Computational technique development in NMR-based metabolomic analysis and applications on the study of metabolomic alteration in pregnancy under normal and challenged conditions

by

Sicong Zhang

(Under the Direction of Arthur S. Edison)

Abstract

Metabolomics is the study of small molecules in a biological system. It has aided in research on characterizing health conditions, screening out important pathways, and disentangling associations between metabolites and pathways. Because of the diversity and complexity of metabolites, data processing and analysis are important steps for metabolomics research. Nuclear magnetic resonance (NMR) spectroscopy is one of the most powerful analytical approaches to metabolomics. However, the difference in sample properties such as pH may cause great variations in chemical shifts of some NMR resonances. Although efforts have been made to control this variation, this phenomenon still challenges NMR-based metabolomics research. In addition, metabolomics data can be integrated with other types of data for systematic analysis, but techniques for data integration have not been well explored yet. Therefore, there is a demand for advancing techniques for NMR-based metabolomic analysis.

In this dissertation, I demonstrated two novel computational tools I developed for NMR-based untargeted metabolomics. These include a data integration technique for analyzing relationships between NMR-measured metabolomics, MS-measured glycomics, and two-dimensional flow-cytometric data; and a spectral alignment algorithm, named pHIT, for processing NMR peaks with high chemical shift variations.

By applying these tools to the study of *C. elegans* development, and pregnancy under normal and challenged conditions, I discovered new biological phenomena. I identified development-associated metabolite-glycan correlations in *C. elegans*. I analyzed urine metabolome from pregnant mothers with virus infection and found disturbed metabolites and pathways. I then used a sheep model to further investigate fetal and neonatal metabolic change with maternal stressed conditions as well as their metabolic transitions after birth. I identified preterm-birth-associated metabolic change with maternal chronic cortisol treatment and altered metabolites and pathways after birth in neonates.

My work is expected to contribute to NMR-based metabolomics analysis, as well as the understanding of metabolic change in animal development and gestation.

INDEX WORDS: [Metabolomics, NMR, spectral alignment, data integration, pregnancy]

Computational technique development in NMR-based metabolomic analysis and applications on the study of metabolomic alteration in pregnancy under normal and challenged conditions

by

Sicong Zhang

B.S., China Agricultural University, China, 2016

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

DOCTOR of PHILOSOPHY

Athens, Georgia

2022

Computational technique development in NMR-based metabolomic analysis and applications on the study of metabolomic alteration in pregnancy under normal and challenged conditions

by

Sicong Zhang

Major Professor: Arthur S. Edison

Committee: Jonathan Arnold
Maria B. Cassera
Ted M. Ross
Ying Xu

Electronic Version Approved:

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

Dedication

To my family.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Arthur S. Edison for leading me to this metabolomics field. I would not have been a bioinformatics person if it was not him kept convincing me that I can do computational work. I would also thank him for inspiring me on research ideas and providing guidance when I trapped myself in some specific details. His brave and sincere personality has impacted me towards a more optimistic attitude.

I would thank my advisory committee members, Dr. Jonathan Arnold, Dr. Maria Belen Cassera, Dr. Ted M. Ross, and Dr. Ying Xu for providing valuable suggestions and feedback on my research and career development. I also want to thank my department and UGA graduate school for providing a variety of resources and opportunities for me to develop my career and scientific skills.

I would like to thank all my lab mates. They are always supportive academically and mentally. I would specially thank Karen S. Howard, who is always helping me out from the stuff that I feel most stressed about. I would like to thank all my collaborators, especially Dorothy Ellis, Dr. Susmita Datta, and Dr. Maureen Keller-Wood for their help and advice. I have learned a lot from them.

Finally, I would like to thank my family and friends for keeping supporting me during my Ph.D. journey. I would thank my mother Jinsong Zhang, my father Ying Zhang, my grandparents Kun Yang and Zhensheng Zhang for always caring about my happiness and encouraging me to achieve my goal. I would specially thank my sister Sipeng Zhang for helping me without any hesitation in my tough time. Last but not least, I thank my friends for sharing their happiness, and anxiety with me. This made us all grow stronger and stronger.

Contents

Chapter 1
INTRODUCTION AND LITERATURE REVIEW

1.1 Overview of Metabolomics and NMR

1.1.1 Metabolomics

In biological systems, besides the widely investigated macromolecules DNA, RNA and proteins, small molecules such as carbohydrates, amino acids and organic acids are also important as they are involved in a variety of reactions and functions. Small molecules, usually defined < 1.5 kDa, are collectively called metabolites.

Metabolites are regulated by an organism's genome and may have direct contact with their environment. They are, therefore, considered to be the closest characteristics to phenotypes or are even phenotypes themselves [1].

For centuries, metabolites have been used to diagnose disease and to aid in our understanding of biological systems. For example, it can be traced back to the 5th/6th century BC when people used urine sugar levels to diagnose diabetes. However, metabolic levels can be very dynamic and are subject to change by both genetic and environmental factors. This makes understanding change in a single metabolite difficult. With the development of 'omics' research, we are now able to study metabolites at a systems level.

The set of small molecules in a biological system is called the metabolome, and the study of the metabolome is defined as metabolomics. Metabolomics has given us the power to characterize health conditions by a group of metabolites [2], [3], [4], screen out important pathways for certain conditions [4], and to disentangle associations between metabolites and pathways [5].

Metabolomics strategies generally can be divided into two categories: targeted and untargeted approaches. Targeted metabolomics, by definition, studies a target list of metabolites and aims to verify specific hypotheses. On the other hand, untargeted metabolomics profiles all detectable metabolites in a biological system, therefore generating hypotheses. Targeted approaches can be very accurate in quantification, but they also require more established knowledge. Untargeted approaches are advantageous as they are not biased by previous findings, but they can be harder to be biologically interpreted. One

big challenge hindering interpretation in untargeted metabolomics is to annotate signals from analytical chemistry measurements to metabolites, or to match the same signal of the same metabolite across different samples. Another challenge is to find the connections between measured metabolites since they are not chosen because of a known pathway.

1.1.2 Nuclear Magnetic Resonance Spectroscopy

When measuring a metabolome, the most widely used techniques are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) [6]. Other techniques include, but are not limited to, Raman, infrared, and ultraviolet-visible spectroscopies [7]. NMR measures the intensity of signals from nuclei that do not have even numbers of both protons and neutrons [8]. Commonly used nuclides for the NMR analysis of biological samples include ^{1}H, ^{13}C, ^{15}N, and ^{31}P.

The signal resonance of a nucleus depends on the magnetic field strength it is in. Besides the static magnetic field where the sample is placed, electrons can also affect the magnetic field strength around a nucleus within a certain spatial range. Because electrons are charged and moving in the orbitals around the nucleus, they produce a magnetic field that slightly changes the magnetic field strength of the nucleus. Electrons in different chemical groups have different densities; nuclei in these chemical groups, therefore, are under different magnetic field strengths. These nuclei thus resonant at different frequencies. The difference in resonances caused by this phenomenon is called chemical shift. Usually, signal frequencies are divided by the static magnetic field strength to make the results measured under different field strength comparable, so the unit of chemical shift is parts-per-million (ppm) [8]. This property makes metabolites in a mixture distinguishable from each other on NMR spectra. The intensity of an NMR signal is proportional to the amount of the corresponding nucleus so it can be used for quantitative measurements.

NMR has good analytical reproducibility for metabolomics samples since reported technical variations are generally <10% [9], [10] and the instrument may only have minimal impact on the results [10], [11]. This establishes a solid foundation for statistical analysis when examining metabolomic samples. For example, we can use the correlation between signals on the same set of samples to find nuclei from the same metabolite, thus facilitating signal annotation. This technique is called Statistical TOtal Correlation SpectroscopY (STOCSY) [12]. Along with its technical stability, NMR measurements also have the advantage of being able to reflect structural information of compounds. This information can be used for identifying unknown compounds and distinguishing structural isomers such as glucose and fructose. Practically, sample preparation for NMR-based metabolomics is relatively easy. For example, aqueous samples such as urine and serum may not need extraction and can be measured with a single centrifugation-and-buffering step [13]. High-resolution magic-angle-spinning (HRMAS) NMR can be used to collect NMR spectra on intact tissue samples or on intact small organisms [14], [6], [15]. NMR measurements are non-destructive, thus NMR-measured samples can be re-used for other analytical techniques [16]. This feature makes multi-analytical-platform integration possible on the same sample. NMR-measured metabolomics also has good compound coverage on major chemical superclasses [16], thus reducing potential bias for untargeted analysis.

However, NMR also has its own limitations. Compared to MS, the detection sensitivity of NMR is lower. NMR sensitivity is usually at μM levels, while MS can reach nM levels. This results in a total number of features detected by NMR on the 1000s scale, while MS detects on the 100000s scale.

On the other hand, although chemical shifts can be used for differentiating metabolites, certain peaks can also be affected by sample matrix effects [17]. Therefore, peaks from the same nucleus may not be at the same chemical shift from different samples. This phenomenon can impede peak annotations because the chemical shift from the sample may not match the frequency in the database. The displacement of signals also means peaks on the same chemical shift may not come from the same metabolite, which is unfavorable for statistical analysis. Therefore, efforts need to be made to control for this chemical shift variation.

1.2 Recent Advances in NMR Chemoinformatics

The phenomena discussed above is an example showing the necessity of NMR data processing in metabolomics. Computational techniques have been developed to facilitate the extraction of information from NMR data for statistical analysis and spectral annotation. The following section discusses recent chemoinformatic advances in NMR-based metabolomics. It is part of a review I published with Arthur S. Edison, Maxwell Colonna, Goncalo J. Gouveia, Nicole R. Holderman, Michael T. Judge, and Xunan Shen in Analytical Chemistry entitled 'NMR: Unique Strengths That Enhance Modern Metabolomics Research' [6]. I researched and wrote the section that is reprinted here in sections 1.2.1-1.2.4 and also adpated Figures 1.1 below. All are included with permission from the publisher.

1.2.1 Spectral Processing

The physical and chemical properties of a sample can affect the chemical shifts of some NMR peaks. This makes it hard to compare peaks across spectra, so alignment and/or division of NMR spectra into smaller regions (binning) is usually applied to manage this problem.[18] However, results from this step are not always optimal, especially in complicated samples. Takis et al. used modeling strategies for this problem by considering the chemical shift of a signal as the function of a mixture's total chemical composition, pH and temperature.[19] They built a model including sample pH, temperature, concentrations of 11 ions, and chemical shifts and concentrations of 40 abundant metabolites to estimate chemical shifts of these metabolites on 4000 artificial urine samples. The algorithm begins by matching five navigating signals, then exports estimations of chemical shifts and concentrations of the targeted metabolites and ions. The algorithm demonstrated high predictive accuracy in real urine samples. It also deconvoluted overlapped peaks, thus improving annotation and quantification.

As another alternative to the "align and/or bin" strategy, "speaq 2.0" used wavelets to extract features from raw spectra.[20] In this method, Mexican hat wavelets were used for peak picking because they are robust to baseline distortions. Picked peaks were grouped for signals from the same nuclei across different spectra. In contrast to using peak integrals for quantifications, wavelet coefficients were used

here to represent the abundance of picked peaks for later analysis. This method showed tolerance to small chemical shift variations and could effectively extract features from simulated and published dataset.

Due to the abundance of signals, peaks are often overlapped in a 1D ^{1}H metabolomics spectrum. 2D NMR experiments could better separate overlapped signals, but due to long acquisition times, they are usually used only for peak annotation or on a small sample set. With the development of fast 2D NMR experiments,[21], [22] 2D spectra have the potential to be used for relative quantification. Therefore, tools for quantifying 2D peaks have recently been improved. Two-dimensional spectra can be vectorized[23] or projected on one dimension[24] to suit both 1D spectra processing methods and statistical analysis methods. For example, the projection of JRES spectra on the chemical shift dimension (pJRES) can be binned by JBA (pJRES Binning Algorithm).[24] JBA extends the concept of statistical recoupling of variables[25] by using the collinearity of adjacent points to help define bin boundaries and is able to retain small signals more efficiently.

Two-dimensional peaks can also be binned[26], [27] or line-shape fitted[28] directly for quantification and matrix size reduction. While non-uniform binning can better quantify peaks than uniform binning, the non-uniform binning algorithms are under-developed for 2D spectra. The binning step in HATS-PR (Hierarchical Alignment of Two-dimensional Spectra-Pattern Recognition) can adjust bins by combining multiplets and extending uniform bins to the next bin or to the maximum user defined length.[29] A more flexible multidimensional binning algorithm, Generalized Adaptive Intelligent (GAI) binning, was recently proposed.[26] It extended adaptive intelligent binning from one-dimensional to multi-dimensional data so that 2D spectra could be binned with flexible bin sizes automatically.

1.2.2 Extracting Information for Peak Annotations

Grouping signals from the same metabolites for database matching can improve the accuracy of annotation. A spectrum from a pure compound can be directly queried and matched in the database. However, for mixtures, peaks from the same metabolite need to be found before querying. Because signals from the same compound should be highly linearly correlated with each other, Statistical TOtal Correlation SpectroscopY (STOCSY) uses Pearson correlation coefficients to gather signals from the same metabolite and builds a pseudospectrum for database query.[30] However, STOCSY performance may be compromised where peaks overlap. JRES can efficiently reduce overlap by separating chemical shifts and multiplicities to two dimensions, but JRES databases are limited and difficult for peak matching.[31] The Hoijemberg group introduced two strategies to circumvent this challenge by querying peaks from the projection of JRES for 1D databases (Figure 1.1).[31], [32] The first strategy (Figure 1.1A) uses projection of STOCSY traces from tilted and symmetrized JRES (p-(JRES-STOCSY)) as pseudospectra.[31] Because projection of tilted and symmetrized JRES (pJRES) spectra differ from 1D spectra in terms of multiplicity, pJRES spectra cannot be matched directly to 1D databases. Therefore, they built a library (Chemical Shift Multiplet Database) using curated pJRES spectra and their traces on the J-coupling dimension obtained from the Birmingham Metabolite Library.[33] They also built a tool for querying this database. This tool allows for repeated use and includes correlated but small peaks in the query list to avoid false-negative matching. Correlated but unmatched peaks can also be queried against this database for unravelling biological asso-

ciations. The second strategy (Figure 1.1B) uses projection of STOCSY results on non-tilted JRES spectra (p-(ntJRES-STOCSY))[32] instead of STOCSY on p-ntJRES ((p-ntJRES)-STOCSY)[34] to mimic 1D spectra. Therefore, the deconvoluting power of JRES is preserved.

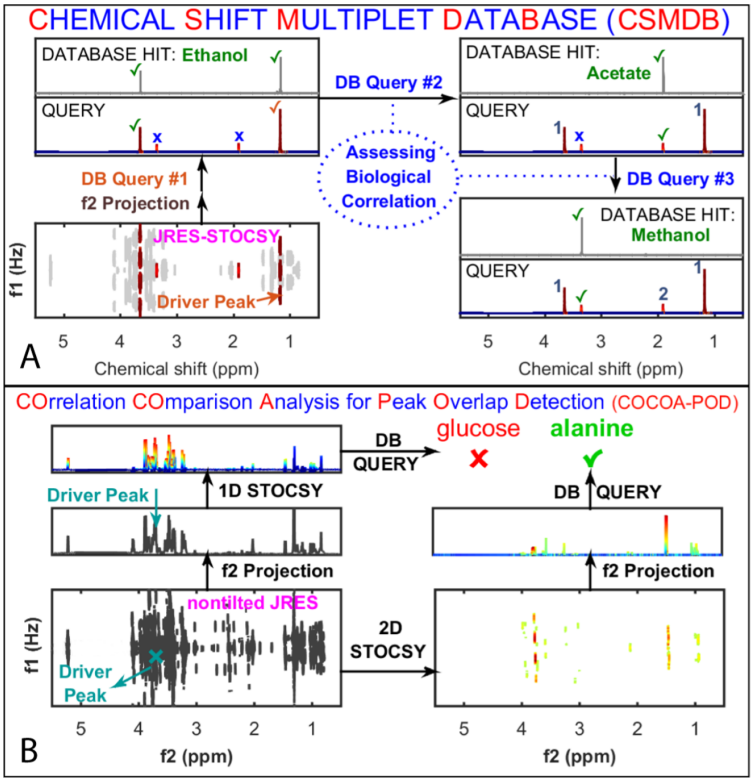


Figure 1.1: Workflows of using projection of STOCSY traces from tilted and symmetrized JRES spectra (A) and non-tilted JRES spectra (B) for database query. (A) Example of consecutively querying Chemical Shift Multiplet Database with p-(JRES-STOCSY) on driver peak at 1.181 ppm on tilted and symmetrized JRES spectra. (B) Example of comparing results of querying database with (p-ntJRES)-STOCSY and p-(ntJRES-STOCSY) on the same driver peak at 3.783 ppm. (A) Adapted with permission from Charris-Molina, A.; Riquelme, G.; Burdisso, P.; Hoijemberg, P. A. J. Proteome Res. 2020, 19(8), 2977-2988 (ref [31]). Copyright 2020 American Chemical Society. (B) Adapted with permission from Charris-Molina, A.; Riquelme, G.; Burdisso, P.; Hoijemberg, P. A. J. Proteome Res. 2019, 18(5), 2241-2253 (ref [32]). Copyright 2019 American Chemical Society.

Although including overlapped peaks in a query list may introduce irrelevant metabolites in the query result, simply discarding the overlapped peak also runs the risk of increasing the false-negative rate. PODCAST (Peak Overlap Detection by Clustering Analysis and Sorting of Traces) took overlapping information from the clustering of all STOCSY traces, which came from each peak used as a driver peak, to complement the peak list for database query.[35]

These computational techniques could be employed with other spectroscopical or physical separating techniques to enhance the efficiency of peak annotation. In a recent approach to compound identification in NMR metabolomics, the Nicholson group proposed a system to sequentially use computational and

experimental annotation strategies for a metabolomics study.[36] This system was shown to efficiently reduce manual input and improve annotation accuracy thus is expected to be generally utilized in the future.

Database coverage also limits NMR peak annotation. With the development of computational biology, structural information can now be predicted from chemical shifts by machine learning at the motif[37],molecular[38], and compound family[39] level. In addition, the 1D spectrum from one metabolite can be simulated under any magnetic field strength according to its spin parametrized system matrix in the GISSMO (Guided Ideographic Spin System Model Optimization) library.[40] These approaches are not only helpful for peak annotations but they also enable the enhancement of current databases with additional putative reference spectra.

1.2.3 Workflows

Owing to the diversity of tools available for NMR spectral analysis, researchers ususally assemble their own workflows to record the functions and parameters they use. There are several general tools which serve as pipelines that include widely used steps and methods, such as NMRProcFlow,[41] ASICS R[42] and Chenomx (Chenomx, Edmonton, Canada). Recently published pipelines include AlpsNMR[43] and PepsNMR[44] for preprocessing, Lipspin[45] for lipids profiling, InterSpin[46] for low-resolution NMR (such as benchtop NMR and solid state NMR), and SigMa[47] for complex spectra processing. For complex spectra, processing regions with different methods then combining them can sometimes provide a better result than processing the entire spectra with one method.[48] SigMa divides a 1D NMR spectrum into three categories: signature signals (SS), signals of unknown spin systems (SUS), and bins (BINS) which are too complicated to align and annotate. SS and SUS are further aligned and quantified by line-shape fitting, but the signals in BINS are integrated directly. By doing this, more information can be extracted from the spectra.

With the rapid development of tools and workflows, the need for workflow management is drawing attention.[49] Verhoeven et al.[50] recently wrote a review on KNIME[51] and Galaxy[52] workflow management platforms, where users can assemble, automatically run and record their workflows. Workflow4Metabolomics (W4M)[53] is another workflow management infrastructure based on Galaxy. Beyond just building and running workflows, this platform is also designed to be a workflow repository. Usually, workflows are deposited with data in data repositories, such as MetaboLights[54] and Metabolomics Workbench[55], but with this platform, workflows can be cited and shared directly. PhenoMeNal[56] is a recently built cloud-based metabolomics analysis e-infrastructure. While it incorporates W4M, it provides a greater variety of established tools than W4M. PhenoMeNal enables calculations to be run on the cloud and therefore makes analysis both time- and resource-efficient. While current pipeline tools aim to provide automatic solutions for analysis, an advanced user may tend to have more flexibility in choosing methods for each step. Therefore, a central language-independent tool repository for metabolomics research would be invaluable for users to learn and explore functions.

1.2.4 Integration of NMR with Other Data

The quantitative and reproducible qualities of NMR spectroscopy make it ideal to use in data integration. Because of the complementary nature of NMR and MS, combining them in a study is advantageous. They can be combined in tandem or in parallel for structure elucidation or for a better metabolite coverage.[57], [58], [59] The Brüschweiler lab developed an approach to integrate NMR and high-resolution MS data called SUMMIT MS/NMR (Structure of Unknown Metabolomic Mixture components by MS/NMR).[60] With high-resolution MS data, it is possible to obtain an accurate molecular formula for an unknown MS feature. NMR chemical shifts are calculated for every possible structure, and the corresponding NMR data are searched for the best match. This conceptually simple approach can become quite complicated with larger values of m/z, which can lead to a large number of structures. It also depends on accurate calculations of NMR chemical shifts, which as noted above are now quite accurate with high-level theory.[61] The SUMMIT approach would nicely complement the metabolite fraction libraries described above.[62]

With recent improvements in quantification accuracy and the development of compatible sample preparation protocols for MS and NMR techniques, inter-platform correlation is reinforced.[57], [63], [64] For example, Clendinen et al. used NMR and LC-MS to look for potential biomarkers of prostate cancer recurrence.[64] They measured polar extractions of human serum samples by both NMR and hydrophilic interaction liquid chromatography (HILIC)-MS and non-polar extractions by reversed phase liquid chromatography (RPLC)-MS. Each platform detected some unique analytes. The authors used correlations between signals across platforms to confirm peak annotations. Also, features from the same metabolites with low or negative inter-platform correlation might indicate unreliable quantifications for either platform; therefore, those features were excluded from statistical analysis. For multivariate statistical analysis, NMR and MS data were concatenated and feature selected. As a result, a set of 20 metabolites (3 from NMR) were reported to be potential biomarkers. In addition, correlations between signals from metabolites measured from different platforms were observed, which supplemented metabolic crosstalk information. Together, these results revealed the strength of inter-platform correlation on improving peak annotation confidence and relative quantification and unravelling biological relationships between metabolites. Moreover, Nagana Gowda and coworkers showed that inter-platform correlation could make absolute quantification in MS samples easier when NMR quantification from the same sample is used as reference.[65]

In many cases, processed NMR and MS data are statistically analyzed separately and integrated on a pathway level. Combining data matrices before multivariate analysis is relatively rare partly due to the matrix size issue.[66] For most cases, the number of variables is much larger than the number of samples, which is not favorable to most statistical analyses. Concatenating matrices will further amplify these differences. The concept of penalized multiblock analysis accompanied with feature selection is suitable for this situation.[66], [67] It also manages the imbalance of signal scales between platforms. Deng et al. reported efficient classification of sample groups in integrated LC-MS and NMR data with feature selection for multiblock PLS-DA.[68] They showed that simply concatenating matrices did not produce better performance than did a single matrix, but integrated matrices with feature selection did outperform

a single matrix with feature selection. The availability of these statistical tools should be recognized for a more thorough usage of information in data integration.

NMR-based metabolomics can also be integrated with other omics, such as genomics, transcriptomics, proteomics, or microbiomics, in order to develop a more comprehensive understanding of biological systems.[69] Sheikh et al. recently introduced metabolomics to glycomics studies in *C. elegans*.[48] Besides glycomics data, the authors also integrated metabolomics data with worm population distribution to analyze the relationship between metabolites, glycans, and size as a proxy for development. In this study, synchronized worm samples were measured for their metabolome by NMR, glycome by LC-MS/MS, and population distribution by large-particle flow cytometry. Different sized worms showed distinct patterns of glycan and metabolite levels. A correlation network between the three data matrices also showed associations between metabolites, glycans, and worm size. Furthermore, NMR-measured metabolites provided a substrate-level detail of glycan modification and glycosylation. For example, the authors observed that phosphocholine was positively correlated to some developmental-stage-specific N-glycans. This result suggested those glycans may be potential substrates for phosphocholine modifications. The correlation between UDP-*N*-acetylglucosamine (UDP-GlcNAc), O-glycans, and worm sizes indicated possible changes in O-glycan utilization with worm growth. Therefore, together with glycan-level changes, metabolomics results shed light on glycan dynamics during worm development. [REFER TO CHAPTER 2]

With the development of statistical methods, data integration is becoming more flexible. For example, Le Moyec et al. used an unsupervised multiblock model to analyze NMR-measured metabolites and biochemical assay results, which contained heterogenous analytes such as specific lipid levels, protein levels and enzymatic activities, for understanding equine energy metabolism during horse racing.[70] Furthermore, data integration can in turn help NMR peak annotation with knowledge from other omics. Wang et al. built a network with NMR-measured metabolite levels, microbiome gene abundance in rumen fluid samples, and compound knowledge in the KEGG database.[71] NMR features were associated with genes through linear and nonlinear correlations. Those genes were mapped in the KEGG database for connected compounds through reaction knowledge. In this way, NMR features were connected to compound names, thus helping to extrapolate peak identity.

The developers of Metabomatching also proposed the idea to annotate metabolites with their associated genetic traits (e.g. SNP).[72] In a mGWAS experimental set, they gathered peaks highly associated with one SNP to generate a pseudospectrum for database query. This approach was tested to work on some known metabolites. While this work may not yet fully replace routine annotation, for example by 2D NMR, it can provide some idea of unknown peaks, as metabolites that generate such signals would be associated with enzymes coded by the genes. However, this kind of annotation technique is limited by the genetic diversity of samples and requires a specific experimental design. It is thus more suitable for an integrated study rather than an unaccompanied metabolomics study.

Statistical methods are developing quickly for multi-omics studies,[67], [73] but methods of integrating metabolomics with other omics data are limited. Using methods developed on other omics-integrated approaches for metabolomics-involved integration is a promising avenue for future research.

1.3 Current Challenges for Data Processing in NMR-based Metabolomics: Chemical Shift Variation in NMR Spectra

1.3.1 Sources of Chemical Shift Variation

Since electrons affect the magnetic field, ionization and complexation, which change the spatial electrical density, can change nuclei resonance frequencies [74], [75]. For acids and bases, their dissociation rates are determined by the pH of the solution. The relationship between dissociation status of an ionization site and pH can be explained by the Henderson–Hasselbalch equation (1.1) [76], [77]. Taking a weak acid (HA) as an example:

$$pH = pK_a + log_{10}\frac{[A^-]}{[HA]} \tag{1.1}$$

where pK_a represents the dissociation constant.

The magnetic field strength around this protonation site is then related to the pH of the sample. For a compound with single ionization capacity, the relationship between its chemical shift and sample pH is expressed in equation (1.2) [78]:

$$\delta_o = \frac{\delta_A + \delta_{HA}(10^{(pH-pK_a)})}{1 + 10^{(pH-pK_a)}} \tag{1.2}$$

where δ_o , δ_A and δ_{HA} represents the chemical shifts of an observed signal of the weak acid, the deprotonated status (A^-) and protonated status (HA), respectively. In another word, δ_o is the weighted average of δ_A and δ_{HA}.

For a metabolite with multiple ionization sites, although the total number of ionized forms can be summed up for showing the relationship between compound ionization status and pH, ionizations on these different sites of the metabolite will affect different signals on the spectra. Ionizations on one site may also affect the electrical density distribution of the other ionization sites, thereby influencing its chemical shift.

When the ionization sites are independent of each other, the relationship between chemical shift and pH can be extended from equation (1.2) as equation (1.3) [78]:

$$\delta_o = \frac{\delta_A + \sum_{i=1}^{n} \delta_{H_iA} 10^{\sum_{j=n-i+1}^{n} pK_j - ipH}}{1 + \sum_{k=1}^{n} 10^{\sum_{l=n-k+1}^{n} pK_l - kpH}} \tag{1.3}$$

where n denotes the number of ionizable sites in the compound.

As shown in equation (1.3), the relationship between chemical shift of one site and sample pH is always monotonic for independent ionization sites, although the curve may have different shapes. But when there is interaction, the relationship is more complicated, and the chemical shift can change non-monotonically with the change in the sample pH [79]. Fortunately, in practice most curves observed in a metabolomics study are monotonic [17], [74].

As ionic strength affects pK_a, the ionic strength is another factor that affect the chemical shift of a nucleus. Observational relationships between chemical shifts and a solution's ionic strength was described before in a study using Na^+ concentration to represent ionic strength [74], where the chemical shift of measured metabolites changed monotonically with Na^+ concentration at a given pH. Chemical shifts affected by complexation also exhibit similar patterns, reflected by the relationship between chemical shifts and the concentration of Ca^{2+} or Mg^{2+} [17].

In summary, ionization and complexation usually affect the chemical shift of a nucleus in a monotonic manner. While the titration curve of a nucleus in an ionization site can be determined using the above-mentioned equations, for experimental samples, the factors that impact ionization and complexation are usually not measured nor are they evenly distributed. Therefore, fitting the curves to equation (1.3) is not practical.

1.3.2 Current Methods for Controlling Chemical Shift Variation

Efforts have been made to control chemical shift variations from matrix effects. One strategy is to control the samples' pH by buffering. However, because the ionic concentration also affects NMR acquisition performance [80], concentrations of buffer need to be limited. Therefore, a balance needs to be determined in what constitutes an appropriate amount of buffering compounds to use. In practice, it is widely reported that the buffering is insufficient. An alternative to buffering samples to neutral physiological conditions [74], [13] is to buffer samples to extreme acidic or basic pH values [81]. Its applications are also limited because the extreme conditions can change the ionization of metabolites and even introduce new reactions that are not observed under certain physiological conditions. Therefore, what we observe under extreme conditions may not reflect the genuine biology. Chelating Ca^{2+} and Mg^{2+} by EDTA has been widely used when analyzing biological samples, but it has been reported that this step can have a great effect on the metabolome [75]. Thus, it is not recommended to chelate Ca^{2+} and Mg^{2+} for metabolomic measurements.

Another strategy is to deal with this problem computationally. One straightforward approach is aligning peaks, which is placing the signal from the same nucleus in the same chemical shift on the spectra to correct the variation [82]. Alignment methods typically use a criterion to match the corresponding peaks. The criterion can be linear correlations, Euclidian distances, or fast Fourier transform (FFT) cross-correlations between peaks across spectra [82]. After peaks are matched, they can be shifted, stretched or compressed in order to be placed on the same position on the spectra. There are a number of alignment methods available for one-dimensional NMR spectra, such as Constrained Correlation Optimized Wrapping (CCOW) [83], Fast Fourier Transform (PAFFT), recursive alignment by FFT (RAFFT) [84], interval-correlation-shifting (icoshift) [85], and Recursive Segment-Wise Peak Alignment (RSPA) [86]. These methods usually can align peaks with small chemical shift variation well but are incapable of processing peaks with great variations. It is important to note, however, that peaks with significant variation are frequently observed and include common metabolites such as histidine and citrate [74].

Another widely used approach is binning peaks. Binning one peak refers to calculating the area under the curve in a small chemical shift region to estimate the integral of the peak [82]. The same region is

used for all spectra, so the region is expected to be broad enough to include all shifting peaks, but not too broad to include other peaks. This can be a tricky task for automatic binning. Therefore, besides using the same width of region for every signal on the spectra (i.e. uniform binning), the width of regions can be optimized for different signals [87]. Binning methods, however, cannot deal with peak variations when the regions of two shifting peaks overlap.

A more accurate but tedious way is to integrate peaks on each spectrum and match corresponding peaks. This can be done manually with computational tools utilized to facilitate integrating peaks when the peaks are annotated. It can be further extended to fit peaks into Gaussian, Lorentzian, or quadratic polynomial [88] curves to estimate the intensity of a signal to deal with overlapped peaks. Since these methods rely on peak annotations to match corresponding peaks, when the peaks are shifting widely, peak annotation may be compromised because of their deviation from those found in chemical shift databases. Therefore, for fit-and-match strategies, matching peaks by only using information from the matrix itself is more favorable. Several methods have attempted to achieve this goal, such as the aforementioned "speaq 2.0" [20], but fitting is not yet the mainstream method used, most likely due to the complexity of spectra in metabolomics.

Other computational methods include modeling the intensity and frequencies of peaks as discussed above [19]. However, this method is context dependent. For samples from different species, for example, a new model is required.

Therefore, developing a tool for aligning peaks with great chemical shift variation is still in demand for NMR-based metabolomics. To fill in this gap, I developed an algorithm to align these peaks utilizing the relationship between chemical shift and pH I discussed in section 1.3.1. I will describe this algorithm in chapter 3.

1.4 Introduction to the Biological Systems

1.4.1 Metabolomics for Pregnancy

Studying pregnancy is an important application field of metabolomics because of the high dynamics in both the fetus and the mother during pregnancy [89], [90], [91]. The fetus grows from a single cell to an infant, so the mother needs to accommodate enough nutrition and space for it. Also, the maternal immune system undergoes a series of changes for adapting to the allogeneic fetus [92]. These biological processes require and lead to systematic changes in the metabolome [89], [90], [91], [93]. Keeping healthy during gestation is critical for the fetus. Prenatal complications such as fetal growth restriction (FGR), organ malformations, preterm birth and fetal mortality, not only threaten fetal safety, but also can impact postnatal health [94]. For example, preterm infants have lower survival rates and may have issues in their respiratory, digestive, neuron, or cardiovascular systems [95], [96]. For mothers, pregnancy is also one of the most dangerous periods in a human life. Maternal complications include gestational diabetes [97], infections [98], and high blood pressure, which is associated with pre-eclampsia [94].

Pregnancy can exacerbate the impact of challenging conditions. Some conditions that an adult can easily deal with may cause serious problems on the mother or on the child. Pathogen infection is a common example. Pregnant mothers are more vulnerable to infections due to changed immune activity[98]. Although the mother usually recovers, the infection may lead to mortality or morbidity in the child [99], [100]. Zika virus infection, as an example, typically causes mild symptoms such as fever and joint pain in adults, but can lead to stillbirth or impaired central neuron system development in the fetuses [101].

Another example of an adverse condition is stress. Stress impacts humans mainly through the hormone cortisol. Cortisol can increase blood glucose levels, blood pressure, and heart rate, as well as dysregulate immune responses, therefore increasing the risk of gestational diabetes, pre-eclampsia, infections, preterm delivery and FGR [102], [103]. Maternal cortisol levels can also directly impact fetal cortisol levels through the placenta [104], [105]. Cortisol is an important factor for fetal organ maturation [106]. Growing fetuses are particularly susceptible to cortisol. Usually, placental enzyme 11β -hydroxysteroid dehydrogenase type 2 deactivates cortisol to restrict its impact on the fetus [107]. However, this barrier can only provide partial protection, especially in early and late gestation [108]. Unusual increased stress also downregulates enzyme activity [108]. Therefore, when the fetal cortisol level is elevated, it promotes the fetus's transition to maturation before being fully developed. There is supporting evidence that prenatal cortisol elevation due to maternal stress also dysregulates the postnatal hypothalamic-pituitary-adrenal (HPA) axis and this effect may lead to lasting neuropsychologic disorders [108], [109], [103].

1.4.2 Considerations of Sample Choices for Pregnancy-related Metabolomics

When studying pregnancy, one big challenge is the inaccessibility of fetal samples to analyze. The only non-invasive measurement on fetal growth is through ultrasound, which provides limited information. Under strict restrictions, amniotic fluid [110], and chorionic villi samples [92] can be collected. For non-human animals, collecting fetal samples usually requires euthanasia, which not only utilizes both time and resources but is subject to critical ethical concerns. Statistically, this limits sample collection to only one time point per animal, which limits longitudinal study design and requires a larger sample size for controlling inter-individual variability. Therefore, a routine choice is to monitor maternal specimens.

Fetuses obtain nutrition, excrete waste, and interact with external environments through their mothers. A study on newborns younger than 48 hours and their mothers during late gestation has shown correlations between the neonatal and maternal urine metabolome [111]. Although not directly comparing fetuses with mothers, this result provided us confidence in assessing fetal metabolic status through maternal specimens. Also, the maternal metabolome has been associated with fetal impairments, such as fetal growth restriction [112], neuron system defects [113], and congenital heart disease [94]. Together, these studies justified assessing the maternal metabolome to study fetal status.

1.4.3 Sample Choices for Human Metabolomics

Urine is one of the most widely used specimen in pregnancy or other metabolomics studies because it can be collected non-invasively [16], [114]. Non-invasive sample collection is not only an advantageous

choice for the mothers, but also an important protection for neonates, especially those under severe health conditions. Comparing to other non-invasive specimen types, such as saliva and hair, urine can be easily collected in sufficient volume. This makes it more favorable for techniques limited by sensitivity. In addition, urine is also suitable for metabolomics studies [16] because the kidney has already filtered out macromolecules in normal conditions. The simplicity of sample preparation not only saves time and resources, but also prevents potential loss of information from extraction.

Another well-recognized choice is blood specimen. Technology development has made blood drops available for metabolomics studies [115]. It has facilitated metabolomics to play an important role in inborn error screenings [116].

1.4.4 Sheep Model as a Good Tool for Fetal Metabolomics

Although maternal and fetal metabolic statuses are interconnected, the fetal metabolome does not totally depend on maternal status because of the barrier that the placenta provides. The placenta is an exchange interface for maternal and fetal gases, nutrients and wastes. Some metabolites, such as glucose [117], diffuse to fetal blood passively with or without the help of a transporter. Therefore, the fetal concentrations of these metabolites are always lower than those found on the maternal side. On the other hand, metabolites such as amino acids are actively transported to the fetus to meet its demands [117]. In addition, the placenta can metabolize some metabolites for supplying fetal growth. For example, the placenta can produce lactate from glucose and transport it to the fetus [117]. It can also produce hormones for regulating the maternal and fetal body, as well as protect fetuses from the maternal hormones as discussed before. Therefore, fetuses can regulate their own metabolome.

With the demand of directly measuring the fetal metabolome, sheep have been used as a good model species for pregnancy studies [118], [119] , [120]. This is because collecting fetal samples in live sheep is feasible. Researchers can surgically place sampling catheters in fetal blood vessels and leave one of the catheter ends out of the maternal flanks [121]. Both sheep mothers and fetuses are resilient enough for this procedure, thus continued sampling from live, unanesthetized fetuses is possible. So far sheep is the only species that has shown this long-term capacity [119]. Catheters can also be placed in sheep in the maternal side of the placenta to directly investigate placental transportation [118].

Sheep also have a long gestation period (140∼150 days for full gestation) [120], which makes more time available for exogenous treatment accumulation or for resting between each step if the experiment involves a series of procedures. However, this also means a long period between generations and may be unfavorable for experiments that require a large number of replicates or generations.

There is a great deal of similarity between sheep and human pregnancies [119], [120]. Sheep have small litter sizes and certain sheep breeds have singleton or twin births, which is the same as human. Sheep newborns have similar birth weights as human newborns. Although the sheep placenta is different than the human placenta, the structure of the circulatory system is similar and a good many metabolites and gases cross the placenta in the same way, although the transporters may be different.

Different organs or systems develop and mature at different times during or after gestation. The timing of fetal development on a lot of organs in sheep are similar to those in humans [119]. In both

sheep and human newborns, the brain and cardiovascular systems are mostly developed at birth, while rats and mice, in comparison, mostly develop their major organs postnatally [120]. Preterm sheep and human newborns have similar issues with underdeveloped respiratory and cardiovascular systems [95] [119]. Sheep and humans also have similar metabolic adaptions after birth [122]. Although mature sheep rely on fatty acids fermented from ruminal bacteria for their energy supply, sheep fetuses still use glucose as a main energy source [122]. Therefore, sheep and human newborns experience similar transitions from taking maternally supplied glucose to drinking fatty-acid-abundant milk.

There have been a number of successful results using the findings from sheep model studies for improving human prenatal health. One extraordinary example of this is through studying preterm sheep lung maturation. Results of these studies have been used to successfully increase the survival rates of preterm human newborns by prenatally treating them with corticosteroids to stimulate lung maturation [123]. The sheep model has also been explored in pregnancy-related metabolomics [124], [125], [126], [127], and has provided interesting information for understanding pregnancy. Therefore, the sheep model is a good choice for investigating fetal metabolomic changes during gestation.

Based on the aforementioned knowledge, I developed computational tools to improve NMR-based metabolomics analysis and applied them to analyze urine metabolome from pregnant mothers with virus infection. I then used a sheep model to further investigate fetal and neonatal metabolic change with maternal stressed condition as well as their metabolic transitions after birth.

Bibliography

[1] L. K. Reed, C. F. Baer, and A. S. Edison, "Considerations when choosing a genetic model organism for metabolomics studies," *Current opinion in chemical biology*, vol. 36, p. 7, Feb. 2017, ISSN: 18790402. DOI: 10.1016/J.CBPA.2016.12.005. [Online]. Available: /pmc/articles/PMC5337163/%20/pmc/articles/PMC5337163/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5337163/.

[2] K. Sinclair and E. Dudley, "Metabolomics and Biomarker Discovery," *Advances in experimental medicine and biology*, vol. 1140, pp. 613–633, 2019, ISSN: 00652598. DOI: 10.1007/978-3-030-15950-4_37. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-15950-4_37.

[3] A. Zhang, H. Sun, G. Yan, P. Wang, and X. Wang, "Metabolomics for Biomarker Discovery: Moving to the Clinic," *BioMed Research International*, vol. 2015, 2015, ISSN: 23146141. DOI: 10.1155/2015/354671.

[4] D. S. Wishart, "Metabolomics for investigating physiological and pathophysiological processes," *Physiological Reviews*, vol. 99, no. 4, pp. 1819–1875, 2019, ISSN: 15221210. DOI: 10.1152/PHYSREV.00035.2018/ASSET/IMAGES/LARGE/Z9J0041929140021.JPEG. [Online]. Available: https://journals.physiology.org/doi/full/10.1152/physrev.00035.2018.

[5] L. Perez De Souza, S. Alseekh, Y. Brotman, and A. R. Fernie, "Network-based strategies in metabolomics data analysis and interpretation: from molecular networking to biological interpretation," *Expert Review of Proteomics*, vol. 17, no. 4, pp. 243–255, Apr. 2020, ISSN: 1478-9450. DOI: 10.1080/14789450.2020.1766975. [Online]. Available: https://doi.org/10.1080/14789450.2020.1766975.

[6] A. S. Edison, M. Colonna, G. J. Gouveia, *et al.*, *NMR: Unique Strengths That Enhance Modern Metabolomics Research*, Jan. 2021. DOI: 10.1021/acs.analchem.0c04414. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/acs.analchem.0c04414.

[7] S. Cardoso, T. Afonso, M. Maraschin, and M. Rocha, "WebSpecmine: A website for metabolomics data analysis and mining," *Metabolites*, vol. 9, no. 10, Oct. 2019, ISSN: 22181989. DOI: 10.3390/metabo9100237.

[8] H. Friebolin and J. K. Becconsall, *Basic One- and Two-Dimensional NMR Spectroscopy*. Wiley, 1998, ISBN: 9783527295135. [Online]. Available: https://books.google.com/books?id=%5C%5C_rLwAAAAMAAJ.

[9] L. Maitre, C. H. E. Lau, E. Vizcaino, *et al.*, "Assessment of metabolic phenotypic variability in children's urine using 1H NMR spectroscopy," *Scientific Reports*, vol. 7, no. October 2016, pp. 1–12, 2017, ISSN: 20452322. DOI: 10.1038/srep46082. [Online]. Available: http://dx.doi.org/10.1038/srep46082.

[10] H. C. Keun, T. M. Ebbels, H. Antti, *et al.*, "Analytical Reproducibility in 1H NMR-Based Metabonomic Urinalysis," *Chemical Research in Toxicology*, vol. 15, no. 11, pp. 1380–1386, Nov. 2002, ISSN: 0893228X. DOI: 10.1021/TX0255774. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/tx0255774.

[11] M. E. Dumas, E. C. Maibaum, C. Teague, *et al.*, "Assessment of Analytical Reproducibility of 1H NMR Spectroscopy Based Metabonomics for Large-Scale Epidemiological Research: the INTERMAP Study," *Analytical chemistry*, vol. 78, no. 7, p. 2199, Apr. 2006, ISSN: 00032700. DOI: 10.1021/AC0517085. [Online]. Available: /pmc/articles/PMC6561113/%20/pmc/articles/PMC6561113/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6561113/.

[12] O. Cloarec, M. E. Dumas, A. Craig, *et al.*, "Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic 1H NMR data sets," *Analytical Chemistry*, vol. 77, no. 5, pp. 1282–1289, 2005, ISSN: 00032700. DOI: 10.1021/ac048630x.

[13] A. C. Dona, B. Jiménez, H. Schafer, *et al.*, "Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping," *Analytical Chemistry*, vol. 86, no. 19, pp. 9887–9894, Oct. 2014, ISSN: 15206882. DOI: 10.1021/ac5025039. [Online]. Available: http://www.ebi.ac..

[14] O. Beckonert, M. Coen, H. C. Keun, *et al.*, "High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues," *Nature Protocols*, vol. 5, no. 6, pp. 1019–1032, 2010, ISSN: 17502799. DOI: 10.1038/nprot.2010.45.

[15] M. T. Judge, Y. We, F. Tayyari, *et al.*, "Continuous in vivo metabolism by NMR," *Frontiers in Molecular Biosciences*, vol. 6, no. APR, p. 26, 2019, ISSN: 2296889X. DOI: 10.3389/FMOLB.2019.00026/BIBTEX.

[16] S. Bouatra, F. Aziat, R. Mandal, *et al.*, "The Human Urine Metabolome," *PLoS ONE*, vol. 8, no. 9, 2013, ISSN: 19326203. DOI: 10.1371/journal.pone.0073076.

[17] G. D. Tredwell, J. G. Bundy, M. De Iorio, and T. M. Ebbels, "Modelling the acid/base1H NMR chemical shift limits of metabolites in human urine," *Metabolomics*, vol. 12, no. 10, pp. 1–10, 2016, ISSN: 15733890. DOI: 10.1007/s11306-016-1101-y.

[18] T. N. Vu and K. Laukens, "Getting your peaks in line: A review of alignment methods for nmr spectral data," *Metabolites*, vol. 3, no. 2, pp. 259–76, 2013, ISSN: 2218-1989 (Print) 2218-1989 (Linking). DOI: 10.3390/metabo3020259. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24957991.

[19] P. G. Takis, H. Schäfer, M. Spraul, and C. Luchinat, "Deconvoluting interrelationships between concentrations and chemical shifts in urine provides a powerful analysis tool," *Nature Communications*, vol. 8, no. 1, p. 1662, 2017, ISSN: 2041-1723. DOI: 10.1038/s41467-017-01587-0. [Online]. Available: https://doi.org/10.1038/s41467-017-01587-0.

[20] C. Beirnaert, P. Meysman, T. N. Vu, *et al.*, "Speaq 2.0: A complete workflow for high-throughput 1d nmr spectra processing and quantification," *PLoS Comput Biol*, vol. 14, no. 3, e1006018, 2018, ISSN: 1553-7358 (Electronic) 1553-734X (Linking). DOI: 10.1371/journal.pcbi.1006018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29494588.

[21] J. Marchand, E. Martineau, Y. Guitton, G. Dervilly-Pinel, and P. Giraudeau, "Multidimensional nmr approaches towards highly resolved, sensitive and high-throughput quantitative metabolomics," *Curr Opin Biotechnol*, vol. 43, pp. 49–55, 2017, ISSN: 1879-0429 (Electronic) 0958-1669 (Linking). DOI: 10.1016/j.copbio.2016.08.004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27639136.

[22] C. Ludwig and M. R. Viant, "Two-dimensional j-resolved nmr spectroscopy: Review of a key methodology in the metabolomics toolbox," *Phytochem Anal*, vol. 21, no. 1, pp. 22–32, 2010, ISSN: 1099-1565 (Electronic) 0958-0344 (Linking). DOI: 10.1002/pca.1186. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19904730%20http://onlinelibrary.wiley.com/doi/10.1002/pca.1186/abstract.

[23] M. Tabatabaei Anaraki, W. Bermel, R. Dutta Majumdar, *et al.*, "1d "spikelet" projections from heteronuclear 2d nmr data-permitting 1d chemometrics while preserving 2d dispersion," *Metabolites*, vol. 9, no. 1, 2019, ISSN: 2218-1989 (Print) 2218-1989 (Linking). DOI: 10.3390/metabo9010016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30654443.

[24] A. Rodriguez-Martinez, R. Ayala, J. M. Posma, *et al.*, "Pjres binning algorithm (jba): A new method to facilitate the recovery of metabolic information from pjres 1h nmr spectra," *Bioinformatics*, vol. 35, no. 11, pp. 1916–1922, 2019, ISSN: 1367-4811 (Electronic) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/bty837. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30351417.

[25] B. J. Blaise, L. Shintu, B. Elena, L. Emsley, M. E. Dumas, and P. Toulhoat, "Statistical recoupling prior to significance testing in nuclear magnetic resonance based metabonomics," *Anal Chem*, vol. 81, no. 15, pp. 6242–51, 2009, ISSN: 1520-6882 (Electronic) 0003-2700 (Linking). DOI: 10.1021/ac9007754. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19585975.

[26] B. Worley and R. Powers, "Generalized adaptive intelligent binning of multiway data," *Chemometr Intell Lab Syst*, vol. 146, pp. 42–46, 2015, ISSN: 0169-7439 (Print) 0169-7439 (Linking). DOI: 10.1016/j.chemolab.2015.05.005. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26052171.

[27] F. Puig-Castellvi, Y. Perez, B. Pina, R. Tauler, and I. Alfonso, "Compression of multidimensional nmr spectra allows a faster and more accurate analysis of complex samples," *Chem Commun (Camb)*, vol. 54, no. 25, pp. 3090–3093, 2018, ISSN: 1364-548X (Electronic) 1359-7345 (Linking). DOI: 10.1039/c7cc09891j. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29411785.

[28] M. Niklasson, R. Otten, A. Ahlner, *et al.*, "Comprehensive analysis of nmr data using advanced line shape fitting," *J Biomol NMR*, vol. 69, no. 2, pp. 93–99, 2017, ISSN: 0925-2738 (Print) 0925-2738. DOI: 10.1007/s10858-017-0141-6.

[29] S. L. Robinette, R. Ajredini, H. Rasheed, *et al.*, "Hierarchical alignment and full resolution pattern recognition of 2d nmr spectra: Application to nematode chemical ecology," *Anal Chem*, vol. 83, no. 5, pp. 1649–57, 2011, ISSN: 1520-6882 (Electronic) 0003-2700 (Linking). DOI: 10.1021/ac102724x. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21314130.

[30] O. Cloarec, M. E. Dumas, A. Craig, *et al.*, "Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic 1h nmr data sets," *Anal Chem*, vol. 77, no. 5, pp. 1282–9, 2005, ISSN: 0003-2700 (Print) 0003-2700 (Linking). DOI: 10.1021/ac048630x. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15732908.

[31] A. Charris-Molina, G. Riquelme, P. Burdisso, and P. A. Hoijemberg, "Consecutive queries to assess biological correlation in nmr metabolomics: Performance of comprehensive search of multiplets over typical 1d 1h nmr database search," *Journal of Proteome Research*, 2020, ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.9b00872. [Online]. Available: https://doi.org/10.1021/acs.jproteome.9b00872.

[32] A. Charris-Molina, G. Riquelme, P. Burdisso, and P. A. Hoijemberg, "Tackling the peak overlap issue in nmr metabolomics studies: 1d projected correlation traces from statistical correlation analysis on nontilted 2d (1)h nmr j-resolved spectra," *J Proteome Res*, vol. 18, no. 5, pp. 2241–2253, 2019, ISSN: 1535-3907 (Electronic) 1535-3893 (Linking). DOI: 10.1021/acs.jproteome.9b00093. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30916564.

[33] C. Ludwig, J. M. Easton, A. Lodi, *et al.*, "Birmingham metabolite library: A publicly accessible database of 1-d 1h and 2-d 1h j-resolved nmr spectra of authentic metabolite standards (bml-nmr)," *Metabolomics*, vol. 8, no. 1, pp. 8–18, 2012, ISSN: 1573-3890. DOI: 10.1007/s11306-011-0347-7. [Online]. Available: https://doi.org/10.1007/s11306-011-0347-7.

[34] C. H. Johnson, T. J. Athersuch, I. D. Wilson, *et al.*, "Kinetic and j-resolved statistical total correlation nmr spectroscopy approaches to structural information recovery in complex reacting mixtures: Application to acyl glucuronide intramolecular transacylation reactions," *Analytical Chemistry*, vol. 80, no. 13, pp. 4886–4895, 2008, ISSN: 0003-2700. DOI: 10.1021/ac702614t. [Online]. Available: https://doi.org/10.1021/ac702614t.

[35] P. A. Hoijemberg and I. Pelczer, "Fast metabolite identification in nuclear magnetic resonance metabolomic studies: Statistical peak sorting and peak overlap detection for more reliable database queries," *J Proteome Res*, vol. 17, no. 1, pp. 392–401, 2018, ISSN: 1535-3907 (Electronic) 1535-3893 (Linking). DOI: 10.1021/acs.jproteome.7b00617. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29135266.

[36] I. Garcia-Perez, J. M. Posma, J. I. Serrano-Contreras, *et al.*, "Identifying unknown metabolites using nmr-based metabolic profiling techniques," *Nat Protoc*, vol. 15, no. 8, pp. 2538–2567, 2020, ISSN: 1750-2799 (Electronic) 1750-2799 (Linking). DOI: 10.1038/s41596-020-0343-3. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32681152.

