12-2020

# The Application and Development of Metabolomics Methodologies for the Profiling of Food and Cellular Toxicity

Jade Woods
*University of Nebraska-Lincoln*, jade.woods@huskers.unl.edu

# THE APPLICATION AND DEVELOPMENT OF METABOLOMICS METHODOLOGIES FOR THE PROFILING OF FOOD AND CELLULAR TOXICITY

by

Jade Woods

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Chemistry

Under the Supervision of Professor Robert Powers

Lincoln, Nebraska

December, 2020

# THE APPLICATION AND DEVELOPMENT OF METABOLOMICS METHODOLOGIES FOR THE PROFILING OF FOOD AND CELLULAR TOXICITY

Jade Woods, M.S.

University of Nebraska, 2019

Advisor: Robert Powers

Metabolomics is a rapidly growing field of study. Its growth reflects advancements in technology and an improved understanding of the impact of the environment on metabolism. As a result, metabolomics is now commonly employed to investigate and characterize human and plant metabolism. The first chapter of this thesis provides an introduction to metabolomics and an overview of the protocols for sample preparation, data collection and statistical analysis. The second thesis chapter describes in explicit detail the step-by-step process of extracting and analyzing metabolites collected from mammalian cells, specifically brain tissue with a focus on Parkinson's disease. The chapter highlights important factors to consider including experiment design, sample collection, and data processing. Chapters 3 and 4 include the application of metabolomics to evaluate how the metabolome responds to the environment. Chapter 3 focuses on the neuronal response to the xenobiotic arsenic. It demonstrates how astrocytes increase glutathione production through an up regulation of the citric acid cycle and glycolytic processes. Arsenic was also observed to decreases related metabolites including citrate and lactate. These metabolites are important intermediates to ATP production and illustrate the interconnection of metabolomic processes. Chapter 4 shows how metabolite profiles can be used to evaluate the impact of environmental conditions on wines. Metabolite profiles of Pinot Noir derived from the same scion clone (Pinot noir 667) and grown in different regions along the Pacific

coast were compared. NMR and a differential sensing array were used to profile the chemical composition of the samples. We observed how environmental conditions resulted in different metabolite profiles in the various wine samples. This thesis aims to highlight the application of metabolomic to various biological studies in order to evaluate the impact of external stimuli.

**Acknowledgements**

I would like to thank my mother, Teri Woods, for supporting me throughout my college career. Her support allowed me to experience both undergraduate and graduate school. I would like to thank Dr. Robert Powers for the opportunity to work in his lab. I am honored I could contribute to the many remarkable projects. I would also like to thank my lab mates, Shulei Lei, Fatema Bhinderwala, Tessa Andrews, Eli Riekeberg, Heidi Roth, Allison Parrett, Isin Sakallioglu, and Alex Crook for their support and advice.

This work would not have been possible without the contributions of my collaborators. Cell culturing and cell work were performed by the students from Dr. Rodrigo Franco Cruz's lab at the University of Lincoln-Nebraska. The Pinot Noir Project was a collaboration with the Freshmen Research Initiative at the University of Texas at Austin. Wine samples, as well as assay data, were provided by Dr. Diana Zamora Olivares and Dr. Eric V. Anslyn from The University of Texas at Austin.

## Preface

This thesis is a result of collaboration with multiple labs. In chapter 3, the cell cultures and cell work were performed by Jordan Rose in Dr. Rodrigo Franco Cruz's lab at the University of Nebraska-Lincoln. In chapter 4, wine samples and assay data were provided by Dr. Diana Zamora Olivares and Dr. Eric V Anslyn at the University of Texas at Austin. Chapter 2 was adapted from the published chapter, "Metabolomic Analyses from Tissues in Parkinson`s Disease", In Metabolomics. Methods in Molecular Biology, Ed. S. Bhattacharya, Humana Press, New York, Vol. 1966, Chapt. 19, pp. 217-257 PMC6625357.

**Table of Contents**

**Table of Figures**



**1 Chapter 2**

**Table of Tables**

**Chapter 4**

# Chapter 1

## 1. Introduction

### 1.1 The Omics Field

The ‘omics’ field refers to a data driven approach to study entire biological systems by observing the totality of the system rather than individual aspects of interest. Omics utilizes large quantities of biological data made available from the development of high throughput technologies, which includes the ability to quantify total levels of DNA, RNA, and proteins in a given system. By quantifying the complete set of biomolecules, a global overview of the molecular processes present in a system can be ascertained, and, accordingly, allows us to investigate and understand the organism in its entirety. Conversely, a traditional reductionist approach is likely to miss important relationships when only a single part is analyzed in isolation. Rather than studying individual genes or proteins, the 'omics' approach takes a holistic view of a biological system to identify significant variations in structure, function and/or biological activity.

The earliest ‘omics’ studies took advantage of large amounts of data from DNA sequencing technology. For example, Sanger sequencing, allowed for the sequencing of entire organisms [1]. In this regard, genomics studies the entire genome rather than focusing on a specific set of genes of interest. The usefulness and popularity of genomics is directly correlated with the rapid analysis of genetic material from multiple organisms. Genomics allowed for the identification of novel genes associated with diseases and disorders, and streamlined the investigation of the cellular and functional role of specific genes [2]. Similarly, sequencing and microarrays allowed for the rapid quantification of RNA. Transcriptomics measures the total cellular levels of RNA to study gene expression and its role in disease [3]. Likewise, total cellular levels of proteins or the proteome is

identified and quantified with mass spectrometry (MS) or, occasionally, through assays [4]. Proteomics allows for a complete view on how proteins interact with drugs and the study of the role of proteins in disease mechanisms [5].

## 1.2 The Metabolome and Metabolomics

The strength of 'omics' studies is the ability to evaluate the overall activity of a cell or organism as a result of a disease state, environmental stressor, or genetic mutation. Transcriptomics identifies changes in the transcription of genes. Proteomics analyzes changes in which RNA sequences are translated into proteins. Figure 1.1 shows the central dogma of biology, which illustrates how information flows from DNA to RNA to proteins. Metabolomics is the next logical step in the 'omics' cascade. Metabolomics is an analytical science that focuses on the study of metabolism. The discipline utilizes multiple analytical techniques and methods to quantify and identify metabolites. Metabolites are the small molecular-weight (< 1,000 Da) compounds found within a biofluid (*e.g.,* serum, urine, *etc.*), cell, tissue, organ, or organism. Metabolites are intermediates and products of numerous cellular processes, which includes energy production, molecular and biomolecule synthesis, and signaling. Initially, metabolomics quantified metabolites as an extension of functional genomics. Measured metabolite concentrations provides a snapshot of the active metabolic processes which are then leveraged to identify the specific metabolic pathways affected by disease, environmental stressors, or gene deletion [6]. The observed metabolic dysfunction illustrates the downstream effects of a change in gene expression, RNA translation, or protein activity. A gene deletion or mutation may result in protein inactivation. Similarly, a disruption or a malfunction in transcription and/or translation may result in a decrease

in protein concentrations [7]. Of course, these processes may also result in an increase in protein activity or concentrations. Cellular processes are also regulated by various enzymes and proteins. Thus, quantifying disruptions in metabolite levels may serve as a proxy of genetic variation as well as a cellular response to external stressors [8]. External stressors, including toxins and bacteria, may damage DNA or alter protein and/or enzyme function. The application of metabolomics attempts to understand the biological response to various stressors. Metabolomics has increased exponentially over the last decade, and is now routinely applied to a wide variety of scientific concerns, including, food, nutrition, climate and environmental issues, human and livestock diseases, personalized medicine, drug development, and disease diagnosis [9].

## 1.3 Application of Metabolomics

### 1.3.1 Metabolomics as a Tool to Studying Disease

Metabolomic dysfunction has been associated with human diseases for hundreds of years [10]. For example, diabetes mellitus is marked by the dysfunction in the production or functionality of the hormone insulin [11]. Insulin plays multiple roles in carbohydrate metabolism including defusing glucose into muscle and fat, and increasing the amount of glucose in the bloodstream. Thus, insulin dysfunction may also result in hyperglycemia. Diabetes mellitus shares similar symptoms with diabetes insipidus, which is marked by a dysfunction of the antidiuretic hormone or receptor [12]. Biomarkers, in combination with clinical symptoms, are useful tools for disease diagnosis and monitoring disease progression [13]. For example, biomarkers can be useful for predicting the onset of neurodegenerative diseases such as Alzheimer's disease [14]. Biomarkers have been observed in blood, serum, and cerebral spinal fluid [15-17]. Metabolomics allows for

the efficient discovery of novel biomarkers as well as linking diseases with metabolic dysfunction(s). In addition to biomarkers, metabolomics may also provide insights into the mechanism of a disease. Disease-associated metabolites or metabolic pathways may be used to identify new therapeutic targets for drug development or to provide insights into drug resistance. In fact, the more we understand about the underpinning processes of human diseases, the greater the appreciation we obtain regarding the importance of metabolism in disease.

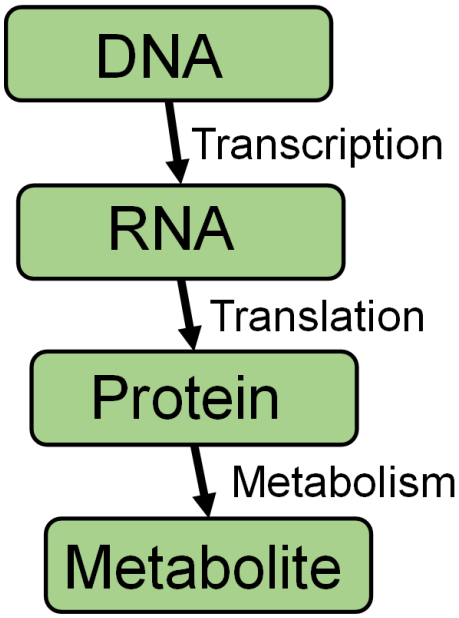



**Figure 1.1**: The central dogma of biochemistry illustrates the cascade of how information flows down from DNA to proteins. DNA is transcribed into RNA, which codes for proteins. External stressors, as well as inherited genetic variation result in downstream modification. Metabolomics, therefore, can capture dysfunction through altered metabolite levels.

Cancer is a disease that effected roughly 12 million Americans in 2008 with numbers expected to grow [18]. Cancer is characterized by metabolic dysfunction, which is used to differentiate cancer cells from healthy cells. For example, the “Warburg Effect” is a hallmark of cancer cells, which exhibit higher levels of glycolysis and a higher consummation of glucose [19]. Notably, a more efficient ATP production occurs through mitochondrial respiration and the citric acid (TCA) cycle then glycolysis [20]. One potential explanation is the upregulation of hexokinase II in cancer cells, which has been shown to be important for fixing glucose into glycolysis compared to oxidative phosphorylation [21].

Diseases of the central nervous system (CNS) are also being investigated by leveraging complex metabolomic processes in the brain. In this regard, metabolomics may provide an overview of brain function or activity by monitoring various cellular processes. Neurodegenerative diseases, which include Parkinson's Disease (PD), Alzheimer’s, and Huntington's Disease, are characterized by the progressive death of neurons and the loss of neuronal activity. For example, PD is the result of the death of dopaminergic neurons in the *substantia nigra*. The symptoms are progressive, start small, and build over time. The major symptoms of PD are motor disorders including stiffness and rigidity, which often start on one side of the body and then spreading throughout. There are also multiple non-motor symptoms that include depression, sleep behavior disorders, and nausea [22]. PD is difficult to diagnose and treat since symptoms appear during the mid- to late-stages of disease progression [23]. In 2010, roughly 630,000 individuals in America were diagnosed with PD with an estimated yearly medical cost of 14.4 billion dollars [24].

The complex nature of the brain and the limited access to tissue samples complicates the diagnosis of PD. Clinical tests that include neurological scans and a physician’s assessment of a response to treatment are typical methods used to diagnose PD [25]. An early diagnosis of PD and the

immediate initiation of treatment may result in reduced symptoms for patients. Early drug-intervention has been suggested to slow down disease progression [26]. However, PD diagnosis is difficult, requires the identification of specific symptoms, and often requires multiple physician visits. Understanding how PD alters metabolism may be beneficial to the early diagnosis and for monitoring disease progression through the use of biomarkers.

The pathogenesis of PD is not fully understood. However, a few risk factors have been linked to the development of PD, which include age, genetics, and exposure to environmental toxins [27]. For example, exposure to copper or lead have been shown to be a high-risk factor for PD [28], which have been linked to oxidative stress [29]. Metabolomics enables understanding how these risk factors alter brain metabolism (*i.e.*, neurons and astrocytes), and how dysregulated metabolism is correlated with the onset and progression of PD [30].

Oxidative stress is a popular investigative target in CNS neurological disorders [31]. Oxidative stress is regulated by metabolomic processes involving oxidative species and antioxidants. Reactive oxygen species (ROS) include radical and non-radical compounds. ROS is generated mainly through aerobic metabolism, but can be induced by other ion transferring reactions. At high levels, ROS can damage lipids, proteins, and DNA [32]. However, ROS is also necessary and is important for the regulation of signaling pathways, including apoptosis. ROS activities occur through oxidation and reducing reactions [33]. ROS cellular levels are kept in balance with antioxidants, one of the most important antioxidants is glutathione (GSH) [34]. Thus, oxidative stress is an imbalance between pro-oxidants and antioxidants [35]. The resting brain consumes about 20% of the body's oxygen [36], which results in a high production of ROS. Accordingly, the brain is highly susceptible to oxidative stress.

### 1.3.2 Metabolomics in Food science

Food science studies the physical, biological, and chemical processes involved in food. This includes determining the authenticity, contamination, nutritional content, quality, and safety of food. Foods can be evaluated and compared by measuring levels of macro and micronutrients. The amount of water, carbohydrates, proteins, and lipids will affect the flavor, structure and nutritional content of food [37]. Metabolomics is a valuable tool to generate a chemical profile for different foods. These chemical profiles will list the identification and quantification of key metabolites. Alternatively, an entire spectrum may provide a chemical fingerprint. These chemical or metabolomic profiles may be useful for the traceability of food or beverages, or for evaluating quality [38]. This is particularly useful for high-cost items, such as honey, oil and wine, where authentication is useful to avoid or prevent fraud [39].

For plant-based beverages, a major factor impacting the quality and value of the product are the environment and the weather. Essentially identical food crops grown in different regions will exhibit a locality-specific metabolite profile. Accordingly, the variable chemical composition will impact the taste, smell and texture of the beverage. Wine, in particular, is often measured by the quality of the grapes. To address this issue, metabolite levels were measured in grape pulp skins and seeds from different regions of South Korea. A specific set of metabolites including sugars and proline were observed to increase in areas with high sun exposure and lower water levels. There was also a decrease of malate, citrate, and alanine [40]. Notably, these metabolomic differences extended to all stages of grape development. $^{1}$H NMR was used to identify region specific isopentanol and isobutanol compounds from wines and grapes from regions in Rioja [41]. The processing of the grapes also impacts the wine's metabolome. For example, during fermentation, yeast consumes sugars and produce a variety of metabolites. Different strains of

yeast produced variable levels of succinate and glycerol [42]. Thus, metabolomics may help monitor and analyze the fermentation processes and asses its quality and verify its origin. A change in a metabolomics profile may easily identify the substitution of a cheaper, lower quality vintage or type of wine for a higher priced product [43]. Metabolomic studies follow a general protocol as outlined in figure 1.2. From a given biological sample, metabolites are extracted. Once extracted, the metabolites are identified and quantified. The resulting data is than analyzed and evaluated, typically with univariate and multivariate statistical methods.

## 1.4 Protocols and Procedures

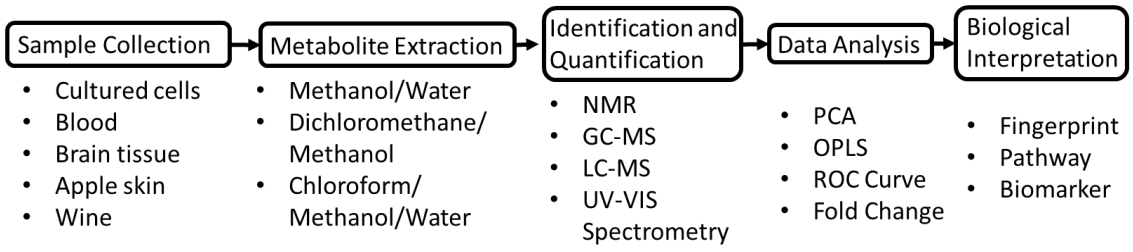



**Figure 1.2**: The metabolomic process from sample collection to biological interaction. Samples can be collected from a wide variety of sample types.

### 1.4.1 Sample Collection and Processing

Metabolomics requires the collection of all the available metabolites from a biological sample to provide a complete view of the state of the system. Metabolite extraction is a very important step of the protocol since it determines what parts of the metabolome are studied [44]. Thus, metabolomics requires specialized and targeted extraction techniques to ensure the preservation of

the metabolites and the prevention of chemical (e.g., oxidation, degradation, etc.) or enzymatic transformation. An efficient extraction protocol is fast with minimal sample preparation. Minimal sample preparation is typically needed for samples that are already in a liquid state. Conversely, solid tissues or cells are typically subjected to sample homogenization or mechanical cell lysis [45]. The quenching of biochemical reactions is also necessary to obtain a correct view of the metabolome at the time of extraction. Different extraction procedures will highlight or emphasize different biochemical pathways as well as impacting the percentage of each metabolite successfully extracted [46]. In addition to chemical stability, physical properties such as solubility and polarity will drastically impact which metabolites are maintained for a given extraction solvent. Methanol and/or water will extract polar metabolites, such as amino acids. Chloroform and/or methanol will extract nonpolar metabolites [47], while dichloromethane and methanol has been used for lipid extraction [48]. Given these physical-chemical constraints, it is not possible to harvest the entire metabolome utilizing a single extraction technique. Instead, extraction methods typically focus on collecting a specific subset of the metabolome. Multiple subsets of metabolites can be extracted with repeated extractions [49]. An additional goal of the extraction process is to remove biomolecules while preserving the metabolites of interest. Reducing error and variability is key to choosing an optimized extraction protocol.

Metabolites comprise a very diverse set of molecules that includes amino acids, vitamins, and polyols. Accordingly, metabolites exhibit a wide range of chemical stability, and are variably susceptible to changes in temperature, pH, ionic strength, oxidation and enzymatic activity. Furthermore, the type of biological sample, such as a cell lysate or a urine sample, may also differentially impact the chemical stability and the detection of specific metabolites. Sample transport and storage is important concern since many human clinical samples cannot be tested as

soon as collected. Most investigations of the impact of storage conditions on metabolite stability recommend storage of samples at or below –20°C to preserve as many metabolites as possible [50]. Freezing of the samples preserves the greatest number of metabolites. However, there is some degradation during an extended storage of more than 5 years [51]. Notably, amino acids are more prone to degradation compared to other metabolites [52]. Before samples are collected it is important to consider how samples will be processed and stored.

Once the metabolome is collected, the metabolites need to be quantified and identified. Nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are the analytical methods ordinarily used to characterize a metabolomics sample. Both methods have inherent strengths and weaknesses in regards to their ability to quantify and analyze different samples and detect different metabolites.

MS is highly versatile and is used in a majority of metabolomic studies. MS detects ions based on their mass to charge ratio, which is used to quantify and identify metabolites and fragments based on their known mass. An advantage of MS is its highly sensitive and universal detection of all ionizable metabolites. Sensitivity and selectivity of the mass spectrometry experiment is determined by the detection method. Mass spectrometers can be single or tandem instruments. Triple-quadrupole and triple-quadrupole ion trap MS are highly sensitive and are typically used in experiments where specific metabolites are targeted. Quadrupole-time of flight, linear-quadrupole ion trap-orbitrap, and Fourier transform ion cyclotron resonance are typically used for global profiling [53].

While MS can detect and identifying multiple metabolites, the use of chromatographic separation helps avoid ion-suppression and addresses the low-mass range of metabolites. Of course, matrix

effects may still interfere with the detection of the metabolites [54]. Liquid chromatography and gas chromatography are the most common chromatographic techniques used in metabolomics. Gas chromatography provides high separation and is less prone to ion suppression, but often requires chemical modification to form ions in the gas phase [55]. Liquid chromatography does not require chemical modification, but may still modulate the composition of the metabolome due to variable recovery of the metabolite from the column, metabolite decomposition or chemical modification, or ion-suppression due to co-eluting matrix compounds [56].

Ionization occurs typically after separation and is key for metabolite detection and quantification. Electron impact is common for GC-MS. It is a harder ionization method and tends to lead to sample fragmentation; however, it tends to avoid matrix effects [54]. Electrospray ionization is commonly used in LC-MS. It is a softer technique and is less prone to fragmentation, but ion suppression is common. Ion suppression occurs when charges on some molecules are lost due to the presence of endogenous compounds that are more efficient in acquiring a charge. This negatively impacts the reproducibility and accuracy of the metabolomics experiments. Ion suppression can be reduced during experimentation design through extraction of only molecules of interest or separating potentially competing molecules [57]. The use of multiple ionization methods has been suggested to expand the number of detectable metabolites. [58]

Mass to charge ratio is used to identify the metabolite. However, the reliably identification of metabolites by exact mass alone is challenging given the narrow mass distribution of metabolites and the fact that a large number of molecules have the same molecular formula and mass. Combining exact mass with retention time may improve the accuracy of metabolite identification. Chromatography provides the experimental retention times. Software programs, such as Progensis QI, are routinely used to identify metabolites based on mass and retention time. However, manual

confirmation of metabolites is necessary to avoid misidentification with molecules with similar mass and retention time. MS/MS fragmentation patterns can be matched with data from the PRIMe website (http://prime.psc.riken.jp/) to further improve the assignment confidence [59].

Nuclear Magnetic Resonance (NMR) is also widely used in metabolomics. NMR detects the absorbance of radio-frequency (RF) energy by specific nuclei in a magnetic field. NMR can detect RF absorbance by $^{1}H$, $^{13}C$, and $^{31}P$ nuclei. NMR samples require minimal sample preparation. This is often limited to adding a deuterated buffer or solvent, and an internal standard. NMR chemical shift standards include sodium trimethylsilylpropanesulfonate (DSS), 3-(Trimethylsilyl)propionic-2,2,3,3 acid sodium salt (TMSP), and trimethylsilylpropanoic acid (TSP), which are also critical for quantitation [60]. Buffers are used sample-dependent pH variations, which is a common problem for clinical samples. For example, a phosphate buffer is typically added to urine samples to maintain a pH of 7.

The most common NMR experiment used in metabolomics detects protons or $^{1}H$. $^{1}H$ is an NMR active nuclei, is a very common atom in organic molecules, and has a natural abundance of 99.98%. NMR experiment acquisition time is directly proportional to the desired sensitivity. For $^{1}H$ NMR metabolic experiments, high sensitivity or signal-to-noise (S/N) can be obtained with acquisition times of 5 minutes or less. Conversely, the natural abundance of $^{13}C$ is only 1.1% requiring significantly longer experimental times (> hours) to achieve the same relative S/N. Despite 100% natural abundance, $^{31}P$ NMR experiments are not as common, but are increasing in popularity. For example, phosphorylation is important in various biological reactions that include glycolysis [61]. Metabolomics typically rely on one-dimensional (1D) $^{1}H$ NMR experiments or two-dimensional (2D) $^{1}H$-$^{13}C$ correlated experiments. 1D NMR experiments allow for easy quantification of metabolites as intensities are directly correlated with metabolite concentrations. 2D NMR

experiments, while more time consuming, allow for easy identification of metabolites by utilizing multiple correlated chemical shifts. NMR is not as sensitive as MS, but can detect metabolites with concentrations as low as ~ 3 μM. NMR also requires a larger sample volume, typically from 500 μL, for common 5 mm NMR tube, to 35 μL for a 1.7 mM NMR tube.

Metabolite identification is accomplished by matching experimental chemical shifts to chemical shifts from standard spectra in NMR databases. For example, the ECMDB (*Escherichia coli* metabolome database, http://www.ecmdb.ca) is a compressive database that contains information about the genome and metabolome of *E. coli*. The database contains 3760 compounds [62]. Chemical shifts are very sensitive to variations in the chemical environment, such as differences in pH, ionic strength and temperature. Accordingly, databases routinely contain chemical shifts collected at a pH of 7 and a temperature of 25 $^{o}$C. Thus, an improvement in assignment accuracy is obtained by matching the databases' experimental parameters. Nevertheless, a chemical shift error tolerance of 0.05 and 0.5 ppm is commonly required for $^{1}$H and $^{13}$C chemical shifts, respectively.

### 1.4.2 Statistical Analysis of Metabolomics Data

Once NMR or MS spectra have been successfully collected, the data set is subjected to a variety of statistical analysis to identify the underlying metabolic differences. Typical data analysis methods focus on observing patterns in multivariate data sets. Proper data analysis requires extracting the significant spectral data that defines the group separation while accounting for non-biological sources of variance. Cell growth or sample collection, as well as sample preparation and data collection can introduce variation and bias in the data set. Clinical samples have intrinsically

large biological variations due to numerous factors such as age, diet, ethnicity, gender, physical activity, race, and weight.

Principal component analysis (PCA), partial least squares regression (PLS), and orthogonal projections to latent structures (OPLS) are multivariate statistical models commonly used in metabolomics. PCA is an unsupervised technique, where sample group membership is not identified. OPLS and PLS are supervised techniques where group membership is defined. As a result, OPLS and PLS suffer from over-fitting the data and require extensive validation of the resulting models. These multivariate statistics methods identify class separation by reducing data to a few components and determining the source of the variance in the data set. The goal is to identify changes in metabolite levels in response to a treatment or stressor. PCA is primarily used to identify without bias the presence of group separation or variance in the data. OPLS is used to identify the spectral signals or metabolites that define the group-based variation.

After a statistical model is generated, it is important to validate the model. Without proper validation, erroneous metabolic perturbations may be improperly assigned to falsely differentiated groups. Model quality and validation is commonly assed using $R^2$, $Q^2$, and p-values. $R^2$ (ranges from 0 to 1) is the measure of the degree of fit to the data. $Q^2$ (ranges from 0 to 1) is a quality assessment corresponding to the measure of the degree of fit for the data left out. A p-values < 0.05 from CV-ANOVA or a cross-permutation test provide model validation [63]. In practice, good PCA or OPLS models yield extremely small p-values, much lower than 0.001.

### 1.5 Summary of Work

Chapter 2 provides a detailed description of the metabolomics protocol developed and employed to investigate PD and other neurodegenerative diseases. The protocol provides a detailed step-by-step description for acquiring metabolomics data from human cell cultures and mouse brain tissues. The protocol includes the step by step direction for NMR and MS metabolite collection and analysis. It also illustrates several statistical analysis methods for both multivariate and univariate approaches. The chapter also covers the process of metabolite identification, a key step to understand what cellular process were altered. Finally, it includes guides on univariate and multivariate statistical methods and ways to validate significance.

Chapter 3 focuses on the effect of xenobiotic arsenic, a common metalloid associated with Parkinson's' disease. Cultured astrocytes were treated with and without arsenic to evaluate changes in cellular metabolism, especially in regards with glycolysis. Arsenic treatment was observed to induce the production of potentially neurotoxic glutamate and a reduction in lactate and citrate. Glycolysis may be upregulated to produce glutamate for glutathione (GSH), and as a byproduct disrupted the production of other metabolites.

Chapter 4 examined the effect of different environments on grapes used in Pinot Noir wine. Untargeted 1D 1H NMR metabolomics and a targeted differential sensing (DS) array were combined to characterize the chemical profile of Pinot Noir wines due to different environments. Each wine was differentiated by variable combinations of NMR and assay features. The NMR data provided a comprehensive coverage of the metabolome, while the DS array targeted phenolic compounds. Thus, the DS assay and the NMR data likely detect a distinct set of metabolites and provided a complementary characterization of the wines.

**1.6 References**

1. Morozova, O. and M.A. Marra, Applications of next-generation sequencing technologies in functional genomics. Genomics, 2008. 92(5): p. 255-64.

2. Morton, C.C., Genetics, genomics and gene discovery in the auditory system. Human Molecular Genetics, 2002. 11(10): p. 1229-1240.

3. Lowe, R., et al., Transcriptomics technologies. PLoS Comput Biol, 2017. 13(5): p. e1005457.

4. Gershon, D., proteomics technologies: Probing the proteome. Nature, 2003. 424(6948): p. 581-583.

5. Graves, P.R. and T.A. Haystead, Molecular biologist's guide to proteomics. Microbiol Mol Biol Rev, 2002. 66(1): p. 39-63; table of contents.

6. Oliver, S.G., et al., Systematic functional analysis of the yeast genome. Trends Biotechnol, 1998. 16(9): p. 373-8.

7. DeBerardinis, R.J. and C.B. Thompson, Cellular metabolism and disease: what do metabolic outliers teach us? Cell, 2012. 148(6): p. 1132-44.

8. Raamsdonk, L.M., et al., A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nature Biotechnology, 2001. 19(1): p. 45-50.

9. Putri, S.P., et al., Current metabolomics: Practical applications. Journal of Bioscience and Bioengineering, 2013. 115(6): p. 579-589.

10. Trowell, H.C., Ants distinguish diabetes mellitus from diabetes insipidus. British medical journal (Clinical research ed.), 1982. 285(6336): p. 217-217.

11. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care, 2010. 33(Supplement 1): p. S62-S69.

12. Robertson, G.L., Diabetes Insipidus. Endocrinology and Metabolism Clinics of North America, 1995. 24(3): p. 549-572.

13. Monteiro, M.S., et al., Metabolomics analysis for biomarker discovery: advances and challenges. Curr Med Chem, 2013. 20(2): p. 257-71.

14. Graham, S.F., et al., Investigation of the Human Brain Metabolome to Identify Potential Markers for Early Diagnosis and Therapeutic Targets of Alzheimer's Disease. Analytical Chemistry, 2013. 85(3): p. 1803-1811.

15. Serkova, N.J., T.J. Standiford, and K.A. Stringer, The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses. Am J Respir Crit Care Med, 2011. 184(6): p. 647-55.

16. Wishart, D.S., et al., The human cerebrospinal fluid metabolome. J Chromatogr B Analyt Technol Biomed Life Sci, 2008. 871(2): p. 164-73.

17. Psychogios, N., et al., The human serum metabolome. PLoS One, 2011. 6(2): p. e16957.

18. Yabroff, K.R., et al., Economic burden of cancer in the United States: estimates, projections, and future research. Cancer Epidemiol Biomarkers Prev, 2011. 20(10): p. 2006-14.

19. Ward, Patrick S. and Craig B. Thompson, Metabolic Reprogramming: A Cancer Hallmark Even Warburg Did Not Anticipate. Cancer Cell, 2012. 21(3): p. 297-308.

20. Warburg, O., On the origin of cancer cells. Science, 1956. 123(3191): p. 309-14.

21. Mathupala, S.P., Y.H. Ko, and P.L. Pedersen, The pivotal roles of mitochondria in cancer: Warburg and beyond and encouraging prospects for effective therapies. Biochimica et Biophysica Acta (BBA) - Bioenergetics, 2010. 1797(6): p. 1225-1230.

22. Chaudhuri, K.R., et al., Non-motor symptoms of Parkinson's disease: diagnosis and management. Lancet Neurol, 2006. 5(3): p. 235-45.

23. Varanese, S., et al., Treatment of advanced Parkinson's disease. Parkinsons Dis, 2011. 2010: p. 480260.

24. Kowal, S.L., et al., The current and projected economic burden of Parkinson's disease in the United States. Movement Disorders, 2013. 28(3): p. 311-318.

25. Jankovic, J., et al., The Evolution of Diagnosis in Early Parkinson Disease. JAMA Neurology, 2000. 57(3): p. 369-372.

26. Murman, D.L., Early treatment of Parkinson's disease: opportunities for managed care. Am J Manag Care, 2012. 18(7 Suppl): p. S183-8.

27. Goldman, S.M., Environmental toxins and Parkinson's disease. Annu Rev Pharmacol Toxicol, 2014. 54: p. 141-64.

28. Gorell, J.M., et al., Multiple risk factors for Parkinson's disease. Journal of the Neurological Sciences, 2004. 217(2): p. 169-174.

29. Chen, P., M.R. Miah, and M. Aschner, Metals and Neurodegeneration. F1000Res, 2016. 5.

30. Powers, R., et al., Metabolic Investigations of the Molecular Mechanisms Associated with Parkinson's Disease. Metabolites, 2017. 7(2).

31. Patel, M., Targeting Oxidative Stress in Central Nervous System Disorders. Trends Pharmacol Sci, 2016. 37(9): p. 768-778.

32. CROSS, C.E., et al., Oxygen Radicals and Human Disease. Annals of Internal Medicine, 1987. 107(4): p. 526-545.

33. Ray, P.D., B.W. Huang, and Y. Tsuji, Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. Cell Signal, 2012. 24(5): p. 981-90.

34. Wilson, J.X., Antioxidant defense of the brain: a role for astrocytes. Can J Physiol Pharmacol, 1997. 75(10-11): p. 1149-63.

35. Jones, D.P., Redefining Oxidative Stress. Antioxidants & Redox Signaling, 2006. 8(9-10): p. 1865-1879.

36. Mink, J.W., R.J. Blumenschine, and D.B. Adams, Ratio of central nervous system to body metabolism in vertebrates: its constancy and functional basis. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 1981. 241(3): p. R203-R212.

37. Pomeranz, Y., Functional properties of food components. 2nd ed. Food science and technology. 1991, San Diego: Academic Press. ix, 569 p.

38. Castro-Puyana, M., et al., Application of mass spectrometry-based metabolomics approaches for food safety, quality and traceability. TrAC Trends in Analytical Chemistry, 2017. 93: p. 102-118.

39. Ogrinc, N., et al., The application of NMR and MS methods for detection of adulteration of wine, fruit juices, and olive oil. A review. Analytical and Bioanalytical Chemistry, 2003. 376(4): p. 424-430.

40. Son, H.-S., et al., Metabolomic Studies on Geographical Grapes and Their Wines Using 1H NMR Analysis Coupled with Multivariate Statistics. Journal of Agricultural and Food Chemistry, 2009. 57(4): p. 1481-1490.

41. López-Rituerto, E., et al., Investigations of La Rioja Terroir for Wine Production Using 1H NMR Metabolomics. Journal of Agricultural and Food Chemistry, 2012. 60(13): p. 3452-3461.

42. Son, H.-S., et al., 1H NMR-Based Metabolomic Approach for Understanding the Fermentation Behaviors of Wine Yeast Strains. Analytical Chemistry, 2009. 81(3): p. 1137-1145.

43. Fotakis, C., et al., NMR metabolite fingerprinting in grape derived products: An overview. Food Research International, 2013. 54(1): p. 1184-1194.

44. Mushtaq, M.Y., et al., Extraction for metabolomics: access to the metabolome. Phytochem Anal, 2014. 25(4): p. 291-306.

45. Beckonert, O., et al., Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nature Protocols, 2007. 2(11): p. 2692-2703.

46. Canelas, A.B., et al., Quantitative Evaluation of Intracellular Metabolite Extraction Techniques for Yeast Metabolomics. Analytical Chemistry, 2009. 81(17): p. 7379-7389.

47. Sapcariu, S.C., et al., Simultaneous extraction of proteins and metabolites from cells in culture. MethodsX, 2014. 1: p. 74-80.

48. Bligh, E.G. and W.J. Dyer, A Rapid Method Of Total Lipid Extraction and Purification. Canadian Journal of Biochemistry and Physiology, 1959. 37(1): p. 911-917.

49. Coman, C., et al., Simultaneous Metabolite, Protein, Lipid Extraction (SIMPLEX): A Combinatorial Multimolecular Omics Approach for Systems Biology. Molecular & cellular proteomics : MCP, 2016. 15(4): p. 1453-1466.

50. Rotter, M., et al., Stability of targeted metabolite profiles of urine samples under different storage conditions. Metabolomics : Official journal of the Metabolomic Society, 2017. 13(1): p. 4-4.

51. Haid, M., et al., Long-Term Stability of Human Plasma Metabolites during Storage at −80 °C. Journal of Proteome Research, 2018. 17(1): p. 203-211.

52. Breier, M., et al., Targeted metabolomics identifies reliable and stable metabolites in human serum and plasma samples. PLoS One, 2014. 9(2): p. e89728.

53. Gowda, G.A. and D. Djukovic, Overview of mass spectrometry-based metabolomics: opportunities and challenges. Methods Mol Biol, 2014. 1198: p. 3-12.

54. Smeraglia, J., S.F. Baldrey, and D. Watson, Matrix effects and selectivity issues in LC-MS-MS. Chromatographia, 2002. 55(1): p. S95-S99.

55. Sparkman, O.D., Z. Penton, and F.G. Kitson, Gas chromatography and mass spectrometry : a practical guide. 2011.

56. Vuckovic, D., Current trends and challenges in sample preparation for global metabolomics using liquid chromatography–mass spectrometry. Analytical and Bioanalytical Chemistry, 2012. 403(6): p. 1523-1548.

57. Furey, A., et al., Ion suppression; a critical review on causes, evaluation, prevention and applications. Talanta, 2013. 115: p. 104-22.

58. Nordström, A., et al., Multiple Ionization Mass Spectrometry Strategy Used To Reveal the Complexity of Metabolomics. Analytical Chemistry, 2008. 80(2): p. 421-429.

59. Sawada, Y., et al., Widely Targeted Metabolomics Based on Large-Scale MS/MS Data for Elucidating Metabolite Accumulation Patterns in Plants. Plant and Cell Physiology, 2008. 50(1): p. 37-47.

60. Rundlöf, T., et al., Survey and qualification of internal standards for quantification by 1H NMR spectroscopy. Journal of Pharmaceutical and Biomedical Analysis, 2010. 52(5): p. 645-651.

61. Duboc, D., et al., Phosphorus NMR spectroscopy study of muscular enzyme deficiencies involving glycogenolysis and glycolysis. Neurology, 1987. 37(4): p. 663-663.

