Studying the temporal dynamics of the gut microbiota using metabolic stable isotope labeling and metaproteomics

by

Patrick Smyth

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Master's of Science Degree of
Biochemistry with Specialization in Bioinformatics

Department of Biochemistry, Microbiology and Immunology
Faculty of Medicine
University of Ottawa

Abstract

The gut microbiome and its metabolic processes are dynamic systems. Surprisingly, our understanding of gut microbiome dynamics is limited. Here we report a metaproteomic workflow that involves protein stable isotope probing (protein-SIP) and identification/quantification of partially labeled peptides. We also developed a package, which we call MetaProfiler, that corrects for false identifications and performs phylogenetic and time series analysis for the study of microbiome dynamics. From the stool sample of five mice that were fed with ^{15}N hydrolysate from *Ralstonia eutropha*, we identified 15,297 non-redundant unlabeled peptides of which 10,839 of their heavy counterparts were quantified. These peptides revealed incorporation profiles over time that were different between and within taxa, as well as between and within clusters of orthologous groups (COGs). Our study helps unravel the complex dynamics of protein synthesis and bacterial dynamics in the mouse microbiome.

Acknowledgements

I would like to thank Xu Zhang, Zhibin Ning, Janice Mayne, Mathieu Lavallée-Adam, and Daniel Figeys for providing critical feedback on my research. I would also like to give further thanks to Mathieu Lavallée-Adam for his supervision on writing the MetaProfiler software; Daniel Figeys for providing supervision on data analysis; Zhibin Ning for his help with the bioinformatic workflow; and Xu Zhang for designing the experimental protocol.

I would also like to thank my significant other, Jasmine Séguin for her love and support throughout my research work, as well for designing the MetaProfiler logo.

This work was supported by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-156), the Natural Sciences and Engineering Research Council of Canada (NSERC, grant no. 210034), and the Ontario Ministry of Economic Development and Innovation (ORF-DIG-14405), as well as a NSERC Discovery Grant to M. Lavallée-Adam and the stipends provided by TECHNOMISE and the Ontario Graduate Scholarship (OGS).

Statement of Contributions

MetaProfiler was written by P. Smyth, with supervision from M. Lavallée-Adam and Z. Ning. The KNIME pipeline was developed by the OpenMS team, and modified by P. Smyth and Z. Ning. Data analysis was performed by P. Smyth with supervision from D. Figeys, Z. Ning, and X. Zhang. Experimental protocol was designed by X. Zhang. Mouse experiment, stool sample collection, protein extraction, trypsin digestion, and mass spec analysis were performed by X. Zhang, J. Mayne, J. I. Moore, and K. Walker.

List of Acronyms

BH Benjamini-Hochberg

CCAC Canadian Council on Animal Care

CI Confidence Interval

CID Collision Induced Dissociation

COG Category of Orthologous Groups

DE Differential Evolutionary

DNA Deoxyibonucleic Acid

DTT Dithiothreitol

ESI Electrospray Ionization

FFT Fast Fourier Transforms

FPR False Positive Rate

HCD Higher-energy collisional dissociation

HP Heavy Peptides

HPLC High-Pressure Liquid Chromatography

IAA Iodoacetamide

IBD Inflammatory Bowel Disorder

iTRAQ Isobaric Tags for Absolute and relative Quantification

KDE Kernel Density Estimation

LFDR Local False Discovery Rate

LFQ Label Free Quantification

LOOCV Leave One Out Cross Validation

LR Labeling Ratio

MS Mass Spectrometry

NOI Naturally Occurring Isotopes

PEP Posterior Error Probability

PSM Peptide-Spectrum Match

RIA Relative Isotopic Abundance

RITZ Relative Intensity From Time Zero

RMSD Root Mean Square Deviation

RMSE Root Mean Square Error

RNA Ribonucleic Acid

RT Retention Time

SDS Sodium Dodecyl Sulfate

SILAC Stable Isotope Labeling with Amino acids in Cell culture

SIP Stable Isotope Probing

TMT Tandem Mass Tags

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Microbiome

While there are several definitions given for microbiomes, in this manuscript, we will be using the definition provided by Berg et al. (2020), where they gathered 40 experts in diverse microbiome areas and hundreds more around the globe through online surveys to discuss current gaps in the frame of the European-funded MicrobiomeSupport project. The definition provided from this assembly is a modification from the first definition provided for microbiomes by Whipps et al. (1988).

The microbiome refers to the microorganisms (bacteria, eubacteria, fungi, virus etc.) living in a contiguous habitat, which has distinct physio-chemical properties. It also refers to the microbial, internal, and external structural elements compromising the microbiome, which brings about explicit ecological niches. The inhabitants are constantly interacting with one another creating a micro-ecosystem that changes in time and scale. This network

is also integrated within a larger macro-ecosystem that is crucial for maintaining health and function within both systems.

Another term that is often used synonymously with microbiome is microbiota. However, this term refers only to the microorganisms that encompass the microbiome and it does not include the structural elements. Researchers will often look for core microbiota within similar environments to look for microorganisms that are crucial for maintaining the microbiome, such as keystone species.

1.2 Gut Microbiome

Much like other microbial communities, the human gut microbiome is an interactive and dynamic ecosystem that, when healthy, forms a symbiotic relationship with the host. The microbes are provided a nutrient rich environment and the host benefits from important physiological and metabolic functions, such as nutrient and mineral absorption; synthesis of enzymes, vitamins, amino acids, and short-chain fatty acids; and protection against pathogens, just to name a few(Rowland et al., 2018). The mucosal surfaces of the intestines provide the ecological niche needed for the gut microbiota to flourish.

Whereas composition can drastically change between individuals, core functionalities that are needed to establish a symbiotic relationship are conserved(Heintz-Buschart and Wilmes, 2018). When these functions are disturbed, serious health complications can arise, a state we call dysbiosis. Despite their level of importance, however, the influence of the microbiome on human health is largely understudied. Part of the reason for this is that research in this field is heavily driven by technological advancements. Fortunately,

significant strides in genomic sequencing have been observed in recent years. Shafin et al. (2020) were able to sequence the entire genome of eleven human cohorts in just nine days, in contrast to the 13 years that it took for the human genome project to finish in 2003. With this technology, we can measure bacterial abundance and gene expression by counting the number of reads that were generated from their corresponding DNA or RNA sequence.(Conesa et al., 2016; McClure et al., 2013) We can then use this information to establish their functional roles within the gut microbiome. Genomic studies of gut microbial composition has increased our awareness of the impact of their alteration on human diseases, which include obesity(Maruvada et al., 2017), type II diabetes(Komaroff, 2017), inflammatory bowel disease (IBD)(Franzosa et al., 2019), cancer(Helmink et al., 2019), neurological disorders(Tremlett et al., 2017) and cardiovascular issues(Tang and Hazen, 2017)(see Figure 1.1), and has revealed that lifestyle, diet, and therapeutics can drastically change the composition and diversity of the gut microbiome(Shafin et al., 2020).

Many models and systems are available to study the gut microbiome(Starr et al., 2018), each providing advantages and limitations. Early research related to microbiome was limited by the fact that many microbiota could not be cultured in nutrient rich media. Today, however, there are techniques available that have been successful in culturing bacteria, which were previously thought to be unculturable(Lagier et al., 2018). *In vitro* monoculturing methods provide direct assessments of interactions between the host and microbome. In addition, since most of what we know about microorganisms is based on cultures, efforts are being made in uncovering the still remaining unknown species in the human gut. However, monoculturing methods cannot capture the complex interactions occur-

ring with other neighbouring microbes. Another method is using *ex vivo* systems, such as gut on a chip(Kim et al., 2016), organoids(Sato and Clevers, 2013), continuous flow culturing(McDonald et al., 2015), RapidAIM(Li et al., 2020), etc. These are cost effective methods that can simulate the gut microbiome environment. However, these methods cannot be used to analyze microbiome flux as they cannot simulate the internal and external factors governing substrate break down. For this purpose, animal models provide an alternative method for analyzing microbiome. Rodent models are a common choice as they have key similarities with human gut microbiota at the phylum and family levels, as well as sharing 99% genomic similarity(Starr et al., 2018). The disadvantage of using animal models are the cost, facilities, and time needed to accommodate them.

While useful, the bulk of microbiome research is based on genomic sequencing and there is a pressing need to move beyond studying compositional changes. The genetic material cannot tell which pathways are active. It also cannot differentiate between dead, active, and dormant microbes as the presence of a gene does not mean that it is expressed. Researchers have alternatively used transcriptomics to assess the expressed genes of the gut microbiome(Lavelle and Sokol, 2018). The move from metagenomics to metatranscriptomics has been fairly straightforward as many of the software used for metagenomics are applicable to metatranscriptomics as well. In fact, the combination of these two techniques have been shown to improve gene prediction and genome assembly(Heintz-Buschart et al., 2016). More recently, metaproteomics, which uses tandem mass spectrometry (MS/MS) to identify and quantify proteins, has also been used to assess the expressed proteins of the gut microbiome(Zhang et al., 2018b). Compared to metagenomics and metatranscriptomics,

metaproteomics provides a deeper insight into the functionality of the gut microbiome as not all genes and transcripts are translated into proteins. However, the use metaproteomics in microbiome research has not been as straightforward since early research was encumbered by the lack of bioinformatics tools capable of analyzing mass spectrometry data from microbiomes. Fortunately, several software packages have recently emerged to address this issue, such as MetaLab(Cheng et al., 2017), MetaProteomeAnalyzer(Muth et al., 2018), Galaxy-P(Jagtap et al., 2015), trans-proteomic pipeline(Pedrioli, 2010), and OpenMS(Röst et al., 2016). These platforms provide tools that can be easily integrated into a pipeline for pre-processing and post-processing mass spectrometry data. By studying the proteome in addition to the genome, it allows to get a clearer picture of which proteins are behind key processes and to determine the active microbes and their functional roles within the gut microbiome. Moreover, the dynamicity of the microbiome also needs to be characterized as it is an important determinant of the structure of the microbiota. Metaproteomics in Microbiome Research

Metaproteomics involves analyzing proteins from microbial communities. By studying the microbes in their natural environment, we can capture the complex interactions occurring between host and microbes. The typical procedures for proteomics involve bacterial harvesting, protein extraction, denaturation, and digestion, followed by liquid chromatography (separation). The peptides is then separated by High Pressure Liquid Chromatography (HPLC), ionized by Electrospray Ionization (ESI), and analyzed by MS/MS. As peptides elute from the HPLC column, the mass spectrometer continuously cycles between ion selections and ion fragmentation. The purpose of the cycle is to select ions that are

coming from peptides and to identify them via their fragmentation pattern. This is because many of the charged molecules will be the result of contaminants and obtaining the mass of the peptide alone is not sufficient to identify its sequence. Briefly, In the first part of the cycle, the mass spectrometer detects the ionized tryptic peptides by their mass-to-charge ratio (m/z), which are represented by a mass spectrum called MS1 or simply MS. Next, a series of ions are selected based on their intensity (i.e. current generated from the ions) for further sequential analysis. Each ion is isolated by the mass spectrometer and fragmented into smaller pieces, which is commonly done using Collision Induced Dissociation (CID)(Sleno and Volmer, 2004) or Higher-energy C-trap dissociation (HCD)(Olsen et al., 2007). The fragments are then detected by the mass spectrometer, resulting in the MS2 spectrum or MS/MS spectrum. To identify the MS/MS spectra, we can perform a database search, wherein spectra are compared against theoretical patterns generated from known sequences. The list of peptides available for matching are generated from *in silico* digestion of a protein database. Peptides associated with an MS2 spectrum are often called peptide-spectrum matches (PSMs). There are two main strategies for scoring PSMs: deterministic and probabilistic models. Software tools like SEQUEST(Eng et al., 1994) uses a deterministic approach where they calculate a cross-correlation score between the observed and theoretical spectra. The main drawback of this approach is that the cross-correlation scores are difficult to interpret and takes a lot of computational resources to calculate. However, Fast-Fourier Transforms (FFT) can be used to accelerate computation time. It is also heavily affected by the quality of the spectra as well as the length of the peptide, selected modifications, and charge state.(Verheggen et al., 2020) Mascot(Perkins et al., 1999) and Andromeda(Cox et al., 2011), on the other hand, use a