[37] C. Wang, B. Zhang, I. Timari, *et al.*, "Accurate and efficient determination of unknown metabolites in metabolomics by nmr-based molecular motif identification," *Anal Chem*, vol. 91, no. 24, pp. 15 686–15 693, 2019, ISSN: 1520-6882 (Electronic) 0003-2700 (Linking). DOI: 10.1021/acs.analchem.9b03849. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31718151.

[38] S. Kang, Y. Kwon, D. Lee, and Y.-S. Choi, "Predictive modeling of nmr chemical shifts without using atomic-level annotations," *Journal of Chemical Information and Modeling*, 2020, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.0c00494. [Online]. Available: https://doi.org/10.1021/acs.jcim.0c00494.

[39] C. Zhang, Y. Idelbayev, N. Roberts, *et al.*, "Small molecule accurate recognition technology (smart) to enhance natural products research," *Sci Rep*, vol. 7, no. 1, p. 14 243, 2017, ISSN: 2045-2322 (Electronic) 2045-2322 (Linking). DOI: 10.1038/s41598-017-13923-x. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29079836.

[40] H. Dashti, J. R. Wedell, W. M. Westler, *et al.*, "Applications of parametrized nmr spin systems of small molecules," *Analytical Chemistry*, vol. 90, no. 18, pp. 10 646–10 649, 2018. DOI: 10.1021/acs.analchem.8b02660.

[41] D. Jacob, C. Deborde, M. Lefebvre, M. Maucourt, and A. Moing, "Nmrprocflow: A graphical and interactive tool dedicated to 1d spectra processing for nmr-based metabolomics," *Metabolomics*, vol. 13, no. 4, p. 36, 2017, ISSN: 1573-3882 (Print) 1573-3882 (Linking). DOI: 10.1007/s11306-017-1178-y. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28261014.

[42] G. Lefort, L. Liaubet, C. Canlet, *et al.*, "Asics: An r package for a whole analysis workflow of 1d 1h nmr spectra," *Bioinformatics*, vol. 35, no. 21, pp. 4356–4363, 2019, ISSN: 1367-4811 (Electronic) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btz248. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30977816.

[43] F. Madrid-Gambin, S. Oller-Moreno, L. Fernandez, *et al.*, "Alpsnmr: An r package for signal processing of fully untargeted nmr-based metabolomics," *Bioinformatics*, vol. 36, no. 9, pp. 2943–2945, 2020, ISSN: 1367-4811 (Electronic) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btaa022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31930381.

[44] M. Martin, B. Legat, J. Leenders, *et al.*, "Pepsnmr for (1)h nmr metabolomic data pre-processing," *Anal Chim Acta*, vol. 1019, pp. 1–13, 2018, ISSN: 1873-4324 (Electronic) 0003-2670 (Linking). DOI: 10.1016/j.aca.2018.02.067. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29625674.

[45] R. Barrilero, M. Gil, N. Amigo, *et al.*, "Lipspin: A new bioinformatics tool for quantitative (1)h nmr lipid profiling," *Anal Chem*, vol. 90, no. 3, pp. 2031–2040, 2018, ISSN: 1520-6882 (Electronic) 0003-2700 (Linking). DOI: 10.1021/acs.analchem.7b04148. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29293319.

[46] S. Yamada, K. Ito, A. Kurotani, Y. Yamada, E. Chikayama, and J. Kikuchi, "Interspin: Integrated supportive webtools for low- and high-field nmr analyses toward molecular complexity," *ACS Omega*, vol. 4, no. 2, pp. 3361–3369, 2019, ISSN: 2470-1343 (Electronic) 2470-1343 (Linking). DOI: 10.1021/acsomega.8b02714. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31459550.

[47] B. Khakimov, N. Mobaraki, A. Trimigno, V. Aru, and S. B. Engelsen, "Signature mapping (sigma): An efficient approach for processing complex human urine (1)h nmr metabolomics data," *Anal Chim Acta*, vol. 1108, pp. 142–151, 2020, ISSN: 1873-4324 (Electronic) 0003-2670 (Linking). DOI: 10.1016/j.aca.2020.02.025. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32222235.

[48] M. O. Sheikh, F. Tayyari, S. Zhang, *et al.*, "Correlations between lc-ms/ms-detected glycomics and nmr-detected metabolomics in caenorhabditis elegans development," *Frontiers in molecular biosciences*, vol. 6, pp. 49–49, 2019, ISSN: 2296-889X. DOI: 10.3389/fmolb.2019.00049. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31316996%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6611444/.

[49] K. O'Shea and B. B. Misra, "Software tools, databases and resources in metabolomics: Updates from 2018 to 2019," *Metabolomics*, vol. 16, no. 3, p. 36, 2020, ISSN: 1573-3890 (Electronic) 1573-3882 (Linking). DOI: 10.1007/s11306-020-01657-3. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32146531.

[50] A. Verhoeven, M. Giera, and O. A. Mayboroda, "Scientific workflow managers in metabolomics: An overview," *Analyst*, vol. 145, no. 11, pp. 3801–3808, 2020, ISSN: 1364-5528 (Electronic) 0003-2654 (Linking). DOI: 10.1039/d0an00272k. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32374793.

[51] A. Fillbrunn, C. Dietz, J. Pfeuffer, R. Rahn, G. A. Landrum, and M. R. Berthold, "Knime for reproducible cross-domain analysis of life science data," *J Biotechnol*, vol. 261, pp. 149–156, 2017, ISSN: 1873-4863 (Electronic) 0168-1656 (Linking). DOI: 10.1016/j.jbiotec.2017.07.028. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28757290.

[52] E. Afgan, D. Baker, B. Batut, *et al.*, "The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update," *Nucleic Acids Res*, vol. 46, no. W1, W537–W544, 2018, ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gky379. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29790989.

[53] Y. Guitton, M. Tremblay-Franco, G. Le Corguille, *et al.*, "Create, run, share, publish, and reference your lc-ms, fia-ms, gc-ms, and nmr data analysis workflows with the workflow4metabolomics 3.0 galaxy online infrastructure for metabolomics," *Int J Biochem Cell Biol*, vol. 93, pp. 89–101, 2017, ISSN: 1878-5875 (Electronic) 1357-2725 (Linking). DOI: 10.1016/j.biocel.2017.07.002. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28710041.

[54] K. Haug, K. Cochrane, V. C. Nainala, *et al.*, "Metabolights: A resource evolving in response to the needs of its scientific community," *Nucleic Acids Res*, vol. 48, no. D1, pp. D440–D444, 2020, ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gkz1019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31691833.

[55] M. Sud, E. Fahy, D. Cotter, *et al.*, "Metabolomics workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools," *Nucleic Acids Res*, vol. 44, no. D1, pp. D463–70, 2016, ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gkv1042. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26467476.

[56] K. Peters, J. Bradbury, S. Bergmann, *et al.*, "Phenomenal: Processing and analysis of metabolomics data in the cloud," *Gigascience*, vol. 8, no. 2, 2019, ISSN: 2047-217X (Electronic) 2047-217X (Linking). DOI: 10.1093/gigascience/giy149. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30535405.

[57] D. D. Marshall and R. Powers, "Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics," *Prog Nucl Magn Reson Spectrosc*, vol. 100, pp. 1–16, 2017, ISSN: 1873-3301 (Electronic) 0079-6565 (Linking). DOI: 10.1016/j.pnmrs.2017.01.001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28552170.

[58] J. Srinivasan, F. Kaplan, R. Ajredini, *et al.*, "A blend of small molecules regulates both mating and development in caenorhabditis elegans," *Nature*, vol. 454, no. 7208, pp. 1115–8, 2008, ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: 10.1038/nature07168. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18650807.

[59] Q. Shou, L. Feng, Y. Long, *et al.*, "A hybrid polyketide-nonribosomal peptide in nematodes that promotes larval survival," *Nat Chem Biol*, vol. 12, no. 10, pp. 770–2, 2016, ISSN: 1552-4469 (Electronic) 1552-4450 (Linking). DOI: 10.1038/nchembio.2144. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27501395.

[60] K. Bingol, L. Bruschweiler-Li, C. Yu, A. Somogyi, F. Zhang, and R. Bruschweiler, "Metabolomics beyond spectroscopic databases: A combined ms/nmr strategy for the rapid identification of new metabolites in complex mixtures," *Anal Chem*, vol. 87, no. 7, pp. 3864–70, 2015, ISSN: 1520-6882 (Electronic) 0003-2700 (Linking). DOI: 10.1021/ac504633z. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25674812%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5035699/pdf/nihms803711.pdf.

[61] S. Das, A. S. Edison, and J. Merz K. M., "Metabolite structure assignment using in silico nmr techniques," *Anal Chem*, vol. 92, no. 15, pp. 10 412–10 419, 2020, ISSN: 1520-6882 (Electronic) 0003-2700 (Linking). DOI: 10.1021/acs.analchem.0c00768. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32608974%20https://pubs.acs.org/doi/10.1021/acs.analchem.0c00768.

[62] L. Whiley, E. Chekmeneva, D. J. Berry, *et al.*, "Systematic isolation and structure elucidation of urinary metabolites optimized for the analytical-scale molecular profiling laboratory," *Anal Chem*, vol. 91, no. 14, pp. 8873–8882, 2019, ISSN: 1520-6882 (Electronic) 0003-2700 (Linking). DOI: 10.1021/acs.analchem.9b00241. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31188566.

[63] X. Li, H. Luo, T. Huang, L. Xu, X. Shi, and K. Hu, "Statistically correlating nmr spectra and lc-ms data to facilitate the identification of individual metabolites in metabolomics mixtures," *Anal Bioanal Chem*, vol. 411, no. 7, pp. 1301–1309, 2019, ISSN: 1618-2650 (Electronic) 1618-2642 (Linking). DOI: 10.1007/s00216-019-01600-z. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30793214.

[64] C. S. Clendinen, D. A. Gaul, M. E. Monge, *et al.*, "Preoperative metabolic signatures of prostate cancer recurrence following radical prostatectomy," *J Proteome Res*, vol. 18, no. 3, pp. 1316–1327, 2019, ISSN: 1535-3907 (Electronic) 1535-3893 (Linking). DOI: 10.1021/acs.jproteome.8b00926. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30758971.

[65] G. A. Nagana Gowda, D. Djukovic, L. F. Bettcher, H. Gu, and D. Raftery, "Nmr-guided mass spectrometry for absolute quantitation of human blood metabolites," *Anal Chem*, vol. 90, no. 3, pp. 2001–2009, 2018, ISSN: 1520-6882 (Electronic) 0003-2700 (Linking). DOI: 10.1021/acs.analchem.7b04089. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29293320.

[66] J. Boccard and S. Rudaz, "Harnessing the complexity of metabolomic data with chemometrics," *Journal of Chemometrics*, vol. 28, no. 1, pp. 1–9, 2014, ISSN: 08869383. DOI: 10.1002/cem.2567.

[67] A. Csala and A. H. Zwinderman, "Multivariate statistical methods for high-dimensional multiset omics data analysis," in *Computational Biology*, H. Husi, Ed. Brisbane (AU), 2019, ISBN: 9780994438195. DOI: 10.15586/computationalbiology.2019.ch5. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31815402.

[68] L. Deng, H. Gu, J. Zhu, *et al.*, "Combining nmr and lc/ms using backward variable elimination: Metabolomics analysis of colorectal cancer, polyps, and healthy controls," *Analytical chemistry*, vol. 88, no. 16, pp. 7975–7983, 2016, ISSN: 1520-6882 0003-2700. DOI: 10.1021/acs.analchem.6b00885. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/27437783%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5450811/.

[69] F. R. Pinu, D. J. Beale, A. M. Paten, *et al.*, "Systems biology and multi-omics integration: Viewpoints from the metabolomics research community," *Metabolites*, vol. 9, no. 4, 2019, ISSN: 2218-1989 (Print) 2218-1989 (Linking). DOI: 10.3390/metabo9040076. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31003499.

[70] L. Le Moyec, C. Robert, M. N. Triba, *et al.*, "A first step toward unraveling the energy metabolism in endurance horses: Comparison of plasma nuclear magnetic resonance metabolomic profiles before and after different endurance race distances," *Front Mol Biosci*, vol. 6, p. 45, 2019, ISSN: 2296-889X (Print) 2296-889X (Linking). DOI: 10.3389/fmolb.2019.00045. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31245385.

[71] M. Wang, H. Wang, H. Zheng, R. Dewhurst, and R. Roehe, "A knowledge-driven network-based analytical framework for the identification of rumen metabolites," *IEEE Trans Nanobioscience*, vol. 19, no. 3, pp. 518–526, 2020, ISSN: 1558-2639 (Electronic) 1536-1241 (Linking). DOI: 10.1109/TNB.2020.2991577. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32356756.

[72] R. Rueedi, R. Mallol, J. Raffler, *et al.*, "Metabomatching: Using genetic association to identify metabolites in proton nmr spectroscopy," *PLoS computational biology*, vol. 13, no. 12, e1005839–e1005839, 2017, ISSN: 1553-7358 1553-734X. DOI: 10.1371/journal.pcbi.1005839. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29194434%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5711027/.

[73] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: Recent progress in multi-omics data integration methods," *Front Genet*, vol. 8, p. 84, 2017, ISSN: 1664-8021 (Print) 1664-8021 (Linking). DOI: 10.3389/fgene.2017.00084. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28670325.

[74] C. Xiao, F. Hao, X. Qin, Y. Wang, and H. Tang, "An optimized buffer system for NMR-based urinary metabonomics with effective pH control, chemical shift consistency and dilution minimization †," 2009. DOI: 10.1039/b818802e. [Online]. Available: www.rsc.org/analyst.

[75] L. Jiang, J. Huang, Y. Wang, and H. Tang, "Eliminating the dication-induced intersample chemical-shift variations for NMR-based biofluid metabonomic analysis," *Analyst*, vol. 137, no. 18, pp. 4209–4219, Aug. 2012, ISSN: 1364-5528. DOI: 10.1039/C2AN35392J. [Online]. Available: https://pubs.rsc.org/en/content/articlehtml/2012/an/c2an35392j%20https://pubs.rsc.org/en/content/articlelanding/2012/an/c2an35392j.

[76] K. A. Hasselbalch, *Die Berechnung der Wasserstoffzahl des Blutes aus der freien und gebundenen Kohlens??ure desselben, und die Sauerstoffbindung des Blutes als Funktion der Wasserstoffzahl*, German. Berlin: Julius Springer, 1916.

[77] L. J. Henderson, "CONCERNING THE RELATIONSHIP BETWEEN THE STRENGTH OF ACIDS AND THEIR CAPACITY TO PRESERVE NEUTRALITY," *American Journal of Physiology-Legacy Content*, vol. 21, no. 2, pp. 173–179, Mar. 1908, ISSN: 0002-9513. DOI: 10.1152/ajplegacy.1908.21.2.173. [Online]. Available: https://doi.org/10.1152/ajplegacy.1908.21.2.173.

[78] Z. Szakács, G. Hägele, and R. Tyka, "1H/31P NMR pH indicator series to eliminate the glass electrode in NMR spectroscopic pKa determinations," *Analytica Chimica Acta*, vol. 522, no. 2, pp. 247–258, 2004, ISSN: 0003-2670. DOI: https://doi.org/10.1016/j.aca.2004.07.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003267004008384.

[79] A. Onufriev, D. A. Case, and G. M. Ullmann, "A Novel View of pH Titration in Biomolecules," *Biochemistry*, vol. 40, no. 12, pp. 3413–3419, Mar. 2001, ISSN: 0006-2960. DOI: 10.1021/bi002740q. [Online]. Available: https://doi.org/10.1021/bi002740q.

[80] M. W. Voehler, G. Collier, J. K. Young, M. P. Stone, and M. W. Germann, "Performance of cryogenic probes as a function of ionic strength and sample tube geometry," *Journal of magnetic resonance (San Diego, Calif. : 1997)*, vol. 183, no. 1, p. 102, Nov. 2006, ISSN: 10907807. DOI: 10.1016/J.JMR.2006.08.002. [Online]. Available: /pmc/articles/PMC4852285/%20/pmc/articles/PMC4852285/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4852285/.

[81] A. Beneduci, G. Chidichimo, G. Dardo, and G. Pontoni, "Highly routinely reproducible alignment of 1H NMR spectral peaks of metabolites in huge sets of urines," *Analytica Chimica Acta*, vol. 685, no. 2, pp. 186–195, 2011, ISSN: 00032670. DOI: 10.1016/j.aca.2010.11.027. [Online]. Available: http://dx.doi.org/10.1016/j.aca.2010.11.027.

[82] T. N. Vu and K. Laukens, "Getting your peaks in line: A review of alignment methods for NMR spectral data," *Metabolites*, vol. 3, no. 2, pp. 259–276, 2013, ISSN: 22181989. DOI: 10.3390/metabo3020259.

[83] N.-P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping," *Journal of Chromatography A*, vol. 805, no. 1-2, pp. 17–35, May 1998, ISSN: 0021-9673. DOI: 10.1016/S0021-9673(98)00021-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021967398000211.

[84] J. W. H. Wong, C. Durante, and H. M. Cartwright, "Application of fast fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets," *Analytical Chemistry*, vol. 77, no. 17, pp. 5655–5661, 2005, ISSN: 00032700. DOI: 10.1021/ac050619p.

[85] F. Savorani, G. Tomasi, and S. Engelsen, "icoshift: A versatile tool for the rapid alignment of 1D NMR spectra," *Journal of Magnetic Resonance*, vol. 202, no. 2, pp. 190–202, Feb. 2010, ISSN: 1090-7807. DOI: 10.1016/J.JMR.2009.11.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1090780709003334?via%5C%3Dihub.

[86] K. A. Veselkov, J. C. Lindon, T. M. D. Ebbels, *et al.*, "Recursive Segment-Wise Peak Alignment of Biological 1 H NMR Spectra for Improved Metabolic Biomarker Recovery," *Analytical Chemistry*, vol. 81, no. 1, pp. 56–66, Jan. 2009, ISSN: 0003-2700. DOI: 10.1021/ac8011544. [Online]. Available: https://pubs.acs.org/doi/10.1021/ac8011544.

[87] S. A. Sousa, A. Magalhães, and M. M. C. Ferreira, "Optimized bucketing for NMR spectra: Three case studies," *Chemometrics and Intelligent Laboratory Systems*, vol. 122, pp. 93–102, Mar. 2013, ISSN: 01697439. DOI: 10.1016/j.chemolab.2013.01.006.

[88] J. W. H. Wong, G. Cagney, and H. M. Cartwright, "SpecAlign-processing and alignment of mass spectra datasets," *BIOINFORMATICS APPLICATIONS NOTE*, vol. 21, no. 9, pp. 2088–2090, 2005. DOI: 10.1093/bioinformatics/bti300. [Online]. Available: http://ptcl.chem.ox.ac.uk/.

[89] S. O. Diaz, A. S. Barros, B. J. Goodfellow, *et al.*, "Following healthy pregnancy by nuclear magnetic resonance (NMR) metabolic profiling of human urine.," eng, *Journal of proteome research*, vol. 12, no. 2, pp. 969–979, Feb. 2013, ISSN: 1535-3907 (Electronic). DOI: 10.1021/pr301022e. [Online]. Available: https://pubs.acs.org/sharingguidelines.

[90] K. L. Lindsay, C. Hellmuth, O. Uhl, *et al.*, "Longitudinal Metabolomic Profiling of Amino Acids and Lipids across Healthy Pregnancy.," *PloS one*, vol. 10, no. 12, e0145794, 2015, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0145794. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26716698%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4699222.

[91] L. Liang, M. L. H. Rasmussen, B. Piening, *et al.*, "Metabolic Dynamics and Prediction of Gestational Age and Time to Delivery in Pregnant Women," *Cell*, vol. 181, no. 7, p. 1680, Jun. 2020, ISSN: 10974172. DOI: 10.1016/J.CELL.2020.05.002. [Online]. Available: /pmc/articles/PMC7327522/%20/pmc/articles/PMC7327522/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7327522/.

[92] G. Monni, F. Murgia, V. Corda, *et al.*, "Metabolomic Investigation of β-Thalassemia in Chorionic Villi Samples," *Journal of Clinical Medicine*, vol. 8, no. 6, Jun. 2019, ISSN: 20770383. DOI: 10.3390/JCM8060798. [Online]. Available: /pmc/articles/PMC6616561/%20/pmc/articles/PMC6616561/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616561/.

[93] M. Wang, W. Xia, H. Li, *et al.*, "Normal pregnancy induced glucose metabolic stress in a longitudinal cohort of healthy women Novel insights generated from a urine metabolomics study," 2018. DOI: 10.1097/MD.0000000000012417. [Online]. Available: http://dx.doi.org/10.1097/MD.0000000000012417.

[94] G. Monni, L. Atzori, V. Corda, *et al.*, "Metabolomics in Prenatal Medicine: A Review," *Frontiers in Medicine*, vol. 8, p. 771, Jun. 2021, ISSN: 2296858X. DOI: 10.3389/FMED.2021.645118/BIBTEX.

[95] J. G. Bensley, R. De Matteo, R. Harding, and M. J. Black, "The effects of preterm birth and its antecedents on the cardiovascular system," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 95, no. 6, pp. 652–663, Jun. 2016, ISSN: 1600-0412. DOI: 10.1111/AOGS.12880. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/aogs.12880%20https://onlinelibrary.wiley.com/doi/abs/10.1111/aogs.12880%20https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.12880.

[96] M. J. Platt, "Narrative Review Outcomes in preterm infants," 2014. DOI: 10.1016/j.puhe.2014.03.010. [Online]. Available: http://dx.doi.org/10.1016/j.puhe.2014.03.010.

[97] Q. Chen, E. Francis, G. Hu, and L. Chen, "Metabolomic profiling of women with gestational diabetes mellitus and their offspring: Review of metabolomics studies," *Journal of Diabetes and its Complications*, vol. 32, no. 5, pp. 512–523, May 2018, ISSN: 1056-8727. DOI: 10.1016/J.JDIACOMP.2018.01.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1056872717316562?via%5C%3Dihub.

[98] B. Abu-Raya, C. Michalski, M. Sadarangani, and P. M. Lavoie, "Maternal Immunological Adaptation During Normal Pregnancy," *Frontiers in Immunology*, vol. 11, p. 2627, Oct. 2020, ISSN: 16643224. DOI: 10.3389/FIMMU.2020.575197/BIBTEX.

[99] S. Giakoumelou, N. Wheelhouse, K. Cuschieri, G. Entrican, S. E. Howie, and A. W. Horne, "The role of infection in miscarriage," *Human Reproduction Update*, vol. 22, no. 1, p. 116, Jan. 2016, ISSN: 14602369. DOI: 10.1093/HUMUPD/DMV041. [Online]. Available: /pmc/articles/PMC4664130/%20/pmc/articles/PMC4664130/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4664130/.

[100] S. Aleem and Z. A. Bhutta, "Infection-related stillbirth: an update on current knowledge and strategies for prevention," *https://doi.org/10.1080/14787210.2021.1882849*, vol. 19, no. 9, pp. 1117–1124, 2021, ISSN: 17448336. DOI: 10.1080/14787210.2021.1882849. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/14787210.2021.1882849.

[101] D. Baud, D. J. Gubler, B. Schaub, M. C. Lanteri, and D. Musso, "An update on Zika virus infection.," eng, *Lancet (London, England)*, vol. 390, no. 10107, pp. 2099–2109, Nov. 2017, ISSN: 1474-547X (Electronic). DOI: 10.1016/S0140-6736(17)31450-2.

[102] M. Keller-Wood, X. Feng, C. E. Wood, *et al.*, "Elevated maternal cortisol leads to relative maternal hyperglycemia and increased stillbirth in ovine pregnancy," *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, vol. 307, no. 4, R405, Aug. 2014. DOI: 10.1152/AJPREGU.00530.2013. [Online]. Available: /pmc/articles/PMC4137155/%20/pmc/articles/PMC4137155/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4137155/.

[103] M. E. Coussons-Read, "Effects of prenatal stress on pregnancy and human development: mechanisms and pathways," *Obstetric Medicine*, vol. 6, no. 2, p. 52, 2013, ISSN: 17534968. DOI: 10.1177/1753495X12473751. [Online]. Available: /pmc/articles/PMC5052760/%20/pmc/articles/PMC5052760/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5052760/.

[104] R. Gitau, A. Cameron, N. M. Fisk, *et al.*, "Fetal exposure to maternal cortisol Symptomatic restenosis after carotid percutaneous transluminal angioplasty," *The Lancet*, vol. 352, no. 9129, pp. 707–708, 1998.

[105] H. S. Kane, C. Dunkel Schetter, L. M. Glynn, C. J. Hobel, and C. A. Sandman, "Pregnancy Anxiety and Prenatal Cortisol Trajectories," vol. 64, no. 12, pp. 2391–2404, 2008, ISSN: 15378276. DOI: 10.1038/jid.2014.371. arXiv: NIHMS150003.

[106] G. C. Liggins, "The Role of Cortisol in Preparing the Fetus for Birth*," *Review Reprod. Fertil. Dev*, vol. 6, pp. 141–50, 1994.

[107] R. Benediktsson, A. A. Calder, C. R. W. Edwards, and J. R. Seckl, "Placental 11β-hydroxysteroid dehydrogenase: a key regulator of fetal glucocorticoid exposure," *Clinical Endocrinology*, vol. 46, no. 2, pp. 161–166, Feb. 1997, ISSN: 0300-0664. DOI: https://doi.org/10.1046/j.1365-2265.1997.1230939.x. [Online]. Available: https://doi.org/10.1046/j.1365-2265.1997.1230939.x.

[108] M. A. Howland, C. A. Sandman, and L. M. Glynn, "Developmental origins of the human hypothalamic-pituitary-adrenal axis," *Expert review of endocrinology & metabolism*, vol. 12, no. 5, p. 321, Sep. 2017, ISSN: 17448417. DOI: 10.1080/17446651.2017.1356222. [Online]. Available: /pmc/articles/PMC6334849/%20/pmc/articles/PMC6334849/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6334849/.

[109] A. Harris and J. Seckl, "Glucocorticoids, prenatal stress and the programming of disease," *Hormones and Behavior*, vol. 59, no. 3, pp. 279–289, Mar. 2011, ISSN: 0018-506X. DOI: 10.1016/J.YHBEH.2010.06.007.

[110] J. Shan, T. Xie, J. Xu, H. Zhou, and X. Zhao, "Metabolomics of the amniotic fluid: Is it a feasible approach to evaluate the safety of Chinese medicine during pregnancy?" *Journal of Applied Toxicology*, vol. 39, no. 1, pp. 163–171, Jan. 2019, ISSN: 1099-1263. DOI: 10.1002/JAT.3653. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/jat.3653%20ht

tps://onlinelibrary.wiley.com/doi/abs/10.1002/jat.3653%20https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jat.3653.

[111] S. Perrone, E. Laschi, G. De Bernardo, *et al.*, "Newborn metabolomic profile mirrors that of mother in pregnancy," 2019. DOI: 10.1016/j.mehy.2019.109543. [Online]. Available: https://doi.org/10.1016/j.mehy.2019.109543.

[112] D. F. B. Leite and J. G. Cecatti, "New Approaches to Fetal Growth Restriction: The Time for Metabolomics Has Come," *Revista brasileira de ginecologia e obstetricia : revista da Federacao Brasileira das Sociedades de Ginecologia e Obstetricia*, vol. 41, no. 7, pp. 454–462, 2019, ISSN: 1806-9339. DOI: 10.1055/S-0039-1692126. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31250420/.

[113] X. Zheng, M. Su, L. Pei, *et al.*, "Metabolic Signature of Pregnant Women with Neural Tube Defects in Offspring," *Journal of Proteome Research*, vol. 10, no. 10, pp. 4845–4854, Oct. 2011, ISSN: 1535-3893. DOI: 10.1021/pr200666d. [Online]. Available: https://doi.org/10.1021/pr200666d.

[114] A. Noto, V. Fanos, and A. Dessì, *Metabolomics in Newborns*, 1st ed. Elsevier Inc., 2016, vol. 74, pp. 35–61, ISBN: 9780128046890. DOI: 10.1016/bs.acc.2015.12.006. [Online]. Available: http://dx.doi.org/10.1016/bs.acc.2015.12.006.

[115] V. V. Hernandes, C. Barbas, and D. Dudzik, "A review of blood sample handling and pre-processing for metabolomics studies," *ELECTROPHORESIS*, vol. 38, no. 18, pp. 2232–2241, Sep. 2017, ISSN: 1522-2683. DOI: 10.1002/ELPS.201700086. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/elps.201700086%20https://onlinelibrary.wiley.com/doi/abs/10.1002/elps.201700086%20https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/elps.201700086.

[116] R. D. Beger, W. Dunn, M. A. Schmidt, *et al.*, "Metabolomics enables precision medicine: "A White Paper, Community Perspective"," *Metabolomics*, vol. 12, no. 10, 2016, ISSN: 15733890. DOI: 10.1007/s11306-016-1094-6.

[117] T. McNanley and J. Woods, "Placental Physiology," *The Global Library of Women's Medicine*, 2009. DOI: 10.3843/GLOWM.10195.

[118] J. S. Barry and R. V. Anthony, "The Pregnant Sheep as a Model for Human Pregnancy," *Theriogenology*, vol. 69, no. 1, p. 55, Jan. 2008, ISSN: 0093691X. DOI: 10.1016/J.THERIOGENOLOGY.2007.09.021. [Online]. Available: /pmc/articles/PMC2262949/%20/pmc/articles/PMC2262949/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2262949/.

[119] J. L. Morrison, M. J. Berry, K. J. Botting, *et al.*, "Improving pregnancy outcomes in humans through studies in sheep," *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*, vol. 315, no. 6, R1123–R1153, Dec. 2018, ISSN: 15221490. DOI: 10.1152/AJPREGU.00391.2017/ASSET/IMAGES/LARGE/ZH60091895460007.JPEG. [Online]. Available: https://journals.physiology.org/doi/abs/10.1152/ajpregu.00391.2017.

[120] M. D. Andersen, A. K. O. Alstrup, C. S. Duvald, *et al.*, "Animal Models of Fetal Medicine and Obstetrics," *Experimental Animal Models of Human Diseases - An Effective Therapeutic Strategy*, Mar. 2018. DOI: 10.5772/INTECHOPEN.74038. [Online]. Available: https://www.intechopen.com/chapters/60219.

[121] B. J. Allison, K. L. Brain, Y. Niu, *et al.*, "Fetal in vivo continuous cardiovascular function during chronic hypoxia," *The Journal of Physiology*, vol. 594, no. 5, p. 1247, Mar. 2016, ISSN: 14697793. DOI: 10.1113/JP271091. [Online]. Available: /pmc/articles/PMC4771786/%20/pmc/articles/PMC4771786/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4771786/.

[122] J. Girard, P. Ferre, J. P. Pegorier, and P. H. Duee, "Adaptations of glucose and fatty acid metabolism during perinatal period and suckling-weaning transition," *Physiological Reviews*, vol. 72, no. 2, pp. 507–562, 1992, ISSN: 00319333. DOI: 10.1152/PHYSREV.1992.72.2.507.

[123] E. McGoldrick, F. Stewart, R. Parker, and S. R. Dalziel, "Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth," *Cochrane Database of Systematic Reviews*, vol. 2021, no. 2, Dec. 2020, ISSN: 14651858. DOI: 10.1002/14651858.CD004454.PUB4/MEDIA/CDSR/CD004454/IMAGE_N/NCD004454-CMP-002.03.SVG. [Online]. Available: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD004454.pub4/full%20https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD004454.pub4/abstract.

[124] J. M. Walejko, A. Antolic, J. P. Koelmel, T. J. Garrett, A. S. Edison, and M. Keller-Wood, "Chronic maternal cortisol excess during late gestation leads to metabolic alterations in the newborn heart," *American Journal of Physiology - Endocrinology and Metabolism*, vol. 316, no. 3, E546–E556, Mar. 2019, ISSN: 15221555. DOI: 10.1152/ajpendo.00386.2018. [Online]. Available: /pmc/articles/PMC6459297/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6459297/.

[125] S. Joseph, J. M. Walejko, S. Zhang, A. S. Edison, and M. Keller-Wood, "Maternal hypercortisolemia alters placental metabolism: A multiomics view," *American Journal of Physiology - Endocrinology and Metabolism*, vol. 319, no. 5, E950–E960, Nov. 2020, ISSN: 15221555. DOI: 10.1152/AJPENDO.00190.2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7790119/.

[126] J. M. Walejko, J. P. Koelmel, T. J. Garrett, A. S. Edison, and M. Keller-Wood, "Multiomics approach reveals metabolic changes in the heart at birth," *American Journal of Physiology - Endocrinology and Metabolism*, vol. 315, no. 6, E1212–E1223, Dec. 2018, ISSN: 15221555. DOI: 10.1152/ajpendo.00297.2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6336953.

[127] S. Joseph, M. Li, S. Zhang, *et al.*, "Sodium dichloroacetate stimulates cardiac mitochondrial metabolism and improves cardiac conduction in the ovine fetus during labor," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 322, no. 1, R83–R98, 2022. DOI: 10.1152/ajpregu.00185.2021. [Online]. Available: https://doi.org/10.1152/ajpregu.00185.2021.

Chapter 2

CORRELATIONS BETWEEN LC-MS/MS-DETECTED GLYCOMICS AND NMR-DETECTED METABOLOMICS IN *CAENORHABDITIS ELEGANS* DEVELOPMENT

[1]

[1] S. Zhang.*, M.O. Sheikh*, F. Tayyari*, M.T. Judge, D.B. Weatherly, F.V. Ponce, L. Wells, and A.S. Edison. 2019. *Frontiers in Molecular Biosciences*. 6.
Reprinted here with permission from the publisher. *: co-first authors.

Author contributions

This paper was published in Frontiers in Molecular Biosciences. I am the co-first author with M. Osman Sheikh and Fariba Tayyari. I developed algorithms for Biosorting correlation analysis and conducted Cytoscape analysis, M. Osman Sheikh conducted the glycan measurements, Fariba Tayyari conducted NMR measurements and analysis, Michael T. Judge developed interactive NMR binning and assisted with Cytoscape analysis, Francesca V. Ponce made worm samples, and D. Brent Weatherly analyzed the glycomics data. Lance Wells, Arthur S. Edison, M. Osman Sheikh, Fariba Tayyari, and I advised on design and interpretation. M. Osman Sheikh, Lance Wells, and Arthur S. Edison wrote the initial manuscript draft, which was edited by all authors and finalized by Lance Wells and Arthur S. Edison.

Abstract

This study examined the relationship between glycans, metabolites, and development in *C. elegans*. Samples of N2 animals were synchronized and grown to five different time points ranging from L1 to a mixed population of adults, gravid adults, and offspring. Each time point was replicated seven times. The samples were each assayed by a large particle flow cytometer (Biosorter) for size distribution data, LC-MS/MS for targeted *N*- and *O*-linked glycans, and NMR for metabolites. The same samples were utilized for all measurements, which allowed for statistical correlations between the data. A new protocol was developed to correlate Biosorter developmental data with LC-MS/MS data to obtain stage-specific information of glycans. From the five time points, four distinct sizes of worms were observed from the Biosorter distributions, ranging from the smallest corresponding to L1 to adult animals. A network model was constructed using the four binned sizes of worms as starting nodes and adding glycans and metabolites that had correlations with $r \geq 0.5$ to those nodes. The emerging structure of the network showed distinct patterns of *N*- and *O*-linked glycans that were consistent with previous studies. Furthermore, some metabolites that were correlated to these glycans and worm sizes showed interesting interactions. Of note, UDP-GlcNAc had strong positive correlations with many *O*-glycans that were expressed in the largest animals. Similarly, phosphorylcholine correlated with many *N*-glycans that were expressed in L1 animals.

2.1 Introduction

This paper presents a new approach to evaluate the relationship between Caenorhabditis elegans development, glycan abundance, and metabolites. Regardless of the organism, glycomics and metabolomics are generally conducted independently, but metabolism and glycan biosynthesis are intimately related [1]. For example, *O*-linked β-*N*-acetylglucosamine (*O*-GlcNAc)—a type of posttranslational glycosylation of nuclear and cytoplasmic proteins—acts as a sensor of nutrition and cellular stress [2], [3]. The addition of *O*-GlcNAc to proteins is catalyzed by a single enzyme, *O*-GlcNAc transferase (OGT), which relies on the availability of the sugar-nucleotide donor substrate, UDP-GlcNAc via the hexosamine biosynthetic pathway (HBP) [4]. Through the HBP, concentrations of UDP-GlcNAc are modulated by the

metabolism of glucose, fatty acids, amino acids, and nucleotides. This vital glycosylation precursor is not only utilized by OGT to modify thousands of proteins with *O*-GlcNAc [2], [3], but also by many other glycosyltransferases to generate more elaborate types of *N*- and *O*-linked glycans [5], [6]. Better approaches of associating glycomics and metabolomics would be valuable to gain a deeper understanding of their interactions.

C. elegans has become an important model organism for chemical signaling [7], [8], [9], metabolism [10], [11], [12], and glycomics [13]. *C. elegans* has a surprising diversity of many of these groups. Both ascaroside [11], [12] and *O*-GlcNAc [2] biosynthesis incorporate many of the same primary metabolic pathways in *C. elegans*. Therefore, *C. elegans* is a good model organism to study the interactions between glycomics and metabolomics. *C. elegans* develops from egg to adult in about 3 days through 4 distinct larval stages (L1-L4), young adult, adult. When resources are limited or the population of worms too high, *C. elegans* enters the dauer stage, which can persist for several months and is specialized for dispersal [14]. *C. elegans* development has been studied for decades, including the seminal study that mapped the entire cell lineage of post-embryonic animals and led to the discovery of apoptosis [15]. Gene expression has been linked with development in *C. elegans* through the use of green fluorescent protein (GFP), the first application of this important technique in animals [16]. Metabolites [17], [9], [18] and glycans [19], [20], [21] have been implicated in playing major roles in different stages of development. However, because there are no simple tools such as the use of a fluorescent reporter of gene expression, it is still extremely difficult to relate metabolites and glycans to development.

In this study, we used LC-MS/MS to quantify the expression profiles of both *N*- and *O*-linked glycans in *C. elegans* as a function of development. We developed a novel approach to statistically correlate LC-MS/MS glycomics data with Biosorting data, which provides a population distribution of the samples. To our knowledge, until now nobody has statistically associated molecular data such as glycans and metabolites with population distribution data through a large-particle flow cytometer. This allows a direct and unbiased association between glycans and developmental stage. We also collected untargeted NMR data on the same samples used for glycomics and Biosorting. Statistical correlations between LC-MS/MS glycomics and NMR data provided links between specific resonances in the NMR data with specific groups of glycans, which allowed us to begin to interpret the interplay between metabolites and glycans through development in *C. elegans*. Finally, we constructed a correlation network of binned sizes of worms from Biosorter distributions, LC-MS/MS glycomics, and NMR metabolomics data, which exposes some unique interactions between these three distinct types of data.

2.2 Materials and Methods

All data reported in this study have been deposited in the Metabolomics Workbench (doi: 10.21228/M8240W) [22].

2.2.1 Reagents

PNGase A (Protein *N*-Glycosidase A, Calbiochem) was purchased from MilliporeSigma (St. Louis, MO, USA). Sodium hydroxide (50%) was purchased from Fisher Scientific. Sep-Pak C18 disposable extraction columns were obtained from Waters Corporation (Milford, MA, USA). AG-50W-X8 cation exchange resin (H^+ form) was purchased from Bio-Rad and trifluoroacetic acid from Pierce. Ultra Pure UDP-GlcNAc was purchased from Promega Corporation (Madison, WI, USA). Trypsin, Chymotrypsin, and all other chemical reagents were purchased from Sigma-Aldrich/MilliporeSigma (St. Louis, MO, USA).

2.2.2 *C. elegans* Sample Preparation

This study used N2, the laboratory reference strain of *C. elegans*, which was obtained from the Caenorhabditis Genetics Center (CGC). We followed the general protocol published previously for obtaining liquid cultures of synchronized worms [9], [17]. This defines our biological replicate: A single L1 animal from a synchronized culture was placed onto an agar plate seeded with *E. coli* MG1655. This plate was grown until there were a large number of young gravid adult hermaphrodites (about 48 h at ~24 °C). The plate was then washed into a 15 mL tube with M9 buffer, rinsed 3× with M9, and lysed with an alkaline hypochlorite solution until about 50% of the worms were dissolved (no more than 5 min). Then, M9 buffer was added to dilute the lysing solution, and the liquid was removed after gentle centrifugation at 580 g for 2 min to pellet the eggs without breaking them. This step was repeated 3× to completely remove the lysis solution. After the final rinse, eggs were resuspended in sterile water before a sucrose gradient to remove cellular debris and bacteria. An equal volume (5 mL) of 60% sucrose was added to the eggs in water and centrifuged at 350 g for 4 min. The eggs were rinsed to remove residual sucrose and once they hatched, approximately 200,000 animals were transferred to 20 mL of S-complete with 2 mL of 50% MG1655. This material was grown to the desired developmental stage and prepared as described below.

We started every culture with synchronized L1 arrested animals. To collect animals at different stages, we relied on an approximate number of hours to estimate the stage of the worm population. However, other than L1, we observed that the population had lost some synchrony over time and all worms were not in the same developmental stage at the moment of collection. Isolating large quantities of worms (e.g. 200,000) with uniform stages is not trivial, since there can be residual bacteria and debris after the synchronization step. Therefore, we report results using these time points rather than developmental stages, since they are not all pure stage cultures. The first time, T1, was collected immediately after hatching and were synchronized L1 arrested animals. Therefore, results from T1 can be directly related to L1 stage animals. The relationship between developmental stage and other time points is less clear. Indeed, as time progressed the cultures became more mixed, as shown in Figure A.1. The subsequent samples were collected at 22, 36, 49, and 90 hours (T2, T3, T4, and T5, respectively) after feeding the cultures. Based on timing and literature values of N2 development, T2 is early larval stages, T3 mid-larval, T4 late-larval to adult, and T5 adults, gravid adults with mixed-stage offspring. These estimates were qualitatively confirmed by visual inspection. To estimate the size of each worm in each sample, we utilized specific ranges of time of flight (TOF) and extinction coefficient (EXT), measured by a large particle flow cytometer called a

Biosorter (Union Biometrica). We obtained Biosorter data on each individual sample before homogenization (Figure A.1). As described below, we have developed a protocol to recover size-specific information, even from samples that have lost synchrony. This information provides a population distribution and count for each sample, because the location of individual data points in a Biosorter dataset is related to the size and optical density of each worm. This information was then statistically correlated with glycomics and NMR data, as described below (Scheme 2.1).

2.2.3 Biosorting

Following worm growth and clean up with sucrose gradient, 2.5% of the cultures were counted and sized using a Union Biometrica Biosorter-Pro large particle flow cytometer using a 250 μm flow cell. Small animals have a shorter time of flight and are optically lighter than larger animals. Therefore, L1 animals are on the lower left of a Biosorter curve and large adults are on the upper right. The Biosorter was calibrated with fluorescent control particles before the runs, as an internal standard. The sheath flow rate was kept constant at 10 mL per minute to decrease variability in length measurements as much as possible. All data were collected using the 488 nm laser with the power set at 50 mW. The signal threshold has set to 500 mV and time of flight minimum was set to 40. Green, yellow, and red photomultiplier tubes were set to 350, 400, and 650 PMT Volts, respectively. All signal gains were set to 1.0. Each Biosorter growth curve is provided in Supplemental Figure A.1.

2.2.4 Homogenization

The remaining 97.5% of the worm pellets were bead homogenized with 80% methanol/20% water using a FastPrep-24 (MP Biomedicals) for 5 cycles of 60 sec. The tubes were then centrifuged at high speed for 20 min to separate the supernatant from the beads. This process was repeated twice, and the supernatants were combined. The supernatant was then placed in a Labconco SpeedVac until no liquid was observed in the sample. The dried supernatant was then stored at -80 °C until NMR analysis. The pellets with beads were frozen at -80 °C until glycomics analysis.

2.2.5 NMR Data Collection and Analysis

NMR spectroscopy is a relatively low sensitivity measurement that requires samples with concentrations greater than about 10 μM for detection. Early larval stage animals are considerably smaller than adults and contribute much less mass per worm. Therefore, the supernatants of the 7 T1 time points were combined into one sample for the NMR analysis. This resulted in 29 dried extracted-worm pellets, which were dissolved in 600 μL NMR buffer (0.1 M sodium phosphate buffer in D_2O with a final concentration of 0.33 mM of DSS) and mixed well using a vortex mixer. 590 μL of each sample were added to 5 mm NMR tubes.

NMR data were collected at 600 MHz on a Bruker AVIII-HD console in a magnet equipped with a 5 mm CryoProbe and a SampleJet autosampler, which cooled samples to 6 °C while waiting in the

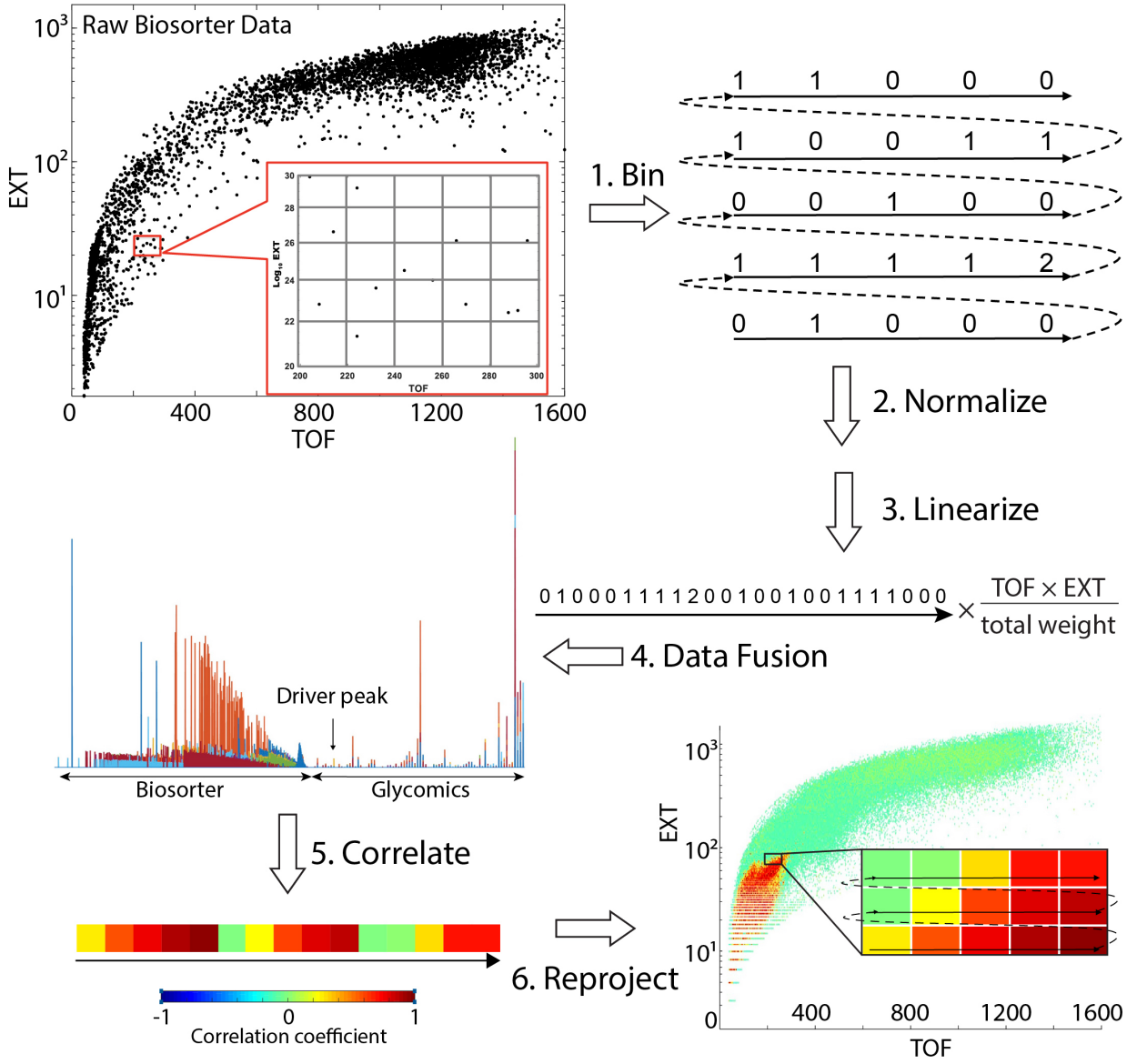


Figure 2.1: Steps involved in obtaining statistical correlations between mass spectrometry glycomics and Biosorter data. Refer to section 2.2.11 for details.

queue. All NMR acquisitions were performed at 27 °C (300 K). One dimensional (1D) nuclear Overhauser enhancement spectroscopy with presaturation (NOESY-PR) was obtained on each sample, and two-dimensional ^{13}C-^{1}H heteronuclear single quantum coherence (HSQC) and ^{13}C-^{1}H heteronuclear single quantum coherence-total correlation spectroscopy (HSQC-TOCSY) were obtained from one representative sample from each time point for compound identification.

NMR data were processed using NMRPipe using standard parameters [23]. For post processing, we used a MATLAB metabolomics toolbox developed in the Edison laboratory [24]. Spectra were referenced to the DSS resonance at 0.0 ppm, and regions corresponding to water and methanol were removed. Aligning NMR peaks can be a challenge. Chemical shifts are sensitive to several factors including pH, sample matrix, and/or ion contents. Different alignment algorithms were compared to find a proper alignment but none of them was perfect for all the regions of the spectra. Therefore, we combined results of three different peak alignment algorithms. We normalized each alignment using probabilistic quotient normalization (PQN) [25]. Constrained correlation optimized warping (CCOW) [26] using spearman cluster method was best for the first region from 0 to 3.3566 ppm, CCOW with correlation cluster method was best for the region between 3.3566 to 4.7651ppm, and Fast Fourier transform (PAFFT) [27] with correlation cluster method best for the remainder. The best regions from each method were then combined. Statistical total correlation spectroscopy (STOCSY) was used both to find the correlation between the intensities of the different peaks across the whole sample to aid in NMR identification [28] and also to correlate specific NMR peaks with glycans [29].

For metabolite identification we used a combination of COLMARm [30] and spiking. For spiking experiments, we first obtained a spectrum of the sample before spiking, along with a separate spectrum of the authentic metabolite standard. Then a small quantity of the authentic metabolite was added into the sample and NMR data recorded. The compound ID was verified if the authentic and putative NMR signals add together with no additional resonances in the spiked spectrum. We used a confidence scale for metabolite ID that ranges from 1 (lowest) to 5 (highest): 5) verified by spiking; 4) Matches in COLMARm using both HSQC and HSQC-TOCSY; 3) COLMARm matches using HSQC but not HSQC-TOCSY; 2) matches from 1D NMR to literature and/or database libraries such as BMRB [31] or HMDB [32]; 1) for putatively characterized compounds or compound classes.

2.2.6 Interactive Binning of NMR Data

For the creation of the correlation network between NMR resonances, glycans, and worm sizes (Figure 2.6B, results), the results were too complicated using full-resolution NMR data. Since we were interested in correlations between specific NMR features and the network, automatic binning using a uniform bin size across the spectrum would not yield the desired results. Therefore, we developed a new workflow in MATLAB that allows for interactive binning of features of interest from an NMR dataset. This results in features of arbitrary width and with bins with user-specified boundaries. An example region of this interactive binning is shown in Figure A.3.

Starting from a working directory with the spectral matrix and ppm vector loaded in MATLAB, the first block of code initializes parameters and creates directories to store results. The second block is then run. If there is no figure already provided as input in "figureFileNames", a new overlay plot is generated from the matrix and ppm vectors provided. The user zooms to the next peak of interest, then clicks a button in the window to draw rectangular boundaries for a feature, which is then highlighted. Boundaries from different features can overlap partially or completely, and only the left and right boundaries are considered. A prompt allows the user the option of choosing another feature or ending the workflow.

When all features have been selected, the user responds "N", and the program saves the result as a .fig file as well as in an updated variable called "ROIs". In addition, a running list of .fig filenames is kept by the script. For speed purposes (as well as security of saving progress), we save after every ∼30-50 features and recommend going across the spectrum in a linear fashion. The second block is run repeatedly until all features have been recorded (for ∼700 features, typically 3-5h). The third block of code compiles the features from all the figures and produces a .fig file containing these. Regions are integrated by summing all intensities between feature boundaries and are reported in the "features" structure object. These are overlaid on the spectra and used in downstream statistical analyses. Each feature is also assigned a unique name, which is the closest unused ppm value to the maximum within the feature boundaries. This workflow is available on the Edison Lab GitHub site (`https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA`).

2.2.7 Glycomics Sample Preparation

Using the frozen extracted pellets from homogenization described above, we further delipidated by resuspending the pellets in chloroform/methanol/water (4:8:3, v/v/v) as described previously [33]. Insoluble proteins were pelleted by centrifugation, and protein pellets were washed twice with ice-cold acetone. Finally, protein power was dried under a nitrogen evaporator.

Preparation of glycopeptides and release of *N*–linked glycans was performed as described previously [33]. Briefly, approximately 5 mg of protein powder from each sample was resuspended in 500 μL of 40 mM NH_4HCO_3, 1 M urea, 20 μg/mL trypsin, and 20 μg/mL chymotrypsin and incubated overnight (16-18 h) at 37 °C. The glycopeptide mixture was boiled for 5 min and adjusted to 5% AcOH (acetic acid) prior to a Sep-Pak C18 cartridge column clean up. Glycopeptides were eluted stepwise in 20% isopropanol in 5% AcOH, 40% isopropanol in 5% AcOH, and 100% isopropanol. The eluates were pooled and evaporated to dryness. Dried glycopeptides were resuspended in 50 mM citrate phosphate buffer (pH 5.0) for digestion with peptide:*N*-glycosidase A (PNGase A) and incubated for 18 h at 37 °C. We chose to utilize PNGase A for the release of *N*-linked glycans since its substrate-specificity is less stringent than the commonly used PNGase F. Specifically, PNGase A is capable of releasing *N*-glycan species containing α1-3-linked fucose on the chitobiose core, known to be synthesized by *C. elegans* [13], [34]. PNGase A-released oligosaccharides were separated from residual peptides by another round of Sep-Pak C18 cartridge clean-up, and the glycan flow-through was collected. Released *N*-glycans were dried down using a SpeedVac.