62. Sajed, T., et al., ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. Nucleic Acids Res, 2016. 44(D1): p. D495-501.

63. Eriksson, L., J. Trygg, and S. Wold, CV-ANOVA for significance testing of PLS and OPLS® models. Journal of Chemometrics, 2008. 22(11-12): p. 594-600.

# Chapter 2

## 2. Metabolomics Analyses from Tissues in Parkinson's disease

### 2.1 Introduction

Parkinson's disease (PD), the second most common neurodegenerative disorder worldwide, is characterized by the selective loss of dopaminergic neurons of the substantia nigra pars compacta (SNpc) [1]. There is no current treatment to stop neuronal cell death progression or to cure PD. Thus, to find neuroprotective strategies, a clear understanding of the mechanism(s) involved in dopaminergic cell death is needed. Mitochondrial dysfunction and the concomitant alterations in redox homeostasis and bioenergetics (energy failure) are thought to be a central component of PD [2-4]. One means of analyzing the state of a biological system is by monitoring the metabolome, *i.e.*, all the metabolites present in a cell, biofluid, tissue, organ, or organism [5, 6]. In this regard, metabolomics is the study of the changes in the concentration and the identity of these metabolites that result from environmental or genetic stress, or from a disease state or drug treatment. A better understanding of the biological phenotype during disease development and progression may be achieved by identifying and quantifying variations in metabolite levels. In essence, metabolomics provides a top-down view of complex biological systems. Accordingly, metabolomics has evolved to become an important resource for systems biology and a valuable tool to study disease states [7]. Metabolomics has been successfully applied to study neurological and neurodegenerative disorders [8]. Indeed, previous studies have demonstrated the applicability of metabolomics in: 1) the identification of potential biomarkers of PD diagnosis, onset and progression [9-11]; 2) the identification of novel mechanisms of disease progression [12-15]; and 3) the assessment of treatment prognosis and outcome [16]. Using metabolomics, we and others have established a link

between the alterations in central carbon metabolism induced by PD risk factors, redox homeostasis and bioenergetics and their contribution to the survival or death of dopaminergic cells [2].

Unlike other OMICs techniques, the composition of the metabolome can easily change from the processing, handling and storage of samples [17]. Metabolites may chemically transform or degrade due to residual enzymatic activity, from oxidation, from low chemical stability, or from other chemical activity. Thus, robust and reproducible isolation of metabolites is a key step in the metabolomics workflow. Univariate and multivariate statistical analysis are also an important aspect of a metabolomics study [18]. But, the incorrect application of statistical techniques, the insufficient preprocessing, the lack of proper model validation, or the over-interpretation of models and outcomes are all common concerns that often lead to erroneous or misleading biological insights from metabolomics data [18]. Metabolomics has commonly relied on mass spectrometry (MS) [19] or nuclear magnetic resonance (NMR) [20] as the primary analytic source for sample analysis. Again, a successful metabolomics investigation is dependent on appropriate protocols for data collection, processing and analysis. To address these issues, we have provided a detailed, step-by-step description of a metabolomics workflow specifically applicable to the analysis of brain cell cultures and tissues used in our research using PD-experimental models (*see* Figure 2.1). We describe methods to assist in the efficient cell culturing, metabolite extraction, and data collection and analyses. Alongside, we discuss a combined NMR and MS approach to improve metabolome coverage, which allows for the identification of key neurological metabolites. While the protocols outlined in this chapter have been developed using PD-experimental models, most of the methodology may be universally applied to any metabolomics study.

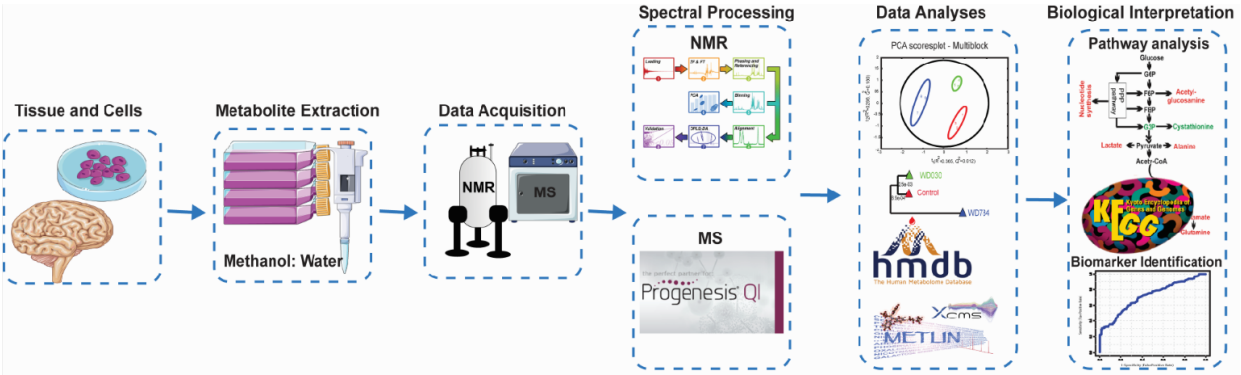



**Figure 2.1**: A schematic diagram is shown that outlines the overall metabolic workflow used in the analysis of brain cell cultures and tissues from experimental PD-models. Only the major protocol steps are highlighted in the flow diagram. The figure was generated using free medical images from Servier Medical Art (https://smart.servier.com/) under the Creative Commons License Attribution 3.0 Unported (CC BY 3.0).

## 2.2 Materials

Prepare all aqueous solutions and buffers with either Nanopure $H_2O$ or deuterated water ($D_2O$). Please follow all safety regulations in regards to handling biological samples and the disposal of both chemical and biological waste. A valuable rule-of-thumb in the handling of all tissues, biofluids (*e.g.*, blood, urine, etc.) and cell lines is to assume a contamination with a virus, pathogen or toxin and to handle the samples accordingly.

### 2.2.1 Laboratory Equipment

1. Bruker AVANCE III HD 700 MHz NMR spectrometer equipped with a 5 mm quadruple resonance QCI-P cryoprobe ($^{1}H$, $^{13}C$, $^{15}N$, and $^{31}P$) with z-axis gradients, an automatic tune and match system (ATM), and a SampleJet automated sample changer system with Bruker ICON-NMR software (Bruker Biospin, Billerica, MA)

2. Synapt G2 HDMS quadrupole time-of-flight (TOF) MS instrument equipped with an ESI source (Waters, Milford, MA).
3. Waters ACQUITY M-class Xevo G2-XS QToF MS instrument equipped with an ESI source (Waters, Milford, MA).
4. BSL-2 biosafety level grade hood (*e.g.*, Biological Safety Cabinet, LF BSC class 2 type A, Thermo Fischer Scientific, Waltham, MA).
5. Nanopure ultra water system (Barnstead Inc., Dubuque, IA)
6. Lab Armor bead bath (Chemglass Life Sciences, Vineland NJ)
7. Incubator capable of maintaining physiological temperature and proper carbon dioxide levels (*e.g.*, HERA CELL VIOS 250i $CO_2$ Incubator, Thermo Fischer Scientific, Waltham, MA).
8. pH meter and probe
9. Refrigerated centrifuge capable of speeds up to 13000 rpm (*e.g.*, SORVALL micro 21R centrifuge, Thermo Fischer Scientific, Waltham, MA).
10. Speed Vac for solvent removal (*e.g.*, SAVANT SC210A SpeedVac concentrator, Thermo Fischer Scientific, Waltham, MA)).
11. Freeze dryer to remove water (*e.g.*, FreeZone 4.5, LABCONCO, Kansas City, MO)).
12. 1000 µL to 1 µL pipettes
13. FastPrep-96 homogenizer (MP Biomedicals, Santa Ana, CA) for brain tissue analysis, uses Lysing Matrix D.
14. ACCU-SCOPE 3030ph microscope (Commac, NY)
15. cryogenic storage container (Taylor Wharton, Theodor, Al)
16. -80°C freezer

### 2.2.2 Disposable supplies

1. 1 mL to 1 μL pipette tips
2. 10 mL aspirating pipettes
3. 15 mL Falcon tubes
4. 2 mL Eppendorf tubes
5. 1 mL screw-cap microcenterfuge tubes
6. LC-MS certified total recovery vial (Waters, Milford, MA)

### 2.2.3 Isotopically labeled solvents and reagents (*see* **Notes 1** and **2)**

1. Deuterium oxide ($D_2O$, 99.8 atom %D)
2. 3-(trimethylsilyl) propionic-2,2,3,3-$d_4$ acid sodium salt (TMSP-$d_4$, 99.8 atom % D)
3. Dimethyl sulfoxide-$d_6$ (DMSO-d6, 99.8 atom %D
4. $^{13}C_6$-glucose (99% $^{13}C$)
5. $^{13}C_2$-acetate (99% $^{13}C$)
6. Other potential $^{13}C$-carbon labeled or $^{15}N$-nitrogen labeled reagents

### 2.2.4 Buffers

1. Wash buffer, phosphate-buffered saline (PBS) at pH 7.4: 137 mM NaCl, 2.7 mM KCl, 10 mM $Na_2HPO_4$ and 2 mM $KH_2PO_4$. To prepare 1 L PBS buffer at pH 7.4, add 8.0 g of NaCl, 0.2 g of KCl, 2.68 g of $Na_2HPO_4{\cdot}7H_20$ and 0.24 g of $KH_2PO4$ to a final volume of 1 L of Nanopure water.

2. NMR buffer: 50 mM phosphate buffer at pH 7.2 (uncorrected, *see* **Note 3**) in 600 μL of 99.8% $D_2O$. Add 50 μM (one-dimensional [1D] NMR experiment) or 500 μM (two-dimensional [2D] NMR experiment) TMSP-$d_4$ as an internal chemical shift reference.
3. MS extraction buffer: Mix 20 mL LC-MS grade water with 80 mL LC-MS grade methanol. Store at –40°C.
4. MS reconstitution solution: LC-MS grade water with 0.1% LC-MS grade formic acid.
5. LC mobile phase A: LC-MS grade water with 0.1% LC-MS grade formic acid.
6. LC mobile phase B: LC-MS grade acetonitrile/methanol with 0.1% LC-MS grade formic acid.

### 2.2.5 Cell lines and media

1. For cell cultures, we have used human dopaminergic neuroblastoma cell lines such as SK-N-SH (HTB-11, ATCC, Manassas, VA) [15], SH-SY5Y (CRL-2266, ATCC), N27 immortalized rat dopaminergic cells (SCC048, EMD Millipore, Temecula, CA) [31], human immortalized midbrain neuronal precursors LUHMES (CRL-2927, ATCC) and primary rat/mouse astrocytes [32] following the specifications of the commercial providers or published protocols.
2. Cell culture media and supplements are obtained from commercial vendors such as GIBCO/Life Technologies (Grand Island, NY), Fisher Scientific, Hyclone (GE Healthcare, Logan, UT) and Atlanta Biologicals (Flowery Branch, GA).

### 2.2.6 Software and Databases

**1.** Bruker ICON-NMR software for automated NMR data acquisition (Bruker Biospin, Billerica, MA).

2. MVAPACK metabolomics toolkit for processing and analyzing chemometric data (http://bionmr.unl.edu/mvapack.php) [21].
3. PCA/PLS-DA utilities for quantifying separation in PCA, PLS-DA and OPLS-DA scores plots (http://bionmr.unl.edu/pca-utils.php) [22].
4. NMRPipe software for processing and visualizing NMR data (https://www.ibbr.umd.edu/nmrpipe/install.html) [23].
5. NMRViewJ software for processing and visualizing NMR data (One Moon Scientific, Inc. Westfield, NJ; https://nmrfx.org/) [24].
6. MassLynx V4.1(Waters Corp., Milford, MA) for mass spectral data processing (http://www.waters.com/waters/en_US/MassLynx-Mass-Spectrometry-Software-/nav.htm?locale=en_US&cid=513164).
7. Progenesis QI (version 2.0, Nonlinear Dynamics, Newcastle, UK) for processing and analysis of LC-MS data (http://www.nonlinear.com/progenesis/qi/)
8. R statistical package (https://www.r-project.org/) [25].
9. Chenomx (Chenomx, Inc., Edmonton, AB, Canada) software for automated metabolite assignment and quantification from 1D $^{1}$H NMR spectra (https://www.chenomx.com/).
10. Mzmine software (http://mzmine.github.io/download.html) for metabolite identification from MS data [26].
11. MetaboAnalyst software for the statistical, functional and integrative analysis of metabolomics data (http://www.metaboanalyst.ca/) [26].
12. ChemSpider chemical structure database http://www.chemspider.com/ [27].
13. Human Metabolomics Database (HMDB) of reference NMR and mass spectral data for known metabolites (http://www.hmdb.ca/) [28].

14. Biological Magnetic Resonance Data Bank (BMRB) of reference NMR data for known metabolites http://www.bmrb.wisc.edu/metabolomics/) [29].
15. Non-uniform schedule (NUS) generator (http://bionmr.unl.edu/dgs-gensched.php) for NUS NMR data acquisition [30].

## 2.3 Methods

### 2.3.1 Experimental PD models

The etiology of PD has yet to be clearly established. The major risk factor identified for PD is aging as its prevalence and incidence increases exponentially from ages 65 to 90 [31]. A fraction of PD occurrence (~10%) is related to mutations in genes such as those encoding α-synuclein (*SNCA/PARK1-4*), DJ-1 (*PARK7*), PTEN-induced putative kinase 1 (*PINK1/PARK6*), leucine-rich repeat kinase 2 (*LRRK2/PARK8*) and parkin (*PARK2*) [32, 33]. However, over 85% of PD occurs in a sporadic (idiopathic) form without a clearly defined genetic basis. Epidemiological studies suggest that lifestyle, occupational and environmental exposures can increase the risk of developing PD [34-36]. Thus, it is thought that PD arises from the convergence of genetic susceptibility, environmental exposures, and aging.

Cellular and animal disease models based on both genetic-, toxin- or stress-induced neurodegeneration have been used to understand PD pathogenesis [35, 37] (*see* Figure 2.2. However, not all experimental models recapitulate all PD hallmarks in their entirety. Genetically engineered PD mouse models have been developed for the overexpression of mutant genes [35, 33]. However, only marginal or null dopaminergic cell death has been observed in genetic-based animal models. Recent advances in mammalian genome engineering technology have led to the generation of rat PD models that seem to better reproduce PD hallmarks including progressive loss

of dopaminergic neurons, locomotor behavior deficits, and age-dependent formation of abnormal α-synuclein protein aggregates (Lewy bodies) [38].

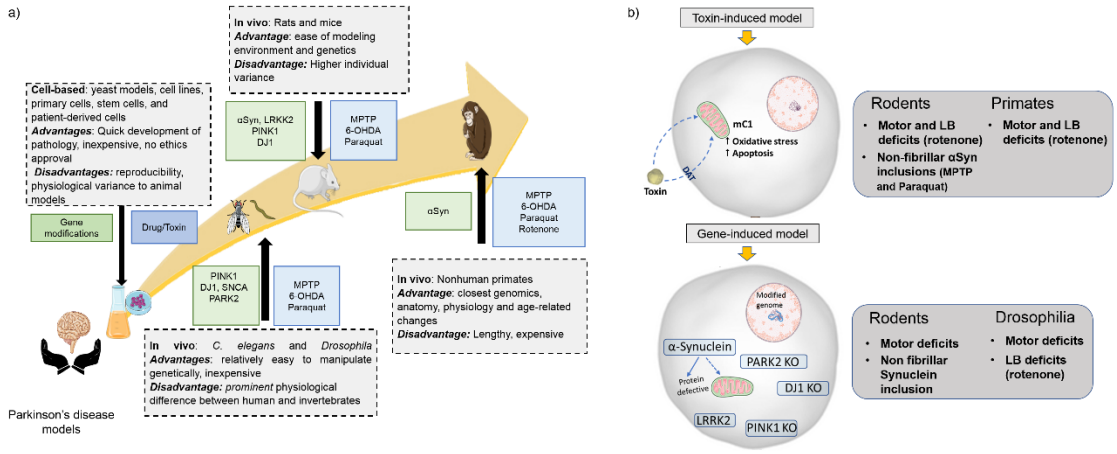



**Figure 2.2**: Common models of PD. **(A)** A summary of advantages and disadvantages of common models of PD. **(B)** A List of some model specific characteristics observed for different PD models. The figure was generated using free medical images from Servier Medical Art (https://smart.servier.com/) under the Creative Commons License Attribution 3.0 Unported (CC BY 3.0).

On the other hand, the use of mitochondrial/environmental toxins such as 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP, or its active metabolite 1-methyl-4-phenylpyridine $MPP^+$) and the pesticides rotenone and paraquat that induce dopaminergic cell death *in vitro* and *in vivo*, is supported by clinical and epidemiological studies [35]. Several other toxicants, such as metals, diverse pesticides, polychlorinated biphenyls, diet as well as inflammatory processes have been implicated as PD risk factors [39, 40]. However, it is clear that not a single environmental exposure is responsible for all PD cases nor are they the single cause for PD. Accordingly, new models studying gene-environment interactions have also emerged.

For the most part, experimental PD-models are design to reproduce one or more key aspects of PD pathogenesis including: genetic modifications, mitochondrial dysfunction, oxidative stress, accumulation of misfolded aggregates and impaired proteostatic processes, alterations in dopamine metabolism and inflammation [35]. Experimental PD models have helped to identify important mechanisms regulating dopaminergic cell death and survival, and they should continue to enhance our understanding of PD pathogenesis. In our metabolomics investigations, we have used neuronal-like cell cultures of neuroblastoma cells and immortalized midbrain dopaminergic cells from rats and humans exposed to PD-related insults and gene-environment interactions. In addition, we have also evaluated changes in the metabolome of mice exposed to pesticides and heavy-metals liked to PD or parkinsonisms [41, 15]. The protocol described below is a general protocol for isolating and characterizing changes in the metabolome applicable to different types of cell cultures and brain.

### 2.3.2 Cell culture

Cell culture procedures must follow published guidelines to avoid misidentification and contamination [42]. We recommend to start with one 100 $mm^2$ dish of 90% confluent cells per sample/replica, but if the metabolite is abundant enough, this can be reduced to a smaller sample size.

For PD-related insults, cells can be treated with mitochondrial toxins ($MPP^+$ or rotenone), pesticides (paraquat or dieldrin) or the overexpression of PD-related genes (WT or mutant forms of α-synuclein via viral vectors or conventional transfection procedures), as explained in our

previous publications [41, 45]. The exact dose and time course must be determined empirically, but we recommend to work with a dose that will induce cell death of ~50% within no less than 48 h as neurodegeneration is a slow process, and evaluate changes in the cellular metabolome prior to any detection of cell death (~24 h of treatment) (*see* Note 4-6 for considerations in regards to cell survivability, sample handling and randomization).

### 2.3.3 Unlabeled Metabolomics Sample Number/Replicas

Use the maximal number of replicates per group that is possible (*see* Note **7**). A typical number of replicate cultures per group is ten. Adjust the number of replicates given practical considerations, such as the number of groups, but the number of replicates per group should not be below six.

### 2.3.4 Isotopically Labeled Metabolomics Samples

Identify the $^{13}C$-, $^{15}N$ or other isotopically labeled tracer. The tracer should be in accordance to the metabolic pathway of interest and expected to be affected by the experimental treatment. $^{13}C_6$-glucose is a common choice for a tracer since it highlights central carbon metabolism (glycolysis and TCA cycle), but a variety of other tracers may be used. Equimolary supplement culture media with the appropriate $^{13}C$-carbon labeled source. (*see* **Note 8**).

### 2.3.5. Extracting Water Soluble Metabolites from PD Cell Cultures

All samples should be kept on ice or at 4 $^{o}C$ during sample preparation or handling. Samples should be stored at -80 $^{o}C$, but, ideally, samples should be immediately analyzed. In addition to keeping

samples cold, there are four other issues that are critical to the successful preparation of metabolomics samples: (1) speed, (2) consistency, (3) random processing of samples, and (4) the efficient removal of all biomolecules and cell debris [6]. The processing of all metabolomics samples should proceed as quickly as possible while minimizing any loss in quality. Metabolites can chemically degrade or transform within milliseconds due to enzymatic activity, oxidation, chemical instability, or any number of other chemical processes [43]. Accordingly, rapidly inactivating and removing all biomolecules and cell debris (usually through methanol/ethanol precipitation) that may transform or bind a metabolite is a necessary step of the protocol (*see* **Notes 9**).

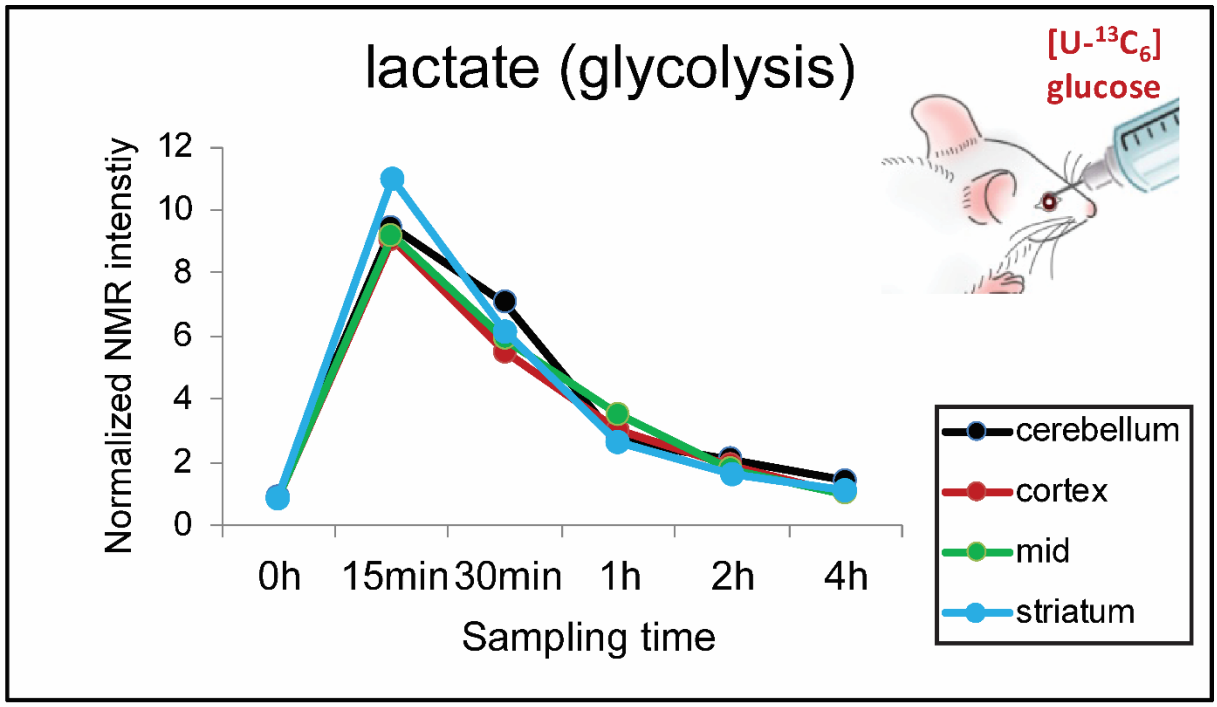
1. Collect 1 mL of the media for metabolomic analysis. In addition to the cell extract, the media may should also be analyzed for metabolomics changes as many metabolites get exchanged or effluxed outside of the cell. In this regards, the media is treated simply as another cell extract.
2. Wash the cells twice with 5 mL of PBS to remove debris. Discard the wash.
3. Lyse and quench cells with 1 mL of pre-chilled methanol at -20 $^{o}$C. Incubate cells at -80 °C for 15 min.
4. Using a cell scraper, detach and collect cell debris and methanol in a 2 mL microcentrifuge tube. Confirm cell detachment using a microscope and repeat lyse and quenching if necessary.
5. Centrifuge the 2 mL microcentrifuge tube for 5 min at 15,000 g and 4 °C to pellet the cell debris.
6. Collect the supernatant and transfer to a new 2 mL microcentrifuge tube.

7. Repeat the metabolome extraction by adding 0.5 mL of an 80%/20% mixture of methanol/water kept at -20°C to the cell pellet.
8. Centrifuge the cell pellet with the extraction solvent for 5 min at 15,000 g at 4 °C to pellet the cell debris.
9. Collect the supernatant and transfer it to the 2 mL microcentrifuge tube containing the original methanol extract. Combine the two extraction supernatants into a single tube.
10. Repeat the metabolome extraction a third time by adding 0.5 mL of ice cold water to the cell pellet.
11. Centrifuge the cell pellet with the extraction solvent for 5 min at 15,000 g at 4 °C to pellet the cell debris.
12. Collect the supernatant and transfer it to the 2 mL microcentrifuge tube containing the two previous extraction supernatants. Combine the three extraction supernatants into a single tube.
13. Split the sample into two 2 mL Eppendorf tube. Aliquot 100 µL for MS analysis and the remainder of the sample is used for NMR analysis.
14. Use a SpeedVac or a rotary evaporator to remove the methanol.
15. Flash-freeze the samples in liquid nitrogen.
16. Remove the water and bring to dryness using a lyophilizer.
17. Repeat **steps 1** to **16** for each replicate and for each group (*see* **Note 6**).
18. Store samples in a -80 $^{o}$C freezer or proceed to preparing the NMR and/or MS samples (*see* **sections 3.7** and **3.8**).

### 2.3.6 Extracting Water Soluble Metabolites from Mouse Brain Tissue

1. Similar to cell culture treatments, a number of experimental paradigms have been used to model PD *in vivo* [35, 44]. We have used the subchronic exposure to pesticides and metals [15], but the protocol described can be applied to all murine animal models.
2. We have successfully used 200 mg/kg of $^{13}C_6$-glucose at a total volume of 100 µL administered to fasted mice (overnight) via intra-orbital injection to label metabolites extracted from mouse brain tissue (**Figure 2.3**).
3. Harvest and dissect the mice brain tissue (15 to 20 min after the injection of $^{13}C$-labeled tracer if used, see **Figure 2.3**).
4. Transfer the tissue to a 2 mL microcentrifuge tube containing Lysing Matrix D and weigh the amount of tissue harvested from the mice, and immediately freeze the tissue with liquid nitrogen.
5. Extract the tissue with a 1:1 mixture of methanol and water prechilled to -20 $^{o}$C. The volume of the extraction solvent depends upon the weight of the tissue.
6. Homogenize the sample in a FastPrep with lysing Matrix D at 1300 rpm for 20 seconds, and for two cycles.
7. Incubate the tissue at -80 $^{o}$C for 10 min to extract the metabolome.
8. Centrifuge at 1000 g for 10 min at 4 $^{o}$C to remove tissue debris
9. Collect the supernatant and transfer to a new 2 mL microcentrifuge tube.
10. Repeat the metabolome extraction by adding 0.7 ml of 1:1 mixture of methanol and water prechilled to -20 $^{o}$C to the tissue pellet.
11. Repeat **steps 6** to **8** and combine the supernatant with the previous extract.

12. Normalize the metabolomics sample to the tissue weight by diluting all of the samples to a final volume of 1.5 mL. Add as much of a 1:1 mixture of methanol and water prechilled to -20 $^{o}$C as needed to achieve a final volume of 1.5 mL.
13. Split the sample into two 2 mL Eppendorf tube. Aliquot 100 μL for MS analysis and the remainder of the sample is used for NMR analysis.
14. Use a SpeedVac or a rotary evaporator to remove the methanol.
15. Flash-freeze the samples in liquid nitrogen.
16. Remove the water and bring to dryness using a lyophilizer.
17. Repeat **steps 3** to **16** for each replicate and for each group (*see* **Note 6**).
18. Store samples in a -80 $^{o}$C freezer or proceed to preparing the NMR and/or MS samples (*see* **sections 2.3.7** and **2.3.8**).



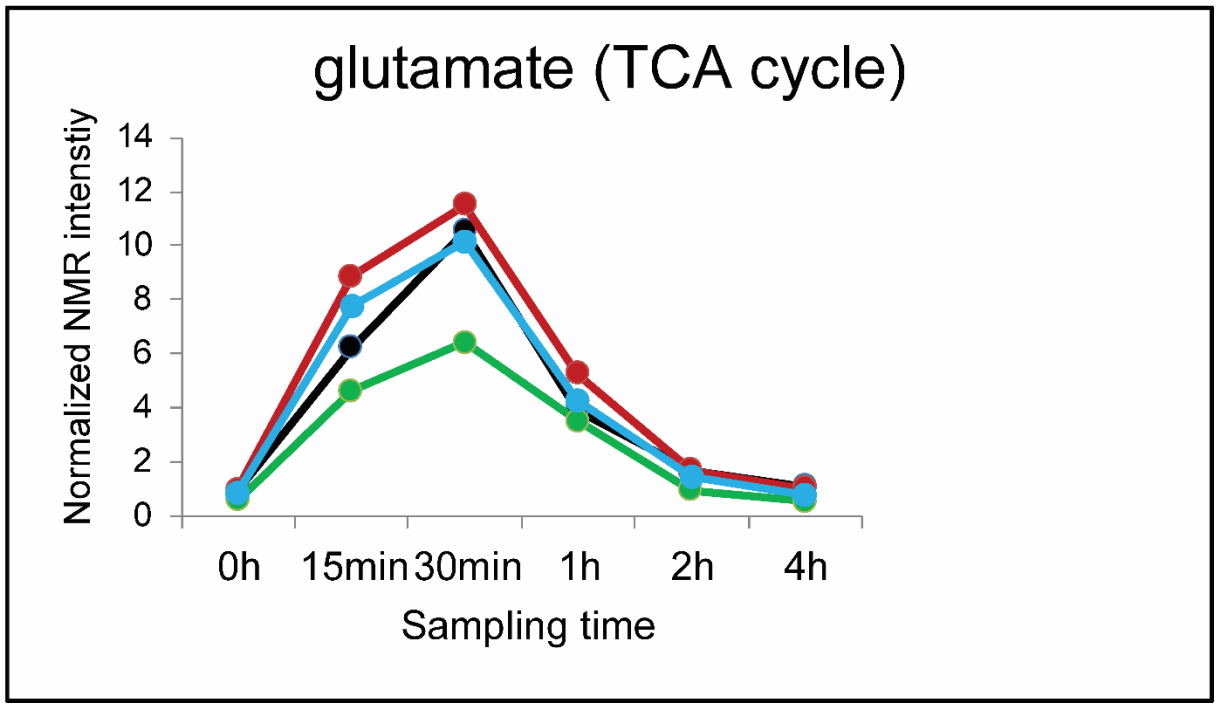




**Figure 2.3**: In vivo evaluation 13C6-glucose metabolism. Fasted mice (overnight) were administered 13C-glucose (200 mg/kg body weight , 100 µl) via retro-orbital injection and brain regions were dissected at the time indicated for NMR analysis.

### 2.3.7 Preparation of NMR Samples

1. For one-dimensional (1D) NMR experiments, lyophilized cell-free lysates or tissue extracts are suspended in 600 µL of 100% 50 mM $D_2O$ phosphate buffer (uncorrected pH 7.2) with 50 µM 3-(trimethylsilyl) propionic-2,2,3,3-$d_4$ acid sodium salt (TMSP-$d_4$)
2. For two-dimensional (2D) NMR experiments, lyophilized cell-free lysates or tissue extracts are suspended in 600 µL of 100% 50 mM $D_2O$ phosphate buffer (uncorrected pH 7.2) with 500 µM TMSP-$d_4$.
3. Centrifuge the sample at 14,000 g for 10 min to remove any particulates.
4. The sample is transferred to a 4" 5 mM SampleJet NMR tube with a pipette (*see* **Note 10**).
5. Repeat **steps 1** to **4** for each replicate and for each group (*see* **Note 6**)
6. Each sample is added to a 96 well plate SampleJet configuration equilibrated to 4 °C to prevent metabolite degradation (*see* **Figure 2.4**).

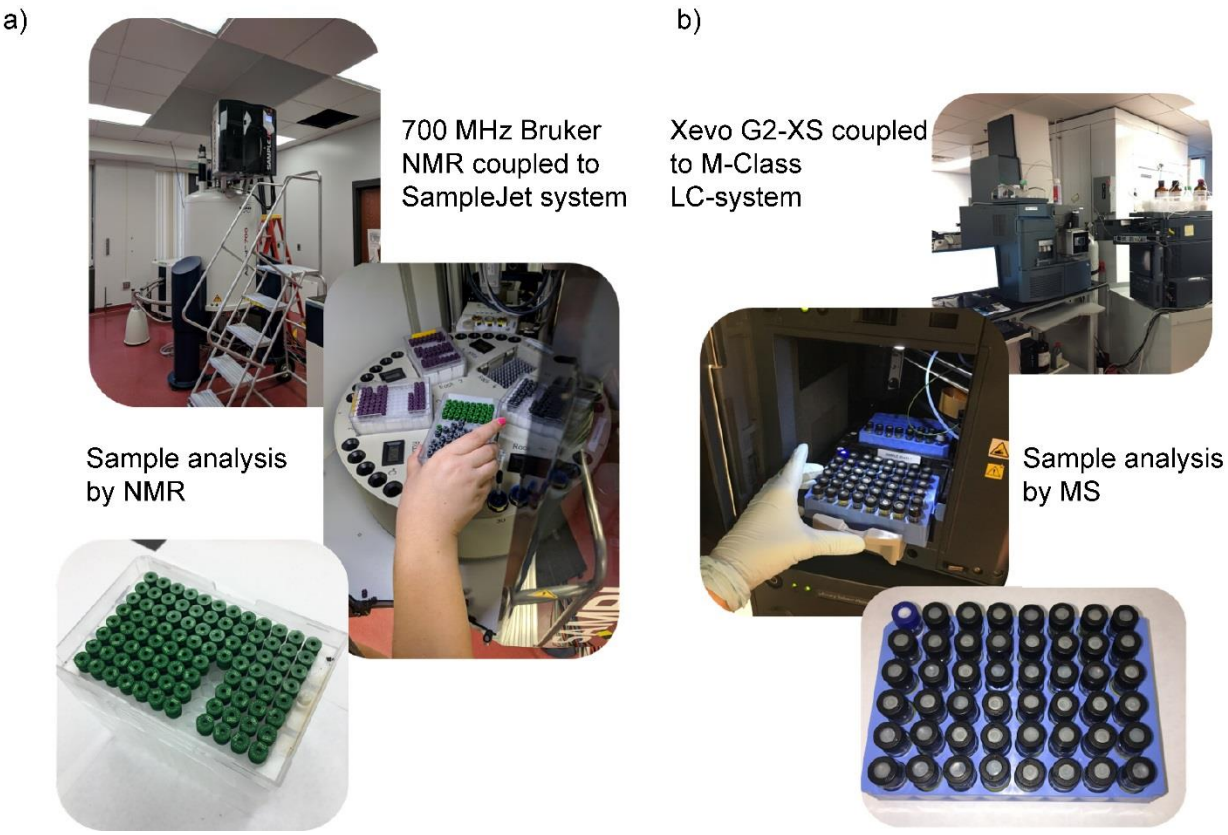



**Figure 2.4**: High-throughput sample preparation. Images illustrating the loading of replicate metabolomics samples into the (A) 96 well plate SampleJet configuration and (B) the LC-MS autosampler.

### 2.3.8 Preparation of Mass Spectrometry Samples

1. Dissolve lyophilized cell-free lysates or tissue extracts in 20 µL of reconstitution solution and vortex for 30 s.
2. Centrifuge the solution at 14,000 g for 10 min to remove any particulate matter.
3. Transfer the supernatant to LC vials and keep them in wet ice.
4. Repeat **steps 1** to **3** for each replicate and for each group (*see* **Note 6**)
5. Prepare quality control (QC) samples by pooling a 1 µL aliquot from each biological sample and transferring to a new LC vial labeled as QC.

6. Place all vials into the autosampler equilibrated to 4 ℃ to prevent metabolite degradation (*see* **Figure 2.4**).

### 2.3.9 NMR Data Collection

All NMR experiments are conducted at 298 K using a Bruker AVANCE III HD 700 MHz spectrometer equipped with a 5 mm quadruple resonance QCI-P cryoprobe ($^{1}H$, $^{13}C$, $^{15}N$, and $^{31}P$) with z-axis gradients. An automatic tune and match system (ATM), and a SampleJet automated sample changer system with Bruker ICON-NMR software were used to automate the NMR data collection (*see* **Figure 2.5**).

