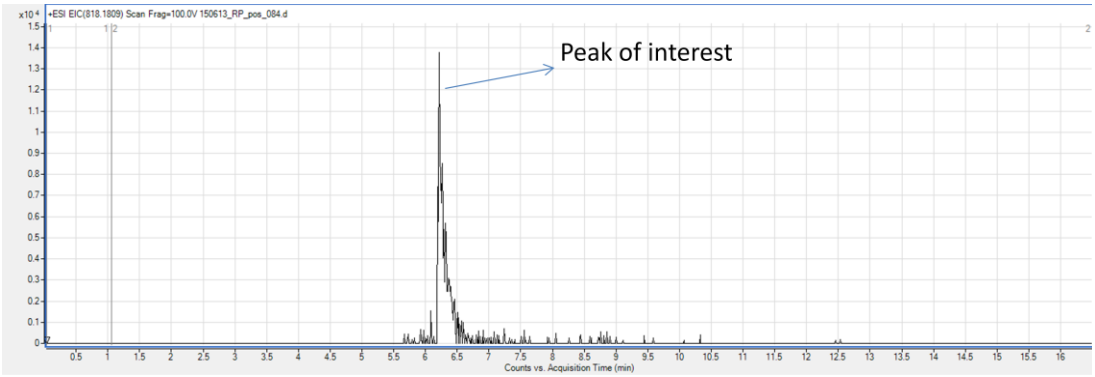


Figure 73. EIC of a molecular feature with a noisy topology.
Depicted in the chromatogram is the signal produced by a molecular ion attributed to a peak of interest. The EIC shows a peak with a topology reflective of background noise (a jagged appearance). Hence, the molecular feature is removed from the data set.

3. Examine the TIC (total ion chromatogram) at the specified retention time. Retain molecular features appearing within well-defined chromatographic peaks (rather than in the baseline). See Figure 74 for an example.

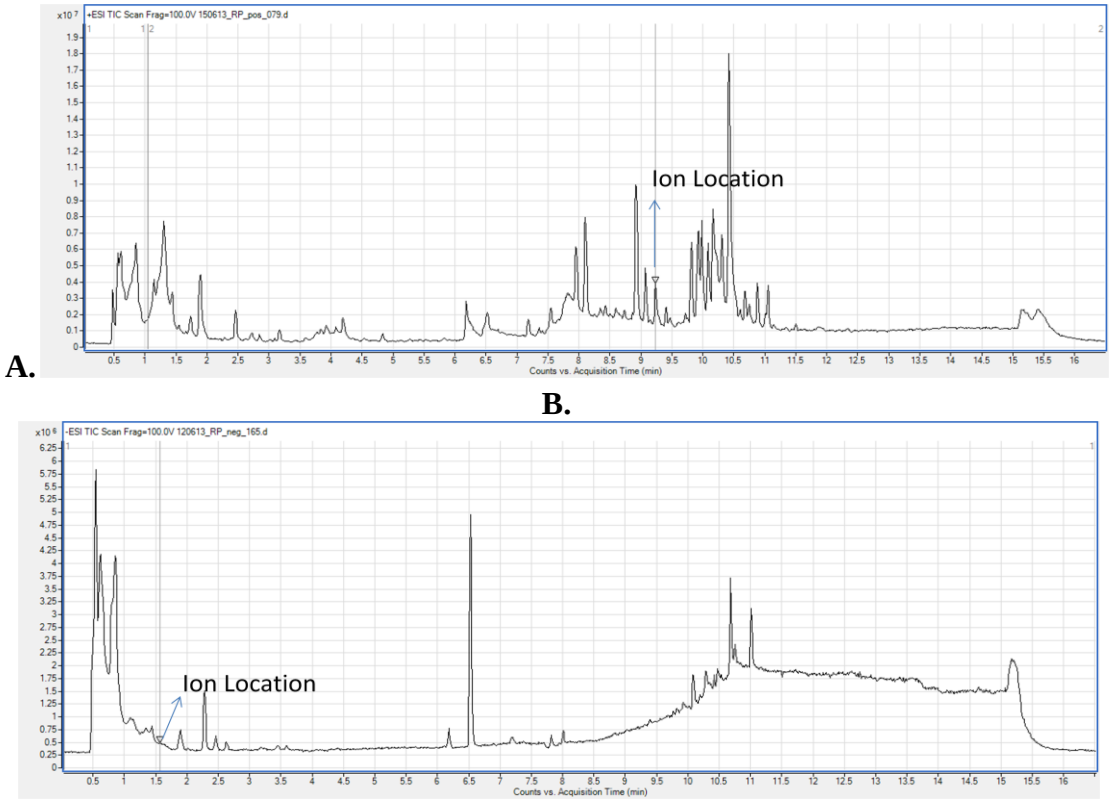


Figure 74. TIC indicating the location of the molecular ion associated with the molecular feature of interest. Depicted in chromatograms A and B is the TIC (the total ion count; the summation of the intensities of all ions acquired at the indicated time point during the sample run). The molecular ion associated with the molecular feature of interest is indicated on the plot and labeled as "ion location". Figure A illustrates an example of the molecular ion of interest appearing within a well-defined chromatographic peak. This molecular feature will be retained in the analysis. Conversely, Figure B demonstrates an example of a molecular ion of interest that does not appear within a well-defined chromatographic peak. This molecular feature will be removed from the analysis.

4. At the apex of the EIC, examine the mass spectra for the indicated retention time of the molecular feature. Rank the molecular feature by the ion's abundance compared to the other ions present. A rank value of 1 indicates the most abundant ion at the specified retention time. Retain the molecular features which are within the top 10 most abundant ions.

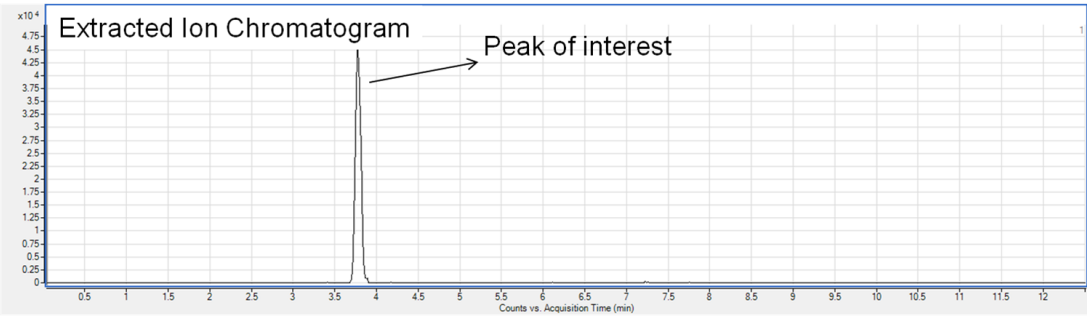
- Note: An easy way to do this in MassHunter Qualitative Analysis is by right clicking on the mass spectra and selecting MS Peaks One. This will provide you with a table indicating the abundance values for all ions within the mass spectra.

APPENDIX 8 - SPECIFIC AIM 2.3: EIC AND TIC COMPILATION OF BASAL VS BASAL+PROBIOTIC AND HIGH FAT VS HIGH FAT+PROBIOTIC

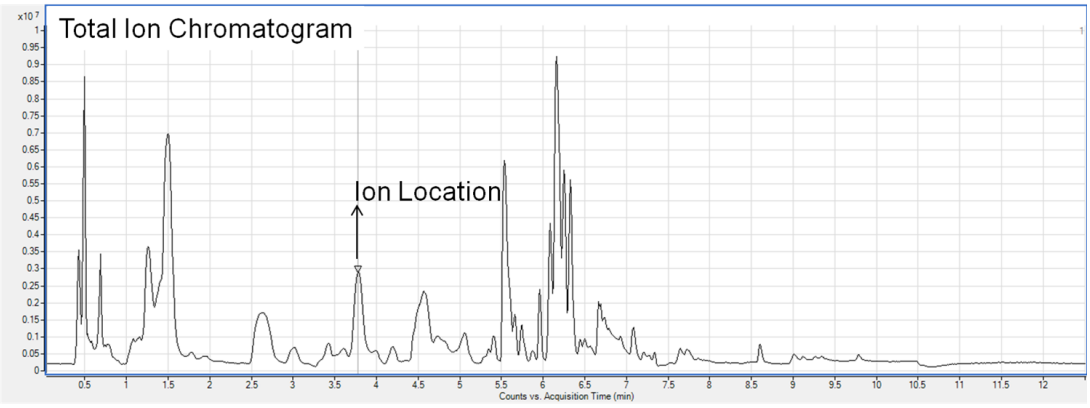
Cortex Basal vs Basal+Probiotic

Metabolite Name: Ion Mass_273.12

HILIC Pos – Ion Rank: #8



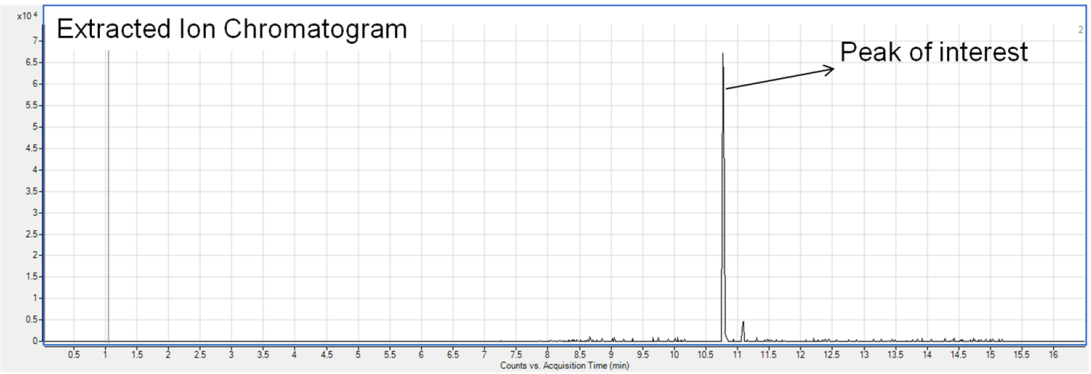
Retention Time:	Ion mass:
Apex of EIC: 3.778 min	Apex of EIC: 273.1207
Spreadsheet: 3.80112 min	Spreadsheet: 273.1213



Cortex High Fat vs High Fat+Probiotic

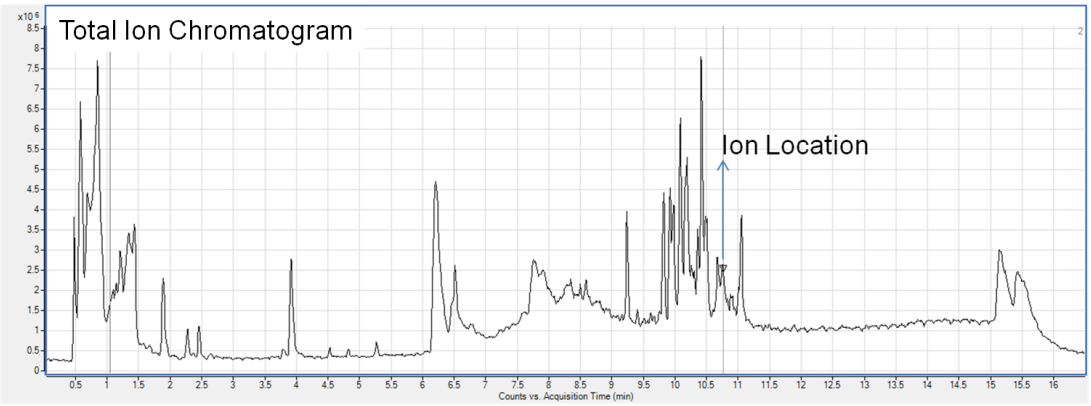
Metabolite Name: Ion Mass_355.26

RP Pos – Ion Rank: #2



Retention Time:
Apex of EIC: 10.769 min
Spreadsheet: 10.423 min

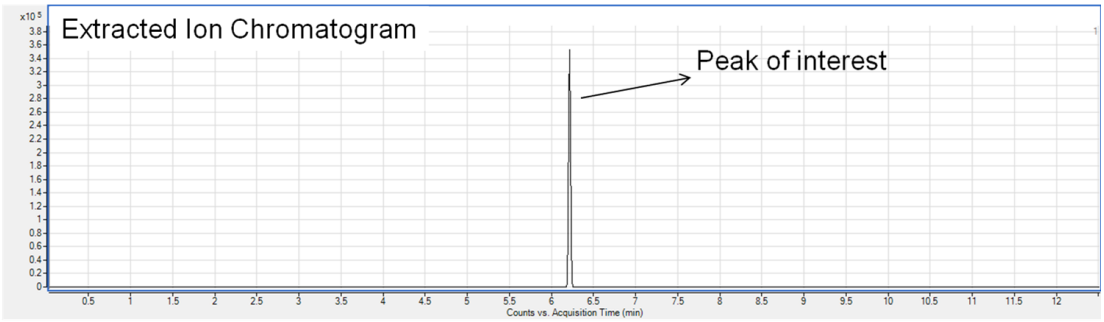
Ion mass:
Apex of EIC: 355.2606
Spreadsheet: 355.2603



Heart High Fat vs High Fat+Probiotic

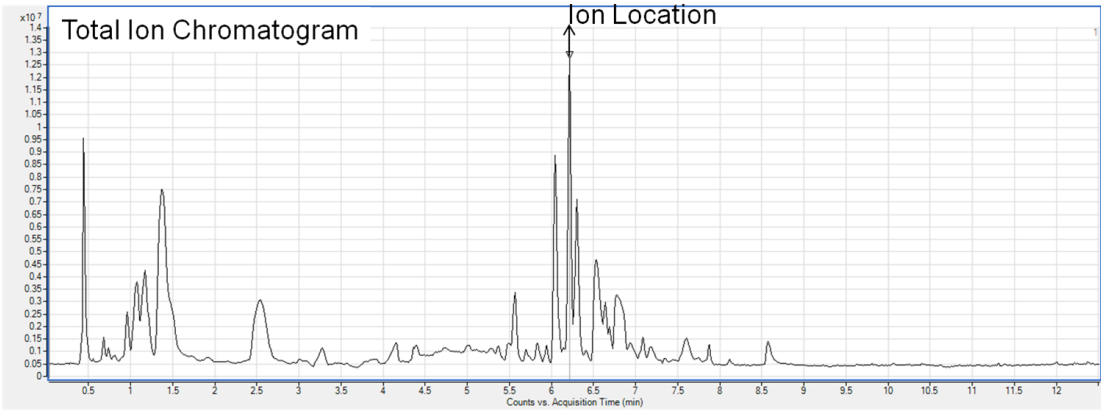
Metabolite Name: Ion Mass_348.15

HILIC Neg – Ion Rank: #8



Retention Time:
Apex of EIC: 6.174 min
Spreadsheet: 6.218 min

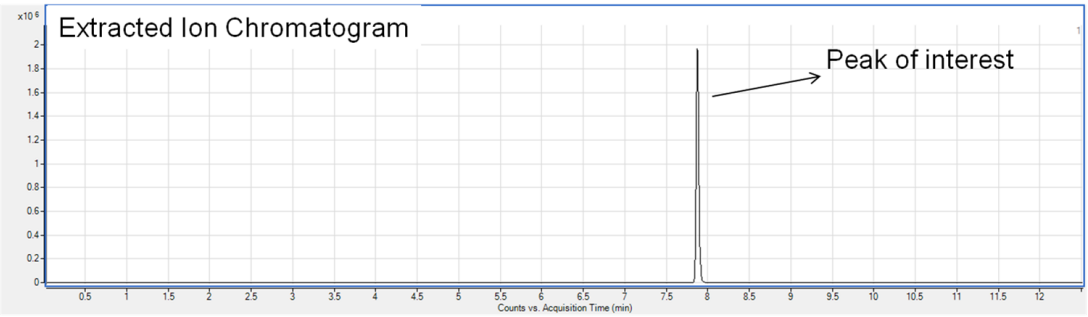
Ion mass:
Apex of EIC: 348.1424
Spreadsheet: 348.14236