Since there is currently no available enzyme for the comprehensive release of *O*-linked glycans, we employed a commonly used chemical release strategy via reductive β-elimination using NaOH and $NaBH_4$ [11]. Approximately 5 mg of protein powder from each sample was processed for reductive β-elimination to release *O*-linked glycan alditols as described previously [35]. Briefly, protein powder was resuspended in 100 mM NaOH containing 1 M $NaBH_4$ and incubated for 18 h at 45 °C in a glass tube sealed with a teflon-lined screw top. Following incubation, the protein concentration of the reaction mixture was determined via absorbance at 280 nm using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE). For normalization of all samples, 2 mg of the reaction mix was neutralized with 10%

acetic acid on ice and desalted using a AG-50W-X8 (H^+ form) column (1 mL bed volume) prior to borate removal and Sep-pack C18 cartridge clean-up. Released *O*-glycans were dried down using a SpeedVac.

Both *N*- and *O*-glycans were permethylated to introduce hydrophobicity, fragmentation, and facilitate in-line separation by reverse-phase (C18) chromatography prior to detection by mass spectrometry (MS) [5], [6]. Furthermore, permethylation of glycans greatly improves MS ionization efficiency resulting in improved quantification. All released *N*- and *O*-linked glycans were permethylated prior to MS analysis according to the method by Anumula and Taylor [36].

2.2.8 NanoLC-MS/MS of Permethylated Glycans

Dried down neutral/non-sulfated permethylated glycans were resuspended in 100 μL of 100% methanol. Samples were prepared by combining 4 μL of resuspended glycans with 4 μL of an internal standard [^{13}C-Permethylated isomaltopentaose (DP5)] at a final concentration of 0.2 pmol/μL and 32 μL of LC-MS Buffer A (1 mM LiOAc and 0.02% acetic acid). For each LC-MS/MS analysis, 5 μL of each prepared sample was injected for liquid chromatography separation using an Ultimate 3000 RSLC (ThermoFisher Scientific/Dionex) equipped with a PepMap Acclaim analytical C18 column (75 μm $\times$ 15 cm, 2 μm pore size) coupled to a ThermoScientific Velos Pro Dual-Pressure Linear Ion Trap mass spectrometer. The HPLC column oven temperature was set to 60 °C to achieve optimal separation of permethylated glycans. After equilibrating the column in 99% LC-MS Buffer A for 5 mins, separation was achieved using a linear gradient from 30% to 70% LC-MS Buffer B over 150 mins at a flow rate of 300 nL/min. The analytical column was regenerated after each run by washing in 99% LC-MS Buffer B for 10 minutes and then returning to 99% LC-MS Buffer A for re-equilibration. Spray into the mass spectrometer using nanospray ionization in positive ion mode was via a stainless-steel emitter with spray voltage set to 1.8 kV and capillary temperature set at 210 °C. The MS method consisted of first collecting a full ITMS (MS1) survey scan, followed by MS2 fragmentation of the Top 10 most intense peaks using CID at 50% collision energy using an isolation window of 2 m/z. Dynamic exclusion parameters were set to exclude ions for fragmentation for 15 sec if they were detected and fragmented 5 times in 15 sec.

2.2.9 Analysis of NanoLC-MS/MS Data

Lists of candidate *N*- and *O*-glycan compositions known to be expressed in *C. elegans* were generated based previous reports [21], [37], [38], [39], [40], [41], [13], [34]. Glycan compositions known to be natively methylated in *C. elegans* were intentionally omitted in our targeted database as that information would be lost through permethylation. Phosphorylcholine-modified *N*-glycans were not considered in our study for simplicity. Glycan isomer abundances of a specific composition were summed as a single species. For each MS run, only scans after the 20 min mark (after column equilibration and sample loading) were considered. For each candidate glycan, MS/MS scans were identified where the precursor m/z was within 3 Da of the candidate m/z, considering charge states (z) of +1, +2, and +3. The background intensity of the precursors was calculated by first determining the max precursor intensity (maxPreInt) or all precursors matching a particular candidate glycan (numPrecursors), binning the precursor intensities to create an

intensity distribution where the number of bins was equal to maxPreInt / numPrecursors, and then determining the value at the 15th percentile of the distribution to represent the background intensity (bgInt). All MS/MS scans with precursor intensity less than 1.5 × bgInt were discarded. The total ion count (TIC) for each glycan was calculated by summing the intensities from each peak in the assigned MS/MS scans in each MS run (glycanTIC).

A two-fold normalization method was utilized across the 35 runs for each sample type (*N*- and *O*-glycans) as follows. For each MS run, the sum of the TIC for all glycans was calculated (repSumTIC). The max replicate TIC sum (maxRepSumTIC) was determined for each time point. A normalization factor (repNormFactor) was determined for each replicate as repSumTIC / maxRepSumTIC for each time point. For the internal standard glycan (^{13}C-permethylated isomaltopentaose, DP5), the intensity for each replicate was set to the maximum DP5 glycanTIC intensity over all replicates for each time point, under the assumption that an equal amount of that glycan is present in each replicate. For the first normalization, the glycanTIC was multiplied by the repNormFactor (glycanTIC × repNormFactor) for all experimental glycan assignments for each replicate to calculate the normalized glycan TIC (glycanNormTIC). For the second normalization, the final glycan intensity value (glycanNormTICFinal) was calculated by dividing the normalized glycan TIC intensity by the standard normalized intensity (glycanNormTIC / stdNormTIC). Symbol and Text nomenclature for representation of glycan structures is displayed according to the Symbol Nomenclature for Glycans (SNFG) [42].

2.2.10 Glycan Clustering

Hierarchical Clustering Analysis (HCA) of glycans was performed using the MATLAB built-in function 'clustergram'. Mean quantities of glycans across replicates of the same time point were calculated and imported for clustering analysis. Data were clustered in both dimensions using Euclidean distance. Linkages were calculated according to the average Euclidean distance of the new cluster. Data were standardized only on the row dimension.

2.2.11 Correlation Analysis Between Glycomics and Biosorter Data

Schematic 2.1 provides a visual summary of the steps involved in correlating analytical LC-MS/MS (or NMR) data with worm size in *C. elegans*. The overall strategy is to calculate the statistical correlation of a specific analytical feature (e.g. glycan) to the Biosorter data (or vice versa) across all samples. The data required include 1) raw Biosorter data, 2) normalized analytical data, and 3) the sample run order that relates the two datasets. We created a new MATLAB workflow that is freely available through the Edison lab GitHub site for this analysis. The steps below correspond to the steps in Schematic 2.1.

1) **Bin Biosorter data.** Raw data from the Biosorter were read and processed in MATLAB by an in-house script. The Biosorter instrument measures time of flight (TOF), which can be interpreted as the worm's length, and extinction coefficient (EXT), or optical density. Each individual event (e.g. worm) in a Biosorter plot is a separate point in the 2D TOF vs. Log_{10} EXT plots. Thus, in addition to each worm's length and optical density, these plots provide an accurate count of the number of worms in each

sample. The individual Biosorter plots for each sample in this study are shown in Figure A.1. The first step was to bin the Biosorter plots in both TOF and EXT dimensions simultaneously for each sample so that each bin counted the number of worms in both dimensions. The bin size can be adjusted, but we have found consistently good results with 1 unit as the binning size in both dimensions, the settings that are hardcoded into the script. For a given study, the bin sizes are constant. This step yields an m by n matrix for each sample containing the number of worms counted, where m is the number of bins along the TOF axis and n the number of bins along the EXT axis. We excluded Biosorter data that were beyond 1600 units in both dimensions, because larger numbers are artifacts caused by clumps of material or multiple worms.

2) **Normalize by worm mass.** Mass spectrometry or NMR spectroscopy are both dependent on sample mass, and the mass of an individual worm varies considerably from L1 to adult. Therefore, we normalized the counts in each bin by multiplying the estimated mass of the worms in each bin. There are several possible ways of doing this, including making detailed experimental mass measurements of known numbers of worms at each developmental stage. We chose to use a simpler approach: neglecting bent worms traveling through the flow cell, the TOF measurements are proportional to the length of each worm. Without a specific fluorescent marker, the EXT measurement is proportional to the thickness of the worm (larger worms are less transparent than smaller worms). We multiplied the means of TOF and EXT for each bin and assume that this is proportional to the mass of each worm. Then, to correct for differences in total numbers of worms between replicates, we also normalized each sample by total mass. To do this, we calculated the total estimated worm weight in each sample by summing up the estimated worm weights across all bins. The normalized counts were divided by the total weight of the worms in that sample.

3) **Linearize the Biosorter 2D matrix.** To statistically correlate the Biosorter with glycomics data, we first converted the binned 2D (TOF vs EXT) matrix to a 1D vector. This step was previously developed by the Edison lab for general 2D NMR multivariate analysis [24]. This step simply involves extracting rows from the binned and normalized 2D matrix and combining them as indicated in Schematic 2.1.

4) **Data fusion.** The linearized 1D vectors from step 3 were joined with the corresponding glycomics LC-MS vectors to make a single vector with Biosorter and LC-MS data from the same sample. For this step, any type of quantitative analytical data can substitute for the glycomics data used here. The overall concept for this step is an extension of statistical heterospectroscopy (SHY) [29].

5) **Correlate the data.** STOCSY[28] is a statistical method that essentially correlates all points in a sample set of vectors with a specific point along the vectors, which is termed the "driver peak". We modified the standard STOCSY script in our MATLAB metabolomics toolbox to perform the correlation analysis. The linearized vectors generated in step 4 were imported, along with a modified X-axis vector that assigns an arbitrary scale to specify the driver peak. We also specify a threshold for the resulting correlation coefficients. We then calculated Pearson correlation coefficients between the driver peak and the rest of the data. The inverse correlation can be also calculated by specifying a 2D region of interest in the Biosorter plot as the driver peak to find all glycans that correlate to a specific developmental stage.

6) **Reproject the statistical correlations onto a 2D Biosorter plot.** The output of step 5 is a single vector with correlation values to the driver peak. To easily visualize the correlations, the vector was first separated into the Biosorter and glycomics components and the linearization in step 3 reversed. This results in a plot that superficially resembles the original Biosorter data (Figure A.1) but now represents the Pearson correlation values to the driver peak from a specific glycan. Thus, the final output from this procedure is a 2D Biosorter map that is the statistical correlation between a specific glycan driver peak and the worm population. In the example figure shown in Scheme 2.1, the bright red region corresponds to small animals and indicates that the selected driver peak from glycan analysis is highly correlated to that size of worms.

2.2.12 Building a Correlation Network

We used Cytoscape (v. 3.7.1) [43] to generate a correlation network between NMR features, glycans, and approximate sizes of worms. We found that if we included all the full-resolution NMR features, the network was uninterpretable. Therefore, we used the interactive binning algorithm for NMR data described above. Even with the binned data, the highly correlated NMR data dominated the network, so we first filtered the NMR bins for those that had robust correlations between specific glycans. This was done using SHY[29], similar to the Biosorter correlations with glycans described above. Select NMR statistical correlations to each glycan are provided in Figure A.2. We picked the highly correlated NMR features to include as binned data in the Cytoscape network.

As described above, without detailed image analysis, it is difficult to assign a specific developmental stage of a worm to a region of a Biosorter plot. Therefore, for this study we hand-selected different non-overlapping regions from a Biosorter distribution that correspond to different sizes (TOF) and optical densities (Log_{10} EXT). We overlaid the Biosorter time points of Figure A.1 as a guide to bin four distinct regions that corresponded to different sizes of worms (results, Figure 2.6A). We summed the total normalized worm counts within each of these regions to include in the Cytoscape network.

Pearson correlation coefficients were calculated in MATLAB between each glycan, the binned NMR features, and the sum of worm counts in each binned Biosorter region. Because we were evaluating correlations between different classes of molecular species along pathways and associated with size, we explored a range of values and empirically found that a threshold of $|r| \geq 0.5$ provided a network with multiple nodes and edges that was simple enough to interpret. The p values for this network ranged from 0.005745 to $3.3^{*}10^{-21}$, indicating that all were statistically significant correlations. The correlation coefficients table with absolute values greater than or equal to 0.5 was exported from MATLAB to Cytoscape (3.7.1). We started by coloring all edges red for positive and blue for negative values of Pearson correlations and setting the linewidths to 0.5. We then highlighted interactions to each Biosorter region by manually selecting one of the four Biosorter worm sizes, automatically selecting its nearest neighbors, and setting the linewidth to 10.0 of the edges between this subnetwork. After doing this for all four Biosorter regions, we could easily visualize direct correlations from all nodes to each Biosorter region. We then manually organized the network by clustering directly correlated nodes. We also highlighted specific interactions of two NMR metabolites, phosphorylcholine and UDP-GlcNAc, as described in the caption for Figure 2.6B.

2.3 Results

In an effort to establish novel connections between the glycome and metabolome of *C. elegans* at defined worm sizes, we developed a strategy to utilize a single biological replicate for all stages of sample preparation, data collection, and analysis (Figure 2.2).

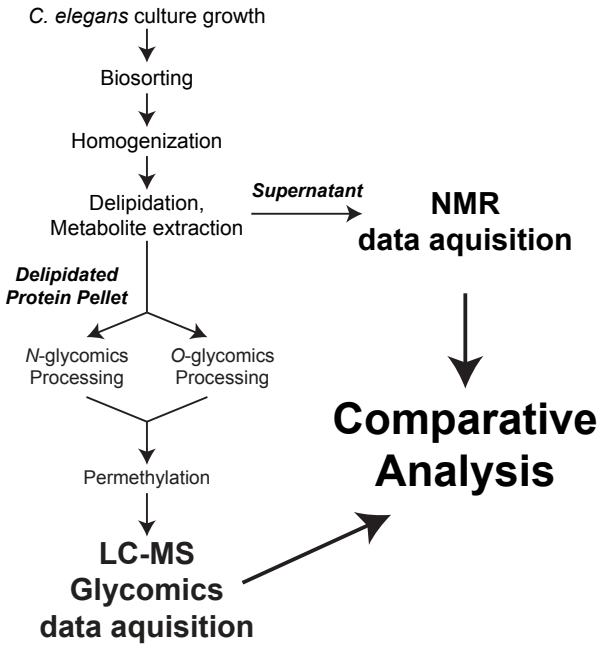


Figure 2.2: Sample Preparation Workflow. For each biological replicate, metabolites were extracted for NMR analysis while *N*- and *O*-linked glycans were released from total protein and permethylated for LC-MS analysis.

2.3.1 *N*-glycan Abundances Cycle During *C. elegans* Development.

Using a targeted approach, we generated a list of known *N*-glycans reported in the literature for the wild-type N2 strain [21], [37], [44], [45], [40], [13], [34]. Glycan compositions known to be natively methylated in *C. elegans* were omitted in our targeted list as native methylation would be masked when glycans are derivatized through permethylation. Additionally, while *C. elegans* is known to modify the core and termini of hybrid- and complex-type *N*-glycans with phosphorylcholine (PC) [21], [37], these low abundance structures were omitted from the target list for simplicity of analysis. To the best of our knowledge, previously reported *N*-glycomic data [reviewed in [13]] derives from mixed stage worms except for one study by Cipollo et al. [21]. Thus, our study aims to identify global variations in the glycome that may exist between different sizes of worms.

The criteria for identification and quantification of candidate glycans by LC-MS/MS are presented in Materials and Methods, and glycan isomers of a specific composition were treated as a single species. For

each biological replicate, glycoprotein starting material was normalized by mass prior to glycan release and derivatization. Glycan compositions identified in this study that are most consistent with specific glycan subgroups are summarized in Table 2.1 and Figure 2.3. Identification numbers were assigned arbitrarily for ease of presentation. Since epimers of carbohydrate residues have indistinguishable masses, we have reported glycan compositions detected by mass spectrometry using the generalized convention as follows, with the named epimers (and abbreviations in parentheses) specific to what has been documented for *C. elegans* previously: Hex [hexose, either glucose (Glc), galactose (Gal), or mannose (Man)], HexNAc [*N*-acetylhexosamine, either *N*-acetylglucosamine (GlcNAc) or *N*-acetylgalactosamine (GalNAc)], dHex [deoxyhexose, namely fucose (Fuc)], and HexA [hexuronic acid, namely glucuronic acid (GlcA)].

Table 2.1: Glycan compositions identified in this study

Identifier	Glycan composition	Glycan Subgroup	Calculated Mass (Permethylated)	Observed m/z (Permethylated, Lithiated)
O-Glycans (reduced reducing end)				
1	$Hex_1HexNAc_1$	Neutral	511.3	518.4
2	$Hex_2HexNAc_1$		715.4	722.4
3	$Hex_3HexNAc_1$		919.5	926.5
4	$Hex_4HexNAc_1$		1123.6	568.7
5	$HexA_1Hex_1HexNAc_1$	Charged	729.4	736.4
6	$HexA_1Hex_2HexNAc_1$		933.5	940.6
7	$HexA_1Hex_3HexNAc_1$		1137.6	575.6
8	$HexA_1Hex_4HexNAc_1$		1341.7	677.6
9	$dHex_1Hex_1HexNAc_1$	Neutral + Fucose(s)	685.4	692.5
10	$dHex_1Hex_2HexNAc_1$		889.5	896.5
11	$dHex_1Hex_4HexNAc_1$		1297.7	655.6
12	$dHex_2Hex_4HexNAc_1$		1471.8	742.7
13	$dHex_2Hex_5HexNAc_1$		1675.9	565.4
14	$dHex_2Hex_2HexNAc_2$		1308.7	661.5
15	$dHex_1Hex_3HexNAc_1$		1093.6	553.7
16	$HexA_1dHex_1Hex_1HexNAc_1$	Charged + Fucose(s)	903.5	910.5
17	$HexA_1dHex_1Hex_2HexNAc_1$		1107.6	560.7
18	$HexA_1dHex_1Hex_3HexNAc_1$		1311.7	662.7
19	$HexA_1dHex_1Hex_2HexNAc_2$		1352.7	683.5
20	$HexA_1dHex_1Hex_3HexNAc_2$		1556.8	785.6
21	$HexA_1dHex_2Hex_1HexNAc_2$		1322.7	668.5
22	$HexA_1dHex_2Hex_2HexNAc_2$		1526.8	770.6
23	$HexA_1dHex_3Hex_2HexNAc_2$		1700.9	574.1
24	$HexA_1dHex_3Hex_3HexNAc_2$		1905	959.8

25	$HexA_1dHex_4Hex_3HexNAc_2$		2079.1	1046.8
26	$HexA_1dHex_4Hex_4HexNAc_2$		2283.2	767.9
27	$HexA_1dHex_4Hex_5HexNAc_2$		2487.3	836.4
28	$HexA_1dHex_5Hex_4HexNAc_2$		2457.3	826.3
29	$HexA_1dHex_5Hex_5HexNAc_2$		2661.4	894.2
N-Glycans (free reducing end)				
30	$Hex_5HexNAc_2$	Oligomannosidic	1556.8	785.6
31	$Hex_6HexNAc_2$		1760.9	887.3
32	$Hex_7HexNAc_2$		1965	989.7
33	$Hex_8HexNAc_2$		2169.1	1091.8
34	$Hex_9HexNAc_2$		2373.2	798.4
35	$Hex_{10}HexNAc_2$		2577.3	866.2
36	$Hex_2HexNAc_2$	Paucimannosidic (+ 1 or 2 Fucoses)	944.5	951.6
37	$Hex_2dHex_1HexNAc_2$		1118.6	566.5
38	$Hex_2dHex_2HexNAc_2$		1292.7	653.5
39	$Hex_3HexNAc_2$		1148.6	581.4
40	$Hex_3dHex_1HexNAc_2$		1322.7	668.5
41	$Hex_3dHex_2HexNAc_2$		1496.8	755.6
42	$Hex_4HexNAc_2$		1352.7	683.6
43	$Hex_4dHex_1HexNAc_2$		1526.8	770.6
44	$Hex_4dHex_2HexNAc_2$		1700.9	857.7
45	$Hex_5dHex_1HexNAc_2$		1730.9	872.7
46	$Hex_5dHex_2HexNAc_2$		1905	959.8
47	$Hex_3dHex_3HexNAc_2$	Fucose-Rich	1670.9	842.7
48	$Hex_4dHex_3HexNAc_2$		1875	944.7
49	$Hex_4dHex_4HexNAc_2$		2049.1	1031.5
50	$Hex_5dHex_3HexNAc_2$		2079.1	700
51	$Hex_5dHex_4HexNAc_2$		2253.2	758.1
52	$Hex_6dHex_1HexNAc_2$		1935	974.7
53	$Hex_6dHex_3HexNAc_2$		2283.2	768.3
54	$Hex_6dHex_4HexNAc_2$		2457.2	826.4
55	$Hex_7dHex_1HexNAc_2$		2139.1	720.3
56	$HexNAc_1Hex_3HexNAc_2$	Truncated Complex	1393.7	703.7
57	$HexNAc_1Hex_3dHex_1HexNAc_2$		1567.8	529.3
58	$HexNAc_2Hex_3HexNAc_2$		1638.8	826.7
59	$HexNAc_2Hex_3dHex_1HexNAc_2$	Complex or Hybrid with additional HexNAc(s)	1812.9	611.4
60	$HexNAc_3Hex_3dHex_1HexNAc_2$		2058.1	693.3

61	$HexNAc_4Hex_3HexNAc_2$	2129.1	716.5
62	$HexNAc_5Hex_3HexNAc_2$	2374.2	798.4
63	$HexNAc_1Hex_4dHex_1HexNAc_2$	1771.9	597.4
64	$HexNAc_1Hex_4dHex_2HexNAc_2$	1946	655.4
65	$HexNAc_1Hex_4HexNAc_2$	1597.8	539.4
66	$HexNAc_2Hex_4HexNAc_2$	1842.9	621.3
67	$HexNAc_2Hex_4dHex_1HexNAc_2$	2017	679.5
68	$HexNAc_3Hex_4dHex_2HexNAc_2$	2436.3	818.8
69	$HexNAc_3Hex_4HexNAc_2$	2088.1	703.2
70	$HexNAc_4Hex_4dHex_2HexNAc_2$	2681.4	900.5
71	$HexNAc_1Hex_5HexNAc_2$	1801.9	607.4
72	$HexNAc_1Hex_5dHex_1HexNAc_2$	1976	665.5
73	$HexNAc_1Hex_5dHex_2HexNAc_2$	2150.1	723.5
74	$HexNAc_3Hex_5HexNAc_2$	2292.2	770.8
75	$HexNAc_3Hex_5dHex_2HexNAc_2$	2640.4	886.9
76	$HexNAc_1Hex_6HexNAc_2$	2006	675.5
77	$HexNAc_1Hex_6dHex_1HexNAc_2$	2180.1	733.4
78	$HexNAc_1Hex_7dHex_1HexNAc_2$	2384.2	801.6

Hex, Hexose; HexNAc, *N*-acetylhexosamine; dHex, Deoxyhexose; HexA, Hexuronic acid.
N-glycans are presented with the chitobiose core disaccharides ($GlcNAc_2$ or displayed in the table as $HexNAc_2$) written towards the right.

The most abundant structures identified belong to the paucimannosidic ($dHex_{0-2}Hex_{2-4}HexNAc_2$) and oligomannosidic (also known as "high-mannose" containing $Hex_{5-10}HexNAc_2$) subgroups, consistent with previous reports (Figure 2.3A,B) [13], [37], [40], [38], [45]. Of the paucimannose structures, which are atypical in vertebrates, the trimannosyl core structure (#39) and its monofucosylated derivative (#40) are of the greatest abundance. High levels of $Hex_5HexNAc_2$ (#30) were detected, with lesser amounts of the larger $Hex_{6-10}HexNAc_2$ (#31-35) species present, consistent with glucose and mannose trimming via the activities of α-glucosidases and α1,2-mannosidases, respectively, in the endoplasmic reticulum and with the $Man_5GlcNAc_2$ structure being a major checkpoint in *N*-glycan processing [13], [34], [46]. While much of the *N*-glycosylation machinery is conserved, the fucosylation patterns of *C. elegans* are noteworthy as this organism is predicted to express nearly 30 unique fucosyltransferases capable of decorating paucimannose- and oligomannose-type structures not usually observed in higher organisms [13], [21], [37], [44], [45], [47]. Of this type, we detected nine low abundance structures that we have classified as the fucose-rich subgroup ($Hex_{3-7}dHex_{1-4}HexNAc_2$, Figure 2.3B). Also lower in abundance were truncated complex- and hybrid/complex-type structures that could also be elaborated with one or more fucose residues (Figure 2.3C). Finally, to assess the comprehensive changes in the *N*-glycome with development, relative glycan abundances were summed for each time point (Figure 2.3D).

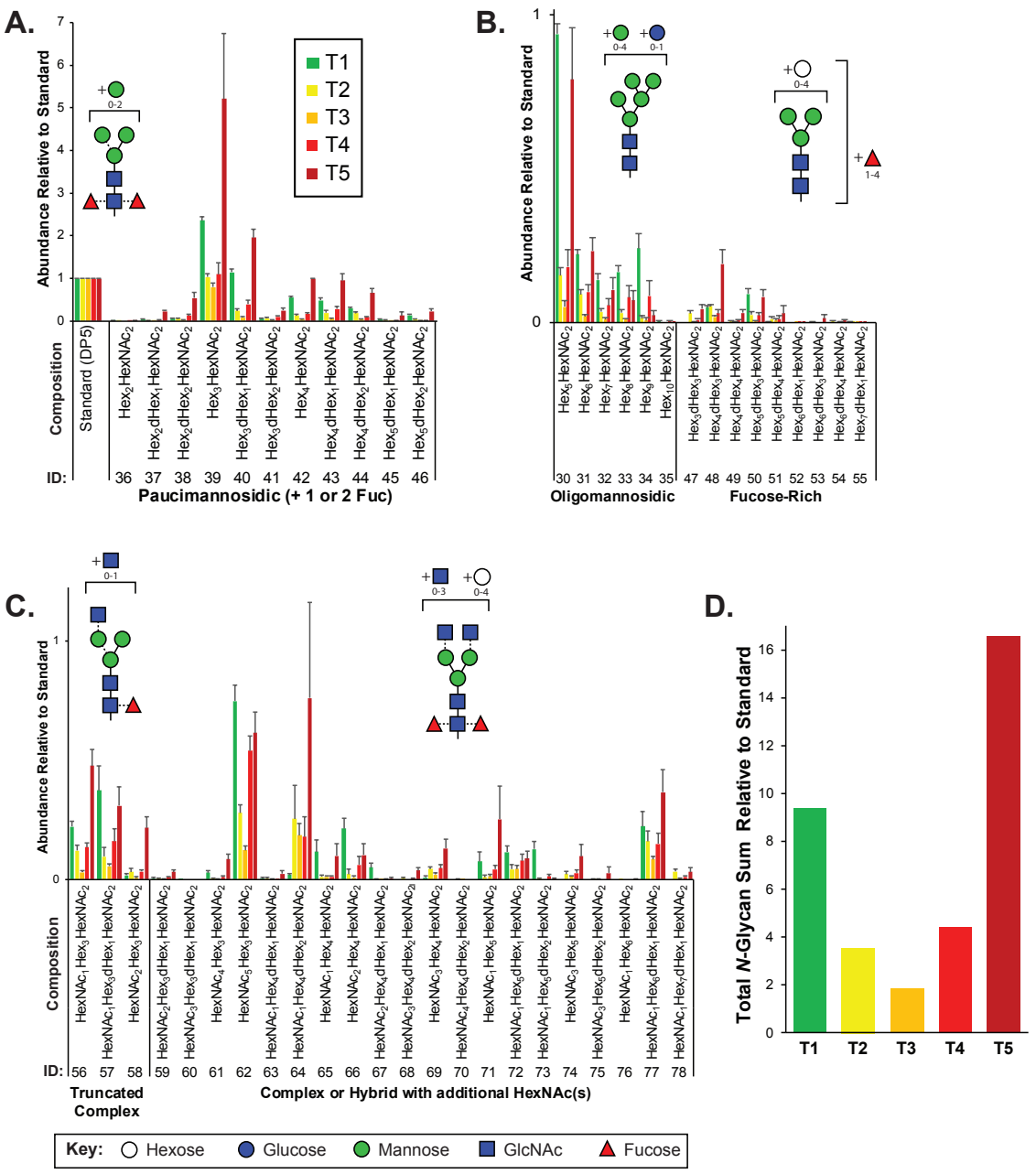


Figure 2.3: Sample Targeted *N*-glycomics of *C. elegans*. at various developmental time points. Presented are the relative abundances of indicated *N*-glycan compositions. Specific *N*-glycan subgroups are indicated in the following panels: (A) Paucimannosidic (B) Oligomannosidic/Fucose-rich (C) Truncated complex/Complex/Hybrid. Representative cartoon examples of each subgroup structure are shown as insets and displayed in accordance with the SNFG guidelines [42]. Glycan IDs are presented in Table 2.1. *N*-glycans are presented with the chitobiose core disaccharides ($GlcNAc_2$ or displayed as $HexNAc_2$) written towards the right. The sum for the relative *N*-glycan abundances at each time point is presented in Panel D. Each bar represents the sample mean relative to the internal standard (^{13}C-permthylated isomaltopentaose, DP_5) where n=7 and the error bars represent Standard Error of the Mean (SEM).

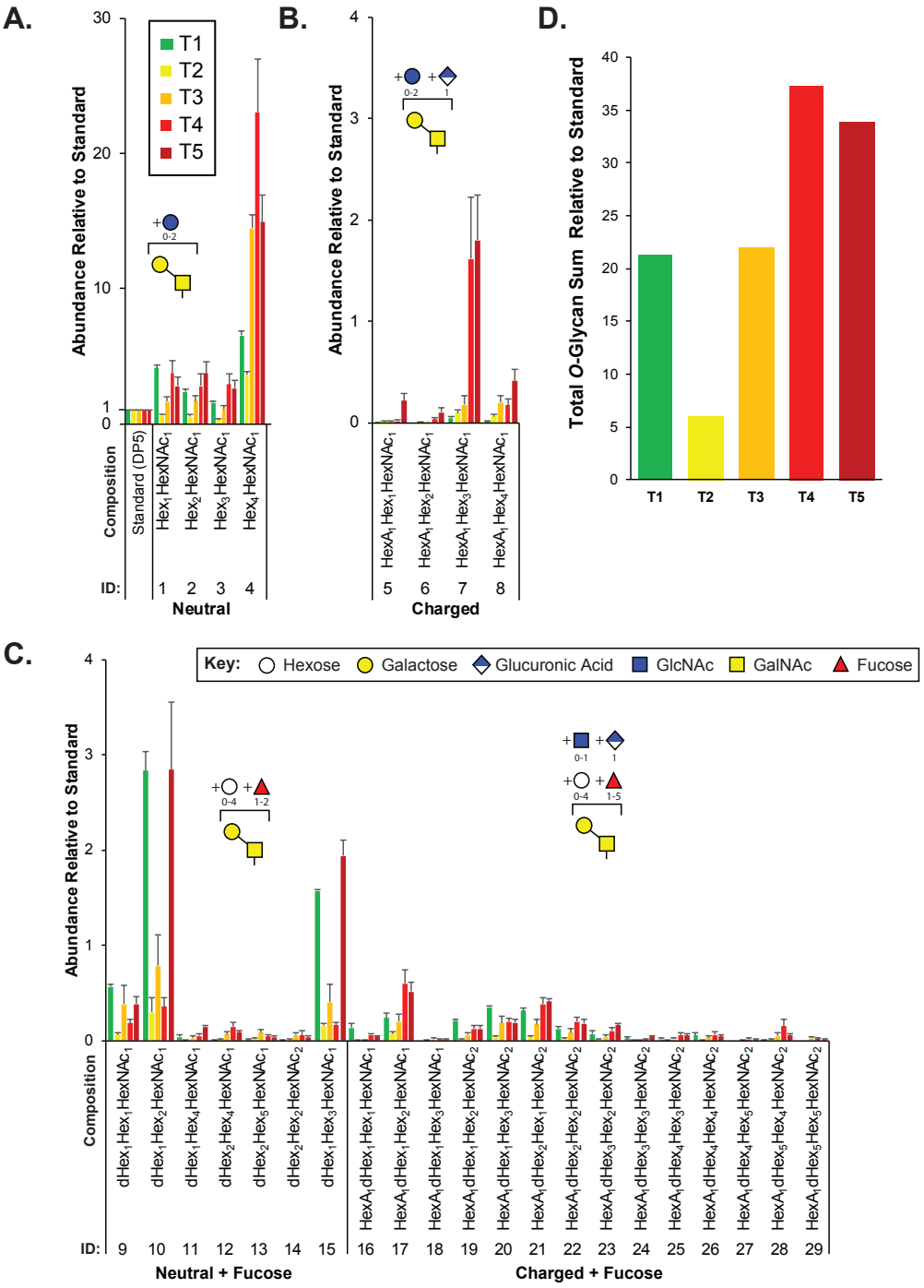


Figure 2.4: Targeted *O*-glycomics of *C. elegans* at various developmental time points. Presented are the relative abundances of indicated *O*-glycan compositions. Specific *O*-glycan subgroups are indicated in the following panels: (A) Neutral (B) Charged (C) Neutral or Charged + additional Fucoses. Representative cartoon examples of each subgroup structure are shown as insets and displayed in accordance with the SNFG guidelines [42]. Glycan IDs are presented in Table 2.1. The sum for the relative *O*-glycan abundances at each time point is presented in Panel D. Each bar represents the sample mean relative to the internal standard (^{13}C-permthylated isomaltopentaose, DP5) where n=7 and the error bars represent SEM.

2.3.2 Charged *O*-Glycan Expression Patterns Peak With Large Worms

We next examined changes in the *O*-glycome with development by using a similar targeted approach by generating a list of expected *O*-glycans reported previously for the analysis of LC-MS/MS data [41], [39]. Previous *O*-glycomic data for *C. elegans* were generated from mixed stage nematodes. Thus, we report the first analysis of *O*-glycans with respect to discrete sizes of worms. Glycan compositions identified in this study that are most consistent with *O*-glycan subgroups are summarized in Table 2.1 and Figure 2.4.

In agreement with previous studies, of the greatest abundance were the neutral, mucin-type core 1 *O*-glycans ($Hex_{1-4}HexNAc_1$, Figure 2.4A), followed by neutral *O*-glycans substituted with one or more fucoses (Figure 2.4C). Relatively minor amounts of charged *O*-glycan species defined by containing HexA (presumably GlcA, Figure 2.4B, C) with and without additional fucoses or extended are observed. Interestingly, most non-fucosylated charged *O*-glycans, which are lowest in abundance at T1 unlike most of the other glycans identified, appeared to peak in abundance at T4 and/or T5 that contain adult nematodes (Figure 2.4B). Finally, an evaluation of the total *O*-glycome is presented in Figure 2.4D.

2.3.3 Glycans Have Specific Developmental Patterns

The heatmap shown in Figure 2.5A represents an average over each sample within each timepoint (columns) for each glycan (rows). The colors in the heatmap indicate the degree of association between specific glycans and time points. Dark red indicates high levels at that time. We also provide supplementary data (Figure A.4) similar to Figure 2.5A that corresponds to individual replicates before averaging. The tree from HCA on the top of 2.5A shows that T5 is most closely related to T1, because the T5 samples contained offspring. The other time points group as expected. The tree on the left of 2.5A groups glycans and indicates structures that are most closely related through biosynthetic steps. Schematic 2.1 provides detailed steps used to create Biosorter correlation maps with specific glycans measured from the same samples (Figure 2.5). The distributions shown in Figure 2.5B were obtained by using the glycans indicated by numbers as driver peaks. The numbers for each glycan are provided in Table 2.1 and are also shown on the right-hand side of the glycan heatmap in Figure 2.5A. We also have shown the driver peaks as small white dots on 2.5A.

The color scale for the Biosorter correlation maps ranges from 1 to -1, which are Pearson correlation coefficients between the glycan driver and Biosorter position. The regions that are dark red in Figure 2.5B correlate most highly with that specific driver peak. For example, glycan 73 is only correlated with time point T1, and Table 2.1 indicates that glycan 73 is $HexNAc_1Hex_5dHex_2HexNAc_2$. The color-coding of the numbers to the right of the heatmap in 2.5A indicate that glycan 73 is *N*-linked. We tested all of the glycans as driver peaks and found similar distribution patterns within each of the dark red clusters shown in Figure 2.5A, but we are showing only the specific glycans for clarity.

Clearly, most of the glycans in our study show changes between time points, and we can use the Biosorter distributions to identify glycans that appear in different sizes of worms. For example, glycan 40 ($Hex_3dHex_1HexNAc_2$) is *N*-linked and is most strongly associated with T1 (L1 stage). However, there is also some correlation with larger animals, and this may indicate that this glycan is expressed embryonically

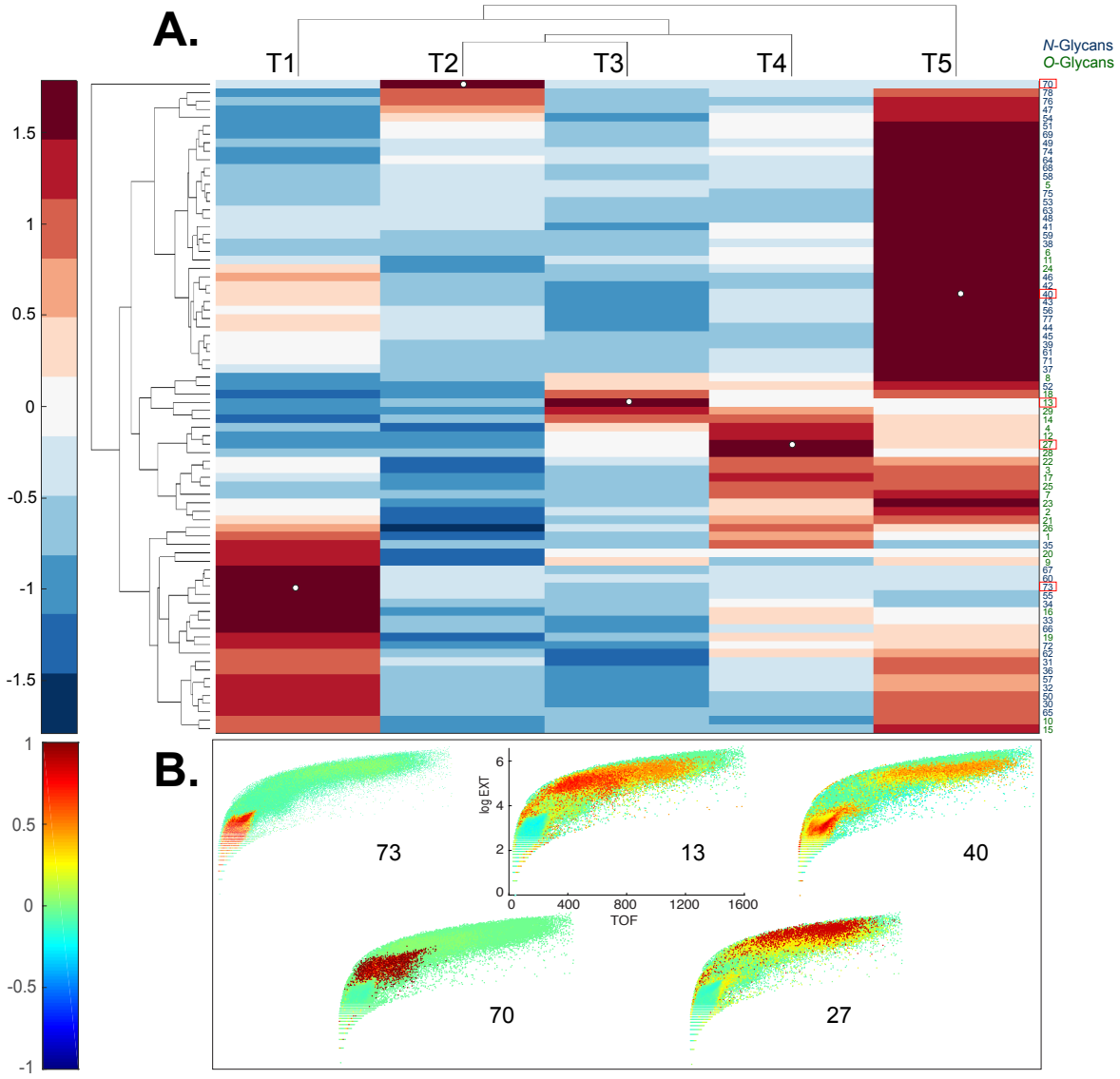


Figure 2.5: Developmental patterns of glycans in C.elegans. (A) Upper panel shows heatmap of glycan abundances, together with dendrogram of glycans (rows) and sample timepoints (columns). Glycan abundances were averaged over replicates in the same time point. A color bar of the heatmap is shown on the left. *N*-glycans IDs are shown in black and *O*-glycans are shown in red. (B) Lower panels show projections of Pearson correlation coefficients between normalized worm counts and glycan abundances on Biosorter maps. Number on each panel corresponds to the unique glycan, bolded and outlined in red in Panel A, that is used to calculate the correlation. A color bar of the correlation coefficients is shown on the left.

in the gravid adults. Glycan 70 ($HexNAc_4Hex_4dHex_2HexNAc_2$) is *N*-linked and is specifically expressed in T2. This glycan has a very low abundance but is clearly associated with a narrow stage of development.

2.3.4 Correlation Network Between Size, Glycans, and NMR Features

The supernatants of each sample from glycan analyses were also analyzed for metabolites by NMR spectroscopy. As noted above, the 7 replicates from T1 (L1 stage) were combined for NMR, but the other replicates have a one-to-one correspondence with LC-MS and Biosorting data. In this study, we focus exclusively on NMR resonances that statistically correlate with glycan data, as described in methods and shown in Figure A.2.

Similar to the approach described in Schematic 2.1, we first fused the normalized NMR and glycan data in MATLAB and performed SHY analysis [29] by systematically using all of the glycans as driver peaks for correlations in the NMR data (Figure A.2). The NMR resonances that correlated with glycans were then binned for network analysis. We attempted to use all the NMR data for this step but were unable to unravel the extensive correlation network that NMR data introduced (data not shown). We used a combination of 2D NMR with COLMARm for database matching and spiking with authentic samples for compound annotation and identification. Table 2.2 lists NMR metabolites and confidence scores reported in this study.

Table 2.2: NMR

Name	Cytoscape label[a]	Confidence[b]	^{1}H ppm (COLMARm or 1D)[c]	^{13}C ppm (COLMARm)[d]
UDP-GlcNAc	5.51	5	5.51, 7.93, 5.96, 5.97, 3.73, 4.36, 4.27, 2.06	
Cystathionine	3.13, 3.11	4	2.16, 2.73, 3.10, 3.85, 3.95	32.93, 29.88, 34.76, 56.48, 56.28
Trehalose	3.43, 3.65, 3.87	5	3.44, 3.64, 3.76, 3.84,3.84, 5.19	72.4, 73.76, 63.28, 63.27, 75.21, 95.99
Lactate	4.11	4	4.1, 1.32	71.23, 22.81
Glycerol	3.57, 3.54, 3.63	4	3.56, 3.63, 3.77	65.26, 65.26, 74.85
2-Aminoadipate	3.73	2	3.73, 2.23, 1.88, 1.82, 1.61, 1.65	N/A
Betaine	3.26, 3.90	4	3.26, 3.9	56.05, 68.84
UK-1	5.84		5.83	
Guanosine	5.91	2	5.9. 7.99	
UK-2	5.25, 5.38, 5.55,3.47, 3.70		5.25, 5.38, 5.54, 3.46, 3.69	N/A
Asparagine	2.89	4	2.88, 2.93, 2.99	37.33, 37.33, 54.01
UK-3	5.79		5.79	
Phosphorylcholine	4.17	4	3.19, 3.59, 4.16	56.67, 68.63, 60.83
UK-4	5.38		5.38	
NAD+	9.34	2	9.32, 9.12, 8.82, 8.4,8.19,8.16, 6.08, 6.02	
Glucose-6-phosphate	4.65, 5.23	2	4.64, 5.23, 4.0	
UK-5	9.57		9.57	

[a] Obtained from the centers of the interactively binned NMR data

[b] Confidence scale: 5) verified by spiking; 4) Matches in COLMARm using both HSQC and HSQC-TOCSY; 3) COLMARm matches using HSQC but not HSQC-TOCSY; 2) matches from 1D NMR to literature and/or database libraries; 1) for putatively characterized compounds or compound classes.

[c] For confidence level 4 and trehalose, ^{1}H chemical shift values were from COLMARm matched searches; For UDP-GlcNAc, ^{1}H shifts from synthetic, spiked standard; confidence level 2, ^{1}H shifts from 1D spectra.

[d] For confidence level 4 and trehalose, ^{13}C chemical shift values were from COLMARm matched searches.

Figure 2.6A shows the regions from the Biosorter distributions that were binned to represent different worm sizes (WS). This binning of sizes allows us to directly compare sizes rather than time points, which contain different mixtures of sizes. As described above, these binned regions cannot be ascribed to specific developmental stages without a more detailed image analysis of animals from each region of the Biosorter distributions, which is beyond the scope of this study. The smallest bin was chosen to overlap with L1 distributions shown in Figure A.1, and larger bins correspond to unique Biosorter regions from each time point.

These 3 sets of data—binned Biosorter sizes, glycans, and NMR features that correlate with glycans—were analyzed in Cytoscape to give the correlation network shown in Figure 2.6B. We adjusted the correlation value and found that r = 0.5 allowed for an interpretable number of nodes at each time point. We organized the network around Biosorter size nodes and colored positive and negative correlations as red and blue edges, respectively.

Figure 2.6B shows the correlation network of worm sizes (green hexagons), *N*-glycans (teal diamonds), *O*-glycans (teal hexagons), and NMR features (red hexagons) that correlate with glycans. The glycans have labels "NG_X" or "OG_Y" for *N*-glycan X or *O*-glycan Y, with numbers of the glycans from Table 2.1 or Figure 2.5A. The NMR resonances are labeled with chemical shift values; assignments (when known) and confidence scores provided in Table 2.2.

Several aspects of Figure 2.6B are noteworthy. First, with the exception of NG_62 (*N*-glycan 62 in Table 2.1), all *N*-glycans positively correlate with the smallest worm size (WS1), which corresponds to L1 animals. Only three *O*-glycans (#10, 15, 16) positively correlate with WS1. Three NMR features negatively correlate with WS1, including cystathionine (3.13 and 3.11 ppm) and lactate (1.33 ppm). Three NMR features positively correlate with WS1, glucose-6-phosphate (5.23 ppm) (overlapped and lower confidence score; Table 2.2) and two unknowns at 5.25 and 5.55 ppm. The second body size, WS2, has a striking pattern, because with the exception of UK-5 (9.57 ppm), all glycans and NMR features negatively correlate with this size. This includes a large number of *O*-glycans (#1, 3, 4, 17, 19, 20, 21, 22, 23), all of which except OG_20 also positively correlate to WS4. A group of NMR features negatively correlate with WS2, including UDP-GlcNAc (5.51 ppm), glucose-6-phosphate (5.23 ppm), betaine (3.90 ppm), trehalose (3.87 ppm), 2-aminoadipate (3.73 ppm), glycerol (3.57 ppm), and five unknowns (3.47, 3.70, 5.84 and 5.55 ppm). WS3 is very sparse in the network and only negatively correlates with two *N*-glycans (#31 and 62). WS4 is the largest body size and largely corresponds with adult animals. WS4 is positively correlated with many of the same *O*-glycans and NMR features that were negatively correlated to WS2. In addition, WS4 positively correlates to *N*-glycan 62, two specific *O*-glycans (#7 and 12), along with several NMR features, including betaine (3.26 ppm), trehalose (3.43, 3.65 ppm), glycerol (3.54, 3.64 ppm), lactate (4.11 ppm), glucose-6-phosphate (5.23 ppm), and guanosine (5.91 ppm).

We highlighted interactions involving two specific NMR features in Figure 2.6B, UDP-GlcNAc (5.51 ppm) and phosphorylcholine (PC: 4.17 ppm) by bolding the outlines of neighboring nodes and applying thick dashed or zig-zag lines for edges, respectively. UDP-GlcNAc (verified by spiking, Figure A.5) positively correlates with several *O*-glycans (#21, 19, 17, 16, 4, and 1) and a single *N*-glycan (#31). UDP-GlcNAc negatively correlates with WS2 but positively correlates with WS4 and positively correlates with

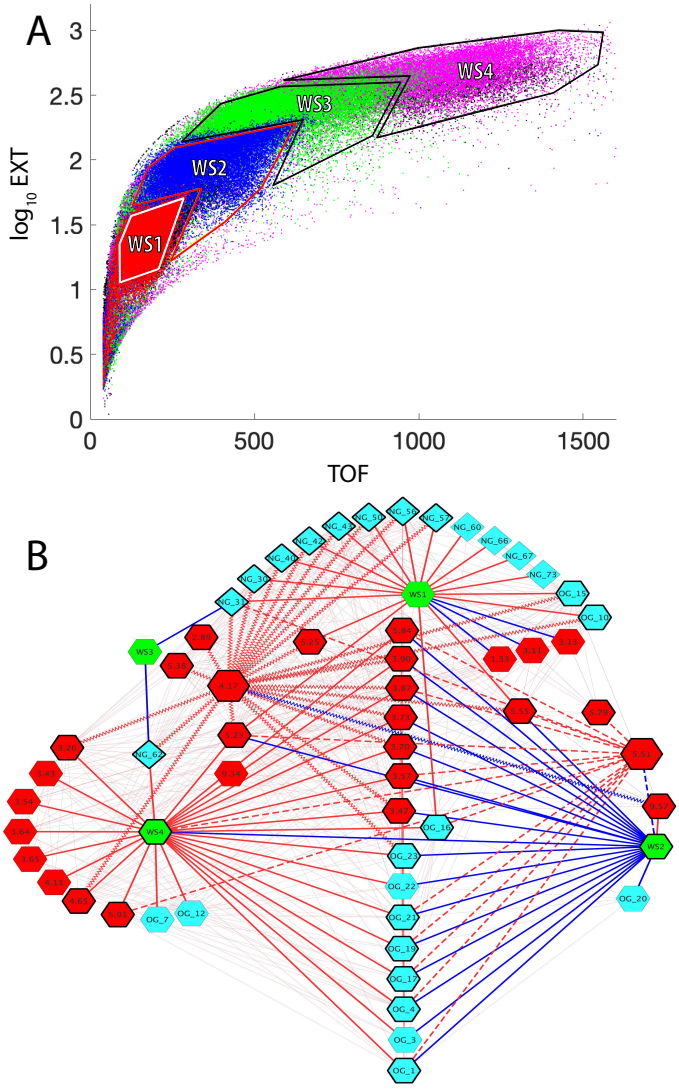


Figure 2.6: Correlation network between worm sizes, LC-MS-measured glycans, and NMR-measured metabolites. A) Superposition of Biosorter data (Supplementary Figure A.1) for all samples and time points (T1=red, T2=blue, T3=green, T4=magenta, T5=black). The data were plotted from T5 on the bottom to T1 on the top to emphasize unique regions. The four distinct resulting regions were binned and labeled for the worm size (WS1-WS4). WS1 corresponds to L1 arrested animals, while the other three stages are less precise. WS4 corresponds to the largest animals in the study and thus are primarily adult. B) Cytoscape correlation network between worm size (WS) regions from A (green hexagons), *N*-glycans (teal diamonds), *O*-glycans (teal hexagons), and NMR data (red hexagons). The numbers for the glycans correspond to Table 2.1. The numbers for NMR features correspond to chemical shift values from Table 2.2. Red and blue edges are positive and negative correlation values with a lower threshold of 0.5. The faint lines in the background include all correlations in the network. Solid bold edges connect direct neighbors from WS nodes to glycans and NMR nodes. Wavy bold lines are edges between phosphorylcholine (PC) (4.17 ppm, larger hexagon) and direct neighbors. Dashed bold lines are edges between UDP-GlcNAc (5.51 ppm, larger hexagon) and direct neighbors. The nodes in direct contact with either PC or UDP-GlcNAc are outlined with a bold black line.

several other unknown NMR features (5.79, 5.55, 5.84 ppm) as well as glycerol (3.57 ppm), guanosine (5.91 ppm), trehalose (3.87 ppm) and glycose-6-phosphate (5.23 ppm).

PC (4.17 ppm) also has interesting correlation patterns. With the exception of a negative correlation to UK-5 (9.57 ppm), all other correlations are positive. Most notable are several positive correlations to *N*-glycans (#30, 31, 40, 42, 43, 50, 56, and 57) that are associated with WS1 (L1 animals). It also positively correlates with *N*-glycan 62, which negatively correlates with WS3 and positively correlates with WS4. PC also positively correlates with the two *O*-glycans (#10 and 15) that positively associate with WS1. PC positively correlates with several unknown NMR features (5.25, 5.38, 3.70, 3.47, and 5.55 ppm), asparagine(2.89 ppm), glucose-6-phosphate (4.65 and 5.23 ppm), betaine (3.26 and 3.90 ppm), trehalose (3.87 ppm), and 2-aminoadipate (3.73 ppm).

2.4 Discussion

Our study has combined three different types of data collected from the same samples of *C. elegans*: Biosorter data, LC-MS-detected glycans, and NMR-detected metabolites. Important technical steps were implementing a protocol that samples a small percentage of a culture for Biosorting and utilizing the pellet after homogenization for glycan analysis. This provides the ability to conduct statistical correlations of different data types using the same biological replicates. Although interesting data were obtained individually, the most important findings were when data were combined.