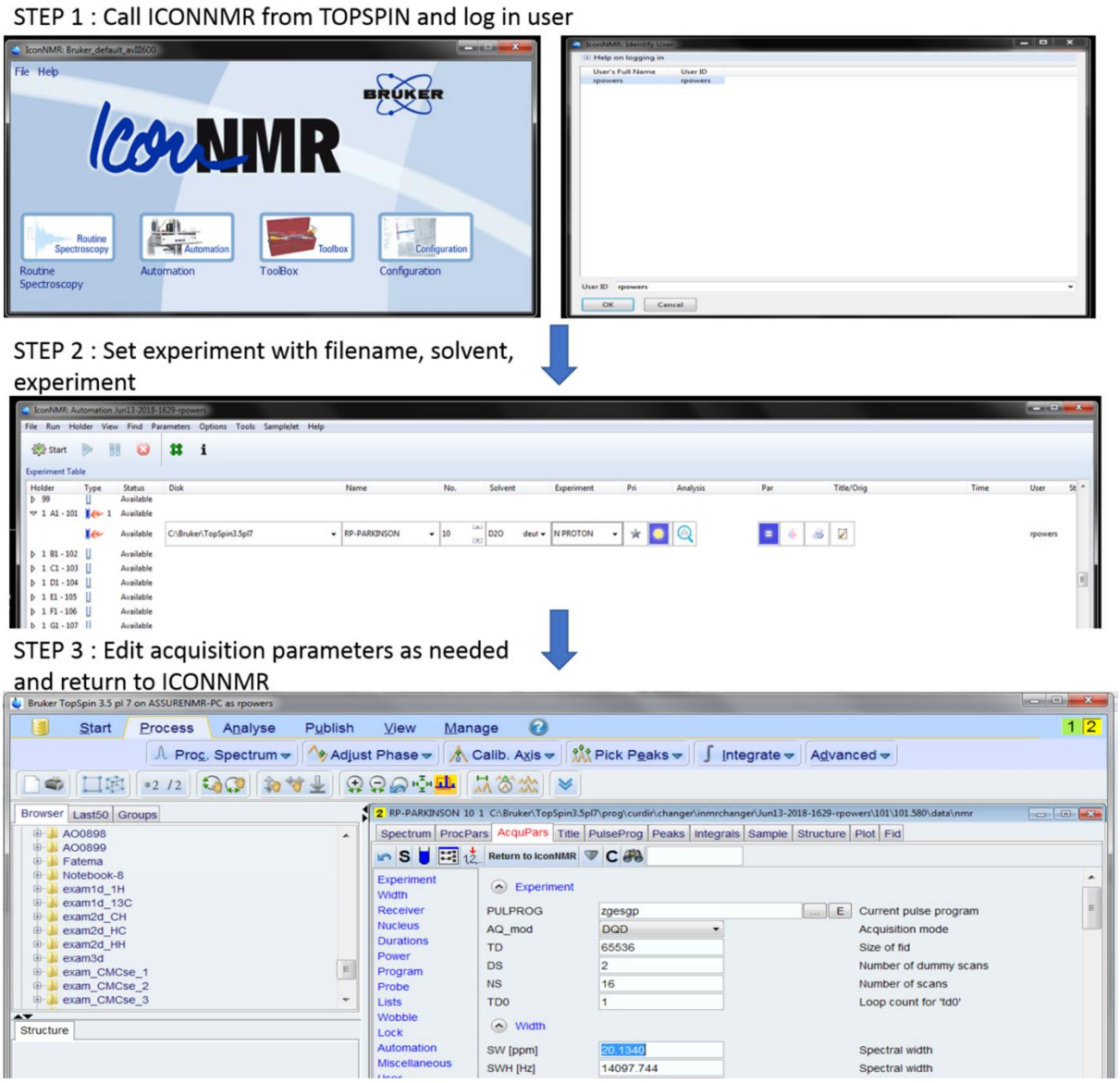



**Figure 2.5**: High-throughput NMR data collection. ICONNMR screenshots illustrating the stepwise workflow for setting-up a high-throughput 1D 1H NMR metabolomics screen.

### 2.3.9.1 1D $^{1}$H NMR

1. Load the NMR metabolomics samples into the SampleJet automated sample changer system (*see* **Figure 2.4**). Check that the SampleJet is in the correct mode (*i.e.*, 5 mm tubes)

2. Log into an account on the spectrometer workstation and start the Topspin software.

3. The first NMR sample is lowered into the magnet using the Bruker command, *sx 101*, where 101 corresponds to sample one in rack one.
4. The spectrometer is locked onto the $D_2O$ solvent frequency using the Bruker command, *lock $D_2O$*.
5. The NMR sample is shimmed for optimal signal and suppression of the water signal by typing the Bruker command *topshim*. This will initiate an automated gradient shimming procedure, which may take a few min to complete (*see* **Note 11**).
6. The sample is automatically tuned and matched using the ATM system by typing the Bruker command, *atma*.
7. The 90-degree pulse length (μs) is determined by measuring a null spectrum with an approximate 360-degree pulse using the Bruker **zg** pulse sequence (*see* **Note 12**).
8. A 1D $^1$H NMR spectrum is obtained for each sample using a standard excitation sculpting water suppression pulse program (Bruker **zgesgp** pulse sequence) that provides optimal suppression of the residual water signal while maintaining a flat baseline (*see* **Note 13**).
9. Typical experimental parameters for a 1D $^1$H NMR spectrum obtained on a Bruker 700 MHz spectrometer with a cryoprobe correspond to 128 scans, 16 dummy scans, 32,768 data points, a spectral width of 11,160.7 Hz, and a relaxation delay of 1.5 (*see* **Note 14**).
10. Automated data collection of the entire set of metabolomics samples is accomplished using Bruker ICONNMR 5 (*see* **Figure 2.5**).
11. The sample filename, solvent, pulse program and temperature parameters are all defined in Bruker ICONNMR 5 (*see* **Notes 15** to **17**).
12. Collect the 1D $^1$H NMR spectrum for each replicate and each group (*see* **Note 2.6**).

13. The data is processed initially with Topspin to verify spectral quality, but exported for further analysis (*see* **Figure 2.6A**).

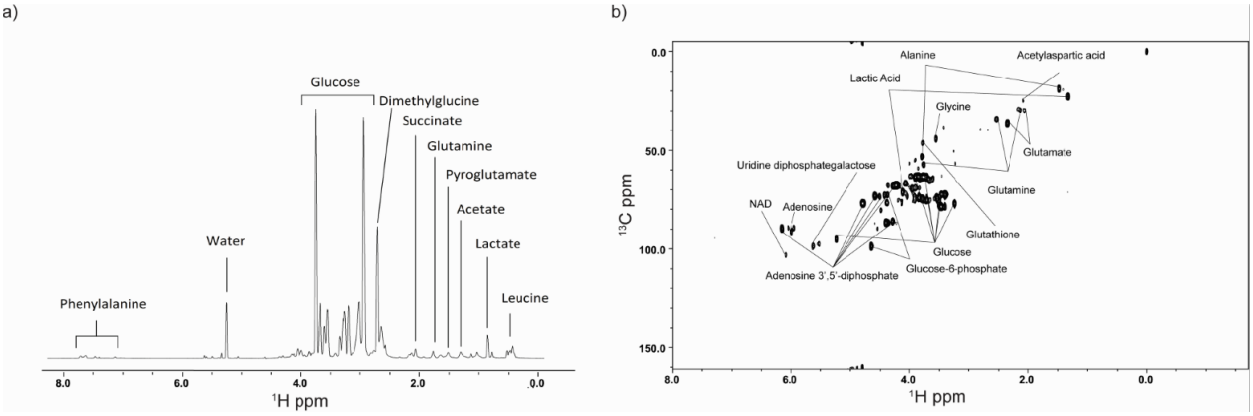



**Figure 2.6**: NMR metabolomics spectral data. Examples of a typical (A) 1D 1H NMR spectrum and a (B) 2D 1H-13C HSQC spectrum acquired from PD metabolomics samples.

### 2.3.9.2 2D $^{1}$H-$^{13}$C-HSQC NMR (*see* **Note 18**)

1. Follow steps 1 to 7 from section 3.9.1.
2. Using ICONNMR 5, the sample filename, solvent, pulse program and temperature parameters are adjusted (*see* Notes 15 to 17).
3. The ICONNMR setup is similar to a 1D $^{1}$H NMR data collection as shown in Figure 5.
4. A standard 2D $^{1}$H-$^{13}$C-HSQC experiment (Bruker hsqcetgpsisp2 pulse program) is used to determine the $^{1}$H-$^{13}$C chemical shift correlations for all $^{13}$C-labeled metabolites in the metabolomics sample (*see* Note 19).
5. Typical experimental parameters for a 2D $^{1}$H-$^{13}$C-HSQC NMR spectrum obtained on a Bruker 700 MHz spectrometer with a cryoprobe correspond to 128 scans, 32 dummy scans, and a 1.0 s relaxation delay. The spectrum is collected with 2 K data

points and a spectrum width of 4,734 Hz in the direct dimension and 64 data points and a spectrum width of 18,864 Hz in the indirect dimension (*see* Note 14).

6. Implementation of fast NMR methods that includes non-uniform sampling significantly decreases data acquisition time and/or increases spectral resolution, but may introduce artifacts (*see* Note 20).
7. Collect the 2D $^{1}$H-$^{13}$C-HSQC NMR spectrum for each replicate and each group (*see* Note 6).
8. The data is processed initially with Topspin to verify spectral quality, but exported for further analysis (*see* Figure 2.6B).

### 2.3.10 Mass Spectrometry Data Collection

#### 2.3.10.1 Direct-Injection (DI) Mass Spectrometry

1. Positive-ion direct infusion electrospray ionization mass spectrometry (DI-ESI–MS) data are collected on a Synapt G2 HDMS quadrupole time-of-flight (TOF) MS instrument equipped with an ESI source.
2. The mass spectrometry experiments are carried out at a flow rate of 10 μL/min for 1 min.
3. The mass spectra are acquired in positive ion and negative mode over a mass range of *m/z* 50 to 1200.
4. Mass spectra are acquired for 0.5 min using the following optimized source conditions: 2.5 kV for ESI capillary voltage, 60 V for sampling cone voltage, 4.0 V for extraction voltage, 80 °C for source temperature, 250 °C for desolvation temperature, 500 L/h for desolvation gas, and 15 μL/min flow rate of injection

5. Collect the DI-ESI–MS mass spectral data for each replicate and each group (*see* Note 6 and Figure 2.7).

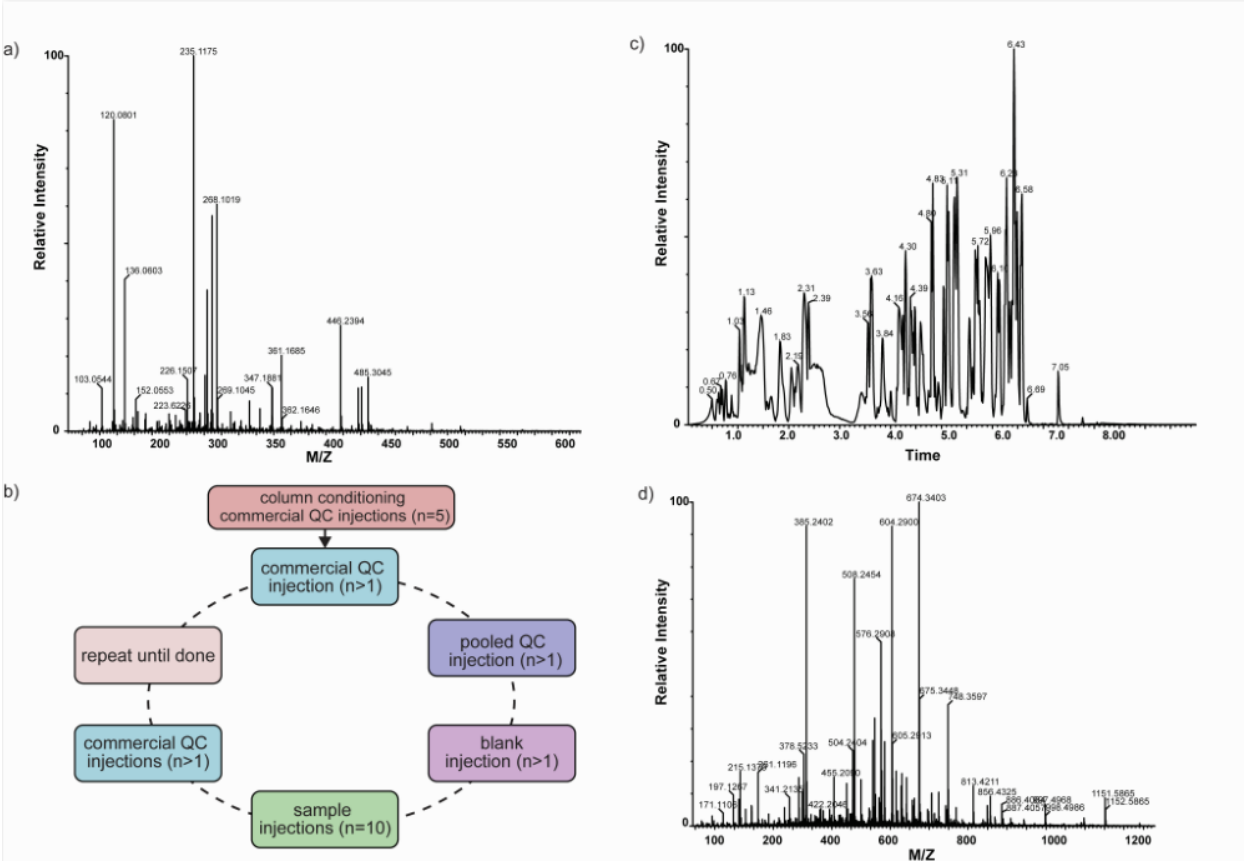



**Figure 2.7**: MS metabolomics data. Examples of typical (A) DI-ESI-MS spectrum, (B) typical analysis sequence, (C) LC base peak chromatograms and (D) MS spectrum acquired from metabolomics samples.

### 2.3.10.2 Liquid Chromatography - Mass Spectrometry

1. Liquid chromatography - mass spectrometry (LC–MS) data are collected on a Waters ACQUITY M-class Xevo G2-XS QToF MS instrument equipped with an ESI source.

2. The LC-MS mass spectrum are acquired with the following system parameters:

    - LC System: Waters ACQUITY M-class
    - Column:  ACQUITY HSS T3 Column 1.0 mm x 100 mm

- Mobile phase A: 0.1% Formic Acid in Water
- Mobile phase B: 0.1% Formic Acid in Acetonitrile
- Flow rate: 70 µL/min
- Run time: 10 min
- Injection volume: 2 µL
- MS system: Xevo G2-XS QtoF
- Ionization mode: ESI + and –
- Capillary voltage: 2.8 kV
- Cone voltage: 30 V
- Source temp: 120 °C
- Desolvation temp: 500 °C
- Cone gas flow: 18 L/h
- Lock mass:
  - Positive mode: Leukin- Enkephalin, *m/z* 556.2771
  - Negative mode: Leukin- Enkephalin, *m/z* 554.2615
- Acquisition mode: MSE
- Acquisition range: 50 to 1200 m/z
- Collision energy (LE): 6 eV
- Collision energy (HE): 20 to 40 eV

3. The temperature for the LC column and auto sampler is set to 40 °C and 4 °C, respectively.
4. Create a sample analysis sequence and inject the QC samples five times for column conditioning. After second QC injection, monitor peak area (<25% RSD), retention

time (+/- 0.05 min), and mass accuracy (+/- 3ppm) until the end of the fifth injection. If the QC samples pass the minimal system performance parameters, then acquire data. If not, do not collect data until the issue has been resolved and the QC samples pass the minimal system performance parameters.

5. Collect the LC-MS mass spectral data for each replicate and each group (*see* **Note 6** and **Figure 2.7**).

### 2.3.11 NMR Data Processing (*see* **Note 21**)

All NMR data is processed and analyzed with our MVAPACK software [21], our PCA/PLS-DA utilities [22], NMRPipe [23], and NMRViewJ [24]. See example processing scripts at http://bionmr.unl.edu/wiki/Scripts.

#### 2.3.11.1 1D $^{1}$H NMR (*see* **Figure 2.8A)**

1. A 1.0-Hz exponential apodization function is applied to the FID.
2. Fourier transform the FID.
3. The resulting NMR spectrum is automatically simultaneously phased corrected and normalized with the phase-scatter correction algorithm [45].
4. The NMR spectrum is referenced to the peak of TMSP-$d_4$ (0.0 ppm).
5. Noise and solvent regions are manually removed.

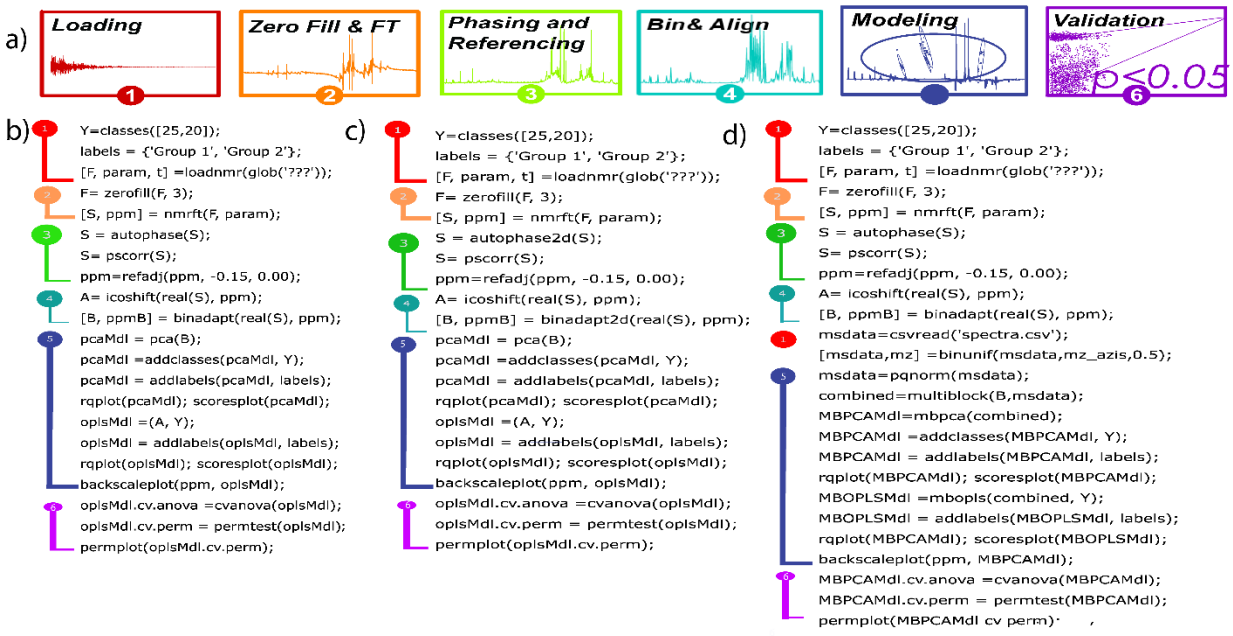



**Figure 2.8**: MVAPACK processing scripts. **(A)** Schematic illustration of the major processing steps. Examples MVAPACK processing script for **(B)** 1D 1H NMR dataset, **(C)** 2D 1H-13C HSQC dataset, and **(D)** combined NMR and MS datasets. The numbered steps in the flow diagram correspond to the numbered lines in the processing scripts.

#### 2.3.11.2 2D $^{1}$H-$^{13}$C-HSQC NMR (*see* Figure 2.8B)

1. A sine-bell apodization function is applied to the $t_2$ dimension.
2. The $t_2$ dimension is zero filled three times.
3. The $t_2$ dimension is Fourier transformed, manually phase corrected and the imaginary data deleted.
4. The matrix is transposed.
5. A sine-bell apodization function is applied to the $t_1$ dimension.
6. The $t_1$ dimension is zero filled three times.
7. The $t_1$ dimension is Fourier transformed and manually phase corrected.

8. The NMR spectrum is referenced in both dimensions to the peak of TMSP-$d_4$ (0.0 ppm).

### 2.3.12 Mass Spectrometry Data Processing - DI-ESI–MS (*see* Figure 2.8C)

1. Mass spectral data processing is first performed using MassLynx V4.1.
2. A background subtraction is performed on all spectra using appropriate reference spectra, such as a free drug or toxin used to treat a cell culture. The background subtraction of each spectrum is performed in a class-dependent manner (*i.e.*, only the MS reference spectrum of the drug/toxin used to treat the cell culture is used for background subtraction). Accordingly, mass spectral signals from the drug/toxin treatments are guaranteed to not influence subsequent analyses. An example of a typical MS spectrum from a metabolomics sample is shown in **Figure 2.7.**
3. The background-subtracted mass spectra are then loaded into MVAPACK as a text file for binning and normalization.

### 2.3.13 Mass Spectrometry Data Processing - LC-MS (*see* Figure 2.9)

All LC-MS data is processed and analyzed with Progenesis QI (version 2.0.). Please see the Progenesis QI user guide (http://storage.nonlinear.com/webfiles/progenesis/qi/v2.2/user-guide/Progenesis_QI_User_Guide_2_2.pdf) for detailed instructions.


**Figure 2.9**: LC-MS processing protocol. The small molecule discovery workflow using the Progenesis QI software is diagrammed. (top left) Summary of the major steps in the LC-MS processing protocol, which also describes each figure block in order starting from middle-left to bottom-right. Images are screenshots from the Progenesis QI software.

#### 2.3.13.1 Data Upload

1. Go to *File* and create a new experiment. Select a location to store the experiment file. Click *Next*.
2. Select the machine type (*i.e.*, high resolution mass spectrometer) and the polarity used to collect the mass spectrum (*i.e.*, positive or negative). Click *Next*.
3. Select the expected adducts [*e.g.*, $M+Na^+$ (+), $M+H+CH_3OH^+$ (+,-), $M+H+CH_3N^+$(+,-), $M+H_3O^+$ (+)] and click *Create experiment*.

4. Go to *Select your run data*, choose the MS *Format* and click *Import*. Click *Next*.
5. Apply Lock mass calibration. Click *Next*.
6. Select *Import*.

#### 2.3.13.2 Perform Automatic Processing

1. Click on *Start automatic processing*.
2. Select an alignment reference by choosing *Use the most suitable run from candidates that I select*. Click *Next*.
3. Select all QC runs. Click *Next*.
4. Select *Yes, automatically align my runs*. Click *Next*. Click *Next* again
5. After processing is complete, click *Section Complete* to move forward to the *Review Alignment* stage.

#### 2.3.13.3 Review Alignment

1. Interrogate the number of vectors and alignment scores.
2. Examine the distribution of green (good alignment), yellow (acceptable alignment) and red (needs review) alignments present in the ion intensity map.
3. As necessary, manually edit the alignments. Make sure that each ion is properly aligned across all replicates and to the reference mass spectrum. This is accomplished by interactively adjusting the alignment vector positions.
4. After processing is complete, click *Section Complete*.

#### 2.3.13.4 Create Experiment Design

1. Choose the type of experiment and click ***Create*** (*see* **Note 22**).
2. Click *Add condition* and rename it according to the groups in the study (e.g., control, treated, *etc.*).
3. Drag and drop each replicate mass spectrum into each of the defined groups from **2**.
4. After processing is complete, click *Section Complete* to move forward to the *Peak Picking* stage.

#### 2.3.13.5 Peak Picking

1. Click *Change parameters.*
2. Go to the *Peak picking limits* grid and define a minimum peak width to reject noise spikes. A typical minimum peak width is 0.05 min.
3. Click *Start peak picking*.
4. After the process is completed, go to *Review normalization* and choose the normalization method. Normalize to all metabolites is the default. A preferred choice is to normalize to an internal standard (*e.g.*, reserpine).
5. After processing is complete, click *Section Complete* to move forward to the *Deconvolution Review* stage.

#### 2.3.13.6 Review Deconvolution (*see* Note 23)

1. Go to the *Deconvolution Review* grid.
2. On the left panel, choose *organize the compound features by adducts*.
3. Click over an ion metabolite to review its adducts (*see* **Note 24**).

4. To remove an adduct assigned to a metabolite, *right click* on the peak in the adduct panel and click *Remove from compound*.
5. After the processing is complete, click *Section Complete* to move forward to the *Compound Statistics* stage.

### 2.3.14 NMR Data Preprocessing for Multivariate Modeling

In order to obtain an accurate and reliable multivariate statistical model, it is essential that the data set is properly preprocessed to remove normal systematic variations resulting from biological variability, instrument instability, and inconsistency in sample handling and preparation. Key preprocessing steps include: (1) alignment, (2) normalization, (3) binning, and (4) scaling, which is illustrated in Figure 2.8. Examples of results from a variety of statistical models are shown in Figure 2.10. All NMR datasets are processed with our MVAPACK software [21] and our PCA/PLS-DA utilities [22]. See example MVAPACK scripts at http://bionmr.unl.edu/wiki/Scripts.

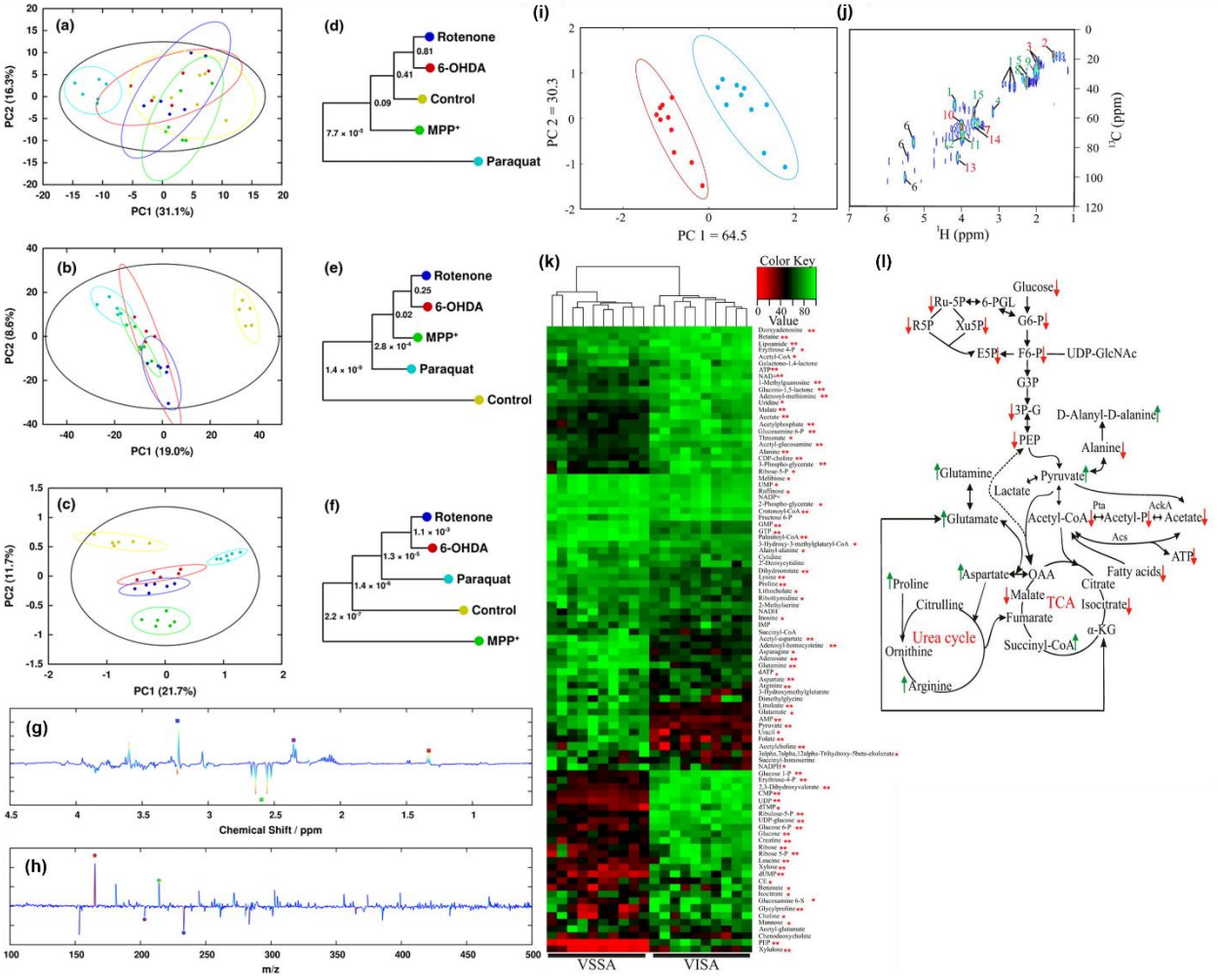



**Figure 2.10**: Univarate and multivariate statistical models. Example of PCA scores plot and the associated metabolomics tree diagram for **(A,D)** 1D 1H NMR dataset, **(B,E)** DI-ESI-MS dataset, and **(C,F)** combined 1D 1H NMR DI-ESI-MS dataset. **(G)** NMR and **(H)** MS back-scaled loadings from an OPLS model generated from combined 1D 1H NMR DI-ESI-MS dataset. Reproduced with permission from [61]. **(I,J)** PCA scores plot and OPLS back-scaled loadings generated from 2D 1H-13C HSQC NMR data set. Reproduced with permission from [1]. **(K)** Example heat-map with hierarchal clustering summarizing specific metabolite changes per replicate and the relative clustering of each individual replicate. Reproduced with permission from **[L]**. Example metabolic pathway summarizing the major metabolite changes between the two groups. Reproduced with permission from [1].

#### 2.3.14.1 1D $^{1}$H NMR

1. Spectra may first be normalized based on either the total cell count or the total protein concentration using the BCA (Biscinchronic Acid) protein estimation assay using parallel dishes treated similarly on the same day.
2. Spectra are normalized with the PSC algorithm [46].
3. Spectra are aligned and/or binned. For principal component analysis (PCA),use the following parameters:
    - The spectral data are globally aligned to the peak of TMSP-$d_4$ at 0.0 ppm.
    - The spectral data are regionally aligned using the icoshift algorithm [47].
    - The spectral data are binned using the adaptive, intelligent binning algorithm [48].

    For orthogonal projection to latent structures (OPLS), use the following parameters:
    - The spectral data are globally aligned to the peak of TMSP-$d_4$ at 0.0 ppm.
    - The spectral data are regionally aligned using the icoshift algorithm [47].
    - The spectral data is not binned. Instead, the full-resolution spectral data is used to build the model.
4. Solvent peaks and noise regions are manually removed.
5. The data set is scaled using Pareto scaling.
6. A PCA or OPLS model is generated from the data matrix.

#### 2.3.14.2 2D $^{1}$H-$^{13}$C-HSQC NMR

1. Spectra may first be normalized based on either the total cell count or the total protein concentration as explained above.
2. The spectral data is normalized using standard normal variate normalization.
3. The spectral data is binned using a generalized adaptive, intelligent binning algorithm [48].
4. The data are Pareto-scaled.
5. A PCA or OPLS model is generated from the data matrix.

### 2.3.15 Mass Spectrometry Data Preprocessing for Multivariate Modeling

LC-MS datasets need to be preprocessed in a similar manner to NMR spectra. The LC-MS datasets are processed with Progenesis QI (version 2.0.).

#### 2.3.15.1 DI-ESI–MS

1. All mass spectra are linearly re-interpolated onto a common axis that spanned from *m/z* 50 to 1,200 in 0.003 *m/z* steps, resulting in 383,334 variables prior to processing.
2. The mass range *m/z* 1,100 to 1,200 is removed prior to binning because of the low probability of observing a metabolite in this region.
3. The mass spectra are uniformly binned using a bin width of 0.5 *m/z*, resulting in a data matrix of 2,095 variables.
4. The MS data matrix is normalized using probabilistic quotient (PQ) normalization.
5. The MS data matrix is then scaled to unit variance prior to modeling.

6. A PCA or OPLS model is generated from the data matrix.

### 2.3.15.2 LC-MS (*see* **Note 25 and Figure 9)**

1. The LC-MS datasets are processed with Progenesis QI (version 2.0.). Please see the Progenesis QI user guide (http://storage.nonlinear.com/webfiles/progenesis/qi/v2.2/user-guide/Progenesis_QI_User_Guide_2_2.pdf) for detailed instructions.
2. Right click on the *Compounds table* and select *Quick Tags.*
3. Set the ANOVA cutoff value to 0.05.
4. Click *Create tag*.
5. All metabolites with an ANOVA p-value ≤ 0.05 will be marked with a red tag.
6. Repeat the process to create a tag for fold change (*see* **Note 29**). Right click on the *Compounds table* and select *Quick Tags*
7. Set the fold change cutoff value to 2.
8. All metabolites with a fold change greater than 2 will be marked with a green tag.
9. Create a filter to show only tagged metabolites. Click *Create* on *Filter* grid to open the filter dialog box.
10. Select the tags and then drag to the box *Show compounds that have all these tags*. Click *OK*.
11. Only the metabolites that match the criteria are showed and will be used for metabolite identification.
12. Go to the *Compound statistics* grid. The statistical analysis is available as a PCA scores plot. A statistically relevant dataset is indicated by replicate samples

clustering together in the scores plot. Furthermore, the set of control and treated replicates form distinct clusters from each other.

13. Go to *File* and select *export all measurements.* A comma-separated value (csv) file will be created with a list of several values per metabolite: (1) metabolite identification, (2) *m/z* value, (3) charge, (4) retention time, (5) relative abundance, (6) ANOVA value and other parameters.

### 2.3.16 Statistical Analysis (*see* Figure 10)

Data sets are analyzed with our MVAPACK software [21], our PCA/PLS-DA utilities [22], and R [49]. See example MVAPACK and R scripts at http://bionmr.unl.edu/wiki/Scripts.

A major challenge in the analysis of metabolomics datasets, and a common source of error, is the incorrect application of statistics. This results from a number of prevailing misconceptions within the metabolomics community. For example, a multivariate model, especially supervised methods such as PLS or OPLS, needs to be properly validated. Validation can be accomplished with CV-ANOVA [50] and/or response permutation testing [51]. Conversely, the resulting $R^2$ and $Q^2$ values only provides a measure of the model fit to the original data and an internal measure of consistency between the original and cross-validation predicted data, respectfully. $R^2$ and $Q^2$ values do not provide for model validation without a proper standard of comparison.

PCA, PLS, and OPLS are routinely used to model metabolomics data. Nevertheless, there are misconceptions regarding the proper application and interpretation of the resulting models,

especially in regards to comparing PCA, PLS, and OPLS models. For example, PCA finds the largest source of variance in the dataset irrespective of the intent of the study. So, an observed separation between treated and untreated groups in a PCA scores plot may have nothing to do with the treatment if some other larger variant is present in the dataset. Supervised methods, like PLS and OPLS, address this issue by aggressively forcing group separation based on the defined group membership. Hence, PLS and OPLS models almost always yield separated groups, as by design! As a result, PLS and OPLS models are easily over-fitted, especially for metabolomics data sets since the number of variables (*e.g.*, metabolites) are typically larger than the number of replicates. Again, model validation is essential for PLS and OPLS.

Another serious misconception is the false belief that PLS/OPLS is a *better* method than PCA and may find group differences when PCA fails to expose group separation. Instead, PCA, PLS and OPLS are simply different models that extract different information and achieve different goals. Thus, if PCA fails to identify group separation it is unlikely that PLS/OPLS will yield a valid model [51]. Remember, PCA finds the largest source of variance. If PCA doesn't find any major variance, then there cannot be any smaller group-specific variance.

PLS and OPLS appear to provide similar models. In fact, a comparison of PLS and OPLS scores plots generated from the same data set may suggest the only difference is a relative rotation of the group-defined ellipses. Nevertheless, this apparently subtle change highlights a critical difference. Simply, OPLS places group-independent variance (*e.g.*, confounding factors such as differences in diet, age, race, *etc.*) orthogonal to group-dependent variance. Conversely, PLS entangles both

group-independent and group dependent variance. In this regards, a metabolite identified as a major contributor to an OPLS model is strictly the result of the defined group difference. For PLS, metabolite changes may be a result of the group difference, a confounding factor or a combination of both. Accordingly, a PLS identified metabolite may be of little interest to the intent of the study. In this regard, we strongly recommend always using OPLS instead of PLS.

#### 2.3.16.1 Univariate Analysis

1. Relative metabolite abundances are inferred from NMR and/or mass spectral peak heights and/or peak volumes.
2. Relative metabolite abundances are then normalized on a per spectrum basis. One common approach is to convert the absolute peak intensities (arbitrary units) to a Z-score:

$$Z = \frac{I_i - \bar{I}}{\sigma} \quad (1)$$

where $\bar{I}$ is the average peak intensity for the spectrum, $I_i$ is the intensity of peak *i*, and $\sigma$ is the standard deviation of peak intensities. Peak intensities may also be normalized to the total number of cells, to the total protein concentration (*see* **section 2.3.14.1**), to the average spectral noise, or to an internal standard (*see* **Note 26**). Relative metabolite abundances may also be converted to fold-changes:

$$F = \frac{I_i}{I_o} \quad (2)$$

where $I_i$ is the normalized peak intensity of metabolite *i* from a treated spectrum and $I_o$ is the normalized peak intensity of metabolite *i* from the control or untreated spectrum.

3. A standard Student's t-test is commonly used to determine statistical significance only for a pairwise comparison of metabolite changes based on either fold-changes or normalized peak intensities (*see* **Note 27**). A statistically significant difference is typically identified by a p-value < 0.05.
4. A Student's t-test is insufficient for the multiple comparisons that are common to a metabolomics study [51, 52]. In order to identify the set of metabolites that exhibit a statistically significant change, a multiple hypothesis test correction method such as a Benjamini-Hochberg [53] or a Bonferroni [54] correction must be applied (*see* **Note 28**).
5. A heat-map with hierarchical clustering (*see* **Figure 2.10k**) is commonly generated from the fold-changes or normalized peak intensities using R (*see* example R script at bionmr.unl.edu/wiki/scripts). The heat-map may contain relative metabolite abundances for each individual replicate in the study or simply the replicate-averages for each group (*see* **Note 29**).

#### 2.3.16.2 Multivariate Analysis

1. Generate a PCA and or OPLS model from the data matrix.
2. Fractions of explained variation ($R^2_X$ and $R^2_Y$) are computed during PCA or OPLS model training.

3. The PCA or OPLS model is internally cross-validated using seven-fold Monte Carlo cross-validation [55] to compute $Q^2$ values (*see* **Note 30**).
4. For an OPLS model, the $Q^2$ value is compared against a distribution of null model $Q^2$ values in 1000 rounds of response permutation testing [51]. Group membership is randomly reassigned to generate the set of null models. A p-value is calculated from a comparison of the true $Q^2$ value to the set of null model $Q^2$ values (*see* **Note 31**).
5. The model is further validated using CV-ANOVA significance testing, which is used to calculate another model p-value [50] (*see* **Note 31**).
6. Scores plots (*see* **Figure 2.10a,b,c,i**), back-scaled loadings plots (*see* **Figure 2.10g,h,j**), S-plots and/or SUS-plots are often generated from OPLS models.
7. PCA/PLS-DA utilities [22] is used to define group membership by drawing an ellipse per group onto the scores plots (*see* **Figure 2.10a,b,c,i**). Each ellipse corresponds to 95% confidence interval for a normal distribution. The PCA/PLS-DA utilities also generates a metabolomics tree diagram that identifies the statistical significance (p-value and/or bootstrap value) and the relative similarity of each group in the scores plot (*see* **Figure 2.10d,e,f**). The p-value or bootstrap number from the pairwise comparison is labeled at each node in the tree.