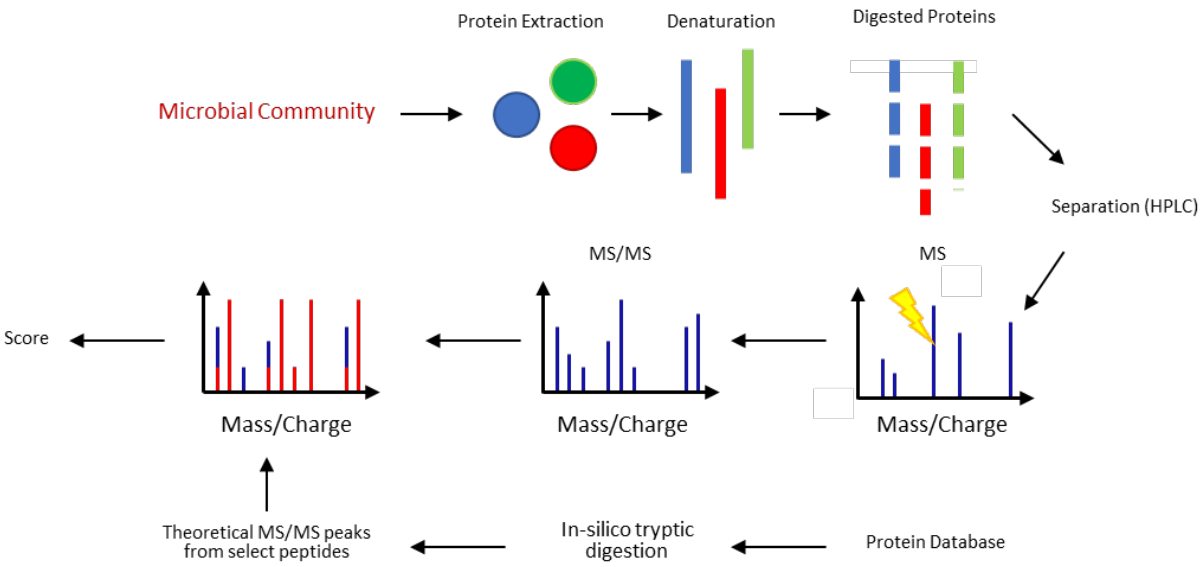

Protein Extraction
Denaturation
Digested Proteins
Microbial Community
Separation (HPLC)
MS/MS
MS
Score
Mass/Charge
Mass/Charge
Mass/Charge
Theoretical MS/MS peaks from select peptides
In-silico tryptic digestion
Protein Database

Figure 1.1: An example of a typical metaprotemics experiment. The steps follow: bacterial harvesting, protein extraction, denaturation, and digestion, liquid chromatography, and mass spectrometry. The circle represents the folded protein, the line represents the denatured protein, and the dashed line represents the digested protein. Database search is then used to identify the peptides in the sample.

probabilistic model. In this strategy, the algorithm calculates the probability of the PSM being from a false discovery by using decoys that were assigned with peptide ids. The advantage of this approach is that it is easier to interpret than the deterministic approach. However, in contrast, the probabilistic model suffers from reproducibility and is heavily influenced by the size of the search space.(Verheggen et al., 2020) As the database size grows, the distribution of false discoveries widens and the difference in score between false and true discoveries decreases(Verheggen et al., 2020). Thus, this results in fewer number of identified peptides. To decrease the search space, our lab recently developed an approach, as part of the package suite MetaLab, that generates a sample specific database by using delegate spectra to search against the large protein database and then extracting the protein sequence according to the resulting peptide identification list to construct the sample-specific database(Xiao et al., 2015).

1.3 Temporal Dynamics of Microbiome

Most disease association studies on gut microbiome have focused on microbiome composition. However, there have been studies suggesting that microbiome dynamics can be a feature in clinical outcome(Gajer et al., 2012; Livanos et al., 2016). Gajer et al. showed that microbial fluctuation in vaginal bacterial communities was different in women with bacterial vaginosis(Gajer et al., 2012). Another study showed that mice administered periodically with antibiotics showed a higher rate of change in microbiota and an increased susceptibility to type 1 diabetes.(Livanos et al., 2016)

The first dense time-series analysis of the human microbiome was conducted by Capo-

raso et al. (2011), where they monitored two individuals, one for six months and the other for fifteen, from four body sites over 396 timepoints. The longitudinal study demonstrated that only 5% of the microbial species were observed in 130 samples in the gut and oral microbiome, representing the core microbiota. Another study monitored the skin, gut, and oral microbiome of 85 individuals between the ages of 17 to 21 on a weekly basis for a period of three months(Flores et al., 2014). While some individuals saw no changes in microbial composition, others saw rapid fluctuations. Interestingly, the use of antibiotics during and before the experiment did not affect the rate of change in microbiome composition. Despite these studies showing differences in the rate of compositional changes, each still retained a personalized microbiome. These finding emphasizes the need to characterize the temporal dynamics of the gut microbiome as both microbial composition and their associated rate of change represent personalized features of human microbiome. By understanding the properties behind microbiome dynamics and the causes behind structural changes in microbiota, we would be able provide further insights into the causal relationships between species and disease. It would also have clinical implications as the rate of recovery of patients undergoing procedure that affect the gut microbiome could be improved.

1.4 Protein-Based Stable Isotope Probing

While metaproteomics is a powerful method in determining the functional repertoire of the gut microbiome, it cannot assess the activity of the microbes since old and newly translated proteins cannot be differentiated. To address this, metabolic incorporation of heavy stable isotopes (usually ^{13}C or ^{15}N) can be used to track the newly synthesized

biomolecules(Jehmlich et al., 2016). By using stable isotopes, microorganisms that assimilate these isotopes can be determined. In addition, by quantifying the proteins that were derived from the labeled substrate, rate of synthesis and, by using phylogenetic information, activity of the microbe can be determined.

Two approaches are currently available for metabolic protein labelling. One uses substrates or nutrients such as labeled glucose, ammonium, or protein hydrolysates, which become incorporated into amino acids, a technique called Stable Isotope Probing (SIP). The second uses a growth medium that contains specifically labelled amino acids (usually lysine and or arginine), which is referred as Stable Isotope Labeling with Amino acids in Cell culture (SILAC). Since the isotopes used are heavier than their most abundant version, a shift in mass will be observed by mass spectrometry, which can be used to detect their incorporation into substrate-derived biomolecules. SIP has become an appealing approach, since two important parameters can be extracted: the average proportion of isotope incorporated in a peptide of interest, termed relative isotopic abundance (RIA), and the relative ratio between the light peptide and the estimated intensity of the heavy peptide, which is termed labeling ratio (LR). LR measures the proportion of proteins that are produced using the isotopes from diet. By monitoring this value over time, the rate of newly synthesized proteins can be evaluated. Labeling ratio is calculated using equation 1.1

$$\mathrm{LR} = \frac{I_h^p}{I_l^p + I_h^p}, \tag{1.1}$$

where I_l^p is the sum of the intensities of the light peaks for peptide p in the spectra and I_h^p is the sum of the intensities of the heavy peaks for p in the spectra.

On the other hand, RIA measures the elemental flux of the isotope. By characterizing the functional and taxonomic origin of the peptide, it gives insight on how the stable isotopic substrate is being converted into biomass and which taxa assimilates this substrate. Thus, measuring RIA can predict which taxa this substrate is limited to. RIA is calculated using equation 1.2

$$\text{RIA} = \frac{H^p - M^p}{F^p - M^p}, \tag{1.2}$$

where H is the m/z position at the center of the theoretical isotopic distribution of the heavy$p$, M is the monoisotopic peak of the light$p$, and F is the m/z position of the fully labeled peptide p. An example of how RIA and LR is calculated is demonstrated in Figure 1.2.

This study proposes using stable isotopes combined with proteomics to provide a deeper understanding of the mechanistic processes governing microbiome dynamics using the properties behind the assimilation of labeled substrates by the gut microbiota.

1.5 Available Isotopes for Protein-SIP

As shown from the average composition for amino acids, $C_{4.7}H_{7.8}N_{1.4}O_{1.5}S_{0.04}$ (also known as an averagine)(Senko et al., 1995), hydrogen is the most abundant element in proteins. However, the stable isotope, deuterium (2H), has not been used in SIP as there is a high degree of hydrogen exchange with the aqueous medium. This is especially true for hydrogen atoms in amide groups and protein side chains. Thus, since exchange will occur during sample preparation, the task of tracking the microorganism that assimilated the

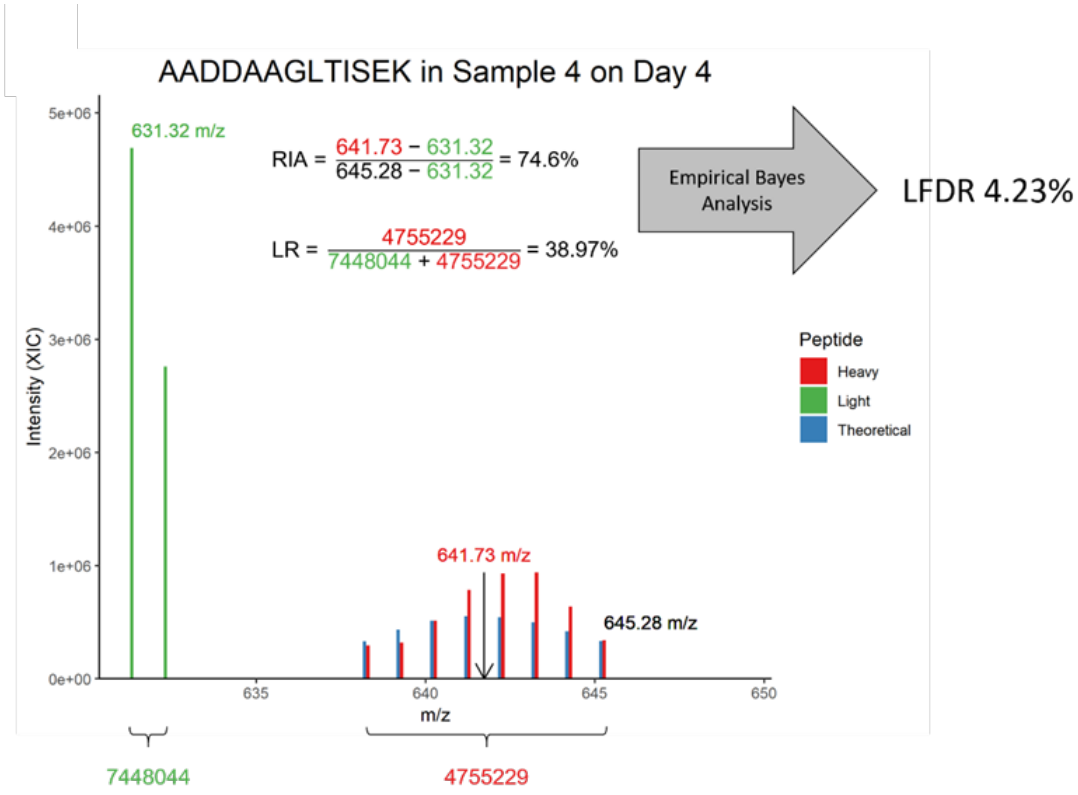

AADDAAGLTISEK in Sample 4 on Day 4
631.32 m/z
RIA = 641.73 − 631.32 / 645.28 − 631.32 = 74.6%
LR = 4755229 / 7448044 + 4755229 = 38.97%
Empirical Bayes Analysis
LFDR 4.23%
Peptide
Heavy
Light
Theoretical
Intensity (XIC)
5e+06
4e+06
3e+06
2e+06
1e+06
0e+00
641.73 m/z
645.28 m/z
635
640
645
650
m/z
7448044
4755229