The protocol to correlate Biosorter data with analytical data provided a key link in our study. Importantly, it is not limited to the LC-MS/MS glycan data. Any quantitative analytical data can be substituted for the glycan data outlined in Schematic 2.1. We think that the approach correlating Biosorter-based population distribution data will allow much more detailed omics studies in worms or other organisms like zebrafish or drosophila embryos, which can also be Biosorted. Although we have not yet examined the detailed biological replicate requirements, we are confident that developmental stage information can be extracted from mixed cultures by combining image analysis of worms isolated from different Biosorter regions, binning those Biosorter regions, and using them as driver peaks for a variety of quantitative omics data, including RNAseq, proteomics, and untargeted LC-MS metabolomics. Moreover, we see no reason why this approach could not also be used in flow cytometry analysis of cells, which would make it even more broadly relevant.

By utilizing previously identified glycan compositions reported in *C. elegans*, we have carefully analyzed released and permethylated *N*- (Figure 2.3) and *O*-glycans (Figure 2.4) using a high-throughput strategy by LC-MS/MS. The relative abundances for certain *N*-glycans compositions were consistent with previous reports, as described above [37], [38], [45], [40], [13]. In particular, to assess global changes in the total *N*-glycome with development, relative glycan abundances were summed for each developmental time point (Figure 2.3D). In agreement with Cipollo et al. [21], *N*-glycan abundances follow the trend of being highest in the early L1 larval stages (T1 and mix with T5) with relatively lower amounts within the intermediate time points (T2, T3, and T4). Thus, our results recapitulate the dynamic *N*-glycan utilization during development with approximately 4- to 8-fold differences in total *N*-glycan abundances when

comparing L1 to intermediate stages. This trend is even more striking when we use unique regions of the Biosorter distribution to correlate worm size with glycans (Figure 2.6B), because nearly every *N*-glycan is positively correlated with the smallest L1 animals.

An NMR resonance from phosphorylcholine (PC) (4.17 ppm, Figure 2.6B) was correlated with the *N*-glycans (#30, 31, 40, 42, 43, 50, 56, and 57) that were correlated to L1 animals (WS1, Figure 2.6B) and the single *N*-glycan (#62) that was correlated with the largest size worms (WS4, Figure 2.6B). It has been previously demonstrated that *N*-Glycans of *C. elegans* can be elaborated by PC, and that this modification is stage specific, being detected in L1, L4, adult, and dauer [48], [21], [37], in good agreement with our study. While we did not search for any low abundance PC-substituted *N*-glycans, PC does cluster with the *N*-glycan subgroups that may be modified with it, which suggests that the highly correlated *N*-glycans we identified by LC-MS/MS may be potential acceptor substrates in which PC modifies by a yet-to-be-identified transferase. Furthermore, the developmental PC correlation may be significant to the modification of glycolipids, which were not analyzed in this report, but have been implicated in specific stages of development and embryogenesis [49].

The total *O*-glycan relative abundances followed a different pattern than *N*-glycans during development, with the approximate pattern $T4 \approx T5 > T1 \approx T3 > T2$ (Figure 2.4D). As this pattern would be mostly dominated by the most abundant structures, superficial inspection shows the general trend holds true with most glycans identified in this study, except for several low abundance fucosylated neutral glycans (#9, 10, 13, and 15) that followed a pattern with the greatest levels in T1, T3, and T5 (Figure 2.4C). The pattern of *O*-glycan correlation with worm sizes in Figure 2.6B shows a complex pattern in which several *O*-glycans are negatively correlated with WS2 but positively correlated with WS4. This pattern suggests a switch during development from smaller to larger worms. In most animals, a null deletion of OGT—the glycosyltransferase responsible for adding *O*-GlcNAc to proteins [50], [51]—is embryonic lethal. However, in *C. elegans*, *ogt-1* null animals are viable, and these animals accumulate UDP-GlcNAc [52], [53], which is the substrate for OGT. Moreover, the modENCODE [54] expression of *ogt-1* RNA reported on WormBase [55] shows a steady decline of expression from the highest levels in early embryonic stages to the lowest in L4 and young adult. These observations are both consistent with the network in Figure 2.6B, which shows UDP-GlcNAc levels measured by NMR positively correlate with many of the same *O*-glycans that have the reciprocal correlation pattern to WS2 and WS4. This suggests a relationship between increases in UDP-GlcNAc and decreases in *ogt-1* gene expression. Perhaps this represents a switch from primary utilization of *O*-GlcNAc for dynamic protein *O*-GlcNAc-ylation mediated by OGT at earlier stages to the biosynthesis of complex mucin-like *O*-glycans at adult stages. In more complex *O*-glycan biosynthesis, UDP-GlcNAc is converted to UDP-GalNAc by UDP-*N*-acetylglucosamine 4'-epimerase (GalE). We examined our NMR data for UDP-GalNAc but were unable to unambiguously assign it. A likely NMR resonance of guanosine (5.91 ppm, Figure 2.6B; Table 2.2) positively correlates with large worms (WS4) and UDP-GlcNAc, suggesting a relationship between the *O*-glycans that positively correlate with the same large worms. GDP-fucose is the sugar nucleotide donor for fucosyltransferases that generate these structures. GDP-fucose is synthesized from GDP-mannose in *C. elegans* [56]. Fucosylation is required for normal development [57].

Future studies will aim to adapt our workflow to analyze natively methylated structures by using ^{13}C-permethylation. In this report, we have broadly established the dynamic *N*- and *O*-glycome, and these data could be utilized in tracking dynamic changes in the glycoproteome or of key glycoproteins throughout developmental transitions including *O*-GlcNAc modified proteins. While the work presented here represents a pilot study for combining different omic data to better understand development in *C. elegans*, it could easily be expanded/adapted to capture additional –omic datasets (transcriptomic, proteomic, and lipidomic) and/or applied to studies aimed at uncovering the impact of genetic and environmental perturbations.

Bibliography

[1] H. H. Freeze, G. W. Hart, and R. L. Schnaar, "Glycosylation precursors," in *Essentials of Glycobiology*, rd, A. Varki, R. D. Cummings, *et al.*, Eds. Cold Spring Harbor (NY), 2015, pp. 51–63. DOI: 10.1101/glycobiology.3e.005. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28876856.

[2] N. E. Zachara and G. W. Hart, "O-glcnac a sensor of cellular state: The role of nucleocytoplasmic glycosylation in modulating cellular function in response to nutrition and stress," *Biochim Biophys Acta*, vol. 1673, no. 1-2, pp. 13–28, 2004, ISSN: 0006-3002 (Print) 0006-3002 (Linking). DOI: 10.1016/j.bbagen.2004.03.016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15238246%20https://ac.els-cdn.com/S0304416504001023/1-s2.0-S0304416504001023-main.pdf?_tid=87f97c13-db14-43c3-9af1-84bbb4afce73&acdnat=1545147537_a32c2d3b9c88eab7e6b5f0e713a0d0a2.

[3] N. E. Zachara, "Critical observations that shaped our understanding of the function(s) of intracellular glycosylation (o-glcnac)," *FEBS Lett*, vol. 592, no. 23, pp. 3950–3975, 2018, ISSN: 1873-3468 (Electronic) 0014-5793 (Linking). DOI: 10.1002/1873-3468.13286. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30414174.

[4] K. Vaidyanathan and L. Wells, "Multiple tissue-specific roles for the o-glcnac post-translational modification in the induction of and complications arising from type ii diabetes," *J Biol Chem*, vol. 289, no. 50, pp. 34 466–71, 2014, ISSN: 1083-351X (Electronic) 0021-9258 (Linking). DOI: 10.1074/jbc.R114.591560. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25336652.

[5] P. Stanley, N. Taniguchi, and M. Aebi, "N-glycans," in *Essentials of Glycobiology*, rd, A. Varki, R. D. Cummings, *et al.*, Eds. Cold Spring Harbor (NY), 2015, pp. 99–111. DOI: 10.1101/glycobiology.3e.009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28876855.

[6] I. Brockhausen and P. Stanley, "O-galnac glycans," in *Essentials of Glycobiology*, rd, A. Varki, R. D. Cummings, *et al.*, Eds. Cold Spring Harbor (NY), 2015, pp. 113–123. DOI: 10.1101/glycobiology.3e.010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28876832.

[7] S. H. von Reuss, N. Bose, J. Srinivasan, *et al.*, "Comparative metabolomics reveals biogenesis of ascarosides, a modular library of small-molecule signals in c. elegans," *Journal of the American Chemical Society*, vol. 134, no. 3, pp. 1817–1824, 2012. DOI: 10.1021/ja210202y. [Online].

Available: http://www.ncbi.nlm.nih.gov/pubmed/22239548%20http://pubs.acs.org/doi/pdfplus/10.1021/ja210202y.

[8] A. H. Ludewig and F. C. Schroeder, "Ascaroside signaling in c. elegans.," *WormBook : the online review of C. elegans biology*, pp. 1–22, 2013. DOI: 10.1895/wormbook.1.155.1. [Online]. Available: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23355522&retmode=ref&cmd=prlinks.

[9] J. Srinivasan, F. Kaplan, R. Ajredini, *et al.*, "A blend of small molecules regulates both mating and development in caenorhabditis elegans," *Nature*, vol. 454, no. 7208, pp. 1115–8, 2008, ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: 10.1038/nature07168. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18650807.

[10] M. Witting, J. Hastings, N. Rodriguez, *et al.*, "Modeling meets metabolomics-the wormjam consensus model as basis for metabolic studies in the model organism caenorhabditis elegans," *Front Mol Biosci*, vol. 5, p. 96, 2018, ISSN: 2296-889X (Print) 2296-889X (Linking). DOI: 10.3389/fmolb.2018.00096. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30488036.

[11] S. H. von Reuss and F. C. Schroeder, "Combinatorial chemistry in nematodes: Modular assembly of primary metabolism-derived building blocks," *Nat Prod Rep*, vol. 32, no. 7, pp. 994–1006, 2015, ISSN: 1460-4752 (Electronic) 0265-0568 (Linking). DOI: 10.1039/c5np00042d. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26059053.

[12] J. Srinivasan, S. H. von Reuss, N. Bose, *et al.*, "A modular library of small molecule signals regulates social behaviors in caenorhabditis elegans," *PLoS Biol*, vol. 10, no. 1, e1001237, 2012, ISSN: 1545-7885 (Electronic) 1544-9173 (Linking). DOI: 10.1371/journal.pbio.1001237. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22253572.

[13] K. Paschinger, M. Gutternigg, D. Rendic, and I. B. Wilson, "The n-glycosylation pattern of caenorhabditis elegans," *Carbohydr Res*, vol. 343, no. 12, pp. 2041–9, 2008, ISSN: 0008-6215 (Print) 0008-6215 (Linking). DOI: 10.1016/j.carres.2007.12.018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18226806.

[14] P. J. Hu, "Dauer," *WormBook : the online review of C. elegans biology*, pp. 1–19, 2007. DOI: 10.1895/wormbook.1.144.1. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17988074.

[15] J. E. Sulston and H. R. Horvitz, "Post-embryonic cell lineages of nematode, caenorhabditis-elegans," *Developmental Biology*, vol. 56, no. 1, pp. 110–156, 1977.

[16] M. Chalfie, Y. Tu, G. Euskirchen, W. W. Ward, and D. C. Prasher, "Green fluorescent protein as a marker for gene expression," *Science*, vol. 263, no. 5148, pp. 802–805, 1994. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8303295.

[17] F. Kaplan, D. V. Badri, C. Zachariah, *et al.*, "Bacterial attraction and quorum sensing inhibition in caenorhabditis elegans exudates," *J Chem Ecol*, vol. 35, no. 8, pp. 878–92, 2009, ISSN: 1573-1561 (Electronic) 0098-0331 (Linking). DOI: 10.1007/s10886-009-9670-0. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19649780.

[18] F. Kaplan, J. Srinivasan, P. Mahanti, *et al.*, "Ascaroside expression in caenorhabditis elegans is strongly dependent on diet and developmental stage," *PLoS One*, vol. 6, no. 3, e17804, 2011, ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0017804. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21423575.

[19] H. Morio, Y. Honda, H. Toyoda, M. Nakajima, H. Kurosawa, and T. Shirasawa, "Ext gene family member rib-2 is essential for embryonic development and heparan sulfate biosynthesis in caenorhabditis elegans," *Biochem Biophys Res Commun*, vol. 301, no. 2, pp. 317–23, 2003, ISSN: 0006-291X (Print) 0006-291X (Linking). [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/12565862.

[20] N. Kanaki, A. Matsuda, K. Dejima, *et al.*, "Udp-n-acetylglucosamine-dolichyl-phosphate n-acetylglucosaminephos is indispensable for oogenesis, oocyte-to-embryo transition, and larval development of the nematode caenorhabditis elegans," *Glycobiology*, 2018, ISSN: 1460-2423 (Electronic) 0959-6658 (Linking). DOI: 10.1093/glycob/cwy104. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30445613.

[21] J. F. Cipollo, A. M. Awad, C. E. Costello, and C. B. Hirschberg, "N-glycans of caenorhabditis elegans are specific to developmental stages," *J Biol Chem*, vol. 280, no. 28, pp. 26 063–72, 2005, ISSN: 0021-9258 (Print) 0021-9258 (Linking). DOI: 10.1074/jbc.M503828200. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15899899.

[22] M. Sud, E. Fahy, D. Cotter, *et al.*, "Metabolomics workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools," *Nucleic Acids Res*, vol. 44, no. D1, pp. D463–70, 2016, ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gkv1042. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26467476.

[23] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax, "Nmrpipe - a multidimensional spectral processing system based on unix pipes," *Journal of Biomolecular NMR*, vol. 6, no. 3, pp. 277–293, 1995.

[24] S. L. Robinette, R. Ajredini, H. Rasheed, *et al.*, "Hierarchical alignment and full resolution pattern recognition of 2d nmr spectra: Application to nematode chemical ecology," *Anal Chem*, vol. 83, no. 5, pp. 1649–57, 2011, ISSN: 1520-6882 (Electronic) 0003-2700 (Linking). DOI: 10.1021/ac102724x. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21314130.

[25] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn, "Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in h-1 nmr metabonomics," *Analytical Chemistry*, vol. 78, no. 13, pp. 4281–4290, 2006, ISSN: 0003-2700. DOI: 10.1021/ac051632c. [Online]. Available: %3CGo%20to%20ISI%3E://WOS:000238665200012.

[26] N. P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping," *Journal of Chromatography A*, vol. 805, no. 1-2, pp. 17–35, 1998, ISSN: 0021-9673. DOI: Doi10.1016/S0021-9673(98)00021-1. [Online]. Available: %3CGo%20to%20ISI%3E://WOS:000073710200002.

[27] J. W. Wong, C. Durante, and H. M. Cartwright, "Application of fast fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets," *Anal Chem*, vol. 77, no. 17, pp. 5655–61, 2005, ISSN: 0003-2700 (Print) 0003-2700 (Linking). DOI: 10.1021/ac050619p. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16131078.

[28] O. Cloarec, M. E. Dumas, A. Craig, *et al.*, "Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic 1h nmr data sets," *Analytical Chemistry*, vol. 77, no. 5, pp. 1282–1289, 2005. DOI: 10.1021/ac048630x. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15732908%20http://pubs.acs.org/doi/pdfplus/10.1021/ac048630x.

[29] D. J. Crockford, E. Holmes, J. C. Lindon, *et al.*, "Statistical heterospectroscopy, an approach to the integrated analysis of nmr and uplc-ms data sets: Application in metabonomic toxicology studies," *Analytical Chemistry*, vol. 78, no. 2, pp. 363–371, 2006, ISSN: 0003-2700. DOI: 10.1021/ac051444m. [Online]. Available: %3CGo%20to%20ISI%3E://WOS:000234826400001.

[30] K. Bingol, D. W. Li, B. Zhang, and R. Bruschweiler, "Comprehensive metabolite identification strategy using multiple two-dimensional nmr spectra of a complex mixture implemented in the colmarm web server," *Anal Chem*, vol. 88, no. 24, pp. 12 411–12 418, 2016, ISSN: 1520-6882 (Electronic) 0003-2700 (Linking). DOI: 10.1021/acs.analchem.6b03724. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28193069.

[31] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, *et al.*, "Biomagresbank," *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D402–8, 2008. DOI: 10.1093/nar/gkm957. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17984079.

[32] D. S. Wishart, D. Tzur, C. Knox, *et al.*, "Hmdb: The human metabolome database," *Nucleic Acids Res*, vol. 35, pp. D521–D526, 2007. [Online]. Available: http://nar.oxfordjournals.org/cgi/content/abstract/35/suppl_1/D521.

[33] K. Aoki, M. Perlman, J. M. Lim, R. Cantu, L. Wells, and M. Tiemeyer, "Dynamic developmental elaboration of n-linked glycan complexity in the drosophila melanogaster embryo," *J Biol Chem*, vol. 282, pp. 9127–9142, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17264077.

[34] H. Schachter, "Paucimannose n-glycans in caenorhabditis elegans and drosophila melanogaster," *Carbohydr Res*, vol. 344, no. 12, pp. 1391–6, 2009, ISSN: 1873-426X (Electronic) 0008-6215 (Linking). DOI: 10.1016/j.carres.2009.04.028. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19515361.

[35] K. Aoki, M. Porterfield, S. Lee, *et al.*, "The diversity of o-linked glycans expressed during drosophila melanogaster development reflects stage- and tissue-specific requirements for cell signaling," *J Biol Chem*, vol. 283, pp. 30 385–30 400, 2008.

[36] K. Anumula and P. Taylor, "A comprehensive procedure for preparation of partially methylated alditol acetates from glycoprotein carbohydrates," *Analytical Biochemistry*, vol. 203, pp. 101–108, 1992.

[37] J. F. Cipollo, C. E. Costello, and C. B. Hirschberg, "The fine structure of caenorhabditis elegans n-glycans," *J Biol Chem*, vol. 277, no. 51, pp. 49 143–57, 2002, ISSN: 0021-9258 (Print) 0021-9258 (Linking). DOI: 10.1074/jbc.M208020200. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/12361949.

[38] H. Geyer, M. Schmidt, M. Muller, R. Schnabel, and R. Geyer, "Mass spectrometric comparison of n-glycan profiles from caenorhabditis elegans mutant embryos," *Glycoconj J*, vol. 29, no. 2-3, pp. 135–45, 2012, ISSN: 1573-4986 (Electronic) 0282-0080 (Linking). DOI: 10.1007/s10719-012-9371-8. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22407488.

[39] Y. Guerardel, L. Balanzino, E. Maes, *et al.*, "The nematode caenorhabditis elegans synthesizes unusual o-linked glycans: Identification of glucose-substituted mucin-type o-glycans and short chondroitin-like oligosaccharides," *Biochem J*, vol. 357, no. Pt 1, pp. 167–82, 2001, ISSN: 0264-6021 (Print) 0264-6021 (Linking). [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11415447.

[40] S. Natsuka, J. Adachi, M. Kawaguchi, *et al.*, "Structural analysis of n-linked glycans in caenorhabditis elegans," *J Biochem*, vol. 131, no. 6, pp. 807–13, 2002, ISSN: 0021-924X (Print) 0021-924X (Linking). [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/12038976.

[41] L. M. Parsons, R. M. Mizanur, E. Jankowska, *et al.*, "Caenorhabditis elegans bacterial pathogen resistant bus-4 mutants produce altered mucins," *PLoS One*, vol. 9, no. 10, e107250, 2014, ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0107250. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25296196.

[42] A. Varki, R. D. Cummings, M. Aebi, *et al.*, "Symbol nomenclature for graphical representations of glycans," *Glycobiology*, vol. 25, no. 12, pp. 1323–4, 2015, ISSN: 1460-2423 (Electronic) 0959-6658 (Linking). DOI: 10.1093/glycob/cwv091. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26543186.

[43] P. Shannon, A. Markiel, O. Ozier, *et al.*, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res*, vol. 13, no. 11, pp. 2498–504, 2003, ISSN: 1088-9051 (Print) 1088-9051 (Linking). DOI: 10.1101/gr.1239303. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/14597658.

[44] S. M. Haslam and A. Dell, "Hallmarks of caenorhabditis elegans n-glycosylation: Complexity and controversy," *Biochimie*, vol. 85, no. 1-2, pp. 25–32, 2003, ISSN: 0300-9084 (Print) 0300-9084 (Linking). [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/12765772.

[45] S. M. Haslam, D. Gems, H. R. Morris, and A. Dell, "The glycomes of caenorhabditis elegans and other model organisms," *Biochem Soc Symp*, no. 69, pp. 117–34, 2002, ISSN: 0067-8694 (Print) 0067-8694 (Linking). [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/12655779.

[46] I. B. Wilson, "The class i alpha1,2-mannosidases of caenorhabditis elegans," *Glycoconj J*, vol. 29, no. 4, pp. 173–9, 2012, ISSN: 1573-4986 (Electronic) 0282-0080 (Linking). DOI: 10.1007/s10719-012-9378-1. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22535467.

[47] F. Altmann, G. Fabini, H. Ahorn, and I. B. Wilson, "Genetic model organisms in the study of n-glycans," *Biochimie*, vol. 83, no. 8, pp. 703–12, 2001, ISSN: 0300-9084 (Print) 0300-9084 (Linking). [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11530201.

[48] J. F. Cipollo, A. Awad, C. E. Costello, P. W. Robbins, and C. B. Hirschberg, "Biosynthesis in vitro of caenorhabditis elegans phosphorylcholine oligosaccharides," *Proc Natl Acad Sci U S A*, vol. 101, no. 10, pp. 3404–8, 2004, ISSN: 0027-8424 (Print) 0027-8424 (Linking). DOI: 10.1073/pnas.0400384101. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/14993596.

[49] S. Gerdt, R. D. Dennis, G. Borgonie, R. Schnabel, and R. Geyer, "Isolation, characterization and immunolocalization of phosphorylcholine-substituted glycolipids in developmental stages of caenorhabditis elegans," *Eur J Biochem*, vol. 266, no. 3, pp. 952–63, 1999, ISSN: 0014-2956 (Print) 0014-2956 (Linking). [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/10583390.

[50] R. S. Haltiwanger, W. G. Kelly, E. P. Roquemore, *et al.*, "Glycosylation of nuclear and cytoplasmic proteins is ubiquitous and dynamic," *Biochem Soc Trans*, vol. 20, no. 2, pp. 264–9, 1992, ISSN: 0300-5127 (Print) 0300-5127 (Linking). [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/1397609.

[51] R. S. Haltiwanger, G. D. Holt, and G. W. Hart, "Enzymatic addition of o-glcnac to nuclear and cytoplasmic proteins. identification of a uridine diphospho-n-acetylglucosamine:peptide beta-n-acetylglucosaminyltransferase," *J Biol Chem*, vol. 265, no. 5, pp. 2563–8, 1990, ISSN: 0021-9258 (Print) 0021-9258 (Linking). [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/2137449%20http://www.jbc.org/content/265/5/2563.full.pdf.

[52] S. K. Ghosh, M. R. Bond, D. C. Love, G. G. Ashwell, M. W. Krause, and J. A. Hanover, "Disruption of o-glcnac cycling in c. elegans perturbs nucleotide sugar pools and complex glycans," *Front Endocrinol (Lausanne)*, vol. 5, p. 197, 2014, ISSN: 1664-2392 (Print) 1664-2392 (Linking). DOI: 10.3389/fendo.2014.00197. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25505447.

[53] M. M. Rahman, O. Stuchlick, E. G. El-Karim, R. Stuart, E. T. Kipreos, and L. Wells, "Intracellular protein glycosylation modulates insulin mediated lifespan in c.elegans," *Aging (Albany NY)*, vol. 2, no. 10, pp. 678–90, 2010, ISSN: 1945-4589 (Electronic) 1945-4589 (Linking). DOI: 10.18632/aging.100208. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20952811.

[54] S. E. Celniker, L. A. Dillon, M. B. Gerstein, *et al.*, "Unlocking the secrets of the genome," *Nature*, vol. 459, no. 7249, pp. 927–30, 2009, ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: 10.1038/459927a. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19536255%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2843545/pdf/nihms135877.pdf.

[55] L. Stein, P. Sternberg, R. Durbin, J. Thierry-Mieg, and J. Spieth, "Wormbase: Network access to the genome and biology of caenorhabditis elegans," *Nucleic Acids Res*, vol. 29, no. 1, pp. 82–6, 2001, ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11125056.

[56] S. Rhomberg, C. Fuchsluger, D. Rendic, *et al.*, "Reconstitution in vitro of the gdp-fucose biosynthetic pathways of caenorhabditis elegans and drosophila melanogaster," *FEBS J*, vol. 273, no. 10, pp. 2244–56, 2006, ISSN: 1742-464X (Print) 1742-464X (Linking). DOI: 10.1111/j.1742-4658.2006.05239.x. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16650000.

[57] O. Menzel, T. Vellai, K. Takacs-Vellai, *et al.*, "The caenorhabditis elegans ortholog of c21orf80, a potential new protein o-fucosyltransferase, is required for normal development," *Genomics*, vol. 84, no. 2, pp. 320–30, 2004, ISSN: 0888-7543 (Print) 0888-7543 (Linking). DOI: 10.1016/j.ygeno.2004.04.002. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15233996.

Chapter 3
PHIT: A NOVEL ALGORITHM FOR IMPROVING NMR SPECTRAL ALIGNMENT BY SPECTRAL REORDERING AND CURVE TRACING

3.1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is widely used for metabolomics studies. Nuclei resonate at characteristic frequencies in NMR spectra as a result of the different microenvironments made by their surrounding electrons, resulting in distinguishable and largely stable NMR signals across different samples. However, matrix effects such as pH and ionic strength may cause additional perturbations on certain resonances through changes of ionization and other factors. Therefore, peaks from the same metabolite may have different frequencies in different samples if the samples vary in pH or salt. To ensure we are comparing the same features when analyzing spectra from multiple samples, we need to either correct for this variation or match peaks across samples.

Current widely used strategies for solving this problem include aligning and binning. Aligning methods aim to adjust peaks from the same nucleus into the same position for all spectra. Popular alignment strategies include Pearson correlation-based alignment (e.g. Constrained Correlation Optimized Warping (CCOW)[1]), FFT cross-correlation based alignment (e.g. PAFFT and RAFFT[2]), distance-based alignment, and others[3]. On the other hand, binning methods tackle this problem by reducing data resolution. They calculate the integration of peaks in a certain region. By expanding the region to encompass all variable peaks in a given region, the differences in peak position are eliminated. The bin sizes can be uniform or adapted for each signal, but the bins are consistent across spectra.

Aligning and binning methods can usually deal with peaks with small variations, e.g. the variation is smaller than half of the peak width. However, their performances are not satisfactory when peaks are highly variable or when the spectra are crowded. Even when efforts are made to buffer the samples, highly

variable peaks are often observed in experimental datasets. This fact may cause artifacts in processed data or cause researchers to exclude these peaks from analyses. This will then undoubtably introduce errors or lose information for the study.

The greatest contributor to the variation of chemical shifts in experiments are pH and the concentration of Ca^{2+} and Mg^{2+}[4] [5] [6][7]. Although buffers are added to the samples, it is often not adequate to prevent inconsistencies across an entire metabolomics study. This phenomenon particularly exists in urine samples. Unlike blood or tissue samples that require homeostasis, urine itself does not utilize physiological buffers. Also, since urine normally does not contain macromolecules, extraction and reconstitution are not needed for sample preparation for these samples. While extraction and reconstitution may buffer the samples well, these steps make sample preparation more complicated and introduce additional variation. In addition, urine samples produce a large number of NMR peaks, thus increasing the difficulty of alignment and binning. Therefore, alignment problems are common among urine samples, and there are currently no ideal solutions.

While it is difficult to eliminate peak variation across samples, we can utilize the same variation for matching peaks. Peak positions change monotonically with pH, the concentration of Ca^{2+}, and Mg^{2+} when there is no interaction between nuclei on different ionization sites, which is often the case[8] [4] [9]. Therefore, if the chemical shifts of two signals are affected by the same factor (such as pH), the changing relationship between the two signals will also be monotonic, as long as other factors are not contributing to the variation. If the spectra are reordered according to the position of one of these signals (here named the guiding signal), then other signals responding to the same variable (named the responding signal) will also have a monotonically changing curve across the spectra. This curve will then make the responding peaks distinguishable from other surrounding peaks and will make peak alignment possible. By reordering the spectra according to the position of one internal peak rather than according to some defined factors (such as pH), we can avoid extra measurements on these factors.

L. Csenki et al. [10] and M. Liebeke et al. [7] have utilized this strategy to extract features from NMR spectra. By using a generalized fuzzy Hough transform (GFHT), Csenki et al. (2007) aligned peaks that were close to the linearly scaled shape of the guiding peak. However, the shape of the guiding and responding peaks may not be the same because of different pK_a values [9]. Although Alm et al. further extended this method to use the linear combination shape of several guiding peaks, this method still relies on the shape of the guiding peak to align the responding peak [11]. On the other hand, Liebeke et al. (2013) used computer-assisted manual tracing of the curve *a prior* to align the responding peaks[7]. While this method does not rely on the shape of the guiding peaks, it does require more user input.

Based on the issues detailed above, I introduce an algorithm that we are calling pH-guided Iterative Tracing (pHIT) alignment that traces and aligns the responding signal on the reordered spectra by using the information only from the responding curve. It detects and removes off-curve outliers by density-based clustering, then detects missing peaks inside a 95% confidence interval according to the positions of the on-curve peaks. I demonstrate that this approach can effectively distinguish neighboring surrounding peaks and align the responding signal on several different examples of urine NMR metabolomics datasets.

3.2 Methods

3.2.1 Dataset

Three experimental one-dimensional ^{1}H NMR datasets were used to test and evaluate the performance of this algorithm. These include two experimental datasets from human and dog urine metabolomics samples and one published dataset[12] from wine samples.

Experiment 1. This dataset is from the experiment discussed in Chapter 4. Briefly, it contained 121 ^{1}H NMR human urine spectra measured at 600 MHz magnetic field strength. The experiment was designed to study the longitudinal metabolic differences of pregnant women with and without Zika virus infection (metabolomics workbench DOI: http://dx.doi.org/10.21228/M8B13G). This dataset contained spectra from 104 experimental samples obtained from 20 individuals, 8 pooled internal control samples, as well as 9 external control samples from commercially available quality control human urine (Golden West Biologicals, Inc). Details of spectra acquisition and processing can be found in Chapter 4.

Experiment 2. Because the human urine dataset does not contain a good example of a variable multiplet, I used another urine dataset to evaluate the algorithm performance on multiplet peaks. This dataset contained 134 ^{1}H NMR dog urine spectra collected at 600 MHz magnetic field strength. The experiment was designed to study the impact of food type on the canine metabolome (unpublished). Four groups of dogs were fed with 4 types of diets. They changed diets every 4 weeks following a Latin square design. Urine samples were centrifuged and 540 μl supernatants from each sample were added with 60 μl NMR buffer (1.5 M KH_2PO_4 buffer with 1/9 mM DSS when mixed with sample) for NMR spectra acquisition. The spectra were one-dimensional nuclear Overhauser enhancement spectroscopy (1D-NOESY PR) with water suppression. The spectra were phased, baseline-corrected, referenced to the DSS resonance at 0 ppm, and removed of water regions and ends using NMRPipe [13] software or Bruker Topspin 4.0.7 software, and in-house MATLAB scripts (`https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA`).

Experiment 3. A publicly available dataset that other alignment or binning methods [14], [15], [16] had used for evaluation was used here for algorithm validation. This dataset contained 40 ^{1}H NMR spectra measured at 400 MHz magnetic field strength from 40 different table wines[12]. D_2O with TSP-d4 (3-trimethyl-silyl-[2,2,3,3-$^{2}H_4$]propionic acid, 5.8 mM) were added to wine samples in 1:9 ratio (D_2O:wine) for NMR analysis and spectra were referenced to TSP-d4 resonance at 0.0 ppm before alignment. Spectra were processed in Xwin-NMR and Matlab in-house scripts [12].

3.2.2 Algorithm

Input spectra are referenced but not aligned and are in their experimental run order. Metabolomics run order is usually randomized before sample preparation to managing uncontrolled variables.

(1) **Choose a guiding signal.** The pHIT algorithm plots a figure showing stacked spectra in experimental order to facilitate users to identify an appropriate guiding signal. The user will choose a signal that shows obvious variations and is isolated from its neighboring signals from the spectra as a guiding signal.

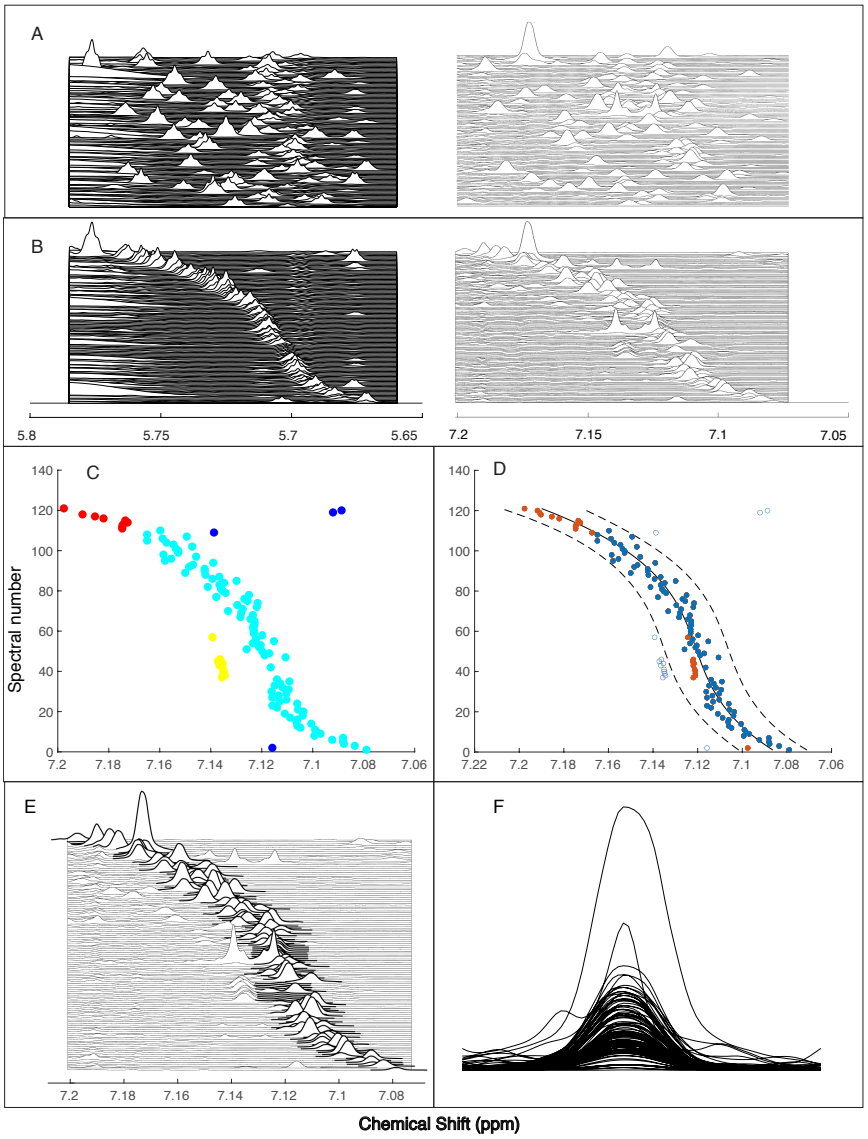


Figure 3.1: Performance of each step of the pHIT algorithm on an example human urine dataset. (A) Stacked spectral of a guiding signal (left, 5.56-5.8 ppm) and a responding signal (right, 7.07-7.20 ppm) in experimental order. (B) Stacked spectra of the guiding and responding signal after reordering spectra by the location of the guiding signal. (C) Off-curve peak detection result on the responding peak in panel B. Each dot represents the highest local maximum on each spectrum inside this region. Colors indicate clusters, and dark blue dots are outliers that do not belong to any clusters. (D) Peak detection result after removing off-curve outliers on the responding peak in panel B. Solid blue circles represent peaks from the biggest cluster in panel C. Open blue circles represent off-curve peaks. Solid red circles represent peaks detected in confidence intervals when the spectra have off-curve outliers. The solid line represents the polynomial curve fitted from the biggest cluster in panel C. Dashed lines indicate the 95% confidence interval of the fitted line. (E) Stacked reordered spectra showing the responding signal from panel B with regions that will be aligned bolded. (F) Aligned responding signal from panel B.

Figure 3.1A shows an example of guiding signal. The algorithm uses the position of peak maximum as the position of the guiding signal. If the guiding signal is a multiplet, the algorithm uses its center as its position. This process can reduce random errors, so a clear-patterned multiplet usually performs better than a singlet as the guiding signal.

(2) **Reorder spectra and choose regions to align.** The spectra are reordered according to the position of the guiding signal, such that spectrum with the smallest chemical shift of the guiding signal is placed as the first spectrum and spectrum with the largest chemical shift of the guiding signal is placed as the last spectrum. A plot of stacked, reordered spectra is displayed to facilitate users to find responding signals (i.e., the signals showing clear curves. Figure 3.1B shows example of a responding signal on reordered spectra. The user will record the left and right boundaries of each responding curve and the number of peaks of the responding signal, then input a list of these boundaries and peak numbers to the algorithm.

(3) **Align the guiding signal.** The algorithm aligns the peaks by putting their peak crests at the same chemical shift. Because the region may contain other peaks, if I put the aligned peaks in their original locations (for example, use the median value of the varied peaks as the new chemical shift of the aligned peaks) the aligned peaks may overlap with neighboring peaks. Therefore, the algorithm places aligned peaks outside of the spectrum and removes them from their original positions. The boundaries of peaks to be moved are determined by the position of nearest local minima to the peak crest. I use the median distances between peak crests and valleys across spectra when moving peaks to avoid cutting off parts of peaks in the event some local minimums are on the peaks. For users to use the aligned spectra, the algorithm saves information including the original and new positions of the peaks in a new variable.

(4) **Align responding signals.**

a. Detect peaks: For each responding signal, the pHIT algorithm detects the highest local maximum within the user-inputted left and right boundaries.

b. Remove neighboring peaks from peak list: As the highest local maxima may include some neighboring signals, like those shown in Figure 3.1C, the algorithm uses a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [17] to detect and exclude the off-curve neighboring peaks. The peak matrix is scaled on both chemical shift and spectral number dimensions to the same scale before clustering. As shown in Figure 3.1C, peaks on the responding curve are clustered into the biggest cluster. Other surrounding peaks are grouped into other clusters or as outliers. These peaks, collectively are defined as off-curve outliers, are then removed from the peak list for alignment.

c. Find peaks on outlier-removed spectra: For spectra whose highest local maxima are off-curve outliers, the algorithm looks for the signal of interest around the responding curve. This is done by first fitting the peaks from the biggest cluster into a polynomial curve then calculating the 95% confidence interval of the curve. The algorithm picks the highest local maxima inside the confidence intervals as the signal of interest (Figure 3.1D).

d. Align detected peaks: The algorithm aligns responding peaks in the same way as aligning guiding peaks in step 3. Because the highest signal in the region is removed, smaller peaks in the same region are then detectable and tracible (Figure 3.2).

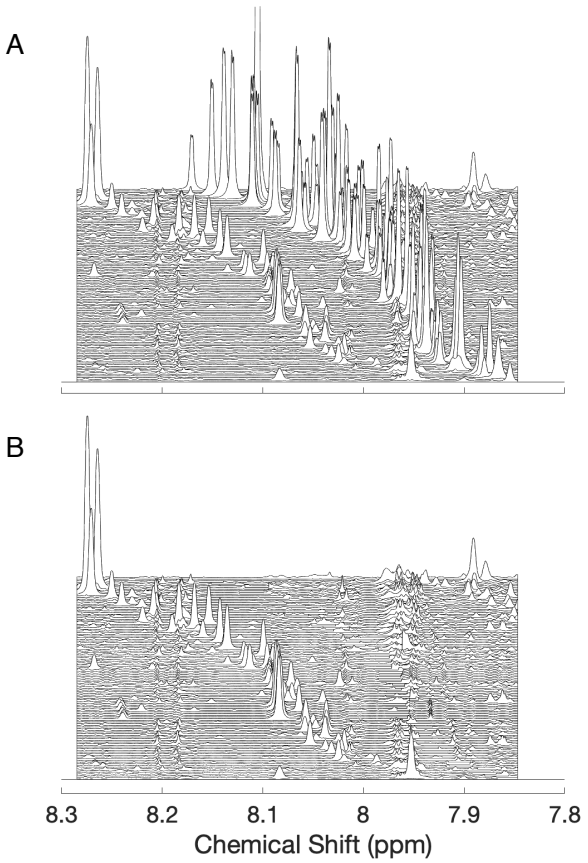


Figure 3.2: Stacked spectra showing 2 responding signals in the human urine dataset, spectra are reordered by the guiding peak at 5.56-5.8 ppm shown in Figure 3.1. (A) Before removing the higher signal at 7.8-8.2 ppm. (B) After removing the higher signal at 7.8-8.2 ppm.

e. Multiplet cases: Sometimes there are multiple parallel curves and the algorithm can align them together no matter if they are peaks from the same metabolite or not. This process can better preserve the pattern of the multiplet and improve alignment performance.

In this case, in addition to detecting outliers, the DBSCAN algorithm can also separate peaks from different curves. If the curves are sparse, they will form different clusters (Figure 3.3). Because some peaks may fall into clusters from other curves, the algorithm calculates the confidence interval of each cluster of peaks, then uses the cluster showing the best fitting performance to align the whole set of peaks. Median distances between this signal and the furthest two detected signals on each spectrum are calculated. This value plus the median distances between the outside peaks and their corresponding outer valleys is the length of peaks to be cut and moved to the end of the spectra.

If the curves are close (e.g., the higher signal on Figure 3.2A), they will fall into the same cluster. In such cases, the algorithm tests if the majority of left-most peaks are inside the confidence interval of

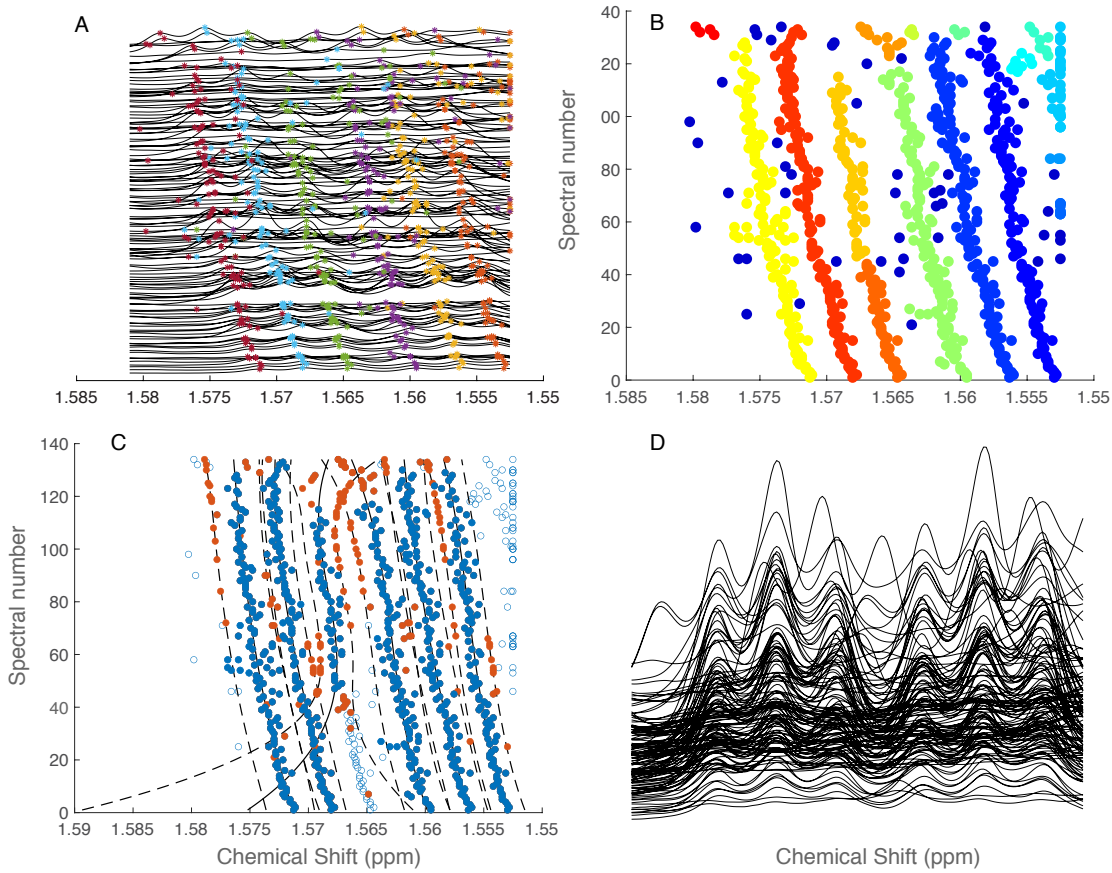


Figure 3.3: Example of pHIT performance on a multiplet from the dog urine dataset. Spectra are reordered according to a guiding peak at 1.45-1.5 ppm (not shown). (A) Stacked spectra of a responding multiplet. Colors indicate 6 different peaks. (B) Off-curve peak detection result. Each dot represents the highest local maximum on each spectrum inside this region. Colors indicate clusters, and dark blue dots are outliers that do not belong to any clusters. (C) Peak detection result after removing off-curve outliers. Solid blue circles represent peaks from the 6 biggest clusters in panel B. Open blue circles represent off-curve peaks. Solid red circles represent peaks detected in confidence intervals when the spectra have off-curve outliers. The solid line represents polynomial curves fitted from each of the 6 biggest clusters in panel B. Dashed lines indicate the 95% confidence intervals of the fitted lines. (D) Aligned responding multiplet.

the right-most peaks. If not, the algorithm processes the peaks like those for multiple peaks. Otherwise, because the same peak (the highest peak) will be detected in all confidence intervals in most cases, the remaining peaks will need to be found in another way. If the peaks are very close, they may not exhibit all local maxima when the signal is small. Therefore, I use the median distances between peaks to locate the remaining peaks. Note that since the distances are calculated based on the location of peaks detected before clustering and re-detecting, they are likely to be affected by off-curve outliers. However, since most of the peaks are close, taking the median value should eliminate this impact.

(5) **Choose another guiding signal and repeat steps 2-4.** Because more than one factor can impact chemical shift values, it is possible that some signals may not exhibit clear curves after reordering using the first guiding signal. In such cases, the user may select another guiding signal from these signals. Repeat steps 2-4 for each guiding signal until all signals are processed. Alm et al. performed a principal component analysis on 10 shifting peaks and found that 3 to 4 guiding peaks should cover most of the shifting sources[11].

(6) **Order back.** Place the spectra back in its original run order so group information matches the spectra.

(7) **Expected output.** The expected output of the workflow includes aligned spectra together with a corresponding elongated chemical shift vector. In the aligned spectra all processed peaks are moved to the end and their original positions are replaced by zeros. A structure variable called "AlignedPeaks" is also generated. It contains information for each region that has been processed. The information includes both the original and new peak positions, the median of the original positions, the number of signals, and the intensity matrix.

3.3 Results

I aligned 16 peaks in the human urine dataset and 29 peaks in the dog urine dataset by pHIT. Figure 3.4 shows example of alignment performance of pHIT and a widely used algorithm CCOW on a peak at 7.07-7.20 ppm in the urine dataset. The pHIT algorithm appropriately aligned it (Figure 3.4A) whereas the CCOW algorithm aligned this peak into several peaks (Figure 3.4B).

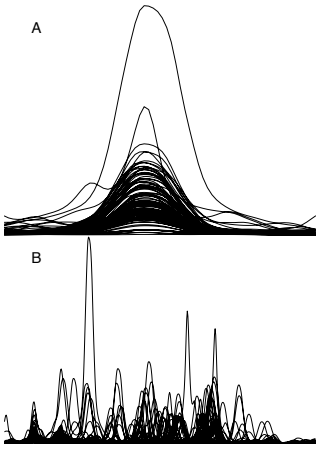


Figure 3.4: Alignment results on an example responding signal at 7.07-7.20 ppm. (A) Aligned by pHIT. (B) Aligned by CCOW.

When using the human urine dataset for statistical modeling, we found that 3 out of 16 reorder-aligned peaks exhibited significant differences between groups, and 2 of the 3 peaks were identifiable, thus showing that the reorder-align algorithm is beneficial for quantitative analysis.

Aligning these widely shifting peaks with pHIT also helped with peak annotation The aligned peak at 7.07-7.20 ppm has a high correlation with another shifting peak at around 7.85-8.28 ppm, and a moderate correlation with another slightly varied peak at around 4.0 ppm (Figure 3.5). This 1D pattern helped us identify this peak as histidine. When I checked the 2D HSQC spectrum, I found the peak at around 4.0 ppm matched the database, but the peaks at 7.1 and 7.9 ppm shifted away from the database peak position. This led to a low matching ratio for histidine, so histidine was not identified when analyzing the 2D data.

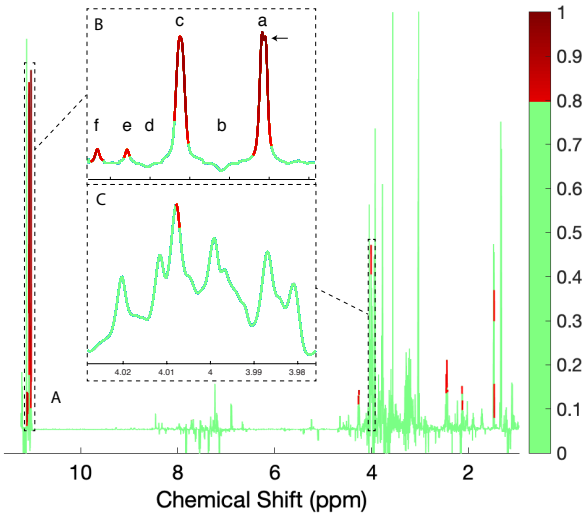


Figure 3.5: Statistical total correlation spectroscopy (STOCSY) plot for a pHIT-aligned signal pointed by an arrow. The color bar shows correlation coefficients. Regions with correlation coefficients lower than 0.8 are colored in green. (A) Full spectrum. (B) Aligned signals, labeled by letters (C) Regions between 3.98 – 4.02 ppm, showing one peak with high correlation coefficients.

In addition, the alignment process were also beneficial for annotating and analyzing neighboring peaks. For example, trigonelline has a peak at 8.08 ppm. When I aligned the spectra by CCOW, the peak at 8.08 ppm did not show a high correlation with other trigonelline peaks (Figure 3.6A) because the histidine peak is nearby. However, when the histidine peak is aligned and cleaned from the region, other trigonelline peaks showed high correlations with the peak at 8.08 ppm (Figure 3.6B).

I further evaluated this algorithm with a publicly available wine dataset[12], which was broadly used as the example for showing binning and alignment performance, such as optimized bucketing[14], icoshift[15], and CluPA alignment performance[16]. The pHIT algorithm also aligned shifting peaks well on this dataset (Figure 3.7).

3.4 Discussion

Our algorithm is designed for significantly shifting peaks. The performance of the algorithm on peaks with small variation may not be ideal because these peaks are not sensitive to the reordering of spectra. When random variation is the main contributor to the total variation, the curves will always be fuzzy.

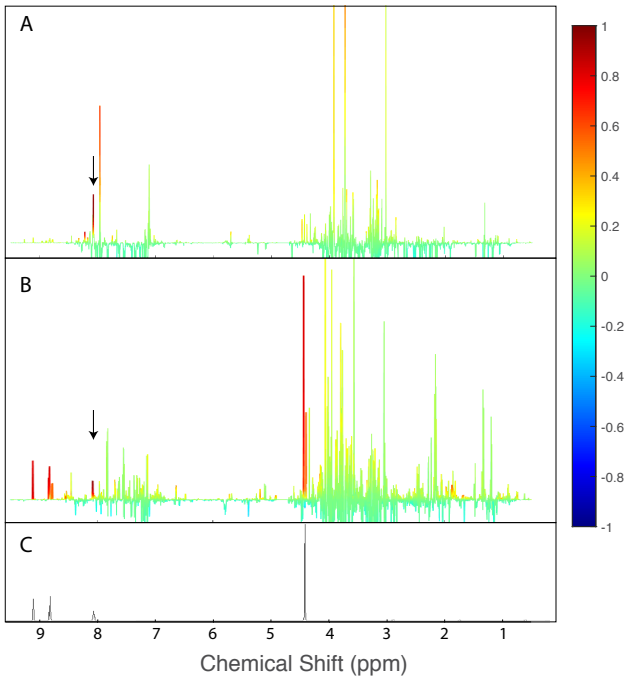


Figure 3.6: STOCSY plot for the signal at 8.08 ppm pointed by arrows on spectra aligned by CCOW (A) and pHIT (B). The color bar shows correlation coefficients. (C) Experimental ^{1}H NMR spectrum of trigonelline under 600 MHz from Human Metabolome Database[18].

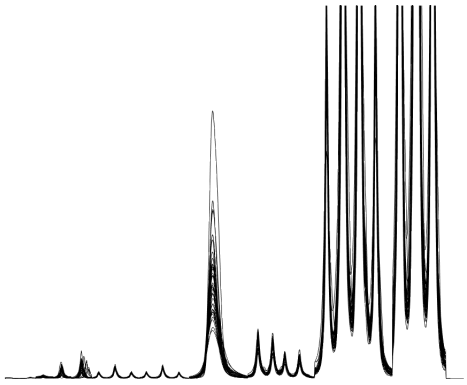


Figure 3.7: Performance of pHIT alignment on the wine dataset.