Kidney Basal vs Basal+Probiotic

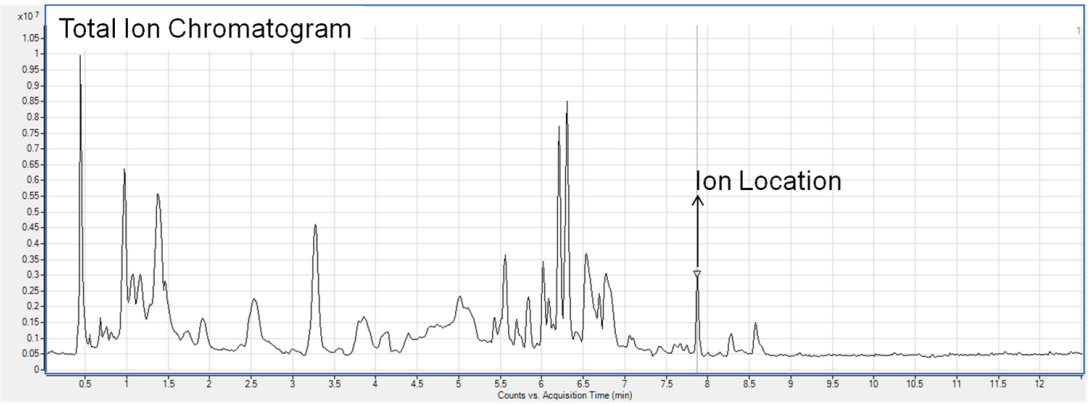
Metabolite Name: Uridine diphosphate-N-acetylglucosamine

HILIC Neg – Ion Rank: #1



Retention Time:
Apex of EIC: 7.874 min
Spreadsheet: 7.897 min

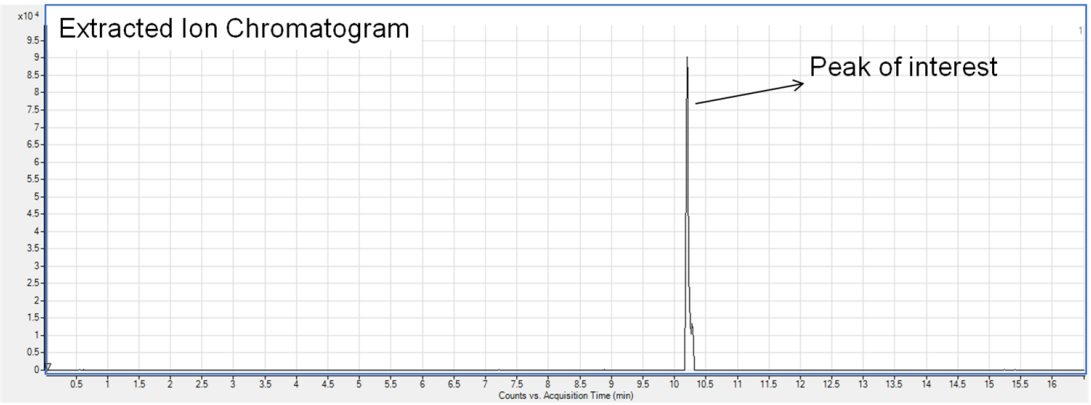
Ion mass:
Apex of EIC: 606.0763
Spreadsheet: 606.07587



Kidney High Fat vs High Fat+Probiotic

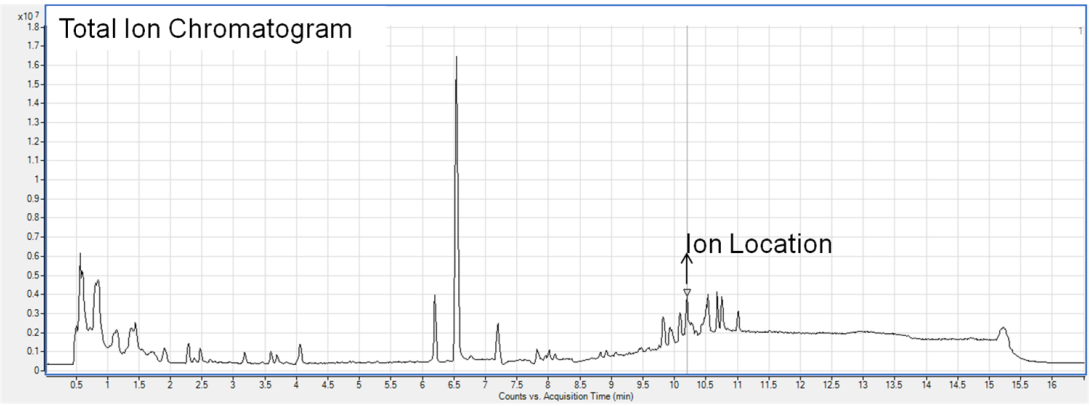
Metabolite Name: Ion Mass_498.19

RP Neg – Ion Rank: #9



Retention Time:
Apex of EIC: 10.2 min
Spreadsheet: 10.918 min

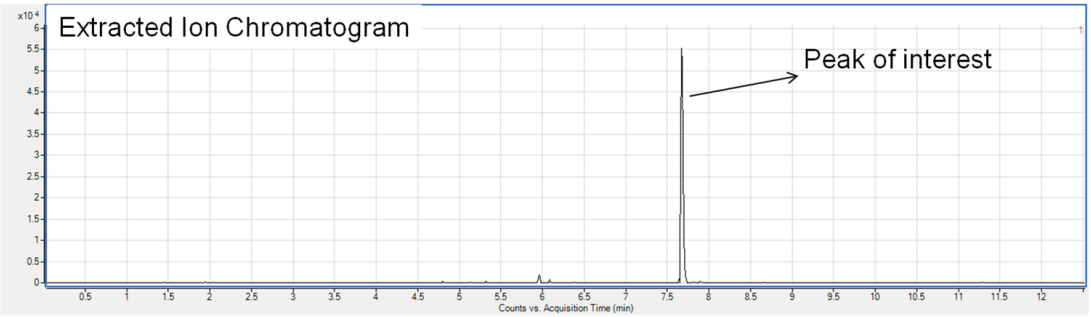
Ion mass:
Apex of EIC: 498.1928
Spreadsheet: 498.193



Liver Basal vs Basal+Probiotic

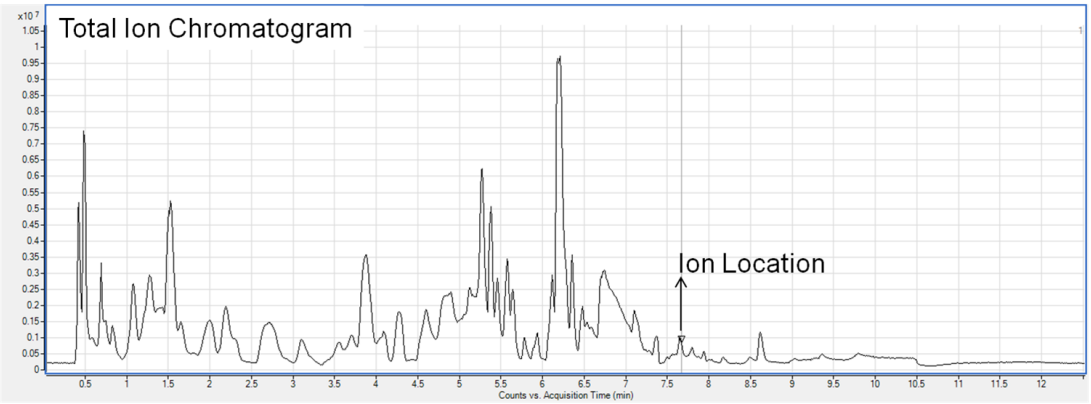
Metabolite Name: Saccharopine

HILIC Pos – Ion Rank: #3



Retention Time:
Apex of EIC: 7.671 min
Spreadsheet: 7.62612 min

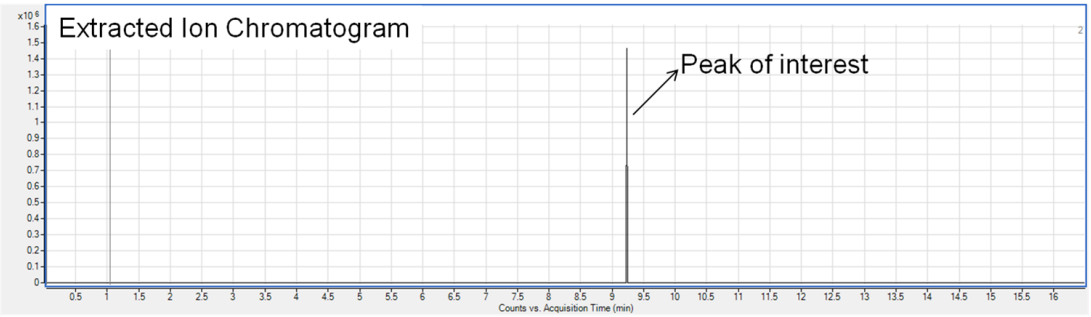
Ion mass:
Apex of EIC: 277.1387
Spreadsheet: 277.13974



Liver High Fat vs High Fat+Probiotic

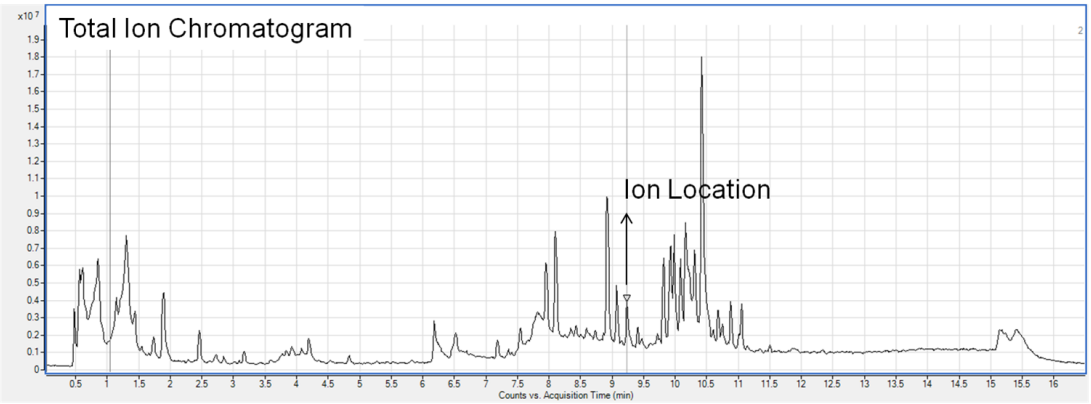
Metabolite Name: Ion Mass_279.16

RP Pos – Ion Rank: #1



Retention Time:
Apex of EIC: 9.236 min
Spreadsheet: 9.2378 min

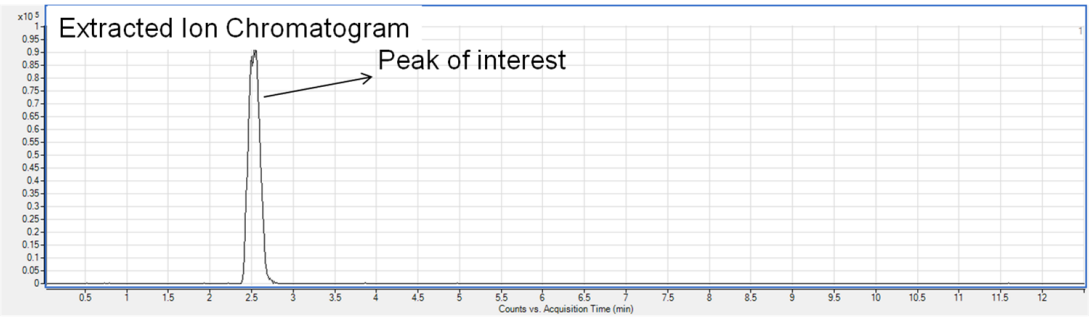
Ion mass:
Apex of EIC: 279.1656
Spreadsheet: 279.16388



Muscle High Fat vs High Fat+Probiotic

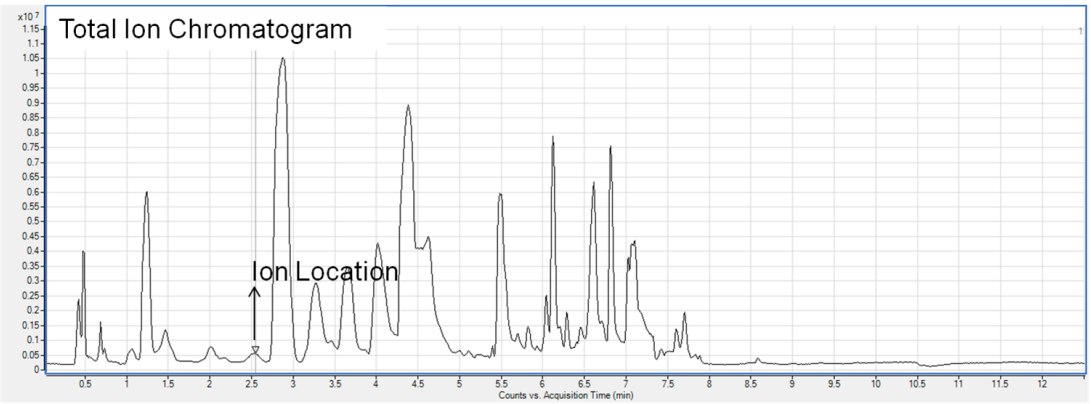
Metabolite Name: Ion Mass_291.07

HILIC Pos – Ion Rank: #1



Retention Time:
Apex of EIC: 2.550 min
Spreadsheet: 2.79323 min

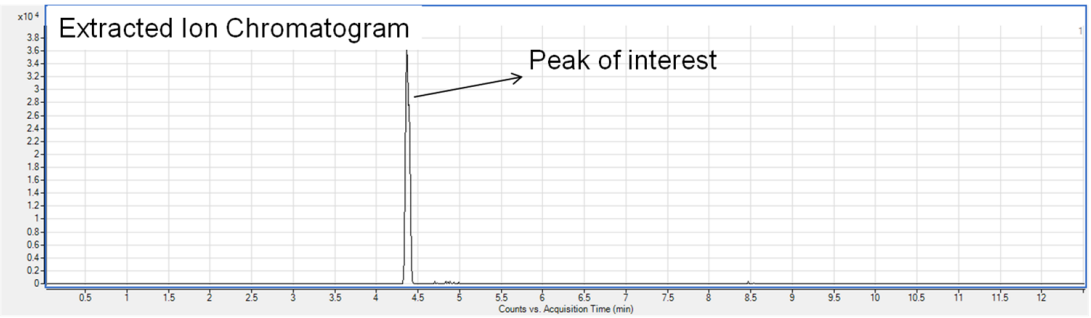
Ion mass:
Apex of EIC: 291.0703
Spreadsheet: 291.07022