### 2.3.17 Data Analysis - Metabolite Assignment from 1D $^1$H NMR Data

All NMR data is analyzed with NMRPipe [23], NMRViewJ [24], and Chenomx. See example scripts at http://bionmr.unl.edu/wiki/Scripts.

1. The identification of metabolites in a 1D $^{1}$H NMR spectrum is performed with software programs such as Chenomx. Chenomx matches the experimental 1D $^{1}$H NMR spectrum to a database of 1D $^{1}$H NMR spectra of known metabolites. Chenomx attempts to explain or describe the experimental NMR spectrum by combining or summing as many of the individual reference metabolite NMR spectra as needed. In addition to metabolite identification, Chenomx also provides an estimate of the metabolite concentration (*see* **Note 32**).
2. Upload the 1D $^{1}$H NMR spectrum for processing. The NMR spectra can be batch processed or processed one at a time.
3. The 1D $^{1}$H NMR spectrum is phased.
4. The 1D $^{1}$H NMR spectrum is calibrated and reference to TMSP-$d_4$, using the known concentration of TMSP-$d_4$.
5. The properly phased and calibrated 1D $^{1}$H NMR spectrum is then sent to the Chenomx profiler where the spectrum is compared against the metabolite library.
6. Chenomx will overlay a 1D $^{1}$H NMR reference spectrum for each metabolite identified in the experimental 1D $^{1}$H NMR spectrum. The spectral overlay needs to be manually adjusted to optimize the alignment of the experimental 1D $^{1}$H NMR spectrum with the reference spectrum. **Figure 2.6A** shows an example of a labeled 1D $^{1}$H NMR spectrum.
7.

### 2.3.18 Data Analysis - Metabolite Assignment from 2D $^{1}$H-$^{13}$C-HSQC NMR Data

All NMR data is analyzed with NMRPipe [23], NMRViewJ [24], and Chenomx. See example scripts at http://bionmr.unl.edu/wiki/Scripts.

#### 2.3.12.1 NMRPipe Processing to Obtain *.ft2* and *.nv* Files.

1. The data files from ICONNMR can be used directly by NMRPipe to process the 2D $^{1}H$-$^{13}C$-HSQC spectra.
2. On a Linux workstation, open a terminal and go to the directory that contains the NMR data. Type *bruker* to start the NMRpipe software.
3. Read in the experimental parameters file by clicking *Read Parameters* and verify that all of the parameters have been correctly updated. Confirm that the mode of data collection has been set to *echo-antiecho* if the NMR spectrum was collected with the **hsqcetgpsisp2** pulse program.
4. Click *Update Script* to save an NMRPipe processing script *fid.com* file in the working directory.
5. Type *./fid.com* to start the NMRPipe processing script.
6. When the NMRPipe processing has finished, type *nmrDraw* to view the processed NMR spectrum. Please see the NMRPipe and nmrDraw tutorial (https://spin.niddk.nih.gov/NMRPipe/doc1/) for detailed instructions.
7. Phase the NMR spectrum in NMRpipe and note the p0 and p1 values for both the $^{1}H$ and $^{13}C$ dimensions.
8. Edit the NMRPipe processing script *hsqcproc.com* and replace the parameters associated with the NMRPipe phase correction command, *ps*, with the p0 and p1 values obtained from **step 7**.
9. Type *./hsqcproc.com* to start the NMRPipe processing script

10. Repeat **steps 3** to **9** for each 2D $^{1}$H-$^{13}$C-HSQC NMR spectrum in the dataset. This produces a set *.ft2* files. One *.ft2* file is created for each 2D $^{1}$H-$^{13}$C-HSQC NMR spectrum collected for each replicate from each group.
11. Copy all of the *.ft2* files into a new folder and use the NMRPipe script *addnmr.com* to generate NMRviewJ files from the *.ft2* files. A *.nv* file will be generated for each individual spectrum (*.ft2* file) with an numerically incremented root name of "*Final_*". In addition, the script will combine all of the NMR spectra together into a single file called *results.nv*. The script will also generate the text file, *rate.txt*, that lists all of the individual *.nv* files (*Final_*).

#### 2.3.18.2 Peak Picking and Peak Integration of 2D $^{1}$H-$^{13}$C-HSQC Spectra in NMRviewJ.

1. Type *nmrviewj* to start NMRviewJ. Please refer to NMRViewJ documentation (http://docs.nmrfx.org/) for more details.
2. From the *Dataset* toolbar in the *main* window, use the *Open and Draw Datasets* function to select the *result.n*v file.
3. Right click and select *attributes* to open the *attributes* window.
4. In the *attributes* window, select the *PeakPick* tab.
5. In the blank *Lists* field in the *attribute* window, type a filename (i.e., *lists*) for the new peak pick list. Click the *Pick* button. The software will automatically peak pick the displayed spectrum and populate *lists* with the peak ID number, chemical shifts, and intensity.

6. Choose *Show Peak Table* from the *Peak* toolbar on the *main* window. A *peak table* window will open that lists the peak ID, peak intensity and the peak chemical shifts.
7. Manually edit the peak list and remove solvent peaks, noise peaks or other spectral artefacts. Peaks are deleted from the peak table by using the delete function in the *PeakPick* tab in the *attributes* window along with the *spectrum display* window. In the *spectrum display* window, use the mouse to position the two cursors around any peak or spectral region to form a box. Then, click the *Delete* button under the *PeakPick* tab in the *attributes* window to remove the peak(s).
8. After the peak table has been completely edited, on the *peak table* window choose the *Edit* tab and select *Compress & Degap.* Answer *yes* to the pop-up question. This will finalize changes to the peak list and prevent any further edits.
9. On the *peak table* window choose the *Edit* tab and then select *Save Table*. A file browser window will open in order to choose a name and location to save the new peak list file. The saved peak pick file can be viewed and edited by Excel.
10. In order to obtain peak intensities across the entire set of NMR spectra in the dataset, click on the *Analysis* tab on the main window and select *Rate Analysis*. A set-up window for the *Rate Analysis* will open.
11. In the *Rate Analysis* set-up window:
    - set the *Prefix for matrix numbers* field to *Final_*
    - set the *Peaklist* field to *lists* (defined in **section 2.3.18.2 step 5**).
    - make sure *Auto fit* is checked.
    - use all other default settings.
    - Click *Load time file.*

- In the file browser window, select *rate.txt* (created in **section 2.3.18.1 step 11**).
- Click *Measure All.* The software automatically populates the table in the *Rate Analysis* set-up window with all of the peak intensities across the entire NMR dataset.
- Click *Save Table.* In the file browser window, save the peak intensities table to a new filename (*i.e., intensities*).

12. The peak list (*i.e., list*) and the peak intensities (*i.e., intensities*) files are merged in Microsoft Excel using the common peak ID column. The *ppm1* ($^{1}$H ppm) and *ppm2* ($^{13}$C ppm) columns are added to the peak intensities columns to generate a complete matrix of NMR peaks and intensities across the entire data set.
13. The merged Excel file is saved to a new filename.

#### 2.3.18.3 Metabolite Assignments from 2D $^{1}$H-$^{13}$C-HSQC Peak Lists

1. The complete list of peaks obtained from the NMRviewJ analyses is searched using NMR metabolomics databases such as HMDB [28], BMRB [29], or other databases (*see* **Note 33**).
2. On the HMDB homepage, choose the *Search* tab and select *2D NMR Search*.
3. From the *Spectra Library* pull-down menu, choose *13C HSQC.*
4. Cut and paste the 2D $^{1}$H-$^{13}$C-HSQC peak lists into the *Peak List* field. One set of $^{1}$H and $^{13}$C chemical shifts, respectively, per line. Chemical shift values should only be separated by white space.
5. Set the $^{1}$H chemical shift error tolerance to 0.05 ppm (*X-axis Peak Tolerance* ± field) and the $^{13}$C chemical shift error tolerance to 0.10 ppm (*Y-axis Peak Tolerance* ± field).

6. Click the *Search* button. Depending on the size of the peak list, the software will return a ranked-order list of possible metabolites based on the number of chemical shift matches to reference spectrum.
7. Manually curate the list of potential metabolite assignments based on the number of chemical shift assignments, the quality of the spectral overlap (i.e., chemical shift match), number of other metabolites in the same metabolic pathway, and the biological system (i.e., is it a reasonable or possible metabolite for the organism),
8. Obtain additional NMR (*e.g.*, HMBC, HSQC-TOCSY) and/or MS spectral data to confirm or refute the assignment.
9. An assigned 2D $^{1}$H-$^{13}$C-HSQC spectrum is shown in **Figure 2.6B**.

### 2.3.19 Data Analysis - Metabolite Assignments from LC-MS Data

The identification process is accomplished using the Progenesis QI (version 2.0.) software **(***see* **Note 34** and **Figure 2.9).** Please see the Progenesis QI user guide for detailed instructions (http://storage.nonlinear.com/webfiles/progenesis/qi/v2.2/user-guide/Progenesis_QI_User_Guide_2_2.pdf).

#### 3.19.1 Identification of Compounds (*see* **Note 36**)

1. Make sure the filter created in **section 2.3.15.2 step 9** is applied and then proceed to *Identity Compounds* grid.
2. At the left panel, define the method to be used. In this case, select *Progenesis MetaScope.*

3. Choose the search parameter, in this case choose HMDB (*see* **Note 35**).
4. Click *Search for identifications*.
5. After few min, a dialogue box will open identifying the number of metabolites identified. Click *ok* to close.
6. All ions with possible identifications will presented as a solid gray icon on the left side.

#### 2.3.19.2 Incorporation of Theoretical Fragmentation (*see* Note 36).

1. On the left panel, select *ChemSpider* [27] as the identification method (*see* **Note 37**).
2. In the *Choose search parameter* field, choose *default* and then click *edit*.
3. Set the following parameters:
    - Select name as *theoretical fragmentation.*
    - Set *precursor tolerance* to 5 ppm.
    - Tick *Perform theoretical fragmentation* box.
    - Set the *Fragment tolerance* to 5 ppm.
4. Click *Save*.
5. Click *Search for identifications*.
6. After a few minutes, a dialogue box will open displaying the number of metabolites identified. Click *ok* to close.

#### 2.3.19.3 Accepting Compounds Assignment

1. Proceed to *Review Compounds* grid.

2. Go to the option *Choose the correct identification* and set a threshold of 45. The choice of a threshold is empirical and may need refinement based on the specific properties of the dataset. The higher the threshold setting, the more confident are the assignments, but the more restrictive analysis may result in a lower number of assignments.
3. Click *Accept identifications.* All identifications with a score of 45 or above will be accepted automatically.

#### 2.3.19.4 Review and Accept the Identifications Manually (*see* Note 38)

1. Select a metabolite from the list.
2. Go to the *Possible identifications* grid.
3. In the bottom panel, select the desired identification threshold for the metabolite.

## 2.4. Notes

1. Isotopically-labeled reagents commonly used for NMR are not radioactive and do not require special handling or safety precautions. However, gloves and eye protection are standard safety protocol for preparing all types of metabolomics samples.
2. Deuterated solvents, such as $D_2O$ or DMSO, are very hygroscopic and require storage in a drybox and need to remain sealed until used.
3. The pH of a 100% $D_2O$ sample using a standard pH probe may not report the correct pH. A standardly applied correction is: pD = pH + 0.4. Conversely, a recent study by K. A. Rubinson [56] suggests the variance is not as significant, especially for a phosphate buffer, and a correction may not be required.

4. Complete cell survivability in each group is essential to a successful metabolomics study. This may be particularly challenging in a study that involves treating cells with a drug, toxin or some other condition (including nutrient depletion or supplementation) that is expected to alter cell viability. Thus, the goal is to identify a dosage and time for the experimental paradigm that will stress the cells, prior to the induction of cell death. In this regard, the observed metabolomic changes will be a result of the cell's immediate response to the mechanism of action of the experimental condition, or the adaptation of the cell to the stress, and not a general cell death response. We typically identify the dosage by collecting a series of growth curves over a range of drug/toxin concentrations and compare them to a growth-curve from untreated cell culture.
5. The resulting composition of the metabolome is easily perturbed by any difference in the protocol. Thus, it is essential that every sample is handled in exactly the same manner as reasonably as possible. Bias can be induced if cells are cultured in different incubators or shakers, if cells are handled by different personnel, if cells are treated with a different wash, buffer or media (even if it is the exact same recipe as prepared by the same individual), or if the time to process the cells differ, *etc*. In essence, any source of variance (regardless of how slight) may lead to a significant biologically-irrelevant change in the metabolome. As a result, an important aspect of the protocol is to randomize the processing of each sample to minimize any bias induced by sample order. The order of sample processing should change at each step of the protocol. It is especially critical to randomly interleave replicate samples from each group.
6. Randomization of samples throughout the protocol is essential to avoid the introduction of bias. For example, if all of the control samples are processed together and first, and all of

the treated samples are processed second, a difference between the controls and treated samples may be due to the processing order instead of the expected response to treatment. Consider another example consisting of a set of twenty samples numbered 1 to 20. If the samples are always processed in the order of the sample number, then a time-bias will be induced across the entire dataset. Sample 20 will always be processed after a maximal wait-time and sample 1 will always be processed the quickest. Accordingly, biologically-relevant differences in the metabolomes will accumulate between the samples due to the difference in processing time. Instead, if the order is constantly changed at each step, the processing time and any impact on the metabolome will be randomized, which in turn will minimize or eliminate any bias.

7. The number of replicates per group will have a significant impact on the quality of the study and the statistical validity of the outcomes. In general, it is best to maximize the number of replicates per group, within reason, with a typical target of ten replicates per group. A variety of experimental considerations may impact the number of replicates that are practical for a given study. For example, a large number of groups may require a reduction in the number of replicates per group. Another consideration is the impact of the number of replicates on the quality of the metabolomics samples. Sacrificing quality for a greater number of replicates will not likely lead to a successful outcome. Conversely, a limited number of replicates < 4 per group will likely provide meaningless results.
8. Other studies have used a combination of isotopically labeled and non-labeled carbon sources. The conditions of optimal labeling should be standardized for every cell line/type used for experimentation considering the composition/recipe of the culture media and the required carbon sources (glucose, pyruvate or glutamine) for cell growth. A time course

between 1-48 h should be performed to assess the rate of carbon consumption. Examples of media used for $^{13}$C-carbon labeled metabolomics are Dulbecco's Modified Eagle Medium DMEM (11966-025, 10938-025, 11960-044 and A14430-01) and RPMI (11879020) from GIBCO/Life Technologies

9. Removal of proteins and other biomolecules by methanol or ethanol precipitation is preferred over mechanical filtration methods or the application of Carr–Purcell–Meiboom–Gill (CPMG) NMR $T_2$ filtering techniques. Filtering techniques may remove metabolites that bind to biomolecules leading to biologically-irrelevant group differences [111].
10. Smaller diameter NMR tubes of 3 mm (160 μL) or 1.7 mm (35 μL) may be needed if the available metabolomics sample is limited. Filling of these smaller diameter NMR tubes may require a liquid handling robot, such as a Gilson 215 Liquid Handler. In addition, the NMR acquisition parameters will likely need to be adjusted to account for the lower sensitivity due to the lower number of nuclei in the samples.
11. Topshim requires the sample contains either a $D_2O$ or $H_2O$ solvent. It is advisable to create a shim file with a parameter set that produces an optimal set of shims for your sample type. Read in a shim file using the Bruker command *rsh* and select the appropriate Topshim shim file. If you are doing this for the first time, complete the command *topshim*, if you are not satisfied with the shim performance use command *topshim tuneb tunea* to obtain an improved set of shims. Write the shim set parameters with the Bruker command *wsh* and save it to a new file name for future reference.
12. The 90 degree pulse length is commonly measured by incrementing the P1 pulse in the **zg** pulse program by 1 μs or smaller increments; and by plotting the relative peak heights or intensities. A maximum peak height should be observed at the pulse length corresponding

to the 90-degree pulse. Conversely, a minimum or null spectrum should be observed at the pulse length corresponding to the 360-degree pulse length. In practice, a more accurate measure of the 90-degree pulse is obtained by measuring the 360-degree pulse length and dividing by four to obtain the 90-degree pulse length. A typical 90-degree pulse length for a metabolomics sample ranges from approximately 8 μs to 13 μs or longer. Among other factors, the relative salt concentration of the metabolomics sample affects the 90-degree pulse, in which a higher salt concentration results in a longer 90-degree pulse. Other factors also contributed to the observed 90-degree pulse, so it is always necessary to experimentally determine the 90-degree pulse for each sample or set of samples.

13. Excitation Sculpting parameters (**zgesgp**) - 32768 data points (TD), SW = 12.02 ppm, O1P (transmitter offset) = 4.70 ppm, D1= 1 second, NS (number of scans) = 128, DS (dummy/steady state scans) = 16, P1 = 9.5 -13.5 us, SPNAM (shaped pulse for water suppression) = SINC1.1000 at 26.39 dB or 0.00228 W.
14. The NMR data acquisition parameters need to be adjusted to compensate for differences in the field strength and sensitivity of the NMR spectrometer actually used for the data collection. Specifically, the number of scans, the number of data points, the sweep-width (13.79 ppm, $^{1}$H frequency range) and the frequency-offset (centered on water peak at 4.70 ppm) need to be adjusted according to the type and configuration of the NMR spectrometer used for the study.
15. For high throughput NMR data collection please refer to the Bruker ICONNMR manual to explore various configuration options. For example, composite experiments allow for the collection of multiple 1D and 2D experiments for the same metabolomics sample. An experimental set consisting of a 1D $^{1}$H, a 2D $^{1}$H-$^{13}$C HSQC, and 2D $^{1}$H-$^{13}$C HMBC

experiment may be subsequently collected for the same sample before moving to the next sample in queue.

16. It is imperative that NMR data is collected at the same temperature for a queue of metabolomics samples. ICONNMR assists this by allowing for a temperature delay when a large number of samples are in the SampleJet queue. For example, a 15 to 60 second delay may be inserted prior to data acquisition to allow each sample to equilibrate to the probe temperature. We recommend a 60 second delay for both pre- and post- sample insertion to prevent any temperature variation.
17. Parameters to check before you queue experiments in ICONNMR for 1D $^{1}$H NMR are: number of scans *ns*, number of dummy scans *ds*, 90 degree pulse *p1*, delay *d1*, sweep width *sw*, receiver gain *rg*, experiment temperature *te*, and automation setup *aunm*. We recommend using *au_zgonly* as the automation setup. This will collect all samples at the same receiver gain, which will avoid peak intensity variation across the dataset.
18. In addition to 2D $^{1}$H-$^{13}$C-HSQC NMR experiments, NMR metabolomics studies may make use of HMBC, TOCSY, HSQC-TOCSY, 2D J-Resolved spectra, or other experiments. Similarly, $^{15}$N, $^{31}$P and other isotope-labeled metabolites may be detected in addition to $^{1}$H- and $^{13}$C-labeled metabolites. Accordingly, experimental parameters, data processing and preprocessing methods, and data analysis techniques all need to be adjusted to accommodate the specifics of each NMR experiment. Nevertheless, there is enough similarity that the detail discussion of the application of 2D $^{1}$H-$^{13}$C-HSQC NMR experiments may provide a useful initial guide to the application of other NMR experiments to metabolomics.

19. 2D $^{1}$H-$^{13}$C-HSQC parameters (**hsqcetgpsisp2**) - 1024 data point in F2 and F1, Non-Uniform Sampling at 25%, O1P = 4.7 ppm, O2P (offset for $^{13}$C) = 75 ppm, NS = 64, DS = 16, d1 =2, P1 = 10 - 13 μs depending on salinity, CPDPRG2 = garp (decoupling program), PCPD2 = 55 us at PLW12 = 4.09 W.
20. Non-uniform sampling of 2D $^{1}$H $^{13}$C HSQC data can be performed on metabolomics samples. We have successfully acquired data at 20% sparsity using a burst augmented scheduler available from http://bionmr.unl.edu/dgs-gensched.php [30]. Download the sampling schedules as a text file for Topspin.
21. A minimalistic approach to the processing of NMR and mass spectrometry data is optimal for a metabolomics analysis utilizing multivariate statistics such as PCA and OPLS. The resulting multivariate statistical model is dependent on the choice of processing and preprocessing protocols. In effect, a different statistical model is likely to be obtain based on the presence (or absence) of baseline correction and the type of baseline correction method chosen. Similarly, the type of weighting (apodization) function, the type of spectral alignment or referencing, the resulting phase correction or phase correction algorithm, the number of zero-fills or the application of linear-prediction, or any other data manipulation method will affect the outcome of the statistical model. Accordingly, it is best to avoid any unnecessary data processing steps since it is difficult to ascertain if the data processing induced a biologically-irrelevant bias to the data or actually improved the model.
22. Before proceeding to statistical analysis it is necessary to create an experiment design. Progenesis QI supports *Between-subject design* and *Within-subject design*. *Between-subject design* separates samples according to the experimental condition (control vs treated) for the statistical comparison. *Within-subject design* is a repeated-measures study

design where the same subject (i.e., cell, animal, or human) is compared across the full range of experimental conditions (before treatment and after treatment; different time points, etc.).

23. The ions and adducts for a compound are automatically recombined by Progenesis QI, but it is advisable to review the deconvolution results. It is important to make sure the same pattern of adducts are assigned equally across all replicates and between all groups. Progenesis compares each detected ion with each of its co-eluting ions. If by the chance, their mass difference matches the difference between two adduct masses (i.e., from the previously chosen list), then it is probably an adducted form of the same compound. Progenesis groups the two ions as the same compound and automatically assigns the ions to the respective adduct. However, if an interesting compound is identified in the sample, it is important to review the deconvolution process to make sure all of the ions grouped together are actually adducts of the same compound.
24. Adducts assigned to a compound should have the same retention time as the compound. Thus, compare the chromatograms from the potential adduct with the compound to determine how well the chromatograms overlay. If a poor match is observed, then remove the adduct.
25. A primary goal of the LC-MS data analysis is to identify metabolites that exhibit significant concentration differences between groups. This is accomplished in the Progenesis software by creating tags to identify metabolites that exhibit a statistically significant (ANOVA [110] p-value $< 0.05$) difference in relative abundance between the groups. Progenesis relies on PCA for this analysis.

26. For NMR, relative peak intensities are averaged across all replicates per group and also for each NMR peak assigned to the metabolite. Most metabolites will have more than one peak in an NMR spectrum and all NMR peaks should be incorporated into an average relative peak intensity. Please note, NMR peaks may need to be scaled by the number of attached hydrogens, since peak intensity is proportional to the number of nuclei.
27. Of course, there are a variety of options beyond the standard Student's t-test such as: Mann–Whitney U test [57], Welch's t-test [58], Hotelling's t-squared statistic [59], and one-way analysis of variance [51], among others. The proper choice of a statistical test depends on a number of factors, which is well-beyond the scope of this protocol review. For an introduction to the topic, please see *A Biologist's Guide To Statistical Thinking And Analysis* [60].
28. In effect, the uncertainty in each pairwise comparison (as determined by the Student's t-test) is compounded with the addition of each metabolite to a set. The actually p value for a set of metabolites is defined as:

$$p = 1-(1-\alpha)^m \quad (3)$$

where $m$ is the number of hypotheses (metabolites) and $\alpha$ is typically defined as 0.05. Accordingly, a set of 10 metabolites becomes an insignificant $p = 0.401$ even though each individual metabolite is statistically significant based on a pairwise Student's t-test with a $p < 0.05$.

29. A heat-map displaying all of the replicates from each group is preferred to only a group-average plot. Specifically, the hierarchical clustering of each replicate is indicative of the relative group separation and provides further confirmation of an observed group separation from a PCA, PLS or OPLS scores plot.

30. A valid PCA, PLS or OPLS model typically has $R^2$ values > $Q^2$ values, and $Q^2$ values > 0.4.
31. While p-values < 0.05 are typically acceptable, more often than not, high quality PLS/OPLS models from metabolomics data sets yield p-values << 0.001.
32. Chenomx maintains a series of 1D $^1$H NMR databases for a variety of NMR field strengths and sample pH. Use the database that matches the experimental conditions of the dataset being analyzed.
33. Most NMR metabolomics databases function in a similar manner to HMDB [28]. Simply upload a peak list with a set of chemical shift tolerances and obtain a list of potential matches.
34. The identification process is also available in open source software such as Mzmine [25] or web based tools such as MetaboAnalyst [26].
35. It is also possible to create or select your own search parameter. Click on *Edit* and select *Create New*. Select a database file in Structure Data Format (SDF) as input.
36. The possible compound assignments are based on an overall *score* determined by the *mass error*, *retention time error*, *isotope similarity*, *fragmentation score* and, if available, the *collision cross section*. The confidence of the identification may be increased by including *theoretical fragmentation* (*see* **Note 37**).
37. ChemSpider is a database comprised of 67 million compounds, and accordingly, is not restricted to known metabolites [27]. But, Progenesis can use the ChemSpider database for *in silico* prediction of fragmentation patterns. Progenesis cannot do this with HMDB [28].
38. **Section 2.3.18.3** sets a global threshold setting for all metabolites. Sometimes this may be too restrictive for specific metabolites, where a lower global threshold setting may cause a

large number of erroneous assignments. **Section 2.3.18.4** describes a manual approach to adjust the threshold settings for individual metabolites to recover incorrectly missed assignments while avoiding a high false assignment rate.

## 2.5. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Number (1660921). This work was supported in part by funding from the Redox Biology Center (P30 GM103335, NIGMS); and the Nebraska Center for Integrated Biomolecular Communication (P20 GM113126, NIGMS). The research was performed in facilities renovated with support from the National Institutes of Health (RR015468-01). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

**2. 6 References**

1. Gardner, S.G., et al., Metabolic Mitigation of Staphylococcus aureus Vancomycin Intermediate-Level Susceptibility. Antimicrob Agents Chemother, 2018. 62(1).

2. Anandhan, A., et al., Metabolic Dysfunction in Parkinson's Disease: Bioenergetics, Redox Homeostasis and Central Carbon Metabolism. Brain Res Bull, 2017. 133: p. 12-30.

3. Camandola, S. and M.P. Mattson, Brain metabolism in health, aging, and neurodegeneration. EMBO J, 2017. 36(11): p. 1474-1492.

4. Powers, R., et al., Metabolic Investigations of the Molecular Mechanisms Associated with Parkinson's Disease. Metabolites, 2017. 7(2).

5. Gebregiworgis, T. and R. Powers, Application of NMR metabolomics to search for human disease biomarkers. Comb Chem High Throughput Screen, 2012. 15(8): p. 595-610.

6. Halouska, S., et al., Revisiting Protocols for the NMR Analysis of Bacterial Metabolomes. J Integr OMICS, 2013. 3(2): p. 120-137.

7. Powers, R., The current state of drug discovery and a potential role for NMR metabolomics. J Med Chem, 2014. 57(14): p. 5860-70.

8. Botas, A., et al., Metabolomics of neurodegenerative diseases. Int Rev Neurobiol, 2015. 122: p. 53-80.

9. Han, W., et al., Profiling novel metabolic biomarkers for Parkinson's disease using in-depth metabolomic analysis. Mov Disord, 2017. 32(12): p. 1720-1728.

10. Luan, H., et al., Comprehensive urinary metabolomic profiling and identification of potential noninvasive marker for idiopathic Parkinson's disease. Sci Rep, 2015. 5: p. 13888.

11. Roede, J.R., et al., Serum metabolomics of slow vs. rapid motor progression Parkinson's disease: a pilot study. PLoS One, 2013. 8(10): p. e77629.

12. Poliquin, P.O., et al., Metabolomics and in-silico analysis reveal critical energy deregulations in animal models of Parkinson's disease. PLoS One, 2013. 8(7): p. e69146.

13. Chen, X., et al., Longitudinal Metabolomics Profiling of Parkinson's Disease-Related alpha-Synuclein A53T Transgenic Mice. PLoS One, 2015. 10(8): p. e0136612.

14. Lewitt, P.A., et al., 3-hydroxykynurenine and other Parkinson's disease biomarkers discovered by metabolomic analysis. Mov Disord, 2013. 28(12): p. 1653-60.

15. Lei, S., et al., Alterations in energy/redox metabolism induced by mitochondrial and environmental toxins: a specific role for glucose-6-phosphate-dehydrogenase and the pentose phosphate pathway in paraquat toxicity. ACS Chem Biol, 2014. 9(9): p. 2032-48.

16. Havelund, J.F., et al., Changes in kynurenine pathway metabolism in Parkinson patients with L-DOPA-induced dyskinesia. J Neurochem, 2017. 142(5): p. 756-766.

17. Breier, M., et al., Targeted metabolomics identifies reliable and stable metabolites in human serum and plasma samples. PLoS One, 2014. 9(2): p. e89728.

18. Worley, B. and R. Powers, Multivariate Analysis in Metabolomics. Curr Metabolomics, 2013. 1(1): p. 92-107.

19. Dettmer, K., A. Aronov Pavel, and D. Hammock Bruce, Mass spectrometry-based metabolomics. Mass Spectrometry Reviews, 2006. 26(1): p. 51-78.

20. Nicholson, J.K., J.C. Lindon, and E. Holmes, 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica, 1999. 29(11): p. 1181-9.

21. Worley, B. and R. Powers, MVAPACK: a complete data handling package for NMR metabolomics. ACS Chem Biol, 2014. 9(5): p. 1138-44.

22. Worley, B., S. Halouska, and R. Powers, Utilities for quantifying separation in PCA/PLS-DA scores plots. Anal Biochem, 2013. 433(2): p. 102-4.

23. Delaglio, F., et al., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR, 1995. 6(3): p. 277-93.

24. Johnson, B.A., Using NMRView to Visualize and Analyze the NMR Spectra of Macromolecules, in Protein NMR Techniques, A.K. Downing, Editor. 2004, Humana Press: Totowa, NJ. p. 313-352.

25. Pluskal, T., et al., MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics, 2010. 11(1): p. 395.

26. Xia, J., et al., MetaboAnalyst 3.0-making metabolomics more meaningful. Nucleic Acids Res., 2015. 43(W1): p. W251-W257.

27. Pence, H.E. and A. Williams, ChemSpider: An Online Chemical Information Resource. Journal of Chemical Education, 2010. 87(11): p. 1123-1124.

28. Wishart, D.S., et al., HMDB 3.0-The Human Metabolome Database in 2013. Nucleic Acids Res., 2013. 41(D1): p. D801-D807.

29. Markley, J.L., et al. New bioinformatics resources for metabolomics. 2007. World Scientific Publishing Co. Pte. Ltd.

30. Worley, B. and R. Powers, Deterministic multidimensional nonuniform gap sampling. J. Magn. Reson., 2015. 261: p. 19-26.

31. Poewe, W., et al., Parkinson disease. Nat Rev Dis Primers, 2017. 3: p. 17013.

32. Bras, J., R. Guerreiro, and J. Hardy, SnapShot: Genetics of Parkinson's disease. Cell, 2015. 160(3): p. 570-570 e1.

33. Klein, C. and A. Westenberger, Genetics of Parkinson's disease. Cold Spring Harb Perspect Med, 2012. 2(1): p. a008888.

34. Goldman, S.M., Environmental toxins and Parkinson's disease. Annu Rev Pharmacol Toxicol, 2014. 54: p. 141-64.

35. Cannon, J.R. and J.T. Greenamyre, Gene-environment interactions in Parkinson's disease: specific evidence in humans and mammalian models. Neurobiol Dis, 2013. 57: p. 38-46.

36. Franco, R., et al., Molecular mechanisms of pesticide-induced neurotoxicity: Relevance to Parkinson's disease. Chem Biol Interact, 2010. 188(2): p. 289-300.

37. Falkenburger, B.H., T. Saridaki, and E. Dinter, Cellular models for Parkinson's disease. J Neurochem, 2016. 139 Suppl 1: p. 121-130.

38. Creed, R.B. and M.S. Goldberg, New Developments in Genetic rat models of Parkinson's Disease. Mov Disord, 2018. 33(5): p. 717-729.

39. Mosley, R.L., et al., Inflammation and adaptive immunity in Parkinson's disease. Cold Spring Harb Perspect Med, 2012. 2(1): p. a009381.

40. Ascherio, A. and M.A. Schwarzschild, The epidemiology of Parkinson's disease: risk factors and prevention. Lancet Neurol, 2016. 15(12): p. 1257-1272.

41. Anandhan, A., et al., Glucose Metabolism and AMPK Signaling Regulate Dopaminergic Cell Death Induced by Gene (alpha-Synuclein)-Environment (Paraquat) Interactions. Mol Neurobiol, 2017. 54(5): p. 3825-3842.

42. Geraghty, R.J., et al., Guidelines for the use of cell lines in biomedical research. Br J Cancer, 2014. 111(6): p. 1021-46.

43. Westerhoff, H.V. and Y.D. Chen, How do enzyme activities control metabolite concentrations? An additional theorem in the theory of metabolic control. Eur J Biochem, 1984. 142(2): p. 425-30.

44. Blesa, J. and S. Przedborski, Parkinson's disease: animal models and dopaminergic cell vulnerability. Front Neuroanat, 2014. 8: p. 155.

45. Siegel, M.M., The use of the modified simplex method for automatic phase correction in fourier-transform nuclear magnetic resonance spectroscopy. Analytica Chimica Acta, 1981. 133(1): p. 103-108.

46. Worley, B. and R. Powers, Simultaneous Phase and Scatter Correction for NMR Datasets. Chemometr Intell Lab Syst, 2014. 131: p. 1-6.

47. Savorani, F., G. Tomasi, and S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. J Magn Reson, 2010. 202(2): p. 190-202.

48. De Meyer, T., et al., NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. Anal Chem, 2008. 80(10): p. 3783-90.

49. Development Core Team, R., R: A Language and Environment for Statistical Computing. Vol. 1. 2011.

50. Eriksson, L., J. Trygg, and S. Wold, CV-ANOVA for significance testing of PLS and OPLS® models. Journal of Chemometrics, 2008. 22(11-12): p. 594-600.

51. Triba, M.N., et al., PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters. Mol Biosyst, 2015. 11(1): p. 13-9.

52. Goodacre, R., et al., Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics, 2007. 3(3): p. 231-241.

53. Benjamini, Y. and Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 1995. 57(1): p. 289-300.

54. Bland, J.M. and D.G. Altman, Multiple significance tests: the Bonferroni method. BMJ, 1995. 310(6973): p. 170.

55. Xu, Q.-S. and Y.-Z. Liang, Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems, 2001. 56(1): p. 1-11.

56. Rubinson, K.A., Practical corrections for p(H,D) measurements in mixed H2O/D2O biological buffers. Anal Methods, 2017. 9(18): p. 2744-2750.

57. Ruxton, G.D., The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology, 2006. 17(4): p. 688-690.

58. Fay, M.P. and M.A. Proschan, Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Statistics surveys, 2010. 4: p. 1-39.

59. Hotelling, H., The Economics of Exhaustible Resources. Journal of Political Economy, 1931. 39(2): p. 137-175.

60. Fay, D.S. and K. Gerow A biologist's guide to statistical thinking and analysis. WormBook : the online review of C. elegans biology, 2013. 1-54 DOI: 10.1895/wormbook.1.159.1.

61. Marshall, D.D., et al., Combining DI-ESI–MS and NMR datasets for metabolic profiling. Metabolomics, 2015. 11(2): p. 391-402.

# CHAPTER 3

## 3. Arsenic and Neurodevelopmental Disorders

### 3.1 Heavy Metal and Arsenic Toxicity, an Environmental Danger

Heavy metals are naturally occurring elements with high atomic weights and densities five times heavier than water. They are typically metals and metalloids such as mercury, lead, chromium, and arsenic [1]. Industrial wastewater and emissions often lead to accumulation of heavy metals in soil [2]. In addition to environmental exposure, use of products containing heavy medals lead to human exposure. Pesticides and fertilizers often contain trace levels of heavy metals including iron, cadmium, and cobalt [3]. Consumer products such as cosmetics may contain trace levels of heavy metals including iron and chromium, which may be absorbed through the skin [4]. Heavy metals are highly toxic, and many are known carcinogens. The threat level posed by a given heavy metal is highly dependent on the level and nature of the exposure.

One of the higher risk heavy metals to humans is arsenic. Arsenic is a common metalloid typically exposed to humans through contaminated water supplies [5]. Production of tube wells in Bangladesh with arsenic contamination put between 35 to 77 million people at risk [6]. This contaminated water may even be used to water crops, resulting in arsenic build up in the food supply. In south east Asia, arsenic contaminated wells led to arsenic build up in rice [7]. In the United States, arsenic exposure is typically through food and water. Roughly 2.1 million Americans use water that contains 10 μg/L of arsenic [8].

Arsenic is a known carcinogen, and has been associated with skin, lung, and bladder cancer [9]. Unlike other cancers, epideological studies rather than animal models were used to establish a dose-response relation. Specifically, wells in Taiwan with high levels of arsenic were used to establish cancer rates [10]. After establishing arsenic as a carcinogen, the exposure limit was

lowered to 10 μg/L from 50 μg/L [11]. However, the risks of lower levels of exposure are difficult to measure.

Neurological disorders are often associated with low levels of exposure often through water and food. Neurological symptoms are typically observed in children or pregnant women. Studies have linked high levels of arsenic exposure to hindered learning development. A study on school children observed arsenic levels in urine at 50 μg/L was corrolated with lower scores in language and memory tests [12]. Another study measured arsenic in the blood of children. Children found with arsenic levels of 147 μg/L scored lower in cognative behavior tests [13]. Children exposed through arsenic in drinking water at levels of 106 to 142 mg/L had lower IQ scores [14]. Nevertheless, these studies have raised concerns and suggest arsenic exposures levels, particularly for products intended to be used by children and pregnant women, should be reevaluated. Establishing a lower limit of arsenic exposure is difficult since epidemiological studies have not yielded consistant results [15]. Additionally, epidemiological studies do not provide a dose-response relationship between arsenic exposure and neurological effects. Further complicating the situation is the fact that other nutritional defects, or additional neurotoxins may contribute to the neurological disorders attributed to arsenic [16]. Recently, the FDA proposed a standard of 10 ppb of arsenic in apple juice; however, more research is needed before a safe level of exposure can be determined [17].