Figure 1.2: Example of how RIA and RITZ were calculated. Once the light peptide feature was extracted from the raw MS file, a theoretical isotopic distribution (blue lines) was determined and compared against the observed peaks to extract the heavy features. Once a match was found, the m/z value from the center of the theoretical distribution was chosen for RIA calculation. RITZ is the sum of the peaks of the heavy peptide (red) divided by the total sum of the light (blue) and heavy features.

substrate would become nearly impossible. This is further complicated by the fact that high concentration of deuterium inhibits enzymatic activity due to kinetic isotope effects. Deuterium also causes significant changes in retention time, making identification of labeled proteins difficult.

The second most abundant element in proteins is carbon. Since carbon is the main component of bacterial biomass, tracking this atom provides an efficient way to assess metabolic activity of microbial communities. Protein-SIP was originally used to identify bacteria that can degrade environmental pollutants, which are mostly carbon-based. For this reason, most SIP experiments are based on ^{13}C.

Oxygen is the third most abundant element in amino acids. Surprisingly, there has not been any protein-SIP experiments based on ^{18}O labeled complex substrates or nutrients. The moderate cost and high abundance of oxygen in amino acids would seem to be an attractive option for protein-SIP. In addition, there is negligible effects of ^{18}O in reverse phase liquid chromatography. However, its limited use is due to the exchange of ^{18}O with ^{16}O-water by enzymatic reactions, which would reduce its direct incorporation from labeled substrate into biomass(Jehmlich et al., 2016).

Nitrogen ranks fourth in abundance for proteogenic amino acids. This atom has recently gained more use in protein-SIP experiments(Jehmlich et al., 2016). While sensitivity is sacrificed compared to carbon, proteins are enriched in nitrogen compared to overall bacterial biomass. This means that there are fewer spurious peaks on the mass spectrum, in contrast to ^{13}C protein-SIP. It also does not get exchanged with aqueous media like oxygen or hydrogen does.

Sulphur is the element with the lowest abundance in peptides as it is only present in cysteine and methionine. While tracking the incorporation of Sulphur would provide valuable information, labeling with ^{36}S isotope is expensive. However, as high-resolution mass spectrometry become more available, tracking with the cheaper ^{34}S isotope would be possible.(Jehmlich et al., 2016)

1.6 Other Stable Isotope Probing Approaches

Aside from protein-SIP, other biomolecules have been used with SIP to study microbial communities(Egert et al., 2007; Herrmann et al., 2017a,b; Kovatcheva-Datchary and Egert, 2009; Young et al., 2015). In RNA- or DNA-based SIP, the incorporation of isotopes into nucleic acids are investigated. By identifying the microbes with heavy DNA or RNA, the cells that are actively replicating can be determined. Since this approach uses metagenomics or metatranscriptomics, there has been more research conducted with DNA and RNA-SIP than with protein SIP. However, the drawback from this approach is that it requires a high degree of labeling as the resolution for observing the change in mass is low. There is also fatty acid-SIP, which uses LC-MS/MS, but phospholipid fatty acids do not contain rich phylogenetic information(Berry and Loy, 2018). Since proteins can be identified via LC-MS/MS, protein-SIP offers an accurate and sensitive way to determine the incorporation of stable isotopes into proteins. In addition, identified peptide and protein sequences can be used to determine taxa of analyzed organisms.

1.7 Applications of Stable Isotope Probing

Previous research using DNA- and RNA-SIP have been successful in uncovering interactions and roles of gut microbiota(Berry and Loy, 2018). By taking samples at different time points during feeding, we can see which bacteria becomes labeled first to elucidate the effects of cross-feeding. Indeed, this strategy combined with RNA-SIP has been applied in several studies to identify key utilizers of glucose, sialic acid, and resistant starch in gut. Egert et al. (2007) administered ^{13}C-labeled glucose to an in vitro human gut model reactor and demonstrated that *Streptococcus bovis* and *Clostridium perfringens* were the principal glucose fermenters. Labeled glucose has also been applied to murine gut microbiota where Allobaculum spp. was identified as the active glucose fermenter, with lactate, acetate, and propionate being the principal end-products(Herrmann et al., 2017a). Using ^{13}C labelled potato starch and RNA-SIP, another study analyzed murine stool to show that members of *Prevotellaceae* and *Ruminococcaceae* were the main assimilators of resistant starch(Herrmann et al., 2017b). Using the same precursor, one study used an in vitro human model reactor to demonstrate *Ruminococcus bromii* as the main degrader of starch(Kovatcheva-Datchary and Egert, 2009). RNA-SIP using 13-C sialic acid was also used to identify *Prevotella spp.* as the key sialic acid utilizer in the piglet cecal microbiota(Young et al., 2015).

While fewer in number than DNA- and RNA-SIP, protein-SIP has also been successful in identifying microbes involved in metabolic processes, as well as contributing insight into the progress behind certain gut dysbiosis. One study intravenously administered ^{13}C-labeled leucine to mice and demonstrated that members of *Enterobacteriaceae* use amino

acids derived from blood, which may contribute to their growth during small intestinal inflammation associated with parenteral nutrition(Ralls et al., 2016). A similar study, where $^{13}C/^{15}N$-labeled threonine was intravenously administered instead, showed that *Akkermansia muciniphila* and *Bacteroides acidifaciens* were the most abundant utilizers of secreted host proteins(Berry et al., 2013). The first study to use total ^{15}N labeled diet to assess microbial activity in the gut was by Oberbach et al. (2017). In that study, they used protein-SIP to quantify the incorporation of 303 bacterial peptides in colonic mucus of rats fed either a control diet or a high-fat diet with ^{15}N(Oberbach et al., 2017). They found that *Verrucomicrobiaceae* and *Desulfovibrionaceae* were enriched in the active fraction of the community in the high-fat diet. The advantage of using total labeled diet over complex compounds is that analysis is not limited to select microbes as all will intake the labeled diet for growth, which allows to assess the general activity of the microbiome.

1.8 Challenges in Protein-SIP

Protein-SIP has not seen as much use in research as DNA-SIP or RNA-SIP due in large part in the limited technologies available for accommodating protein-SIP experiments. Other labeling strategies such as stable isotope labeling of mammals (SILAM; heavy peptides have single labeled lysine when fully digested by trypsin)(Rauniyar et al., 2013), tandem mass tags (TMT)(Thompson et al., 2003), and isobaric tags for absolute and relative quantification (iTRAQ)(Aggarwal et al., 2006), have defined mass shifts, which can be easily implemented as a variable or fixed modification in most search engines. However, in protein-SIP, a single peptide can contain multiple sites that may or may not be labeled.

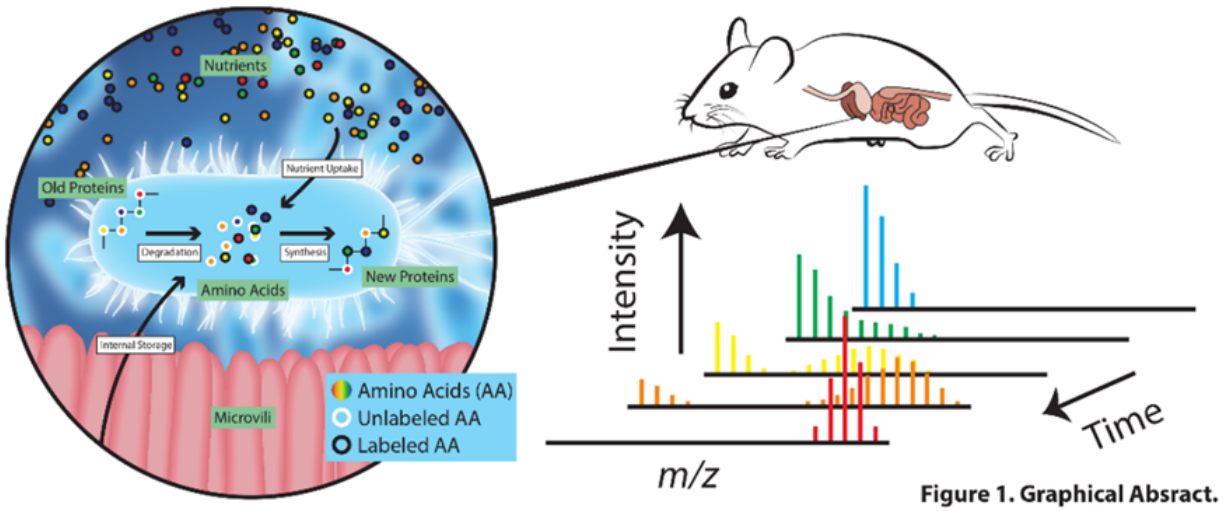


Figure 1. Graphical Absract.

Figure 1.3: An example of partial labeling in protein SIP experiments. The left figure demonstrates the possible source of amino acid for protein synthesis in gut microbes. The right figure shows an example of a mass spectra containing peaks from a light and heavy peptide over time. The extent of labeling is represented the color change from light blue to red.

The concept of partial labeling is in part due to processes like host protein degradation, de novo synthesis, and amino acid interconversion,(Fan et al., 2016) which causes the pool of amino acid to sequentially grow towards heavy isotopes as time progresses. An illustration of this concept is shown in figure 1.3. Therefore, since the number of heavy isotopes within a peptide may not be known at one particular time, common database searches would need to accommodate all possible iterations of labeling for each peptide in the database. This can sometimes be feasible when the number of potential labeling sites are low. However, when using substrates like total ^{15}N labeled diet where all nitrogens in the peptide are potential labeling sites, the search space would be far too big to compute the PSMs in a reasonable amount of time. In addition, due to the large increase in database size, the discrimination power between false and true discoveries will be low, resulting in fewer number of identifications. For this reason, common tools available for labeled peptide identification and quantification, such as MaxQuant(Cox and Mann, 2008), Census(Park et al., 2014), and pQuant(Liu et al., 2014), are not suitable for protein-SIP experiments as they rely on defined mass shifts to generate the list of peptide candidates.