Pancreas Basal vs Basal+Probiotic

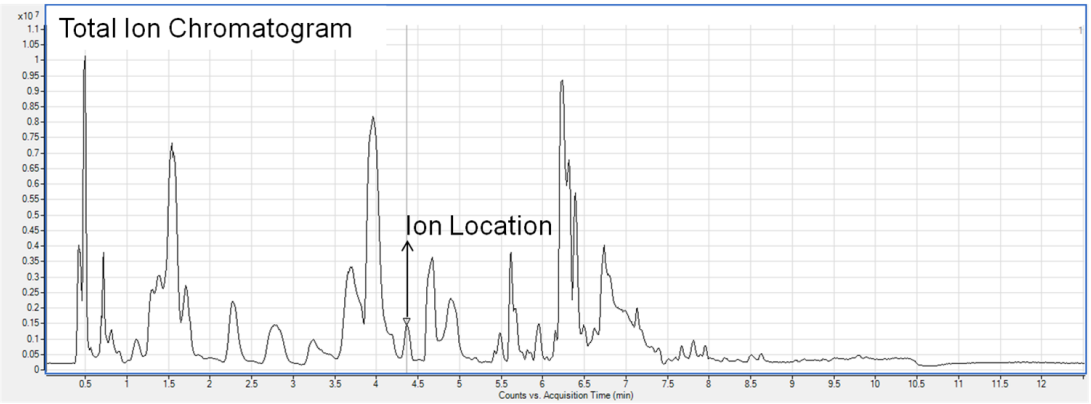
Metabolite Name: Ion Mass_322.05

HILIC Pos – Ion Rank: #7



Retention Time:
Apex of EIC: 4.363 min
Spreadsheet: 4.32996 min

Ion mass:
Apex of EIC: 322.0544
Spreadsheet: 322.0545



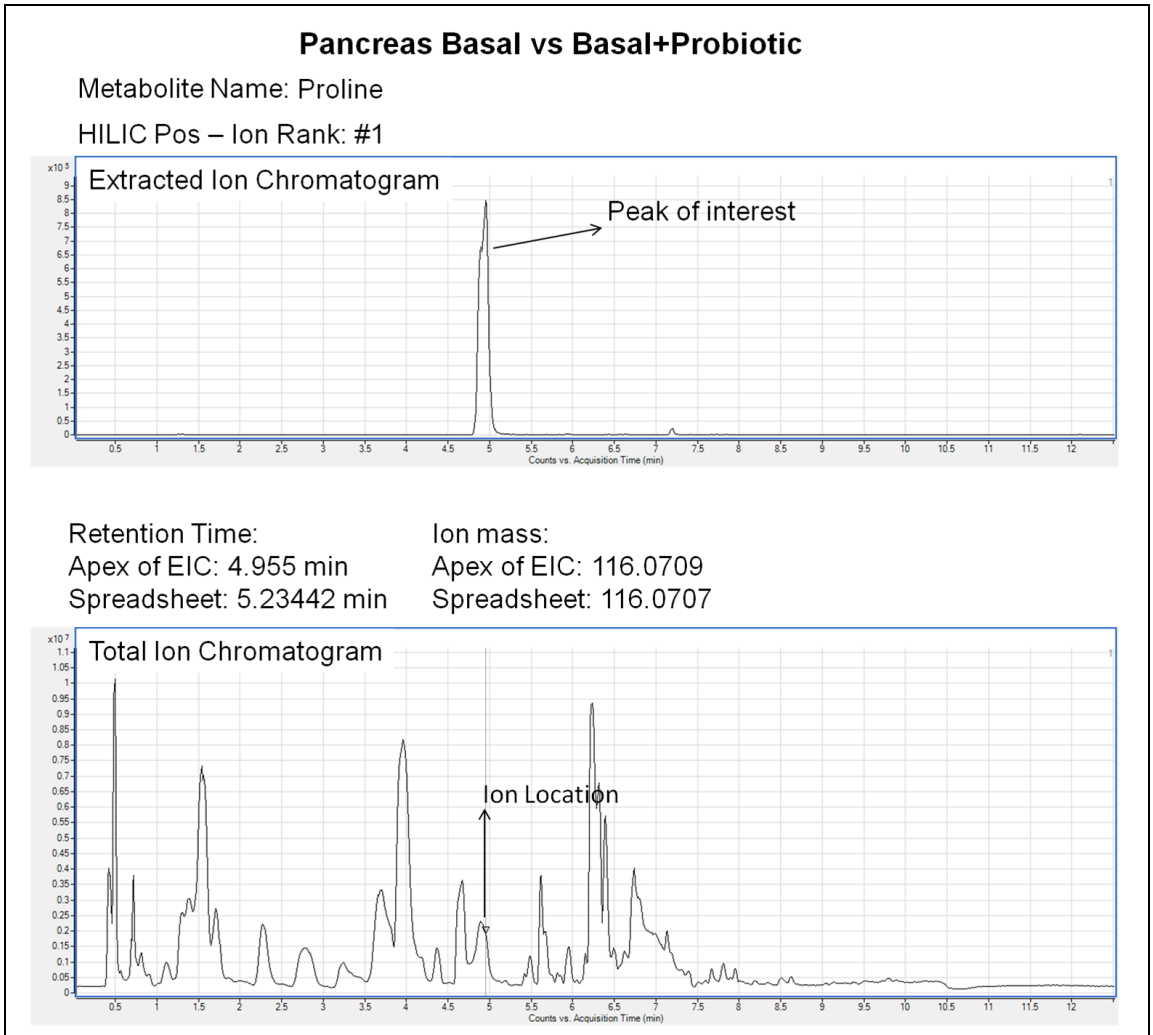


Figure 75. Extracted Ion Chromatograms (EIC) and Total Ion Chromatograms (TIC) of the Top Molecular Features associated with probiotic induced alterations on a nutritionally balanced basal diet or a high fat diet.

The provided chromatograms depict the EIC and TIC of each molecular feature deemed statistically significant in my analysis. Using the method described in Appendix 7, I used the chromatograms shown to refine my data set to include the top 10 molecular features depicted here.

APPENDIX 9 - SPECIFIC AIM 2.3: MS/MS FRAGMENTATION PATTERN MATCHING FOR BASAL VS BASAL+PROBIOTIC AND HIGH FAT VS HIGH FAT+PROBIOTIC

1. Uridine Diphosphate-N-Acetylglucosamine

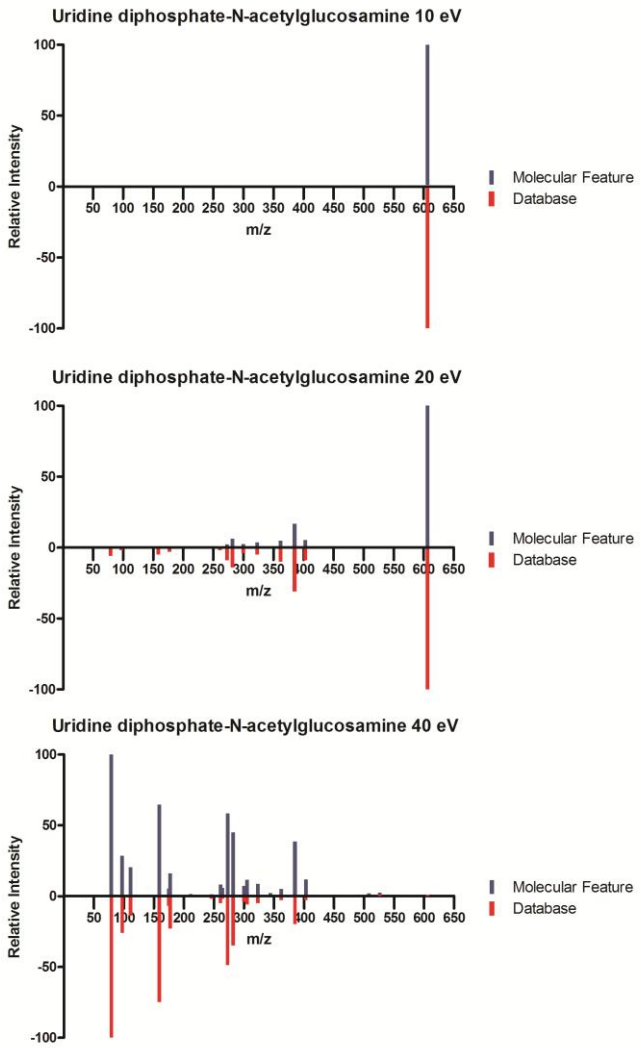


Figure 76. MS/MS comparative analysis for Uridine diphosphate-N-acetyl glucosamine using data acquired by our LC-QToF against data acquired from the Metlin Database at collision energies 10 eV, 20 eV, and 40 eV.

2. *Saccharopine*

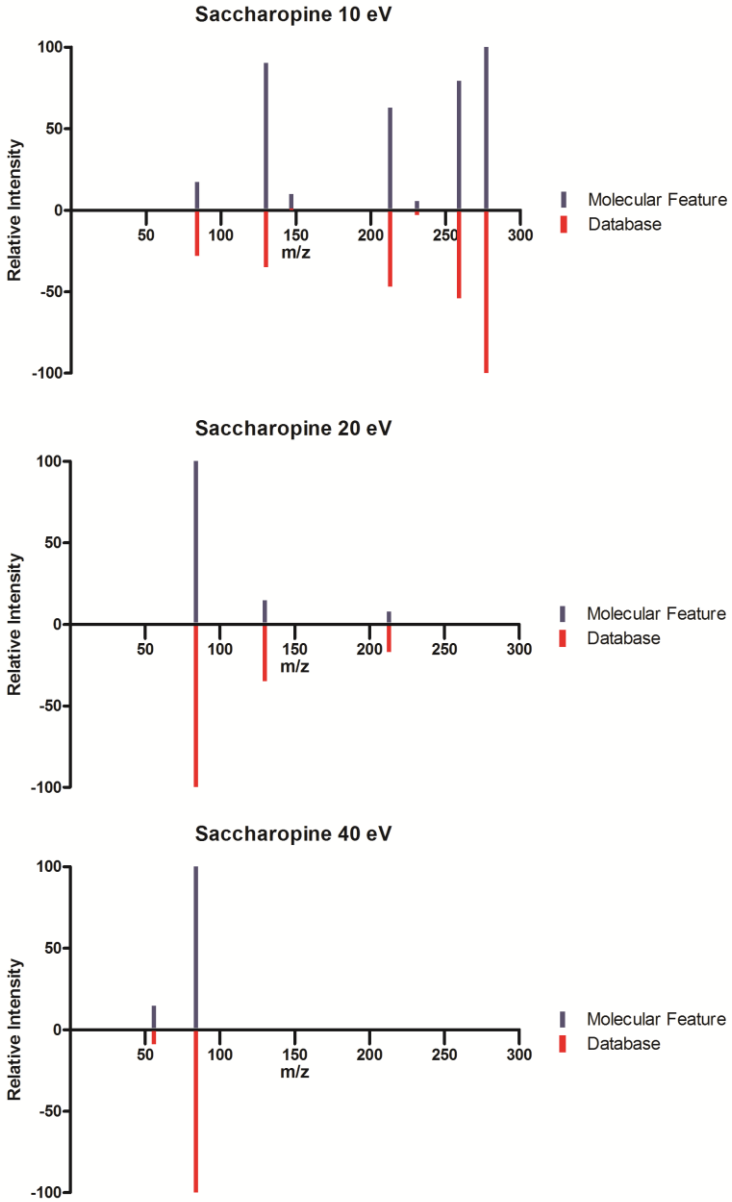


Figure 77. MS/MS comparative analysis for Saccharopine using data acquired by our LC-QToF against data acquired from the Metlin Database at collision energies 10 eV, 20 eV, and 40 eV.

3. Proline

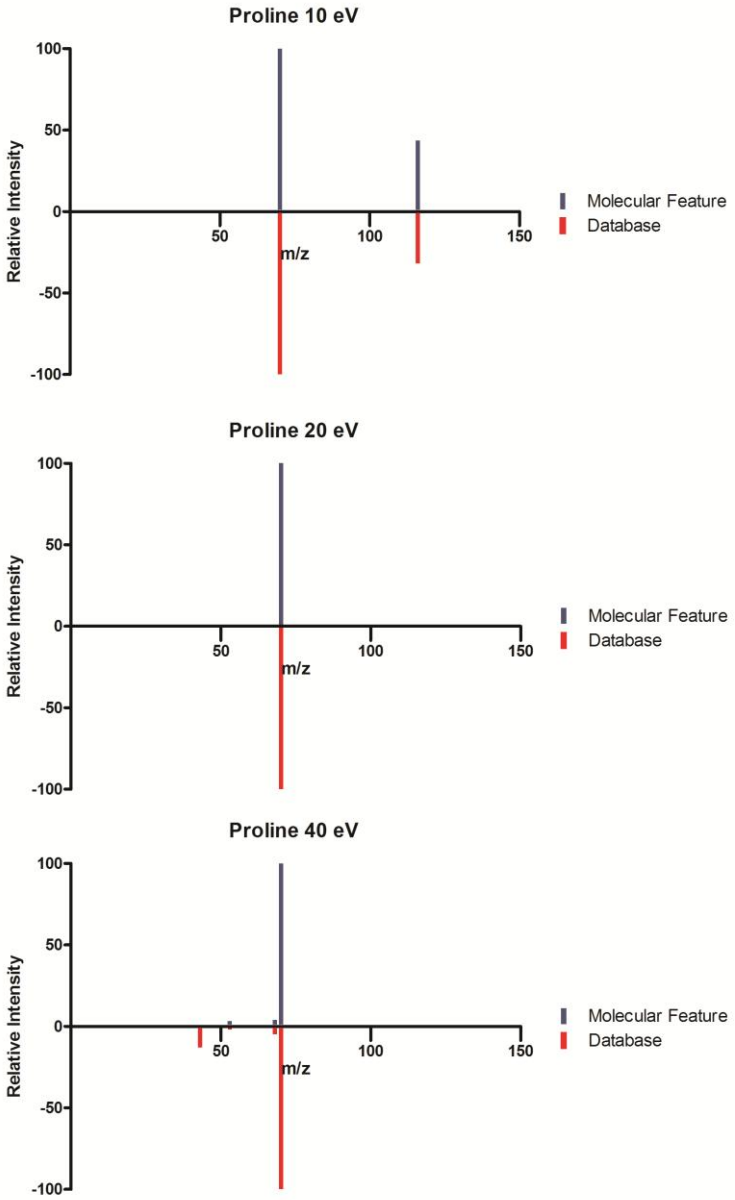


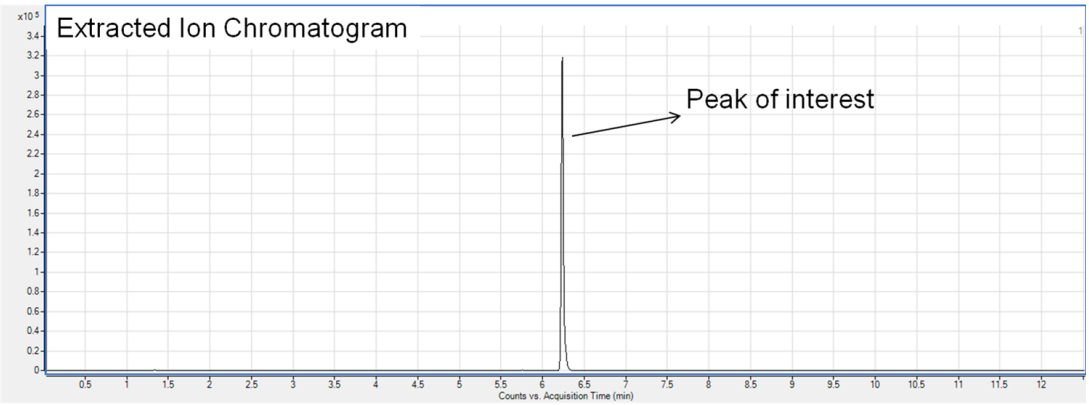
Figure 78. MS/MS comparative analysis for Proline using data acquired by our LC-QToF against data acquired from the Metlin Database at collision energies 10 eV, 20 eV, and 40 eV.