Meanwhile, the confidence intervals of peaks with small random variation are narrow so they may be inadvertently clustered as outliers and excluded from the alignment. In addition, when a signal only

shows up in part of the spectra, it will be difficult to use this algorithm because its confidence interval can be very broad.

Although cleaning big peaks usually helps to align neighboring small peaks, sometimes moving the peaks may remove part of the small peaks when they are very close, thus impacting the alignment of the smaller peaks. Users should be aware that it is possible the cut edge on the remaining spectra may also impact analysis when the peak is not moved completely.

Therefore, the pHIT algorithm can be used jointly with other alignment methods. For example, highly varied peaks can be aligned and moved by this algorithm, and then the rest of the spectra can be aligned by another algorithm and concatenated with the previous aligned peaks that have been put at outside of the spectra. The algorithm can also be used with or followed by binning.

3.5 Conclusions

The pHIT algorithm presented here utilized the source of peak position variation to match peaks across spectra. Without measuring the pH or salt concentrations of the sample, ordering spectra by the position of one shifting internal signal can make other shifting signals traceable. The algorithm automatically traces the shifting signal within confidence intervals of off-curve-outlier-removed peaks. This strategy does not rely on the similarity of the shapes between the responding and guiding curves and can be suitable for generalized cases.

Evaluating the pHIT algorithm on experimental human and dog urine datasets shows that it can sufficiently distinguish neighboring peaks and align single and multiple peaks. This alignment then improved both quantitative analysis and spectra annotation. I expect the algorithm will be used with other alignment or binning methods to facilitate extracting more information from NMR spectra in metabolomics research.

Bibliography

[1] N.-P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping," *Journal of Chromatography A*, vol. 805, no. 1, pp. 17–35, 1998, ISSN: 0021-9673. DOI: https://doi.org/10.1016/S0021-9673(98)00021-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021967398000211.

[2] J. W. H. Wong, C. Durante, and H. M. Cartwright, "Application of fast fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets," *Analytical Chemistry*, vol. 77, no. 17, pp. 5655–5661, 2005, ISSN: 00032700. DOI: 10.1021/ac050619p.

[3] T. N. Vu and K. Laukens, "Getting your peaks in line: A review of alignment methods for NMR spectral data," *Metabolites*, vol. 3, no. 2, pp. 259–276, 2013, ISSN: 22181989. DOI: 10.3390/metabo3020259.

[4] G. D. Tredwell, J. G. Bundy, M. De Iorio, and T. M. Ebbels, "Modelling the acid/base1H NMR chemical shift limits of metabolites in human urine," *Metabolomics*, vol. 12, no. 10, pp. 1–10, 2016, ISSN: 15733890. DOI: 10.1007/s11306-016-1101-y.

[5] P. G. Takis, H. Schäfer, M. Spraul, and C. Luchinat, "Deconvoluting interrelationships between concentrations and chemical shifts in urine provides a powerful analysis tool," *Nature Communications*, vol. 8, no. 1, p. 1662, 2017, ISSN: 2041-1723. DOI: 10.1038/s41467-017-01587-0. [Online]. Available: http://dx.doi.org/10.1038/s41467-017-01587-0%20https://doi.org/10.1038/s41467-017-01587-0.

[6] L. Jiang, J. Huang, Y. Wang, and H. Tang, "Eliminating the dication-induced intersample chemical-shift variations for NMR-based biofluid metabonomic analysis," *Analyst*, vol. 137, no. 18, pp. 4209–4219, Aug. 2012, ISSN: 1364-5528. DOI: 10.1039/C2AN35392J. [Online]. Available: https://pubs.rsc.org/en/content/articlehtml/2012/an/c2an35392j%20https://pubs.rsc.org/en/content/articlelanding/2012/an/c2an35392j.

[7] M. Liebeke, J. Hao, T. M. D. Ebbels, and J. G. Bundy, "Combining spectral ordering with peak fitting for one-dimensional NMR quantitative metabolomics," *Analytical Chemistry*, vol. 85, no. 9, pp. 4605–4612, 2013, ISSN: 00032700. DOI: 10.1021/ac400237w.

[8] A. Onufriev, D. A. Case, and G. M. Ullmann, "A Novel View of pH Titration in Biomolecules," *Biochemistry*, vol. 40, no. 12, pp. 3413–3419, Mar. 2001, ISSN: 0006-2960. DOI: 10.1021/bi002740q. [Online]. Available: https://doi.org/10.1021/bi002740q.

[9] C. Xiao, F. Hao, X. Qin, Y. Wang, and H. Tang, "An optimized buffer system for NMR-based urinary metabonomics with effective pH control, chemical shift consistency and dilution minimization †," 2009. DOI: 10.1039/b818802e. [Online]. Available: www.rsc.org/analyst.

[10] L. Csenki, E. Alm, R. J. Torgrip, *et al.*, "Proof of principle of a generalized fuzzy Hough transform approach to peak alignment of one-dimensional 1H NMR data," *Analytical and Bioanalytical Chemistry*, vol. 389, no. 3, pp. 875–885, 2007, ISSN: 16182650. DOI: 10.1007/s00216-007-1475-9.

[11] E. Alm, R. J. Torgrip, K. M. Åberg, I. Schuppe-Koistinen, and J. Lindberg, "A solution to the 1D NMR alignment problem using an extended generalized fuzzy Hough transform and mode support," *Analytical and Bioanalytical Chemistry*, vol. 395, no. 1, pp. 213–223, 2009, ISSN: 16182650. DOI: 10.1007/s00216-009-2940-4.

[12] F. H. Larsen, F. Van Den Berg, and S. B. Engelsen, "An exploratory chemometric study of 1H NMR spectra of table wines," *Journal of Chemometrics*, vol. 20, no. 5, pp. 198–208, May 2006. DOI: 10.1002/CEM.991.

[13] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax, "NMRPipe: a multidimensional spectral processing system based on UNIX pipes.," eng, *Journal of biomolecular NMR*, vol. 6, no. 3, pp. 277–293, Nov. 1995, ISSN: 0925-2738 (Print). DOI: 10.1007/BF00197809.

[14] S. A. Sousa, A. Magalhães, and M. M. C. Ferreira, "Optimized bucketing for NMR spectra: Three case studies," *Chemometrics and Intelligent Laboratory Systems*, vol. 122, pp. 93–102, Mar. 2013, ISSN: 01697439. DOI: 10.1016/j.chemolab.2013.01.006.

[15] F. Savorani, G. Tomasi, and S. Engelsen, "icoshift: A versatile tool for the rapid alignment of 1D NMR spectra," *Journal of Magnetic Resonance*, vol. 202, no. 2, pp. 190–202, Feb. 2010, ISSN: 1090-7807. DOI: 10.1016/J.JMR.2009.11.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1090780709003334?via%5C%3Dihub.

[16] T. N. Vu, D. Valkenborg, K. Smets, *et al.*, "An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data," *BMC Bioinformatics*, vol. 12, no. 1, p. 405, 2011, ISSN: 14712105. DOI: 10.1186/1471-2105-12-405. arXiv: 1609.09627. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/405.

[17] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Tech. Rep., 1996. [Online]. Available: www.aaai.org.

[18] D. S. Wishart, C. Knox, A. C. Guo, *et al.*, "HMDB: A knowledgebase for the human metabolome," *Nucleic Acids Research*, vol. 37, no. SUPPL. 1, 2009, ISSN: 03051048. DOI: 10.1093/NAR/GKN810.

Chapter 4

EFFECTS OF ZIKA VIRUS INFECTION ON THE METABOLOME OF PREGNANT WOMEN: A LONGITUDINAL STUDY

[1]

[1] S. Zhang.*, D.D. Ellis*, J.M. Walejko, M.T. Arévalo, M. Welton, J.F. Cordero, T.M. Ross, S. Datta, and A.S. Edison. To be submitted to *Scientific Reports*. *: co-first authors.

Author contributions

This paper will be submitted to the Scientific Reports journal soon. I am the co-first author with Dorothy D. Ellis. I conducted NMR experiments with Jacquelyn M. Walejko. I processed and annotated NMR data. I advised on data analysis and interpretation with Arthur S. Edison, Susmita Datta, Dorothy D. Ellis. I wrote the main manuscript text with Dorothy D. Ellis and Arthur S. Edison. Ted M. Ross, José F. Cordero, Michael Welton, and Maria T. Arévalo designed the sample collection protocol. Maria T. Arévalo, Michael Welton, José F. Cordero, and Jacquelyn M. Walejko collected the samples and tested samples for infection. Dorothy D. Ellis conducted statistical analysis.

Abstract

Zika virus (ZIKV) is a mosquito-borne +ssRNA virus that can cause abnormal development in the human fetal central nervous system and even lead to stillbirth. Despite numerous studies in this area, there is currently still no sufficient treatment for it. Knowledge on how ZIKV infection impact human metabolism is still lacking. Untargeted metabolomics can profile the overall change in metabolites after infection, thus provide hypotheses for specific investigations. We performed a Nuclear Magnetic Resonance spectroscopy (NMR)-based untargeted case-control metabolomics study on urine of ZIKV-infected pregnant women. We collected samples monthly from the first trimester for 6 months of ZIKV-infected and non-infected individuals and modelled the longitudinal data. We identified 3-aminoisobutyrate and trigonelline with significantly higher levels in the ZIKV-infected patients, while 11 metabolites (fucose, 2-hydroxyglutarate, N-acetyl-glutamine, dimethylamine, 4-hydroxyphenethyl alcohol, creatinine, lactate, threonine, histidine, pseudouridine, and 1-methylnicotinamide) had significantly lower levels. We also identified 2 metabolites, including glucose and 1-methylnicotinamide, where the trends over time of the intensity levels between the two groups were significantly different. These metabolites suggested further study on tryptophan, NAD+, pyrimidine, and glucose metabolic pathways. These metabolomic changes may lead us to a better understanding of mechanisms that cause poor fetal outcomes as well as effects of virus infection on human pregnancy.

4.1 Introduction

Zika virus (ZIKV) is a positive, single-stranded RNA virus that belongs to the *Flaviviridae* virus family. This family consists of four genera: *Flavivirus*, which includes ZIKV, yellow fever virus, West Nile virus, and dengue virus, *Hepacivirus*, which includes hepatitis C virus, *Pegivirus* and *Pestivirus* [1], [2]. ZIKV can be transmitted by at least 31 different species of mosquito, with the *Aedes* genus accounting for the largest portion of ZIKV-infected mosquitoes in both urban and non-urban environments [3]. ZIKV can also be transmitted through sex[4] and from pregnant mothers to fetuses[5]. From its discovery in 1947 until the 2015 Brazilian outbreak[6], ZIKV did not receive much public attention. During the 2015 Brazilian outbreak, reports of severe complications, including Guillain-Barré syndrome in adults[7] and microcephaly in neonates[8], were highly associated with ZIKV infection. So far, ZIKV has spread to 5 continents. Though there is not a current outbreak, it still threatens 2 billion people in tropical and

subtropical regions[6]. In the United States, 5-10% of newborns born to infected mothers have ZIKV-associated birth defects[9], [10]. Due to the severe impact and rapid spread of ZIKV, the World Health Organization (WHO) added ZIKV to its list of Public Health Emergency of International Concern in April 2016[11] where it remained until November 2016[12]. There is still no effective, approved therapy for ZIKV infection; the most advanced vaccine is still in Phase I clinical trials[13]. Several FDA-approved small molecule drugs have exhibited an effect *in vitro*, but none of these have proved to be efficacious *in vivo*[14], [15], [16], [17]. Furthermore, the pathogenesis of this disease is far from clear. Current research mostly focuses on the immunological and virological aspects[18], [19], [20]; the metabolic pathways involved in ZIKV infection have rarely been investigated.

To provide future directions for targeted research and biomarker discovery, an untargeted metabolomics study is suitable for an initial exploration of the metabolic pathways involved in ZIKV infection. There have been several metabolomics studies on human samples, but most of them focus on lipid-resolved metabolites[21], [22], [23], [24], [25], [26]; analysis on hydrophilic metabolomics is lacking. In addition, although some of these studies used samples from newborn/infants[22], [23], [26] or placenta[24], a direct assessment on pregnant women is also missing. For the measurement of hydrophilic metabolite levels, urine samples are ideal since the kidneys have already filtered proteins and large lipids. Furthermore, urine-metabolomics are good ways to monitor pregnancy status[27], [28]. Therefore, we used urine samples in pregnant human to fill this gap.

For our analysis of urine-metabolomics, we used nuclear magnetic resonance (NMR) spectroscopy, which is a powerful and popular analytical platform for metabolomics research[29]. Compared to the other popular platform mass spectrometry, NMR is non-destructive, more quantitatively reproducible, and is capable of distinguishing between isomers, which enables it to provide structural information[30]. Although the detection sensitivity of NMR is lower than mass spectrometry, NMR can still detect low micromolar concentrations of metabolites, which cover most of the major chemical superclasses[31], [32].

Here, we performed an NMR-based longitudinal case-control study on the urine metabolome of ZIKV-infected pregnant women by comparing the urine metabolome of these women to a control group of non-infected women. Because the urine metabolome changes over the course of pregnancy[27], [33], we collected multiple samples from each individual in the case and control groups over time. We collected serum and urine samples from 275 pregnant women monthly from their first trimester in Puerto Rico, used serum samples to indicate recent infection of ZIKV, and analyzed the metabolome of the corresponding urine samples via NMR. Due to limited detection sensitivity[6], we define any participants with at least one ZIKV-non-negative serum sample as ZIKV+. We determined the control group by matching the same number of ZIKV-negative participants from the cohort by age and gestational age of the fetus and denote this group as ZIKV-. Each group was comprised of 10 individuals for a total of 20 individuals. By using a nonparametric longitudinal model on the reduced-dimension NMR spectra, we identified 13 metabolites for which the concentration was significantly different between the two groups and 2 metabolites where the trends over time of the intensity levels between the ZIKV+ and ZIKV- individuals were significantly different.

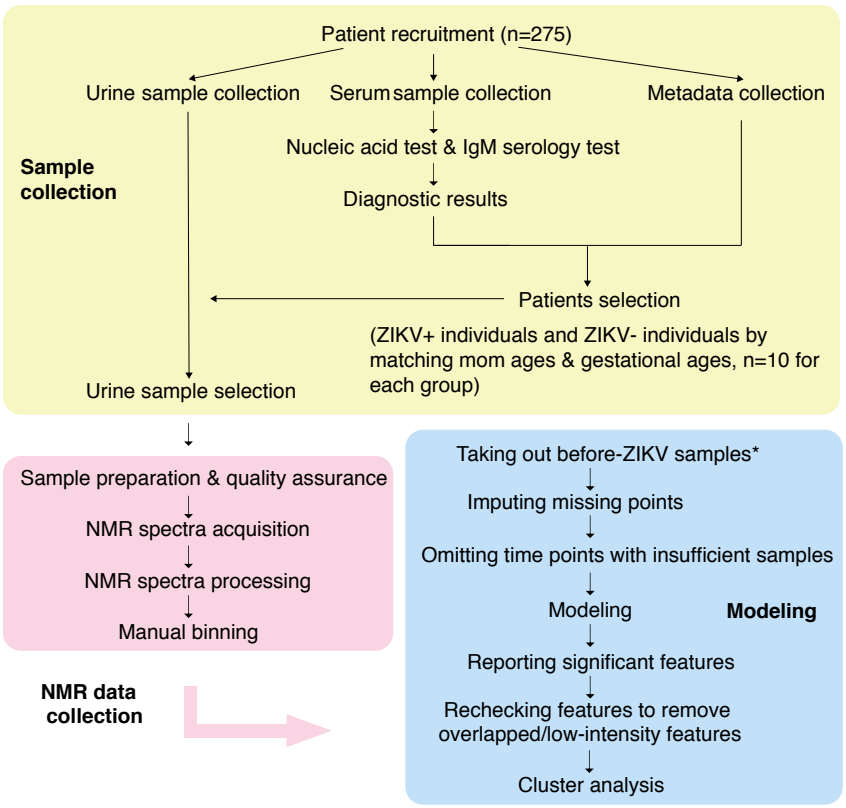


Figure 4.1: Workflow of data collection and analysis. *We remove all ZIKV-negative samples from ZIKV+ individuals and treat these observations as missing.

4.2 Results

4.2.1 Description of Dataset

Between 2016 and 2017, through an IRB approved study on human pregnancy, we recruited 275 pregnant women during their first trimester in Puerto Rico, a ZIKV affected region. At approximately 4-week intervals, serum and urine samples were simultaneously collected from the participants. We tested for the presence of ZIKV RNA or IgM antibodies in serum samples to indicate recent infection of ZIKV. Of these 275 women, ten individuals, each of whom had at least one non-negative serum test result for ZIKV over the course of the study, comprised the case (ZIKV+) group. We then selected 10 individuals with no positive serum test results from the total cohort to serve as the control group by selecting individuals who had similar maternal and gestational ages to the individuals in the ZIKV+ group (Supplementary Table A.1). We denote this group as control (ZIKV-) group. Figure 4.1 shows the workflow of data collection and processes. Within these 20 participants, most individuals joined during their first trimester; however,

one individual joined during the second trimester at around the thirteenth week of gestation. Among the ZIKV+ group, 6 individuals already had ZIKV antibodies detected in their sera at their first prenatal visit. The remaining 4 people each had one ZIKV-negative sample before their first ZIKV-non-negative sample. Of these four individuals, one provided a sample that indicated acute infection (ZIKV RNA detected). Two subjects became infected with ZIKV before the second time point, one became infected between the third and fourth time point, and one became infected before the fifth time point visit. At the last time point, ZIKV IgM was not detected in the sera of nine of the individuals; only one individual still had detectable levels of IgM. Table 4.1 summarizes the sample conditions at each time point. All ZIKV- patients, except for one individual who is missing a urine sample for the fourth time point, provided at least one urine sample for each time point. In total, we analyzed 104 urine samples from 20 people, including 37 samples from ZIKV+ patients. Samples were collected from the first, second, and early third trimesters of pregnancy. Of the ten infected women, two gave birth preterm and one had a stillbirth. All newborns from the control group were full-term live births (one individual withdrew from the study).

Table 4.1: Number of samples collected at each time point

Time point (Estimated trimester)	Samples from ZIKV+ participants (n=10)				Samples from control participants (n=10)
	Before infection	ZIKV-non-negative	Post infection	Total	
1 (1st)	1	6	0	7	10
2 (1st & 2nd)	1	5	4	10***	12**
3 (2nd)	1	4	4	9**	11*
4 (2nd)	1	3	0	4	10*
5 (2nd)	0	2	3	5	12**
6 (3rd)	0	0	2	2	12**
Total	4	20	13	37	67

*One participant has an additional sample at this time point.
**Two participants have an additional sample at this time point.
***Three participants have an additional sample at this time point.

We collected one-dimensional proton NMR spectra on each urine sample. We identified 69 metabolites from these NMR spectra (Supplementary Table A.2). However, because not all peaks from the spectra were annotated and those unannotated peaks may include important information, we performed statistical analysis in an untargeted manner. To reduce the dimension of the data matrix for statistical analysis, we manually binned the normalized full resolution spectra into buckets independently from annotation. Because this was done independently of peak annotation, we did not assume peaks from a multiplet were from one metabolite, but placed them in separate buckets. We also gave a quantification confidence score for each feature to indicate our confidence of treating the binned peak as a real signal for

metabolites (details in Methods). In the end, we binned the spectra into 649 bins for statistical analysis. For each bin, the chemical shift value identified represents the mode of the binned peak and the intensity represents the area under the curve. We will refer to these bins as peaks and the area under these curves as intensities throughout the rest of the paper.

4.2.2 Modeling Longitudinal Data to Discover Discriminative Metabolites Between ZIKV+ and ZIKV- Women

To control for the effects of pregnancy progression and individual variance on a set of data that do not follow a known linear or nonlinear pattern, we used a non-parametric mixed effects model to examine the metabolomic differences between controls and patients who had had an active or recent ZIKV infection. Since some individuals who entered in the study tested negative for ZIKV for their initial visit, we removed the ZIKV-negative samples from individuals in the ZIKV+ group from the data set, grouped samples into 6 time points, and imputed the intensities of the peaks at the missing time points using classification and regression tree (CART)[34]. Since we needed to impute more than 50% of the intensities of the peaks for the samples in the ZIKV+ group for time points 4 and 6, we excluded these time points from the final model. We treated each peak as independent, fit 649 non-parametric mixed effects models (one model for each peak), and performed Benjamini-Hochberg false discovery rate (FDR)[35] correction. Features with FDR-adjusted p-values lower than 0.15 were treated as significant. The adjusted p-value of sodium trimethylsilylpropanesulfonate (DSS), our reference substrate, was 0.16.

We found 58 significant peaks, 39 of which had quantification confidence scores at the highest level. Of these 39 peaks, we identified 13 metabolites that corresponded to 20 of the peaks (Table 4.2). We provide information about unidentified features in Supplementary Table A.3. Figure 4.2 shows the trajectories (fitted Loess curves[36]) of annotated significant metabolites during pregnancy. Supplementary Fig. A.6 shows the trajectories of all significant features. Since we observed similar patterns of the trajectories on some metabolites, we conducted Hierarchical Clustering Analysis[37], [38] (HCA) on these metabolites and unknown features. We performed clustering on the median intensity of the imputed data. This analysis resulted in the identification of 7 major clusters of metabolites (Figure 4.3). Not all metabolites are included in these 7 major clusters since their nodes are relatively far away.

When discussing the average intensity scores in the interpretation of the graphs for this section, we refer to a robust average as plotted by the Loess curves rather than to the mean intensity score. Of the 7 clusters, the metabolites belonging to the first 6 clusters had, on average, lower intensity levels in the ZIKV+ group than in the ZIKV- group. For metabolites belonging to the final cluster, the intensity levels of metabolites were on average higher in the ZIKV+ group. Threonine levels, on average, decreased from the first time point over the course of the pregnancy in the ZIKV+ group. In ZIKV- group, it increased from the 1st time point to the 2nd time point and then decreased slightly; however, overall, later time points had higher levels of threonine than the 1st time point in the ZIKV- group. 1-Methylnicotinamide (1-MNA) generally increased over the course of pregnancy, but at time point 3, the average intensity levels of both groups were similar. Cluster 2, which includes fucose, 2-hydroxyglutarate, and histidine,

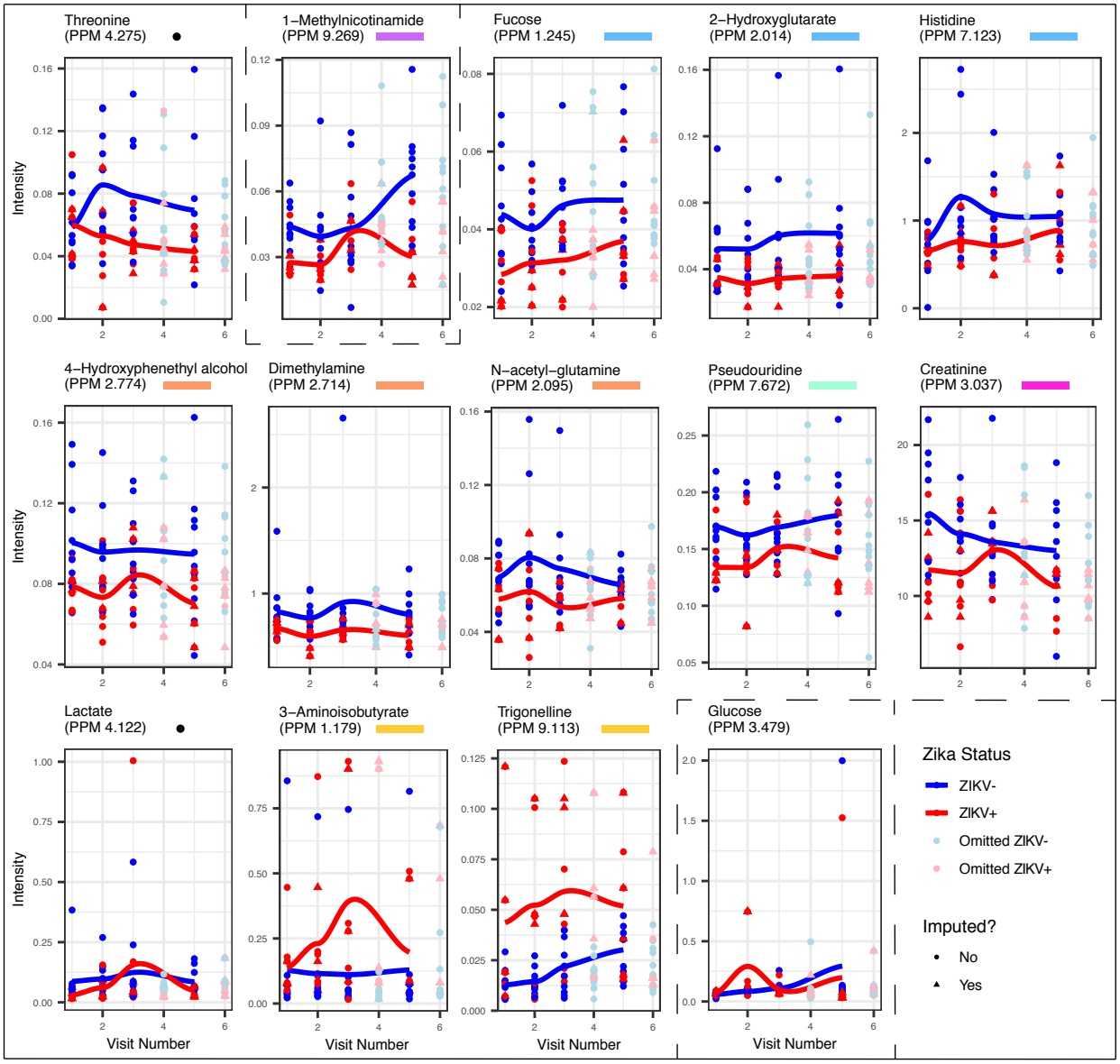


Figure 4.2: Graphical representation of intensity scores for annotated features. If multiple features correspond to the same metabolite, we select a representative feature. The fitted curves are not based on the nonparametric model; instead, they are fitted Loess curves to allow easier interpretation of the differences in intensity scores between the ZIKV+ and ZIKV- populations at each feature. The nonparametric model identifies a significant difference in intensity scores for all time points for all listed features except for glucose, which is listed last, between the ZIKV+ and ZIKV- populations. For glucose and 1-methylnicotinamade, outlined by a dashed line, the nonparametric model identifies the interaction between ZIKV-infection status and the time point as significant, which means the pattern of the intensity score is different over time for these two groups. Colors after feature annotations correspond to clusters in Figure 4.3. Black dots indicate features are not included in any clusters.

Table 4.2: Significantly different metabolites between ZIKV+ and control individuals. Each metabolite is represented by one of its peaks.

No.	Annotation	Chemical Shift (ppm)	FDR adjusted p-value	Confidence Score*
1	3-Aminoisobutyrate	1.179	1.06E-01	4
2	Fucose	1.245	9.34E-02	3
3	2-Hydroxyglutarate	2.014	1.21E-01	3
4	N-acetyl-glutamine	2.095	1.06E-01	4
5	Dimethylamine	2.714	1.06E-01	3
6	4-Hydroxyphenethyl alcohol	2.774	9.15E-02	3
7	Creatinine	3.037	1.21E-01	3
8	Lactate	4.122	1.21E-01	4
9	Threonine	4.275	1.21E-01	4
10	Histidine	7.123	1.25E-01	2
11	Pseudouridine	7.672	9.84E-02	3
12	Trigonelline	9.113	9.68E-02	3
13	1-Methylnicotinamide	9.269	9.15E-02	3

*Confidence score is defined as follows: 1) putatively characterized compounds or compound classes, 2) 1D NMR matches to literature and/or database (BMRB and/or HMDB), 3) HSQC matches on COLMARm or AssureNMR, 4) HSQC and HSQC-TOCSY match on COLMARm, 5) verified by spiking

increased slightly during pregnancy and had similar patterns in the two groups. Cluster 3, including 4-hydroxyphenethyl alcohol, dimethylamine, and N-acetyl-glutamine had similar patterns between groups as well, but their levels were neither increasing nor decreasing overall. For pseudouridine and creatinine, the average intensity levels spiked in the 3rd time point for the ZIKV+ group, but the average intensity levels for the ZIKV- group changed relatively smoothly. Pseudouridine increased as pregnancy progressed whereas creatinine decreased. The average intensity levels in cluster 6 remained relatively constant in the ZIKV+ group but were higher at early time points (time point 1 and 2) in the ZIKV- group. The patterns of lactate in Figure 4.2 and Figure 4.3 are different; this is likely due to the extreme values of time point 3. However, in general, the patterns of lactate level change were similar in the ZIKV- and ZIKV+ group, and lactate had higher intensity levels in the ZIKV- group. Metabolites in cluster 7, including 3-aminoisobutyrate and trigonelline, had higher values on average in the ZIKV+ group, especially at time point 3; this was a different pattern from other clusters where the ZIKV+ group had, on average, lower intensity levels.

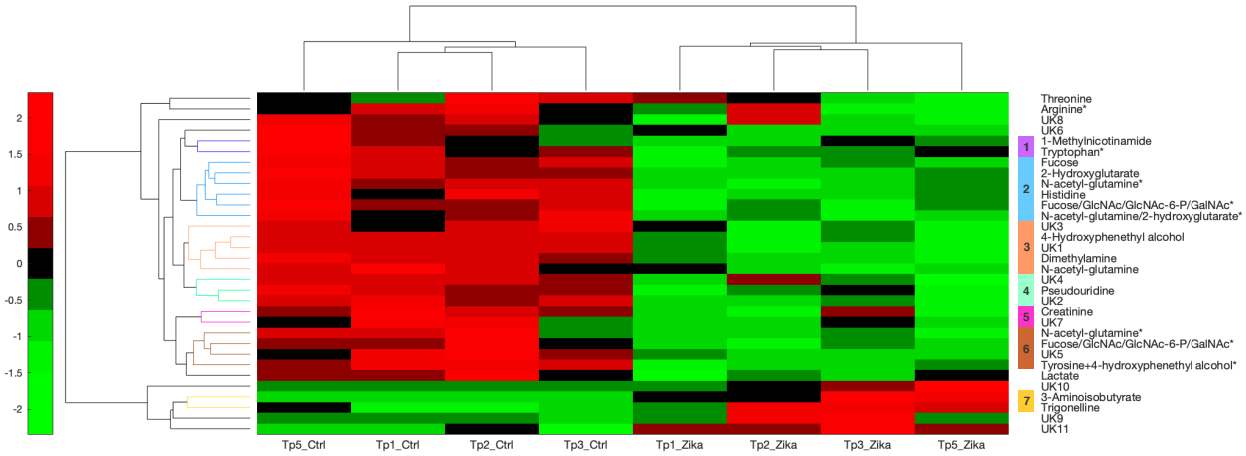


Figure 4.3: Hierarchical clustering analysis of significant metabolites and unknown features (rows) and sample time points (columns). Colors on dendrogram and before row labels show unique clusters (linkages below 1.6). Unique numbers are assigned to unique clusters. Row labels: UK: unknown feature. GlcNAc: N-acetylglucosamine; GlcNAc-6-P: N-acetyl-glucosamine 6-phosphate; GalNAc: N-acetylgalactosamine. Features with * indicate there are more than one metabolites assigned to these features, or the features are overlapped with some unknown metabolites if not specified. Mean values were used for features from the same metabolites. Features with the same annotation but have a * were not averaged, because they have different degrees of overlaps (no linear correlations above 0.8 were observed in either pair). Column labels: Tp: time point. Ctrl: samples from control patients. Zika: samples from ZIKV+ patients. Median values were taken from each time point in each group for analysis. Color on heatmap indicate z-scored intensities and colorbar is shown on the left.

Clustering on the sample groups showed a clear separation between samples from ZIKV+ patients and ZIKV- patients. For both the ZIKV- group and the ZIKV+ group, closer sample time points were grouped together. We also included an interaction term to identify differences over time between the two groups. We were only able to detect two features where there was a significant difference, likely due to low power from small sample size. These two features corresponded to a glucose peak at 3. 479 ppm (confidence score =3) and a 1-MNA peak at 9.269 ppm, with the same FDR-corrected p-value 0.0461. We described this 1-MNA peak above as it was significant without the interaction term. The glucose peak had rising trends during pregnancy in both groups. At time point 2, the glucose levels were higher in the ZIKV+ group, but then at later time points, the ZIKV- group had higher levels.

4.3 Discussion

Our study reflects long-term rather than acute change after infection in pregnant women's bodies. Only one of the four subjects whose ZIKV-infection status changed over the course of the study had ZIKV RNA detected in her serum at the time of urine sample collection. The presence of ZIKV RNA points to a relatively clear time of infection[39]. For the other 3 patients infected over the course of the study, their ZIKV-positive serum samples were at least 3 weeks after their previous ZIKV-negative visit. How long a viral infection affects the metabolome may vary from pathogen to pathogen. Additionally, the duration of an immune response effect varies from pathway to pathway[40]. Zhou et al. have reported that influenza immunizations exhibit different patterns of effect in different pathways[40]. Also, although adults usually return to normal quickly after ZIKV infection[41], the effects of infection on pregnant individuals may persist longer since there is a potential for congenital defect or infant death. ZIKV has been reported to be able to be transmitted and replicate in the fetus, and persist in the placenta for several months even until parturition[42], [43], [44]; therefore, the virus may have a longer-lasting effect on the mother's metabolome than it would on the metabolome of a non-pregnant adult. Because of these potential considerations, we treat all after-infection samples as the ZIKV+ group.

From early to late gestation, urinary histidine, 1-MNA, and trigonelline levels were reported to increase[33], [45], while creatinine was reported to decrease[27], [33]. These results are in agreement with the average longitudinal trends of these metabolites in the control group of our study. In our study, control individuals had relatively stable average intensity levels for dimethylamine, N-acetyl-glutamine and lactate urine secretion. However, dimethylamine was previously reported to increase[33], N-acetyl-glutamine was reported to decrease[33], and lactate was reported to be either constant or increasing slightly during pregnancy[27], [46]. Threonine is generally reported to increase during pregnancy, but this increase is not significant from the first to the second trimester[27], [46]. However, in our study, the level of threonine in control individuals was on average lowest at their first time point, increased in the second time point, and then had a decreasing trend for subsequent time points. Overall, though we did not statistically compare the levels of these metabolites during pregnancy, their changing trends in the control group were consistent with previous findings, adding to the confidence of our observations in spite of the small sample size and variable sampling times.

Among the significant features, some could be assigned to more than one metabolite, while other features were made up of overlapped signals, as shown in Supplementary Table A.3. Clustering these features with metabolites that had high confidence scores for their annotations may potentially give us more information about the identity of the features. For example, the feature at 2.322 ppm (labeled as N-acetyl-glutamine/2-hydroxyglutarate* in Figure 4.3) is grouped more closely with 2-hydroxyglutarate than with N-acetyl-glutamine, we can infer that this feature is likely 2-hydroxyglutarate, although N-acetyl-glutamine is also overlapped.

The feature at 7.311 ppm was annotated as a tryptophan peak but was overlapped with some other metabolite (Tryptophan* in Figure 4.3). It belonged to the same cluster as 1-MNA in our clustering analysis. 1-MNA is the major nicotinamide catabolite[47]. It can be further oxidized to N^1-methyl-2-

pyridone-5-carboxamide and N^1-methyl-4-pyridone-3-carboxamide, and all three of these metabolites are excreted in urine. Interestingly, tryptophan is associated with nicotinamide through nicotinamide adenine dinucleotide (NAD+) metabolism[48]. NAD+ can be synthesized from tryptophan through the kynurenine pathway, from nicotinic acid, or salvaged from nicotinamide, which is also the degradation product of NAD+[49], [50]. Therefore, the observation that this overlapped tryptophan feature was clustered with 1-MNA provides evidence of NAD+ metabolism disturbance. As tryptophan cannot be synthesized in humans, the decreased urinary tryptophan level suggests increased tryptophan consumption in infected patients. This is also supported by the observations that virus infections, including ZIKV infection, and other immune stimulations may boost tryptophan-NAD+ pathway[51], [52], [53], [54], [55], [56] and decrease blood tryptophan level[57], [58]. We cannot deduce the direction of NAD+ change according to 1-MNA and tryptophan level from our data. This is because firstly, nicotinamide, the precursor of 1-MNA, can both produce and consume NAD+, and both pathways can be affected by ZIKV infection[52]; secondly, although infection promotes NAD+ synthesis pathways, NAD+ is reported to be depleted in infected cells, possibly due to increased NAD+ degradation[55], [49]. Recent studies also reported disturbed NAD+ metabolism in ZIKV-infected human cell lines and mouse fetal and neonatal brains[52], [59]. Considering the role tryptophan, NAD+ and their related metabolites play in immune regulation and neuroprotection or neurotoxicity[55], [60], [61], [49], these metabolites and their pathways would be worth investigating further.

Pseudouridine and 3-aminoisobutyrate are both products of pyrimidine metabolism[62]; however, the direction of their average changes are opposite in the ZIKV+ group. 3-Aminoisobutyrate has D- and L-enantiomers in humans. These two enantiomers are involved in different metabolic pathways, have different effects, and are distributed differentially in plasma and urine[62]. The D-isomer is the catabolite of thymine and constitutes more than 90% of total 3-aminoisobutyrate in urine[63], [64]. The L-isomer is involved in valine metabolism[65]. NMR spectroscopy cannot distinguish between enantiomers, so we cannot determine the form of 3-aminoisobutyrate detected here. Though the total urinary 3-aminoisobutyrate is primarily made up of the D-isomer, the ratio between the two isomers (determined by gas chromatography) is reported to be constant among patients with different diseases[66]. Therefore, we should not ascribe the increase of 3-aminoisobutyrate in the ZIKV+ group only to the dynamics of thymine metabolism. Besides being involved in thymine metabolism, 3-aminoisobutyrate also has been reported to have a regulatory effect on energy expenditure, lipid metabolism and anti-inflammatory effect[67], [62].

Pseudouridine comes from uridine nucleotide modification in RNA. After the degradation of RNA, free pseudouridine cannot be salvaged; instead, it is excreted in urine directly[68], [69]. The level of urinary pseudouridine is thought to reflect RNA turnover rate in the body[70], [71]. Urine pseudouridine level was reported to be related to fetal growth rate[72], so the decrease of urinary pseudouridine in the ZIKV+ group may reflect slower fetal developmental growth. It has been reported that stress, including viral infection, can affect RNA pseudouridylation[73]. We did not observe uridine and its degradation end-product β-alanine in the spectra. Although we observed uracil, the uracil peak is overlapped heavily with the urea peak, so we did not include the uracil peak in the statistical analysis. Therefore, we cannot answer

the question of whether ZIKV infection potentially affects pseudouridylation in pregnant mothers or their fetuses here and need more direct experiments to determine whether this effect occurs. On the other hand, pseudouridylation also occurs in viral RNA; McIntyre et al. reported detection of pseudouridine in ZIKV virions collected from culture media[74]. Since virus replication and fetal development both happen during pregnancy, we need to consider the possibility of viral RNA contributing to total RNA when we interpret the data as well.

Interestingly, a metabolomics study on ZIKV-carrying mosquitoes reported pseudouridine, tryptophan, threonine, and histidine levels to be significantly higher, and glucose significantly lower in ZIKV-infected mosquitoes than in controls[75]. The changes in direction of these metabolites except for glucose are all opposite to what we observed here. These results suggest that these metabolite levels may reflect viral activity.

Some inflammatory cytokines induced by virus infection can affect glucose homeostasis[76], [77], [78]. For example, ZIKV infection can elevate INF-γ level in human tissues[79], and studies on mouse and human showed that INF-γ could induce insulin resistance. The effect of this insulin resistance on blood glucose levels could be compensated in healthy individuals by increased insulin secretion, but not in pre-diabetic mice[77]. Although none of the participants in our study were known diabetics, pregnant women experience similar low-grade inflammations to pre-diabetic patients[80], [81]. It is therefore reasonable to consider the contribution of insulin sensitivity change on explaining the rise of glucose level in the infected patients. In this case, if the glucose transport rate is not changed, the maternal glycemic increase would lead to an elevation in the fetal blood glucose level because glucose is facilitated diffused across the placenta[82]. Although glucose transporter expression levels have been reported to be upregulated with ZIKV infection in human cytotrophoblast and human umbilical vein endothelial cell lines[83], [84], direct investigation on how ZIKV impact glucose transplacental rate is still lacking. Further studies on glycemic status of the fetus will aid in better understanding of Zika-related pathogenesis. Interestingly, 3-aminoisobutyrate[62], 1-MNA[49], [85], 4-hydroxyphenethyl alcohol[86], and trigonelline[87] have been reported to suppress insulin resistance. Levels of these metabolites apparently increased at time point 3, and glucose levels dropped at time point 3 in the ZIKV+ group. Therefore, the fluctuation of these metabolites provides possible explanations to the change of glucose levels.

There are limitations to this study due to limited sample size and quality issues in the sample metadata. To align our data longitudinally, we needed to use the visit number rather than the gestational age of the fetus because of data quality issues (we discuss these further in Methods-Statistical Analysis-Data Quality and Processing). Because we did not have information about neonatal outcomes (i.e. developmental issues) except for stillbirth or loss of pregnancy, we were not able to study the relationship between fetal outcomes and metabolomics. The sample size of this study was also small with only 20 total individuals, and it was not well-balanced. The data were much more complete for the individuals in the ZIKV- group because ZIKV- individuals were more likely to have had samples collected for all 6 visits. This required imputation, which had further potential to introduce bias to the results. Urine can be used for diagnostic purposes, but our study does not have a large enough sample size for predictive research. Because of the nature of observational study design, the results from these data cannot be generalized to the population;

however, these results can provide direction for future studies and show that there is potential for research on the metabolites involved in and metabolic pathways of ZIKV-infection.

To conclude, our study recruited 275 pregnant women in Puerto Rico and screened out 10 ZIKV-infected patients. By comparing their longitudinal metabolomic profiles with 10 control mothers from the same cohort, we found 14 significantly changed metabolites. Thirteen of these metabolites, including 3-aminoisobutyrate, fucose, 2-hydroxyglutarate, N-acetyl-glutamine, dimethylamine, 4-hydroxyphenethyl alcohol, creatinine, lactate, threonine, histidine, pseudouridine, trigonelline, and 1-MNA, had different overall levels during pregnancy between groups, while 1-MNA and glucose levels had significantly different trends over time between the groups. These metabolites suggested disturbance in tryptophan, NAD+, pyrimidine, and glucose metabolic pathways. This study, to our knowledge, is the first to view metabolic changes in urine from ZIKV-infected pregnant humans. These results suggest that it is possible to develop a diagnostic and prognostic panel of metabolites from urine, which is readily available and non-invasive. This study included samples from the first to the third trimesters in pregnant women and included post-acute-phase data. By modeling longitudinal data, we were able to find metabolites from ZIKV+ subjects that differed over time compared to normal pregnancy. These findings improve our understanding of ZIKV pathogenesis and could be a foundation for future diagnostics.

4.4 Methods

4.4.1 Study Subjects and Sample Collection

The study subjects were enrolled as part of a prospective international observational cohort study of Zika in infants and pregnancy (ZIP study). The study protocol including study design, patient enrollment, ethical standards, sample collections and handling, diagnosis and analyses have been previously published[88]. Briefly, the ZIP study recruited pregnant women in areas in Latin America with local ZIKV transmission and included Brazil (4 sites), Colombia, Guatemala, Nicaragua, Puerto Rico (2 sites), and Peru. The overall objective of this multi-site ZIP study was to assess the association between ZIKV infection during pregnancy and any adverse maternal, fetal, and infant outcomes. The study aimed to enroll up to 10,000 pregnant women in the first and early second trimesters of pregnancy with monthly follow-ups through delivery and up to 6 weeks post-partum. Enrolled infants would be followed until at least 1 year of age. The University of Georgia partnered with one of the sites in Puerto Rico, which recruited women from the North Karst region. At each visit, women were given a physical examination and blood, urine, saliva, and vaginal fluid samples were collected as per ZIP study protocols. In between monthly visits, women would also collect a urine sample at home to be brought to the next monthly visit or collected at a biweekly visit to the clinic. Testing/diagnosis for Zika infections relied on serological assays to detect Anti-ZIKV IgM antibodies under Emergency Use Authorization (EUA) approval from the United States (U.S.) Food and Drug Administration (FDA) and EUA-approved ZIKV molecular assays for the detection of ZIKV RNA. At the time these assays were the U.S. Centers for Disease Control (CDC) Zika IgM antibody capture enzyme-linked immunosorbent assay (MAC-ELISA) and Trioplex rRT-PCR assays, respectively.

Serological testing was conducted by both the CDC site in Puerto Rico and UGA in, while rRT-PCR assays were conducted by the CDC.

4.4.2 NMR Data Collection and Analysis

Urine Sample Preparation

Urine samples were randomized first and then prepared according to the protocol by Dona *et al.*[89]. In brief, each sample was centrifuged at *3500 g* and then 540 μl of supernatant was added to 60 μl NMR buffer (1.5 M KH_2PO_4 in D_2O with a final concentration of 0.33 mM DSS, pH = 7.4). After this, we centrifuged the sample again at *4000 g* and transferred 590 μl supernatant to the 5 mm NMR tubes in racks of 96 tubes (Bruker Biospin, USA). For quality assurance, we added six buffer blank controls, nine external controls, and eight internal pooled controls to the sample set. We added buffer blank controls at the beginning, in the middle, and at the end of each 96-tube box of samples. External controls were commercially available urine samples (ethanol, drug, and nicotine-free, non-pregnant, females) from Golden West Biologicals, Inc. We prepared internal pooled controls by combining 100 μl of each sample. We randomized internal and external controls together with the clinic samples.

NMR Data Acquisition

NMR spectra were collected at 300 K on a 600 MHz Bruker AVIII-HD instrument equipped with a SampleJet autosampler and 5 mm TCI cryoprobe. We collected one-dimensional proton Nuclear Overhauser Effect Spectroscopy with presaturation (1D ^{1}H NOESY-PR) spectra from each urine sample (Bruker pulse program: noesypr1d). We monitored the DSS peak width at half height to assure the quality of 1D spectra. We reran any sample with a DSS peak-width greater than 1.5 Hz. Then, we collected two-dimensional (2D) ^{1}H-^{13}C Heteronuclear Single Quantum Coherence (HSQC) and two-dimensional ^{1}H-^{13}C Heteronuclear Single-Quantum Correlation–TOtal Correlation SpectroscopY (HSQC-TOCSY) spectra on one internal pooled control sample for peak annotation. We deposited data at the Metabolomics Workbench[90] (`https://www.metabolomicsworkbench.org`) under project PR001295 (DOI: http://dx.doi.org/10.21228/M8B13G).

NMR Spectral Processing and Annotation

We manually phased and referenced one-dimensional NMR spectra on Topspin 3.5pl7 and processed these spectra via an in-house workflow in MATLAB (the workflow is deposited with raw data in Metabolomics Workbench[90], related functions are available at `https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA/tree/master/metabolomics_toolbox`). In MATLAB, we removed the water region (4.693-4.9 ppm) and ends (<-1 ppm or >11 ppm) from the spectra. Then, we corrected the spectral baseline. Because the chemical shift of metabolites can vary greatly in urine samples, we combined results from multiple alignment methods. Most regions are aligned by constrained correlation optimized warping[91], regions between 7.808-7.842, 7.66-7.677 ppm are aligned by fast Fourier

transform cross-correlation[92]. For peaks with the largest variation, we developed an algorithm to align them. Complete details of this algorithm will be published elsewhere. Briefly, we reordered all spectra according to the chemical shift of an internal guiding peak. Other peaks, whose chemical shifts were affected by the same factor (e.g. pH) as this guiding peak, would exhibit clear curves across spectra. These curves made these peaks distinguishable from their neighbor confounding peaks, so we could trace the peaks and align them. Supplementary Fig. A.7 shows alignment performance on these highly varying peaks. We then normalized spectra by using Probabilistic Quotient Normalization (PQN)[93].

We processed the HSQC and HSQC-TOCSY spectra on NMRPipe (Version 9.5 Rev 2018.044.14.11 64-bit) with standard parameters[94] and annotated the spectra in COLMARm[95] or AssureNMR (Bruker Biospin, USA) with the bbiorefcode database. We used Statistical Total Correlation Spectroscopy (STOCSY)[96] to help match 2D peaks to 1D peaks. The annotation of 1D peaks with confidence scores from 1 (lowest) to 5 (highest) are presented in Table 4.2.

One-dimensional NMR Spectral Binning

Using an in-house manual binning algorithm (`https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA/tree/66c5e5ad8077ad49eabb36d83d9873c3f85584b0/metabolomics_toolbox/code/manual_binning`), we manually binned the full-resolution NMR spectra in MATLAB. Briefly, we drag a rectangle from the left to the right valley of a peak and the integral of the region under the curve was calculated automatically. All spectra in the study used the same bucketing boundaries. We binned peaks from multiplets into separate buckets because we performed the binning step independently from peak annotations. Since the manual binning step is subjective, we also gave a quantification confidence score for each significant feature to indicate our confidence of treating the binned peak as a real signal for metabolites. The quantification confidence scores range from 0 (lowest) to 3 (highest). Peaks with a score equal to 3 are peaks where two NMR experts have agreed that this is a true peak, whereas peaks with lower scores are either very small or are located on the shoulder of other peaks. We do not include significant peaks with a quantification confidence score lower than 3 in Table 4.2 and instead show them in Supplementary Table A.3 along with other unidentified peaks. The PCA plots (Supplementary Fig. A.8) from the full-resolution spectra and binned features indicate that the binning step kept most variance of the NMR spectra in the bins.

4.4.3 Statistical Analysis

Data Quality and Processing

Due to the small sample size for Zika-positive individuals, we collapsed the category indicating Zika status from ZIKV-negative, ZIKV-non-negative, and post-ZIKV infection into ZIKV- and ZIKV+ (Table 4.1). The ZIKV+ status includes individuals who have a current infection of ZIKV as indicated by positive PCR or non-negative IgM or who had a positive PCR or non-negative IgM test at an earlier time point in the study. Since all individuals except for one enrolled in the study in the first trimester, we aligned individuals longitudinally by the order of each individual's visit. We aligned visits longitudinally by the

visit number because we discovered some inconsistencies in the recorded "date entered" columns that we were unable to resolve and so were unable to align the visits longitudinally by the gestational age of the fetus. The visit number, on the other hand, was the same across these conflicts. Some individuals had multiple visits within the same visit time point. For example, an individual may have had two visits which would fall into the category of 4 weeks post entry. We averaged the intensity scores at each peak for these individuals at these time points. While some individuals had more than one visit, many individuals had missing visits, especially ZIKV+ individuals at visits 4 and 6. For these individuals, we imputed their missing data using the implementation of the CART algorithm in the R package **mice**[97]. For data imputation, we treated each peak and each time point as independent (i.e. we did not take into account correlation structure within each individual between time points) and imputed based on the Zika status at each peak and at each time point using CART. Finally, some individuals went from testing negative for ZIKV to testing positive for ZIKV. If these individuals did not have a subsequent visit within the same time point where they tested positive, we treated their first observation as missing and imputed the first observation using the **mice** package. For the case when there was a subsequent visit within the same time point where the individual tested positive for ZIKV, we treated this second observation as the only observation within that time point and dropped the ZIKV- observation.

Nonparametric Modeling

Because the pattern of change over time points was not the same across different peaks and did not appear to follow a known linear or non-linear pattern, we fit a non-parametric mixed effect model for factorial designs using the R package **nparLD**[98]. While non-parametric models unfortunately limit interpretation of covariates, since these models use a rank-based methodology, they do not rely on distributional or asymptotic assumptions, are robust to outliers and are thus suitable for small sample size data with unknown distributions[98]. Because the data in this analysis have these properties, we used a non-parametric model to identify differences between the distributions of the intensity scores over time for ZIKV- individuals and ZIKV+ individuals. In this analysis, we treated each peak as independent from all other peaks. We fit 649 non-parametric mixed effects models, one for each peak. Since **nparLD** requires complete data, we needed to impute a large portion of the peak intensities, especially for ZIKV+ individuals at the fourth and sixth visits. To ensure consistency across results, we fit these 649 models to three separate sets of data: the full data set which included all six visit time points, a subset that included the first three visit time points, and a subset that included the first three visit time points as well as the fifth visit time point. Ultimately, we used the third set of data for our analysis to achieve a balance between using as much data as possible while not using too much imputed data. The peaks we identified as significant did not change much across the different subsets of data. Since we had a limited amount of demographic information available to us, for each peak, we used Zika status and visit time point as fixed effects, the ID of each individual as the random effect to capture within-subject variability, and intensity of the binned NMR spectra as the response variable for our model. In statistical terms, we treated each independent random vector $\mathbf{X}_{izk} = \{X_{izkt}\}$, indicating the intensity level of peak i for individual k in infection group z ($z = 0$ for ZIKV- and $z = 1$ for ZIKV+) over time $t = 1, 2, 3, 5$ as an observation from distribution $\mathbf{F}_{iz} = \{F_{izt}\}$,

$t = 1, 2, 3, 5$. We used a non-parametric factorial model as described in **nparLD**[98] to test the following null hypotheses of no main effect Z from ZIKV infection status, i.e. the mean distribution over time for the ZIKV+ and ZIKV- groups are equal, and no interaction effect between the infection status and time, i.e. the distribution at each time point for each ZIKV group is equal to the mean distribution of that ZIKV group plus the mean distribution of that time point minus the total mean distribution across all time points and groups:

$$H_0^F(Z) : \overline{F}_{i0\cdot} = \overline{F}_{i1\cdot}$$
$$H_0^F(ZT) : F_{izt} = \overline{F}_{iz\cdot} + \overline{F}_{i\cdot t} - \overline{F}_{i\cdot\cdot}$$

where $\overline{F}_{i0\cdot} = \frac{1}{4}(F_{i01} + F_{i02} + F_{i03} + F_{i05})$ is the mean distribution at peak i over all times for ZIKV- samples; $\overline{F}_{i1\cdot} = \frac{1}{4}(F_{i11} + F_{i12} + F_{i13} + F_{i15})$ is the mean distribution at peak i over all times for ZIKV+ samples; $\overline{F}_{iz\cdot} = \frac{1}{4}(F_{iz1} + F_{iz2} + F_{iz3} + F_{iz5})$ is the mean distribution at peak i over time for infection group z; $\overline{F}_{i\cdot t} = \frac{1}{2}(F_{i0t} + F_{i1t})$ is the mean distribution at peak i over ZIKV- and ZIKV+ samples for times t; and $F_{i\cdot\cdot} = \frac{1}{8}(F_{i01} + F_{i02} + F_{i03} + F_{i05} + F_{i11} + F_{i12} + F_{i13} + F_{i15})$ is the mean distribution over all observations for peak i. For further details on nonparametric hypothesis tests for repeated measures designs, refer to Akritas and Arnold (1994)[99]. After fitting the 649 models under this methodology and extracting p-values, we then performed Benjamini and Hochberg false discovery rate (FDR) correction [35] and identified significant peaks as peaks that had a lower corrected p-value than 0.15. The reference substrate DSS had a corrected p-value of 0.16.