### 3.2 Metal Xenobiotics and Metabolism

While Epidemiological data is not sufficient to establish a safe limit, understanding the mechanism of toxicity may provide more insights in how arsenic disrupts neurological functions. The toxic effect of arsenic is highly dependent on the dosage and route of exposure. Some heavy metals like

aluminum can be removed from the body while others accumulate and lead to chronic health effects. Heavy metals may act as pseudo elements in the body and interfere with metabolomic processes, generate free radicals, and produce oxidative stress. [18].

Arsenic is known to alter protein and some enzyme function due to a high affinity to sulfhydryl groups and can bind to reduced cysteines [19]. Enzymes that contain hydroxyl and thiol groups, such as pyruvate dehydrogenase (PDH), can be deactivated through arsenic binding [20]. PDH is part of the pyruvate dehydrogenase complex (PDC), which plays an important role by controlling the rate of pyruvate entry into the tricarboxylic acid cycle (TCA) by converting it into acetyl coenzyme A (Acetyl-CoA). PDC is the limiting step in the TCA cycle by funneling pyruvate into the TCA cycle. Arsenic is shown to substitute for the phosphate in ATP, which can potentially lead to a disruption to energy metabolism [21]. These effects, once built up, can lead to cell death or metabolic dysfunction. Energy metabolism is very important in the brain and its dysfunction has been linked to neuronal death [22]. Energy metabolism dysfunction is commonly observed in disorders in the nervous system including psychomotor retardation [23].

### 3.3 Astrocytes and the Brain

In the case of arsenic, it is believed that the developing brain is more susceptible to arsenic-induced toxic damage [24]. To that end, it is important to consider how heavy metals disrupt neuron function or lead to cell death, especially during the developmental stage [25]. Brain development includes prenatal to early childhood. The formation of neurons, axons, and dendrites occur during brain development. It also includes building and pruning of synapses located between neurons [26]. These synapses are key to learning and memory formation. Increase and decrease in synapse strength is critical to memory storage [27].

An increasing focus on glial cells and their role in neurological disorders has emerged [28]. The glial cell astrocytes play an important role in forming and maintaining synapses [29]. Astrocytes are also located between synapses and maintain and regulate synapse function [30]. Astrocytes help maintain synapses by regulating levels of metabolites present in the synapse. They help maintain the synapses between neurons including the removal of excess neurotransmitters, such as glutamate and GABA [31, 32]. Glutamate is taken in by astrocytes and metabolically converted to glutamine through glutamine synthetase. Glutamine is than shuttled to neurons, which is converted back to glutamate in a glutamate/glutamine cycle [33]. Astrocytes also help regulate nutrients that enter the brain. Astrocyte end feet cover 90% of the blood vesicles in the brain and assist with the blood brain barrier (BBB) function [34]. These blood vesicles form the BBB, which is very selective in allowing what crosses into the brain [35].

Astrocytes metabolically process nutrients from the BBB and shuttle them to neurons [36]. Glucose metabolism is of interest in astrocytes. Astrocytes consume glucose glycolytically, blocking glycolysis had minimal effect on ATP production [37]. Astrocytes take in glucose and produce pyruvate, which is pooled into cytosol pyruvate, which is converted to lactate and mitochondrial pyruvate [38]. Astrocytes can also glycolytically produce glycogen and convert it back to lactate for later use [39, 40].

Mitochondrial pyruvate is also funneled into the TCA cycle. Neurons cannot upregulate glycolysis due to a lack of 6-phosphotrocto-2-kinase/fructose 2,6-bisphosphatse isoform 3 enzymes [41]. Both neurons and astrocytes express PDC to generate ATP from glucose. While neurons express PDC at a near maximum capacity, astrocytes tightly regulate PDC [42]. Astrocytes produce and export a number of TCA intermediates, which get taken up by neurons [43]. $^{13}C$ glucose has been used to show that TCA products, including citrate, is funneled to neurons [44].

Due to astrocytes' relationship to the BBB, it is one of the first cell types to encounter and respond to foreign agents. Arsenic has been shown to cross the BBB and the placenta [45]. This suggests that astrocytes are the first neurological cells to encounter xenobiotics like arsenic. Rodent models have shown that arsenic can accumulate in the brain, particularly in the pituitary, hippocampus, thalamus and hypothalamus [46].

Astrocytes have high concentrations of antioxidants, including glutathione and vitamin C, making them highly resistant to oxidative stress induced by xenobiotics [47]. GSH has also been shown to catalyze the reduction of peroxides and to form complexes with xenobiotics [48]. GSH is also used by arsenic (III) methyltransferase (AS3MT), which methylates arsenic in the brain [46]. In addition, GSH has been shown to form complexes with arsenic as part of the arsenic excretion process. However, in addition to supporting cells, astrocytes have been known to induce toxic effects. Astrocytes have been shown to be reactive, including regulating and inducing inflammation, and can induce cell damage as well as cell repair [49].

Cultured astrocytes have been shown to alter their metabolism when exposed to arsenic notably in the production of GSH [51]. Glucose metabolism is important in the production of GSH as well as the formation of many biomolecules. When treated with 3 mM of arsenic for 2 hours there was a decrease in internal GSH and an increase in external GSH [52]. To observe how arsenic altered the glucose metabolism, astrocytes were feed $^{13}C$ labeled glucose. Levels of metabolites derived from glucose were measured to identify metabolomic pathways that were altered by arsenic treatment.

## 3.4 Method and Materials

### 3.4.1 Chemicals and Reagents.

*Astrocyte Cell Cultures.*

All cell culture work was done by Jordan Rose from Dr. Rodrigo Franco Cruz's lab.

Primary astrocytes were cultured in 100 mm dishes for NMR analysis and 6 well plates for flow cytometric analysis and cell-based assays. Primary astrocytes were collected from mouse pups, stored in a cryofreezer, and were thawed on ice before use. Cells were grown in Dulbecco's Modified Nutrient Mixture F-12 (DMEM) media with 10% Fetal Bovine Serum (FBS) and 1% penicillin-streptomycin. The cells were incubated at 37$^{o}$C and 5% $CO_2$ and the media was changed every 2-3 days until cells reached 90% confluence. Once the cells reached confluence, the media was replaced with the treatment media. Treatment media consisted of DMEM with the measured-out arsenic dosage as well as replacing $^{12}$C-glucose with $^{13}C_6$-glucose for $^{13}$C-labeling of the metabolome. The cells were treated for 12 hours followed by extraction of the metabolome.

### 3.4.2 Preparation of Metabolomics Samples for NMR Analysis.

Prior to cellular extraction, 1 mL of media was collected of and used to measure extracellular metabolites. Extracellular metabolites extracted by centrifuging the collected media for 5 minutes at 15,000 g at 4°C and collecting the supernatant. Intracellular metabolites were collected from the cell lysate. The cells were first washed twice with 5 ml of phosphate buffer. 1 mL of methanol at -80°C was used to lyse and quench the cells. Cells were then incubated for 15 minutes at -80°C to facilitate the lysis. Cells were then detached with a cell scraper and confirmed with an inverted microscope. The process was repeated if cells remained attached. The methanol and cell debris were collected in 2 mL microcentrifugation tubes and centrifuged for 5 minutes at 15,000 g at 4 $^{o}$C to pellet the cell debris. The methanol supernatant was collected. The cell debris was extracted with a 80%/20% mixture of methanol/water, centrifuged for 5 minutes at 15,000 g at 4$^{o}$C, and the

supernatant collected. The process was repeated with 100% water. The three supernatants were combined, evaporated in a RotoSpeed vacuum to remove the methanol. and then lyophilized to dryness. The samples were reconstituted for NMR analysis with the addition of 500 uL of 50 mM phosphate buffer in $D_2O$ at pH 7.2 (uncorrected) with 500 μM sodium-3-trimethylsilylpropionate (TMSP) used as an internal chemical shift standard.

### 3.4.3 NMR Data Collection and Processing.

NMR spectra was collected on a Bruker AVANCE III-HD 700 MHz with a 5 mm quadrupole resonance QCI-P cryoprobe ($^{1}H$, $^{13}C$,$^{15}N$ and $^{31}P$) with z-axis gradients. Samples were collected at 300K using a SampleJet automated sample changer and Bruker ICON-NMR software to automate data collection. The 1D $^{1}H$ experiments were collected with 128 scans and 4 dummy scans. There were 32,768 data points collected with a spectral width of 11,160 Hz with a 2s relaxation delay. 2D $^{1}H$-$^{13}C$ HSQC spectra were collected with 128 scans and 16 dummy scans. In the direct dimension, there were 1024 points and a spectral width of 9,090 Hz. In the indirect dimension, there were 128 points and a spectral width of 29,165 Hz. NMR data was Fourier transformed, auto phased, and referenced to TMSP with NMRpipe. NMRViewJ Version 98 was used to peak pick and to quantify peak changes. The NMR data sets was normalized to the total peak intensity.

### 3.4.4 Metabolite identification.

Chemical shifts were assigned using Platform for RIKEN Metabolomics (PRIMe) (http://prime.psc.riken.jp/) with a 0.05 ppm and 0.1 ppm error range for $^{1}H$ and $^{13}C$ chemical shifts, respectively. Metabolite identity was manually confirmed with the Human Metabolome Database (HMDB) (https://hmdb.ca/) [54]. Metabolite concentrations were measured from peak intensities.

All metabolite concentrations were normalized to TMSP (500 μM). For metabolites with multiple peaks, intensities were averaged after normalization.

### 3.4.5 Statistical Analysis and Data Processing.

To measure the effect of arsenic, metabolite levels of treated samples were compared to untreated control samples. Fold changes were calculated between treated and untreated samples. Student's t-test was used to verify statistical significance followed by a Benjamin-Hochberg multiple hypothesis correction to account for false discovery rate [53]. A corrected p-value $< 0.05$ was considered statistically significant.

## 3.5 Results and Discussion

Astrocytes respond to arsenic in a time and concentration dependent fashion. Astrocytes exposed to arsenic have been shown to have decreased viability. Exposure of cultured cells at 1 mM for 24 hours resulted in decreased cell viability [55].

To observe the effect arsenic on glucose metabolites, intermediate and product metabolites were measured by NMR. 13C Glucose was fed to astrocytes and metabolite products were quantified in both the intracellular and extra cellular space. Figure 3.1 is an example HSQC and shows some of the metabolites identified by chemical shifts. We looked for changes which upon treatment showed consistent increases or decrease across replicates. Figure 3.2 shows fold change in metabolites. Upon treatment with arsenic, there was an observed increase in glycolysis intermediates, fructose-6-phosphate at and 3-Phosphate-Glycate. However, there was a decrease in glycolytic products lactate and glycogen precursor Uridine diphosphate glucose (UDP-glucose). Figure 3.2 shows the fold changes for metabolites identified both intracellular and extra cellular.

This suggests that glycolysis is upregulated but funneled into pyruvate production rather than lactate production.

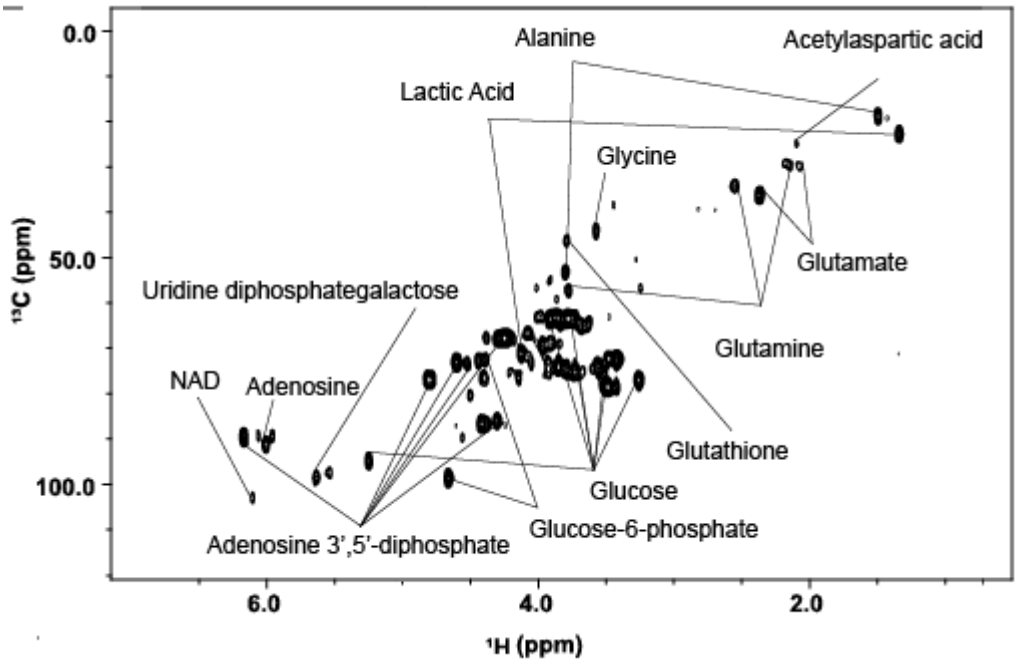



**Figure 3.1**: 2D $^{1}H$-$^{13}C$ HSQC spectra from astrocytes cell extract following treatment with 10 μ of $As_2O_3$. Highlighted are some of the metabolites identified based on chemical shifts.

Decreased lactate production is observed when astrocytes experience oxidative stress. It has been observed that peroxide induced oxidative stress exposure decreased lactate production. Measuring $^{13}C$ enrichment levels show that it was not a shift in carbon sources but a decrease in overall production [56]. At low levels of arsenic exposure, astrocytes with 0.1 mM and 0.3 mM arsenic stimulated GSH export and glycolic flux resulted in excess lactate production [51]. However, we observed it is likely that astrocytes divert glucose from glycolytic lactate production to pyruvate production for use in the TCA cycle at high levels of arsenic exposure. Lactate has been theorized to be funneled into the TCA cycle of neurons as a carbon source [43]. However, lactate also

modulates receptors and channels in neurons [57].

TCA cycle intermediates are difficult to measure after long treatments due to the cyclic nature of metabolic processes. Instead, the focus was shifted to the products. Figure 3.2 shows metabolites with significant fold change in both the intercellular and extracellular space. We observed a major increase in glutamate (2.33, 36.35 ppm) in the extracellular and a matching decrease in the intercellular concentration. We also observed a decrease in extracellular metabolites lactate and citrate at high doses. Figure 3.3 shows where these metabolites are in the glycolic and citric acid cycle. Increases in the export of glutamate due to arsenic exposure has been previously observed [58, 59].

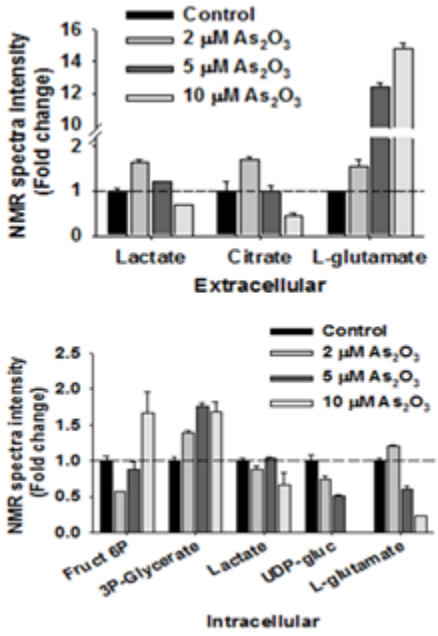



**Figure 3.2**: Relative fold change of metabolites detected in 2D $^{1}H$-$^{13}C$ HSQC spectra. Fold changes represent metabolite levels compared to an untreated control. Error bar represents standard deviation. Astrocytes cells were treated with 0 (black), 2, (grey) 5 (dark grey) and 10 (white) μM of $As_2O_3$. All detected metabolites are derived from $^{13}C_6$-glucose.

After being converted into pyruvate, glucose was funneled into the TCA cycle. The intermediates

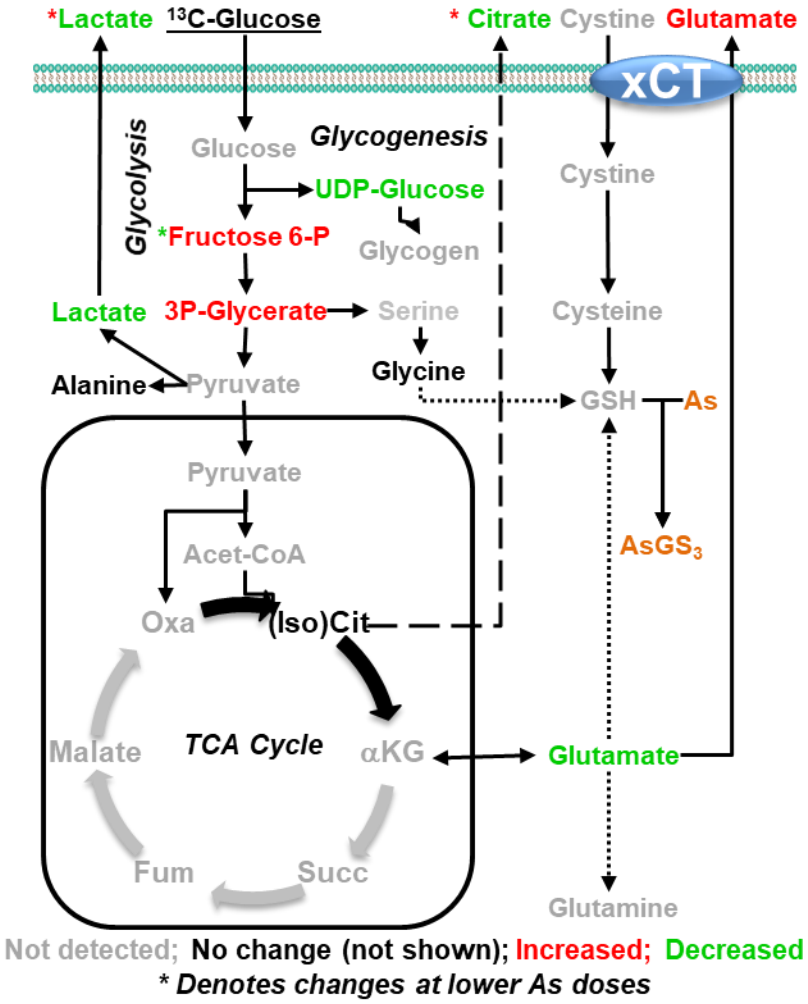



**Figure 3.3**: Summary of change in metabolism of $^{13}C$ glucose in astrocytes following arsenic treatment. The metabolic pathway shows glucose metabolized through glycolysis and the TCA cycle. Gray colored metabolites were not detected. Black colored metabolites were identified but had no significant change. Green colored metabolites had a significant decrease relative to untreated astrocytes. Red colored metabolites had a significant increase relative to untreated astrocytes.

Treating astrocytes with 1.5 to 30 μM of arsenic resulted in decrease in the expression of glutamate transporters in astrocytes, which resulted in a decrease in glutamate uptake [60]. Glutamate is a key component in GSH. It has been previously shown that GSH catalyzes the condensation of ammonia and glutamate, which reduces toxic levels of glutamate and ammonia [61]. Astrocytes exposed to hydrogen peroxide has been shown to reduce glutamine levels, suggesting a disruption

in GSH [62]. Glutamate export is also important to the import of cystine through the xCT transport [63]. Studies have shown that the production of GSH results in the consumption of cystine [64]. Glutamate export is a response to the import of cystine into the cell to upregulate GSH production. This may result in neuron death due to an inability to produce GSH, leading to oxidative stress [65].

A decrease in citrate was also observed at a high level of arsenic exposure. A decrease in citrate production has been observed with aluminum, although this was likely due to the binding to aluminum and prevention of transport [66]. Citrate is a TCA intermediate and could act as a possible carbon source for ATP production in neurons. Citrate has been speculated to chelate calcium and magnesium, and modulate glutamate receptors [67].

### 3.6 Conclusion

Astrocytes being a one of the larges types of glial cells play a large and diverse role in the brain. They have been shown to metabolically support neurons, including maintaining homeostasis by metabolically producing molecules that support neurons. However, they have also been shown to have negative impacts on neurons. Astrocytes have been known to produce inflammatory products such as cytokines in response to stressors [68]. There is also the consideration that exposure to toxins can hijack glucose metabolism to shift focus to protective GSH production.

Overall, we observed a decrease in glucose products specifically lactate and citrate as well as an increase in glutamate. It is possible that the production of GSH alters carbon metabolism away from energy production towards antioxidant production. Decreased TCA intermediates could deprive neurons of a carbon source. Additionally, increased consumption of cystine could reduce the pool available to neurons resulting in oxidative stress [65]. However, it is important to note

that these metabolites also function to modulate receptors. Lactate has been shown to bind to synaptic receptors connected to synaptic plasticity [57]. Lactate has suggested to be a key metabolite in memory formation due to high brain energy demands [69]. Westergaard suggested that citrate plays a role in NMDA glutamate receptor by chelation of $Zn^{2+}$ [67]. Many of these effects could lead to dysfunction or cellular death of neurons.

**3.7 References**

1. Fergusson, J.E., The heavy elements: chemistry, environmental impact and health effects\Jack E. Fergusson. 1990.

2. Wuana, R.A. and F.E. Okieimen, Heavy Metals in Contaminated Soils: A Review of Sources, Chemistry, Risks and Best Available Strategies for Remediation. ISRN Ecology, 2011. 2011: p. 402647.

3. Gimeno-Garcia, E., V. Andreu, and R. Boluda, Heavy metals incidence in the application of inorganic fertilizers and pesticides to rice farming soils. Environ Pollut, 1996. 92(1): p. 19-25.

4. Borowska, S. and M.M. Brzoska, Metals in cosmetics: implications for human health. J Appl Toxicol, 2015. 35(6): p. 551-72.

5. Chung, J.Y., S.D. Yu, and Y.S. Hong, Environmental source of arsenic exposure. J Prev Med Public Health, 2014. 47(5): p. 253-7.

6. Smith, A.H., E.O. Lingas, and M. Rahman, Contamination of drinking-water by arsenic in Bangladesh: a public health emergency. Bull World Health Organ, 2000. 78(9): p. 1093-103.

7. Meharg, A.A., Arsenic in rice--understanding a new disaster for South-East Asia. Trends Plant Sci, 2004. 9(9): p. 415-7.

8. Ayotte, J.D., et al., Estimating the High-Arsenic Domestic-Well Population in the Conterminous United States. Environmental Science & Technology, 2017. 51(21): p. 12443-12454.

9. Hong, Y.S., K.H. Song, and J.Y. Chung, Health effects of chronic arsenic exposure. J Prev Med Public Health, 2014. 47(5): p. 245-52.

10. Smith, A.H., et al., Cancer risks from arsenic in drinking water. Environ Health Perspect, 1992. 97: p. 259-67.

11. EPA. Arsenic: Rule-Making History. 2012 [cited 2019 June 15]; Available from:

http://water.epa.gov/lawsregs/rulesregs/sdwa/arsenic/history.cfm.

12. Rosado, J.L., et al., Arsenic exposure and cognitive performance in Mexican schoolchildren. Environ Health Perspect, 2007. 115(9): p. 1371-5.

13. Roy, A., et al., Association between arsenic exposure and behavior among first-graders from Torreon, Mexico. Environ Res, 2011. 111(5): p. 670-6.

14. Wang, S.X., et al., Arsenic and fluoride exposure in drinking water: children's IQ and growth in Shanyin county, Shanxi province, China. Environ Health Perspect, 2007. 115(4): p. 643-7.

15. Tyler, C.R. and A.M. Allan, The Effects of Arsenic Exposure on Neurological and Cognitive Dysfunction in Human and Rodent Studies: A Review. Curr Environ Health Rep, 2014. 1: p. 132-147.

16. Schmidt Charles, W., Low-Dose Arsenic: In Search of a Risk Threshold. Environmental Health Perspectives, 2014. 122(5): p. A130-A134.

17. Tsuji, J.S., et al., Low-level arsenic exposure and developmental neurotoxicity in children: A systematic review and risk assessment. Toxicology, 2015. 337: p. 91-107.

18. Jaishankar, M., et al., Toxicity, mechanism and health effects of some heavy metals. Interdiscip Toxicol, 2014. 7(2): p. 60-72.

19. Shen, S., et al., Arsenic binding to proteins. Chem Rev, 2013. 113(10): p. 7769-92.

20. M. Schiller, C., B. A. Fowler, and J. Woods, Effects of Arsenic on Pyruvate Dehydrogenase Activation. Vol. 19. 1977. 205-7.

21. Tseng, C.H., The potential biological mechanisms of arsenic-induced diabetes mellitus. Toxicol Appl Pharmacol, 2004. 197(2): p. 67-83.

22. Pathak, D., A. Berthet, and K. Nakamura, Energy failure: does it contribute to

neurodegeneration? Ann Neurol, 2013. 74(4): p. 506-16.

23. Blass, J.P., R.K. Sheu, and J.M. Cedarbaum, Energy metabolism in disorders of the nervous system. Rev Neurol (Paris), 1988. 144(10): p. 543-63.

24. Grandjean, P. and K.T. Herz, Trace elements as paradigms of developmental neurotoxicants: Lead, methylmercury and arsenic. Journal of Trace Elements in Medicine and Biology, 2015. 31: p. 130-134.

25. Chen, P., M.R. Miah, and M. Aschner, Metals and Neurodegeneration. F1000Res, 2016.

26. Tierney, A.L. and C.A. Nelson, 3rd, Brain Development and the Role of Experience in the Early Years. Zero Three, 2009. 30(2): p. 9-13.

27. Mayford, M., S.A. Siegelbaum, and E.R. Kandel, Synapses and memory storage. Cold Spring Harb Perspect Biol, 2012. 4(6).

28. Ndubaku, U. and M.E. de Bellard, Glial cells: old cells with new twists. Acta Histochem, 2008. 110(3): p. 182-95.

29. Chung, W.S., N.J. Allen, and C. Eroglu, Astrocytes Control Synapse Formation, Function, and Elimination. Cold Spring Harb Perspect Biol, 2015. 7(9): p. a020370.

30. Allen, N.J. and C. Eroglu, Cell Biology of Astrocyte-Synapse Interactions. Neuron, 2017. 96(3): p. 697-708.

31. Sofroniew, M.V. and H.V. Vinters, Astrocytes: biology and pathology. Acta Neuropathol, 2010. 119(1): p. 7-35.

32. Weber, B. and L.F. Barros, The Astrocyte: Powerhouse and Recycling Center. Cold Spring Harb Perspect Biol, 2015. 7(12).

33. Meldrum, B.S., Glutamate as a neurotransmitter in the brain: review of physiology and pathology. J Nutr, 2000. 130(4S Suppl): p. 1007S-15S.

34. Jukkola, P. and C. Gu, Regulation of neurovascular coupling in autoimmunity to water and ion channels. Autoimmun Rev, 2015. 14(3): p. 258-67.

35. Squire, L.R., Fundamental neuroscience. 3rd ed. ed. 2008, Amsterdam ; Boston: Elsevier/Academic Press.

36. Tsacopoulos, M. and P.J. Magistretti, Metabolic coupling between glia and neurons. J Neurosci, 1996. 16(3): p. 877-85.

37. Swanson, R.A. and J.H. Benington, Astrocyte glucose metabolism under normal and pathological conditions in vitro. Dev Neurosci, 1996. 18(5-6): p. 515-21.

38. Zwingmann, C., C. Richter-Landsberg, and D. Leibfritz, 13C isotopomer analysis of glucose and alanine metabolism reveals cytosolic pyruvate compartmentation as part of energy metabolism in astrocytes. Glia, 2001. 34(3): p. 200-12.

39. Brown, A.M. and B.R. Ransom, Astrocyte glycogen and brain energy metabolism. Glia, 2007. 55(12): p. 1263-71.

40. Matsui, T., et al., Astrocytic glycogen-derived lactate fuels the brain during exhaustive exercise to maintain endurance capacity. Proceedings of the National Academy of Sciences, 2017. 114(24): p. 6358-6363.

41. Bolanos, J.P., A. Almeida, and S. Moncada, Glycolysis: a bioenergetic or a survival pathway? Trends Biochem Sci, 2010. 35(3): p. 145-9.

42. Halim, N.D., et al., Phosphorylation status of pyruvate dehydrogenase distinguishes metabolic phenotypes of cultured rat brain astrocytes and neurons. Glia, 2010. 58(10): p. 1168-76.

43. Waagepetersen, H.S., et al., Comparison of Lactate and Glucose Metabolism in Cultured Neocortical Neurons and Astrocytes Using <sup>13</sup>C-NMR Spectroscopy. Developmental Neuroscience, 1998. 20(4-5): p. 310-320.

44. Sonnewald, U., et al., First direct demonstration of preferential release of citrate from astrocytes using [13C]NMR spectroscopy of cultured neurons and astrocytes. Neurosci Lett, 1991. 128(2): p. 235-9.

45. Ahmed, S., et al., Arsenic-Associated Oxidative Stress, Inflammation, and Immune Disruption in Human Placenta and Cord Blood. Environmental Health Perspectives, 2011. 119(2): p. 258-264.

46. Sanchez-Pena, L.C., et al., Arsenic species, AS3MT amount, and AS3MT gene expression in different brain regions of mouse exposed to arsenite. Environ Res, 2010. 110(5): p. 428-34.

47. Wilson, J.X., Antioxidant defense of the brain: a role for astrocytes. Can J Physiol Pharmacol, 1997. 75(10-11): p. 1149-63.

48. Dringen, R., et al., Glutathione-Dependent Detoxification Processes in Astrocytes. Neurochem Res, 2015. 40(12): p. 2570-82.

49. Gyurasics, Á., F. Varga, and Z. Gregus, Glutathione-dependent biliary excretion of arsenic. Biochemical Pharmacology, 1991. 42(3): p. 465-468.

50. Colombo, E. and C. Farina, Astrocytes: Key Regulators of Neuroinflammation. Trends Immunol, 2016. 37(9): p. 608-620.

51. Tadepalle, N., et al., Arsenite stimulates glutathione export and glycolytic flux in viable primary rat brain astrocytes. Neurochem Int, 2014. 76: p. 1-11.

52. Meyer, N., et al., Arsenate accumulation and arsenate-induced glutathione export in astrocyte-rich primary cultures. Neurochem Int, 2013. 62(7): p. 1012-9.

53. Benjamini, Y. and Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 1995. 57(1): p. 289-300.

54. Wishart, D.S., et al., HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res, 2018. 46(D1): p. D608-D617.

55. Koehler, Y., et al., Uptake and toxicity of arsenite and arsenate in cultured brain astrocytes. J Trace Elem Med Biol, 2014. 28(3): p. 328-37.

56. Liddell, J.R., et al., Sustained hydrogen peroxide stress decreases lactate production by cultured astrocytes. Journal of Neuroscience Research, 2009. 87(12): p. 2696-2708.

57. Goncalves, C.A., et al., Glycolysis-Derived Compounds From Astrocytes That Modulate Synaptic Communication. Front Neurosci, 2018. 12: p. 1035.

58. Frade, J., et al., Glutamate induces release of glutathione from cultured rat astrocytes – a possible neuroprotective mechanism? Journal of Neurochemistry, 2008. 105(4): p. 1144-1152.

59. Zhao, F., et al., Effects of arsenite on glutamate metabolism in primary cultured astrocytes. Toxicol In Vitro, 2012. 26(1): p. 24-31.

60. Castro-Coronel, Y., et al., Arsenite exposure downregulates EAAT1/GLAST transporter expression in glial cells. Toxicol Sci, 2011. 122(2): p. 539-50.

61. Rose, C.F., A. Verkhratsky, and V. Parpura, Astrocyte glutamine synthetase: pivotal in health and disease. Biochem Soc Trans, 2013. 41(6): p. 1518-24.

62. Brand, A., D. Leibfritz, and C. Richter-Landsberg, Oxidative stress-induced metabolic alterations in rat brain astrocytes studied by multinuclear NMR spectroscopy. Journal of Neuroscience Research, 1999. 58(4): p. 576-585.

63. Ramos-Chávez, L.A., et al., Neurological effects of inorganic arsenic exposure: altered cysteine/glutamate transport, NMDA expression and spatial memory impairment. Frontiers in Cellular Neuroscience, 2015. 9(21).

64. Singh, V., et al., Hijacking microglial glutathione by inorganic arsenic impels bystander

death of immature neurons through extracellular cystine/glutamate imbalance. Scientific Reports, 2016. 6(1): p. 30601.

65. Murphy, T.H., R.L. Schnaar, and J.T. Coyle, Immature cortical neurons are uniquely sensitive to glutamate toxicity by inhibition of cystine uptake. The FASEB Journal, 1990. 4(6): p. 1624-1633.

66. Meshitsuka, S. and D.A. Aremu, (13)C heteronuclear NMR studies of the interaction of cultured neurons and astrocytes and aluminum blockade of the preferential release of citrate from astrocytes. J Biol Inorg Chem, 2008. 13(2): p. 241-7.

67. Westergaard, N., et al., Citrate, a Ubiquitous Key Metabolite with Regulatory Function in the CNS. Neurochemical Research, 2017. 42(6): p. 1583-1588.

68. Sun, X., et al., Arsenic affects inflammatory cytokine expression in Gallus gallus brain tissues. BMC Vet Res, 2017. 13(1): p. 157.

69. Alberini, C.M., et al., Astrocyte glycogen and lactate: New insights into learning and memory mechanisms. Glia, 2018. 66(6): p. 1244-1262.

# Chapter 4

## 4. Geographical Analysis of Wine

### 4.1 Introduction to Wine Science

Since wines discovery, continuous efforts have been made to improve its quality, taste and production levels. Wine is a complex mixture consisting of 86% water, 12% glycerol and polysaccharides, 0.5% acids, and 0.5% volatile compounds [1]. This mixture is shaped by the complex process of wine making, from grape growth to fermentation. For example, phenolic compounds (hydroxy-substituted benzene rings) are developed in grapes and are largely responsible for the taste and color of wine. Each phenolic compound comes from a different part of the grape and plays a unique role in the overall quality of the wine. For example, flavonols are phenolic compounds found in the seed and skin of the grape, where production is stimulated by UV light exposure. Flavonols can increase the wine's color intensity and are highly correlated with the market price of wine. Anthocyanins are another class of phenolic compounds responsible for the color of red wine [2]. Yeast fermentations develops wine further by consuming grape products and converting them into new chemicals. There are two key steps to wine fermentation, alcoholic fermentation and malate fermentation, which may occur simultaneously or sequentially. Alcoholic fermentation converts sugars, mainly glucose and fructose, to ethanol and carbon dioxide. Malic fermentation, while not technically fermentation, converts malic acid to lactic acid [3]. Overall, the taste, smell and texture of wine, and correspondingly its value, is defined by the wine's chemical composition, which is impacted by the environment, climate and wine-making process. Accordingly, there is a long history and interest in correlating the chemical composition of wine with its quality and production.

In 1866, In the late 1970s, Shauils and Smart evaluated canopy management and found ways to increase the production of high quality grapes [4]. As scientific methods improved, wine science was better able to pinpoint the specific compounds that contribute to wine flavor. In the late 19$^{th}$ century, many of the phenolic compounds in wine were identified and the structures determined [5]. Thus, phenolic compounds can now be rapidly quantified from individual wine samples [6]. In this manner, the factors that contribute to high-quality wines may be assessed by studying the wine's phenolic profile. Furthermore, by identifying and quantifying the different compounds in a wine, the resulting chemical profile may be used to ensure wine quality. Alternatively, wine chemical profiles may be used to evaluate different yeast strains to identify which strains produced the desired result [7]. Phenolic compounds are also of interest to human health. Their antioxidant and anti-inflammatory capacities may present potential health benefits [8]. Chemical or phenolic profiles are frequently used to authenticate wines or identify its terrior [9]. Terroir defines all the factors or variables (*i.e.,* soil composition, sun exposure, rainfall, etc.) that impact how the wine was produced by a vineyard. Alternatively, terroir is a measure of the effect of the environment on the wine [10].

The profiling of the chemical composition of wine may be accomplished in several ways. Assay tests are frequently used to quantify a specific class of chemicals or just a few particular compounds. Assays allow for the rapid testing of many wine samples while providing accurate analysis. For example, an assay can monitor the fermentation process by specifically measuring ethanol production [11]. Other assays may be used to identify errors or problems with the fermentation process, such as observing the accumulation of acetic acid. The production of excess acetic acid leads to wine with an undesirable sour taste [12]. Similarly, the Fox-1 assay targets

tannins, which are often associated with wine astringency [12]. Of course, any assay limited to a single molecule will be restricted to what can be learned about the system.

As an alternative to targeted assays, analytical techniques can offer a broader coverage of wine metabolites. The simultaneous quantification and identification of the large number of compounds present in wine has an inherent and distinct value for characterizing authenticity [13]. Mass spectrometry (MS) coupled with liquid chromatography (LC-MS) or gas chromatography (GC-MS) may rapidly quantify the entire metabolic or chemical profile of wine. MS is a popular method for detecting any ionizable compound, which is further enabled by the availability of large reference databases for ready metabolite assignments. The exact mass, retention time and fragmentation patterns are used to identify metabolites by matching experimental values to database reference values. For example, LC-MS has been used to measure tannins in red wine to characterize its age [14]. NMR has also been effectively used to profile wines, grapes, and the products before, during and after fermentation. NMR detects $^{1}H$ and $^{13}C$ chemical shifts, which makes it useful for identifying and quantifying the organic compounds in wine. For example, NMR was used to measure the compounds in the pulp, skin, and the seeds from grapes to evaluate quality. The pulps were differentiated by levels of alanine and citrate [15]. NMR was also used to compare different strains of yeast, it was noted that yeasts that fermented wine faster produced higher levels of succinate and glycerol [16]. Port wines of different ages were compared by NMR and it was observed that aged wines had lower levels of succinate acid, pyruvic acid γ-butyric acid and proline [17].