APPENDIX 10 - SPECIFIC AIM 2.3: EIC AND TIC COMPILATION PROBIOTIC ASSOCIATED REVERSIONS FROM A HIGH FAT DIET PHENOTYPE TO A BASAL DIET PHENOTYPE

Cortex

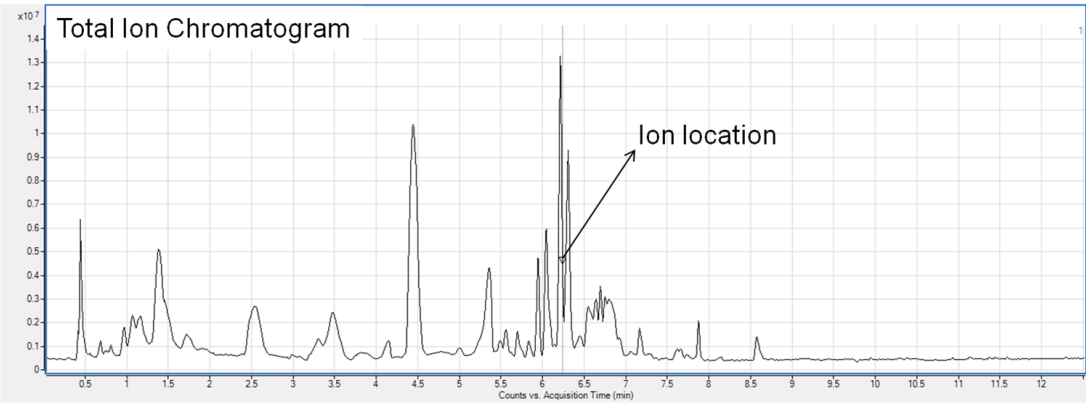
Metabolite Name: S-adenosylhomocysteine

HILIC Neg – Ion Rank: # 3



Retention Time:
Apex of EIC: 6.243 min
Spreadsheet: 6.243 min

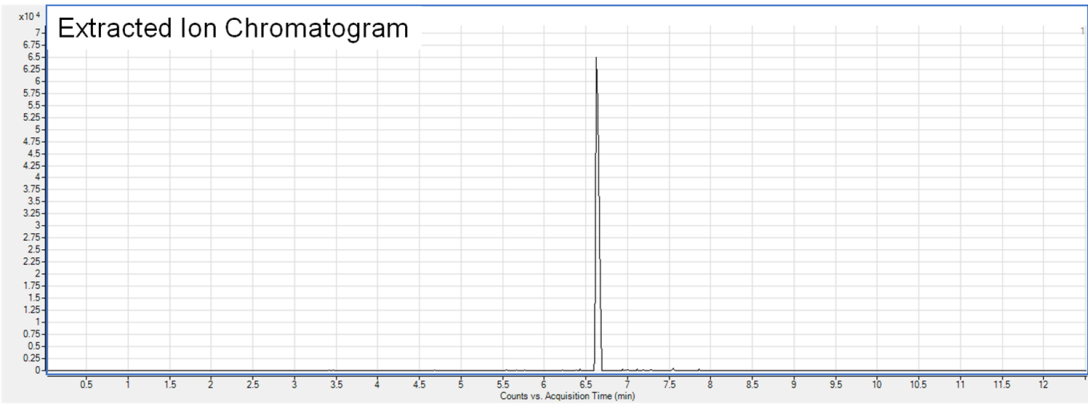
Ion mass:
Apex of EIC: 383.1151
Spreadsheet: 383.11505



Heart

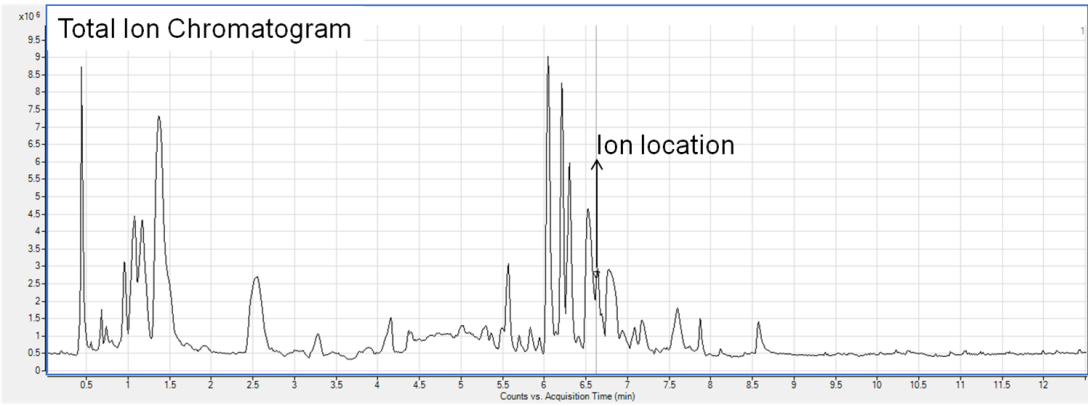
Metabolite Name: CDP-glycerol

HILIC Neg – Ion Rank: # 6



Retention Time:
Apex of EIC: 6.629 min
Spreadsheet: 6.63 min

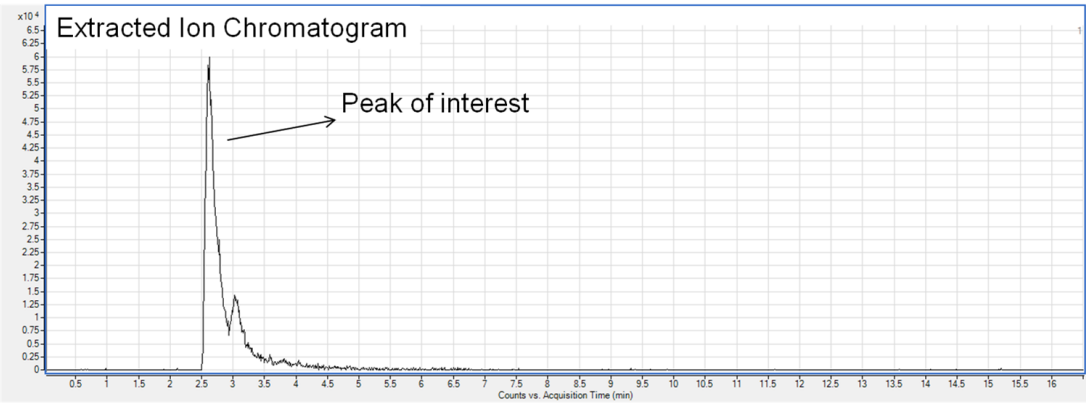
Ion mass:
Apex of EIC: 476.0751
Spreadsheet: 476.07495



Kidney

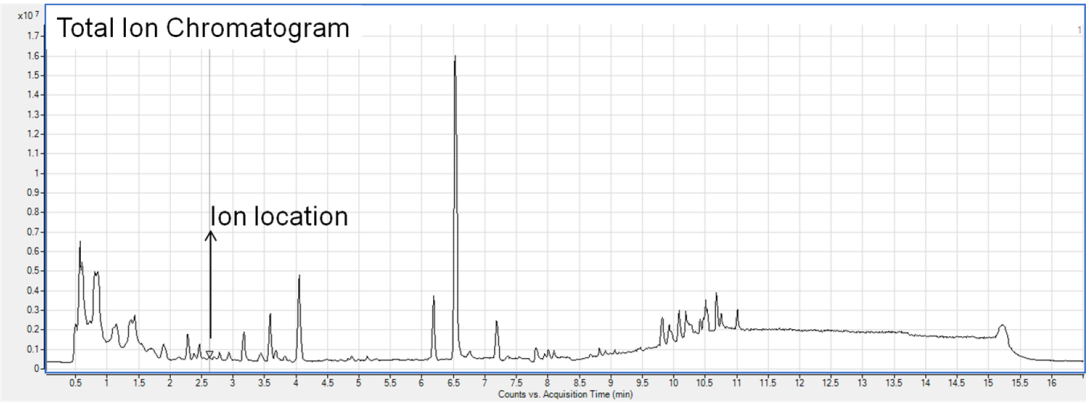
Metabolite Name: Indoxylsulfuric acid

RP Neg – Ion Rank: # 2



Retention Time:
Apex of EIC: 2.622 min
Spreadsheet: 2.619 min

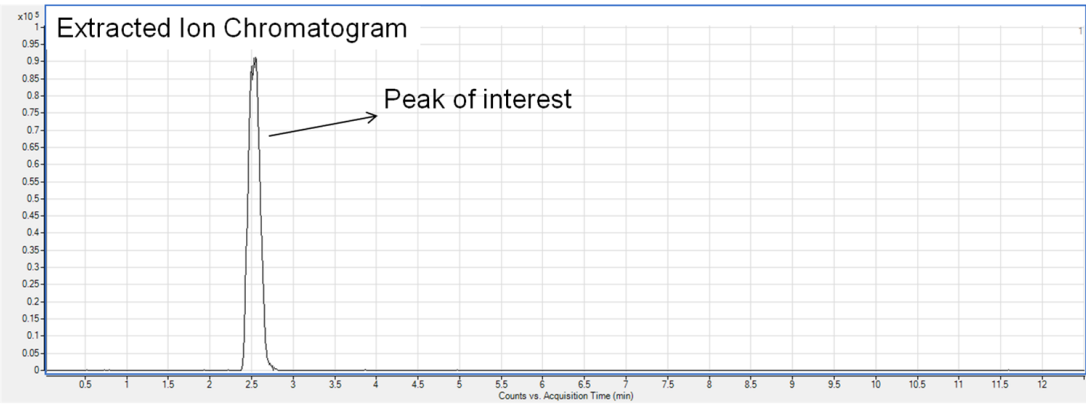
Ion mass:
Apex of EIC: 212.0022
Spreadsheet: 212.00214



Muscle

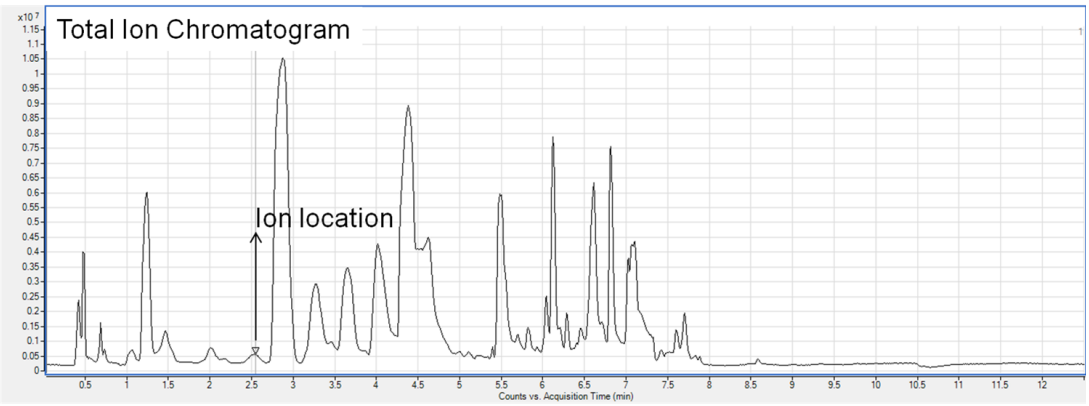
Metabolite Name: D-Sedoheptulose 7-phosphate

HILIC Pos – Ion Rank: # 1



Retention Time:
Apex of EIC: 2.55 min
Spreadsheet: 2.79 min

Ion mass:
Apex of EIC: 291.0703
Spreadsheet: 291.07022



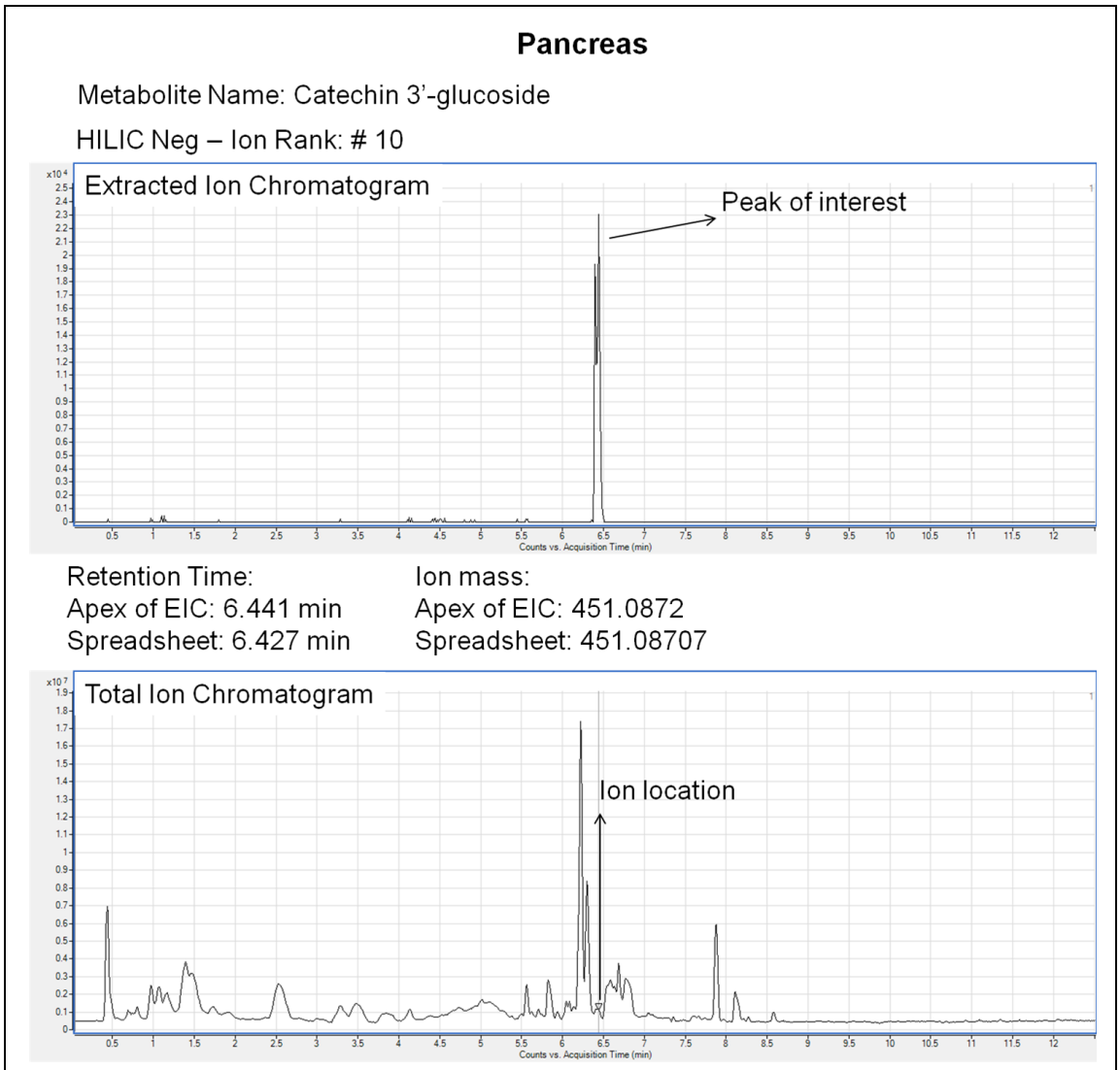


Figure 79. Extracted Ion Chromatograms (EIC) and Total Ion Chromatograms (TIC) of the Top Molecular Features associated with probiotic induced reversions from a metabolome indicative of a pig consuming a high fat diet to one comparable to a pig consuming a on a nutritionally balanced basal diet.
The provided chromatograms depict the EIC and TIC of each molecular feature deemed statistically significant in my analysis. Using the method described in Appendix 7, I used the chromatograms shown to refine my data set to include the top 5 molecular features depicted here.