Clustering Analysis

To identify peaks with similar intensity patterns over time, we hierarchically clustered the intensities of NMR features from time point 1, 2, 3, and 5 (with imputation) via the 'clustergram' function from the MATLAB Bioinformatics Toolbox. We included only significant features with a quantification confidence score of 3 in the analysis. We used the mean value for features from the same metabolites. We collected the median for each time point of each group for the analysis. We standardized data for each row (metabolite) and clustered on both dimensions. The linkage was calculated based on average Euclidean distance, and linkages lower than 1.6 were colored as clusters. This unique color threshold was close to the median linkage (1.5577) in the column dimension.

Bibliography

[1] V. C. Agumadu and K. Ramphul, "Zika virus: A review of literature," *Cureus*, vol. 10, no. 7, 2018. DOI: 10.7759/cureus.3025.

[2] P. Simmonds, P. Becher, J. Bukh, *et al.*, "ICTV virus taxonomy profile: Flaviviridae," *Journal of General Virology*, vol. 98, no. 1, pp. 2–3, Jan. 2017, ISSN: 14652099. DOI: 10.1099/JGV.0.000672/CITE/REFWORKS. [Online]. Available: https://www.microbiologyresearch.org/content/journal/jgv/10.1099/jgv.0.000672.

[3] G. Gutiérrez-Bugallo, L. A. Piedra, M. Rodriguez, *et al.*, "Vector-borne transmission and evolution of zika virus," *Nature Ecology & Evolution*, vol. 3, no. 4, pp. 561–569, Apr. 2019, ISSN: 2397-334X. DOI: 10.1038/s41559-019-0836-z. [Online]. Available: https://doi.org/10.1038/s41559-019-0836-z.

[4] B. D. Foy, K. C. Kobylinski, J. L. Chilson Foy, *et al.*, "Probable non-vector-borne transmission of Zika virus, Colorado, USA.," eng, *Emerging infectious diseases*, vol. 17, no. 5, pp. 880–882, May 2011, ISSN: 1080-6059 (Electronic). DOI: 10.3201/eid1705.101939.

[5] C. Charlier, M. C. Beaudoin, T. Couderc, O. Lortholary, and M. Lecuit, *Arboviruses and pregnancy: maternal, fetal, and neonatal effects*, Oct. 2017. DOI: 10.1016/S2352-4642(17)30021-4. [Online]. Available: http://dx.doi.org/10.1016/.

[6] D. Baud, D. J. Gubler, B. Schaub, M. C. Lanteri, and D. Musso, "An update on Zika virus infection.," eng, *Lancet (London, England)*, vol. 390, no. 10107, pp. 2099–2109, Nov. 2017, ISSN: 1474-547X (Electronic). DOI: 10.1016/S0140-6736(17)31450-2.

[7] V.-M. Cao-Lormeau, A. Blake, S. Mons, *et al.*, "Guillain-Barré Syndrome outbreak associated with Zika virus infection in French Polynesia: a case-control study," *The Lancet*, vol. 387, no. 10027, pp. 1531–1539, Apr. 2016, ISSN: 0140-6736. DOI: 10.1016/S0140-6736(16)00562-6. [Online]. Available: http://www.thelancet.com/article/S0140673616005626/fulltext%20http://www.thelancet.com/article/S0140673616005626/abstract%20https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)00562-6/abstract.

[8] L. C. Rodrigues, *Microcephaly and Zika virus infection*, May 2016. DOI: 10.1016/S0140-6736(16)00742-X. [Online]. Available: http://www.hygiene-publique.gov.pf/IMG/.

[9] C. K. Shapiro-Mendoza, M. E. Rice, R. R. Galang, *et al.*, "Pregnancy Outcomes After Maternal Zika Virus Infection During Pregnancy — U.S. Territories, January 1, 2016–April 25, 2017," *MMWR. Morbidity and Mortality Weekly Report*, vol. 66, no. 23, pp. 615–621, Jun. 2019, ISSN: 0149-21951545-861X. DOI: 10.15585/MMWR.MM6623E1. [Online]. Available: https://www.facebook.com/CDCMMWR.

[10] M. R. Reynolds, A. M. Jones, E. E. Petersen, *et al.*, "Vital Signs: Update on Zika Virus–Associated Birth Defects and Evaluation of All U.S. Infants with Congenital Zika Virus Exposure — U.S. Zika Pregnancy Registry, 2016," *MMWR. Morbidity and Mortality Weekly Report*, vol. 66, no. 13, pp. 366–373, Apr. 2019, ISSN: 0149-21951545-861X. DOI: 10.15585/MMWR.MM6613E1. [Online]. Available: https://www.facebook.com/cdcmmwr.

[11] A. Gulland, *Zika virus is a global public health emergency, declares WHO*, 2016. DOI: 10.1136/bmj.i657. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26839247/.

[12] B. McCloskey and T. Endericks, "The rise of zika infection and microcephaly: What can we learn from a public health emergency?" eng, *Public health*, vol. 150, pp. 87–92, Sep. 2017, 28651111[pmid], ISSN: 1476-5616. DOI: 10.1016/j.puhe.2017.05.008. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28651111.

[13] A. D. T. Barrett, "Current status of Zika vaccine development: Zika vaccines advance into clinical evaluation.," eng, *NPJ vaccines*, vol. 3, no. 1, p. 24, Dec. 2018, ISSN: 2059-0105 (Electronic). DOI: 10.1038/s41541-018-0061-9. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29900012/.

[14] L. Eyer, R. Nencka, I. Huvarová, *et al.*, "Nucleoside inhibitors of zika virus," *Journal of Infectious Diseases*, vol. 214, no. 5, pp. 707–711, Sep. 2016, ISSN: 15376613. DOI: 10.1093/infdis/jiw226. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/27234417/.

[15] M. Xu, E. M. Lee, Z. Wen, *et al.*, "Identification of small-molecule inhibitors of Zika virus infection and induced neural cell death via a drug repurposing screen," *Nature Medicine*, vol. 22, no. 10, pp. 1101–1107, Oct. 2016, ISSN: 1546170X. DOI: 10.1038/nm.4184. [Online]. Available: https://pubchem.ncbi.nlm..

[16] N. J. Barrows, R. K. Campos, S. T. Powell, *et al.*, "A Screen of FDA-Approved Drugs for Inhibitors of Zika Virus Infection," *Cell Host and Microbe*, vol. 20, no. 2, pp. 259–270, Aug. 2016, ISSN: 19346069. DOI: 10.1016/j.chom.2016.07.004. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/27476412/.

[17] R. Y. Cao, Y. fen Xu, T. H. Zhang, *et al.*, "Pediatric drug nitazoxanide: A potential choice for control of Zika," *Open Forum Infectious Diseases*, vol. 4, no. 1, pp. 1–5, 2017, ISSN: 23288957. DOI: 10.1093/OFID/OFX009.

[18] Q. Liang, Z. Luo, J. Zeng, *et al.*, "Zika Virus NS4A and NS4B Proteins Deregulate Akt-mTOR Signaling in Human Fetal Neural Stem Cells to Inhibit Neurogenesis and Induce Autophagy," *Cell Stem Cell*, vol. 19, no. 5, pp. 663–671, 2016, ISSN: 18759777. DOI: 10.1016/j.stem.2016.07.019. arXiv: 15334406. [Online]. Available: http://dx.doi.org/10.1016/j.stem.2016.07.019.

[19] A. I. Chiramel and S. M. Best, "Role of autophagy in Zika virus infection and pathogenesis," *Virus Research*, vol. 254, no. June, pp. 34–40, 2018, ISSN: 18727492. DOI: 10.1016/j.virusres.2017.09.006. [Online]. Available: http://dx.doi.org/10.1016/j.virusres.2017.09.006.

[20] B. Cao, L. A. Parnell, M. S. Diamond, and I. U. Mysorekar, "Inhibition of autophagy limits vertical transmission of Zika virus in pregnant mice," *The Journal of Experimental Medicine*, vol. 214, no. 8, pp. 2303–2313, 2017, ISSN: 0022-1007. DOI: 10.1084/jem.20170957. [Online]. Available: http://www.jem.org/lookup/doi/10.1084/jem.20170957.

[21] C. F. O. Melo, J. Delafiori, D. N. de Oliveira, *et al.*, "Serum metabolic alterations upon ZIKA infection," *Frontiers in Microbiology*, vol. 8, no. OCT, pp. 1–10, 2017, ISSN: 1664302X. DOI: 10.3389/fmicb.2017.01954.

[22] E. d. C. Nunes, A. M. de Filippis, T. D. E. Pereira, *et al.*, "Untargeted metabolomics insights into newborns with congenital zika infection," *Pathogens*, vol. 10, no. 4, p. 468, Apr. 2021, ISSN: 20760817. DOI: 10.3390/PATHOGENS10040468/S1. [Online]. Available: https://www.mdpi.com/2076-0817/10/4/468/htm%20https://www.mdpi.com/2076-0817/10/4/468.

[23] N. R. d. C. Faria, A. B. Chaves-Filho, L. C. J. Alcantara, *et al.*, "Plasma lipidome profiling of newborns with antenatal exposure to Zika virus," *PLOS Neglected Tropical Diseases*, vol. 15, no. 4, e0009388, Apr. 2021, ISSN: 1935-2735. DOI: 10.1371/JOURNAL.PNTD.0009388. [Online]. Available: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0009388.

[24] Q. Chen, J. Gouilly, Y. J. Ferrat, *et al.*, "Metabolic reprogramming by Zika virus provokes inflammation in human placenta," *Nature Communications*, vol. 11, no. 1, Dec. 2020. DOI: 10.1038/S41467-020-16754-Z.

[25] H. C. Leier, J. B. Weinstein, J. E. Kyle, *et al.*, "A global lipid map defines a network essential for Zika virus replication," *Nature Communications*, vol. 11, no. 1, Dec. 2020, ISSN: 20411723. DOI: 10.1038/S41467-020-17433-9. [Online]. Available: /pmc/articles/PMC7374707/%20/pmc/articles/PMC7374707/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7374707/.

[26] D. N. de Oliveira, E. O. Lima, C. F. Melo, *et al.*, "Inflammation markers in the saliva of infants born from Zika-infected mothers: exploring potential mechanisms of microcephaly during fetal development," *Scientific Reports*, vol. 9, no. 1, Dec. 2019, ISSN: 20452322. DOI: 10.1038/s41598-019-49796-5. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31541139/.

[27] S. O. Diaz, A. S. Barros, B. J. Goodfellow, *et al.*, "Following healthy pregnancy by nuclear magnetic resonance (NMR) metabolic profiling of human urine.," eng, *Journal of proteome research*, vol. 12, no. 2, pp. 969–979, Feb. 2013, ISSN: 1535-3907 (Electronic). DOI: 10.1021/pr301022e. [Online]. Available: https://pubs.acs.org/sharingguidelines.

[28] I. F. Duarte, S. O. Diaz, and A. M. Gil, "NMR metabolomics of human blood and urine in disease research," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 93, pp. 17–26, 2014, ISSN: 1873264X. DOI: 10.1016/j.jpba.2013.09.025. [Online]. Available: http://dx.doi.org/10.1016/j.jpba.2013.09.025.

[29] A. S. Edison, M. Colonna, G. J. Gouveia, *et al.*, *NMR: Unique Strengths That Enhance Modern Metabolomics Research*, Jan. 2021. DOI: 10.1021/acs.analchem.0c04414. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/acs.analchem.0c04414.

[30] A.-H. H. Emwas, R. Roy, R. T. McKay, *et al.*, "NMR Spectroscopy for Metabolomics Research.," eng, *Metabolites*, vol. 9, no. 7, Jun. 2019, ISSN: 2218-1989 (Print). DOI: 10.3390/metabo9070123. [Online]. Available: /pmc/articles/PMC6680826/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6680826/.

[31] S. Bouatra, F. Aziat, R. Mandal, *et al.*, "The Human Urine Metabolome," *PLoS ONE*, vol. 8, no. 9, 2013, ISSN: 19326203. DOI: 10.1371/journal.pone.0073076.

[32] A. H. Emwas, E. Saccenti, X. Gao, *et al.*, "Recommended strategies for spectral processing and post-processing of 1D1H-NMR data of biofluids with a particular focus on urine," *Metabolomics*, vol. 14, no. 3, pp. 1–23, 2018, ISSN: 15733890. DOI: 10.1007/s11306-018-1321-4. [Online]. Available: http://dx.doi.org/10.1007/s11306-018-1321-4.

[33] D. Sachse, L. Sletner, K. Mørkrid, *et al.*, "Metabolic Changes in Urine during and after Pregnancy in a Large, Multiethnic Population-Based Cohort Study of Gestational Diabetes," *PLoS ONE*, vol. 7, no. 12, L. K. Rogers, Ed., e52399, Dec. 2012, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0052399. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0052399.

[34] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*, ser. The Wadsworth & Brooks/Cole statistics/probability series. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984. [Online]. Available: https://cds.cern.ch/record/2253780.

[35] Y. Benjamini 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)*, vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: http://www.jstor.com/stable/2346101.

[36] W. S. Cleveland, "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979, ISSN: 0162-1459. DOI: 10.2307/2286407. [Online]. Available: https://www.jstor.org/stable/2286407 (visited on 02/11/2021).

[37] R. R. Sokal, *A statistical method for evaluating systematic relationships*. Univ Kans Sci Bull, Feb. 1958, vol. 38, pp. 1409–1438, ISBN: 0001948000237.

[38] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[39] C. Eppes, M. Rac, J. Dunn, *et al.*, *Testing for Zika virus infection in pregnancy: key concepts to deal with an emerging epidemic*, Mar. 2017. DOI: 10.1016/j.ajog.2017.01.020. [Online]. Available: http://dx.doi.org/10.1016/j.ajog.2017.01.020.

[40] W. Zhou, M. R. Sailani, K. Contrepois, *et al.*, "Longitudinal multi-omics of host–microbe dynamics in prediabetes," *Nature*, vol. 569, no. 7758, pp. 663–671, May 2019, ISSN: 14764687. DOI: 10.1038/s41586-019-1236-x. [Online]. Available: /pmc/articles/PMC6666404/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6666404/.

[41] L. R. Petersen, D. J. Jamieson, A. M. Powers, and M. A. Honein, "Zika Virus," *New England Journal of Medicine*, vol. 374, no. 16, L. R. Baden, Ed., pp. 1552–1563, Apr. 2016, ISSN: 0028-4793. DOI: 10.1056/NEJMra1602113. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMra1602113.

[42] J. Bhatnagar, D. B. Rabeneck, R. B. Martines, *et al.*, "Zika virus RNA replication and persistence in brain and placental tissue," *Emerging Infectious Diseases*, vol. 23, no. 3, pp. 405–414, Mar. 2017, ISSN: 10806059. DOI: 10.3201/eid2303.161499. [Online]. Available: /pmc/articles/PMC5382738/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5382738/.

[43] L. de Noronha, C. Zanluca, M. Burger, *et al.*, "Zika Virus Infection at Different Pregnancy Stages: Anatomopathological Findings, Target Cells and Viral Persistence in Placental Tissues," *Frontiers in Microbiology*, vol. 0, no. SEP, p. 2266, Sep. 2018, ISSN: 1664-302X. DOI: 10.3389/FMICB.2018.02266.

[44] A. J. Hirsch, V. H. J. Roberts, P. L. Grigsby, *et al.*, "Zika virus infection in pregnant rhesus macaques causes placental dysfunction and immunopathology," *Nature Communications 2018 9:1*, vol. 9, no. 1, pp. 1–15, Jan. 2018, ISSN: 2041-1723. DOI: 10.1038/s41467-017-02499-9. [Online]. Available: https://www.nature.com/articles/s41467-017-02499-9.

[45] K. Shibata, T. Fukuwatari, M. Murakami, and R. Sasaki, "Increase in conversion of tryptophan to niacin in pregnant rats," in *Advances in Experimental Medicine and Biology*, vol. 527, Springer, Boston, MA, 2003, pp. 435–441. DOI: 10.1007/978-1-4615-0135-0_51.

[46] A. M. Gil, D. Duarte, J. Pinto, and A. S. Barros, "Assessing Exposome Effects on Pregnancy through Urine Metabolomics of a Portuguese (Estarreja) Cohort," *Journal of Proteome Research*, vol. 17, no. 3, pp. 1278–1289, Mar. 2018, ISSN: 15353907. DOI: 10.1021/acs.jproteome.7b00878.

[47] D. A. Bender, "NIACIN | Physiology," in *Encyclopedia of Food Sciences and Nutrition (Second Edition)*, B. Caballero, Ed., Second Edition, Oxford: Academic Press, 2003, pp. 4119–4128, ISBN: 978-0-12-227055-0. DOI: https://doi.org/10.1016/B0-12-227055-X/00827-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B012227055X008270.

[48] T. Fukuwatari and K. Shibata, *Nutritional aspect of tryptophan metabolism*, 2013. DOI: 10.4137/IJTR.S11588. [Online]. Available: /pmc/articles/PMC3729278/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3729278/.

[49] N. Xie, L. Zhang, W. Gao, *et al.*, "NAD+ metabolism: pathophysiologic mechanisms and therapeutic potential," *Signal Transduction and Targeted Therapy 2020 5:1*, vol. 5, no. 1, pp. 1–37, Oct. 2020, ISSN: 2059-3635. DOI: 10.1038/s41392-020-00311-7. [Online]. Available: https://www.nature.com/articles/s41392-020-00311-7.

[50] D. A. Bender, B. I. Magboul, and D. Wynick, "Probable mechanisms of regulation of the utilization of dietary tryptophan, nicotinamide and nicotinic acid as precursors of nicotinamide nucleotides in the rat," *British Journal of Nutrition*, vol. 48, no. 1, pp. 119–127, Jul. 1982, ISSN: 0007-1145. DOI: 10.1079/bjn19820094.

[51] A. W. S. Yeung, W. Wu, M. Freewan, R. Stocker, N. J. C. King, and S. R. Thomas, "Flavivirus infection induces indoleamine 2,3-dioxygenase in human monocyte-derived macrophages via tumor necrosis factor and NF-κB," *Journal of Leukocyte Biology*, vol. 91, no. 4, pp. 657–666, Apr. 2012, ISSN: 1938-3673. DOI: 10.1189/JLB.1011532. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1189/jlb.1011532%20https://onlinelibrary.wiley.com/doi/abs/10.1189/jlb.1011532%20https://jlb.onlinelibrary.wiley.com/doi/10.1189/jlb.1011532.

[52] Z. Pang, J. Chong, G. Zhou, *et al.*, "MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights," *Nucleic Acids Research*, vol. 49, no. W1, W388–W396, Jul. 2021, ISSN: 0305-1048. DOI: 10.1093/NAR/GKAB382. [Online]. Available: https://academic.oup.com/nar/article/49/W1/W388/6279832.

[53] F. M. Marim, D. C. Teixeira, C. M. Queiroz-Junior, *et al.*, "Inhibition of Tryptophan Catabolism Is Associated With Neuroprotection During Zika Virus Infection," *Frontiers in Immunology*, vol. 0, p. 2843, Jul. 2021, ISSN: 1664-3224. DOI: 10.3389/FIMMU.2021.702048.

[54] K. Schröcksnadel, B. Wirleitner, C. Winkler, and D. Fuchs, "Monitoring tryptophan metabolism in chronic immune activation," *Clinica Chimica Acta*, vol. 364, no. 1-2, pp. 82–90, Feb. 2006, ISSN: 0009-8981. DOI: 10.1016/J.CCA.2005.06.013.

[55] J. R. Moffett and M. A. Namboodiri, "Tryptophan and the immune response," *Immunology and Cell Biology*, vol. 81, no. 4, pp. 247–265, 2003, ISSN: 08189641. DOI: 10.1046/j.1440-1711.2003.t01-1-01177.x.

[56] F. Fallarino, C. K. Lim, R. Grant, *et al.*, "Quinolinate as a Marker for Kynurenine Metabolite Formation and the Unresolved Question of NAD + Synthesis During Inflammation and Infection," *Frontiers in Immunology | www.frontiersin.org*, vol. 11, p. 31, 2020. DOI: 10.3389/fimmu.2020.00031. [Online]. Available: www.frontiersin.org.

[57] A. Cozzi, A. L. Zignego, R. Carpendo, *et al.*, "Low serum tryptophan levels, reduced macrophage IDO activity and high frequency of psychopathology in HCV patients," *Journal of Viral Hepatitis*, vol. 13, no. 6, pp. 402–408, Jun. 2006, ISSN: 1365-2893. DOI: 10.1111/J.1365-2893.2005.00706.X. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2893.2005.00706.x%20https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2893.2005.00706.x%20https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2893.2005.00706.x.

[58] M. L. Harrison, A. S. Wolfe, J. Fordyce, J. Rock, A. A. García, and J. A. Zuñiga, "The additive effect of type 2 diabetes on fibrinogen, von Willebrand factor, tryptophan and threonine in people living with HIV," *Amino Acids*, vol. 51, no. 5, 2019, ISSN: 14382199. DOI: 10.1007/s00726-019-02715-4.

[59] G. Xu, S. Li, X. Liu, *et al.*, "PARP-1 mediated cell death is directly activated by ZIKV infection," *Virology*, vol. 537, pp. 254–262, Nov. 2019, ISSN: 0042-6822. DOI: 10.1016/J.VIROL.2019.08.024.

[60] H. R. Cetina Biefer, A. Vasudevan, and A. Elkhal, *Aspects of tryptophan and nicotinamide adenine dinucleotide in immunity: A new twist in an old tale*, 2017. DOI: 10.1177/1178646917713491. [Online]. Available: /pmc/articles/PMC5476425/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5476425/.

[61] R. hao Mu, Y. zhi Tan, L. li Fu, *et al.*, "1-Methylnicotinamide attenuates lipopolysaccharide-induced cognitive deficits via targeting neuroinflammation and neuronal apoptosis," *International Immunopharmacology*, vol. 77, Dec. 2019, ISSN: 18781705. DOI: 10.1016/j.intimp.2019.105918.

[62] D. A. Tanianskii, N. Jarzebska, A. L. Birkenfeld, J. F. O'sullivan, and R. N. Rodionov, *Beta-aminoisobutyric acid as a novel regulator of carbohydrate and lipid metabolism*, Mar. 2019. DOI: 10.3390/nu11030524. [Online]. Available: /pmc/articles/PMC6470580/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6470580/.

[63] E. Solem, E. Jellum, and L. Eldjarn, "The absolute configuration of β-aminoisobutyric acid in human serum and urine," *Clinica Chimica Acta*, vol. 50, no. 3, pp. 393–403, Feb. 1974, ISSN: 00098981. DOI: 10.1016/0009-8981(74)90159-4.

[64] A. B. P. van KUILENBURG, A. E. M. STROOMER, H. van LENTHE, N. G. G. M. ABELING, and A. H. van GENNIP, "New insights in dihydropyrimidine dehydrogenase deficiency: a pivotal role for beta-aminoisobutyric acid?" *Biochemical Journal*, vol. 379, no. 1, pp. 119–124, Apr. 2004, ISSN: 0264-6021. DOI: 10.1042/BJ20031463. [Online]. Available: /biochemj/article/379/1/119/43200/New-insights-in-dihydropyrimidine-dehydrogenase.

[65] F. P. Kupiecki and M. J. Coon, "The enzymatic synthesis of beta-aminoisobutyrate, a product of valine metabolism, and of beta-alanine, a product of beta-hydroxypropionate metabolism.," *The Journal of biological chemistry*, vol. 229, no. 2, pp. 743–54, Dec. 1957, ISSN: 0021-9258. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/13502336.

[66] A. H. van Gennip, J. P. Kamerling, P. K. de Bree, and S. K. Wadman, "Linear relationship between the R- and S-enantiomers of β-aminoisobutyric acid in human urine," *Clinica Chimica Acta*, vol. 116, no. 3, pp. 261–267, Nov. 1981, ISSN: 00098981. DOI: 10.1016/0009-8981(81)90045-0.

[67] K. Begriche, J. Massart, A. Abbey-Toby, A. Igoudjil, P. Lettéron, and B. Fromenty, "Beta-aminoisobutyric acid prevents diet-induced obesity in mice with partial leptin deficiency.," eng, *Obesity (Silver Spring, Md.)*, vol. 16, no. 9, pp. 2053–2067, Sep. 2008, ISSN: 1930-7381 (Print). DOI: 10.1038/oby.2008.337.

[68] A. Dlugajczyk and J. J. Eiler, "Lack of Catabolism of 5-Ribosyluracil in Man," *Nature*, vol. 212, no. 5062, pp. 611–612, 1966, ISSN: 1476-4687. DOI: 10.1038/212611a0. [Online]. Available: https://doi.org/10.1038/212611a0.

[69] C. W. Gehrke and K. C. Kuo, "Patterns of Urinary Excretion of Modified Nucleosides," *Cancer Research*, vol. 39, no. 4, pp. 1150–1153, 1979, ISSN: 15387445.

[70] G. Sander, H. Topp, G. Heller-Schöch, J. Wieland, and G. Schöch, "Ribonucleic acid turnover in man: RNA catabolites in urine as measure for the metabolism of each of the three major species of RNA," *Clinical Science*, vol. 71, no. 4, pp. 367–374, Oct. 1986, ISSN: 0143-5221. DOI: 10.1042/cs0710367. [Online]. Available: https://doi.org/10.1042/cs0710367.

[71] S. M. M. Orue, J. Balcells, J. A. Guada, and C. Castrillo, "Endogenous purine and pyrimidine derivative excretion in pregnant sows," *British Journal of Nutrition*, vol. 73, no. 3, pp. 375–385, Mar. 1995, ISSN: 0007-1145. DOI: 10.1079/bjn19950040.

[72] G. Schoech, E. Hoting, and F. J. Sohulte, "Excretion of RNA catabolites in pregnancy," *Pediatric Research*, vol. 14, no. 12, p. 1420, 1980, ISSN: 1530-0447. DOI: 10.1203/00006450-198012000-00074. [Online]. Available: https://doi.org/10.1203/00006450-198012000-00074.

[73] R. Netzband and C. T. Pager, "Epitranscriptomic marks: Emerging modulators of RNA virus gene expression," *Wiley Interdisciplinary Reviews: RNA*, vol. 11, no. 3, May 2020, ISSN: 17577012. DOI: 10.1002/wrna.1576. [Online]. Available: /pmc/articles/PMC7169815/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7169815/.

[74] W. McIntyre, R. Netzband, G. Bonenfant, *et al.*, "Positive-sense RNA viruses reveal the complexity and dynamics of the cellular and viral epitranscriptomes during infection," *Nucleic Acids Research*, vol. 46, no. 11, pp. 5776–5791, Jun. 2018, ISSN: 13624962. DOI: 10.1093/nar/gky029. [Online]. Available: https://academic.oup.com/nar/article/46/11/5776/4823209.

[75] M. G. Onyango, G. M. Attardo, E. T. Kelly, *et al.*, "Zika Virus Infection Results in Biochemical Changes Associated With RNA Editing, Inflammatory and Antiviral Responses in Aedes albopictus," *Frontiers in Microbiology*, vol. 11, p. 2456, Oct. 2020, ISSN: 1664302X. DOI: 10.3389/fmicb.2020.559035. [Online]. Available: www.frontiersin.org.

[76] V. A. Koivisto, R. Pelkonen, and K. Cantell, "Effect of Interferon on Glucose Tolerance and Insulin Sensitivity," *Diabetes*, vol. 38, no. 5, pp. 641–647, May 1989, ISSN: 0012-1797. DOI: 10.2337/DIAB.38.5.641. [Online]. Available: https://diabetes.diabetesjournals.org/content/38/5/641%20https://diabetes.diabetesjournals.org/content/38/5/641.abstract.

[77] M. Šestan, S. Marinović, I. Kavazović, *et al.*, "Virus-Induced Interferon-γ Causes Insulin Resistance in Skeletal Muscle and Derails Glycemic Control in Obesity," *Immunity*, vol. 49, no. 1, 164–177.e6, Jul. 2018, ISSN: 1074-7613. DOI: 10.1016/J.IMMUNI.2018.05.005.

[78] A. Toniolo, G. Cassani, A. Puggioni, *et al.*, "The diabetes pandemic and associated infections: Suggestions for clinical microbiology," *Reviews in Medical Microbiology*, vol. 30, no. 1, pp. 1–17, Jan. 2019. DOI: 10.1097/MRM.0000000000000155. [Online]. Available: https://journals.lww.com/revmedmicrobiol/Fulltext/2019/01000/The_diabetes_pandemic_and_associated_infections_.1.aspx.

[79] K. Rabelo, L. J. de Souza, N. G. Salomão, *et al.*, "Zika Induces Human Placental Damage and Inflammation," *Frontiers in Immunology*, vol. 0, p. 2146, Sep. 2020, ISSN: 1664-3224. DOI: 10.3389/FIMMU.2020.02146.

[80] T. Lekva, E. R. Norwitz, P. Aukrust, and T. Ueland, "Impact of Systemic Inflammation on the Progression of Gestational Diabetes Mellitus," *Current Diabetes Reports*, vol. 16, no. 4, pp. 1–11, Apr. 2016, ISSN: 15390829. DOI: 10.1007/S11892-016-0715-9/FIGURES/1. [Online]. Available: https://link.springer.com/article/10.1007/s11892-016-0715-9.

[81] K. Luc, A. Schramm-Luc, T. J. Guzik, and T. P. Mikolajczyk, "Oxidative stress and inflammatory markers in prediabetes and diabetes.," eng, *Journal of physiology and pharmacology : an official journal of the Polish Physiological Society*, vol. 70, no. 6, Dec. 2019, ISSN: 1899-1505 (Electronic). DOI: 10.26402/jpp.2019.6.01.

[82] N. P. Illsley, "CURRENT TOPIC: Glucose Transporters in the Human Placenta," *Placenta*, vol. 21, no. 1, pp. 14–22, 2000, ISSN: 0143-4004. DOI: https://doi.org/10.1053/plac.1999.0448. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0143400499904484.

[83] S. Singh, P. K. Singh, H. Suhail, *et al.*, "AMP-Activated Protein Kinase Restricts Zika Virus Replication in Endothelial Cells by Potentiating Innate Antiviral Responses and Inhibiting Glycolysis," *The Journal of Immunology*, vol. 204, no. 7, pp. 1810–1824, Apr. 2020, ISSN: 0022-1767. DOI: 10.4049/JIMMUNOL.1901310. [Online]. Available: https://www.jimmunol.org/content/204/7/1810%20https://www.jimmunol.org/content/204/7/1810.abstract.

[84] D. Vota, M. Torti, D. Paparini, *et al.*, "Zika virus infection of first trimester trophoblast cells affects cell migration, metabolism and immune homeostasis control," *Journal of Cellular Physiology*, vol. 236, no. 7, pp. 4913–4925, Jul. 2021, ISSN: 1097-4652. DOI: 10.1002/JCP.30203. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/jcp.30203%20https://onlinelibrary.wiley.com/doi/abs/10.1002/jcp.30203%20https://onlinelibrary.wiley.com/doi/10.1002/jcp.30203.

[85] C. Y, Z. J, L. P, L. C, and L. L, "N1-methylnicotinamide ameliorates insulin resistance in skeletal muscle of type 2 diabetic mice by activating the SIRT1/PGC-1α signaling pathway," *Molecular medicine reports*, vol. 23, no. 4, Apr. 2021, ISSN: 1791-3004. DOI: 10.3892/MMR.2021.11909. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33576435/.

[86] R. Chandramohan, L. Pari, A. Rathinam, and B. A. Sheikh, "Tyrosol, a phenolic compound, ameliorates hyperglycemia by regulating key enzymes of carbohydrate metabolism in streptozotocin induced diabetic rats," *Chemico-Biological Interactions*, vol. 229, pp. 44–54, 2015, ISSN: 0009-2797. DOI: https://doi.org/10.1016/j.cbi.2015.01.026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0009279715000320.

[87] J. Zhou, L. Chan, and S. Zhou, "Trigonelline: A Plant Alkaloid with Therapeutic Potential for Diabetes and Central Nervous System Disease," *Current Medicinal Chemistry*, vol. 19, no. 21, pp. 3523–3531, Oct. 2012, ISSN: 09298673. DOI: 10.2174/092986712801323171.

[88] J. F. Lebov, J. F. Arias, A. Balmaseda, *et al.*, "International prospective observational cohort study of Zika in infants and pregnancy (ZIP study): Study protocol," *BMC Pregnancy and Childbirth*, vol. 19, no. 1, pp. 1–10, Aug. 2019, ISSN: 14712393. DOI: 10.1186/S12884-019-2430-4/TABLES/4. [Online]. Available: https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-019-2430-4.

[89] A. C. Dona, B. Jiménez, H. Schafer, *et al.*, "Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping," *Analytical Chemistry*, vol. 86, no. 19, pp. 9887–9894, Oct. 2014, ISSN: 15206882. DOI: 10.1021/ac5025039. [Online]. Available: http://www.ebi.ac..

[90] M. Sud, E. Fahy, D. Cotter, *et al.*, "Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools," *Nucleic Acids Research*, vol. 44, no. D1, pp. D463–D470, 2016, ISSN: 13624962. DOI: 10.1093/nar/gkv1042. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26467476/.

[91] N.-P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping," *Journal of Chromatography A*, vol. 805, no. 1, pp. 17–35, 1998, ISSN: 0021-9673. DOI: https://doi.org/10.1016/S0021-9673(98)00021-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021967398000211.

[92] J. W. H. Wong, C. Durante, and H. M. Cartwright, "Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets.," eng, *Analytical chemistry*, vol. 77, no. 17, pp. 5655–5661, Sep. 2005, ISSN: 0003-2700 (Print). DOI: 10.1021/ac050619p.

[93] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn, "Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics.," *Analytical chemistry*, vol. 78, no. 13, pp. 4281–90, Jul. 2006, ISSN: 0003-2700. DOI: 10.1021/ac051632c. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16808434%20https://pubs.acs.org/doi/10.1021/ac051632c.

[94] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax, "NMRPipe: a multidimensional spectral processing system based on UNIX pipes.," eng, *Journal of biomolecular NMR*, vol. 6, no. 3, pp. 277–293, Nov. 1995, ISSN: 0925-2738 (Print). DOI: 10.1007/BF00197809.

[95] K. Bingol, D. W. Li, B. Zhang, and R. Brüschweiler, "Comprehensive metabolite identification strategy using multiple two-dimensional NMR spectra of a complex mixture implemented in the COLMARm web server," *Analytical Chemistry*, vol. 88, no. 24, pp. 12 411–12 418, 2016, ISSN: 15206882. DOI: 10.1021/acs.analchem.6b03724.

[96] O. Cloarec, M. E. Dumas, A. Craig, *et al.*, "Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic 1H NMR data sets," *Analytical Chemistry*, vol. 77, no. 5, pp. 1282–1289, 2005, ISSN: 00032700. DOI: 10.1021/ac048630x.

[97] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011. [Online]. Available: https://www.jstatsoft.org/v45/i03/.

[98] K. Noguchi, Y. Gel, E. Brunner, and F. Konietschke, "nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments," *Journal of Statistical Software, Articles*, vol. 50, no. 12, pp. 1–23, 2012, ISSN: 1548-7660. DOI: 10.18637/jss.v050.i12. [Online]. Available: https://www.jstatsoft.org/v050/i12.

[99] M. G. Akritas and S. F. Arnold, "Fully nonparametric hypotheses for factorial designs i: Multivariate repeated measures designs," *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 336–343, Mar. 1994. DOI: 10.1080/01621459.1994.10476475.

Chapter 5
METABOLIC ADAPTATIONS AFTER BIRTH: A DIRECT COMPARISON BETWEEN SHEEP FETAL AND NEONATAL METABOLOME

5.1 Introduction

Cortisol is an important hormone for fetal organ maturation and for preparing the fetus for transition to extrauterine life after birth [1]. Treating high-preterm-risk mothers with synthetic glucocorticoid during the peripartum period can stimulate fetal lung maturation, therefore increasing the survival rate of preterm infants [2]. However, excess cortisol can be harmful to fetuses as it stimulates transition from cell proliferation to differentiation and maturation [1],[3], thereby forcing underdeveloped fetuses to the next stage of development. Previous research has found that excess cortisol exposure on healthy pregnant animals during late gestation restricts fetal body growth [3] and has adverse impacts on fetal endocrine [4], cardiovascular [5], [6], and neuron system [4] development. Excess maternal cortisol exposure can elevate preterm [7] and stillbirth rate [8].

Cortisol exposure also changes fetal metabolic status [9]. When investigating this at a system level in sheep, it has been reported that chronic elevation of maternal cortisol levels during late gestation leads to changes in the fetal heart [5] and placenta [10] metabolome. Because prenatal exposure to cortisol has been reported to have long-term impacts on children and even adult offspring [11], it is important to know if elevated maternal cortisol continues to impact the neonatal metabolome postpartum.

Using sheep as a model organism allows us to collect blood samples directly from fetuses non-destructively [2]. The fetal and newborn serum metabolome under cortisol exposure have been previously characterized in sheep [5], as well as in human infant blood [12]. However, these studies focused only on prenatal or on postnatal time points. A direct comparison on the same animals before and after birth has not been explored. This has been generally true regardless of any research questions, even though the transition from in utero growth to independent living is of great importance in newborn animals. Most studies compared before and after birth in different samples [13], [14], [15] due to the destructive manner of fetal

sample collection. This type of comparison is subject to inter-individual variation, which impacts omics statistical results. Also, comparing the metabolic changes before and after birth has mostly been made to a limited number of metabolites, especially those related to energy supply [16], [17], [18]. There are many other metabolites that perform important biological roles, but these have not as yet been investigated in terms of their change with birth.

The sheep fetal growth rate slows down during the last few days of gestation [19], [20]. Concentrations of some characterized metabolites are stable during this time but change abruptly several hours before birth. Several hours after birth, metabolic status reverses back to a relatively stable level [16]. As a result, the neonate transitions to a new free-living life stage. In our study, we used a previously established sheep model of chronic maternal cortisol exposure during late gestation [21] to profile the metabolome of neonates and to compare this with the fetal metabolome of the same animals. We treated 25 ewes with cortisol from gestational age (GA) 115 days ($\sim$0.8 gestation) until labor. Twelve untreated ewes were used as a control group. Fetal serum samples were then collected at GA 136 $\pm$ 2 days, and neonatal serum and heart tissue samples were collected two days after birth. We then used NMR to analyze the metabolome of the intact heart tissue and serum samples. We present here metabolomic comparisons between cortisol treated (CORT) and untreated animals, as well as fetal and neonatal serum samples from the same animals.

5.2 Methods

5.2.1 Experimental Design and Animal Use

The use of all animals in this study was approved by the Institutional Animal Care and Use Committee (IACUC) of the University of Florida (UF). Animals were unrestrained from movement and had free access to food and water during the study. Thirty-seven Dorset cross ewes, including 36 with singletons and one with twins, were randomly assigned to the cortisol treatment group (CORT; n=25, 24 singleton pregnancies and 1 twin pregnancy) and control group (n=12). Each ewe in the CORT group was continuously infused with cortisol (Solu-Cortef; hydrocortisone sodium succinate in sodium phosphate; Pfizer, New York, NY), 1mg/kg/day through an infusion pump (3D Micro Infusion Pump; Strategic Applications Inc., Lake Villa, IL) from gestational day 115 or 116 until delivery. In previous studies, this dose increased maternal plasma cortisol levels to approximately 1.5-fold of that of the control sheep from 115-130 days of gestation [21].

Catheters were installed at 124 $\pm$ 4 gestational days in fetal tibial arteries and maternal femoral arteries and veins. Anesthesia was performed by using isoflurane inhalation (1-2%) and ketamine infusion (4-6 mg/kg/h). This detailed procedure was reported before [22]. We performed preoperative treatment (meloxicam, 1 mg/kg; Cipla USA Inc, Warren, NJ) or intraoperative treatment (flunixin meglumine, 0.5 mg/kg; Merck Animal Health) on all ewes. For postoperative care, the ewes were treated with meloxicam (1 mg/kg *sid* for three days) and/or flunixin meglumine (0.5 mg/kg) and antibiotics (Polyflex, ampicillin, 12.5 mg/kg *bid* for five days; Boehringer Ingelheim Vetmedica, St. Joseph, MO, or Naxcel, ceftiofur

sodium, 2.2 mg/kg/day *sid* for three days; Zoetis, Parsippany-Troy Hills, NJ). We also measured their rectal temperature twice a day for five days.

At gestational age (GA) 136 $\pm$ 2 days, 8 ml fetal blood sample was collected from the catheters of each fetus. Serum was obtained and $\sim$200 μl of it was used for metabolomics. Two days after birth, 8 ml neonatal blood sample was collected from each sheep and 200 μl serum was used for metabolomics analysis. These neonates were then euthanized with an intravenous overdose of euthanasia solution (Euthasol; Fort Worth, TX). Immediately after euthanasia, $\sim$100 mg of heart tissue samples were collected from the left ventricular free wall, right ventricular free wall, and interventricular septum, respectively. Two preterm animals were euthanized one day after birth because the condition of the lamb met criteria (e.g., body temperature, alertness, time to standing) established for euthanasia in consultation with the veterinary staff and approved by the UF IACUC. Fetuses who died before or at birth were excluded from sample collection for the metabolomics study. Samples were snap-frozen in liquid nitrogen, stored at -80 °C after collection, and packed with dry ice for shipment.

5.2.2 Sample Preparation and NMR Acquisition

Sample Preparation

We prepared heart tissue samples of 22 animals from the left ventricular free wall, right ventricular free wall, and interventricular septum independently as three batches for high-resolution magic angle spinning (HRMAS) proton nuclear magnetic resonance (^{1}H-NMR). The sample experimental order within each batch was randomized, with all batches following the same order. For each sample, we cut $\sim$30 mg tissues on a clean weighing boat, then added 30 μl D_2O (Cambridge Isotope Labs, Inc.) with 20/3 mM sodium 3-(trimethylsilyl)propane-1-sulfonate (DSS; Cambridge Isotope Labs, Inc.). The tissue sample and buffer were then transferred to a 4-mm zirconium dioxide rotor with Kel-F cap and polytetrafluoroethylene spacer (Bruker Biospin). Sample weights were 29.41 $\pm$ 3.05 mg for the control group and 29.56 $\pm$ 3.38 mg for the CORT group. We also prepared buffer blank samples approximately every eight samples as well as at the beginning and end of each batch to monitor potential contamination.

Two preterm animals were born later than the other animals, so their heart samples were analyzed in another batch. To monitor the batch effect, we randomly picked and re-analyzed one piece of tissue from previous batches from the preterm and term CORT groups. To directly compare batches and to avoid the impact of the freeze-thaw cycle, we also analyzed another two pieces from the same heart part from the same animal. We analyzed these four pieces of tissues in the same batch with the six pieces of tissues from the two new preterm animals. Their experimental orders were randomized before sample preparation.

Fetal and neonatal serum samples were prepared in different batches. We randomized fetal and neonatal sample order separately. We prepared serum samples according to the protocol of Dona et al [23]. Briefly, we first centrifuged samples at 14000 rpm, 4 °C for five minutes. For quality assurance, we took 50 μl of supernatant from each fetal sample and pooled them to make one internal control sample. We then aliquoted it to two internal control samples. Because a large number of neonatal serum samples had low volume, we only took 40 μl of supernatant from each sample and pooled them to one internal control

sample. We excluded hemolyzed samples from pooling to avoid potential impairment from hemolyzation on spectral annotation. We took 300 μl of supernatant from each sample and mixed them with 350 μl NMR buffer (0.075 M NaH_2PO_4, 4.6 mM TSP, 6.2 mM NaN_3, pH=7.4; NaH_2PO_4: Fisher-Scientific, TSP: 3-trimethyl-silyl-[2,2,3,3-$^2H_4$]propionic acid, Sigma-Aldrich, NaN_3: Sigma-Aldrich). If the sample had less than 300 μl, we added buffer to make the total volume 650 μl. We centrifuged samples again at 14000 rpm, 4 °C for five minutes, and transferred 600 μl supernatant from each sample to 5 mm NMR tubes in racks of 96 tubes (Bruker Biospin, USA). We included blank buffer controls and external controls (human serum quality control samples) before and after each batch and included internal controls in randomized order.

For spectral annotation, we pooled heart tissue NMR samples after HRMAS acquisition by group (control, CORT term, CORT preterm) and extracted them by 80:20 methanol:H_2O. All NMR samples from the CORT preterm group were pooled. However, because control and CORT term groups had excess amounts of samples for extraction, only half of the samples were pooled in these two groups. The samples pooled included one sample from each animal and another five random samples. We included one extraction blank control to check if there was any contamination. We extracted these pooled samples according to the protocol Walejko et al. reported [14]. Briefly, we added 80:20 methanol:H_2O to samples (1 ml/100 mg) and homogenized them with beads. We took supernatants for vacuum concentration and reconstituted them in NMR buffer (1 M sodium phosphate with a final concentration of 0.33 mM DSS, pH = 7.0). After centrifugation, 590 μl of supernatants were transferred to 5 mm NMR tubes. We used two representative unextracted serum samples from fetal and neonatal samples for annotation.

NMR Data Acquisition

Heart tissue samples were measured by HRMAS at 10 °C on a Bruker NEO 600 MHz NMR spectrometer with a 4-mm CMP-HRMAS probe. The rotor spinning rate was set to 6000 Hz. We collected a one-dimensional (1D) ^{1}H nuclear Overhauser effect spectroscopy with water presaturation (noesypr1d) spectrum for each sample. We reran spectra with bad shimming. If we could not shim it well, we then prepared another sample from the original heart tissue. If the spectra contained very high lipid signals, we recollected spectra on another sample from the same tissue.

Serum spectra were collected at 37 °C on a 600 MHz Bruker AVIII-HD instrument equipped with a SampleJet autosampler and 5 mm TCI cryoprobe. We collected PROJECT with water presaturation (PROJECTpr1d) spectra to get good baselines [24].

Extracted heart tissue samples and unextracted serum samples were used for annotation. We collected 2-dimensional (2D) ^{1}H-^{13}C Heteronuclear Single Quantum Coherence (HSQC), 2D ^{1}H-^{1}H TOtal Correlation SpectroscopY (TOCSY) to annotate both heart tissue and serum 1D spectra. We also collected 2D ^{1}H-^{13}C HSQC-TOCSY spectra on heart samples. For quality control, we also measured buffer blanks and collected 1D noesypr spectra on each sample before and after 2D experiments to check for sample degradation.

5.2.3 Data Processing and Statistical Analysis

Data Processing

We manually phased and baseline corrected each 1D spectrum and referenced them to DSS and TSP resonances (0.0 ppm) for the heart and serum samples, respectively, on Topspin 3.5pl7. We further processed these spectra on MATLAB using an in-house developed workflow. For heart samples, we replaced the water region (4.78-5.16 ppm) with zeros and removed spectral ends <-0.5 ppm or >10 ppm. We aligned spectra by Peak Alignment by Fast Fourier Transform (PAFFT) [25] alignment, then normalized them by probabilistic quotient normalization PQN [26]. We bucketed the spectra into 335 bins semiautomatically with an in-house peak picking algorithm and custom optimized bucketing [27] MATLAB scripts. We auto-scaled these bins for multivariate analysis. For serum samples, we replaced the water region (4.56-5.18 ppm) with zeros and removed spectra ends <-0.5 ppm or >10 ppm. We combined regions aligned by recursive alignment by FFT (RAFFT) [25] from -0.5-3.553, 4.289-7.599, 7.899-10 ppm, and regions aligned by Constrained Correlation Optimized Warping (CCOW) [28] 3.533-4.289, 7.599-7.899 ppm. We normalized, bucketed, and scaled them by the same method used for the heart spectra. In the end, we obtained 333 bins from serum spectra for multivariate analysis.

Statistical Analysis

We first used PCA to check if there were any batch effects on samples from the last two preterm animals. We did not observe a batch effect (Figure 5.1), so these samples were statistically analyzed with the other samples (samples used for monitoring the batch effect were not included in the statistical analysis). We also checked batch differences between fetal and neonatal samples. External controls from the two batches were closely distributed on the PCA plot (Figure 5.2), indicating no batch effects. Some serum samples were hemolyzed, but these samples were not significantly different from other samples we observed by PCA (Figure 5.3), PLS-DA (Fig 5.4), or by visual examination. Therefore, these samples were not excluded from the analysis.

The PLS Toolbox (version 881, Eigenvector Research, Inc., Manson, WA) was used to perform orthogonal signal-corrected partial least squares discriminant analysis (OSC-PLS-DA) for comparison between two groups. The matrix containing group information was also auto-scaled. We used random subsets with ten data splits and 20 iterations for cross-validation. For small sample sets (n<20), we used seven data splits and 15 iterations.

We performed two-sample Student's t-tests on each bucket of heart spectra to compare cortisol treatment groups on all or term samples, or to compare preterm vs. term on all or CORT samples. We adjusted the p-values with Benjamini-Hochberg false discovery rate (FDR) and set the threshold of significance to 0.1. We show the results in Table 5.2 with one representative bucket for each metabolite for simplicity.

We performed two-way ANOVA on each serum spectra bucket to analyze the effect of cortisol treatment and sample time point. As no buckets were significantly different between CORT and control serum samples, we included samples from both groups in paired analysis. We paired samples from the same animals (animals with missing fetal or neonatal samples were excluded) and then performed paired

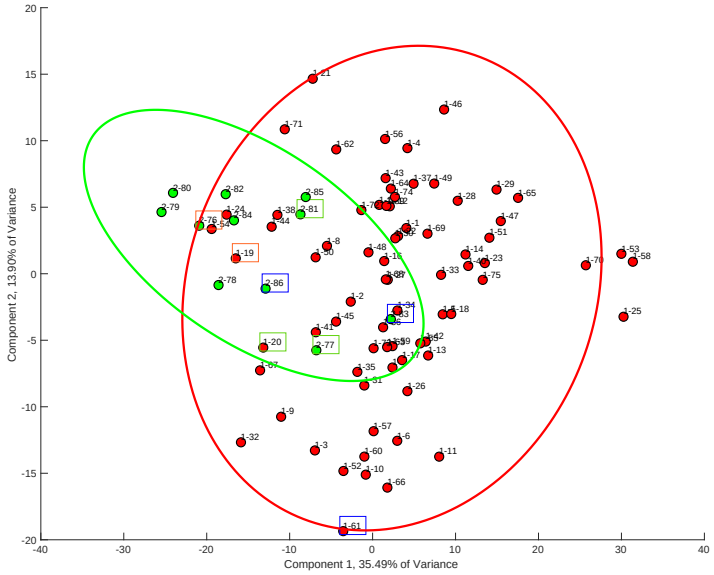


Figure 5.1: Scores plot of principal component analysis (PCA) on neonatal heart samples from all batches. Samples from the two late preterm animals and samples used for monitoring the batch effect are colored in green. The rest of samples are colored in red. NMR samples from the same tissue samples are boxed in the same color.

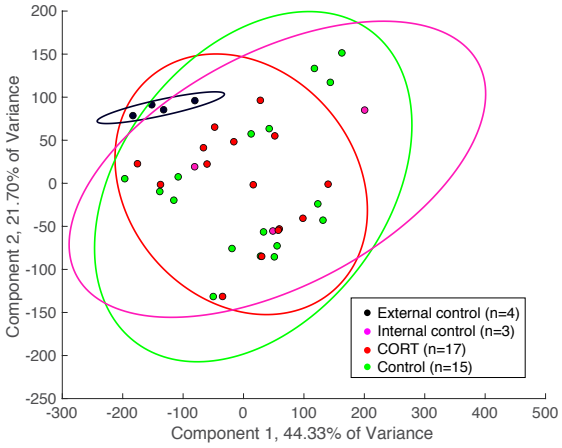


Figure 5.2: Scores plot of PCA on external, internal controls and all serum samples.

t-test on all metabolites. Each metabolite was represented by one bucket containing resonance(s) from the metabolite. All these univariate analyses were controlled for FDR in the same way discussed above, with the same significant threshold.

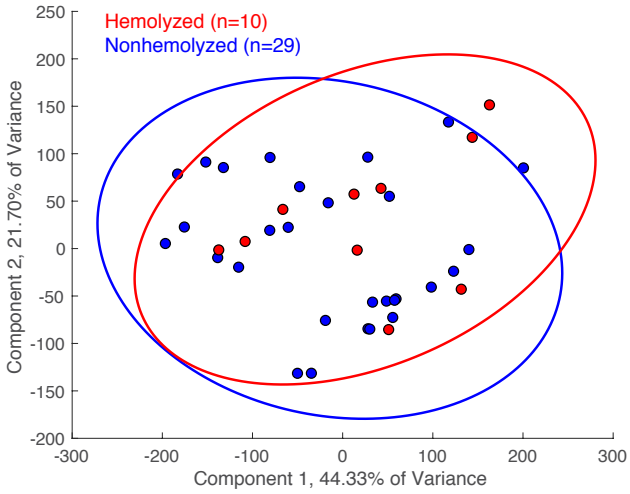


Figure 5.3: Scores plot of PCA on all serum samples comparing hemolyzed and nonhemolyzed samples.