The general strategy to address the issues of partial labeling is to do a database search using only the light peptides and then using their sequence, mass, retention time and charge state to identify their heavy counterparts. This avoids using a very large search space to identify the heavy peptides via MS/MS. Early research using protein-SIP used manual peak extraction and Excel scripts that validated the labeled peptides and isotope patterns(Taubert et al., 2012). However, this process was very laborious and time consuming, often requiring several months of work. Fortunately, several software tools have

been published that offers automated approaches to labeled protein extraction. The first was by Price et al. (2010) where they used a decomposition algorithm to match theoretical isotope distributions to the observed distributions. They applied this software to measure protein turnover in mouse brain tissue. However, its use is only intended for single cultured organisms and not for heterogenous microbial communities. The second was the SIPROS software by Wang et al. (2013), which is used for protein-SIP experiments using ^{15}N-labeled substrates. The drawback to this software is that, compared to the method by Price et al., it requires large amounts of computational resources. Two additional software were published, SIPPER by Slysz et al. (2014) and MetaProSIP by Sachsenberg et al. (2015) within one year of each other. The SIPPER tool is a software package that requires an external input of molecular formulas and feature positions of ^{13}C-labeled peptides. It currently does not support any additional isotopes. MetaProSIP, which is still being updated today, supports ^{15}N, ^{14}C, ^{16,18}O, and ^{2}H based protein-SIP experiments. This software tool uses the same algorithm to Price et al. (2010), but has features compatible with metaproteomic studies. It is also a part of the OpenMS package suite, allowing for fully automated workflows. The package contains many useful tools for processing mass spectrometry data which enables a high degree of flexibility.

In the study by Oberbach et al. (2017), they used MetaProSIP to extract the heavy peptides from their dataset. However, they were only able to identify 303 heavy peptides from 1,850 light peptides due to light peptides disappearing in later timepoints. In this study, I developed a metaproteomic workflow that addresses this issue by spiking unlabeled proteins from the sample at time zero to the pool of labeled proteins to ensure the contin-

uous presence of a light peptide. In addition, I minimized the number of false positives by computing a local false discovery rate for each heavy petpide feature extracted.

1.9 Local False Discovery Rate

One particular issue with doing biological interpretation from proteomic data is that some peptide identifications can be the result of a false positive, which in the case of partial labeling studies can that the form of heavy features quantified from co-eluting peptides and experimental/bioinformatic noise. Thus, confidence scores are often reported with proteomic data so that confident peptide identifications can be obtained. However, no such value is provided for protein-SIP experiments. Thus, in order to estimate a confidence score for protein-SIP data, we can use the idea that the probability of obtaining high RIAs and LRs at early timepoints in a given time series experiment is lower than the probability of obtaining high RIAs and LRs at later timepoints. In addition, we can also use the idea that the distribution of RIA and LR at time zero will reflect the distribution of false discoveries at later timepoints. It is important to note that some deviations between the distribution of RIA and LR at time zero and the distribution of RIA and LR for false discoveries at later timepoints is expected since that, in later timepoints, some of the heavy peptides might co-elute with the light peptide of another, which will be not be reflected at time zero.

1.10 Project Aims

The aims of this project were to (1) develop a bioinformatic workflow that can extract heavy peptides from partially labeled proteomics data; (2) build a tool for analyzing and acquiring confident protein-SIP data, and (3) apply these tools for the analysis of proteome and microbial dynamics in mouse gut microbiome. Namely, in this study, I investigate the extent to which the metaproteome is labeled with ^{15}N to find key species responsible in various metabolic functions as well as finding the principal assimilators of dietary nitrogen. Not only was I successful in aims (1) and (2), the workflow and software greatly enhanced the number of peptides identified and quantified with ^{15}N incorporations from the previous study by Oberbach et al. (2017). This provided deeper insights into the functionality of the gut microbiome, thus fufilling aims (3).

Chapter 2

Materials and Methods

2.1 Mouse experiment and stool sample collection

The animal experiments were performed at the Ottawa Hospital Research Institute by Xu Zhang, Janice Mayne, Jasmine I. Moore, and Krystal Walker. The animal use protocol was approved by the Animal Care Committee at the University of Ottawa and conducted in strict accordance with the guidelines on the Care and Use of Experimental Animals of Canadian Council on Animal Care (CCAC). A total of five male C57BL/6J mice (Charles River, Sherbrooke, QC) were housed individually in the same room at 25 ∘C with a strict 12-h light/dark cycle. Food and water were available *ad libitum*. Mice were acclimatized to the facility for 2 weeks and fed a normal chow diet (containing 18 % fat by energy; Harlan Laboratories, Inc., Madison, WI). The diet was then switched to ^{15}N-labelled SILAM Mouse Diet (Product no.: 231304650; Silantes GmbH, Munich, Germany) for 43 days, which is a ^{15}N labeled hydrolysate from *Ralstonia eutropha*. This diet is a mixture of

protein-free mouse chow with biomass of ^{15}N-labeled *Ralstonia eutropha*. Stool samples were collected at days 0, 1, 2, 4, 8, 12, 19, 29, 34 and 43 and stored at -80 ∘C until analysis.

2.2 Mouse gut microbial protein extraction and trypsin digestion

Approximately 1g of mouse stool sample were suspended in 1.5 mL of ice-cold phosphate-buffered saline (PBS, pH 7.6) with thorough vortexing (3 to 5 glass beads were added to facilitate suspension of stool pellets). The fecal slurries were centrifuged at 300g, 4 ∘C for 5 min. Supernatants were carefully collected, and the pellets were subjected to the above procedure three times. The supernatant for each sample was combined and then followed by three more centrifugations at 300g in 4 ∘C for 5 min to remove debris. The supernatant was then centrifuged at 14,000g in 4 ∘C for 20 min to pellet bacterial cells. The pellet was washed three times by fully re-suspending them in fresh ice-cold PBS. After the last wash, the supernatant was removed and microbial cells were lysed with 4 % (w/v) sodium dodecyl sulfate (SDS) and 6 M urea in 50 mM Tris-HCl buffer (pH 8.0). A tablet of Roche cOmplete ^{TM} mini protease inhibitor tablet was added per 10ml lysis buffer. To promote microbial cell lysis, each sample was subjected to three ultrasonications (30 s each with 1 min interval on ice) using Q125 Sonicator (Qsonica, LLC) with an amplitude of 25%. Cell debris were then removed through centrifugation at 16,000g in 4 ∘C for 10 min. Proteins were precipitated using a 10-fold volume of acidified acetone/ethanol buffer at -20 ∘C overnight and washed three times using cold acetone. The precipitated proteins

were then dissolved in 6 M urea/50 mM ammonium bicarbonate buffer (pH 8) for protein quantitation using DC^{TM} protein assay (Bio-Rad Laboratories) and trypsin digestion.

Aliquots from each Day 0 samples were combined to generate the unlabeled sample. An equal amount of this light sample was spiked into each heavy labeled sample, generating a 1:1 ratio of light and heavy proteins for in-solution trypsin digestion. Trypsin digestion was conducted as described previously(Zhang et al., 2018a). Briefly, 50 μg of proteins were reduced and alkylated with 10 mM dithiothreitol (DTT) and 20 mM iodoacetamide (IAA), respectively. One microgram of trypsin (Worthington Biochemical Corp., Lakewood, NJ) was added to each sample for trypsin digestion at 37 °C overnight with agitation. The tryptic digest was desalted with a 10-μm C18 column and the tryptic peptides were then eluted with 80% (v/v) acetonitrile/0.1% (v/v) formic acid. The eluted peptides were then evaporated using Savant SpeedVac Concentrator and stored at -20∘C for further analysis.

2.3 Mass Spectrometry Analysis

The dried tryptic peptides were dissolved in 0.1% (v/v) formic acid and the peptides equivalent to 2-4 μg of proteins were loaded for mass spectrometry analysis on a Orbitrap Elite mass spectrometer (ThermoFisher Scientific Inc.). The separation of peptides was performed on an analytical column (75 μm x 15 cm) packed with reverse phase beads (1.9 μm; 120-A pore size; Dr. Maisch GmbH, Ammerbuch, Germany). A 2-hr gradient was performed from 5 to 35 % (v/v) acetonitrile containing 0.1 % (v/v) formic acid at a flow rate of 200 nL /min. The instrument method consisted of one full MS scan from 300 to 1800m/z using an Orbitrap mass analyzer followed by data-dependent MS/MS scan of the

20 most intense ions using a LTQ Velos Pro mass analyzer. A dynamic exclusion repeat count of 1 and repeat exclusion duration of 30 s was applied. All data were recorded with the Xcalibur package and exported as RAW format for further analysis. Light and Heavy Peptide Identification and Quantification

For light peptide identification, I identified peptides and assigned them with taxonomic and functional information using MetaLab(Cheng et al., 2017). The platform uses the MaxQuant(Cox and Mann, 2008) package and a target-decoy approach for peptide identification. The protein database used was derived from the catalog of the mouse gut metagenome, by Xiao et al. (2015). A sample specific database was generated using the MS/MS clustering approach from MetaLab(Cheng et al., 2017). This database was *in silico* digested using trypsin, with two allowed missed cleavages and peptide length of 7 to 42 amino acids long. Fixed modification included carbodimethylation of cystein and variable modifications included acetylation (Protein N-term) and oxidation (M). The ppm tolerance was set at 10.

For heavy peptide identification and quantification, I used a modified template pipeline (original template available at `https://sourceforge.net/projects/open-ms/files/Papers/MetaProSIP`) on the KNIME platform(Berthold et al., 2009), which uses tools from the open source package suite, OpenMS(Alka et al., 2019). The difference of our pipeline from the original is that no mass calibration was used and retention time (RT) alignment used MapAlignerIdentification instead of MapAlignerPoseClustering since the light protein spike-in can be used as a point of alignment for each run. In addition, parameters in the tool MetaProSIP were changed so that it would search for ^{15}N labeled peptides

instead of oxygen or hydrogen. The KNIME file describing the pipeline is available at https://github.com/northomics/MetaProfiler.git.

Since the MaxQuant files generated from MetaLab are not compatible with the input files for the OpenMS tools, I developed an R script that can convert result files from MaxQuant into idXML. Feature selection, i.e. the group of peaks along the m/z and RT dimension that belongs to a single peptide, is performed using FeatureFinderMultiplex(Alka et al., 2019), which specializes in finding unlabeled peptide features in labeling experiments. The features were then assigned with a peptide sequence using IDMapper(Alka et al., 2019).

In order to combine all the features from each sample into a single master table, RT alignment was performed using MapAlignerIdentification(Alka et al., 2019). Features were then linked using FeatureLinkerUnlabeledQT(Weisser and Choudhary, 2017). Linked features that contain conflicting peptide sequence information were resolved by keeping the sequence with the best score, which in this case is the peptide with the lowest Posterior Error Probability (PEP)(Cox and Mann, 2008) as reported by MaxQuant. This step was done with IDConflictResolver(Alka et al., 2019). These light features were then used to identify their heavy counterparts using the MetaProSIP tool(Sachsenberg et al., 2015). The correlation threshold was set at 0.2 as it ensured that the heavy peptide feature was selected from each mass spectrum. False discoveries were filtered from data using local false discovery rate. The detailed pipeline is illustrated in Figure 2.1.