APPENDIX 11 - SPECIFIC AIM 2.3: MS/MS FRAGMENTATION PATTERN MATCHING FOR PROBIOTIC ASSOCIATED REVERSIONS FROM A HIGH FAT PHENOTYPE

1. S-adenosylhomocysteine

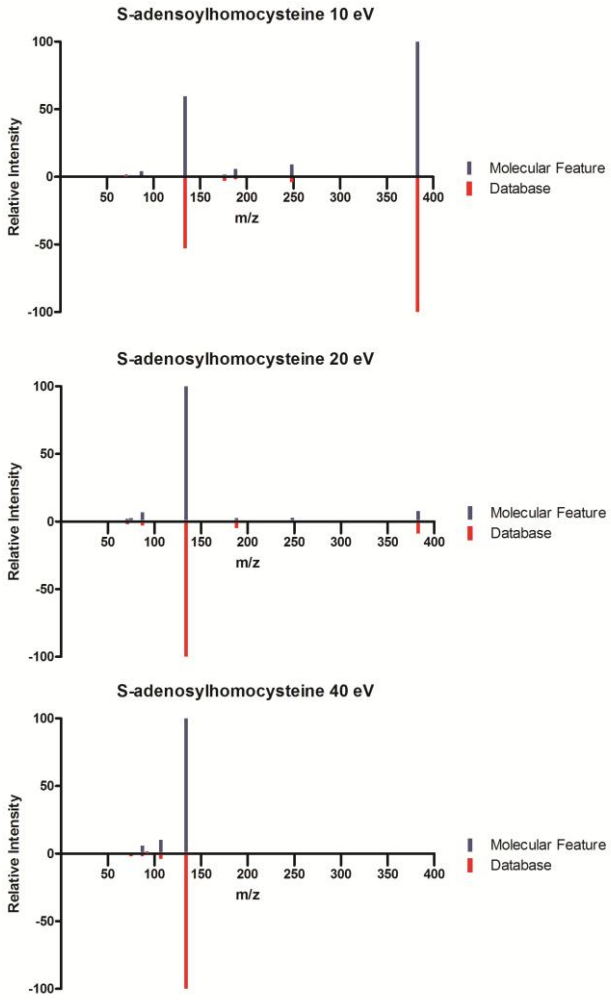


Figure 80. MS/MS comparative analysis for S-adenosylhomocysteine using data acquired by our LC-QToF against data acquired from the Metlin Database at collision energies 10 eV, 20 eV, and 40 eV.

2. Indoxyl sulfate

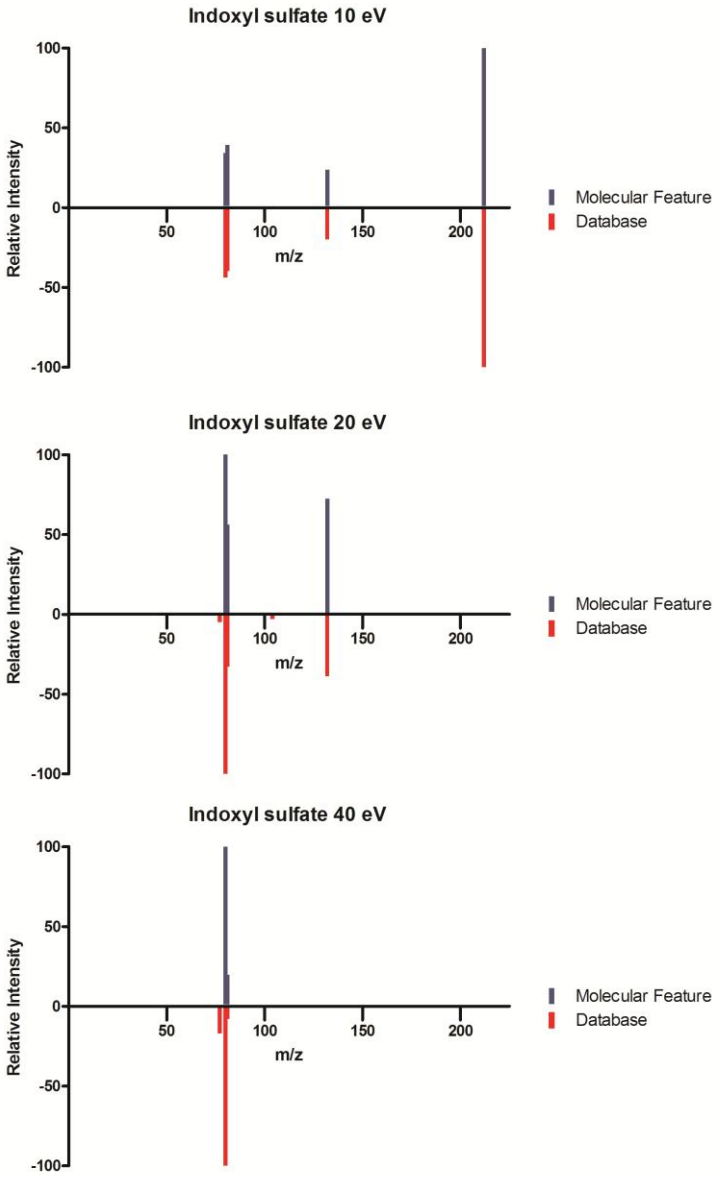
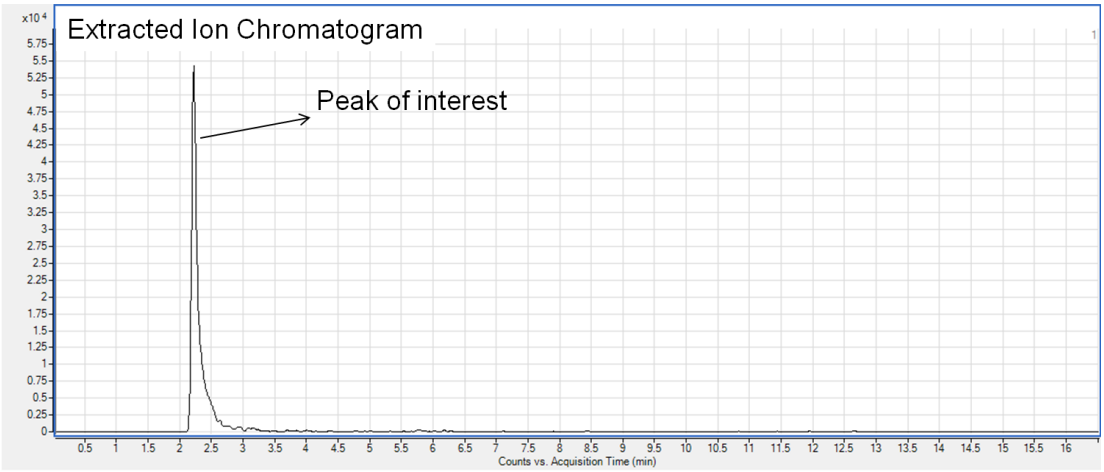


Figure 81. MS/MS comparative analysis for indoxyl sulfate using data acquired by our LC-QToF against data acquired from the Metlin Database at collision energies 10 eV, 20 eV, and 40 eV.

APPENDIX 12 - SPECIFIC AIM 2.4: EIC AND TIC COMPILATION OF TOP METABOLITES ASSOCIATED WITH INFLAMMATORY PATHWAYS

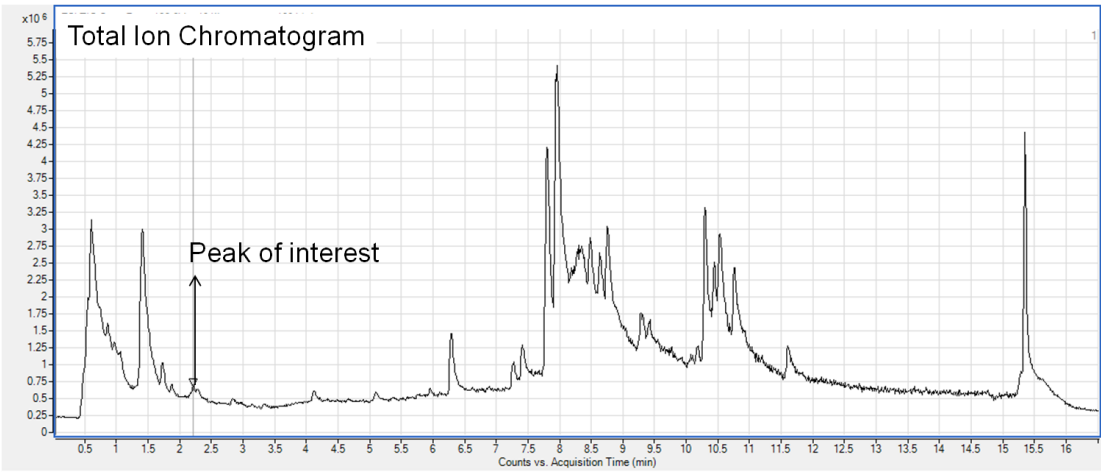
Metabolite Name: Tryptophan

RP Neg – Ion Rank: # 1



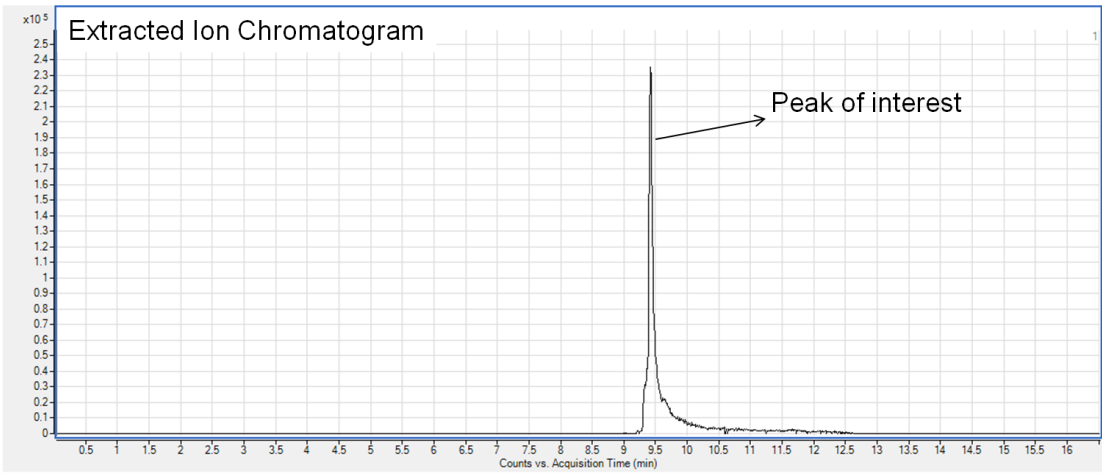
Retention Time:
Apex of EIC: 2.212 min
Spreadsheet: 2.195 min

Ion mass:
Apex of EIC: 203.0816
Spreadsheet: 203.0813



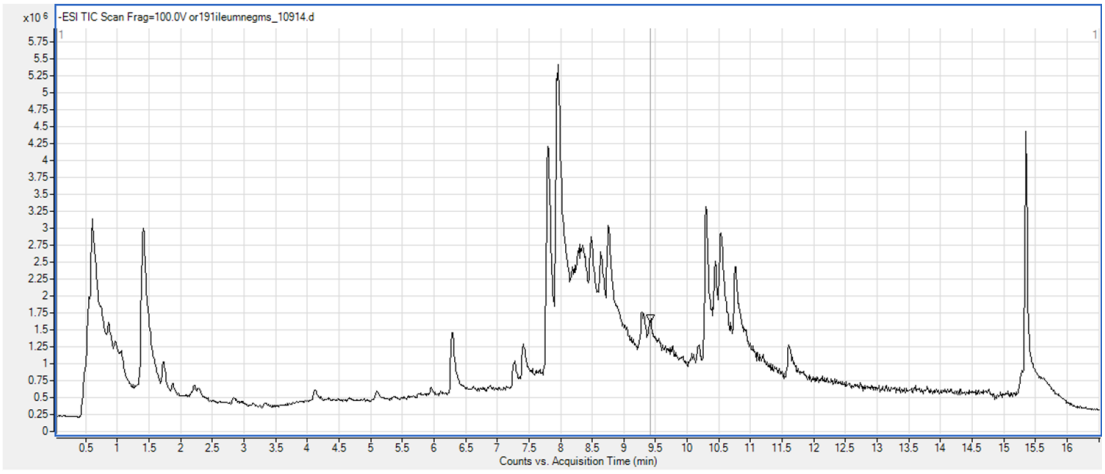
Metabolite Name: 12-HETE

RP Neg – Ion Rank: # 1



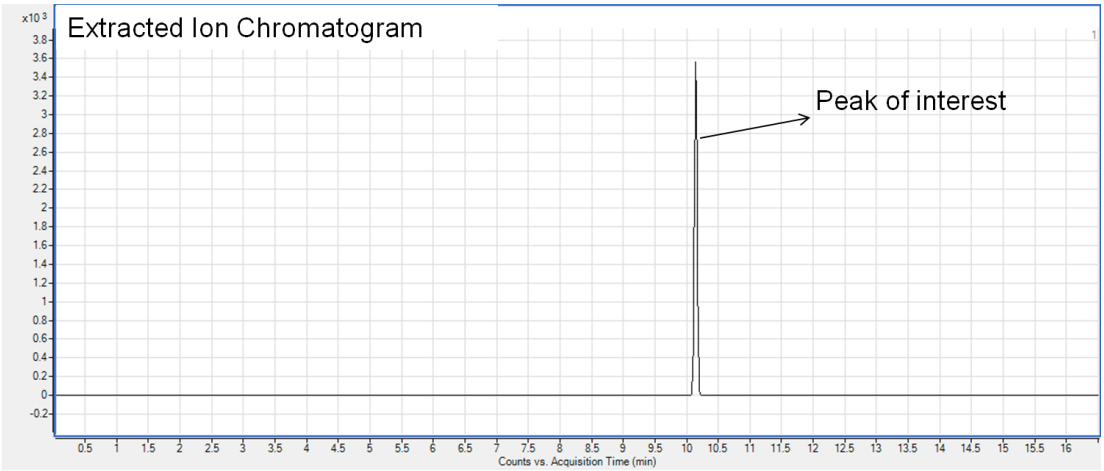
Retention Time:
Apex of EIC: 9.416 min
Spreadsheet: 9.406 min

Ion mass:
Apex of EIC: 319.2255
Spreadsheet: 319.2255



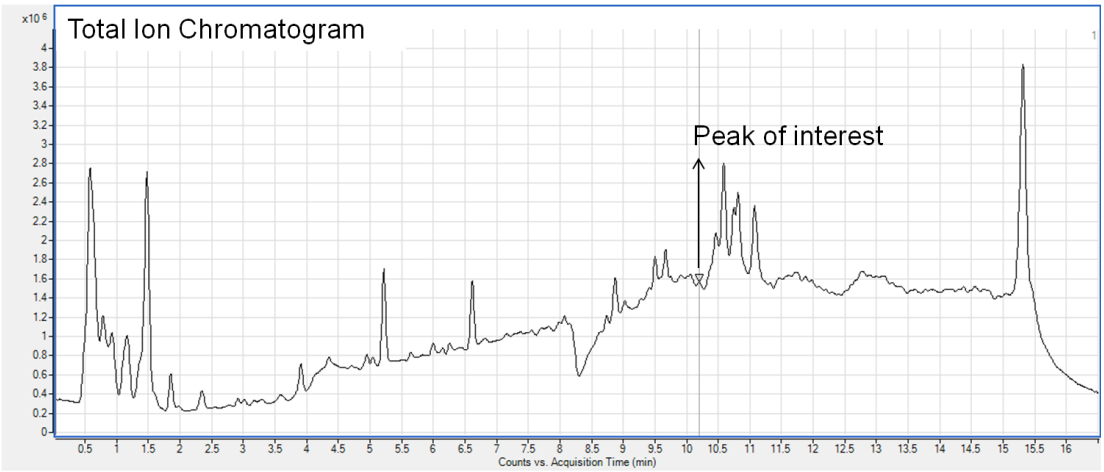
Metabolite Name: 6-Ketoprostaglandin F1a