Chemical or metabolic profiles are useful for evaluating wines, but sample variation is frequently a problem with commercial wines, Herein, we describe the chemical characterization of various

Pinot Noir (PN) wines produced by a number of California and Oregon vineyards. A chemical profile can be used to measure the environmental impact on PN wines and to differentiate different PN wines. A differential sensing assay that provides a phenolic profile was combined with untargeted one-dimensional (1D) $^{1}$H NMR data to compare PN wines from different vineyards, vintages, and wine regions. Wines were successfully identified based on vineyard of origin using a combination of univariate and multivariate statistical analysis.

## 4.2 Materials and Methods

### 4.2.1 Chemicals

Deuterium oxide (99.9% D) was obtained from Sigma Aldrich (Milwaukee, WI). 3-(trimethylsilyl) propionic-2,2,3,3-$D_4$ acid sodium salt (98% D) (TMSP-D4) was purchased from Cambridge Isotopes (Andover, MA). Potassium phosphate dibasic salt (anhydrous, 99.1% pure) and monobasic salt (crystal, 99.8% pure) were purchased from Fisher Scientific (Fair Lawn, NJ).

Wine grapes (*Vitis vinifera* L. cv. Pinot noir clone Dijon 667) from fifteen different vineyard sites along the West Coast of the United States were harvested at similar sugar concentration of 24 Brix between 13 Aug to 15 Sept 2015 and between 25 Aug to 21 Sept 2016 (Table 4.1). Eight American Viticultural Areas, which span a latitudinal distance of approximately 1450 km, are represented in this experiment: Santa Rita Hills (SRH), Santa Maria Valley (SMV), Arroyo Seco (AS), Carneros (CRN), Sonoma Coast (SNC), Russian River Valley (RRV), Anderson Valley (AV), and Willamette Valley (OR). Table 4.1 lists the selected vineyards, the wine region and the nearest county.

| County | Wine Region | Vineyard |
|---|---|---|
| Sonoma | Sonoma Coast | Annapolis |
| Sonoma | Sonoma Carneros | Cloud Landing |
| Sonoma | Sonoma RRV | Carneos Hills West |
| Sonoma | Sonoma RRV | Ross |
| Sonoma | Sonoma RRV | Bones |
| Sonoma | Sonoma RRV | Bloomfield |
| Mendocino | Anderson Valley | Boone Ridge |
| Mendocino | Anderson Valley | Maggy Hawk |
| Monterey | Arroyo | Panarama 5A |
| Monterey | Arroyo | MSA |
| Santa Barbara | Santa Maria Valley | Nielson |
| Santa Barbara | Santa Maria Valley | Rice/Cambria |
| Santa Barbara | Santa Maria Hills | Radian |
| Marin County | Willamette Valley | Gran Moraine |
| Marin County | Willamette Valley | Zenna West |

**Table 4.1: List of wines and vineyards.** The nearest county was used to group wine regions. Counties Sonoma, Mendocino, Monterey, and Santa Barbra are located on the cost of California while Marin County is in Oregon.

### 4.2.2 Winemaking

Grapes were fermented in 200 L stainless steel fermenters at the UC Davis Teaching & Research Winery. Primary fermentation was initiated by inoculating with Lalvin RC212 (Lallemand) after warming the must to 21°C. The fermentation temperature was held a 21°C for two days after inoculation, and subsequently allowed to rise to 27°C where it was held for the remainder of primary fermentation. Wine was separated from the red grape skins by using a basket press on the ninth day after grapes were placed into the fermenter.

### 4.2.3 Differential Sensing Method

The indicators Chrome Azurol S (CAS) (purity 65%), Bromopyrogagllol Red (BPR), and Pyrocatechol Violet (PCV) (purity 100%) were purchased from Sigma-Aldrich (Saint Louis, MO).

Nickel chloride hexahydrate (purity 99.7%), copper (II) sulfate (purity 99.2%), and HEPES buffer were purchased from Fisher Scientific (Hampton, NH). Solid phase peptide synthesis reagents were purchased from P3 BioSystems (Louisville, KY). Peptides were synthesized using standard SPPS and a CEM Liberty Blue Automated Microwave Synthesizer, (Matthews, NC, USA). Absorbance values were recorded using a Spectra Max Plus 384 plate reader (Molecular Device Inc.)

### 4.2.4 Array and Indicator Displacement Assay

A library of nine peptide-based sensors were used for the construction of the differential sensing (DS) array. Each ensemble was composed of a histidine peptide, a divalent metal and a colorimetric indicator. The peptides: WAHEDEFF (TT2), FHFPHHF (SEL1), and WEEHEE (RN8) were used to prepare the same peptide-metal-indicator ensembles (MM1-MM9) and corresponding binding ratios were previously reported [18,19]. Arrays were prepared in Fisher Scientific non-treated 96-well plates with flat bottom and clear polystyrene. Final well-plate solutions of peptide ensembles and wine concentration of 1% (v/v) were prepared using 50 mM HEPES in ethanol, 1:1 (v/v), pH = 7.4). Absorbance endpoint-values due to the displacement of each indicator by the tannins were measured at 430 nm, 444 nm, and 560 nm corresponding to the $\lambda_{max}$ of free CAS, PCV, and BPR, respectively. Eight replicates were performed to ensure reproducibility. Controls consisted of a column of wine alone and a column of the ensemble alone in each plate. Two experimental replicates of the full array were performed in 2017 using the wines from 2015 vintage and in 2018 using wines from the 2016 vintage, respectively.

### 4.2.5 NMR Sample Preparation.

Wine was removed for amber vials via a syringe immediately before preparing the NMR samples. Each NMR sample was prepared by adding 15 µL of 50 mM phosphate buffer prepared in $D_2O$ at pH 7.2 (uncorrected) with the addition of 50 µM of TMSP-D4 as an internal chemical shift standard to 150 µL of wine. Eight analytical replicates were prepared for each wine for a total of 120 NMR samples.

**4.2.6 NMR Data Collection and Processing**.

NMR spectra were collected on Bruker a AVANCE III 700 MHz spectrometer equipped with a 5 mm quadrupole resonance QCI-P cryoprobe™ ($^{1}H$, $^{13}C$, $^{15}N$ and $^{31}P$) with a Z-axis gradient. A SampleJet automated sample changer system with Bruker ICON-NMR™ software and an automatic tuning and matching accessory was used to automate the data collection.

A 1D $^{1}H$ NMR spectrum with a presaturation pulse and a NOESY pulse sequence was collected for each of the 120 wine samples. The Bruker automation program, Multisupp, was used to suppress the multiple solvent peaks resulting from the presence of water and ethanol in the sample. Multisupp automatically identifies and suppresses the most intense peaks in the NMR spectrum. For this experiment, one peak at 4.7 ppm due to water and the three peaks at 1.3, 2.7. and 3.7 ppm due to ethanol were suppressed. The 1D $^{1}H$ NMR spectra were collected at 300K with 65K points, a spectral width of 14705 Hz, 128 scans, 4 dummy scans, and 4s relaxation delay.

1D $^{1}H$ NMR spectra were batch processed and analyzed using our NMR metabolomics toolbox, MVAPACK [20]. The spectra were Fourier transformed, auto phased and referenced to TMSP at 0 ppm. Regions of the spectra containing residual water and TMSP were removed. The spectra were normalized using probabilistic quotient normalization and unit-variance scaled. Uniform bins

of the 1D $^{1}$H NMR dataset were exported as a matrix from MVAPACK.

### 4.3 Statistical Analysis.

#### 4.3.1 PCA analysis

PCA models were generated from the binned NMR data, the DS array data, and the combined NMR and DS array data set. PCA models were generated with MVAPACK [13] using 3 principal components. Dendrogram were created from the associated PCA scores plot with all wines separated by a Mahalanobis distances depicted as p-values [21,22]. A p-value $< 0.05$ was deemed to be statistically significant.

#### 4.3.2 ROC analyses

The combined NMR and DS array data set was processed in MetaboAnalyst 4.04 (https://www.metaboanalyst.ca/) to obtain multivariate ROC curves [23]. The dataset was range scaled and ROC curves generated using a support vector machine algorithm. A one versus all comparison was made for each individual wine.

### 4.4 Results

#### 4.4.1 NMR Data Collection for Wine Samples

To optimize data collection, multiple methods of NMR data collection and data processing were evaluated. Wine contains relatively high concentrations of water and ethanol, while containing low concentrations of other compounds, such as phenolic, which are key to characterizing wines. In order to fully chemically profile each wine, it is essential to detect metabolites with low concentrations in the 1D $^{1}$H NMR spectrum. The large dynamic range between the solvent peaks

and the wine metabolites posed a serious challenge. Therefore, an initial goal was to suppress the high-intensity solvent peaks while preserving the NMR peaks from the low concentration metabolites. Furthermore, the experimental design focused on minimizing sample preparation and handling to reduce the introduction of error.

Solvent signals are often removed by lyophilization. However, we observed residual amounts of solvent after lyophilizing the wine samples and reconstitution into buffered $D_2O$. Therefore, the water and ethanol signals were also suppressed during NMR data collection. The Bruker program Multisupp was used, which suppresses the most intense peaks in the NMR spectrum. Accordingly, the four largest peaks corresponding to the residual water and ethanol resonances were simultaneously suppressed. A dramatic improvement in the spectral quality was achieved as evident by the observation of weak metabolite peaks near the base line that were not visible prior to solvent suppression. Figure 4.1 shows a comparison between the suppressed and unsuppressed 1D $^1$H NMR spectra for the wine from Nielson. In total, 120 NMR spectra were collected. While the baseline in the multi suppressed NMR spectra was not stable across the data set and required baseline corrections, there were significantly more metabolites detectable following solvent suppression. Specifically, an abundance of low level metabolites were observed in the NMR spectral regions between 0 to 3 ppm and between 5.5 to 8 ppm, which corresponds to organic acids and carbohydrates, respectively.

Other experimental protocols were evaluated to assess which approach preserved the most metabolites detected in the 1D $^1$H NMR spectrum. For example, samples were prepared under both atmospheric conditions or a nitrogen atmosphere. No difference was observed between samples

prepared in the presence or absence of oxygen. Wine samples were also adjusted to a common pH by the addition of a $D_2O$ phosphate buffer at pH 7.2. Not surprisingly, the addition of a buffer resulted in a better comparison between NMR spectra. Any residual pH variation was accounted for by aligning the NMR spectra in data processing.

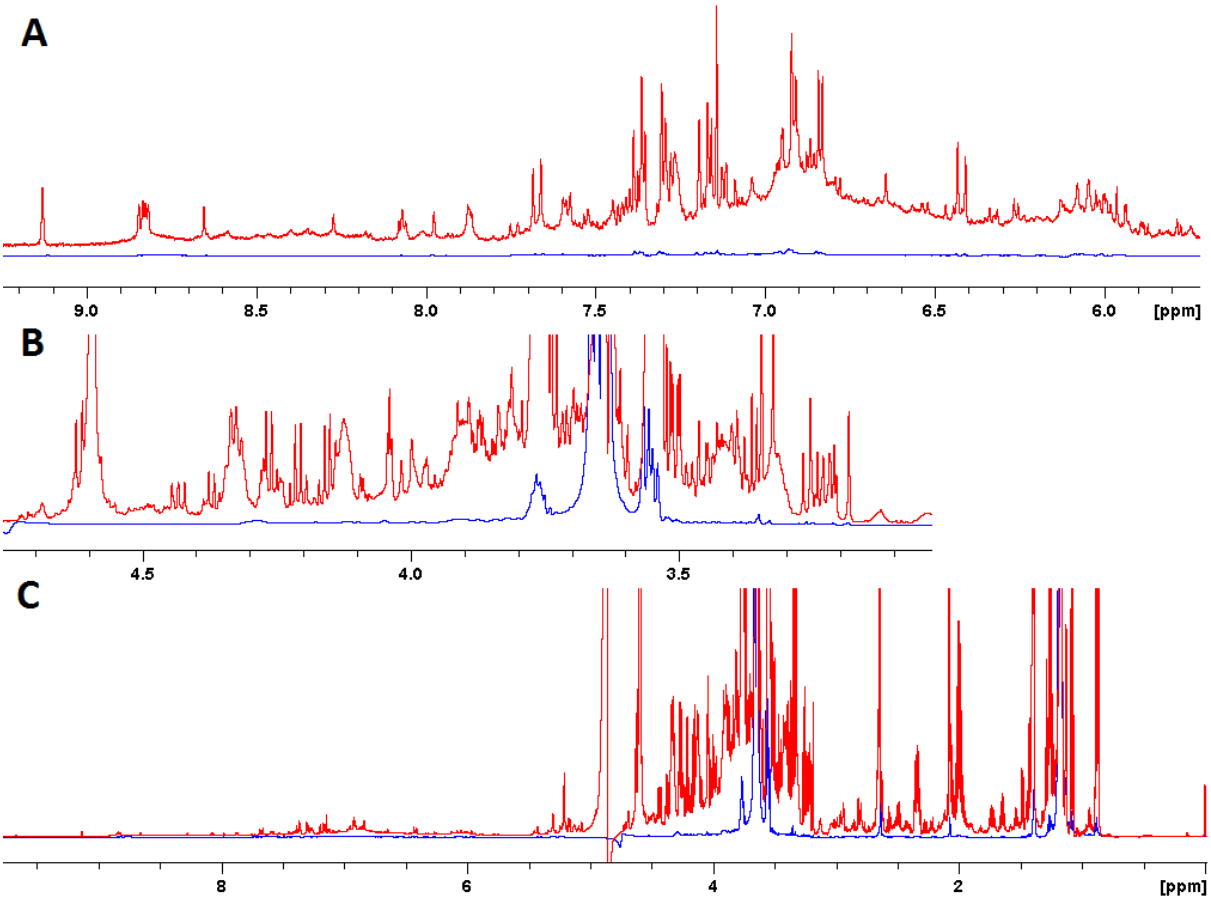



**Figure 4.1:** A comparison between a multi-suppressed 1D $^1$H NMR spectrum (red) and a water suppressed spectrum (blue) for wine from Nielson. (**A**) The expanded region of the 1D $^1$H NMR spectra between 6 ppm and 9 ppm highlights phenolic compounds. (**B**) The expanded region of the 1D $^1$H NMR spectra between 3 ppm and 5 ppm highlights sugar compounds. (**C**) Overlay of the complete 1D $^1$H NMR spectra highlighting the dramatic improvement resulting from successful solvent suppression.

### 4.4.2 Overview of the Statistical Analysis of NMR and DS array data sets

Combining the NMR and DS array data sets provides an expanded view of the entire wine metabolome [24]. Simply, the NMR data set provides an untargeted characterization of the

metabolome; in which, only the relatively high concentrated metabolites (> 3 μM) are detected. Conversely, the DS array data set is only providing a profile of the phenolic compounds present in the wine sample. Thus, the NMR and DS array datasets are highly complementary. The NMR spectral data set was UV scaled and then exported as a matrix to match the structure of the DS assay data. The NMR data matrix was generated by uniformly binning the 1D $^{1}H$ spectra such that the resulting data matrix consisted of an integrated intensity over the binned ppm range. The NMR matrix for each wine sample consisted of 8 technical replicates. The DS assay data set also contains 8 technical replicates. The NMR and DS assay data sets were analyzed individually and as a combined multiblock data set. The combined data set was used to identify which spectral features was best at categorizing each wine.

### 4.4.3 Variations in Vineyard Climates

To study the impact of climate on each wine's metabolome, environmental statistics were obtained from the Everyvine (http://www.everyvine.com/) database. Everyvine provides climate conditions for various wine regions by collecting data from individual vineyards and determining an average from the collated climate data. Climate data was assembled for the wine regions of interest with the assumption that a vineyard in a defined region will have a similar climate. The average sunlight, average high and low growing temperatures, and the total rain fall were measured during each growing season. Heliothermal or Huglin index values were also reported by Everyvine. Huglin index is a vineyard heat index that sums all temperatures above 10 $^{o}C$ between April and September [25]. The Huglin heat sum index relies on the daily median and maximum temperatures, and a parameter (k) based on the latitude of the vineyard (eqn 4.1):

$$\sum_{d=1}^{n} \boldsymbol{max}\left[\frac{T_{min}-10+T_{max}-10}{2}, \mathbf{0}\right] \boldsymbol{k} \qquad (4.1)$$

Equation 2.1: The calulation for the Huglin heat sum index where k factor is adjusted for latitude [26].

However, only averages climate values were available, which does not account for seasonal variation. Since the Pinot noir wines were derived from the same clone (Dijon 667), a major source of metabolome variations may be attributed to climate differences. In this regards, climate data was used to determine if wines that exhibited a similar metabolome also shared similar climate data. Table 4.2 lists the available climate data for the wine regions used in this study. Climate data was divided into low medium and high based on the regions selected. For the Huglin Index low is considered below 2000, with high above 2190. For average high temperature is considered below 24.5°C and high above 25.5. For average low temperature low is considered 8.5°C and high above 9.25C. Rain fall was considered low below 3 in and high above 6in.

| Vineyard | Wine Region | Huglin Index | Average High (°C) | Average Low (°C) | Rain Fall (in) |
|---|---|---|---|---|---|
| Annapolis | Sonoma Coast | 2146.3 | 25.33 | 9.11 | 6.24 |
| Cloud Landing | Sonoma Carneros | 2224.94 | 25.67 | 9.89 | 4.14 |
| Carneos Hills West | Sonoma RRV | 2197.24 | 25.78 | 8.83 | 5.92 |
| Ross | Sonoma RRV | 2197.24 | 25.78 | 8.83 | 5.92 |
| Bones | Sonoma RRV | 2197.24 | 25.78 | 8.83 | 5.92 |
| Bloomfield | Sonoma RRV | 2197.24 | 25.78 | 8.83 | 5.92 |
| Boone Ridge | Anderson Valley | 2185.79 | 25.67 | 8.50 | 7.55 |
| Maggy Hawk | Anderson Valley | 2185.79 | 25.67 | 8.50 | 7.55 |
| Panarama 5A | Arroyo | 2034.7 | 25.00 | 9.17 | 1.96 |
| MSA | Arroyo | 2034.7 | 25.00 | 9.17 | 1.96 |
| Nielson | Santa Maria Valley | 1862.37 | 23.72 | 9.50 | 2.8 |
| Rice/Cambria | Santa Maria Valley | 1862.37 | 23.72 | 9.50 | 2.8 |
| Radian | Santa Maria Hills | 1862.37 | 24.28 | 9.83 | 2.54 |
| Gran Moraine | Willamette Valley | 1749 | 21.78 | 8.11 | 14.13 |
| Zenna West | Willamette Valley | 1749 | 21.78 | 8.11 | 14.13 |

**Table 4.2: Average Climate Conditions** Local environmental and climate information was generated from averaging vineyard climate data submitted to Everyvine. Climate data was collected September of 2019. The Huglin index acts as a measure of heat and is the sum of the temperatures above 10°C during the growing period (see eqn. 4.1).

### 4.4.4 Global Comparison of PN using Metabolic Profiles

An overall PCA model was generated that included all 15 wines. A PCA model reduces a large multivariate data set into a limited number of principle components (usually 2 to 9) and identifies the unique spectral features that distinguishes the groups. A PCA model is a common approach for viewing wine classification data and is frequently used to provide an overview of the global similarity or differences between the individual wines. A PCA model is usually presented as a scores plot where each replicate (*i.e.,* 1D $^{1}$H NMR spectrum or DS assay array) is presented as a single point in the two- or three-dimensional plot. The relative similarity or difference between each replicate and/or group is assessed by how close or how far each data point or group cluster is from each other.

PCA models were generated from the 1D $^{1}$H NMR data set, the DS assay data set, and the combined NMR and DS array data set. A PCA model was generated from the data sets with replicates grouped by the individual vineyards (15 groups) or the vineyards classified by wine region (8 groups, Table 4.2). The quality of the resulting PCA models were assessed by the reported $R^2$ and $Q^2$ values. Typically, a good model has an $Q^2$ and $R^2$ close to 1, with $R^2 > Q^2$. A high $R^2$ value indicates that the data fits the model well. A high $Q^2$ value indicates a good reproducibility of the model. $Q^2$ measures the amount of variance in fitting the held-out data to the model.

PCA models could not be generated using the NMR and/or DS assay data grouped by wine region. This indicates that there was too much variation in the wine region defined data sets to generate a valid PCA plot. However, a PCA model was generated from the NMR and/or DS assay data grouped by individual vineyards. The resulting PCA models could separate each wine.

### 4.4.5 DS Assay PCA Model

The PCA model generated from the DS assay data set (**Figure 4.2A**) indicated that most of the individual vineyards formed a distinct group from the other vineyards in the scores plot. The associated dendrogram (**Figure 4.2B**) also indicated a clear separation between most of the vineyards. Thus, despite the wines being derived from the same Pinot noir clone Dijon 667, environmental and climate factors, among others, define the chemical profile of the wines.

The dendrograms were produced from the associated PCA scores plot based on a matrix of Mahalanobis distances between each group. Each node in the dendrogram was labeled with a p-value indicating the statistical significance of the group separation. Overall, the dendrogram

clusters the vineyards into 4 groups containing between 3 to 5 vineyards within each cluster. A few wines did significantly overlap (p-values > 0.05). Notably, wines from vineyards in the same region did not all cluster together, which is consistent with the failure to generate a PCA model based on wine regions

Wines from Bloomfield, Annapolis, and Cloud Landing did cluster together, which could be due to the fact that these vineyards are all located in Sonoma County. However, the wines are in distinct wine regions. Bones and Carneros Hills West clustered together and do belong to the same wine region, Sonoma RRV. In fact, the wines were statistically indistinguishable (p-value 0.33). However, the other vineyards in Sonoma RRV, Ross and Bloomfield, were found in other clusters. Boone Ridge and Nielson also clustered together and were statistically indistinguishable (p-value 0.23). These wines were not in the same wine region and do not share climate averages as observed from the Everyvine data. This suggests that general location and/or climate data are insufficient to explain relative vineyard groupings. Other terroir factors, such as soil conditions, or wine processing protocols, may better explain the relative clustering of vineyards.

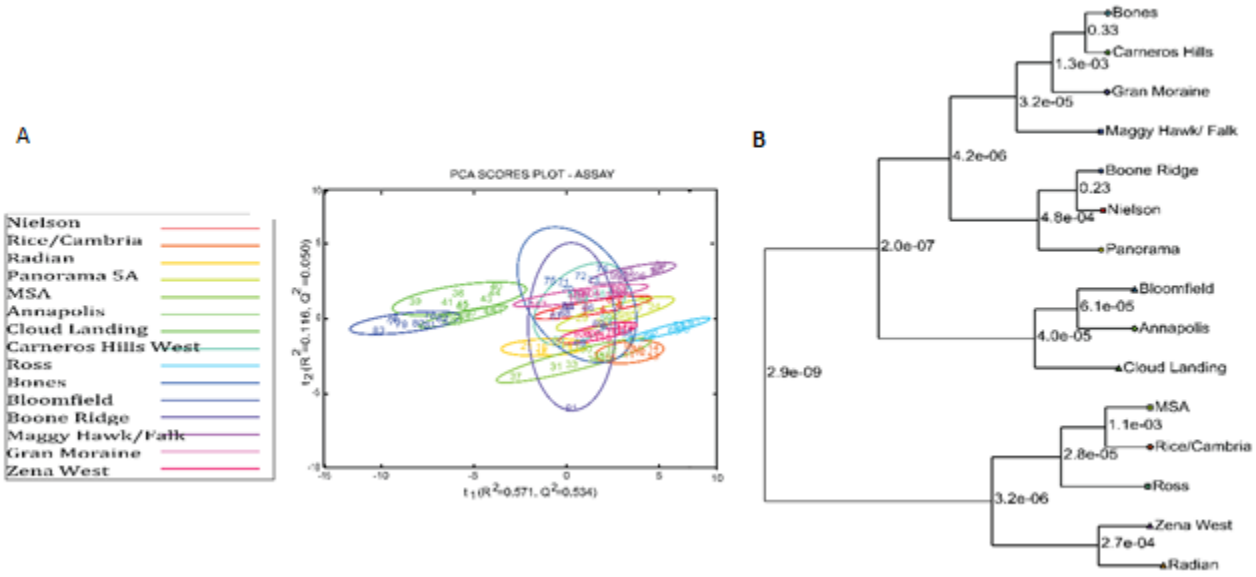



**Figure 4.2:** PCA model ($R^2$ .787, $Q^2$ .634 generated from the DS assay data. (**A**) PCA scores plot showing all 15 wines where Nielson (red), Rice/ Cambria (Orange), Radian (Yellow), Panorama 5A (Lime Green), MSA ( Light Green), Annapolis (Green), Cloud Landing (Green), Carneros Hills West (Teal), Ross (Cyan), Bones (Blue), Bloomfield (Navy Blue), Boone Ridge (Violet), Maggy Hawk (Purple), Gran Moraine (magenta), Zena West (Pink). Each ellipse corresponds to 95% confidence interval for a normal distribution. (**B**) The dendrogram was generated from the PCA scores plot in **A** and is based on a matrix of Mahalanobis distances between each wine group in the PCA scores plot. Each node in the dendrogram is labeled with a pairwise p-value.

### 4.4.6 NMR PCA Model

The PCA model generated from the 1D $^1$H NMR data (**Figure 4.3**) clustered similarly to the DS assay data. Most of the individual vineyards formed a separate and distinct group in the PCA scores plot. The associated dendrogram clustered the wines into 5 distinct groups containing between 2 to 4 wines each. Like the DS assay data, the vineyard grouping was not determined by wine region or climate data. Notably, the relative wine clustering differed between the NMR and DS array data sets. For example, Bones is now nearly overlapped with Zena West (p –value 0.04) instead of

Carneros Hills West. Similarly, Boone Ridge is close to, but not identical (p-value $2.6x10^{-4}$) to Cloud Landing instead of Nielson. While Ross and Bloomfield are still clustered together they are now grouped with Annapolis. They are all from the Sonoma county, but not the same wine region. All of the wines were more distinguishable based on the NMR data set (p-values < 0.05). The pair of wines with the highest p value (p-value 0.05) was Panorama 5A and Radian, which are from the Santa Barbra County, but not from the same wine region.

Wines from Bloomfield, Annapolis, and Cloud Landing did cluster together, which could be due to the fact that these vineyards are all located in Sonoma County. However, the wines are in distinct wine regions. Bones and Carneros Hills West clustered together and do belong to the same wine region, Sonoma RRV. In fact, the wines were statistically indistinguishable (p-value 0.33). However, the other vineyards in Sonoma RRV, Ross and Bloomfield, were found in other clusters. Boone Ridge and Nielson also clustered together and were statistically indistinguishable (p-value 0.23). These wines were not in the same wine region and do not share climate averages as observed form the Everyvine data. This suggests that general location and/or climate data are insufficient to explain relative vineyard groupings. Other terroir factors, such as soil conditions, or wine processing protocols, may better explain the relative clustering of vineyards.

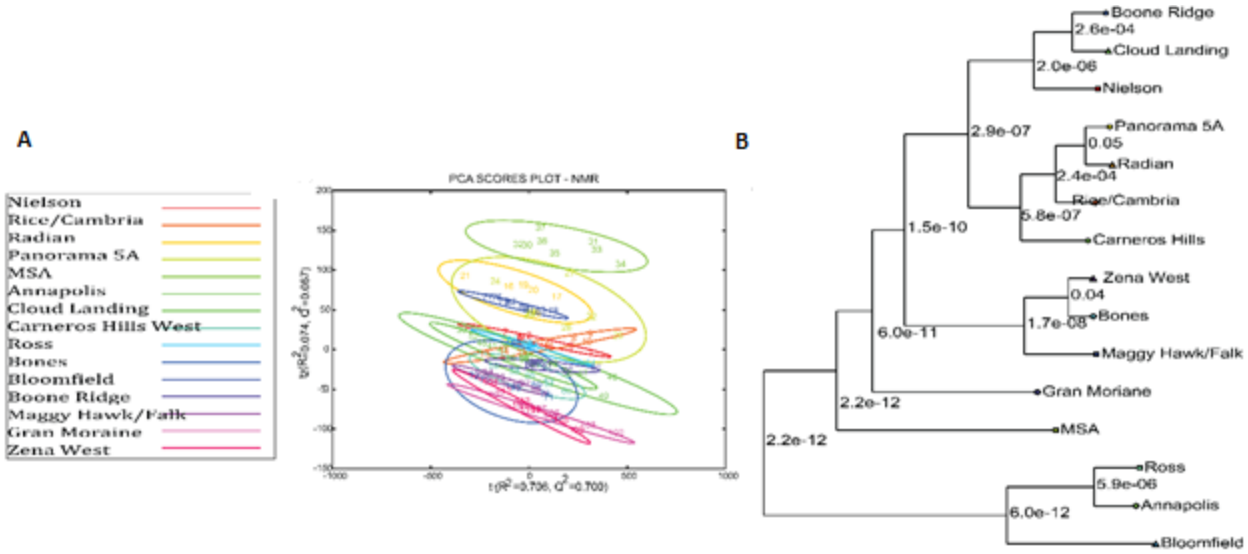



**Figure 4.3:** PCA model ($R^2$ .812 $Q^2$ .77 generated from the 1D $^1$H NMR data. (**A**) PCA scores plot showing all 15 wines where Nielson (red), Rice/ Cambria (Orange), Radian (Yellow), Panorama 5A (Lime Green), MSA ( Light Green), Annapolis (Green), Cloud Landing (Green), Carneros Hills West (Teal), Ross (Cyan), Bones (Blue), Bloomfield (Navy Blue), Boone Ridge (Violet), Maggy Hawk (Purple), Gran Moraine (magenta), Zena West (Pink). Each ellipse corresponds to 95% confidence interval for a normal distribution. (**B**) The dendrogram was generated from the PCA scores plot in **A** and is based on a matrix of Mahalanobis distances between each wine group in the PCA scores plot. Each node in the dendrogram is labeled with a pairwise p-value.

### 4.4.7 Multiblock PCA Model

Unexpectedly, the multiblock-PCA (MB-PCA) model generated from the combined NMR and DS array data did not perform as well as the individual data sets. Very few vineyards were clearly separated in the MB-PCA scores plot (**Figure 4.4A**).

In fact, the pairwise comparison of six vineyards yielded p-values > 0.05 indicating no difference in the chemical profile. Bones, Carneos Hills, and Gran Morane could not be distinguished by the MB-PCA model. The Bone Ridge, Nielson, and Zena West wines also were indistinguishable. These wines do not share similar climate or locational data. The associated dendrogram clustered the wines into 4 clusters containing between 3 to 5 wines each. Again, there was no similarity in

the clustering patterns between the three PCA models (**Figures 4.2** to 4.**4**), suggesting unique, non-overlapping information between the NMR and DS array data sets. Furthermore, since combining the two data sets resulted in a worse differentiation between the vineyards, the discriminating spectral features may be anti-correlated. In essence, variance in the NMR data sets partially cancel variance in the DS array data sets.

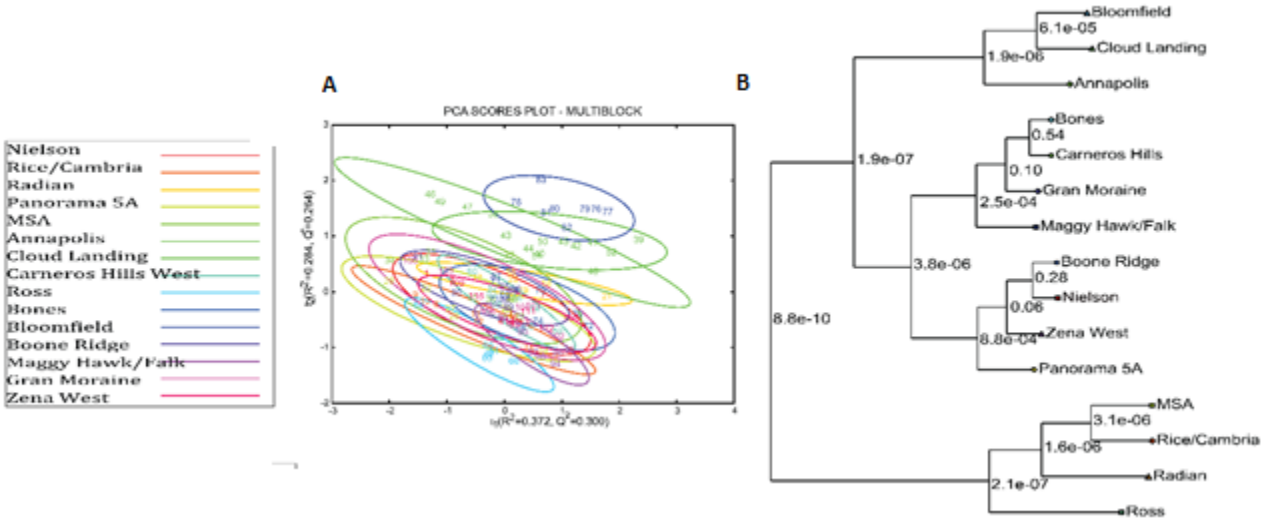



**Figure 4.4:** PCA model ($R^2$ .767 $Q^2$ .623) generated from the combined NMR and DS assay data set. (**A**) PCA scores plot showing all 15 wines where Nielson (red), Rice/ Cambria (Orange), Radian (Yellow), Panorama 5A (Lime Green), MSA (Light Green), Annapolis (Green), Cloud Landing (Green), Carneros Hills West (Teal), Ross (Cyan), Bones (Blue), Bloomfield (Navy Blue), Boone Ridge (Violet), Maggy Hawk (Purple), Gran Moraine (magenta), Zena West (Pink). Each ellipse corresponds to 95% confidence interval for a normal distribution. (**B**) The dendrogram was generated from the PCA scores plot in **A** and is based on a matrix of Mahalanobis distances between each wine group in the PCA scores plot. Each node in the dendrogram is labeled with a pairwise p-value.

### 4.4.8 Wine Classification using a ROC Curve Analysis

Receiver operator characteristic (ROC) curves plot the true positive rate against the false positive rate. In this regard, a ROC curve is used to ascertain the predictive accuracy of a set of signals or spectral features by measuring the area under the curve (AUC). The AUC ranges from 1 for a

perfect prediction to 0.5 for a completely random outcome. A ROC curve was used to identify the spectral features most useful for wine classification. Specifically, which NMR and/or DS assay spectral features were better at classifying each wine to a given vineyard? To accomplish this, a ROC curve was generated that compared each individual wine against the entire collection of PN wines. A total of 15 ROC curves were produced, one for each wine listed in **Table 4.1**. Overall, most ROC curves yielded an AUC close to 1 with the lowest AUC being approximately 0.8. This indicates that the ROC curves have an accuracy of > 80 to 90% in correctly classifying each PN wine from the set of 15 wines. Notably, most ROC curves required only 10 spectral features to achieve the high AUC values. Furthermore, the contribution of NMR and/or DS array spectral features to the ROC curves varied by wine or vineyard. In some cases, the model was dominated by NMR spectral features, in other cases by DS array data, or as a nearly equal combination of both NMR and DS array spectral features. Table 4.3 shows a summary of the ROC curves.

| Wine | Area Under the Curve | Confidence Interval | NMR Components | Assay Components |
|---|---|---|---|---|
| Nielson | .736 | .376 – .979 | 0 | 5 |
| Rice/Cambria | .926 | .614 – .995 | 3 | 2 |
| Radian | .808 | .335 – .978 | 1 | 4 |
| Panorama 5A | .766 | .496 – .937 | 1 | 4 |
| MSA | .942 | .496 – .937 | 5 | 0 |
| Annapolis | .916 | .778 – 1 | 1 | 4 |
| Cloud Landing | .904 | .642 – .991 | 0 | 5 |
| Canneros Hills | .897 | .733 – .933 | 1 | 4 |
| Ross | .967 | .883 – 1 | 4 | 1 |
| Bones | .847 | .35 – 1 | 0 | 5 |
| Bloomfield | .97 | .814 – 1 | 2 | 3 |
| Boone Ridge | .792 | .577 – .938 | 4 | 1 |
| Maggy Hawk/Falk | .907 | .606 – 1 | 0 | 5 |
| Gran Moraine | .909 | .674 – 1 | 1 | 4 |
| Zena West | .86 | .576 – .992 | 2 | 3 |

**Table 4.3**: Shows the summary of the ROC curves for each vineyard and includes the area under the curve and confidence interval for the best performing ROC curve. It also lists the top ten components used to generate the curves and whither they came from the NMR or Assay data.

Assigning the NMR spectral features to a specific metabolite is challenging. Databases of NMR reference spectra are incomplete and typically lack secondary metabolites from plants [27]. Instead, databases are primarily populated with metabolites associated with known metabolic process. So, uniquely modified compounds are lacking in reference databases and NMR chemical shifts can only be estimated based on similarities to known compounds. Spiking NMR samples with a commercially available compound can be used to identify and confirm the presence of a specific metabolite. Of course, there is a very limited availability of known metabolites, especially in regard to secondary metabolites from plants. The typical 1D $^{1}$H NMR spectrum for wine can be divided into three sections. The 0 to 3 ppm region contains organic acids that includes lactic acid, acetic acid, citric acid and malic acid. The 3 to 5.5 ppm region contains carbohydrates that includes

glucose and fructose. The > 5.5 ppm region includes aromatic compounds, such as 2-phenylethanol [17]. While the compounds contributing to each wine's chemical profile can't be individually identified, the class of compounds may be inferred.

### 4.4.9 Wine Regions

#### 4.4.9.1 Santa Maria Valley

Santa Maria Valley is in northern Santa Barbara County and San Luis Obispo County in California. Two vineyards, Nielson and Rice/Cambria, were selected from the Santa Maria Valley wine region for this study.