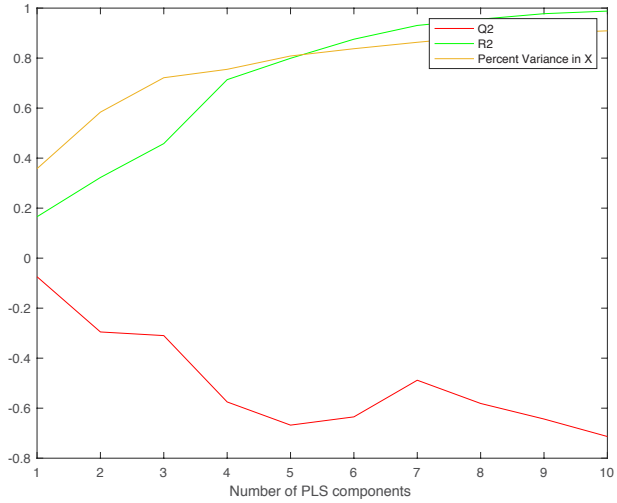


Figure 5.4: Partial least squares discriminant analysis (PLS-DA) model performance on comparing hemolyzed and nonhemolyzed samples.

Finally, we put significantly changed metabolites from the paired t-tests into MetaboAnalyst 5.0 [29] for pathway analysis. We excluded lipids from the list because the identification was not narrowed down to a compound level. We chose the *Bos taurus* (cow) pathway library which contains 81 pathways as the reference library because a sheep library was not included in this tool and cow is the closest species that was available. We used over-representation analysis with a hypergeometric test and determined node centrality by relative betweenness centrality. We define a pathway with an FDR-corrected p-value <0.1 as a significantly changed pathway.

5.3 Results

5.3.1 Cortisol Induced Metabolic Differences in Neonates Are Associated With Preterm Birth

The CORT group had higher rates of preterm and stillbirth. Of the 26 fetuses from the 25 ewes in the CORT group, six fetuses were born alive preterm at gestational day 132-136 (133± 2), eight fetuses died before or at birth and were delivered preterm, and three term fetuses were dead in utero. By contrast, three fetuses in the control group died before birth, two of which were delivered preterm. The other nine control and nine CORT sheep were born alive at full term. Therefore, the CORT group had a 54% preterm rate and 42% stillbirth rate, while the control group had a 17% preterm rate and a 25% stillbirth rate. The pregnancy outcomes are summarized in Table 5.1.

Table 5.1: Fetal birth conditions

	CORT (n=26)	Control (n=12)
Preterm livebirth	6	0
Preterm stillbirth	8	2
Term livebirth	9	9
Term stillbirth	3	1
Preterm birth rate	0.54	0.17
Stillbirth rate	0.42	0.25
Gender of livebirth animals (male/all)	0.47	0.56

Heart tissue from the 24 live birth two-day-old neonates were analyzed by high-resolution magic angle spinning (HRMAS) proton nuclear magnetic resonance (^{1}H-NMR) for metabolomics studies. Using principal component analysis (PCA), no separations were observed between the left and right ventricular free wall and the interventricular septum from these animals. This result is consistent with previous observations [5], so the three parts were treated as replicates.

PCA showed differences between CORT and the control group, but we observed greater differences between preterm and term neonates (Figure 5.5). Considering that all preterm animals were from the CORT group, we compared both CORT vs. control and preterm CORT vs. term CORT by orthogonal signal corrected partial least squares discriminant analysis (OSC-PLS-DA) and found that five of the top ten metabolites that drove the separation between CORT vs. control neonates were also the most important metabolites that differentiate between preterm and term CORT neonates (Figure 5.6). These metabolites were lipids, glutamate, 1,2-propanediol, glycerophosphocholine (GPC), and an unidentified feature (unknown2). This result indicates that cortisol treatment-induced metabolic changes in two-day-old neonatal hearts were associated with preterm changes in this sample set.

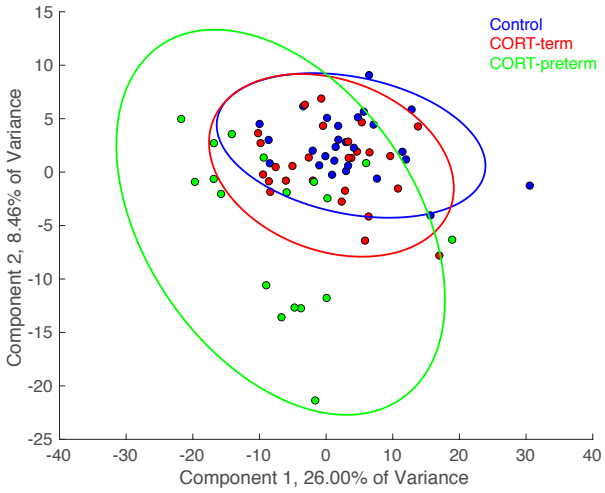


Figure 5.5: Scores plot of PCA on all neonatal heart samples.

We then compared the metabolite concentration levels between CORT vs. control, CORT preterm vs. CORT term, and control vs. CORT term samples. Significant metabolites are listed in Table 5.2. Only aspartate showed a significantly different level between term CORT and control animals, but its concentration was also significantly lower in preterm samples compared to term CORT samples. This explains why aspartate is the most important metabolite that drives the separation between CORT and control samples (Figure 5.6B). Fourteen of the 20 significant metabolites between treatment groups were also significant between term types on CORT animals but did not show significant difference in the term CORT and control animals.

5.3.2 Cortisol Did Not Change The Serum Metabolome Both Before and After Birth

Serum samples were collected at 4 ± 2 days before birth and two days after birth for both groups. Preterm animals did not have serum samples collected before birth because they were born before the scheduled sample collection time and sera was only collected from three preterm neonates. Considering the significant difference between term and preterm animals, serum analysis was not performed on the preterm neonates.

When comparing serum samples, we did not observe separations between CORT and control groups but saw clear differences between fetal and neonatal samples on the PCA scores plot (Figure 5.7). Supervised multivariate analysis by OSC-PLS-DA gave the same results when the model stability was very low (Q^2=-0.1889) in CORT vs. control comparisons, whereas fetus vs. neonates demonstrated stable separation (Q^2=0.7778) (Figure 5.8). When comparing samples at one time point, the model cross-validation results were Q^2=-0.2593 and -0.0433 for fetal and neonatal sera, respectively, which indicates there was still

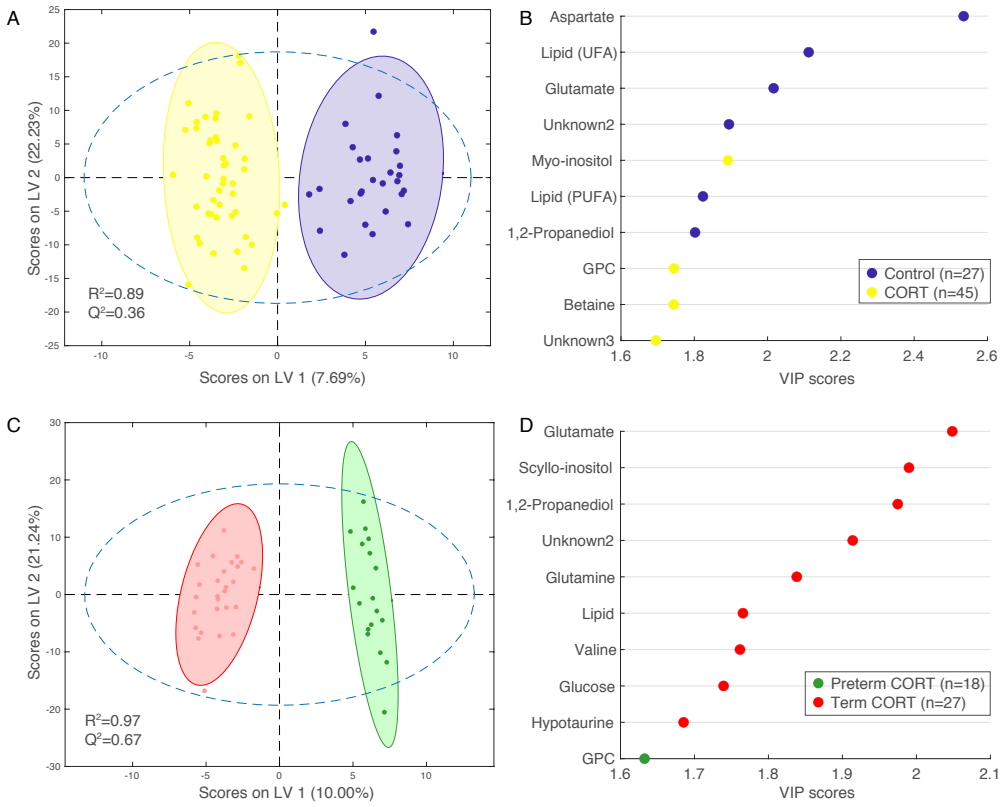


Figure 5.6: Orthogonal signal-corrected PLS-DA (OSC-PLS-DA) between cortisol treated (CORT) vs. control neonatal heart tissues (A, B) and between preterm vs. term CORT neonatal heart tissues (C, D). A: OSC-PLS-DA scores plot showing separation between CORT (n=45 from 15 animals) and control (n=27 from 9 animals) samples. B: variable importance of projection (VIP) score plot showing the 10 most important metabolites that contributed to the separation between CORT and control samples in the OSC-PLS-DA component 1. C: OSC-PLS-DA scores plot showing separation between preterm (n=18 from 6 animals) and term (n=27 from 9 animals) CORT samples. D: VIP score plot showing the 10 most important metabolites that contributed to the separation between preterm and term CORT samples in the OSC-PLS-DA component 1. UFA: unsaturated fatty acid, PUFA: polyunsaturated fatty acid, GPC: glycerophosphocholine.

no separation between the two treatment groups. On the other hand, separation between fetal and neonatal sera were observed in both CORT and the control group (Figure 5.9). To further check if there were any metabolites that have significantly different levels between CORT and control animals, we performed a two-way ANOVA assessing the effect of cortisol treatment and sample time point. No metabolites showed significant differences for cortisol treatment or cortisol treatment × time interaction.

Table 5.2: Significantly different metabolites between cortisol treated and control or between preterm and term cortisol treated neonatal hearts

Metabolite	Relative mean ± std			FDR-adjusted p-value			
	Control	CORT term	CORT preterm	CORT vs. control on all neonates	Preterm vs. term on CORT neonates	CORT vs. control on term neonates	Confidence Score*
Aspartate	5.04 ± 0.68	4.34 ± 0.60	3.98 ± 0.32	2.23E-05	0.09	0.03	4
Lipid (-CH=CH-CH_2-)	17.72 ± 1.52	16.21 ± 2.13	14.35 ± 2.10	1.18E-03	0.04	0.19	1
Glutamate	17.10 ± 2.33	15.58 ± 2.03	12.40 ± 2.90	2.30E-03	0.01	0.23	4
Unknown2	19.48 ± 3.86	17.34 ± 3.80	12.86 ± 3.22	4.80E-03	0.01	0.30	NA
Myo-inositol	16.41 ± 1.65	17.77 ± 2.16	19.83 ± 2.67	4.80E-03	0.04	0.23	4
Lipid (PUFA)	4.88 ± 0.40	4.58 ± 0.48	4.23 ± 0.51	0.01	0.10	0.24	1
1,2-Propanediol	8.62 ± 1.50	7.81 ± 1.66	5.43 ± 2.02	0.01	0.01	0.37	4
Betaine	16.86 ± 3.14	19.45 ± 4.46	22.82 ± 5.50	0.01	0.10	0.24	4
Unknown3	1.43 ± 0.42	1.74 ± 0.43	2.15 ± 0.78	0.01	0.10	0.23	NA
GPC	9.05 ± 1.01	10.05 ± 2.49	11.84 ± 2.57	0.02	0.09	0.34	4
Proline	6.30 ± 0.39	6.10 ± 0.42	5.53 ± 0.74	0.02	0.02	0.37	4
Unknown6	1.34 ± 0.13	1.46 ± 0.21	1.59 ± 0.32	0.02	0.21	0.24	NA
Lysine	7.44 ± 0.62	7.03 ± 0.69	6.76 ± 0.76	0.02	0.37	0.25	4
Unknown4	3.92 ± 1.20	3.45 ± 1.18	2.71 ± 0.94	0.05	0.10	0.52	NA
Hypotaurine	3.67 ± 0.27	3.56 ± 0.33	3.34 ± 0.29	0.06	0.09	0.62	4
Creatine/creatine-phosphate	81.83 ± 12.31	90.85 ± 13.54	88.54 ± 13.98	0.06	0.70	0.23	4/3
Threonine	6.15 ± 0.42	5.88 ± 0.50	5.85 ± 0.53	0.07	0.91	0.27	4
Phenylalanine	3.66 ± 0.48	3.51 ± 0.60	3.08 ± 0.57	0.09	0.09	0.71	3
Lipid (-CH_2CH_2CO-)	12.17 ± 2.30	11.12 ± 2.59	10.32 ± 2.21	0.09	0.43	0.48	1
Lipid (-CH_3)	26.85 ± 4.21	24.66 ± 5.64	23.16 ± 4.76	0.09	0.50	0.47	1
Valine	8.91 ± 0.92	8.77 ± 0.79	7.84 ± 0.89	0.10	0.01	0.87	4
Lipid (glycerol moiety)	6.84 ± 0.86	6.71 ± 0.82	5.94 ± 0.36	0.10	0.01	0.87	1
Unknown7	7.41 ± 0.58	7.33 ± 0.58	6.77 ± 0.57	0.13	0.03	0.88	NA
Glutamine	16.30 ± 2.56	16.18 ± 2.32	13.22 ± 2.77	0.15	0.01	0.94	4
Glutathione-reduced	4.96 ± 0.48	4.90 ± 0.35	4.57 ± 0.41	0.18	0.04	0.87	3
AMP-sulfate	5.06 ± 0.42	5.01 ± 0.38	4.62 ± 0.52	0.18	0.04	0.89	2
Inosine	11.08 ± 1.55	10.96 ± 1.61	9.63 ± 1.19	0.21	0.04	0.92	3
Glucose	5.73 ± 0.58	5.78 ± 0.66	4.87 ± 1.03	0.25	0.01	0.92	4
Phenethylamine	1.35 ± 0.29	1.34 ± 0.21	1.12 ± 0.30	0.32	0.03	0.99	2
Unknown8	4.83 ± 0.48	5.11 ± 0.51	4.78 ± 0.42	0.38	0.10	0.29	NA
Scyllo-inositol	9.40 ± 1.39	10.36 ± 1.09	8.80 ± 1.30	0.49	0.01	0.20	3
Fructose-1,6-bisphosphate	7.46 ± 0.71	7.52 ± 0.54	6.92 ± 1.05	0.50	0.07	0.91	4
UDP-GlcNAc	1.45 ± 0.19	1.44 ± 0.24	1.59 ± 0.21	0.51	0.10	0.94	2

For control and CORT term group, n=27 from 9 animals; for CORT preterm group, n=18 from 6 animals. PUFA: polyunsaturated fatty acid, GPC: glycerophosphocholine, AMP: adenosine monophosphate, UDP-GlcNAc: uridine diphosphate N-acetylglucosamine.

*Confidence score is defined as follows: 1) putatively characterized compounds or compound classes, 2) 1 D NMR matches to literature and/or database (BMRB and/or HMDB), 3) HSQC or TOCSY matches on COLMARm, 4) HSQC and TOCSY match on COLMARm, and 5) verified by spiking. HSQC: heteronuclear single-quantum coherence, TOCSY: total correlation spectroscopy.

5.3.3 Serum Metabolome Changed Significantly After Birth

To control inter-individual variability, we performed a paired t-test on fetal and neonatal samples. Four animals lacking serum samples at one of the two time points were excluded from the paired analysis. Figure 5.10 shows paired t-test results on full-resolution spectra. Comparison between fetal and neonatal serum metabolites is summarized in Table 5.3. In the 39 identified metabolites and chemical groups, concentra-

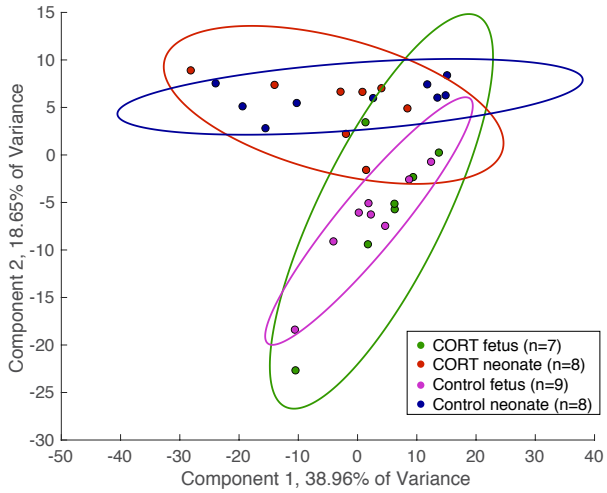


Figure 5.7: Scores plot of PCA on all fetal and neonatal serum samples from CORT and control sheep, external and internal controls not included.

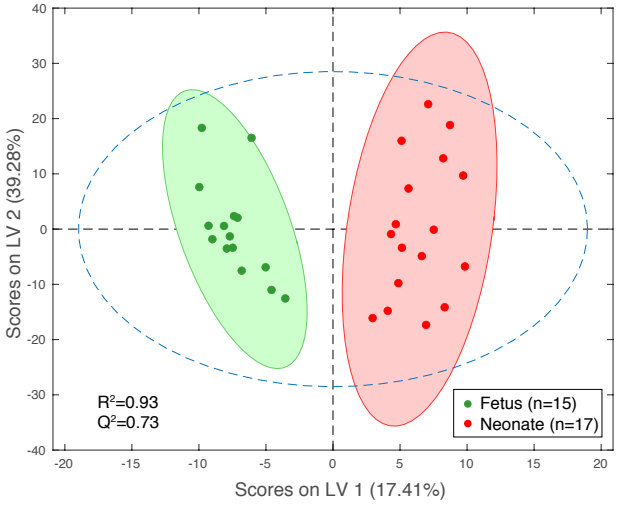


Figure 5.8: OSC-PLS-DA scores plot showing separation between fetal and neonatal serum from all animals.

tion of 15 metabolites increased in neonatal serum. These metabolites include lipid ($(\text{-CH}_2\text{-})_n$), glucose, lipid (-CH_3), lipid (-CH=CH-), N-acetyl-histidine, choline, proline, mannose/mannose-6-phosphate, valine, tyrosine, creatine/creatine-phosphate, leucine, phenylalanine, o-tyrosine, and hydroxyphenyllactate. Seven metabolites, including fructose, myo-inositol, N-acetyl-putrescine, homoarginine, allantoin, 1-methyl-histidine, creatinine, and glycerol, had decreased concentration in neonatal serum. The other 17 metabolites did not show significantly different concentrations in the two time points.

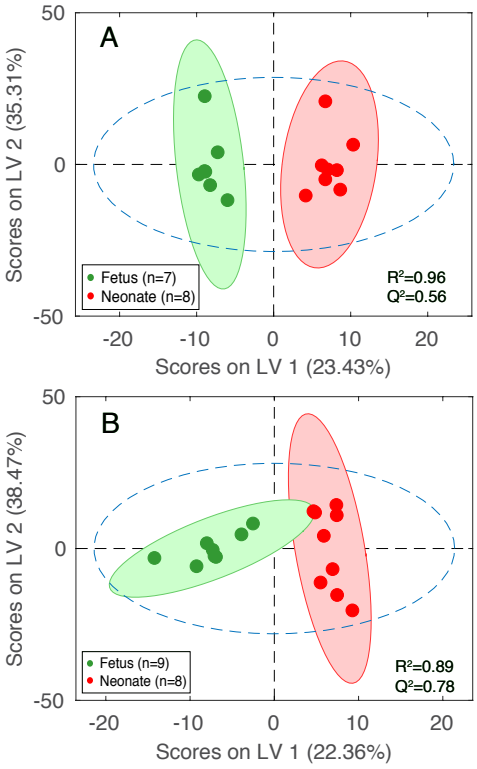


Figure 5.9: OSC-PLS-DA scores plot showing separation between fetal and neonatal serum from CORT (A) and control (B) animals.

Table 5.3: Serum metabolites comparison between fetal and neonatal samples

Annotation	Fetus (mean ± std)	Neonate (mean ± std)	Fold Change (neonate /fetus)	p-value	FDR-corrected p-value	Confidence score*
1-Methyl-histidine	0.10 ± 0.04	0.05 ± 0.03	0.47	1.81E-05	9.80E-04	1
Fructose	1.90 ± 0.85	0.56 ± 0.25	0.30	3.20E-05	1.19E-03	4
Glucose	0.59 ± 0.38	2.91 ± 1.47	4.91	9.00E-05	1.52E-03	3
Homoarginine	0.18 ± 0.05	0.07 ± 0.06	0.42	8.34E-05	1.52E-03	1
Mannose/mannose-6-phosphate	0.06 ± 0.02	0.12 ± 0.05	2.07	2.68E-04	3.19E-03	1
o-Tyrosine	0.04 ± 0.01	0.06 ± 0.02	1.62	3.11E-04	3.50E-03	2
Creatinine	0.55 ± 0.23	0.29 ± 0.32	0.52	9.96E-04	8.30E-03	4
Lipid (-CH_3)	0.29 ± 0.20	1.06 ± 0.66	3.72	1.06E-03	8.37E-03	4
Lipid ((-CH_2-)$_n$)	0.11 ± 0.23	1.53 ± 1.44	14.09	2.21E-03	0.01	2/2
N-acetyl-putrescine	0.58 ± 0.38	0.18 ± 0.13	0.31	2.22E-03	0.01	4

Table 5.3 continued from previous page

Hydroxyphenyllactate	0.03 ± 0.01	0.04 ± 0.01	1.40	3.02E-03	0.02	3
N-acetyl-histidine	0.08 ± 0.04	0.25 ± 0.17	2.98	5.03E-03	0.03	4/3
Allantoin	0.30 ± 0.16	0.14 ± 0.06	0.45	5.43E-03	0.03	4
Creatine/creatine-phosphate	0.57 ± 0.38	1.12 ± 0.51	1.94	0.01	0.05	2
Tyrosine	0.14 ± 0.07	0.29 ± 0.15	2.01	0.01	0.05	3
Lipid (-CH=CH-)	0.05 ± 0.03	0.17 ± 0.16	3.71	0.01	0.06	3
Proline	0.33 ± 0.13	0.80 ± 0.60	2.45	0.02	0.06	2
Valine	1.01 ± 0.41	2.03 ± 1.24	2.01	0.02	0.06	4
Leucine	0.97 ± 0.38	1.84 ± 1.10	1.90	0.02	0.07	4
Myo-inositol	1.14 ± 1.15	0.35 ± 0.15	0.30	0.02	0.07	4
Choline	0.41 ± 0.21	1.09 ± 0.99	2.66	0.03	0.08	4
Phenylalanine	0.13 ± 0.05	0.21 ± 0.11	1.65	0.04	0.10	2
Isoleucine	0.27 ± 0.11	0.49 ± 0.31	1.78	0.04	0.11	2
Citrulline	0.11 ± 0.04	0.17 ± 0.12	1.59	0.07	0.17	3
Lactate	5.51 ± 2.25	8.28 ± 4.06	1.50	0.08	0.18	4
3-Hydroxyisobutyrate	0.19 ± 0.08	0.33 ± 0.27	1.71	0.08	0.18	4
Scyllo-inositol	0.11 ± 0.14	0.05 ± 0.00	0.42	0.09	0.20	3
Betaine	3.67 ± 2.04	5.46 ± 2.48	1.49	0.11	0.22	3
Isovalerate	0.32 ± 0.16	0.25 ± 0.19	0.77	0.11	0.22	4
Glycerol	0.98 ± 0.43	0.71 ± 0.39	0.72	0.14	0.27	3
Glutamate	0.31 ± 0.14	0.50 ± 0.37	1.58	0.15	0.28	4
TSP	1.65 ± 0.86	7.71 ± 16.24	4.67	0.19	0.34	3
Threonine	0.24 ± 0.10	0.34 ± 0.25	1.44	0.20	0.35	3
Lysine	0.28 ± 0.10	0.33 ± 0.15	1.18	0.36	0.49	3
Alanine	0.84 ± 0.36	1.02 ± 0.62	1.21	0.41	0.54	3
Serine	0.52 ± 0.26	0.47 ± 0.31	0.89	0.57	0.68	4
3-Hydroxybutyrate	0.37 ± 0.79	0.48 ± 0.28	1.29	0.63	0.72	2
Glutamine	0.28 ± 0.11	0.25 ± 0.23	0.90	0.69	0.77	4
Glycine	1.84 ± 1.03	1.75 ± 1.19	0.96	0.87	0.90	4

Metabolites are shown in relative concentrations. n=14.

TSP: 3-trimethyl-silyl-[2,2,3,3-$^{2}H_4$]propionic acid.

*Confidence score is defined as follows: 1) putatively characterized compounds or compound classes, 2) 1 D NMR matches to literature and/or database (BMRB and/or HMDB), 3) HSQC or TOCSY matches on COLMARm, 4) HSQC and TOCSY match on COLMARm, and 5) verified by spiking. HSQC: heteronuclear single-quantum coherence, TOCSY: total correlation spectroscopy.

To understand the metabolomic change in a pathway level, an over-representation pathway analysis was performed on significantly changed metabolites. After false discovery rate (FDR)-correction, four pathways were significantly over-represented (Table 5.4). Phenylalanine and tyrosine had an important impact on the phenylalanine, tyrosine and tryptophan biosynthesis pathway, while creatine/creatine-phosphate, proline, and N-acetylputrescine impacted the arginine and proline metabolism.

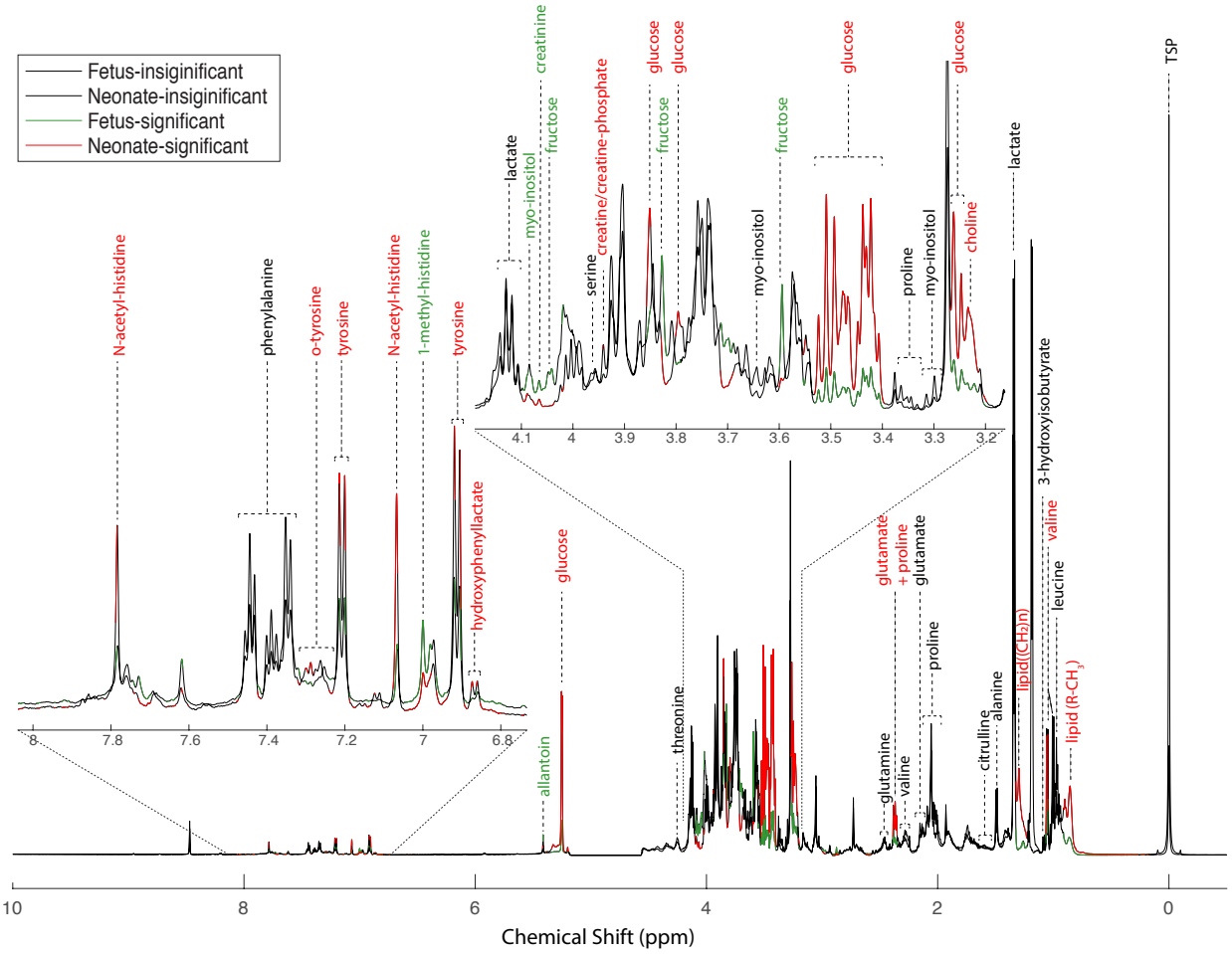


Figure 5.10: Comparison between fetal and neonatal serum spectra from all animals. Mean spectra are presented for visual comparison. Color shows significantly different (p-value <0.01, n=14) regions according to paired t-test. Higher regions for fetuses are colored green and higher regions for neonates are colored red. TSP: 3-trimethyl-silyl-[2,2,3,3-$^{2}H_4$]propionic acid.

5.4 Discussion

Differences in the metabolome between cortisol treated and control two-day-old neonatal hearts were associated with preterm births. This may indicate different adaption capabilities in preterm newborn sheep after increased antenatal cortisol exposure. However, for full term newborns, the maternal cortisol elevation did not have this impact. As discussed in the introduction, external cortisol treatment initiates transition in the fetus to the stage of preparing for parturition and can make the newborn immature at birth [1], [3]. The observation that there were lower levels of valine, glucose, and lipid in fetuses compared

Table 5.4: Significantly changed pathways from fetuses to neonates

	Total	Expected	Hits	Raw p-value	FDR-adjusted p-value	Pathway impact	Changed metabolites*
Aminoacyl-tRNA biosynthesis	48	0.63	5	2.83E-04	0.02	0.00	Phenylalanine, tyrosine, proline, leucine, valine
Phenylalanine, tyrosine and tryptophan biosynthesis	4	0.05	2	9.82E-04	0.03	1.00	Phenylalanine, tyrosine
Arginine and proline metabolism	38	0.50	4	1.23E-03	0.03	0.11	Creatine/creatine-phosphate, proline, N-acetylputrescine
Valine, leucine and isoleucine biosynthesis	8	0.11	2	4.44E-03	0.09	0.00	Leucine, valine
Phenylalanine metabolism	12	0.16	2	1.01E-02	0.17	0.36	Phenylalanine, tyrosine

* All metabolites had higher level in neonatal serum except for N-acetylputrescine

to the levels found in neonates, and in preterm compared to term neonates, demonstrates that the preterm neonates were more "fetal" like and, therefore, less mature than their term peers.

Chronic maternal cortisol treatment changed the heart metabolome of neonates, but the serum metabolome at the same point was not affected. The difference may come from the exclusion of preterm animals from the serum analysis, because all differential metabolites in the neonatal hearts were associated with preterm. Although not all metabolic changes in the CORT group were caused by preterm, the term animals may only represent milder effects compared to the preterm animals.

Significantly changed serum metabolites from four-day-before-birth fetal to two-day old neonatal reflect the metabolic transition from intrauterine to extrauterine conditions. Glucose is the main energy source for fetuses. Amino acids and fatty acids contribute approximately 30% and 5% to the energy supply in sheep fetuses, respectively [18], [2]. Although sheep, as ruminant animals, use fatty acids from microbial fermentation as an important energy source, their placentas have low amounts of fatty acid transporters, so only a small amount of fatty acids is delivered to sheep fetuses [2]. Also, a sheep's placenta can produce fructose from glucose and transport fructose to fetuses to supplement their energy supply [2], [30], [31]. After birth, sheep stop getting nutrition from the placenta and begin receiving nutrition through milk,

which is abundant in fatty acids, low in glucose, and absent of fructose. [32] The significant increase in lipid and choline concentrations and the significant decrease in fructose concentration in the neonatal serum reflect this change of energy source composition. However, serum glucose levels were significantly increased. This may be the result of endogenous synthesis of glucose in sheep neonates. Before birth, fetuses cannot produce glucose so need to receive glucose from their mothers through facilitated diffusion across placenta [2]. Therefore, fetal blood glucose levels are lower than their maternal blood levels. However, immediately after birth, the rate limiting enzyme for gluconeogenesis, phosphoenolpyruvate carboxykinase, starts to be expressed and the newborns can begin using fatty acids to generate glucose [17]. In addition, lactose from the milk is metabolized to glucose. Previous studies have found that during birth, maternal glucose levels increase to ensure an energy supply, thus increasing fetal glucose levels [16]. However, this change is reversed several hours after birth, and fetal glucose levels return back to pre-birth levels. Once the newborn sheep begin feeding, their glucose levels increase to the antenatal maternal levels [16].

Pathway over-representation analysis demonstrated that the phenylalanine, tyrosine and tryptophan biosynthesis pathway is significantly changed after birth. In this pathway, tyrosine is converted from phenylalanine. Interestingly, although not included in the pathway database, o-tyrosine is also a product from phenylalanine, and hydroxyphenyllactate is a metabolite of tyrosine. All these metabolites had higher concentrations in neonatal serum, indicating greater phenylalanine uptake and greater biosynthesis and utilization of tyrosine after birth. Tyrosine is an important metabolite because several neurotransmitters, including epinephrine, norepinephrine, dopamine, and hormones triiodothyronine (T_3), thyroxine (T_4) are derived from it. Therefore, the increased turnover of tyrosine reflects accelerated neuron and endocrine system development after birth.

The other significantly changed pathway, arginine and proline metabolism, has been found to impact a variety of biochemical conditions [33], [34],[35]. However, the significantly changed metabolites from our study, i.e. proline, creatine/creatine-phosphate, and N-acetylputrescine, are not directly connected in this pathway. N-acetylputrescine is a derivative of the polyamine putrescine and is involved in one pathway of the 4-aminobutanoate (GABA) biosynthesis. However, GABA is primarily synthesized by the GABA shunt pathway, which does not include N-acetylputrescine [36]. The increase in proline and creatine/creatine-phosphate and decrease in N-acetylputrescine in neonatal serum compared to fetal serum may reflect a redistribution of amino groups in the arginine and proline metabolism. A deeper investigation in this pathway will expand our understanding of newborn change.

Finally, we observed significant changes in lipid levels after birth. The different fold-change of different lipid chemical groups indicated a mixed composition of lipids. However, it is difficult to distinguish lipids through NMR because of their structural similarity, so we do not have the exact identification of these changed lipids. Since lipid is an important category of metabolites, a lipidomic analysis of these lipids by mass spectrometry is warranted to supplement the profiling of overall changes in the neonatal metabolome.

Bibliography

[1] G. C. Liggins, "The Role of Cortisol in Preparing the Fetus for Birth*," *Review Reprod. Fertil. Dev*, vol. 6, pp. 141–50, 1994.

[2] J. L. Morrison, M. J. Berry, K. J. Botting, *et al.*, "Improving pregnancy outcomes in humans through studies in sheep," *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*, vol. 315, no. 6, R1123–R1153, Dec. 2018, ISSN: 15221490. DOI: 10.1152/AJPREGU.00391.2017/ASSET/IMAGES/LARGE/ZH60091895460007.JPEG. [Online]. Available: https://journals.physiology.org/doi/abs/10.1152/ajpregu.00391.2017.

[3] A. L. Fowden, J. Szemere, P. Hughes, R. S. Gilmour, and A. J. Forhead, "The effects of cortisol on the growth rate of the sheep fetus during late gestation," *Journal of Endocrinology*, vol. 151, no. 1, pp. 97–105, 1996, ISSN: 00220795. DOI: 10.1677/joe.0.1510097.

[4] A. Harris and J. Seckl, "Glucocorticoids, prenatal stress and the programming of disease," *Hormones and Behavior*, vol. 59, no. 3, pp. 279–289, Mar. 2011, ISSN: 0018-506X. DOI: 10.1016/J.YHBEH.2010.06.007.

[5] J. M. Walejko, A. Antolic, J. P. Koelmel, T. J. Garrett, A. S. Edison, and M. Keller-Wood, "Chronic maternal cortisol excess during late gestation leads to metabolic alterations in the newborn heart," *American Journal of Physiology - Endocrinology and Metabolism*, vol. 316, no. 3, E546–E556, Mar. 2019, ISSN: 15221555. DOI: 10.1152/ajpendo.00386.2018. [Online]. Available: /pmc/articles/PMC6459297/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6459297/.

[6] P. W. Nathanielsz, K. A. Berghorn, J. B. Derks, *et al.*, "Life before birth: Effects of cortisol on future cardiovascular and metabolic function," *Acta Paediatrica, International Journal of Paediatrics*, vol. 92, no. 7, pp. 766–772, Jul. 2003, ISSN: 08035253. DOI: 10.1080/08035250310003668.

[7] C. Giurgescu, "Are maternal cortisol levels related to preterm birth?" *JOGNN - Journal of Obstetric, Gynecologic, and Neonatal Nursing*, vol. 38, no. 4, pp. 377–390, 2009, ISSN: 15526909. DOI: 10.1111/j.1552-6909.2009.01034.x.

[8] M. Keller-Wood, X. Feng, C. E. Wood, *et al.*, "Elevated maternal cortisol leads to relative maternal hyperglycemia and increased stillbirth in ovine pregnancy," *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*, vol. 307, no. 4, pp. 405–413, Aug. 2014, ISSN: 15221490. DOI: 10.1152/AJPREGU.00530.2013/ASSET/IMAGES/LARGE/ZH6016148507

0005.JPEG. [Online]. Available: https://journals.physiology.org/doi/abs/10.1152/ajpregu.00530.2013.

[9] L. Bennet, S. Kozuma, H. H. McGarrigle, and M. A. Hanson, "Temporal changes in fetal cardiovascular, behavioural, metabolic and endocrine responses to maternally administered dexamethasone in the late gestation fetal sheep," *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 106, no. 4, pp. 331–339, 1999, ISSN: 14710528. DOI: 10.1111/J.1471-0528.1999.TB08270.X.

[10] S. Joseph, J. M. Walejko, S. Zhang, A. S. Edison, and M. Keller-Wood, "Maternal hypercortisolemia alters placental metabolism: A multiomics view," *American Journal of Physiology - Endocrinology and Metabolism*, vol. 319, no. 5, E950–E960, Nov. 2020, ISSN: 15221555. DOI: 10.1152/AJPENDO.00190.2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7790119/.

[11] A. Harris and J. Seckl, "Glucocorticoids, prenatal stress and the programming of disease," *Hormones and Behavior*, vol. 59, no. 3, pp. 279–289, Mar. 2011, ISSN: 0018-506X. DOI: 10.1016/J.YHBEH.2010.06.007.

[12] P. L. Ballard, D. Torgerson, R. Wadhawan, *et al.*, "Blood metabolomics in infants enrolled in a dose escalation pilot trial of budesonide in surfactant," *Pediatric Research 2021 90:4*, vol. 90, no. 4, pp. 784–794, Jan. 2021, ISSN: 1530-0447. DOI: 10.1038/s41390-020-01343-z. [Online]. Available: https://www.nature.com/articles/s41390-020-01343-z.

[13] G. Chen, Q. Zhang, C. Ai, *et al.*, "Serum metabolic profile characteristics of offspring rats before and after birth caused by prenatal caffeine exposure," *Toxicology*, vol. 427, p. 152 302, Nov. 2019, ISSN: 0300-483X. DOI: 10.1016/J.TOX.2019.152302.

[14] J. M. Walejko, J. P. Koelmel, T. J. Garrett, A. S. Edison, and M. Keller-Wood, "Multiomics approach reveals metabolic changes in the heart at birth," *American Journal of Physiology - Endocrinology and Metabolism*, vol. 315, no. 6, E1212–E1223, Dec. 2018, ISSN: 15221555. DOI: 10.1152/ajpendo.00297.2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6336953.

[15] W. D. Lust, S. Pundik, J. Zechel, Y. Zhou, M. Buczek, and W. R. Selman, "Changing Metabolic and Energy Profiles in Fetal, Neonatal, and Adult Rat Brain," *Metabolic Brain Disease 2003 18:3*, vol. 18, no. 3, pp. 195–206, Sep. 2003, ISSN: 1573-7365. DOI: 10.1023/A:1025503115837. [Online]. Available: https://link.springer.com/article/10.1023/A:1025503115837.

[16] R. S. Comline and M. Silver, "THE COMPOSITION OF FOETAL AND MATERNAL BLOOD DURING PARTURITION IN THE EWE," Tech. Rep., 1972, pp. 233–256.

[17] M. W. Platt and S. Deshpande, "Metabolic adaptation at birth," *Seminars in Fetal and Neonatal Medicine*, vol. 10, no. 4, pp. 341–350, Aug. 2005, ISSN: 1744-165X. DOI: 10.1016/J.SINY.2005.04.001. [Online]. Available: http://www.sfnmjournal.com/article/S1744165X05000181/fulltext.

[18] J. Girard, P. Ferre, J. P. Pegorier, and P. H. Duee, "Adaptations of glucose and fatty acid metabolism during perinatal period and suckling-weaning transition," *Physiological Reviews*, vol. 72, no. 2, pp. 507–562, 1992, ISSN: 00319333. DOI: 10.1152/PHYSREV.1992.72.2.507.

[19] W. W. Green and L. M. Winters, "Prenatal Development of the Sheep," 1945.

[20] H. A. Harris, "The Foetal Growth of the Sheep.," *Journal of anatomy*, vol. 71, no. Pt 4, pp. 516–27, 1937, ISSN: 0021-8782. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17104663%5C%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1252305.

[21] S. A. Reini, G. Dutta, C. E. Wood, and M. Keller-Wood, "Cardiac corticosteroid receptors mediate the enlargement of the ovine fetal heart induced by chronic increases in maternal cortisol," *Journal of Endocrinology*, vol. 198, no. 2, pp. 419–427, Aug. 2008, ISSN: 0022-0795. DOI: 10.1677/JOE-08-0022. [Online]. Available: https://joe.bioscientifica.com/view/journals/joe/198/2/419.xml.

[22] A. Antolic, C. E. Wood, and M. Keller-Wood, "Chronic maternal hypercortisolemia in late gestation alters fetal cardiac function at birth," *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*, vol. 314, no. 3, R342–R352, Mar. 2018, ISSN: 15221490. DOI: 10.1152/AJPREGU.00296.2017/ASSET/IMAGES/LARGE/ZH60121793910005.JPEG. [Online]. Available: https://journals.physiology.org/doi/abs/10.1152/ajpregu.00296.2017.

[23] A. C. Dona, B. Jiménez, H. Schafer, *et al.*, "Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping," *Analytical Chemistry*, vol. 86, no. 19, pp. 9887–9894, Oct. 2014, ISSN: 15206882. DOI: 10.1021/ac5025039. [Online]. Available: http://www.ebi.ac..

[24] A. Le Guennec, F. Tayyari, and A. S. Edison, "Alternatives to Nuclear Overhauser Enhancement Spectroscopy Presat and CarrPurcellMeiboomGill Presat for NMR-Based Metabolomics," *Anal. Chem*, vol. 89, p. 46, 2017. DOI: 10.1021/acs.analchem.7b02354. [Online]. Available: https://pubs.acs.org/sharingguidelines.

[25] J. W. H. Wong, C. Durante, and H. M. Cartwright, "Application of fast fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets," *Analytical Chemistry*, vol. 77, no. 17, pp. 5655–5661, 2005, ISSN: 00032700. DOI: 10.1021/ac050619p.

[26] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn, "Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics.," *Analytical chemistry*, vol. 78, no. 13, pp. 4281–90, Jul. 2006, ISSN: 0003-2700. DOI: 10.1021/ac051632c. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16808434%20https://pubs.acs.org/doi/10.1021/ac051632c.

[27] S. A. Sousa, A. Magalhães, and M. M. C. Ferreira, "Optimized bucketing for NMR spectra: Three case studies," *Chemometrics and Intelligent Laboratory Systems*, vol. 122, pp. 93–102, Mar. 2013, ISSN: 01697439. DOI: 10.1016/j.chemolab.2013.01.006.

[28] N.-P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping," *Journal of Chromatography A*, vol. 805, no. 1-2, pp. 17–35, May 1998, ISSN: 0021-9673. DOI: 10.1016/S0021-9673(98)00021-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021967398000211.

[29] Z. Pang, J. Chong, G. Zhou, *et al.*, "MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights," *Nucleic Acids Research*, vol. 49, no. W1, W388–W396, Jul. 2021, ISSN: 0305-1048. DOI: 10.1093/NAR/GKAB382. [Online]. Available: https://academic.oup.com/nar/article/49/W1/W388/6279832.

[30] H. K. Meznarich, W. W. Hay, J. W. Sparks, G. Meschia, and F. C. Battaglia, "FRUCTOSE DISPOSAL AND OXIDATION RATES IN THE OVINE FETUS," *Quarterly Journal of Experimental Physiology*, vol. 72, no. 4, pp. 617–625, Oct. 1987, ISSN: 1469-445X. DOI: 10.1113/EXPPHYSIOL.1987.SP003102. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1113/expphysiol.1987.sp003102%20https://onlinelibrary.wiley.com/doi/abs/10.1113/expphysiol.1987.sp003102%20https://physoc.onlinelibrary.wiley.com/doi/10.1113/expphysiol.1987.sp003102.

[31] C. E. Trindade, R. C. Barreiros, C. Kurokawa, and G. Bossolan, "Fructose in fetal cord blood and its relationship with maternal and 48-hour-newborn blood concentrations," *Early Human Development*, vol. 87, no. 3, pp. 193–197, Mar. 2011, ISSN: 0378-3782. DOI: 10.1016/J.EARLHUMDEV.2010.12.005.

[32] W. L. Wendorff and G. F. Haenlein, "Sheep milk: Sheep milk - composition and nutrition," *Handbook of Milk of Non-Bovine Mammals: Second Edition*, pp. 210–221, 2017. DOI: 10.1002/9781119110316.ch3.2.

[33] F. Patin, P. Corcia, P. Vourc'h, *et al.*, "Omics to Explore Amyotrophic Lateral Sclerosis Evolution: the Central Role of Arginine and Proline Metabolism," *Molecular Neurobiology*, vol. 54, no. 7, pp. 5361–5374, Sep. 2017, ISSN: 1559182. DOI: 10.1007/S12035-016-0078-X/FIGURES/7. [Online]. Available: https://link.springer.com/article/10.1007/s12035-016-0078-x.

[34] K. D. Quinn, M. Schedel, Y. Nkrumah-Elie, *et al.*, "Dysregulation of metabolic pathways in a mouse model of allergic asthma," *Allergy*, vol. 72, no. 9, pp. 1327–1337, Sep. 2017, ISSN: 1398-9995. DOI: 10.1111/ALL.13144. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/all.13144%20https://onlinelibrary.wiley.com/doi/abs/10.1111/all.13144%20https://onlinelibrary.wiley.com/doi/10.1111/all.13144.

[35] M. F. Raza, Y. Wang, Z. Cai, *et al.*, "Gut microbiota promotes host resistance to low-temperature stress by stimulating its arginine and proline metabolism pathway in adult Bactrocera dorsalis," *PLoS Pathogens*, vol. 16, no. 4, Apr. 2020, ISSN: 15537374. DOI: 10.1371/JOURNAL.PPAT.1008441. [Online]. Available: /pmc/articles/PMC7185725/%20/pmc/articles/PMC7185725/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7185725/.

[36] D. Rashmi, R. Zanan, S. John, K. Khandagale, and A. Nadaf, "γ-Aminobutyric Acid (GABA): Biosynthesis, Role, Commercial Production, and Applications," *Studies in Natural Products Chemistry*, vol. 57, pp. 413–452, Jan. 2018, ISSN: 1572-5995. DOI: 10.1016/B978-0-444-64057-4.00013-2.

Chapter 6
CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Conclusions

In this dissertation, I demonstrated two novel computational tools I developed for NMR-based untargeted metabolomics, including a data integration technique and a spectral processing algorithm. By applying these tools to the study of *C. elegans* development, and pregnancy under normal and challenged conditions, I discovered new biological phenomena. I identified development-associated metabolite-glycan correlations in *C. elegans*. I found disturbed metabolites in pregnant mothers with Zika virus infection. Using a sheep model, I identified preterm-birth-associated metabolic change with maternal chronic cortisol treatment and altered metabolites and pathways after birth in neonates.

6.1.1 Computational Approach for Integrating NMR-measured Metabolomics, LC-MS/MS-detected Glycomics and Two-dimensional Flow-cytometric Data in Studying *C. elegans* Development

I developed an original algorithm for integrating worm sorter data with NMR-measured metabolomics data and MS-measured glycomics data, which are in two different data structures. By binning and vectorizing the two-dimensional worm body size data, this work calculated correlations between the body size data with one-dimensional metabolomics and glycomics data. This work was the first to integrate flow-cytometric-like data with metabolomics data. As discussed in chapter 2, synchronizing *C. elegans* growth to generate a developmental-stage-homogenous sample is hard to achieve experimentally, thus compromising quantitative comparisons between time points. My work utilized the variation between samples to map metabolites and glycans to specific worm sizes, thereby making it possible to find developmentally related metabolites with semi-synchronized worm samples. Though this algorithm was developed for metabolomics data integration with worm size data, it can be adapted to other omics data with similar data structures. This technique can also be extended to flow-cytometry-based analysis on cell line samples for unravelling associations between metabolites and subtypes.

My work also investigated correlations between analytes from different analytical platforms without feature selection by significance. Therefore, we were able to make maximum use of the information. We were able to find direct correlations between metabolites and glycans, even the metabolites (e.g. phosphorylcholine) were not directly correlated with specific worm sizes. With this approach, we discovered potential glycan substrates of phosphorylcholine modification, as well as possible *O*-glycosylation transition during *C. elegans* development. Therefore, this approach and algorithm can be extended to inter-platform integration analysis to facilitate researchers systematically investigating biology.

6.1.2 Development of a Novel Algorithm for Improving NMR Spectral Alignment by Spectral Reordering and Curve Tracing

The NMR spectral alignment algorithm (pHIT) demonstrated in chapter 3 is a novel tool I developed for processing NMR peaks with high chemical shift variation. It utilized the source of chemical shift variation to match peaks from the same nucleus on different spectra. This approach, as an orthogonal strategy of minimizing variations, was able to align those peaks that the existing minimizing-strategy-based methods could not correct, therefore can be used in combination with those methods for improving NMR data quality.

Compared to the other two reorder-align based algorithms discussed in chapter 3 [1], [2], [3], the pHIT algorithm does not rely on curve shape similarity to trace a responding curve, thus is more flexible and widely applicable, while it requires less manual input for proper tracing, so it is more efficient. Since I developed this tool because no publicly available tools had been able to align my ZIKV urine dataset well, this pHIT alignment algorithm should have great potential to aid in urine-based NMR metabolomics analysis.

6.1.3 Profiling Longitudinal Metabolome of ZIKV-infected Pregnant Women

The metabolomics study of ZIKV infection on pregnant women reported here is the first research, to my knowledge, to investigate the urine metabolome of ZIKV infected human patients as well as the first to directly assess metabolome of pregnant women infected with ZIKV.

By profiling the longitudinal urine metabolome of ZIKV-infected and uninfected pregnant women, I identified maternal metabolites associated with ZIKV infection. These metabolites, including 3-aminoisobutyrate, fucose, 2-hydroxyglutarate, N-acetyl-glutamine, dimethylamine, 4-hydroxyphenethyl alcohol, creatinine, lactate, threonine, histidine, pseudouridine, trigonelline, 1-MNA, and glucose, suggested disturbed tryptophan, NAD+, pyrimidine, and glucose metabolic pathways with ZIKV infection. Targeted research on these metabolites and pathways may be conducted to verify our findings and extend understandings of mechanisms pregnant mothers and fetuses use to respond to ZIKV.

6.1.4 Metabolic Alterations After Birth and With Chronic Maternal Cortisol Excess During Late Gestation in Sheep Neonates

Our original metabolomics study of comparing the same sheep before and after birth discovered significant alterations in neonate metabolomic profiles after birth. These changes were not associated with prenatal exposure to excess chronic maternal cortisol. The significant changes in energy supplying metabolites such as glucose, fructose, and lipids reflected the change in nutrition sources after birth. The sheep neonates stopped receiving fructose from maternal blood and started to receive fatty acids from milk after birth. The elevation of glucose level may indicate gluconeogenesis in neonates. Over-representation pathway analysis demonstrated changes in phenylalanine, tyrosine and tryptophan biosynthesis pathways, especially in tyrosine activity after birth, suggesting the importance of tyrosine in the few days after birth.