RP Neg – Ion Rank: # 2



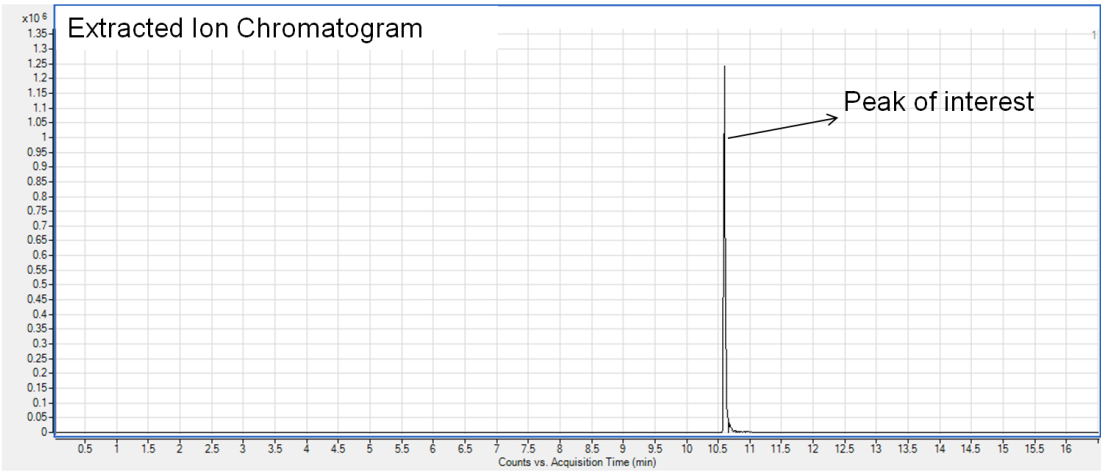
Retention Time:
Apex of EIC: 10.055 min
Spreadsheet: 10.166 min

Ion mass:
Apex of EIC: 369.3020
Spreadsheet: 369.3004



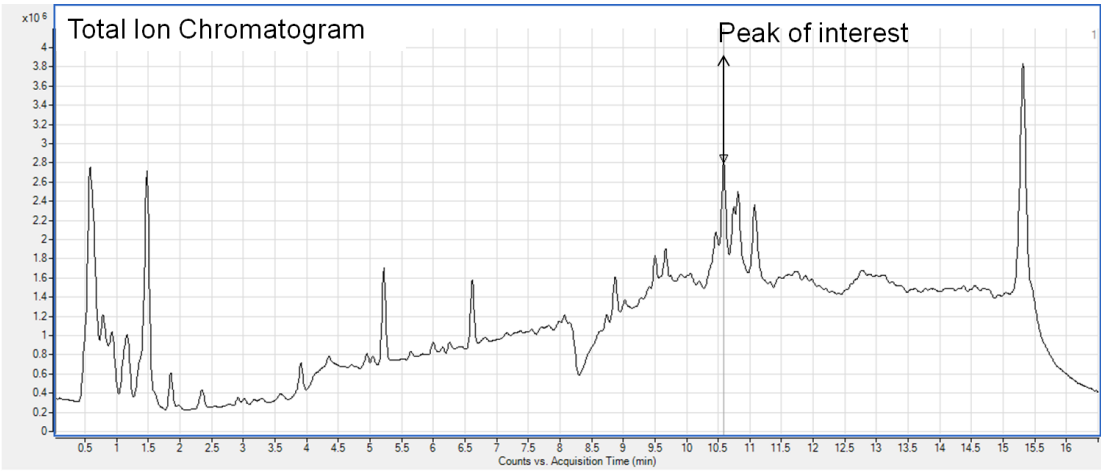
Metabolite Name: 8-iprostane

RP Neg – Ion Rank: # 1



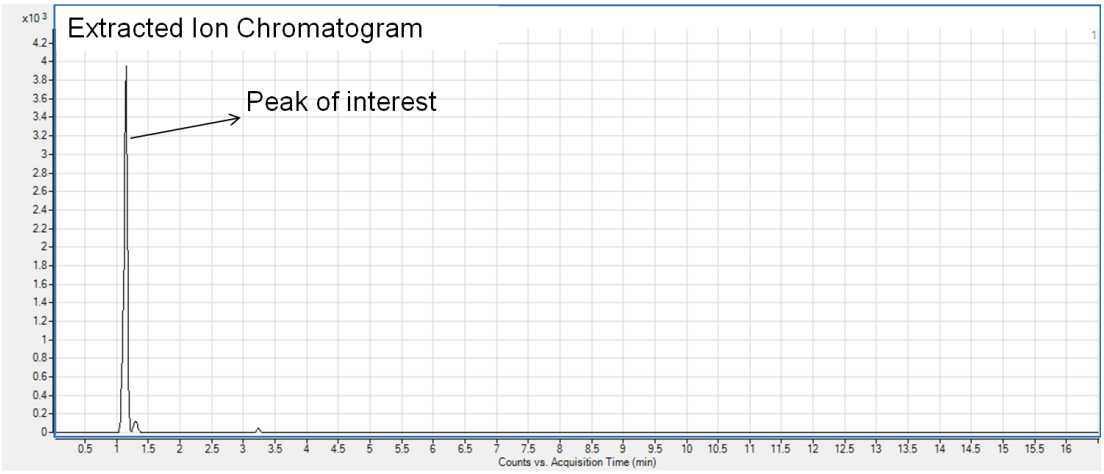
Retention Time:
Apex of EIC: 10.595 min
Spreadsheet: 10.711 min

Ion mass:
Apex of EIC: 279.2349
Spreadsheet: 279.2324



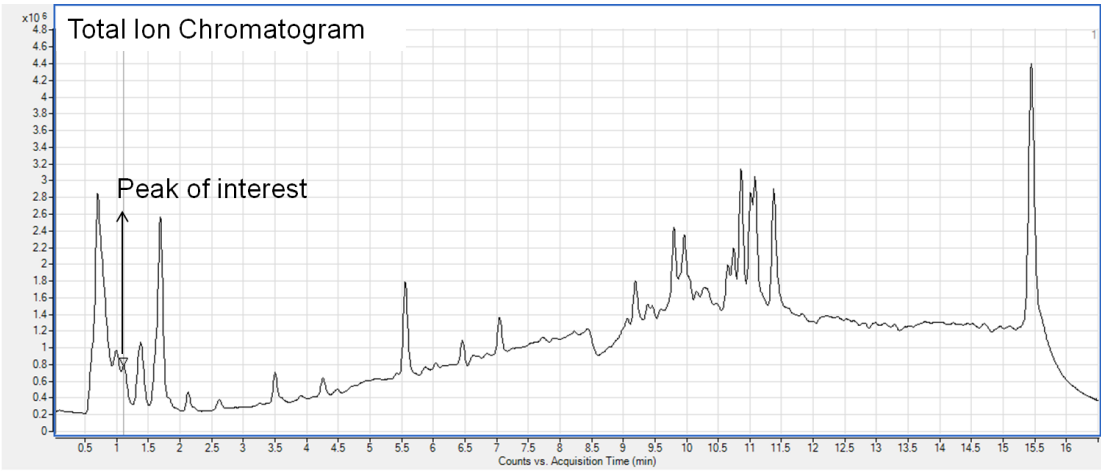
Metabolite Name: Niacin

RP Neg – Ion Rank: # 3



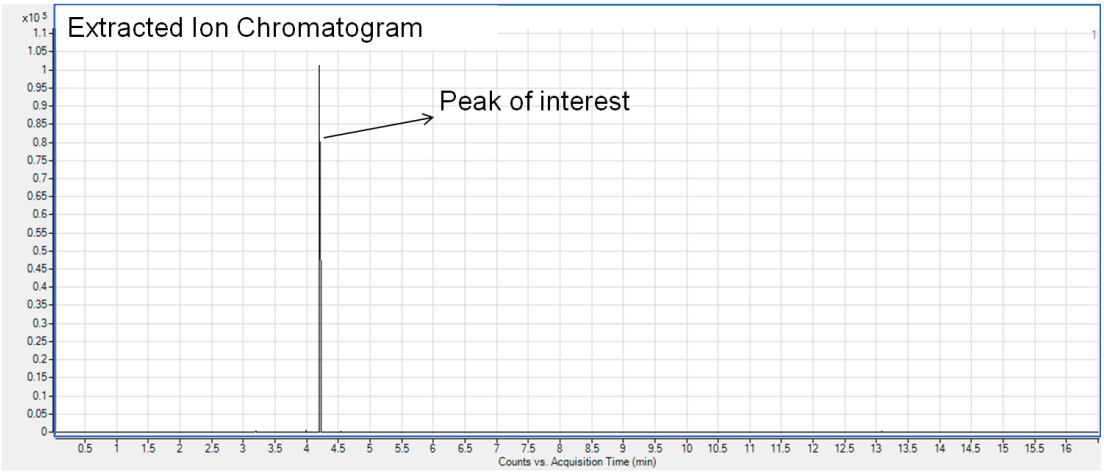
Retention Time:
Apex of EIC: 1.148 min
Spreadsheet: 1.024 min

Ion mass:
Apex of EIC: 122.0253
Spreadsheet: 122.0246



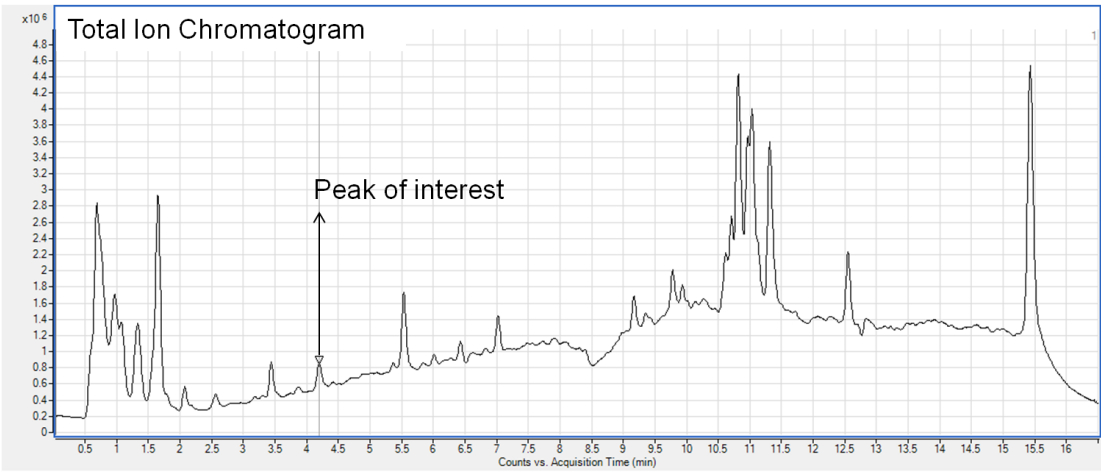
Metabolite Name: 5-hydroxyindoleacetic acid

RP Neg – Ion Rank: # 2



Retention Time:
Apex of EIC: 4.204 min
Spreadsheet: 4.060 min

Ion mass:
Apex of EIC: 190.0513
Spreadsheet: 190.0506



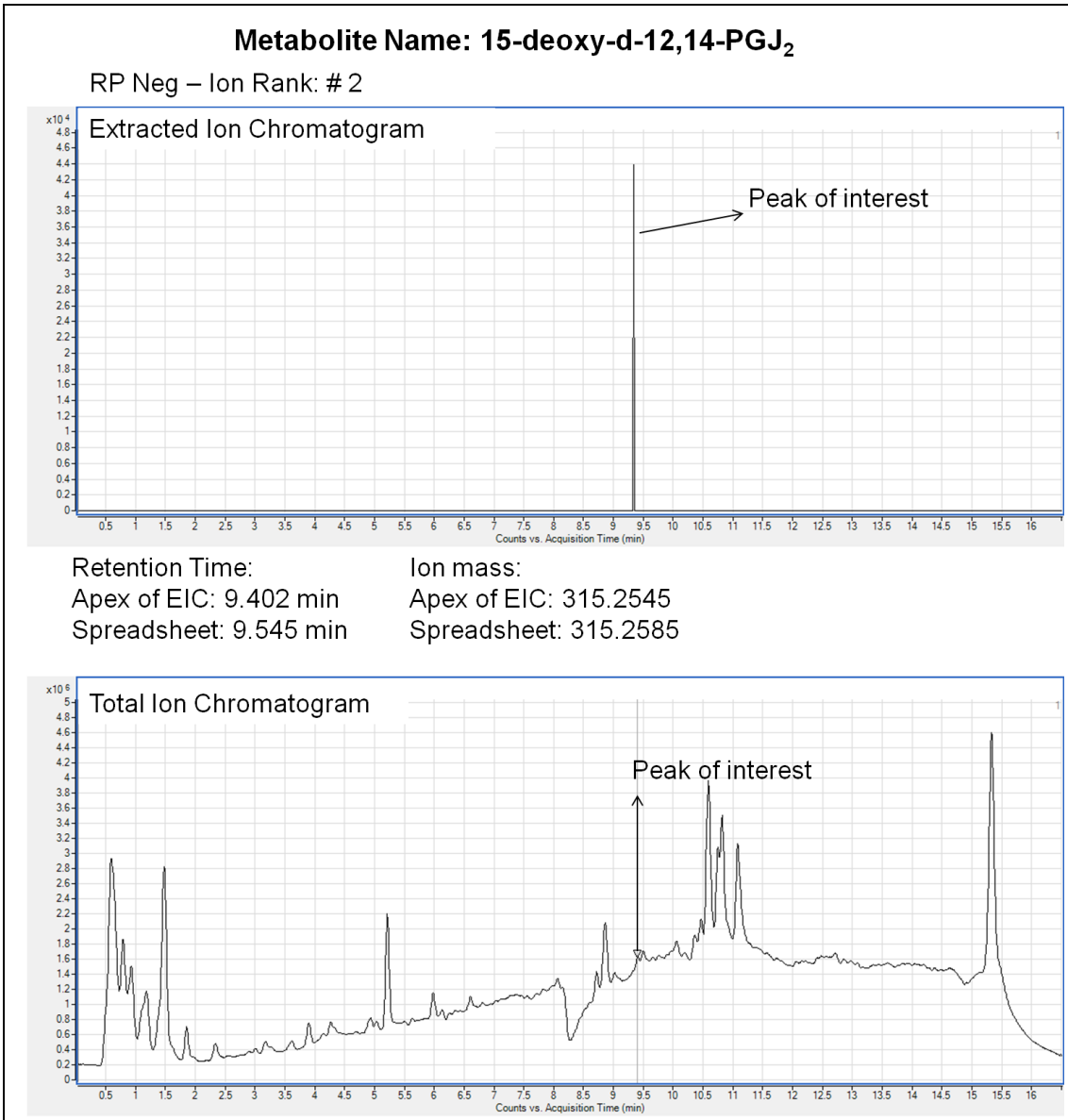


Figure 82. Extracted Ion Chromatograms (EIC) and Total Ion Chromatograms (TIC) of the Top Molecular Features associated with inflammatory pathways present in stage 2 of *T. suis* infection.
The provided chromatograms depict the EIC and TIC of each molecular feature deemed statistically significant in my analysis. Using the method described in Appendix 7, I used the chromatograms shown to refine my data set to include the top 7 molecular features depicted here.

REFERENCES

1. Strimbu, K. & Tavel, J. A. What are Biomarkers? *Curr. Opin. HIV AIDS* **5,** 463–466 (2010).

2. Siraisi, N. G. *Medieval & early Renaissance medicine: an introduction to knowledge and practice*. (c1990.).

3. Putri, S. *et al.* Current Metabolomics: Practical applications. *J. Biosci. Bioeng.* **115,** 579–589 (2013).

4. Putri, S., Yamamoto, S., Tsugawa, H. & Fukusaki, E. Current Metabolomics: Technological Advances. *J. Biosci. Bioeng.* **116,** 9–16 (2013).