For the PCA models using the DS assay data, Nielson clustered with Cloud Landing and Boone Ridge vineyards. Nielson clustered with Boone Ridge and Panorama 5A using the NMR data. Similarly, Nielson clustered with Boone Ridge, Zena West, and Panorama 5A using the MB data set.

For the Nielson ROC curves shown in **Figure 4.5**, the ROC curve with the highest AUC of 0.802 was generated using 10 variables. The AUC of 0.802 indicates a predictive accuracy of 80% when differentiating the Nielson wine from the 14 other wines. The ROC curve was defined predominately with DS array assay data, with MM7 430 being the top feature that distinguished the Nielson wine.

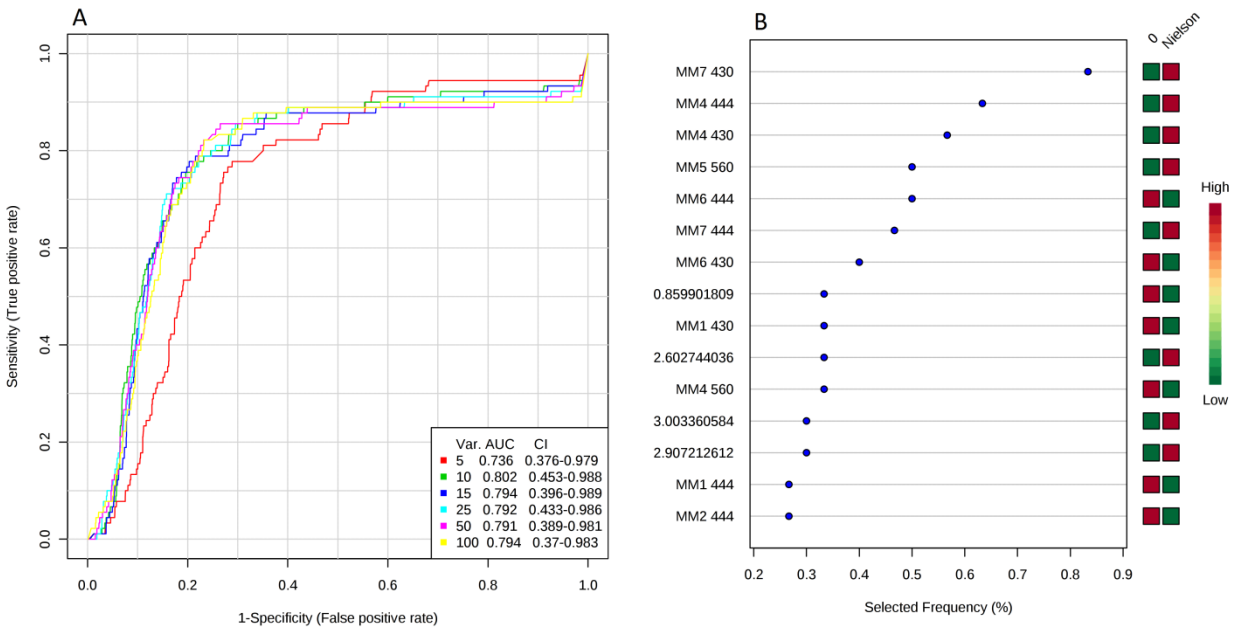



**Figure 4.5:** (**A**) The ROC curves for the Nielson wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

For the PCA models using the DS assay data, Rice/Cambria clustered with MSA and Ross vineyards. Rice/Cambria clustered with Radian, Panorama 5A, and Carneos Hills West using the NMR data. Similarly, Rice/Cambria clustered with Radian, Panorama 5A, and Carneos Hills West using the MB data set.

For the Rice/Cambria ROC curves shown in **Figure 4.6**, the ROC curve with a high AUC of 0.926 was generated using 5 variables. The AUC of 0.926 indicates a predictive accuracy of 93% when differentiating the Rice/Cambria wine from the 14 other wines. The ROC curve used a significant amount of NMR data. Since the majority of the NMR bins were in the range of 2 to 3 ppm, the metabolites potentially corresponded to organic acids.

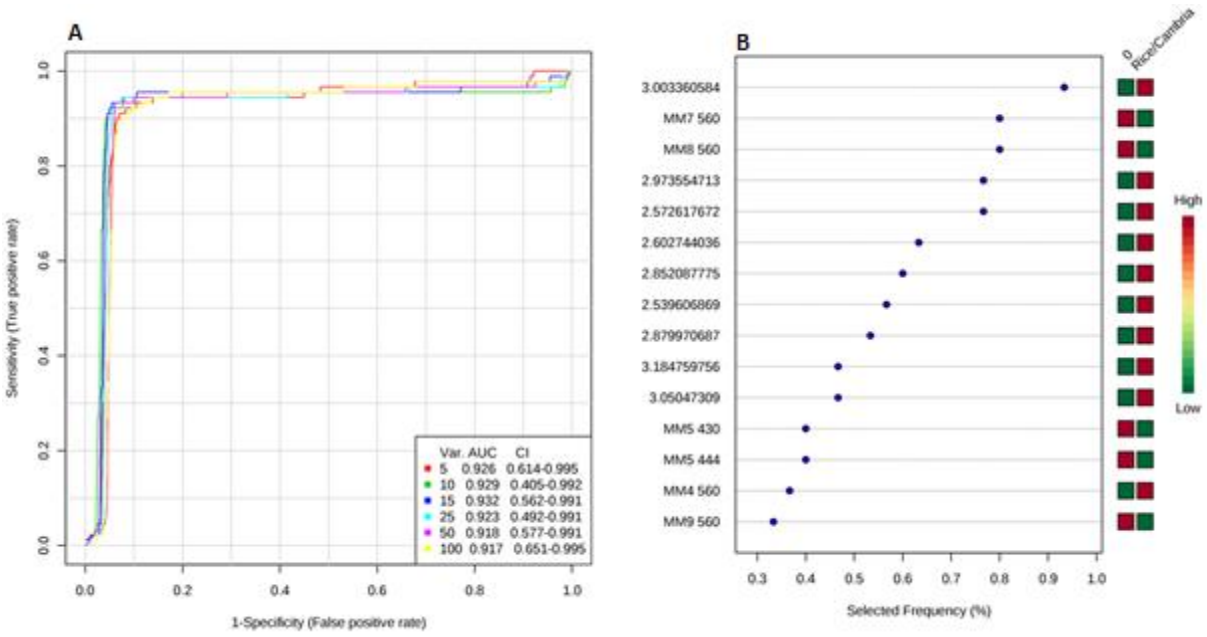



**Figure 4.6:** (**A**) The ROC curves for the Rice/Cambria wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

#### 4.4.9.2 Santa Maria Hills

Santa Maria Hills is a wine region located in the Santa Ynez Valley in California. Santa Rita Hills has the highest level of solar radiation at 149663.28 WH/m$^2$. Only one wine, Radian, was selected from the Santa Maria Hills wine region.

For the PCA models using the DS assay data, Radian clustered with Zena West. Radian clustered with Panorama 5A and Rice/Cambria using the NMR data. Radian clustered with MSA, Rice/Cambria, and Ross using the MB data set. The vineyards did not share a similar climate with Radian.

For the Radian ROC curves shown in **Figure 4.7**, the ROC curve with a high AUC of 0.86 was

generated using 15 variables. The AUC of 0.86 indicates a predictive accuracy of 86% when differentiating the Radian wine from the 14 other wines. The ROC curve was nearly defined by all DS array data, but an NMR bin was the top feature that distinguished the Radian wine. The ppm of .9 suggests that this feature is an organic acid.

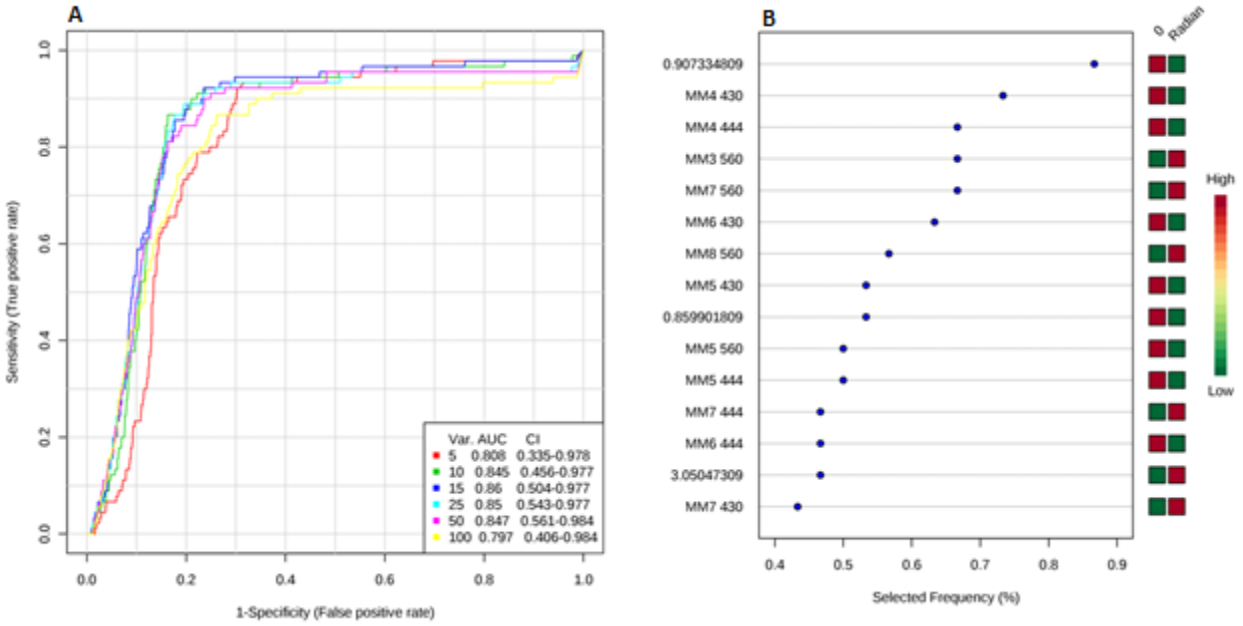



**Figure 4.7:** (**A**) The ROC curves for the Radian wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

#### 4.4.9.3 Arroyo

Arroyo or the Arroyo Grande Valley wine region is located in the San Luis Obispo county of California Of all the wine regions selected for this study, Arroyo has the lowest rain fall at 1.96 inches per year. Two wines, Panarma 5A and MSA, were selected from the Arroyo wine region.

For the PCA models using the DS assay data, Panorama 5A clustered with Nielson and Boone

Ridge. Panorama 5A clustered with Radian, Rice/Cambria and Carneros Hills using the NMR data. Panorama 5A clustered with Zena West, Nielson, Boone Ridge using MB data set. Zena West does not share environmental conditions with Panorama 5A.

For the Panorama 5A ROC curves shown in **Figure 4.8**, the ROC curve with a high AUC of 0.841 was generated using 15 variables. The AUC of 0. 841 indicates a predictive accuracy of 84% when differentiating the Panorama 5A wine from the 14 other wines. The ROC curve used a mixture of NMR and DS array data, with MM7 560 being the top feature that distinguished the Panorama 5A wine. The NMR bins covered a range of chemical shifts from 0.9 to 4.3 ppm, suggesting key metabolites defining Panorama 5A as potentially corresponding to carbohydrates and organic acids.

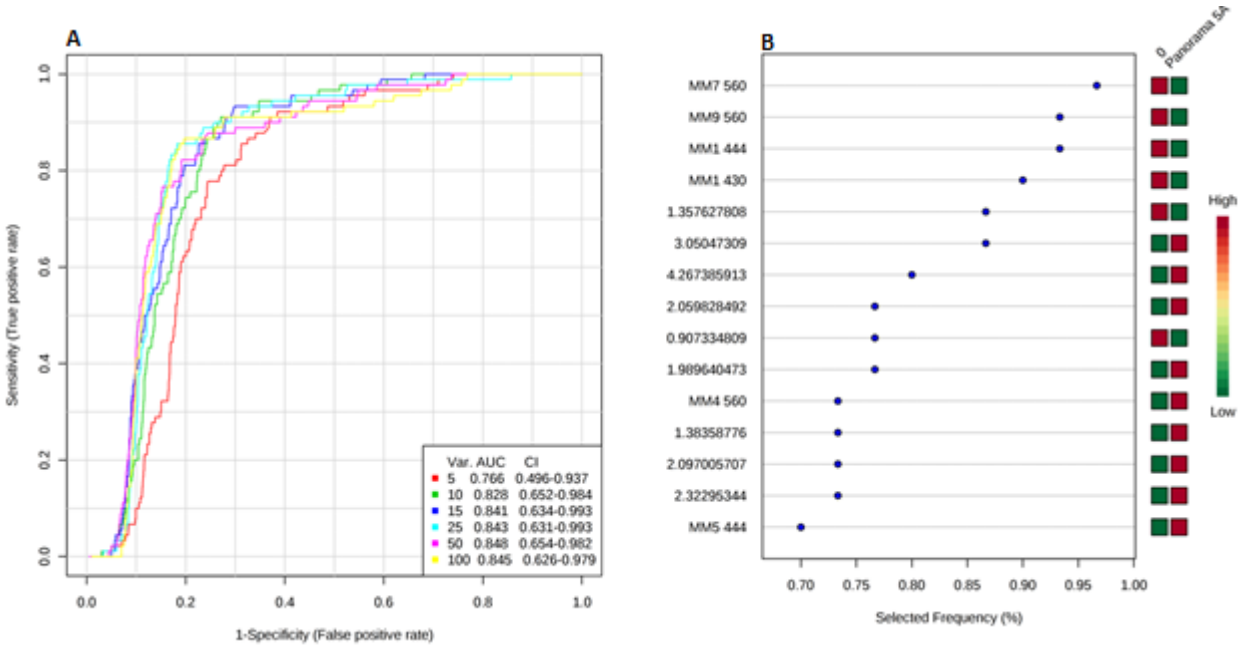



**Figure 4.8:** (**A**) The ROC curves for the Panorama 5A wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

#### 4.4.9.4 MSA

For the PCA models using the DS assay data, MSA clustered with Rice/Cambria and Ross. MSA clustered with Rice/Cambria and Radian using the NMR data. MSA clustered again with Rice/Cambria and Radian. using the MB data set.

For the MSA ROC curves shown in **Figure 4.9**, the ROC curve with a high AUC of 0.96 was generated using 10 variables. The AUC of 0.96 indicates a predictive accuracy of 96% when differentiating the MSA wine from the 14 other wines. The ROC curve was nearly exclusively defined by NMR data. The NMR bins covered a range of chemical shifts from 1.5 to 3.3 ppm, suggesting key metabolites defining MSA are potentially organic acids.

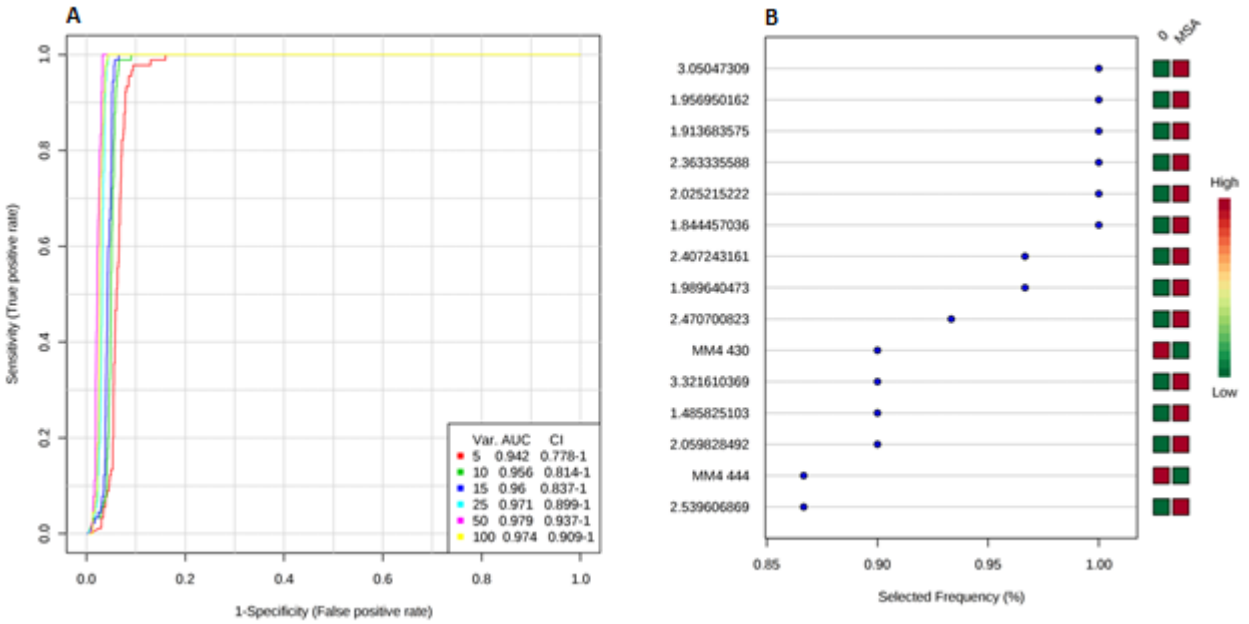



**Figure 4.9:** (**A**) The ROC curves for the MSA wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

#### 4.4.9.5 Sonoma Coast

Only one wine, Annapolis, was selected from Sonoma Coast region. For the PCA models using the DS assay data, Annapolis clustered with Bloomfield and Cloud Landing. Annapolis clustered with Ross and Bloomfield using the NMR data. Annapolis clustered again with Bloomfield and Cloud Landing using the MB data set.

For the Annapolis ROC curves shown in **Figure 4.10**, the ROC curve with a high AUC of 0.92 was generated using 10 variables. The AUC of 0. 92 indicates a predictive accuracy of 92% when

differentiating the Annapolis wine from the 14 other wines. The ROC curve was defined as a mixture of NMR and DS assay features. The NMR bins covered a range of chemical shifts from 1.3 to 3.5 ppm, suggesting key metabolites defining Annapolis are potentially carbohydrates and organic acids.

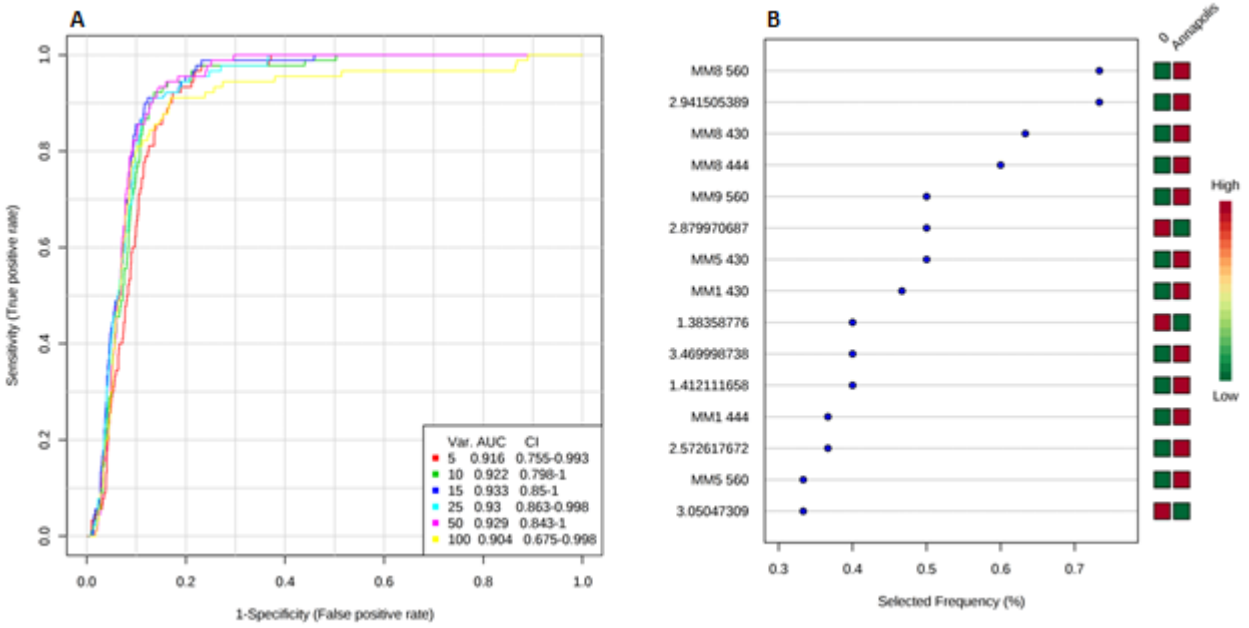



**Figure 4.10:** (**A**) The ROC curves for the Annapolis wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

#### 4.4.9.6 Sonoma Carneros

Sonoma Carneros is a wine region in California. Sonoma Carneros is one of the hottest wine regions included in this study. The Huglin index is the highest at 2224.94. It also experiences the highest growing low temperature at 9.88°C. Only one vineyard, Cloud Landing, was selected from the Sonoma Carneros wine region.

For the PCA models using the DS assay data, Cloud Landing clustered with Annapolis and Bloomfield. Cloud Landing clustered with Boone Ridge and Nielson using the NMR data. Cloud Landing clustered again with Annapolis and Bloomfield using the MB data set.

For the Cloud Landing ROC curves shown in **Figure 4.11**, the ROC curve with a high AUC of 0.936 was generated using 15 variables. The AUC of 0. 936 indicates a predictive accuracy of 94% when differentiating the Cloud Landing wine from the 14 other wines. The ROC curve was defined primarily by DS assay features, with MM2 560 as the top feature distinguishing the Cloud Landing wine.

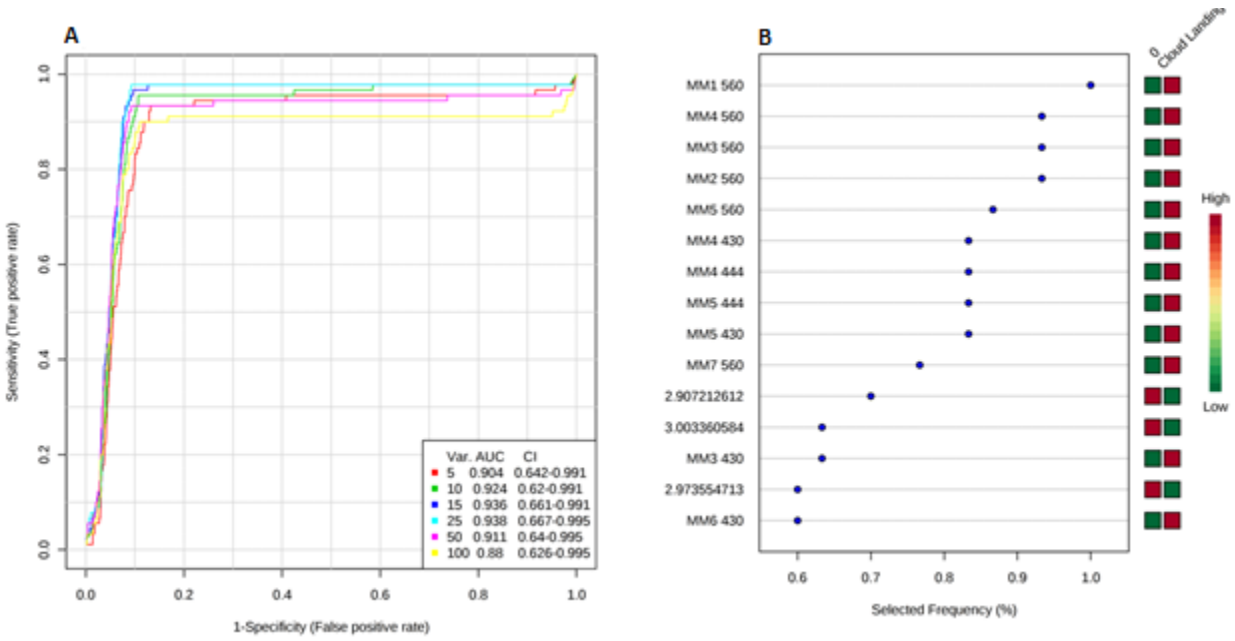



**Figure 4.11:** (**A**) The ROC curves for the Cloud Landing wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

#### 4.4.9.7 Sonoma RRV

Sonoma RRV or the Russian River Valley is located in Sonoma County. It is also a rather hot wine region with the highest high growing temperature of 25.77 $^{o}$C. Sonoma RRV was the wine region with the largest number of vineyards used in this study. Specifically, Sonoma RRV consisted of vineyards: Carneos Hills West, Bloomfield, Bones, and Ross.

For the PCA models using the DS assay data, Carneos Hills West clustered with Bones, Gran Moraine, and Maggy Hawk. Carneos Hills West clustered with Rice/Cambria, Radian and Panorama 5A using the NMR data. Carneos Hills West clustered again with Bones, Gran Moraine, Maggy Hawk/Falk using the MB data set. Carneos Hills West does not share a similar climate with any of these vineyards.

For the Carneos Hills West ROC curves shown in **Figure 4.12**, the ROC curve with a high AUC of 0.935 was generated using 25 variables. The AUC of 0. 935 indicates a predictive accuracy of 94% when differentiating the Carneos Hills West wine from the 14 other wines. The ROC curve was defined by a majority of DS assay features. The NMR features included chemical shift bins of 2.94, 2.91, 2.77, 2.21, and 1.36, which are likely organic acids.

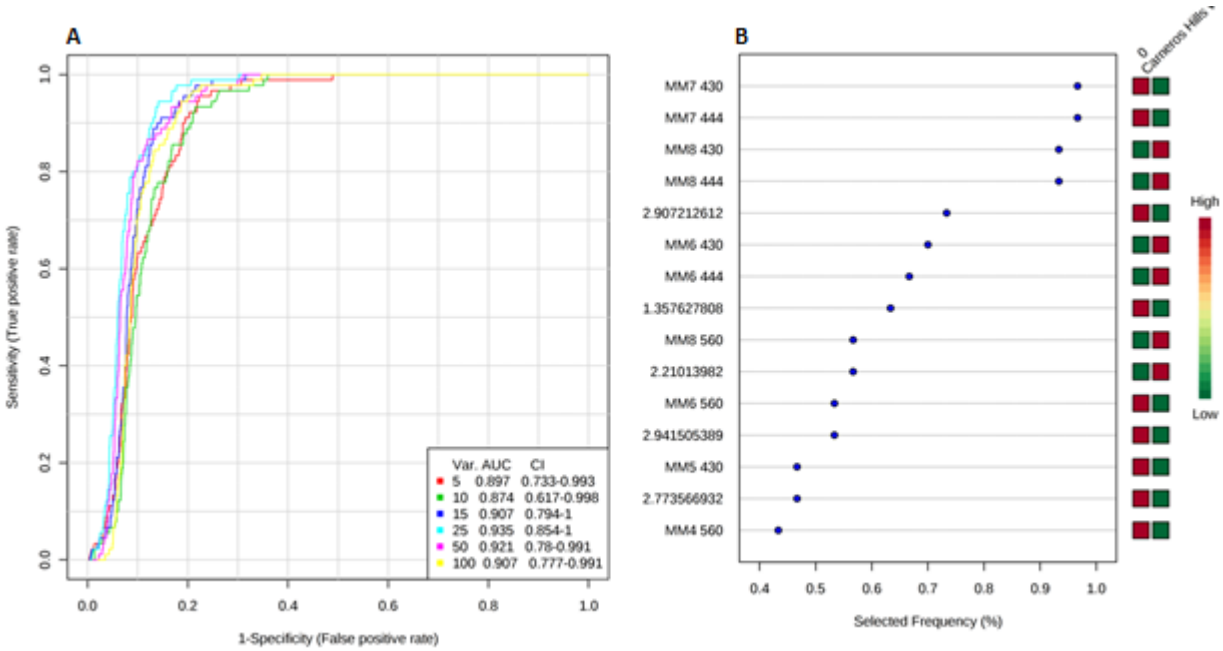



**Figure 4.12**: (**A**) The ROC curves for the Carneos Hills West wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

For the PCA models using the DS assay data, Ross clustered with MSA and Rice/Cambria. Ross clustered with Annapolis, Bloomfield, Zena West, and Radian using the NMR data. Ross clustered with Rice, MSA, and Radian using the MB data set.

For the Ross ROC curves shown in **Figure 4.13**, the ROC curve with a high AUC of 0.983 was generated using 10 variables. The AUC of 0.983 indicates a predictive accuracy of 98% when differentiating the Ross wine from the 14 other wines. The ROC curve was defined by a mixture

NMR and DS assay features. The NMR features included chemical shift bins that ranged from 1.38 to 2.97 ppm, which are likely organic acids.

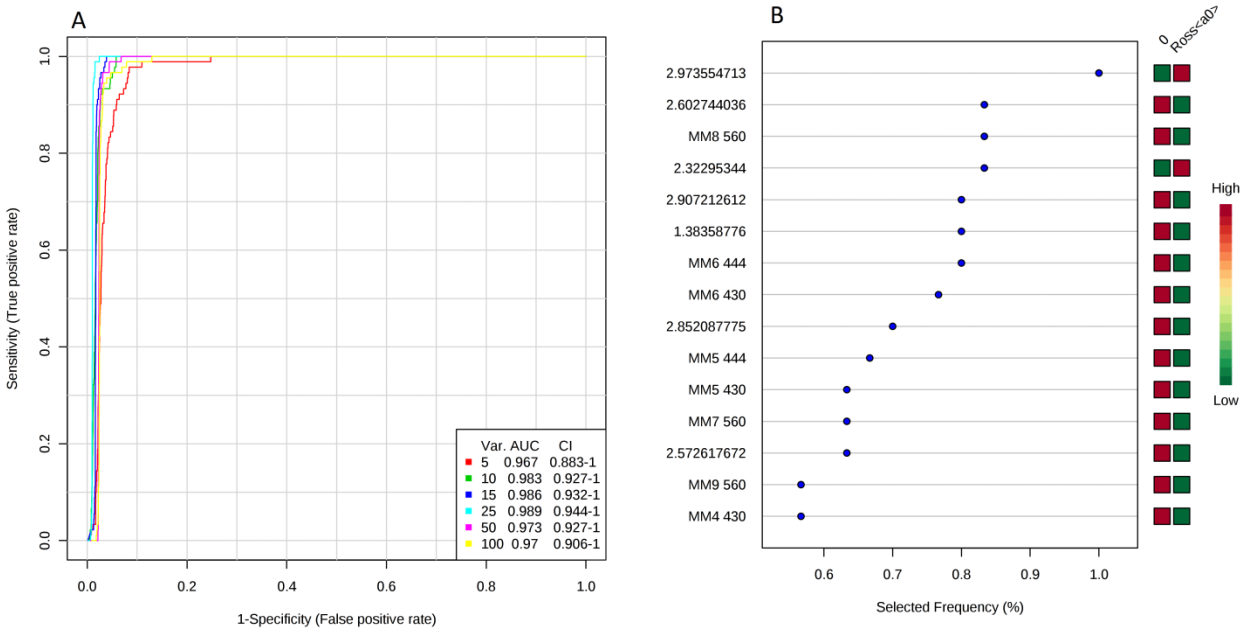



**Figure 4.13**: (**A**) The ROC curves for the Ross wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

For the PCA models using the DS assay data, Bones clustered with Carneos Hills West and Gran Moraine. Bones clustered with Zena West and Maggy Hawk using the NMR data. Bones clustered with Carneos Hills West and Gran Moraine using the MB data set.

For the Bones ROC curves shown in **Figure 4.14**, the ROC curve with a high AUC of 0.836 was generated using 15 variables. The AUC of 0.836 indicates a predictive accuracy of 84% when differentiating the Bones wine from the 14 other wines. The ROC curve was defined by a majority

of DS assay features with MM9 444 and MM9 430 being the top features distinguishing the Bones wine.

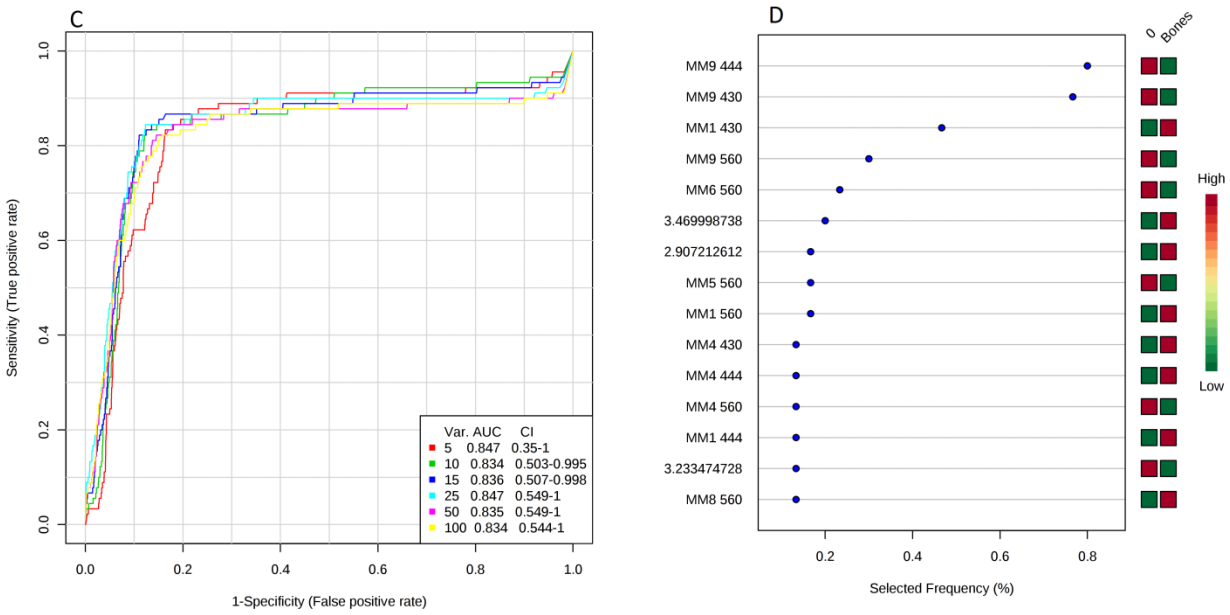



**Figure 4.14**: (**A**) The ROC curves for the Bones wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

For the PCA models using the DS assay data, Bloomfield clustered with Annapolis and Cloud Landing. Bloomfield clustered with Ross and Annapolis using the NMR data.Bloomfield clustered again with Annapolis and Cloud Landing using the MB data set.

For the Bloomfield ROC curves shown in **Figure 4.15**, the ROC curve with a high AUC of 0.98 was generated using 15 variables. The AUC of 0.986 indicates a predictive accuracy of 98% when differentiating the Bloomfield wine from the 14 other wines. The ROC curve was defined by a majority of DS assay features, but an NMR bin was top feature that differentiates Bloomfield from

the other wines.

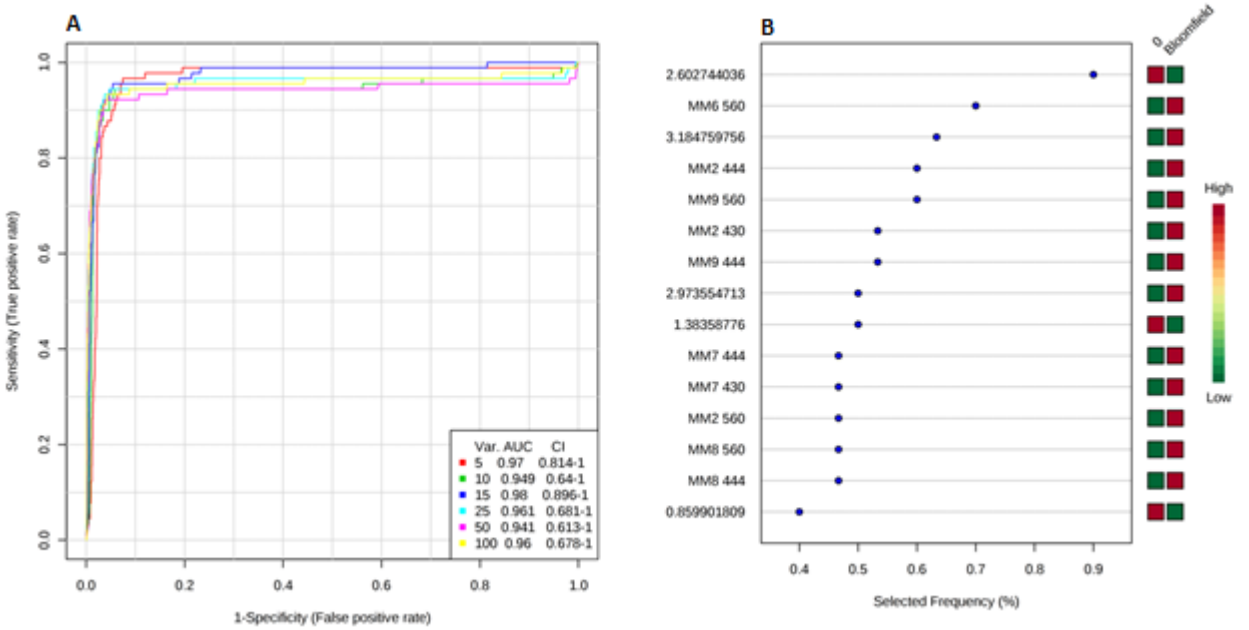



**Figure 4.15**: (**A**) The ROC curves for the Bloomfield wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

#### 4.4.9.8 Anderson Valley

Anderson Valley is a wine region in Mendocino County California. Two vineyards, Boone Ridge and Maggy Hawk, were selected from the Anderson Valley.

For the PCA models using the DS assay data, Boone Ridge clustered with Nielson and Panorama 5A. Boone Ridge clustered with Cloud Landing and Nielson using the NMR data. Boone Ridge clustered with Nielson, Zena West and Panorama 5A using the MB data set.