In this work, I used paired analysis to control the inter-individual variability so that result interpretation can be straightforward. Sample collection on sheep fetuses can be done in a non-destructive manner [4]. However, although the sheep model has been used for pregnancy-related metabolomics, no one has used this system to compare the metabolome on the same samples for investigating neonatal adaptation after birth. Therefore, our work for the first time systematically described sheep metabolic transitions from fetuses to neonates and proved the possibility for further longitudinally profiling fetal-neonatal metabolic change at omics level.

Our study also discovered differences in the cardiac metabolome between chronically maternally cortisol treated and control neonatal sheep. However, we found this difference associated with preterm labor. We found metabolites such as glucose, valine, and lipid had lower levels in the preterm animals when compared with term animals, as well as in the fetal serum compared to neonatal serum. The similarity suggested preterm animals were less mature to adapt to ex utero conditions.

Metabolomics on fetal serum, newborn hearts and placentas (samples collected the same day after birth) with the same chronic excess cortisol gestational model have been conducted previously [5], [6]. Differences have been observed between cortisol treated and control animals. These results together with my results described here suggest the impact of maternal cortisol on fetal metabolome may not be directly extended to neonatal metabolome, but it may appear through preterm birth.

6.2 Future Directions

The pHIT algorithm currently moves the whole peak outside of the spectra after alignment. However, curves of two close responding signals have been observed to merge or cross each other on experimental spectra. Therefore, if one signal is moved, the current method would impair the quantification of the remaining signal. Deconvolution is then needed for such cases. Yue Wu from our lab is developing a spectra deconvolution method based on fitting on time-domain NMR spectra (unpublished). We are working together to combine these tools for processing more complex NMR spectra. We also plan to simulate spectra with larger number and with different levels of peak overlap to model complex data.

The ZIKV metabolomics work can be considered as a pilot study for potential diagnostic studies. Due to the complexity of clinical data as I discussed in chapter 4, I suggest future studies to increase the sample size for robust statistics. Also, pregnancy is a complex process, maternal changes may not fully reflect fetal changes. Our study did not assess fetal health conditions, so it would be nice to analyze the associations between ZIKV-induced maternal metabolic changes with fetal conditions. In addition, most metabolites that changed significantly between ZIKV+ and ZIKV- group had consistently higher or lower values over time, suggesting a long-term effect of ZIKV. Future studies may take this into consideration for study design.

Finally, as I discussed in chapter 5, lipid levels showed significant differences between the fetal and neonatal sheep serum samples. It would be valuable to determine the composition of these lipids. I have sent the samples to our lipidomics expert collaborators for detailed profiling of lipid change after birth. Beyond gaining a more comprehensive coverage of changed compounds, with the data integration approach I demonstrated in chapter 2, we will be able to find correlated metabolites and lipids from the dataset, therefore expanding our knowledge on metabolic transition after birth.

Bibliography

[1] L. Csenki, E. Alm, R. J. Torgrip, *et al.*, "Proof of principle of a generalized fuzzy Hough transform approach to peak alignment of one-dimensional 1H NMR data," *Analytical and Bioanalytical Chemistry*, vol. 389, no. 3, pp. 875–885, 2007, ISSN: 16182650. DOI: 10.1007/s00216-007-1475-9.

[2] E. Alm, R. J. Torgrip, K. M. Åberg, I. Schuppe-Koistinen, and J. Lindberg, "A solution to the 1D NMR alignment problem using an extended generalized fuzzy Hough transform and mode support," *Analytical and Bioanalytical Chemistry*, vol. 395, no. 1, pp. 213–223, 2009, ISSN: 16182650. DOI: 10.1007/s00216-009-2940-4.

[3] M. Liebeke, J. Hao, T. M. D. Ebbels, and J. G. Bundy, "Combining spectral ordering with peak fitting for one-dimensional NMR quantitative metabolomics," *Analytical Chemistry*, vol. 85, no. 9, pp. 4605–4612, 2013, ISSN: 00032700. DOI: 10.1021/ac400237w.

[4] J. L. Morrison, M. J. Berry, K. J. Botting, *et al.*, "Improving pregnancy outcomes in humans through studies in sheep," *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*, vol. 315, no. 6, R1123–R1153, Dec. 2018, ISSN: 15221490. DOI: 10.1152/AJPREGU.00391.2017/ASSET/IMAGES/LARGE/ZH60091895460007.JPEG. [Online]. Available: https://journals.physiology.org/doi/abs/10.1152/ajpregu.00391.2017.

[5] J. M. Walejko, A. Antolic, J. P. Koelmel, T. J. Garrett, A. S. Edison, and M. Keller-Wood, "Chronic maternal cortisol excess during late gestation leads to metabolic alterations in the newborn heart," *American Journal of Physiology - Endocrinology and Metabolism*, vol. 316, no. 3, E546–E556, Mar. 2019, ISSN: 15221555. DOI: 10.1152/ajpendo.00386.2018. [Online]. Available: /pmc/articles/PMC6459297/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6459297/.

[6] S. Joseph, J. M. Walejko, S. Zhang, A. S. Edison, and M. Keller-Wood, "Maternal hypercortisolemia alters placental metabolism: A multiomics view," *American Journal of Physiology - Endocrinology and Metabolism*, vol. 319, no. 5, E950–E960, Nov. 2020, ISSN: 15221555. DOI: 10.1152/AJPENDO.00190.2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7790119/.

APPENDIX A
SUPPLEMENTARY FILES

A.1 Supplementary Files for Chapter 2

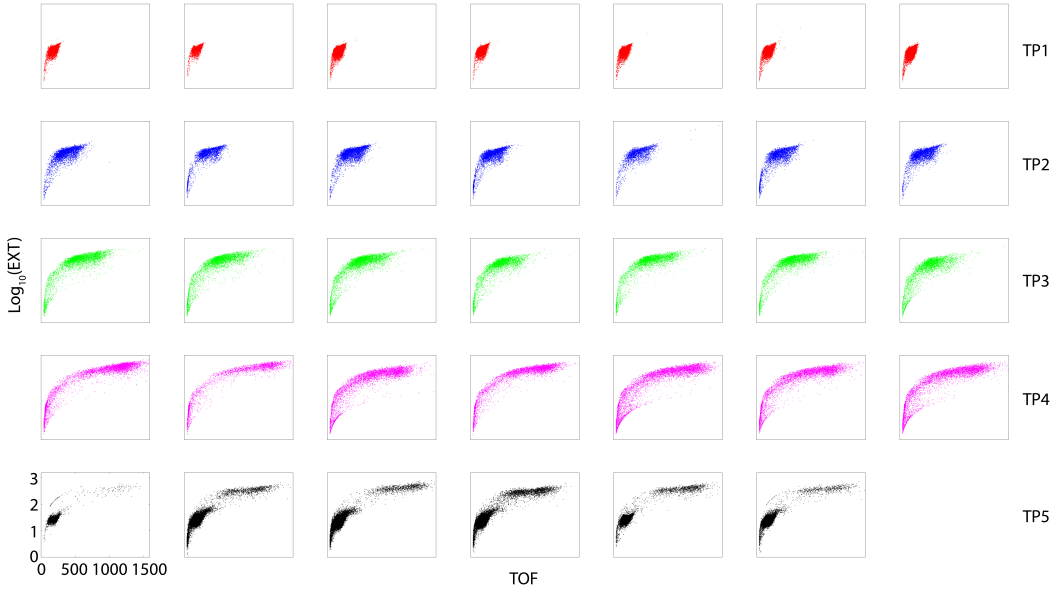


Figure A.1: Individual *C.elegans* distribution plots for each sample. Each subplot shows raw reads from the Union Biometrica Biosorter for each sample. The biological replicates are the columns. Different colors indicate samples collected from different time points (red: TP1, blue: TP2, green: TP3, magenta: TP4, black: TP5). One replicate at TP5 was not biosorted.

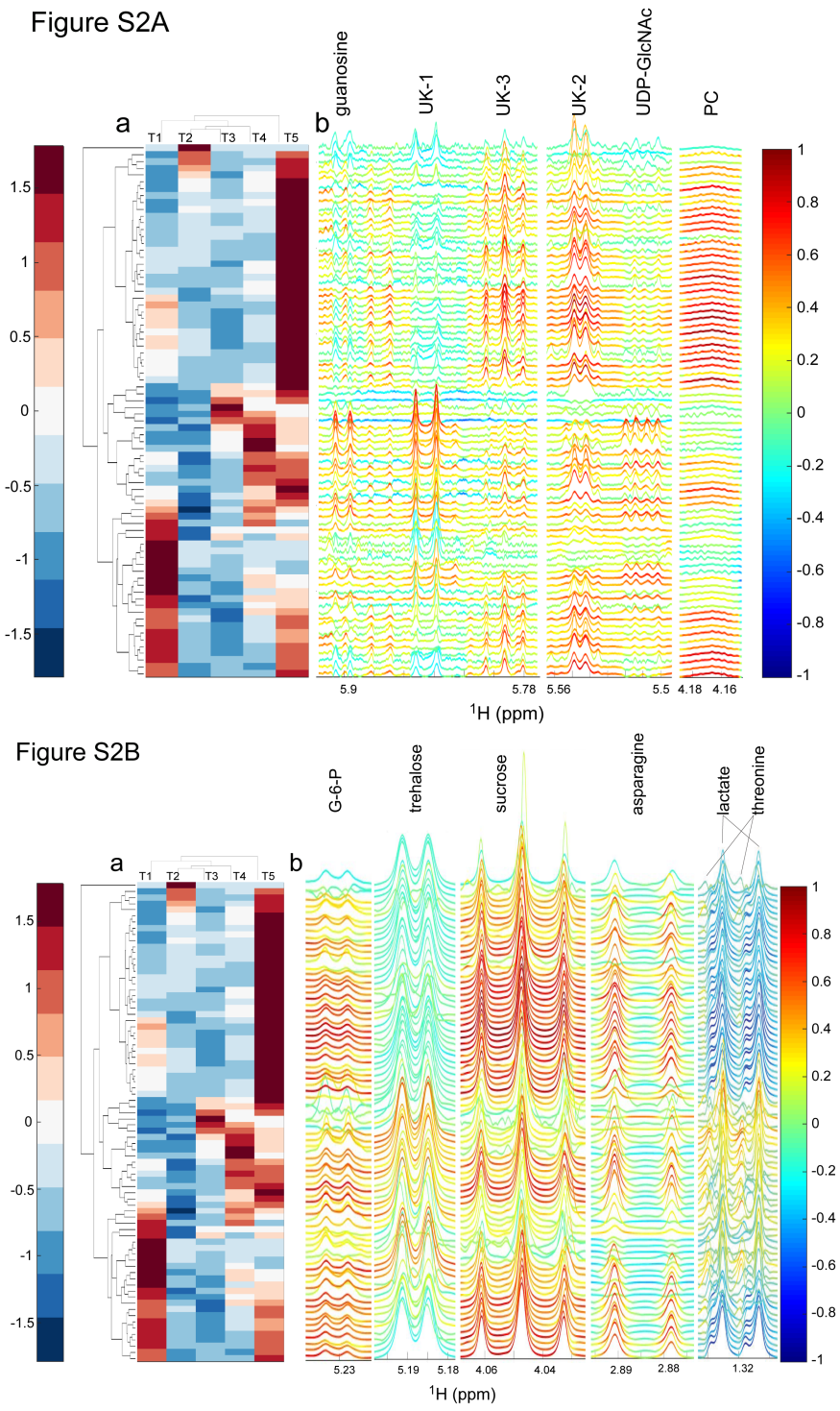


Figure A.2: Correlations between NMR-measured metabolites and LC-MS-measured glycans. (a) Heatmap of glycan abundances and dendrogram of glycans (rows) and sample time points (columns) (same heatmap as used in Figure 2.5). Glycan abundances were averaged over replicates. A color bar of the heatmap is shown on the left. (b) Regions of NMR STOCSYs on glycans. A color bar of the correlation coefficients is shown on the right.

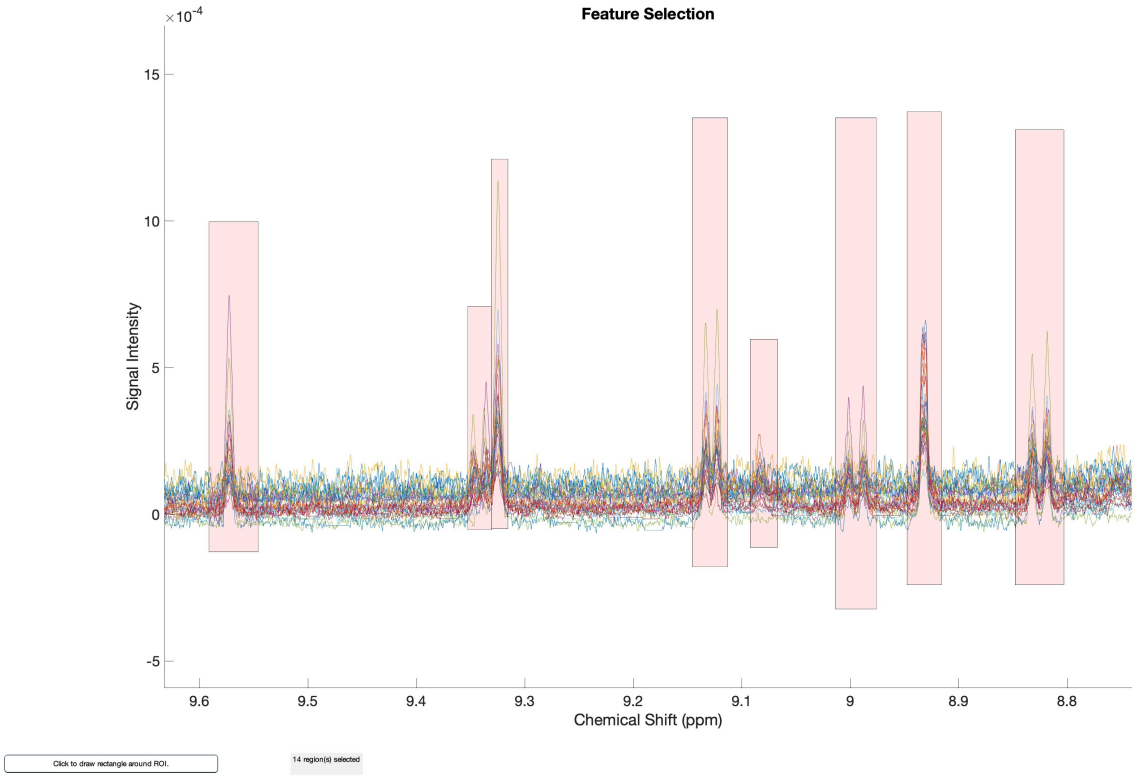


Figure A.3: NMR interactive binning example. The interactive binning algorithm described in 2.2.6 allows for user-defined regions with variable width and range to be specified for binning. This method was used to extract features for Figure 2.6B. The boxes below the figure are part of the interactive GUI. This expansion shows 8 of the 14 total regions selected from this particular session. The complete workflow is available through the Edison lab GitHub site (`https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA`).

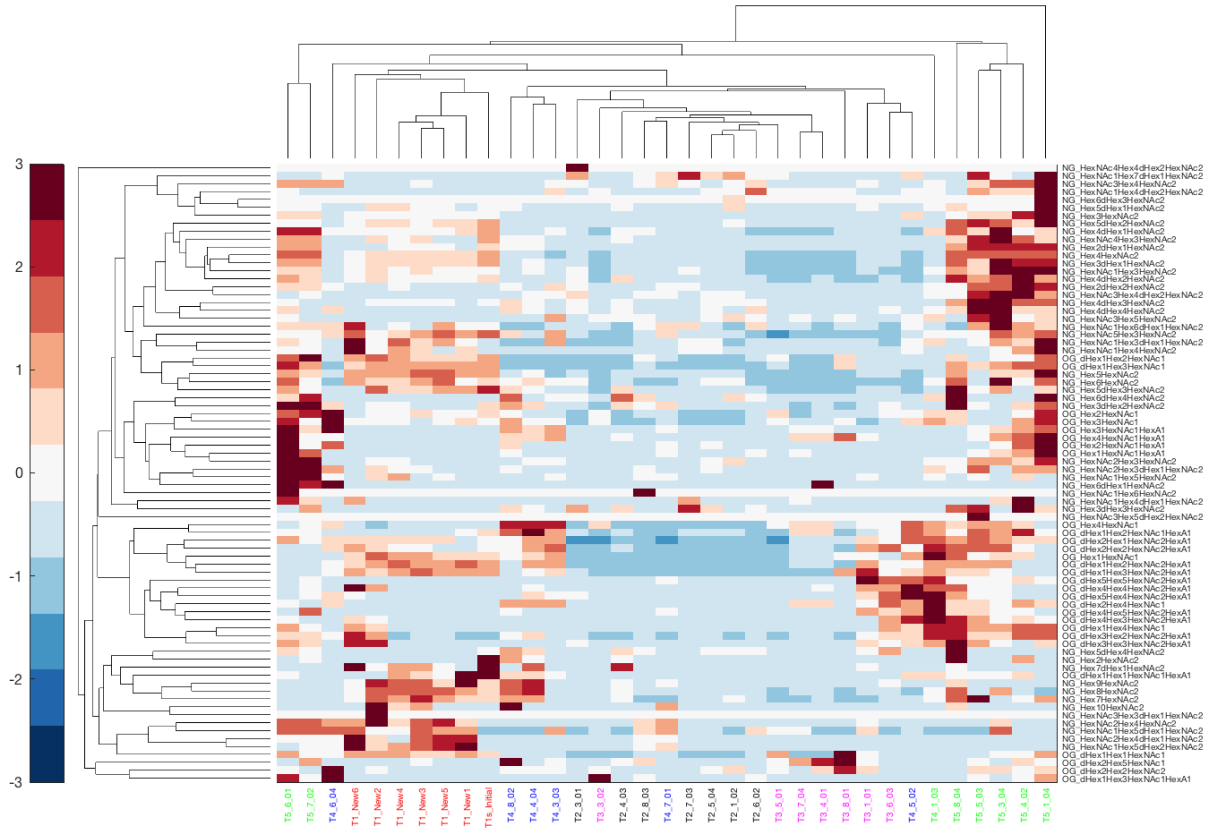


Figure A.4: Developmental patterns of glycans in *C. elegans*. The heatmap shows glycan abundances, together with dendrogram of glycans (rows) and sample timepoints (columns). This is similar to Figure 2.5 in the main text, but each column is a biological replicate rather than the time average. A color bar of the heatmap is shown on the left.

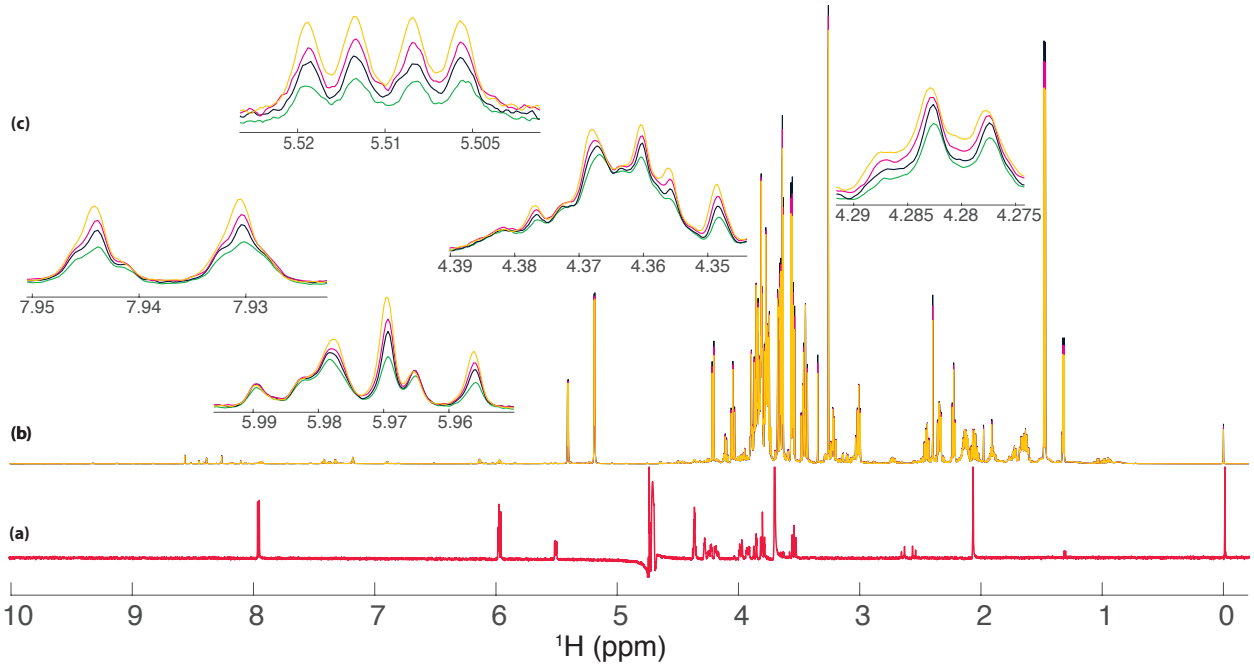


Figure A.5: Spiking of UDP-GlcNAc. Spectrum of synthetic UDP-GlcNAc (a; red); overlay of all 29 experimental worm samples at different time points (b), and different regions after one of the T4 samples was spiked with the UDP-GlcNAc standard solution (green: no spiking; dark blue: first spike; magenta: second spike; yellow: third spike.)

A.2 Supplementary Files for Chapter 4

Table A.1: Characteristics of participants at recruitment

Characteristic	ZIKV+ participants (n=10)	ZIKV- participants (n=10)
Race		
Mestiza	6	4
White	1	1
Hispanic/latina	3	5
Age, years (mean ± std)	22.3 ± 5.3	23.4 ± 6.7
Gestational age, weeks (mean ± std)	9.2 ± 2.3	8.9 ± 2.6
Medical History[a]		
Anemia	0	1

Respiratory disease	0	1
Others	0	2[b]
Prior infections (Nonnegative/total)		
Rubell_IgG_entry	8/8	9/9
Rubell_IgM_entry	0/5	0/9
Rubell_IgG_T3	5/6	10/10
Rubell_IgM_T3	0/6	0/10
Cyto_IgG_entry	5/8	3/9
Cyto_IgM_entry	0/5	0/9
Cyto_IgG_T3	4/5	4/10
Cyto_IgM_T3	0/6	2/10
Herpes1_IgG_entry	5/8	5/9
Herpes1_IgM_entry	1/5	6/9
Herpes1_IgG_T3	4/6	6/10
Herpes1_IgM_T3	3/6	6/10
Herpes2_IgG_entry	1/8	0/9
Herpes2_IgM_entry	1/5	6/9
Herpes2_IgG_T3	0/6	0/10
Herpes2_IgM_T3	4/6	6/9
Toxoplas_IgG_entry	3/8	4/8
Toxoplas_IgM_entry	4/5	4/9
Toxoplas_IgG_T3	4/6	4/10
Toxoplas_IgM_T3	1/6	4/10

T3: third trimester. Cyto: cytomegalovirus.

[a]Zero case for malaria, tuberculosis, cancer, diabetes, parasites, cardiovascular disease history.

[b]One tiroides, one asma.

Table A.2: Identified metabolites from pooled internal control urine sample.

No.	Annotation	Right boundary (ppm)	Left boundary (ppm)	Confidence Score[a]
1	-S-3-Hydroxyisobutyric acid	1.052	1.073	3
2	1-Methylnicotinamide	4.469	4.478	3
2	1-Methylnicotinamide	8.876	8.903	3
2	1-Methylnicotinamide	8.948	8.971	3
2	1-Methylnicotinamide	9.257	9.280	3
3	2-Hydroxyglutaric acid	2.010	2.014	3

4	2-Oxoglutarate	2.421	2.426	4
4	2-Oxoglutarate	2.432	2.438	4
4	2-Oxoglutarate	2.995	3.004	4
5	3-Aminoisobutyric acid	1.191	1.193	4
5	3-Aminoisobutyric acid	2.571	2.663	4
5	3-Aminoisobutyric acid	3.082	3.088	4
6	3-Hydroxypropionic acid	2.421	2.439	3
7	3-Methylhistidine	3.215	3.222	3
7	3-Methylhistidine	3.702	3.710	3
7	3-Methylhistidine	7.029	7.041	3
8	3-Methylxanthine	8.011	8.024	3
9	4-Hydroxyphenethyl alcohol	2.770	2.778	3
10	4-Hydroxyphenylacetic acid	6.846	6.867	4
11	Acetic acid	1.912	1.917	3
12	Acetone	2.224	2.231	3
13	Adipic acid/Pimelic acid/Suberic acid	1.526	1.539	4
14	Alanine	1.463	1.489	4
15	Allantoin	5.379	5.389	3
16	Alpha-hydroxyisobutyric acid	1.348	1.354	3
17	Betaine	3.255	3.261	3
18	Choline	3.192	3.197	4
19	Cis-aconitate	5.660	5.786	2
20	Citrate	2.508	2.554	3
21	Creatine/Creatine phosphate	3.021	3.031	4
21	Creatine/Creatine phosphate	3.918	3.928	4
22	Creatinine	3.033	3.045	3
22	Creatinine	4.038	4.057	3
23	D-Glucose	3.477	3.483	4
23	D-Glucose	3.487	3.498	4
23	D-Glucose	3.881	3.891	4
23	D-Glucose	5.228	5.242	4
24	D-Mannitol[b]			4
25	Dimethylamine	2.708	2.721	3
26	Ethanol	1.164	1.168	3
27	Ethanolamine	3.132	3.138	4
28	Formate	8.446	8.456	3
29	Glycine	3.554	3.567	3
30	Glycolate	3.941	3.947	3
31	Guanidineacetic acid	3.786	3.798	3

32	Hippuric acid	3.953	3.970	4
32	Hippuric acid	7.526	7.563	4
32	Hippuric acid	7.613	7.648	4
32	Hippuric acid	7.812	7.839	4
33	Histamine	8.083	8.096	2
34	Histidine	7.113	7.124	2
34	Histidine	7.966	7.984	2
35	Indoxyl sulfate	7.487	7.51	3
36	L-Arabinose	4.522	4.526	3
37	L-Arginine	1.616	1.637	3
37	L-Arginine	1.660	1.668	3
38	L-Asparagine	2.926	2.945	4
39	L-Fucose	1.233	1.238	3
39	L-Fucose	1.245	1.248	3
39	L-Fucose	4.561	4.566	3
40	L-Glutamine	2.428	2.433	4
41	L-Glutamine	2.453	2.458	4
40	L-Glutamine	2.466	2.476	4
41	L-Phenylalanine	7.323	7.332	3
42	L-Tartaric acid	4.330	4.334	3
43	L-Threitol[b]			3
44	L-Threonine	3.581	3.6	4
44	L-Threonine	4.233	4.279	4
45	L-Tryptophan	7.703	7.729	3
45	L-Tryptophan	7.515	7.526	3
45	L-Tryptophan	7.311	7.322	3
46	L-Valine	0.975	0.980	3
46	L-Valine	1.025	1.032	3
47	Lactic acid	4.102	4.125	4
48	Lactose	4.439	4.452	3
48	Lactose	4.457	4.464	3
48	Lactose	4.664	4.670	3
49	Leucine	0.940	0.946	3
49	Leucine	0.950	0.956	3
50	Lysine	1.693	1.743	4
51	Methylguanidine	2.821	2.826	3
52	N-Acetyl-D-glucosamine/N-acetylgalactosamine /N-Acetyl-D-glucosamine-6-phosphate	5.197	5.211	3
53	N-Acetyl-L-glutamine	2.079	2.114	4

53	N-Acetyl-L-glutamine	2.308	2.323	4
53	N-Acetyl-L-glutamine	4.143	4.198	4
54	N,N-Dimethylglycine	2.919	2.923	3
55	Phenylacetylglycine	7.333	7.344	4
55	Phenylacetylglycine	7.397	7.435	4
56	Pimelic acid/1,11-Undecanedicarboxylic acid /Undecanedioic acid/Suberic acid	1.289	1.304	4
57	Pseudouridine	7.666	7.678	3
58	Quinolinic acid	7.460	7.484	3
59	Scyllo-Inositol	3.347	3.354	3
60	Serine	3.842	3.845	4
61	Succinic acid	2.392	2.407	3
62	Taurine	3.408	3.412	4
63	Threo-isocitric acid	2.972	2.983	4
64	Trigonelline	4.430	4.436	3
64	Trigonelline	8.073	8.084	3
64	Trigonelline	8.809	8.848	3
64	Trigonelline	9.104	9.123	3
65	Trimethylamine	2.870	2.876	3
66	Trimethylamine-N-oxide	3.262	3.270	3
67	Tyrosine	6.880	6.902	4
68	Uracil	5.786	5.807	2
69	Urea	5.743	5.831	1

[a]Confidence scale is defined as follows: 1) putatively characterized compounds or compound classes, 2) 1D NMR matches to literature and/or database (BMRB and/or HMDB), 3) HSQC matches on COLMARm or AssureNMR, 4) HSQC and HSQC-TOCSY match on COLMARm, 5) verified by spiking

[b]No clear one-dimensional peaks.

Table A.3: Significantly different features without specific annotations or high quantification confidence scores between ZIKV+ and ZIKV- individuals.

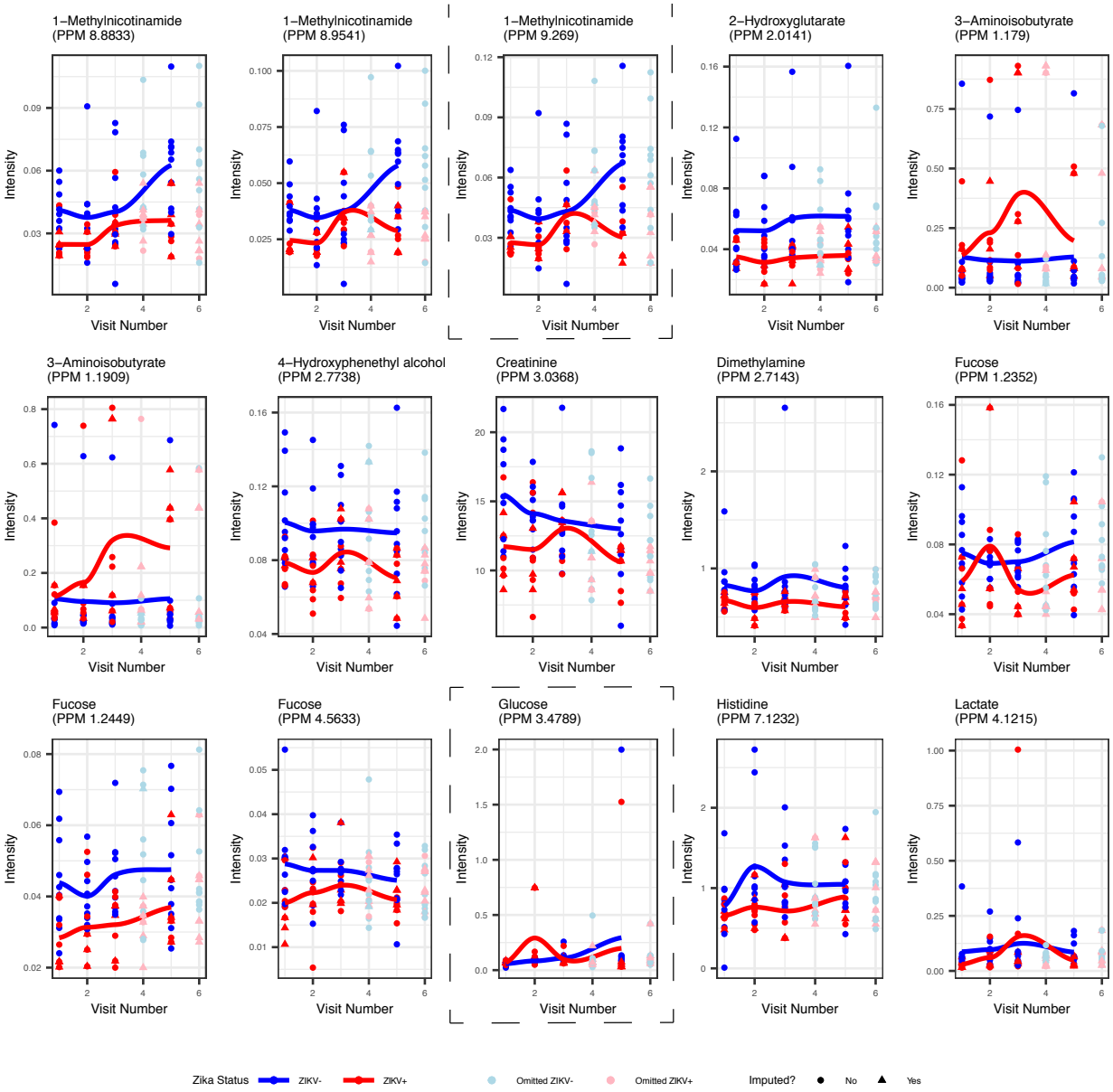
Chemical Shift (ppm)	FDR adjusted p-value	Annotation	Quantification confidence score[a]
1.218	5.25E-02	UK1	3
1.229	6.08E-02	UK2	3
1.657	9.15E-02	Arginine (overlap)	3
1.934	1.21E-01	N-acetyl-glutamine (overlap)	3

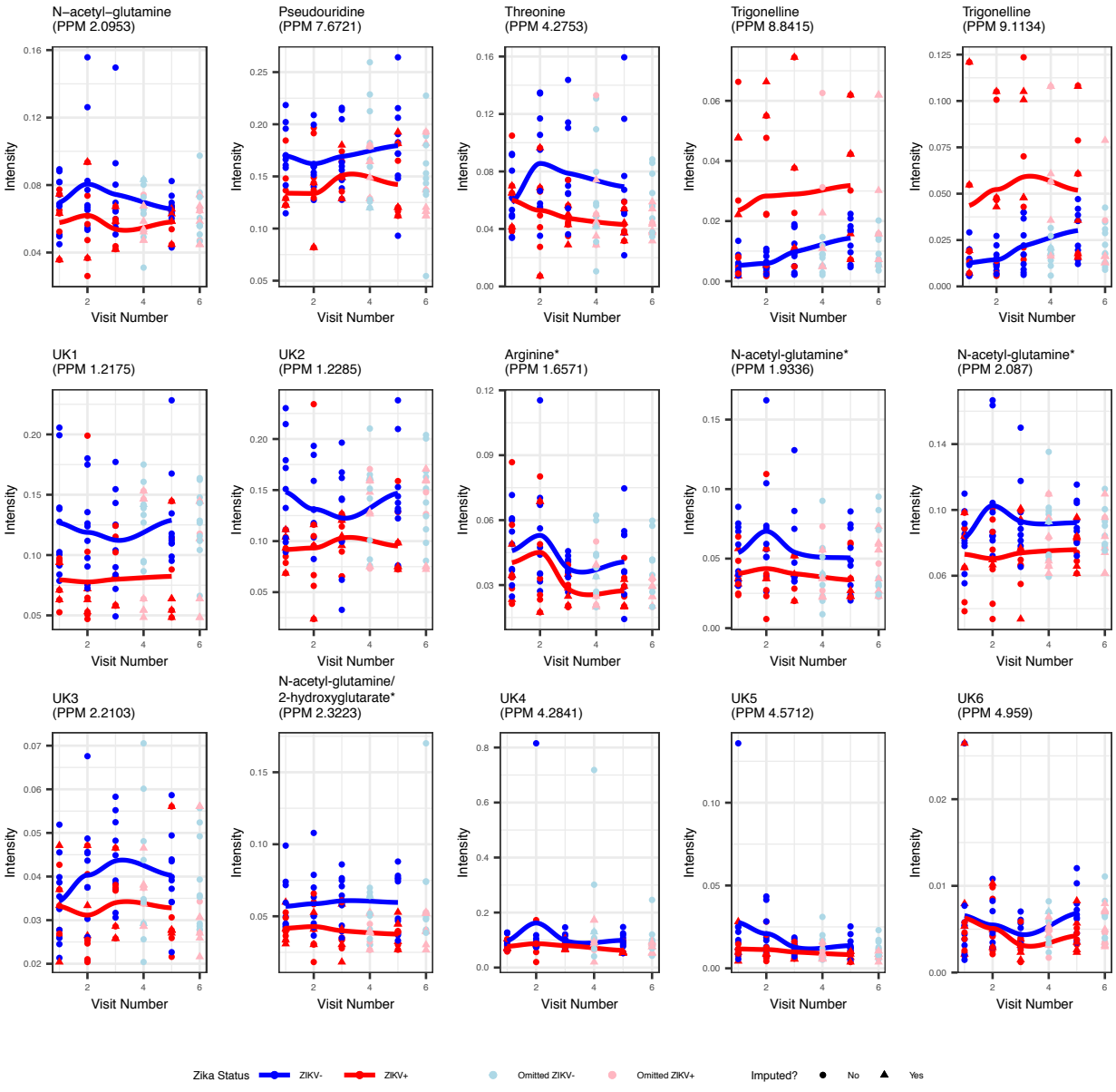
2.087	1.04E-01	N-acetyl-glutamine (overlap)	3
2.210	1.38E-01	UK3	3
2.322	5.25E-02	N-acetyl-glutamine/2-hydroxyglutarate	3
4.284	9.34E-02	UK4	3
4.571	9.15E-02	UK5	3
4.959	1.38E-01	UK6	3
4.984	1.25E-01	UK7	3
5.200	1.06E-01	Fucose/GlcNAc/GlcNAc-6-P/GalNAc	3
5.209	9.84E-02	Fucose/GlcNAc/GlcNAc-6-P/GalNAc	3
5.839	1.38E-01	UK8	3
7.003	9.15E-02	UK9	3
7.190	1.25E-01	Tyrosine+4-hydroxyphenethyl alcohol	3
7.311	1.21E-01	Tryptophan (overlap)	3
8.774	1.06E-01	UK10	3
8.928	1.21E-01	UK11	3
0.948	1.32E-01	UK12	2
1.910	1.21E-01	Lysine (overlap)	2
2.479	1.46E-01	UK13	2
7.251	1.38E-01	Tryptophan/indoxyl sulfate	2
1.165	1.22E-01	Ethanol	1
1.338	1.46E-01	UK14	1
3.453	1.06E-01	Glucose (overlap)	1
7.315	9.34E-02	Tryptophan	1
7.317	1.25E-01	Tryptophan	1
7.327	1.21E-01	Phenylalanine	1
7.430	1.29E-01	Phenylacetylglycine (overlap)	1
7.453	1.46E-01	Quinolinate (overlap)	1
1.327	1.06E-01	UK15	0
1.502	1.46E-01	Lysine (overlap)	0
1.629	5.25E-02	Arginine	0
4.561	1.25E-01	UK16	0
5.836	1.21E-01	UK17	0
6.099	9.34E-02	UK18	0
7.294	9.15E-02	Tryptophan (overlap)	0

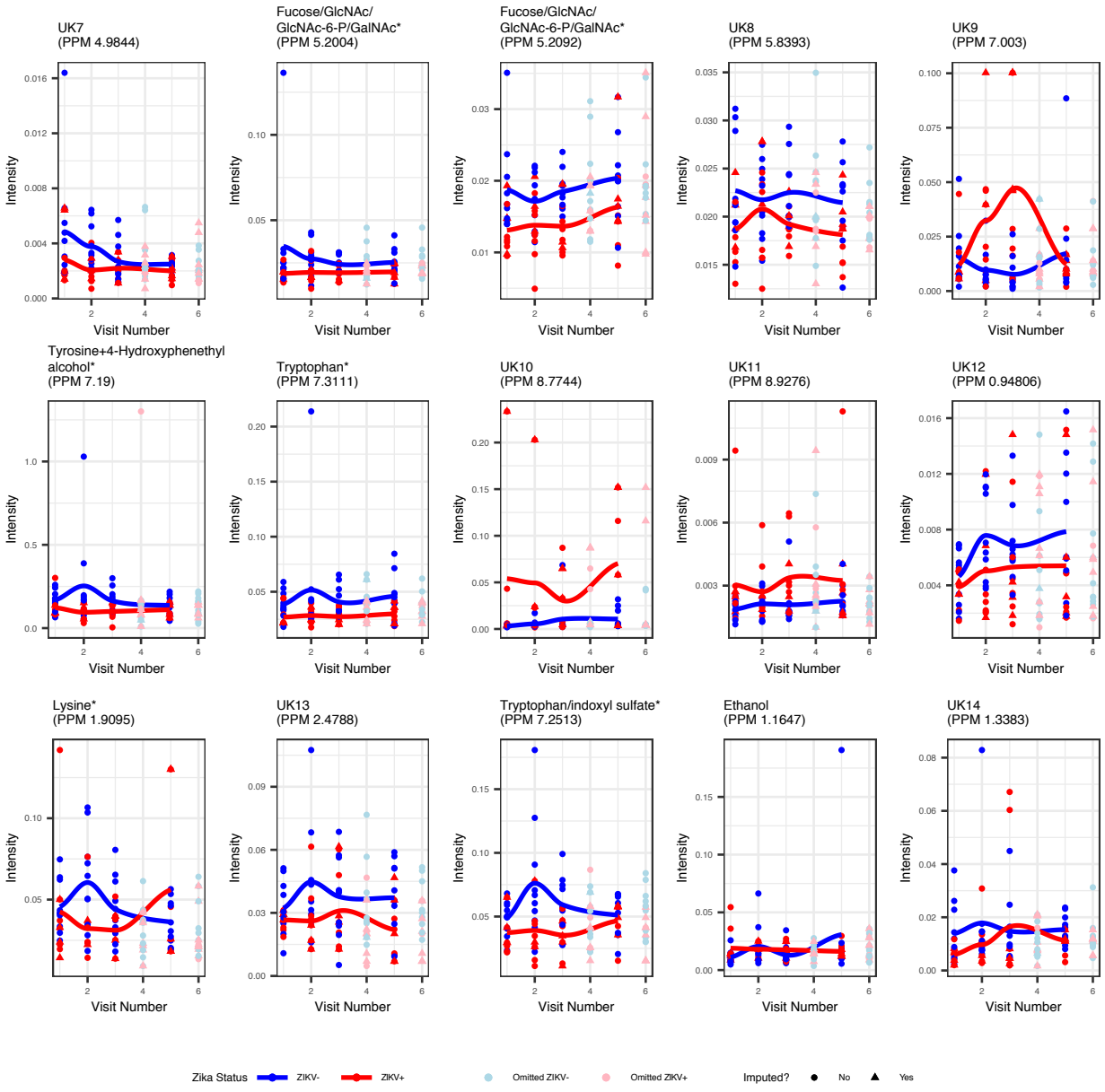
UK: unknown feature. GlcNAc: N-acetylglucosamine; GlcNAc-6-P: N-acetyl-glucosamine 6-phosphate; GalNAc: N-acetylgalactosamine.

[a] Quantification Confidence score is ranged from 0 to 3. It is defined to be 0 for the least quantitively confident peaks, where they are either very small or are located on the shoulder

of other peaks; and to be 3 for the most quantitively confident peaks, where two NMR experts have agreed that this is a true peak.




N−acetyl−glutamine
(PPM 2.0953)
Pseudouridine
(PPM 7.6721)
Threonine
(PPM 4.2753)
Trigonelline
(PPM 8.8415)
Trigonelline
(PPM 9.1134)
UK1
(PPM 1.2175)
UK2
(PPM 1.2285)
Arginine*
(PPM 1.6571)
N-acetyl-glutamine*
(PPM 1.9336)
N-acetyl-glutamine*
(PPM 2.087)
UK3
(PPM 2.2103)
N-acetyl-glutamine/
2-hydroxyglutarate*
(PPM 2.3223)
UK4
(PPM 4.2841)
UK5
(PPM 4.5712)
UK6
(PPM 4.959)
Intensity
Visit Number
Zika Status
ZIKV-
ZIKV+
Omitted ZIKV-
Omitted ZIKV+
Imputed?
No
Yes


UK7
(PPM 4.9844)
Fucose/GlcNAc/
GlcNAc-6-P/GalNAc*
(PPM 5.2004)
Fucose/GlcNAc/
GlcNAc-6-P/GalNAc*
(PPM 5.2092)
UK8
(PPM 5.8393)
UK9
(PPM 7.003)
Tyrosine+4-Hydroxyphenethyl
alcohol*
(PPM 7.19)
Tryptophan*
(PPM 7.3111)
UK10
(PPM 8.7744)
UK11
(PPM 8.9276)
UK12
(PPM 0.94806)
Lysine*
(PPM 1.9095)
UK13
(PPM 2.4788)
Tryptophan/indoxyl sulfate*
(PPM 7.2513)
Ethanol
(PPM 1.1647)
UK14
(PPM 1.3383)
Intensity
Visit Number
Zika Status
ZIKV-
ZIKV+
Omitted ZIKV-
Omitted ZIKV+
Imputed?
No
Yes

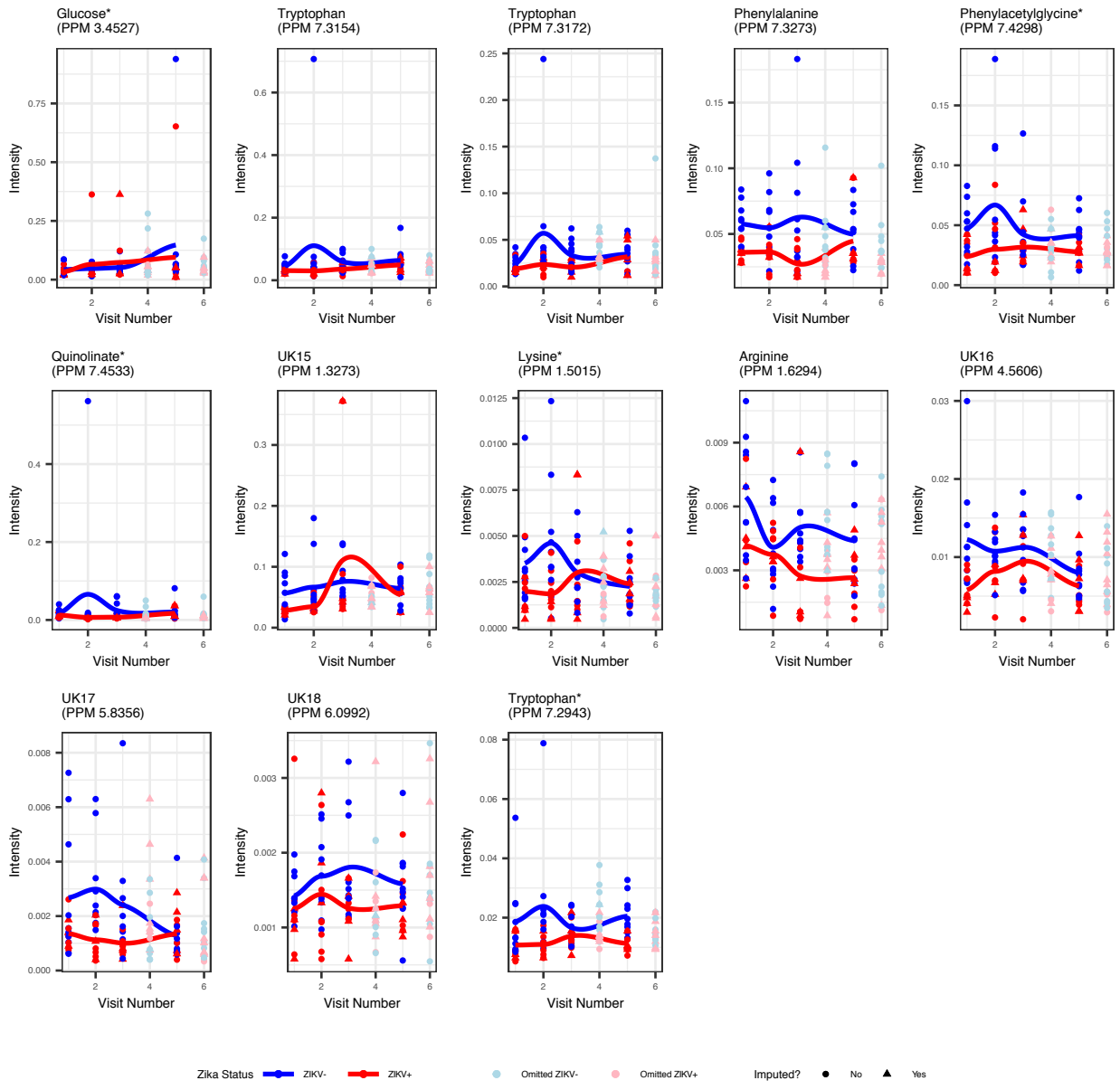


Figure A.6: Graphical representation of intensity scores for all significant features. The fitted curves are not based on the nonparametric model; instead, they are fitted Loess curves to allow easier interpretation of the differences in intensity scores between ZIKV+ and ZIKV+ populations at each feature. The nonparametric model identifies a significant difference in intensity scores for all timepoints for all listed between ZIKV+ and ZIKV- populations. For glucose and 1-methylnicotinamade, outlined by a dotted line, the nonparametric model identifies the interaction between ZIKV-status and the time point as significant, which means the pattern of the intensity score is different over time for these two groups. UK: unknown feature. GlcNAc: N-acetylglucosamine; GlcNAc-6-P: N-acetyl-glucosamine 6-phosphate; GalNAc: N-acetylgalactosamine. Features with * indicate there are more than one metabolites assigned to these features, or the features are overlapped with some unknown metabolites if not specified.

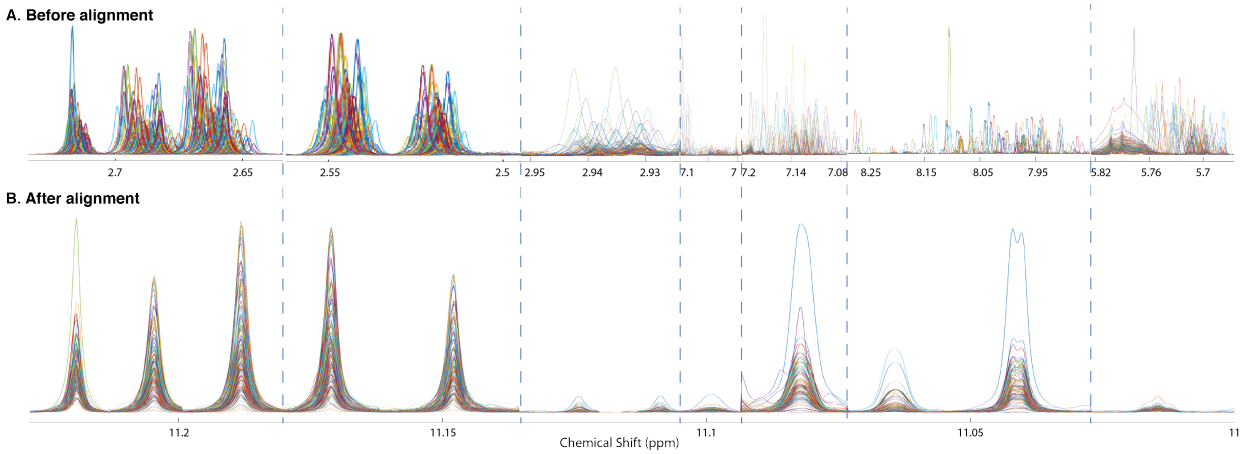


Figure A.7: Performance of alignment on greatly varied peaks. (A) Before alignment. (B) after alignment. Small peaks in panel A were scaled for better visualization.

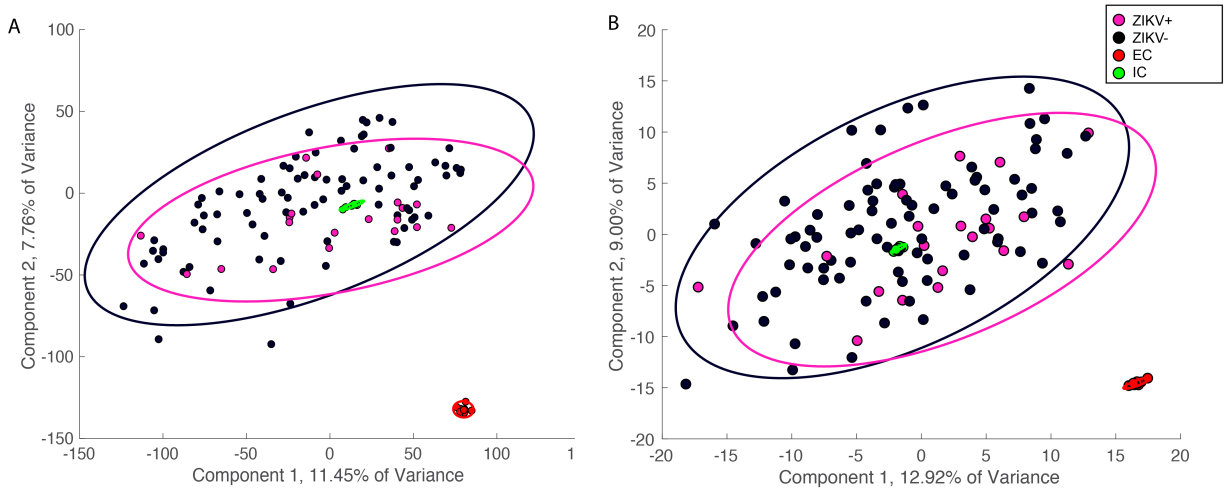


Figure A.8: Score plots of PCA on (A) full-resolution spectra and (B) binned features. EC: external control. IC: internal pooled control

APPENDIX B
BIOGRAPHICAL SKETCH

Sicong Zhang was born and raised in Hebei, China. In 2012, she joined the Resources and Environment department at China Agricultural University (CAU, China) then transferred to Biology department at CAU, where she received her Bachelor of Science degree in Biological Sciences in 2016. She then joined Integrated Life Sciences program at the University of Georgia (USA) at the same year. She joined the Edison lab in 2017. She researched on developing computational tools for NMR-based metabolomics studies and applied these tools to study pregnant-related metabolic changes. She will keep exploring on metabolomics in the future.