5. Bajtarevic, A. *et al.* Noninvasive detection of lung cancer by analysis of exhaled breath. *Biomed Cent. Cancer* **9,** (2009).

6. Buszewski, B. *et al.* Identification of volatile organic compounds secreted from cancer tissues and bacterial cultures. *J. Chromatogr. B* **868,** 88–94 (2008).

7. Zehm, S., Schweinitz, S., Wurzner, R., Colvin, H. P. & Rieder, J. Detection of Candida albicans by Mass Spectrometric Fingerprinting. *Curr. Microbiol.* **64,** 271–275

8. Kafasack, B. & Llinas, M. Eating at the Table of Another: Metabolomics of Host-Parasite Interactions. *Cell Host Microbe* **7,** 90–99 (2010).

9. Vermeersch, K. A. & Styczynski, M. P. Applications of metabolomics in cancer research. *J. Carcinog.* **12,** 9 (2013).

10. Kordalewska, M. & Markuszewski, M. J. Metabolomics in cardiovascular diseases. *Rev. 2015* **113,** 121–136 (2015).

11. Li, R. *et al.* Alterations in the Porcine Colon Microbiota Induced by the Gastrointestinal Nematode Trichuris suis. *Infect. Immun.* **80,** 2150–2157 (2012).

12. Holmes, E. *et al.* Metabolic Profiling of CSF: Evidence That Early Intervention May Impact on Disease Progression and Outcome in Schizophrenia. *PLoS Med.* **3,** e327 (2006).

13. Kaddurah-Daouk, R. *et al.* Metabolomic mapping of atypical antipsychotic effects in schizophrenia. Molecular Psychiatry. *Mol. Psychiatry* **12,** 934–945 (2007).

14. Goodacre, R., Vaidyanathan, S., Dunn, W., Harrigan, G. & Kell, D. B. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* **22,** 245–252 (2004).

15. Kelder, T., Stroeve, J. H. M., Bijlsma, S., Radonjic, M. & Roeselers, G. Correlation network analysis reveals relationships between diet-induced changes in human gut microbiota and metabolic health. *Nutr. Diabetes* **4,** e122 (2014).

16. Kulasingam, V. & Diamandis, E. P. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nat. Clin. Pract. Oncol.* **5,** 588–599 (2008).

17. Grund, B. & Sabin, C. Analysis of Biomarker Data: logs, odds ratios and ROC curves. *Curr. Opin. HIV AIDS* **5,** 473–479 (2011).

18. Adourian, A. *et al.* Correlation network analysis for data integration and biomarker selection. *Mol. Biosyst.* **4,** 249–259 (2008).

19. Allen, E. *et al.* Correlation Network Analysis reveals a sequential reorganization of metabolic and transcriptional states during germination and gene-metabolite relationships in developing seedlings of Arabidopsis. *BMC Syst. Biol.* **4,** 1–16 (2010).

20. Becker, S., Kortz, L., Helmschrodt, C., Thiery, J. & Ceglarek, U. LC–MS-based metabolomics in the clinical laboratory. *LC-MSMS Clin. Chem.* **883–884,** 68–75 (2012).

21. Agilent Technologies. Considerations for Selecting GC/MS or LC/MS for Metabolomics. Available at: https://www.agilent.com/cs/library/selectionguide/Public/5989-6328EN.pdf.

22. Theodoridis, G. A., Gika, H. G., Want, E. J. & Wilson, I. D. Liquid chromatography–mass spectrometry based global metabolite profiling: A review. *Anal. Chim. Acta* **711,** 7–16 (2012).

23. Gika, H. G., Theodoridis, G. A., Plumb, R. S. & Wilson, I. D. Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics. *Rev. Pap. Pharm. Biomed. Anal. 2013* **87,** 12–25 (2014).

24. Deng, C., Yang, X., Li, N., Huang, Y. & Zhang, X. A novel miniaturized flame ionization detector for portable gas chromatography. *J. Chromatogr. Sci.* **43,** 355–357 (2005).

25. Cazes, J. *Encyclopedia of Chromatography (Print)*. (CRC Press, 2001).

26. Dixon, E. *et al.* Solid-Phase Microextraction and the Human Fecal VOC Metabolome. *PLoS ONE* **6,** (2011).

27. Barié, N., Bücking, M., Stahl, U. & Rapp, M. Detection of coffee flavour ageing by solid-phase microextraction/surface acoustic wave sensor array technique (SPME/SAW). *Food Chem.* **176,** 212–218 (2015).

28. Liu, H.-W. *et al.* Process sampling module coupled with purge and trap–GC–FID for in situ auto-monitoring of volatile organic compounds in wastewater. *Talanta* **80,** 903–908 (2003).

29. Ma, W., Gao, P., Fan, J., Hashi, Y. & Chen, Z. Determination of breath gas composition of lung cancer patients using gas chromatography/mass spectrometry with monolithic material sorptive extraction. *Biomed. Chromatogr.* (2014). doi:10.1002/bmc.3385

30. Kramer, R. *et al.* A rapid method for breath analysis in cystic fibrosis patients. *Eur. J. Clin. Microbiol. Infect. Dis.* 1–7 (2014). doi:10.1007/s10096-014-2286-5

31. Robroeks, C. *et al.* Metabolomics of Volatile Organic Compounds in Cystic Fibrosis Patients and Controls. *Pediatr. Res.* **68,** 75–80 (2010).

32. Couch, R. D. *et al.* Alcohol Induced Alterations to the Human Fecal VOC Metabolome. *PLOS ONE* **10,** e0119362 (2015).

33. Couch, R. *et al.* The Approach to Sample Acquisition and Its Impact on the Derived Human Fecal Microbiome and VOC Metabolome. *PLoS ONE* **8,** e81163 (2013).

34. Chao, W. & Na, Z. Analysis of Volatile Flavor Compounds in Wines by Headspace Solid-phase Microextraction Coupled with Gas Chromatography. *Shimadzu*

35. Boots, A. *et al.* The versatile use of exhaled volatile organic compounds in human health and disease. *J. Breath Res.* **6,** 027108 (2012).

36. Niwa, T. Basic Theory of Mass Spectrometry. *Clin. Chim. Acta* **241–242,** 15–71 (1995).

37. Watson, J. T. & Sparkman, O. D. *Introduction to Mass Spectrometry: Instrumentation, Applications, and Strategies for Data Interpretation*. (Wiley, 2007).

38. de Hoffmann, E. & Stroobant, V. *Mass Spectrometry: Principles and Applications*. (Wiley, 2007).

39. Anas El-Aneed, Aljandro Cohen & Joseph Banoub. Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers. *Appl. Spectrosc. Rev.* **44,** 210–230 (2009).

40. Snyder, L. R., Kirkland, J. J. & Dolan, J. W. *Introduction to modern liquid chromatography*. (John Wiley & Sons, 2011).

41. Lough, W. J. & Wainer, I. W. *High performance liquid chromatography: fundamental principles and practice*. (CRC Press, 1995).

42. Swartz, M. E. UPLC^{TM}: An Introduction and Review. *J. Liq. Chromatogr. Relat. Technol.* **28,** 1253–1263 (2005).

43. Jared Anderson, Alain Berthod, Veronica Pino & Apryll M. Stalcup. *Analytical Separation Science*. (Wiley, 2016).

44. Shibdas Banerjee & Shyamalava Mazumdar. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte,. *International Journal of Analytical Chemistry* **2012,** 1–40 (2012).

45. de Hoffmann, E. Mass Spectrometry. in *Kirk-Othmer Encyclopedia of Chemical Technology* (John Wiley & Sons, Inc., 2000).

46. Covey, T. R., Thomson, B. A. & Schneider, B. B. Atmospheric pressure ion sources. *Mass Spectrom. Rev.* **28,** 870–897 (2009).

47. Wilm, M. Principles of Electrospray Ionization. *Mol. Cell. Proteomics MCP* **10,** M111.009407 (2011).

48. Awad, H., Khamis, M. M. & El-Aneed, A. Mass Spectrometry, Review of the Basics: Ionization. *Appl. Spectrosc. Rev.* **50,** 158–175 (2015).

49. Pitt, J. J. Principles and Applications of Liquid Chromatography-Mass Spectrometry in Clinical Biochemistry. *Clin. Biochem. Rev.* **30,** 19–34 (2009).

50. Agilent Technologies. Basics of LC/MS. (1998). Available at: http://www.agilent.com/cs/library/support/documents/a05296.pdf. (Accessed: 24th March 2016)

51. Agilent Technologies. Technical Overview: Time-of-Flight Mass Spectrometery. (2015).

52. Stein, S. E. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.* **10,** 770–781 (1999).

53. Dromey, R. G., Stefik, M. J., Rindfleisch, T. C. & Duffield, A. M. Extraction of mass spectra free of background and neighboring component contributions from gas chromatography/mass spectrometry data. *Anal. Chem.* **48,** 1368–1375 (1976).

54. Sana, T. R., Roark, J. C., Li, X., Waddell, K. & Fischer, S. M. Molecular Formula and METLIN Personal Metabolite Database Matching Applied to the Identification of Compounds Generated by LC/TOF-MS. *J. Biomol. Tech. JBT* **19,** 258–266 (2008).

55. Agilent Technologies. Application Note: Q-TOF LC/MS Screening and Confirming of Non-Targeted Pesticides in a Strawberry Extract. (2009).

56. Bijlsma, S. *et al.* Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation. *Anal. Chem.* **78,** 567–574 (2006).

57. Hodge, V. & Austin, J. *A Survey of Outlier Detection Methodologies*. (Kluwer Academic Publishers, 2004).

58. Barnett, V. & Lewis, T. *Outliers in Statistical Data*. (Sons, 1984).

59. Cho, H., Kim, Y. J., Jung, H. J., Lee, S. W. & Lee, J. W. OutlierD: an R package for outlier detection using quantile regression on mass spectrometry data. *Bioinformatics* **24,** (2008).

60. Eo, S.-H., Pak, D., Choi, J. & Cho, H. Outlier Detection using Projection Quantile Regression for Mass Spectrometry Data with Low Replication. *BMC Res. Notes* **5,** 236 (2012).

61. De Livera, A. *et al.* Normalizing and Integrating Metabolomics Data. *Anal. Chem.* **84,** 10768–10776 (2012).

62. Mohamad, I. B. & Usman, D. Standardization and its effects on k-means clustering algorithm. *Res. J. Appl. Sci. Eng. Technol.* **6,** 3299–3303 (2013).

63. Abdi, H. & Williams, L. J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2,** 433–459 (2010).

64. Bro, R. & Smilde, A. K. Principal component analysis. *Anal. Methods* **6,** 2812–2831 (2014).

65. Jolliffe, I. Principal Component Analysis. in *Wiley StatsRef: Statistics Reference Online* (John Wiley & Sons, Ltd, 2014). doi:10.1002/9781118445112.stat06472

66. Olson, C. F. Parallel algorithms for hierarchical clustering. *Parallel Comput.* **21,** 1313–1325 (1995).

67. Tan, P.-N., Steinbach, M. & Kumar, V. *Introduction to Data Mining*. (Addison-Wesley, 2006).

68. Goodacre, R. *et al.* Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* **3,** 231–241

69. Deza, P. M. M. & Deza, E. Encyclopedia of Distances. in *Encyclopedia of Distances* 1–583 (Springer Berlin Heidelberg, 2009). doi:10.1007/978-3-642-00234-2_1

70. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **58,** 236–244 (1963).

71. Steuer, R. Review: On the analysis and interpretation of correlations in metabolomic data. *Brief. Bioinform.* **7,** 151–158 (2006).

72. McDonald, J. H. *Handbook of Biological Statistics*. (Sparky House Publishing, 2014).

73. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B-Methodol.* **57,** (1995).

74. Bittencourt de Andrade, J. *et al.* Perspectives for Diagnosis and Control of Leishmaniasis Based on Volatile Organic Compounds. in *Leishmaniasis - Trends in Epidemiology, Diagnosis and Treatment* (InTech, 2014).

75. CDC. Emergency Preparedness and Response: Bioterrorism.

76. Lazcka, O., Del Campo, F. J. & Muñoz, F. X. Pathogen detection: A perspective of traditional methods and biosensors. *Biosens. Bioelectron.* **22,** 1205–1217 (2007).

77. Couch, R. D. *et al.* Alcohol induced alterations to the human fecal VOC metabolome. *PloS One* **10,** e0119362 (2015).

78. Couch, R. D. *et al.* The approach to sample acquisition and its impact on the derived human fecal microbiome and VOC metabolome. *PloS One* **8,** e81163 (2013).

79. Li, R. W. *et al.* Alterations in the Porcine Colon Microbiota Induced by the Gastrointestinal Nematode Trichuris suis. *Infect. Immun.* **80,** 2150–2157 (2012).

80. Dixon, E. *et al.* Solid-phase microextraction and the human fecal VOC metabolome. *PloS One* **6,** e18471 (2011).

81. Bos, L. D. J., Sterk, P. J. & Schultz, M. J. Volatile Metabolites of Pathogens: A Systematic Review. *PLoS Pathog.* **9,** e1003311 (2013).

82. Tait, E., Perry, J. D., Stanforth, S. P. & Dean, J. R. Identification of Volatile Organic Compounds Produced by Bacteria Using HS-SPME-GC–MS. *J. Chromatogr. Sci.* **52,** 363–373 (2014).

83. Zhu, J., Bean, H. D., Kuo, Y.-M. & Hill, J. E. Fast Detection of Volatile Organic Compounds from Bacterial Cultures by Secondary Electrospray Ionization-Mass Spectrometry. *J. Clin. Microbiol.* **48,** 4426–4431 (2010).

84. Lim, S. H. *et al.* Colorimetric sensor array allows fast detection and simultaneous identification of sepsis-causing bacteria in spiked blood culture. *J. Clin. Microbiol.* **52,** 592–598 (2014).

85. Boland, W., Ney, P., Jaenicke, L. & Gassmann, G. *Analysis of Volatiles*. (1984).

86. Arthur, C. L. & Pawliszyn, J. Solid phase microextraction with thermal desorption using fused silica optical fibers. *Anal. Chem.* **62,** 2145–2148 (1990).

87. de Fátima Alpendurada, M. Solid-phase microextraction: a promising technique for sample preparation in environmental analysis. *J. Chromatogr. A* **889,** 3–14 (2000).

88. Conchas, R. F. & Carniel, E. A highly efficient electroporation system for transformation of Yersinia. *Gene* **87,** 133–137 (1990).

89. Maier, T. M. *et al.* Construction and Characterization of a Highly Efficient Francisella Shuttle Plasmid. *Appl. Environ. Microbiol.* **70,** 7511–7519 (2004).

90. Chudobova, D. *et al.* The effect of metal ions on Staphylococcus aureus revealed by biochemical and mass spectrometric analyses. *Microbiol. Res.* **170,** 147–156 (2015).

91. Blundell, M. R. & Wild, D. G. Inhibition of bacterial growth by metal salts. A survey of effects on the synthesis of ribonucleic acid and protein. *Biochem. J.* **115,** 207–212 (1969).

92. Correa, R., Coronado, L. M., Garrido, A. C., Durant-Archibold, A. A. & Spadafora, C. Volatile organic compounds associated with Plasmodium falciparum infection in vitro. *Parasit. Vectors* **10,** 215 (2017).

93. Grabowska-Polanowska, B., Skowron, M., Miarka, P., Pietrzycka, A. & Śliwka, I. The application of chromatographic breath analysis in the search of volatile biomarkers of chronic kidney disease and coexisting type 2 diabetes mellitus. *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.* **1060,** 103–110 (2017).

94. Smolinska, A. *et al.* Current breathomics—a review on data pre-processing techniques and machine learning in metabolomics breath analysis. *J. Breath Res.* **8,** 027105 (2014).