Quantification was done through a relative ratio between the light peptide and the estimated intensity of the heavy peptide, which is termed labeling ratio (LR) by MetaProSIP.

However, since proteins from day 0 was spiked in every sample, the labeling ratio in this experiment instead measures the proportion of proteins that are produced using the heavy nitrogen from hydrolysate relative to day 0. Therefore, I will refer to this value as RITZ to differentiate it from LR. By taking this measure over time, the rate of newly synthesized proteins that incorporate the heavy nitrogen from hydrolysate can be estimated. RITZ is calculated the same way as labeling ratio 1.1. An example of how RIA and RITZ is measured is illustrated in Figure 1.2B.

2.4 Extracting the Heavy Peptide Features from Protein-SIP Experiments

Once the features with their assigned IDs are obtained, each sequence is broken down into their elemental composition. Since this study uses total 15N labeled diet, the number of nitrogen atoms defines the maximum number of heavy isotopes that can be artificially introduced (i.e., yield a fully labeled peptide). In addition, the presence of naturally occurring isotopes (NOI) of other elements such as carbon or hydrogen will generate additional peaks. Thus, the maximum number of isotopic peaks that are expected for the fully-labeled peptides would be equal to the number of nitrogen plus the additional peaks from NOIs (which I define as five). With this, to calculate the m/z positions of the isotopic peaks, p_j, one can use equation 2.1(Sachsenberg et al., 2015)

$$p_j = m_k + j\frac{D_e}{z}, \tag{2.1}$$

where m_f is the monoisotopic m/z of the feature k, j is the number of isotopes being considered for the feature, D_e is the mass difference between heavy and light isotope of the labeling element e in the peptides, and z is the charge of the feature. For the observed peak to be extracted, the difference in the m/z and retention time position between the feature and the observed peak must be within a user defined threshold.

When dealing with complex samples, overlap between coeluting peptides or contaminants with the isotopic peaks are going to be expected. To address this, MetaProSIP calculates the Pearson's correlation between the elution profile of the monoisotopic and the putative isotopic peaks in order to detect signals originating from other analytes. If the observed peaks are coming from the real isotopic mass traces, then the Pearson's correlation coefficient will trend towards one. If the peaks are overlapped with spurious signals are it differs in elution profile, then the coefficient will trend towards zero. MetaProSIP allows the users to specify the threshold for the signals to be retained.

2.5 Decomposition algorithm

Once the peaks have been selected, a vector of all isotope intensities, $\mathrm{y} = [\mathrm{y}_1, \ldots, \mathrm{y}_{\mathrm{n+a}}]$, is obtained, where the size will equal to the maximum number of isotopes that can be artificially introduced, n, plus the number of additional peaks corresponding to NOIs, a. In order to resolve the peaks belonging to the heavy peptides, Price et al. (2010) and Sachsenberg et al. (2015) both use a decomposition algorithm to approximate the true isotopic distribution. This involves using the elemental composition of the sequence s to obtain a linear combination of theoretical isotope patterns $\phi(s)$. By multiplying the linear

combination with non-negative weights, β, you obtain the estimated distribution of the heavy isotopes(Sachsenberg et al., 2015).

$$y = \phi(s)\,\beta \equiv \begin{pmatrix} y_0 \\ \vdots \\ y_{n+a} \end{pmatrix} = \begin{pmatrix} \phi_{0,0} & \cdots & \phi_{0,n} \\ \vdots & \ddots & \vdots \\ \phi_{n+a,0} & \cdots & \phi_{n+a,n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{n+a} \end{pmatrix} \tag{2.2}$$

The columns of $\phi(s)$ represents the number of incoporated isotopes being considered and the row represents the peak position. The variable n represents the maximum number of ^{15}N that can be incorporated into peptide s and a represents the number additional peaks corresponding to NOIs. Since there are technical variation and noise, determining an exact solution for the equation would not be possible. Thus, in order to solve for the equation, I minimize the squared residual error between the observed, y, and the calculated, $\phi(s)\,\beta$, intensities. The resulting weights and intensities are then used to calculate the relative isotopic abundance (RIA) and relative intensity (RITZ) of the heavy peptide.

2.6 Filtering Spurious peaks

While MetaProSIP filters out signals that do not match the elution profile of the monoisotopic peaks, some signals originating from irrelevant analytes may still be retained. In order to remove these signals, the tool uses another correlation score based on the shape

of the isotopic distribution.

$$c_r\left(y, \phi_r\left(s\right)\right) = \frac{\sum_{i=0}^{n+a}\left(y_i - \bar{y}\right)\left(\phi_{i,r}\left(s\right) - \bar{\phi}_{i,r}\left(s\right)\right)}{\sqrt{\sum_{i=0}^{n+a}\left(y_i - \bar{y}\right)^2}\sqrt{\sum_{d=0}^{n+a}\left(\phi_{i,r}\left(s\right) - \bar{\phi}_{i,r}\left(s\right)\right)^2}} \tag{2.3}$$

This equation calculates the sample Pearson's correlation coefficient, $c_r\left(y, \phi_r\left(s\right)\right)$, between observed isotope pattern, y, and, the theoretical isotope pattern, $\phi_r\left(s\right)$. The variable $\bar{\phi}_{i,r}$ and $\bar{y}$ corresponds to the mean intensity of the observed and theoretical peak, respectively. If the correlation is not equal or higher than 0.6, the decomposition weights are discarded.

2.7 Filtering Heavy Peptide Features using Local False Discovery Rate

A confidence score can be obtained by using an empirical Bayesian approach to calculate a local false discovery rate (LFDR). LFDR, formulated by Efron et al. (2001), is defined as the probability of the hypothesis being null given the observed data. Let us suppose that, at time t, p_{0_t} is the probability of data, x_t, for false discoveries (the null hypothesis) and p_{1_t} is the probability of data for true discoveries (the alternative hypothesis). In addition, let us suppose that $f_{0_t}\left(x_t\right)$ is the probability density function (pdf) of x_t for the null hypothesis and $f_{1_t}\left(x_t\right)$ is the density of x_t for the alternative hypothesis at time t. Thus, according the law of total probability, the mixture density for both hypotheses, $f\left(x_t\right)$, is measured using equation 2.4

$$f\left(x_t\right) = p_{0_t} f_{0_t}\left(x_t\right) + p_{1_t} f_{1_t}\left(x_t\right). \tag{2.4}$$

By applying Bayes rule to equation 2.4, one can obtain the *a posteriori* probability that the hypothesis H for the random variable X_t with value x_t at time t is null $p_{0_t}(x_t)$ or non-null $p_{1_t}(x_t)$.

$$\begin{aligned} p_{0_t}(x_t) &= P(H = \text{null} | X_t = x_t) = p_{0_t} \frac{f_{0_t}(x_t)}{f(x_t)}, \\ p_{1_t}(x_t) &= P(H = \text{non-null} | X_t = x_t) \\ &= p_{1_t} \frac{f_{1_t}(x_t)}{f(x_t)} \\ &= 1 - p_{0_t} \frac{f_{0_t}(x_t)}{f(x_t)}. \end{aligned} \tag{2.5}$$

Efron et al. specify that, by definition, the LFDR, $P(H = \text{null} | X_t = x_t)$, is equal to $p_{0_t}(x_t)$. Thus, in order to measure LFDR for protein-SIP data, p_{0_t}, $f_{0_t}(x_t)$, and $f_{1_t}(x_t)$ must be measured, where x_t is the observed value of RIA and LR. Since the parameters of the pipeline for quantifying heavy peptide features were the same for each timepoint and that all of the heavy peptide features quantified at time zero are false, the density of RIA and LR for false discoveries can be estimated from time zero using multivariate kernel density estimation to obtain $f_{0_t}(x_t)$. Similarly, the probability of false discoveries at a given timepoint can be estimated by fitting both the distribution of false discoveries and the distribution of true discoveries to the observed mixture density using maximum likelihood estimation. Finally, LFDR can be empirically measured by using equation 2.4 and 2.5.

With all the parameters estimated, a threshold can then be set on the computed LFDR to obtain confident heavy peptide features. It must be noted that the disadvantage to using this approach is that heavy peptide features quantified from unlikely events−for example,

those that have both higher RIA and LR at early timepoints−will often have high LFDR. Thus, these features will be lost in those timepoints when setting the LFDR threshold low.

In our study, the density of RIA and RITZ for false discoveries was empirically estimated from time zero using multivariate kernel density estimation (KDE), performed by the R package ks(Chacón and Duong, 2018), since all of the heavy peptide features quantified at time zero are false. Similarly, the probability of false discoveries at a given timepoint can be estimated by fitting both the distribution of false discoveries, which was measured at time zero, and the distribution of true discoveries, which was assumed to be a bivariate Weibull distribution, to the observed mixture density using maximum likelihood estimation.

With all the parameters set, a threshold can then be set on the computed LFDR to obtain confident heavy peptide features. For this study, features with less than 10% LFDR were used in downstream analysis.

It must be noted that the disadvantage to using this approach is that heavy peptide features quantified from unlikely events−for example, those that have both higher RIA and RITZ at early timepoints−will often have high LFDR. Thus, these features will be lost in those timepoints when setting the LFDR threshold low.

2.8 Data Imputation

In order to predict the RIA or RITZ at missing timepoints, I used equation derived from a three-compartment model and developed by Guan et al. (2012)

$$\gamma\left(t\right) = \left(1 + y_{\mu}e^{-\mu t} + y_{v}e{-}vt + y_{k_{bi}}e^{-k_{bi}t}\right) V_{max} \tag{2.6}$$

where

$$
\begin{aligned}
\mu &= \frac{(k_{st} + k_{0a} + k_{bt}) - \sqrt{(k_{st} + k_{0a} + k_{bt})^2 - (4k_{0a}k_{bt})}}{2}, \\
v &= \frac{(k_{st} + k_{0a} + k_{bt}) + \sqrt{(k_{st} + k_{0a} + k_{bt})^2 - (4k_{0a}k_{bt})}}{2}, \\
y_\mu &= \frac{k_{0a}k_{bi}\,(\mu - k_{bt})}{(\mu - v)\,(\mu - k_{bi})\,\mu}, \\
y_v &= \frac{k_{0a}k_{bi}\,(v - k_{bt})}{(v - \mu)\,(v - k_{bi})\,v}, \\
y_{k_{bi}} &= \frac{k_{0a}\,(k_{bi} - k_{bt})}{(u - k_{bi})\,(v - k_{bi})}.
\end{aligned}
\tag{2.7}
$$

They describe that k_{st} is the total protein synthesis rate constant, k_{bt} is the total protein degradation rate constant, k_{0a} is the amino acid out flow rate constant, and k_{bi}, is the individual protein degradation rate constant. Note that $V_{\max}$, the maximum fraction that can be achieved, is added to the equation. For this study, $V_{\max} = 100\%$ in order to convert the fraction into percentages. The four constants that define the three-compartment model are estimated using the C++ code from RcppDE(Eddelbuettel, 2010), which is a package that uses a differential evolutionary (DE) algorithm for minimizing a cost function. The advantage of using DE was that it struck a balance between the robustness of an exhaustive search and the speed of a gradient descent. The least squares approach was used as the cost function. This approach was evaluated using a leave-one-out cross validation (LOOCV) approach.