For the Boone Ridge ROC curves shown in **Figure 4.16**, the ROC curve with a high AUC of 0.809 was generated using 10 variables. The AUC of 0.809 indicates a predictive accuracy of 81% when differentiating the Boone Ridge wine from the 14 other wines. The ROC curve was defined by a mixture of NMR and DS assay features, where a number NMR bins were top features that differentiates Boone Ridge from the other wines. The NMR bins corresponded to chemical shifts that ranged from 1.49 to 4.06, which are likely carbohydrates and organic acids.

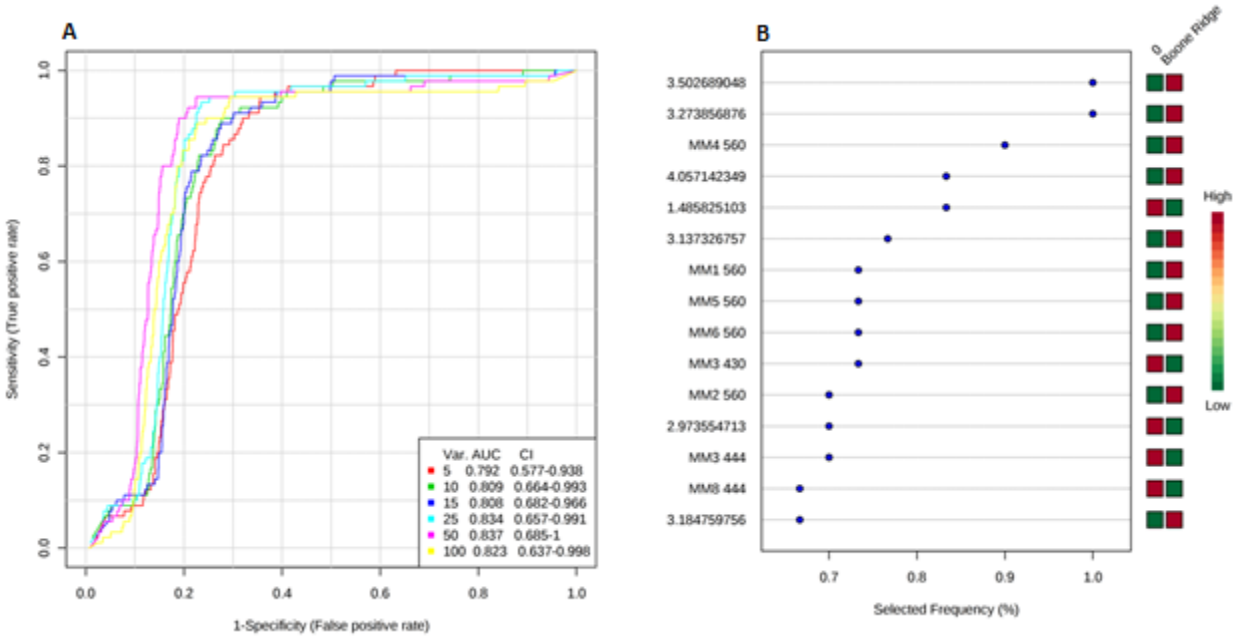



**Figure 4.16**: (**A**) The ROC curves for the Boone Ridge wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

For the PCA models using the DS assay data, Maggy Hawk/Falk clustered with Gran Moraine,

Carneos Hills, and Bones. Maggy Hawk/Falk clustered with Bones and Zena West using the NMR data. Maggy Hawk/Falk clustered with Gran Moraine, Carnos Hills, and Bones using the MB data set.

For the Maggy Hawk/Falk ROC curves shown in **Figure 4.17**, the ROC curve with a high AUC of 0.929 was generated using 10 variables. The AUC of 0.929 indicates a predictive accuracy of 93% when differentiating the Maggy Hawk/Falk wine from the 14 other wines. The ROC curve was defined nearly exclusively by DS assay features, with MM4 560 as the top distinguishing feature.

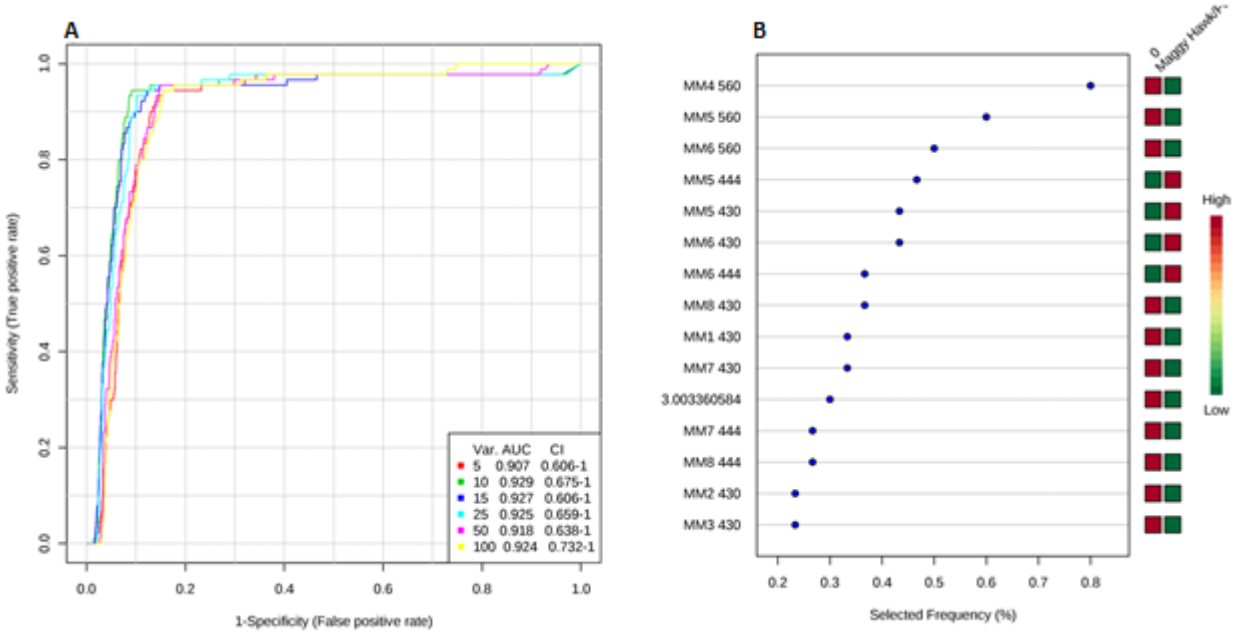



**Figure 4.17**: (**A**) The ROC curves for the Maggy Hawk/Falk wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

Anderson Valley vineyards did not cluster in the PCA models with vineyards that shared similar climate conditions. Interestingly, ROC data curves were split in their utilization of NMR and DS array data. Boone Ridge used a mixture of DS assay and NMR data favoring organic acids, while Maggy Hawk used nearly all DS assay data.

#### 4.4.9.9 Willamette Valley

Willamette Valley is located in Oregon. Willamette Valley is the least sunny and the coldest the wine region used in this study. Its solar radiation is 131092.27 $WH/m^2$ with the smallest Huglin index of 1749. Willamette Valley also has the lowest high and low growing temperatures at 21.9 $^oC$ and 8.1 $^oC$, respectively. Two vineyards, Gran Moraine and Zena West, were selected from the Willamette Valley wine region.

For the PCA models using the DS assay data, Gran Moraine clustered with Bones, Carneos Hills, and Maggy Halk. Gran Moraine clustered with MSA using the NMR data. Gran Moraine again clustered with Bones, Carneos Hills, and Maggy Halk using MB data set.

For the Gran Moraine ROC curves shown in **Figure 4.18**, the ROC curve with a high AUC of 0.976 was generated using 25 variables. The AUC of 0.976 indicates a predictive accuracy of 98% when differentiating the Gran Moraine wine from the 14 other wines. The ROC curve was defined by a majority of NMR features, with only three DS array values. The NMR bins corresponded to chemical shifts that ranged from 1.99 to 4.27, which are likely carbohydrates and organic acids.

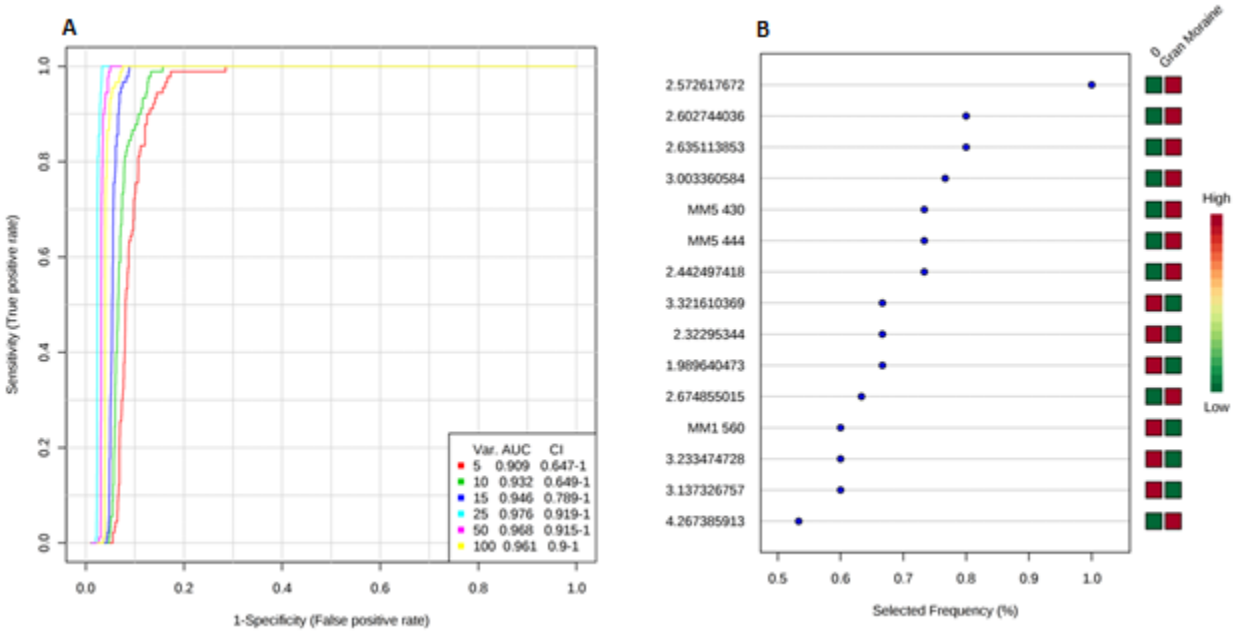



**Figure 4.18**: (**A**) The ROC curves for the Gran Moraine wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

For the PCA models using the DS assay data, Zena West clustered with Radian. Zena West clustered with Nielson, Boone Ridge, and Panarama 5A using the NMR data. Zena West again clustered with Nielson, Boone Ridge, and Panarama 5A using the MB data set.

For the Zena West ROC curves shown in **Figure 4.18**, the ROC curve with a high AUC of 0.95 was generated using 25 variables. The AUC of 0.95 indicates a predictive accuracy of 95% when differentiating the Zena West wine from the 14 other wines. The ROC curve was defined with a mixture of NMR and DS array features. The NMR bins corresponded to chemical shifts that ranged from 0.86 to 3.50, which are likely carbohydrates and organic acids.

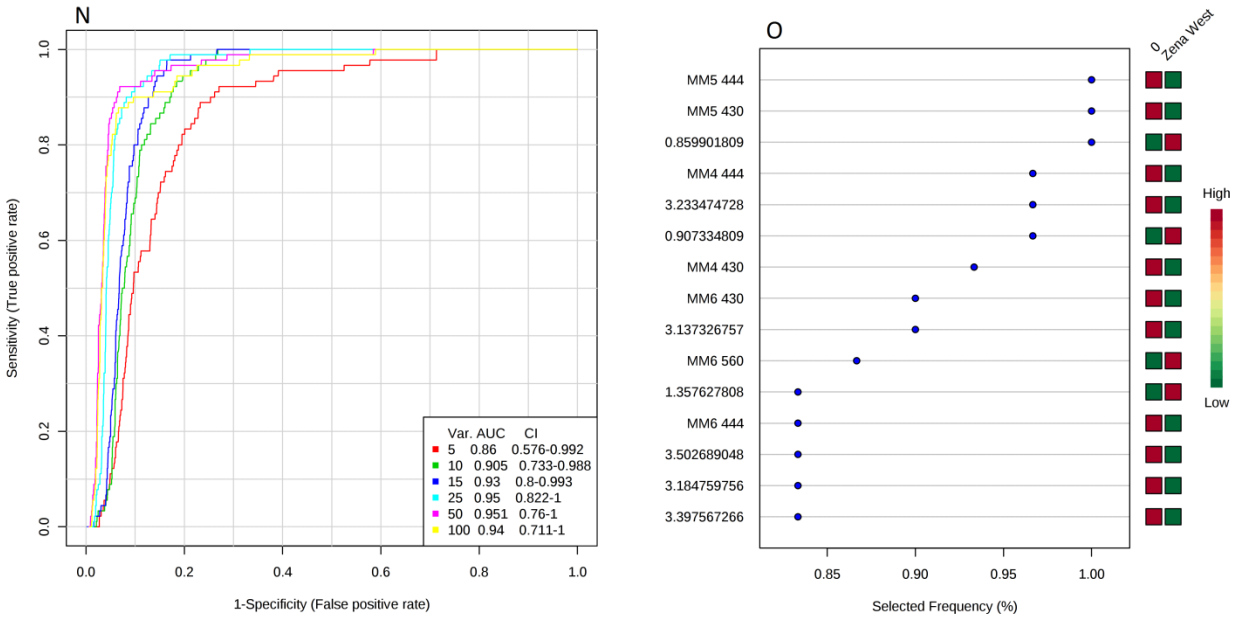



**Figure 4.19**: (**A**) The ROC curves for the Zena West wine compared against the 14 other wine samples. ROC curves were generated with MetaboAnalyst (https://www.metaboanalyst.ca/). The graph shows the ROC curves generated using 5 (orange), 10 (blue), 15 (purple), 25 (teal), 50 (red), and 500 (yellow) variables. (**B**) The top fifteen variables used to generate the ROC curves. The frequency represents how often a variable is used in the ROC curve. MM1 to MM7 represents data from the DS assay, while the numbers (ppm) are binned NMR data.

## 4.5 Conclusion

The goal of this project was to determine if a metabolomics profile can distinguish between different PN wines based on vineyard, and to determine if wines with similar environmental conditions have similar chemical profiles. Based on the PCA models and the ROC curve analysis most PN wines was distinguishable with the either the 1D $^{1}$H NMR or the DS assay data. The PCA models suggested that the NMR data provided a slight improvement over the DS assay data in differentiating between the 15 PN wines, where Panorama 5A and Radian (p-value 0.05) and Zena West and Bones (p-value 0.04) were the closest wine pairs. Surprisingly, a PCA model generated by combining the NMR and DS array data sets resulted in poor group separation. Potentially, the wine-dependent variance in the two datasets were partially anti-correlated. Also, a valid PCA

model could not be generated when the individual wines were grouped according to wine region. Suggesting a larger variance in the chemical profile of the individual wines that cannot be explain by wine region alone. The fact that all 15 PN wines, despite originating from grapes from the same clone (Dijon 667), were uniquely differentiated and could not be classified by wine region, highlights that local environment and climate are key factors for determining the chemical profile of a wine.

The ROC curves generated from the combined NMR and DS array data sets lead to variable contributions of spectral features for the characterization of individual wines. Specifically, depending on the wine, the ROC curves used different combinations of NMR and DS array spectral features. In some cases, a ROC curve was nearly exclusively defined by NMR features, while other wines were defined predominately by DS array data. There were also other cases where an equal combination of NMR and DS array features were used to define a wine. Nevertheless, in all cases the ROC curves yielded AUCs ranging from 0.80 to 0.98 indicating an accuracy of $\geq$ 80% in characterizing the PN wines. In most cases, the ROC curves required only 5 to 10 spectral features. Notably, a majority of the NMR spectral features used in the ROC curves corresponded to chemical shifts in the 0 to 4 ppm region suggesting a potential importance of organic acids and carbohydrates in differentiating the wines. Of course, the DS array data highlights the importance of phenolics to classifying different wines. Our results are consistent with some prior studies in which differences in isopentanol and isobutanol (0.9 ppm region) where observed to be key discriminators of La Rioja wine terroir [29]. Similarly, changes in tannins and other phenolic compounds have been attributed to changes in environmental conditions and grapevine vigor [30, 31].

The second part of the study was to determine if wines harvested under similar climate conditions or in the same wine region shared similar metabolomic profiles. Overall, our results demonstrated that the PN wines were impacted significantly by local variations in environment and climate, and likely other factors. In general, the PN wines could not be grouped by wine region or average climate parameters using metabolomics profiles. Instead, each PN wine yielded a unique metabolic profile. Pinot Noir wine are derived from *vitis vinifera L*. These grapes are associated with the Burgundy region of France, but are grown around the world with the exclusion of hot climates. Pinot Noir is known to take on the characteristics of the environment resulting in a distinct taste [32], which is consistent with our overall findings. Goldman *et al.* examined different varieties of red wine with 1D $^{1}$H NMR and found that PN had a 95% prediction rate, which is comparable to our findings, but high compared to the other wines [33]. Thus, environmental and climate conditions that vary between vineyard or are altered due to human action impact the metabolome of wines.

Nicholas *et al.* evaluated climate variability on PN from the Carneros and Sonoma Valley in California. Phenolic compounds and temperature were measured and correlated. Warm temperatures from budburst to bloom were found to increase phenolic content; however, cooler temperatures from the previous harvest negated this effect [31]. Reynolds *et al.* measured the effect of water stress on PN fruit maturity and vegetative growth. Exposure to reduced water resulted in a decrease in berry weight and an increase in soluble solids found in the grapes. Vegetative growth was reduced with decreases in shoot length, number, and leaf size when exposed to reduced water. This effect was also observed in grapes grown in soils that do not retain water [34]. Cortel and Kennedy measured the effect of sunlight on PN grapes by comparing the flavonoid compounds

found in shaded grapes compared to sunlight exposed grapes. Shaded grapes resulted in lower levels of flavonoids, proanthocyanidins and anthocyanins [35]. Price *et al.* measured quercetin glycosides at different levels of sun exposure and found an increase in grapes exposed to sunlight [36]. Thus, numerous factors, including activities directly controlled by humans, impact how grapes grow, and consequently, the chemical composition of the resulting wine. Overall, this study exposed the complexities of terroir, its impact on metabolic profile of wine, and its utility to accurately characterize PN wines derived from the same PN clone (Dijon 667).

**4.6 References**

1. Markoski, M.M., et al., Molecular Properties of Red Wine Compounds and Cardiometabolic Benefits. Nutr Metab Insights, 2016. 9: p. 51-7.

2. Waterhouse, A.L., Wine Phenolics. Annals of the New York Academy of Sciences, 2002. 957(1): p. 21-36.

3. Zamora, F., Biochemistry of Alcoholic Fermentation, in Wine Chemistry and Biochemistry, M.V. Moreno-Arribas and M.C. Polo, Editors. 2009, Springer New York: New York, NY. p. 3-26.

4. Smart, R.E., et al., Canopy Management to Improve Grape Yield and Wine Quality - Principles and Practices. South African Journal of Enology and Viticulture; Vol 11, No 1 (1990), 2017.

5. Kennedy, J.A., C. Saucier, and Y. Glories, Grape and Wine Phenolics: History and Perspective. American Journal of Enology and Viticulture, 2006. 57(3): p. 239-248.

6. Proestos, C., D. Sereli, and M. Komaitis, Determination of phenolic compounds in aromatic plants by RP-HPLC and GC-MS. Food Chemistry, 2006. 95(1): p. 44-52.

7. Pretorius, I.S., Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking. Yeast, 2000. 16(8): p. 675-729.

8. Kekelidze, I., et al., Phenolic antioxidants in red dessert wine produced with innovative technology. Annals of Agrarian Science, 2018. 16(1): p. 34-38.

9. Sobolev, A.P., et al., Molecular fingerprinting of food authenticity. Current Opinion in Food Science, 2017. 16: p. 59-66.

10. Vaudour, E., The Quality of Grapes and Wine in Relation to Geography: Notions of Terroir at Various Scales. Journal of Wine Research, 2002. 13(2): p. 117-141.

11. Bartowsky, E.J. and P.A. Henschke, Acetic acid bacteria spoilage of bottled red wine—A review. International Journal of Food Microbiology, 2008. 125(1): p. 60-70.

12. Bridi, R., et al., Monitoring peroxides generation during model wine fermentation by FOX-1 assay. Food Chem, 2015. 175: p. 25-8.

13. Amargianitaki, M. and A. Spyros, NMR-based metabolomics in wine quality control and authentication. Chemical and Biological Technologies in Agriculture, 2017. 4(1): p. 9.

14. Revelette, M.R., J.A. Barak, and J.A. Kennedy, High-performance liquid chromatography determination of red wine tannin stickiness. J Agric Food Chem, 2014. 62(28): p. 6626-31.

15. Pereira, G.E., et al., 1H NMR and Chemometrics To Characterize Mature Grape Berries in Four Wine-Growing Areas in Bordeaux, France. Journal of Agricultural and Food Chemistry, 2005. 53(16): p. 6382-6389.

16. Son, H.-S., et al., 1H NMR-Based Metabolomic Approach for Understanding the Fermentation Behaviors of Wine Yeast Strains. Analytical Chemistry, 2009. 81(3): p. 1137-1145.

17. Nilsson, M., et al., High-Resolution NMR and Diffusion-Ordered Spectroscopy of Port Wine. Journal of Agricultural and Food Chemistry, 2004. 52(12): p. 3736-3743.

18. Umali, A.P., et al., Discrimination of flavonoids and red wine varietals by arrays of differential peptidic sensors. Chemical Science, 2011. 2(3): p. 439-445.

19. Ghanem, E., et al., Predicting the composition of red wine blends using an array of multicomponent Peptide-based sensors. Molecules, 2015. 20(5): p. 9170-82.

20. Worley, B. and R. Powers, MVAPACK: A Complete Data Handling Package for NMR Metabolomics. ACS Chemical Biology, 2014. 9(5): p. 1138-1144.

21. Worley, B., S. Halouska, and R. Powers, Utilities for quantifying separation in PCA/PLS-DA scores plots. Anal Biochem, 2013. 433(2): p. 102-4.

22. Werth, M.T., et al., Analysis of metabolomic PCA data using tree diagrams. Anal Biochem, 2010. 399(1): p. 58-63.

23. Xia, J., et al., MetaboAnalyst 3.0--making metabolomics more meaningful. Nucleic Acids Res, 2015. 43(W1): p. W251-7.

24. Bhinderwala, F., et al., Combining Mass Spectrometry and NMR Improves Metabolite Detection and Annotation. Journal of Proteome Research, 2018. 17(11): p. 4017-4022.

25. Neumann, P.A. and A. Matzarakis, Estimation of wine characteristics using a modified Heliothermal Index in Baden-Wurttemberg, SW Germany. Int J Biometeorol, 2014. 58(3): p. 407-15.

26. Hall, A. and G.V. Jones, Spatial analysis of climate in winegrape-growing regions in Australia. Australian Journal of Grape and Wine Research, 2010. 16(3): p. 389-404.

27. Johnson, S.R. and B.M. Lange, Open-access metabolomics databases for natural product research: present capabilities and future potential. Frontiers in bioengineering and biotechnology, 2015. 3: p. 22-22.

28. Swan, F. What You Should Know about Oregon's Willamette Valley AVA. 2018 [cited 2020 2020/05/19]; Available from: https://www.jjbuckley.com/wine-knowledge/blog/what-you-should-know-about-oregons-willamette-valley-ava/888.

29. López-Rituerto, E., et al., Investigations of La Rioja Terroir for Wine Production Using 1H NMR Metabolomics. Journal of Agricultural and Food Chemistry, 2012. 60(13): p. 3452-3461.

30. Song, J., et al., Pinot Noir wine composition from different vine vigour zones classified by remote imaging technology. Food Chemistry, 2014. 153: p. 52-59.

31. Nicholas, K.A., et al., Effect of vineyard-scale climate variability on Pinot noir phenolic composition. Agricultural and Forest Meteorology, 2011. 151(12): p. 1556-1567.

32. Robinson, J., The Oxford companion to wine. 1994, Oxford ; New York: Oxford University Press. xvi, 1088 p.

33. Goldman, S.M., Environmental toxins and Parkinson's disease. Annu Rev Pharmacol Toxicol, 2014. 54: p. 141-64.

34. Andrew, G.R. and P.N. Andrew, `Pinot noir' and `Riesling' Grapevines Respond to Water Stress Duration and Soil Water-holding Capacity. HortScience HortSci, 1994. 29(12): p. 1505-1510.

35. Cortell, J.M. and J.A. Kennedy, Effect of Shading on Accumulation of Flavonoid Compounds in (Vitis vinifera L.) Pinot Noir Fruit and Extraction in a Model System. Journal of Agricultural and Food Chemistry, 2006. 54(22): p. 8510-8520.

36. Price, S.F., et al., Cluster Sun Exposure and Quercetin in Pinot noir Grapes and Wine. American Journal of Enology and Viticulture, 1995. 46(2): p. 187.

# Chapter 5

## 5. Summary and Conclusion

### 5.1 Summary of Work

Metabolomics has seen applications in human diseases, plant genomics, and toxicology, among numerous other research areas [1-3]. In fact, metabolomics is expanding into other areas of investigation, such as the food industry. As a growing field, there is a strong need for standardized methodologies for preparing metabolomics samples, conducting experiments, and analyzing analytical data sets. As such, it is important to consider how samples are properly handled, and to evaluate protocols for data processing, statistical modeling and the identification of metabolites. Metabolites are unstable, and are often prone to oxidation or other forms of modification [4]. Therefore, the proper storage and transport of metabolomics samples is an important point to consider. Metabolomics experiments generate large volumes of data, but commonly have limited number of biological replicates. As a result, statistical and network analysis may be prone to overfitting and over-interpretation. To address these and other issues, my thesis highlights the development of metabolomics procedures and the application of metabolomics to issues related to human health and food integrity.

Chapter 2 provides an exhaustive and detailed description of metabolomics protocols for the investigation of Parkinson's Disease (PD) and other neurological disorders. The described protocols include methods to extract, quantify, and process metabolomic samples obtained from mammalian cells and brain tissues [5]. Proper and complete metabolite extraction from a biological sample is a critical step of the entire metabolomics protocol since it determines how much of the

metabolome is available for study [6]. In this regard, Chapter 2 focuses on the extraction of water-soluble metabolites from cell lysis and the extracellular space or culture medium. Water soluble metabolites include amino acids, sugars, and other organic acids. In addition to sample preparation, Chapter 2 includes details regarding standard data collection for NMR and mass spectrometry, and liquid chromatography. Typical statistical methods for the analysis of metabolomics data that include principal component analysis and orthogonal projection to latent structures were also described. Importantly, the described protocols include methods for validating multivariate statistical models, such as permutation testing. Permutation testing validates supervised statistical models by generating multiple replicate models using different subsets of the data while scrambling group classification [7]. Furthermore, the importance of multiple hypothesis testing, like Bonferroni and Benjamini-Hochberg, which help establish significance by controlling false discovery rates, are also highlighted [8, 9].

Although the metabolomics procedures described in Chapter 2 can be applied to a variety of metabolomic studies, the protocols were focused on investigating PD using neuroblastoma cells and brain tissue. The complex heterogeneous nature of the brain makes metabolomics a useful tool for providing an overview of cellular processes. Also, metabolomic studies are particularly useful for neurological studies where regular access to tissues is difficult. Typical PD studies expose animal models to various conditions to replicate PD symptoms or exposure risks. Similarly, cellular models use neuroblastoma cells to reproduce PD changes and to mimic PD risk factors [10]. Evidence suggests that the interactions between environment and genetics play an important role in PD development [11]. Thus, animal or tissue models exposed to various environmental

stressors are useful for investigating PD, other neurodegenerative diseases, or for monitoring brain or neuronal health at different points in life.

Chapter 3 describes our investigation into astrocytes response to the neurotoxin arsenic. Importantly, the outcome of this study may provide insights into how arsenic exposure may relate to neurodevelopmental disorders in children. The role of astrocytes in the brain is a complex one. Typically, astrocytes play a supportive role by providing neurotransmitter precursors to neurons and by maintaining neuronal synapses [12]. However, astrocytes have been shown to respond to immune or minor trauma in both supportive and harmful ways [13]. Astrocytes are known to be a major source of glutathione, which protects the brain from oxidative stress and xenobiotics [14]. Astrocytes may also play a role in protecting neurons from bystander death (*i.e.*, the death of cells not directly impacted by an injury, toxin or radiation). Glial cells and neurons co-cultured with astrocytes were protected from arsenic treatments [15].

In our metabolomics analysis of astrocytes exposed to arsenic, we observed an upregulation in antioxidant production, which is possibly due to an upregulation of glycolysis and pyruvate carboxylase. The production of glutathione (GSH) is dependent on glucose metabolism; therefore, we focused our metabolomics analysis on the effect of arsenic on glucose metabolism. We observed an increase in glutamate, a potentially neurotoxic neurotransmitter in the extracellular space. Glutamate has been shown to upregulate glycolysis in astrocytes through activation of the Na+ dependent uptake system [16]. We also observed a decrease in the export of lactate and citrate by astrocytes. Lactate is a product of glycolysis and is derived from pyruvate; whereas, citrate is

an intermediate of the tricarboxylic acid (TCA) cycle. Thus, lactate and citrate are known to be energy substrates by entering the TCA cycle to produce ATP. However, recent evidence suggest lactate and citrate may also play a role in modulating receptors like glutamate. Lactate has been shown to modulate receptors linked to memory [17]. Similarly, citrate may modulate the glutamate receptor MNDA by chelating $Zn^{2+}$. Overall, our analysis of astrocytes exposed to arsenic demonstrated a rapid metabolomics response to combat the oxidative damage induced by arsenic.

Metabolomics is valuable approach for evaluating the impact of a wide variety of environmental stressors. Accordingly, metabolomics has been introduced into other areas of research, which includes the food industry. The environment, handling and processing procedures may impact the production of both food and beverages. Metabolomics can be used to produce a chemical profile of a consumable, which allows food scientists to examine the impact of production, transport, and storage on the food product. Wine is a prime example of a chemically complex food with metabolites originating from plants, yeast fermentation, and sample aging [18]. Profiling is where the chemical composition of the wine is measured. This is a useful tool for evaluating the nutrition and health properties of a wine. For high value items like wine, chemical profiles are often used to authenticate the wine's origin for quality purposes. Chapter 4 examined Pinot Noir wines produced from the same scion clone (Pinot noir 667) grown in vineyards across multiple California and Oregon wine regions. Metabolite profiles were built from a combination of untargeted 1D $^{1}H$ NMR spectral data and a targeted differential sensing array data. The profiles were used to evaluate the metabolic differences between the Pinot Noir wines produced by 15 vineyards. A major outcome of the study was the observation that each method has its own advantages and limitations in regard to classifying the wines. NMR captured a wide variety of molecules, such as carbohydrates and

organic acids. Conversely, the DS assay was only used to detect phenolic compounds [19, 20]. The 1D $^{1}$H NMR and DS array data were evaluated individually and as a combined data set. A PCA model produced from the NMR data set provided the best separation of the wines based on vineyard of origin. Conversely, one versus all ROC curves were generated from the combined data set. All of the ROC curves yielded a predictive accuracy of > 80% in distinguishing one wine from the set of Pinot Noir wines. Notably, each resulting ROC curve used a different combination of NMR and DS array spectral features to classify each wine. In some cases, the ROC curve was dominated by either NMR or DS array features. In other cases, the ROC curve used an equal mixture of both. Clearly, combining analytical techniques improved the overall wine classification accuracy. Notably, the wines could not be consistently clustered by either wine region or by average climate data. Thus, wine metabolic profiles are predominantly impacted by the local environment (*i.e.*, terrior), by the handling of the grapes/wine or by the fermentation process.

### 5.2 Future Direction

Metabolomics is a growing field and is constantly developing and refining techniques. As the application of metabolomics continues to expand into new fields, it is important to consider how the experimental design and the protocols will affect the desired results. As shown in chapter 4, the application of different instrumentation or analytical methods will dictate the results or outcomes of a given study. Simply, NMR and the DS array detected a completely different set of metabolites and provided unique, but distinct views of each wine sample. In this regard, the analytical method will determine which metabolites are detected. It will also determine how the samples are prepared and how much sample is needed. Which, in turn, will impact the experimental

time, the metabolome extraction protocol, and numerous other experimental parameters. These are not trivial concerns. Following proper protocols reduces systemic error and exposes more of the metabolome for proper characterization. Chapter 2 described the sample preparation protocol for mammalian cells, but extraction and other procedures will vary based on the sample and the analytical instrument. Thus, as the field of metabolomics continues to evolve, what is critical to its future success is the establishment of a set of standardized protocols. Challengingly, these protocols are likely to be sample and/or study specific.

NMR-based metabolomics often relies on the inclusion of a $^{13}C$-labeled metabolite like $^{13}C_6$-glucose to enhance the sensitivity of the NMR experiment and the detection of metabolites [21]. Many biological and cellular reactions are independent of carbon movement and are therefore difficult to monitor by only following central carbon metabolism. For example, many phosphorylated compounds, which are involved in energy metabolism or regulate signaling pathways, may be missed by relying on only $^{1}H$ -$^{13}C$-NMR [22]. Instead, $^{31}P$ and other nuclei NMR experiments will be needed to expand the coverage of the metabolome by NMR.

The application of metabolomics to clinical samples has further highlighted the important differences between humans and animal models. Metabolomics may be valuable in breaching these differences to facilitate drug discovery and disease diagnosis. For example, metabolomics may be able to correlate disease relevant biomarkers across multiple platforms, from cell system, to animal model, and then to the human patient. Identifying biomarkers for neurological disorders like PD would allow for early diagnosis. An early diagnosis and intervention for PD has been shown to

slow down the progression of symptoms. Leveraging metabolomics to identify risk factors for PD may reduce exposure and prevent the disease. Metabolomics can make equally beneficial contributions to the food industry. Metabolomics may be used to evaluate multiple aspects of the food cycle - from growing food to its efficient processing. In this regard, metabolomics can be used to improve anything from nutritional value to taste.

## 5.3 References

1. Robertson, D.G., P.B. Watkins, and M.D. Reily, Metabolomics in Toxicology: Preclinical and Clinical Applications. Toxicological Sciences, 2010. 120(suppl_1): p. S146-S170.

2. Emwas, A.-H.M., et al., NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. Metabolomics, 2013. 9(5): p. 1048-1072.

3. Sumner, L.W., P. Mendes, and R.A. Dixon, Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry, 2003. 62(6): p. 817-36.

4. Clark, S., et al., Stability of plasma analytes after delayed separation of whole blood: implications for epidemiological studies. International Journal of Epidemiology, 2003. 32(1): p. 125-130.

5. F. Bhinderwala, S.L., J. Woods, J. Rose, D. D. Marshall, F. Bhinderwala, E. Riekeberg, A. De Lima Leite, M. Morton, E. D. Dodds, R. Franco, and R. Powers, In Metabolomics. Methods in Molecular Biology. 2019.

6. Mushtaq, M.Y., et al., Extraction for metabolomics: access to the metabolome. Phytochem Anal, 2014. 25(4): p. 291-306.

7. Triba, M.N., et al., PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters. Molecular BioSystems, 2015. 11(1): p. 13-19.

8. Benjamini, Y. and Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 1995. 57(1): p. 289-300.

9. Bland, J.M. and D.G. Altman, Multiple significance tests: the Bonferroni method. BMJ, 1995. 310(6973): p. 170.

10. Cannon, J.R. and J.T. Greenamyre, Gene-environment interactions in Parkinson's disease: specific evidence in humans and mammalian models. Neurobiol Dis, 2013. 57: p. 38-46.

11. Falkenburger, B.H., T. Saridaki, and E. Dinter, Cellular models for Parkinson's disease. J Neurochem, 2016. 139 Suppl 1: p. 121-130.

12. Allen, N.J. and C. Eroglu, Cell Biology of Astrocyte-Synapse Interactions. Neuron, 2017. 96(3): p. 697-708.

13. Liddelow, S.A. and B.A. Barres, Reactive Astrocytes: Production, Function, and Therapeutic Potential. Immunity, 2017. 46(6): p. 957-967.

14. Wilson, J.X., Antioxidant defense of the brain: a role for astrocytes. Can J Physiol Pharmacol, 1997. 75(10-11): p. 1149-63.

15. Singh, V., et al., Hijacking microglial glutathione by inorganic arsenic impels bystander death of immature neurons through extracellular cystine/glutamate imbalance. Scientific Reports, 2016. 6(1): p. 30601.

16. Pellerin, L. and P.J. Magistretti, Glutamate uptake into astrocytes stimulates aerobic glycolysis: a mechanism coupling neuronal activity to glucose utilization. Proc Natl Acad Sci U S A, 1994. 91(22): p. 10625-9.

17. Alberini, C.M., et al., Astrocyte glycogen and lactate: New insights into learning and memory mechanisms. Glia, 2018. 66(6): p. 1244-1262.

18. Robinson, J., The Oxford companion to wine. 1994, Oxford ; New York: Oxford University Press. xvi, 1088 p.

19. McCloskey, L.P., An Acetic Acid Assay for Wine Using Enzymes. American Journal of Enology and Viticulture, 1976. 27(4): p. 176-180.

20. Mercurio, M.D., et al., High Throughput Analysis of Red Wine and Grape PhenolicsAdaptation and Validation of Methyl Cellulose Precipitable Tannin Assay and Modified Somers Color Assay to a Rapid 96 Well Plate Format. Journal of Agricultural and Food Chemistry, 2007. 55(12): p. 4651-4657.

21. Clendinen, C.S., et al., An overview of methods using 13C for improved compound identification in metabolomics and natural products. Frontiers in Plant Science, 2015. 6(611).

22. Duboc, D., et al., Phosphorus NMR spectroscopy study of muscular enzyme deficiencies involving glycogenolysis and glycolysis. Neurology, 1987. 37(4): p. 663-663.