95. Shestivska, V. *et al.* Quantitative analysis of volatile metabolites released in vitro by bacteria of the genus Stenotrophomonas for identification of breath biomarkers of respiratory infection in cystic fibrosis. *J. Breath Res.* **9,** 027104 (2015).

96. Amorim, L. C. A. & de L. Cardeal, Z. Breath air analysis and its use as a biomarker in biological monitoring of occupational and environmental exposure to chemical agents. *J. Chromatogr. B* **853,** 1–9 (2007).

97. Creative Commons — Attribution 4.0 International — CC BY 4.0. Available at: https://creativecommons.org/licenses/by/4.0/legalcode.

98. NIH. Oral Probiotics: An Introduction. (2012).

99. Sekirov, I., Russell, S. L., Antunes, L. C. M. & Finlay, B. B. Gut Microbiota in Health and Disease. *Physiol. Rev.* **90,** 859–904 (2010).

100. Christensen, H. R., Frøkiaer, H. & Pestka, J. J. Lactobacilli differentially modulate expression of cytokines and maturation surface markers in murine dendritic cells. *J. Immunol. Baltim. Md 1950* **168,** 171–178 (2002).

101. Gibson, G. R. & Roberfroid, M. B. Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics. *J. Nutr.* **125,** 1401–1412 (1995).

102. Varcoe, J. J., Krejcarek, G., Busta, F. & Brady, L. Prophylactic feeding of Lactobacillus acidophilus NCFM to mice attenuates overt colonic hyperplasia. *J. Food Prot.* **66,** 457–465 (2003).

103. Ouwehand, A. C., Salminen, S. & Isolauri, E. Probiotics: an overview of beneficial effects. *Antonie Van Leeuwenhoek* **82,** 279–289 (2002).

104. Corthésy, B., Gaskins, H. R. & Mercenier, A. Cross-talk between probiotic bacteria and the host immune system. *J. Nutr.* **137,** 781S–90S (2007).

105. Rousseaux, C. *et al.* Lactobacillus acidophilus modulates intestinal pain and induces opioid and cannabinoid receptors. *Nat. Med.* **13,** 35–37 (2007).

106. Mohamadzadeh, M. *et al.* Lactobacilli activate human dendritic cells that skew T cells toward T helper 1 polarization. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 2880–2885 (2005).

107. Hong, Y.-S. *et al.* Metabonomic Understanding of Probiotic Effects in Humans With Irritable Bowel Syndrome. *J. Clin. Gastroenterol.* **45,** (2011).

108. Kumar, M. *et al.* Probiotic metabolites as epigenetic targets in the prevention of colon cancer. *Nutr. Rev.* **71,** 23–34 (2013).

109. Madsen, K. Using Metabolomics to Decipher Probiotic Effects in Patients With Irritable Bowel Syndrome. *J. Clin. Gastroenterol.* **45,** (2011).

110. Ma, Y.-Y. *et al.* Effects of probiotics on nonalcoholic fatty liver disease: A meta-analysis. *World J. Gastroenterol. WJG* **19,** 6911–6918 (2013).

111. Cani, P. D. *et al.* Selective increases of bifidobacteria in gut microflora improve high-fat-diet-induced diabetes in mice through a mechanism associated with endotoxaemia. *Diabetologia* **50,** 2374–2383 (2007).

112. Lee, H.-Y. *et al.* Human originated bacteria, Lactobacillus rhamnosus PL60, produce conjugated linoleic acid and show anti-obesity effects in diet-induced obese mice. *Biochim. Biophys. Acta* **1761,** 736–744 (2006).

113. Hanhineva, K. *et al.* Comparative nontargeted profiling of metabolic changes in tissues and biofluids in high-fat diet-fed Ossabaw pig. *J. Proteome Res.* **12,** 3980–3992 (2013).

114. Lee, L. *et al.* Nutritional model of steatohepatitis and metabolic syndrome in the Ossabaw miniature swine. *Hepatol. Baltim. Md* **50,** 56–67 (2009).

115. Perona, J. S. Membrane lipid alterations in the metabolic syndrome and the role of dietary oils. *Biochim. Biophys. Acta BBA - Biomembr.* **1859,** 1690–1703 (2017).

116. Buse, M. G. Hexosamines, insulin resistance and the complications of diabetes: current status. *Am. J. Physiol. Endocrinol. Metab.* **290,** E1–E8 (2006).

117. Fantus, I. G., Goldberg, H. J., Whiteside, C. I. & Topic, D. The Hexosamine Biosynthesis Pathway. in *The Diabetic Kidney* (eds. MD, P. C. & MD, C. E. M.) 117–133 (Humana Press, 2006). doi:10.1007/978-1-59745-153-6_7

118. Bond, M. R. & Hanover, J. A. O-GlcNAc cycling: a link between metabolism and chronic disease. *Annu. Rev. Nutr.* **33,** 205–229 (2013).

119. Maszczak-Seneczko, D. *et al.* UDP-N-acetylglucosamine Transporter (SLC35A3) Regulates Biosynthesis of Highly Branched N-Glycans and Keratan Sulfate. *J. Biol. Chem.* **288,** 21850–21860 (2013).

120. Toma, L., Pinhal, M. A., Dietrich, C. P., Nader, H. B. & Hirschberg, C. B. Transport of UDP-galactose into the Golgi lumen regulates the biosynthesis of proteoglycans. *J. Biol. Chem.* **271,** 3897–3901 (1996).

121. Arias, E. B., Kim, J. & Cartee, G. D. Prolonged Incubation in PUGNAc Results in Increased Protein O-Linked Glycosylation and Insulin Resistance in Rat Skeletal Muscle. *Diabetes* **53,** 921–930 (2004).

122. Pena, I. A. *et al.* Mouse lysine catabolism to aminoadipate occurs primarily through the saccharopine pathway; implications for pyridoxine dependent epilepsy (PDE). *Biochim. Biophys. Acta* **1863,** 121–128 (2017).

123. Papes, F., Kemper, E. L., Cord-Neto, G., Langone, F. & Arruda, P. Lysine degradation through the saccharopine pathway in mammals: involvement of both bifunctional and monofunctional lysine-degrading enzymes in mouse. *Biochem. J.* **344,** 555–563 (1999).

124. Gatrell, S. K. *et al.* Tissue distribution of indices of lysine catabolism in growing swine. *J. Anim. Sci.* **91,** 238–247 (2013).

125. KEGG PATHWAY: Lysine degradation - Reference pathway. Available at: http://www.genome.jp/kegg-bin/show_pathway?map00310+C00449. (

126. Galili, G., Tang, G., Zhu, X. & Gakiere, B. Lysine catabolism: a stress and development super-regulated metabolic pathway. *Curr. Opin. Plant Biol.* **4,** 261–266 (2001).

127. Wu, G. Synthesis of citrulline and arginine from proline in enterocytes of postnatal pigs. *Am. J. Physiol.* **272,** G1382-1390 (1997).

128. Stoll, B. *et al.* Catabolism dominates the first-pass intestinal metabolism of dietary essential amino acids in milk protein-fed piglets. *J. Nutr.* **128,** 606–614 (1998).

129. Hao, X. *et al.* Immunoassay of S-adenosylmethionine and S-adenosylhomocysteine: the methylation index as a biomarker for disease and health status. *BMC Res. Notes* **9,** 498 (2016).

130. Obeid, R. & Herrmann, W. Homocysteine and lipids: S-Adenosyl methionine as a key intermediate. *FEBS Lett.* **583,** 1215–1225 (2009).

131. KEGG PATHWAY: Cysteine and methionine metabolism - Reference pathway. Available at: http://www.genome.jp/kegg-bin/show_pathway?map00270+C00021.

132. Barroso, M. *et al.* Inhibition of cellular methyltransferases promotes endothelial cell activation by suppressing glutathione peroxidase 1 protein expression. *J. Biol. Chem.* **289,** 15350–15362 (2014).

133. Kennedy, P. B. *et al.* Elevated S-adenosylhomocysteine in Alzheimer brain: influence on methyltransferases and cognitive function. *J. Neural Transm.* **111,** 547–567 (2004).

134. Selley, M. L. A metabolic link between S-adenosylhomocysteine and polyunsaturated fatty acid metabolism in Alzheimer's disease. *Neurobiol. Aging* **28,** 1834–1839 (2007).

135. KEGG PATHWAY: Tryptophan metabolism - Homo sapiens (human). Available at: http://www.genome.jp/kegg-bin/show_pathway?hsa00380+1644.

136. Leong, S. C. & Sirich, T. L. Indoxyl Sulfate-Review of Toxicity and Therapeutic Strategies. *Toxins* **8,** (2016).

137. Kim, J., Hong, H., Heo, A. & Park, W. Indole toxicity involves the inhibition of adenosine triphosphate production and protein folding in Pseudomonas putida. *FEMS Microbiol. Lett.* **343,** 89–99 (2013).

138. KEGG PATHWAY: Glycerophospholipid metabolism - Reference pathway. Available at: http://www.genome.jp/kegg-bin/show_pathway?map00564+C00513.

139. Lu, J. *et al.* Comprehensive metabolomics identified lipid peroxidation as a prominent feature in human plasma of patients with coronary heart diseases. *Redox Biol.* **12,** 899–907 (2017).

140. KEGG PATHWAY: Pentose phosphate pathway - Reference pathway. Available at: http://www.genome.jp/kegg-bin/show_pathway?map00030+C05382.

141. KEGG MODULE: M00167. Available at: http://www.genome.jp/kegg-bin/show_module?M00167+C05382.

142. KEGG PATHWAY: Biosynthesis of amino acids - Reference pathway. Available at: http://www.genome.jp/kegg-bin/show_pathway?map01230+C05382.

143. Kelley, D. E., Goodpaster, B., Wing, R. R. & Simoneau, J.-A. Skeletal muscle fatty acid metabolism in association with insulin resistance, obesity, and weight loss. *Am. J. Physiol. - Endocrinol. Metab.* **277,** E1130–E1141 (1999).

144. Devoto, L., Henríquez, S., Kohen, P. & Strauss, J. F. The significance of estradiol metabolites in human corpus luteum physiology. *Steroids* **123,** 50–54 (2017).

145. Almeida-Pereira, G., Rorato, R., Reis, L. C., Elias, L. L. K. & Antunes-Rodrigues, J. The role of estradiol in adrenal insufficiency and its interaction with corticosterone on hydromineral balance. *Horm. Behav.* **64,** 847–855 (2013).

146. López, M. & Tena-Sempere, M. Estradiol and brown fat. *Best Pract. Res. Clin. Endocrinol. Metab.* **30,** 527–536 (2016).

147. Lee, S. H. *et al.* Metabolite profiling of sex developmental steroid conjugates reveals an association between decreased levels of steroid sulfates and adiposity in obese girls. *J. Steroid Biochem. Mol. Biol.* **162,** 100–109 (2016).

148. Miller, C. N. Estradiol reduces inflammation in rats fed a high-fat diet. (The University of North Carolina at Greensboro, 2010).

149. CDC. Inflammatory Bowel Disease. (2014). Available at: http://www.cdc.gov/ibd/index.htm.

150. Hansen, R., Cameron, F. L., Hold, G. L., El-Omar, E. M. & Russell, R. K. Inflammatory Bowel Disease. *Paediatr. Child Health* **20,** 473–478 (2010).

151. Shanahan, F. Inflammatory Bowel Disease: Immunodiagnostics, Immunotherapeutics, and Ecotherapeutics. *Gastroenterology* **120,** 622–635 (2001).

152. Strong, S. Surgical management of Crohn's disease. in *Surgical Treatment: Evidence-Based and Problem-Oriented* (Zuckschwerdt, 2001).

153. Hanauer, S. B. Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities. *Inflamm. Bowel Dis.* **12,** S3–S9 (2006).

154. Elliott, D. E., Urban, J. F., Argo, C. & Weinstock, J. V. Does the failure to acquire helminthic parasites predispose to Crohn's disease? *J. Fed. Am. Soc. Exp. Biol.* **14,** 1848–1855 (2000).

155. Weinstock, J. V. & Elliott, D. E. Translatability of helminth therapy in inflammatory bowel diseases. *Int. J. Parasitol.* **43,** 245–251 (2013).

156. National Institute of Allergy and Infectious Diseases & Coronado Biosciences, INC. Trichuris Suis Ova Treatment in Left-sided Ulcerative Colitis. *ClinicalTrials.gov* (2014). Available at: https://clinicaltrials.gov/ct2/show/NCT01953354.

157. Mucosal Immunity of Ulcerative Colitis Patients Undergoing Therapy With Trichuris Suis Ova - Full Text View - ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT01433471.

158. Efficacy and Safety of Trichuris Suis Ova (TSO) as Compared to Placebo - Full Text View - ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT01576471.

159. Safety and Tolerability of Single Doses Oral CNDO 201 Trichuris Suis Ova in Patients With Crohn's Disease - Full Text View - ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT01434693.

160. Trichuris Suis Ova (TSO) Suspension Versus Placebo in Active Crohn's Disease - Full Text View - ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT01279577.

161. Summers, R. W., Elliott, D. E., Urban, J. F., Thompson, R. & Weinstock, J. V. Trichuris suis therapy in Crohn's disease. *Gut* **54,** 87–90 (2005).

162. Pittman, J., Shepherd, G. & Thacker, B. Trichuris suis in finishing pigs: Case report and review. *J. Swine Health Prod.* **18,** 306–313 (2010).

163. Moore, P. K. Eicosanoids and gastrointestinal motility. in *Eicosanoids and the Gastrointestinal Tract* (ed. Hillier, K.) 145–162 (Springer Netherlands, 1988). doi:10.1007/978-94-009-1281-6_8

164. Eberhart, C. E. & Dubois, R. N. Eicosanoids and the gastrointestinal tract. *Gastroenterology* **109,** 285–301 (1995).

165. Khan, W. I. The Role of 5-HT Dysregulation in Inflammatory Bowel Disease. *Gastroenterol. Hepatol.* **9,** 259–261 (2013).

166. Manocha, M. & Khan, W. I. Serotonin and GI Disorders: An Update on Clinical and Experimental Studies. *Clin Trans Gastroenterol* **3,** e13 (2012).

167. Regmi, S. C., Kang, Y., Park, S., Park, S.-Y. & Kim, J.-A. Role of serotonin in the pathogenesis of inflammatory bowel disease. *FASEB J.* **26,** 1108.4-1108.4 (2012).

168. Keszthelyi, D., Troost, F. J. & Masclee, A. a. M. Understanding the role of tryptophan and serotonin metabolism in gastrointestinal function. *Neurogastroenterol. Motil.* **21,** 1239–1249 (2009).

169. Coquenlorge, S. *et al.* The arachidonic acid metabolite 11β-ProstaglandinF2α controls intestinal epithelial healing: deficiency in patients with Crohn's disease. *Sci. Rep.* **6,** 25203 (2016).

170. Scher, J. U. & Pillinger, M. H. 15d-PGJ2: The anti-inflammatory prostaglandin? *Clin. Immunol.* **114,** 100–109 (2005).

171. Clark, R. B. The role of PPARs in inflammation and immunity. *J. Leukoc. Biol.* **71,** 388–400 (2002).

172. Wu, S. *et al.* Worm burden-dependent disruption of the porcine colon microbiota by Trichuris suis infection. *PLoS ONE* **7,** e35470 (2012).

173. Tautenhahn, R. *et al.* An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat. Biotechnol.* **30,** 826–828 (2012).

BIOGRAPHY

Allyson L. Dailey graduated from Gar-Field High School, Woodbridge, Virginia, in 2007. She received her Bachelor of Science from George Mason University in 2011 and her Master of Science from George Mason University in 2013. Since enrolling in the doctoral program, Allyson has been employed as a Graduate Teaching Assistant and Graduate Research Assistant at George Mason University and the USDA.