2.9 Data Analysis and Visualization

Data analysis was performed using our developed tool MetaProfiler. All plots were generated using ggplot2(Wickham, 2016) or ComplexHeatmap(Gu et al., 2016). The phylogenetic tree was generated with GraPhlAn(Asnicar et al., 2015). The optimal number of clusters from the hierarchical clustering approach was determined by the R package, NbClust(Charrad and Ghazzali, 2014), which reports the best number of clusters by majority vote from 30 clustering indices. To test for over-representation of taxa in each hierarchical cluster, a hypergeometric test was used to assess significance. The total number of non-redundant peptides with more than 4 timepoints were used as background for the hypergeometric test. The p-values were adjusted using the benjamini-hochberg (BH) procedure. The package is available at `https://github.com/psmyth94/MetaProfiler.git`. The mass spectrometry proteomics data have been deposited to the ProteomeXchange with the dataset identifier PXD017451.

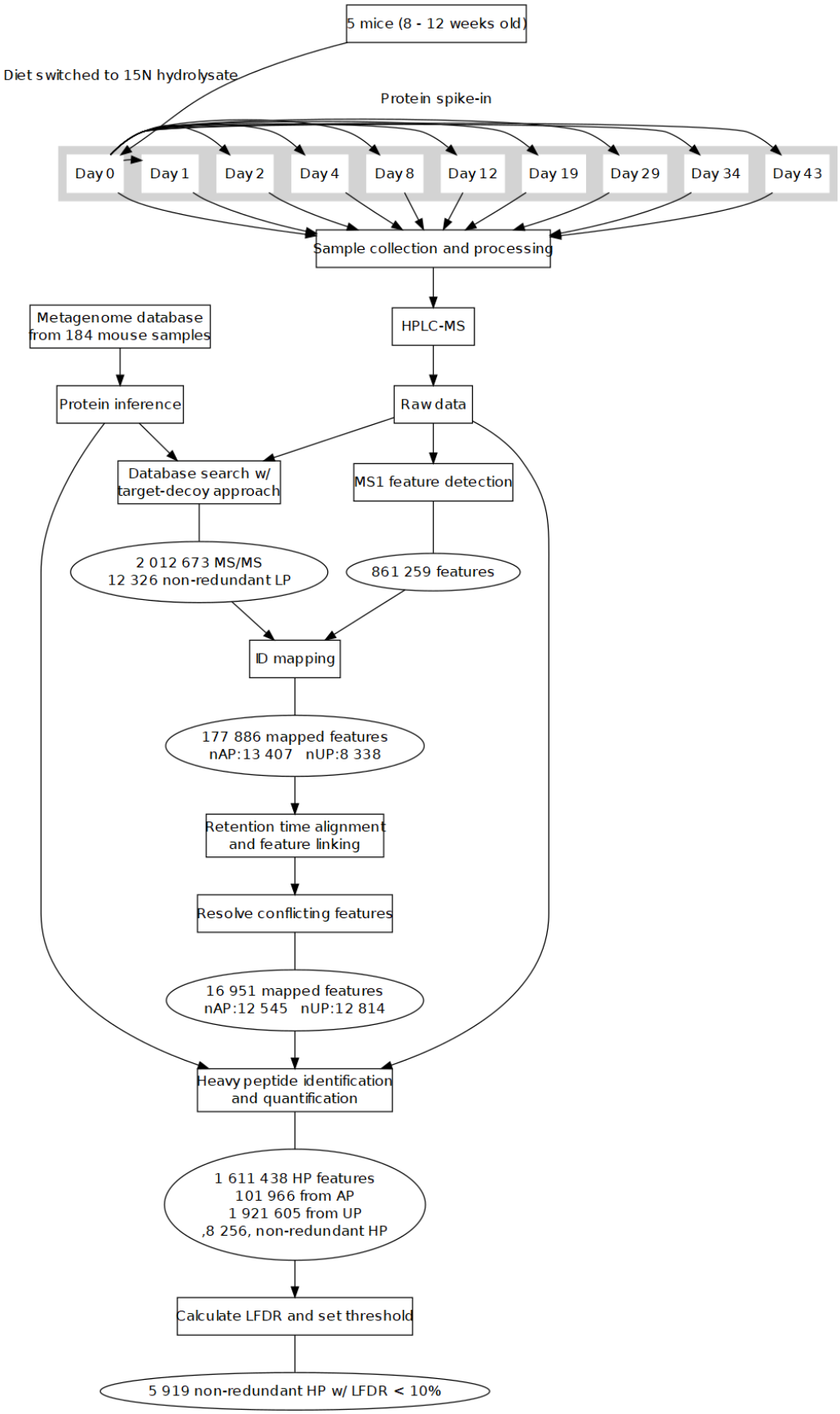

5 mice (8 - 12 weeks old)
Diet switched to 15N hydrolysate
Protein spike-in
Day 0
Day 1
Day 2
Day 4
Day 8
Day 12
Day 19
Day 29
Day 34
Day 43
Sample collection and processing
Metagenome database from 184 mouse samples
HPLC-MS
Protein inference
Raw data
Database search w/ target-decoy approach
MS1 feature detection
2 012 673 MS/MS
12 326 non-redundant LP
861 259 features
ID mapping
177 886 mapped features
nAP:13 407 nUP:8 338
Retention time alignment and feature linking
Resolve conflicting features
16 951 mapped features
nAP:12 545 nUP:12 814
Heavy peptide identification and quantification
1 611 438 HP features
101 966 from AP
1 921 605 from UP
,8 256, non-redundant HP
Calculate LFDR and set threshold
5 919 non-redundant HP w/ LFDR < 10%

Figure 2.1: Detailed Workflow. Flow chart illustrating the pipeline for identifying and quantifying heavy peptides (HP). LP = Light Peptides; AP = Assigned LP; UP = Unassigned LP; nAP = number of non-redundant LP assigned in at least one sample; nUP = number of non-redundant LP unassigned in at least one sample.

Chapter 3

Results

Five mice were fed ^{15}N labeled hydrolysate from *Ralstonia eutropha* for 43 days, stool samples were collected over 10 time points and their microbiomes were analyzed by metaproteomics (Figure 3.1). From the samples, 10,173 non-redundant heavy peptides corresponding to 2,030,462 features were quantified from the 15,297 non-redundant reference peptide identifications. The pipeline for extracting the heavy peptide features is detailed in the methodology chapter as well as summarized in Figure 3.2A. Peptides were kept if they have at least two RIA values: a value below one over the total number of nitrogen in the peptide,(Kim et al., 2016)$N_{\text{max}}{}^{-1}$, which corresponds to the light peptide, and a value above $^{15}N_{\text{max}}{}^{-1}$, which corresponds to the heavy peptide. After filtering at a threshold level of 10% local false discovery rate (LFDR), this resulted in 7,108 non-redundant heavy labeled peptides quantified across all samples.

Pairwise correlation on the label free quantification intensity obtained from maxquant for the samples was performed to determine if the runs can be combined *in silico*. Figure 3.3

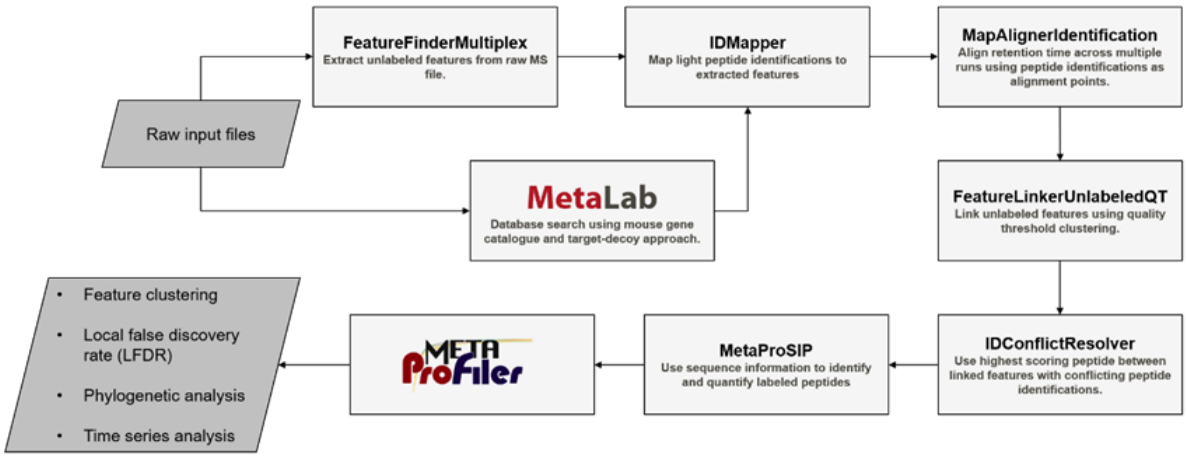

FeatureFinderMultiplex
Extract unlabeled features from raw MS file.
IDMapper
Map light peptide identifications to extracted features
MapAlignerIdentification
Align retention time across multiple runs using peptide identifications as alignment points.
Raw input files
MetaLab
Database search using mouse gene catalogue and target-decoy approach.
FeatureLinkerUnlabeledQT
Link unlabeled features using quality threshold clustering.
IDConflictResolver
Use highest scoring peptide between linked features with conflicting peptide identifications.
MetaProSIP
Use sequence information to identify and quantify labeled peptides
META ProFiler
Feature clustering
Local false discovery rate (LFDR)
Phylogenetic analysis
Time series analysis

Figure 3.1: Bioinformatics Pipeline for Extracting Relative Isotopic Abundance and Relative Intensity from Time Zero. An overview of the workflow for extracting light and heavy peptide features from the raw MS files.

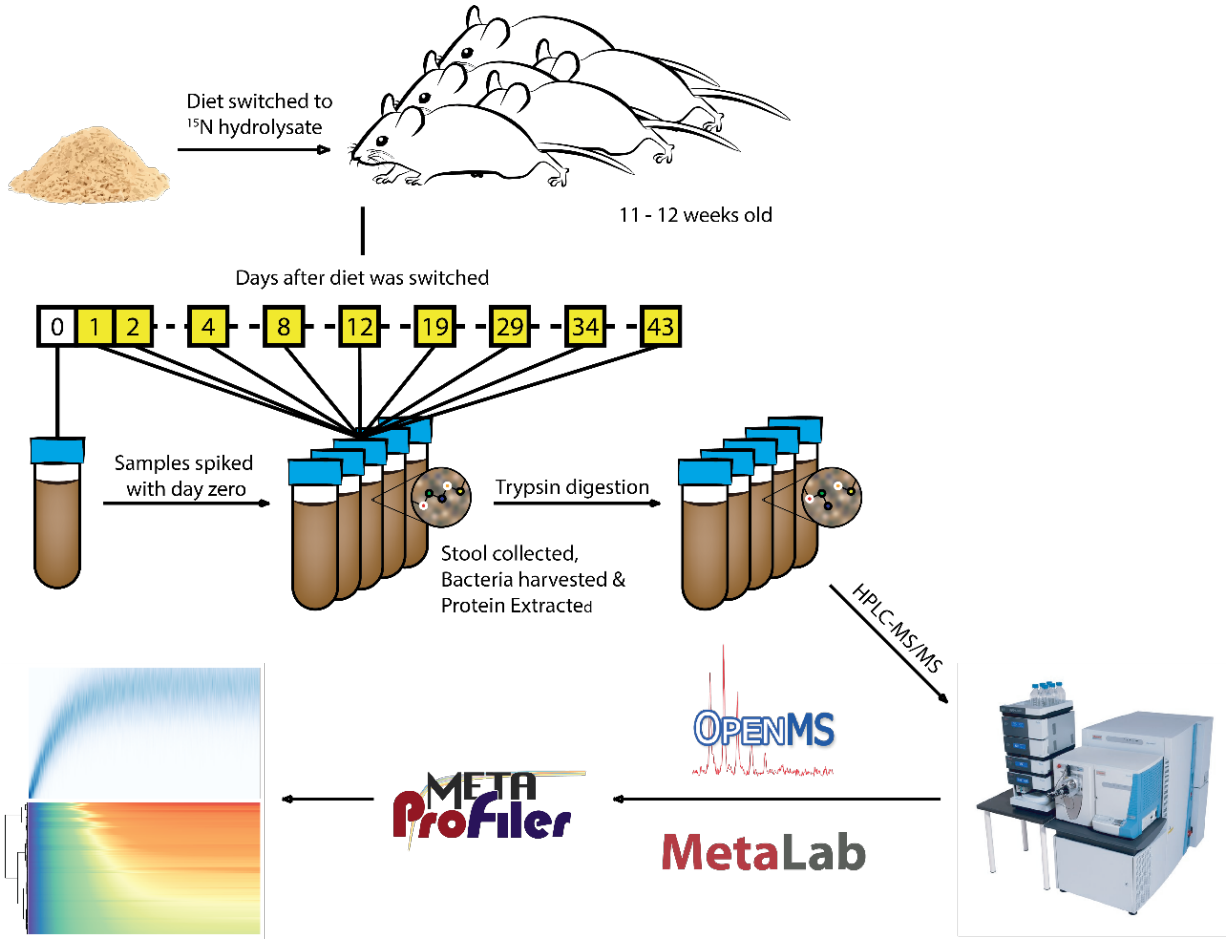

Diet switched to
15N hydrolysate
11 - 12 weeks old
Days after diet was switched
0
1
2
4
8
12
19
29
34
43
Samples spiked
with day zero
Trypsin digestion
Stool collected,
Bacteria harvested &
Protein Extracted
HPLC-MS/MS
OPENMS
META
ProFiler
MetaLab

Figure 3.2: Overview of Experimental Workflow. Five 11-12 weeks old male mice were fed with a 15N labeled mouse diet, where hydrolysate of the chemolithoautotrophic bacteria, Ralstonia eutropha, was the source of heavy nitrogen. Stool samples were collected at 10 different time points, were spiked with day 0 microbiome, and were processed by mass spectrometry. The mass spectrometry data was analyzed by OpenMS, MetaLab, and MetaProfiler to identify and quantify peptides and proteins and to establish profiles of 15N incorporations over time.

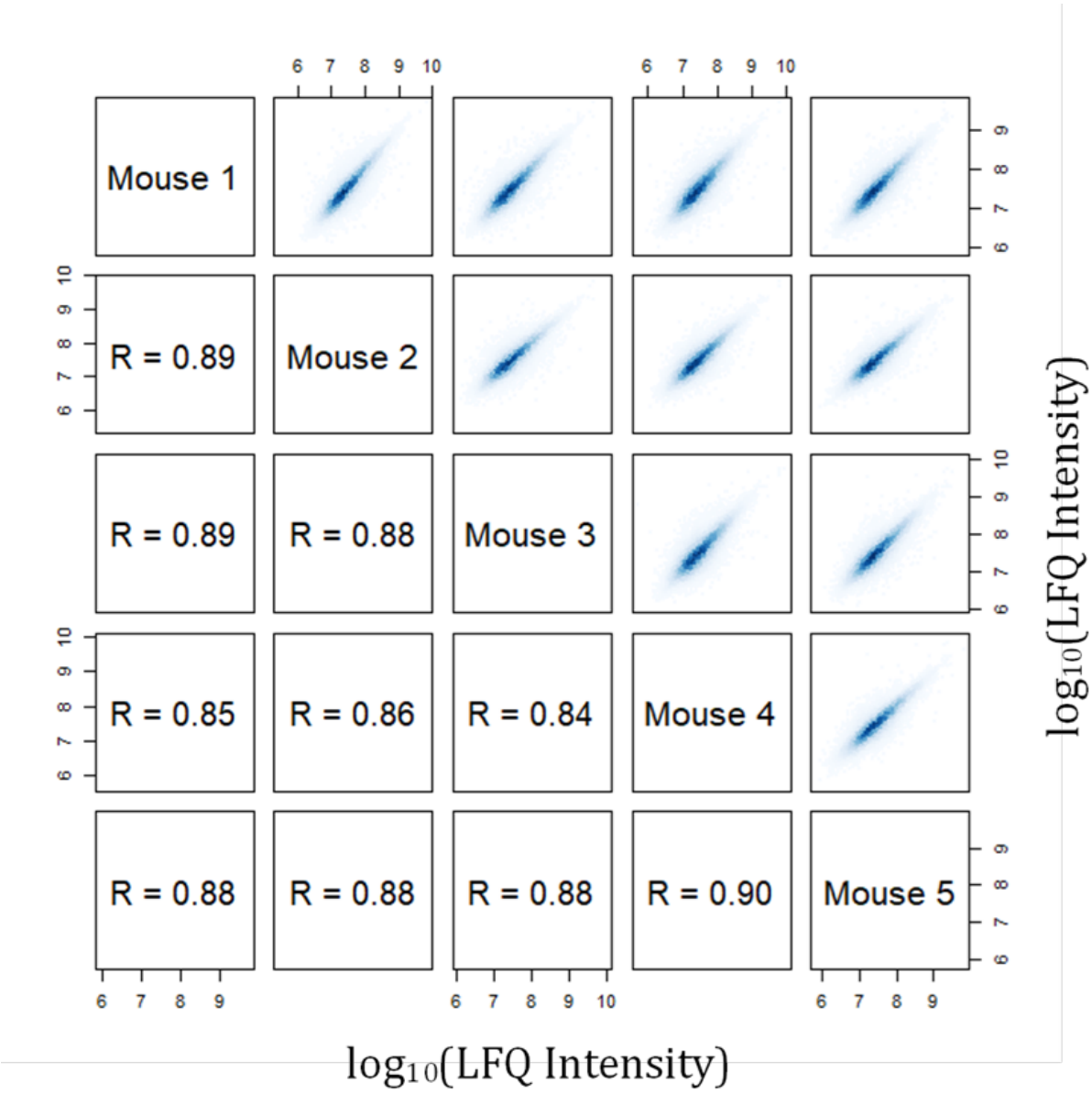
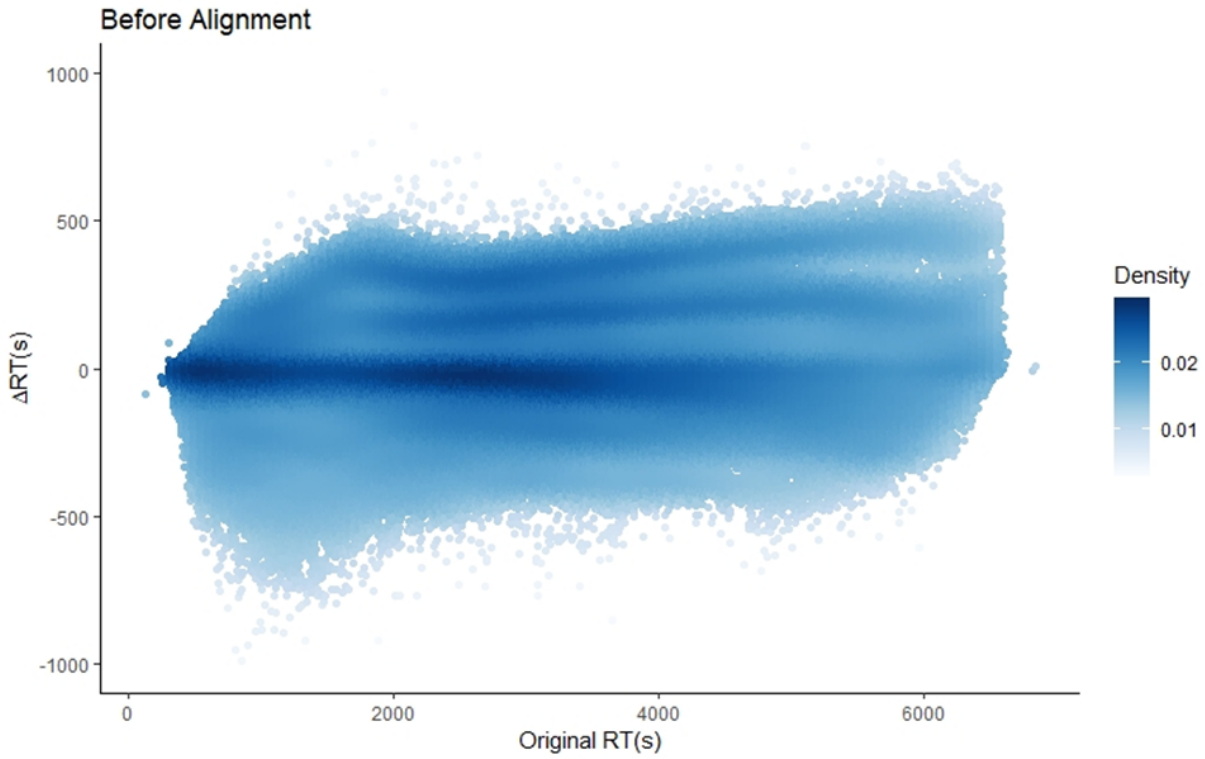

Mouse 1
R = 0.89
Mouse 2
R = 0.89
R = 0.88
Mouse 3
R = 0.85
R = 0.86
R = 0.84
Mouse 4
R = 0.88
R = 0.88
R = 0.88
R = 0.90
Mouse 5
log10(LFQ Intensity)
log10(LFQ Intensity)

Figure 3.3: The lower panels illustrates the Pearson correlation coefficient between corresponding samples. The upper panels are scatter plots where the axes are the label free quantification (LFQ) intensities obtained from MaxQuant. The color of the points corresponds to their density where blue is the highest value and light blue is the lowest value.

illustrates the scatter plot and the Pearson correlation coefficient obtained. The smallest value measured is 0.84 denoting that the mouse samples have similar microbiome profiles.

The performances of different workflows were also assessed. When no pooling is done, 2,003 heavy peptides are obtained. Meanwhile, when the timepoints for each sample are pooled together, a slight improvement is seen (3,125 peptides). However, the biggest increase comes from pooling and aligning the samples together, where 10,173 non-redundant heavy peptides are obtained. The need to align the samples together is clearly demonstrated in Figure 3.4, where the retention time can shift from around 5 minutes in different runs. A demonstration of how retention times are aligned is present in figure 3.5.

I also investigated the intensities (Table 3.1), number of identification (Table 3.2), and RIA (Table 3.3) of peptides identified at time zero and those that were not. As seen from Table 3.1, the average log10 intensity, log10(INT), is higher on day 1 for time zero peptides. In Table 3.2, there is an increase in the number of ids from day 1 to day 2 for non-time zero peptides, while you get ˜2000 less ids for time zero peptides. In addition, ˜50% less non-time zero peptide ids are obtained on day 43 than on day 1, while for time zero, there is 30% less. In Table 3.3, the results demonstrates that RIA was significantly lower for non-time-zero peptides from day 4 to day 43 compared to peptides found at time zero.


Before Alignment
1000
500
0
-500
-1000
ΔRT(s)
0
2000
4000
6000
Original RT(s)
Density
0.02
0.01

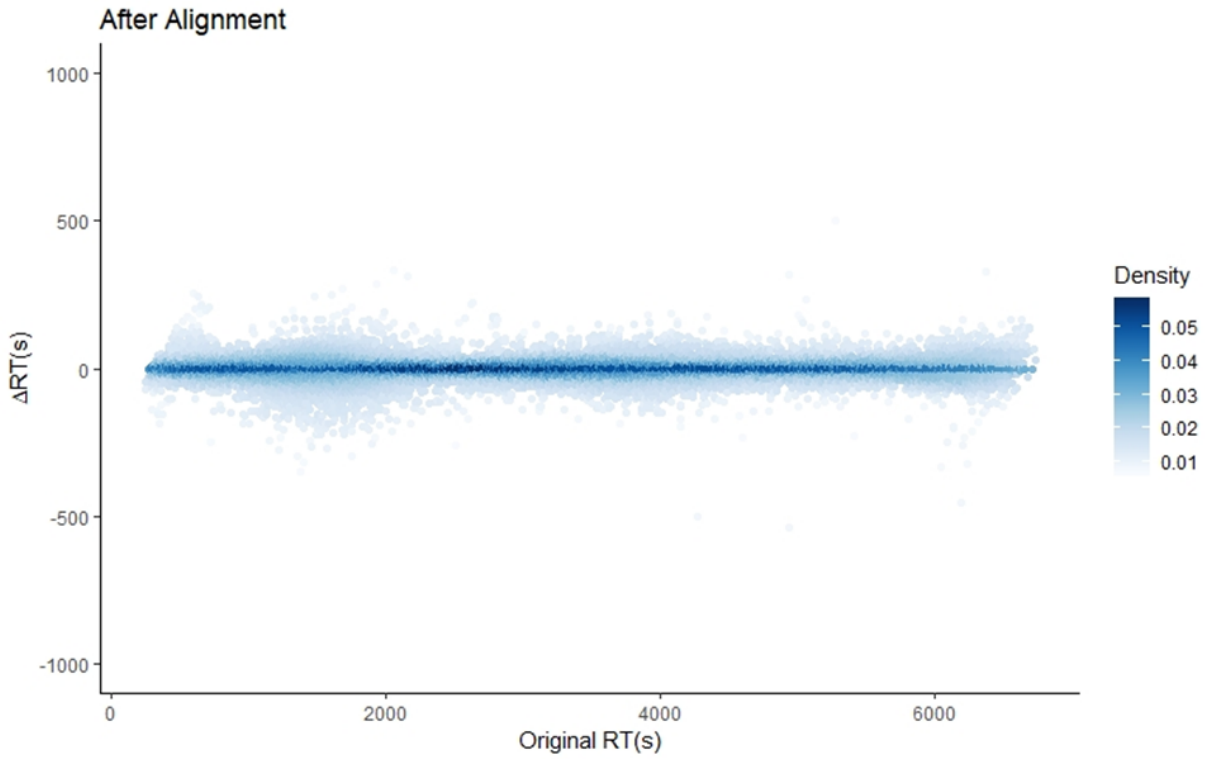
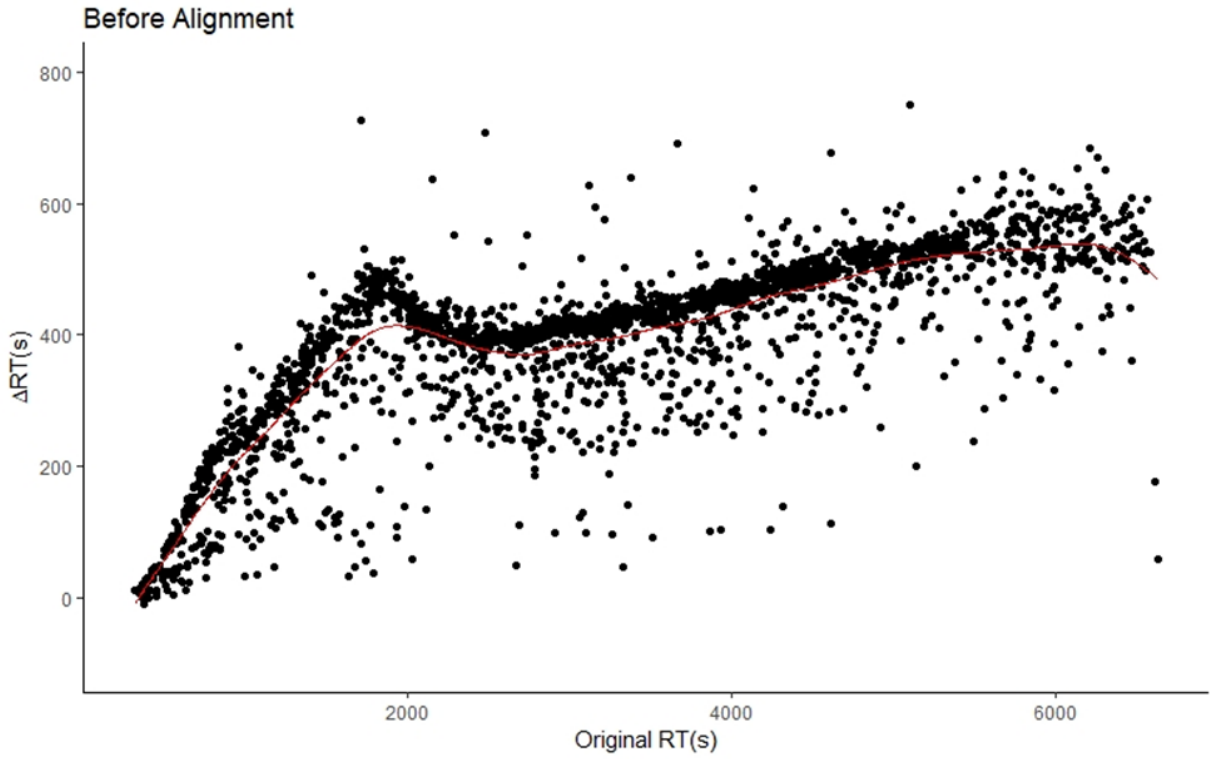

After Alignment
1000
500
0
-500
-1000
ΔRT(s)
0
2000
4000
6000
Original RT(s)
Density
0.05
0.04
0.03
0.02
0.01

Figure 3.4: The overall retention time drift before and after alignment. The y axis is the difference between the median retention time for a given peptide in all sample and the actual retention time observed in the sample. The x axis is the actual retention time for a given peptide in a sample. The color of the points corresponds to their density where blue is the highest value and light blue is the lowest value.


Before Alignment
800
600
400
200
0
ΔRT(s)
2000
4000
6000
Original RT(s)

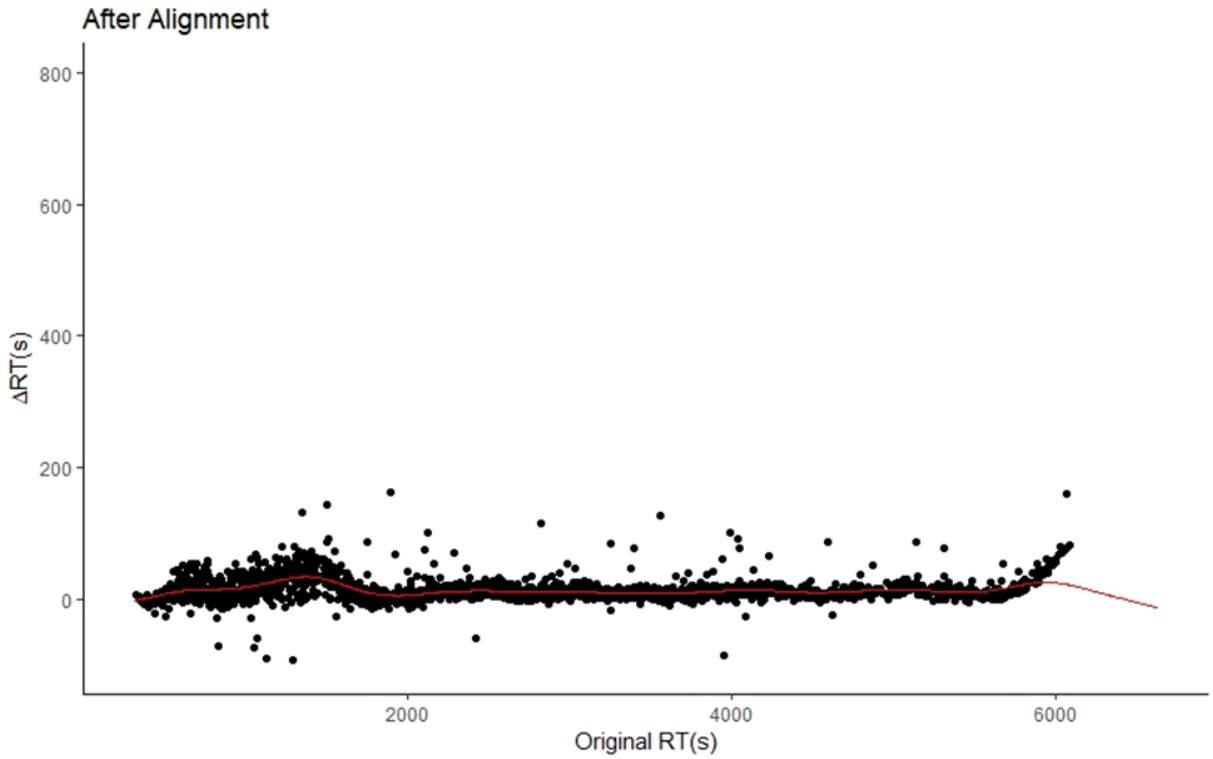

After Alignment
800
600
400
200
0
ΔRT(s)
2000
4000
6000
Original RT(s)

Figure 3.5: The retention time drift before and after alignment for mouse 1 in day 43. The y axis is the difference between the median retention time for a given peptide in mouse 1 and the actual retention time observed in the sample. The x axis is the actual retention time for a given peptide. A line is interpolated using cubic smoothing splines (red line) and then the observed retention time is subtracted by the drift value predicted from the smoothing function.

Day	Found at time zero (log10(INT))	Not found at time zero (log10(INT))
0	7.2706	NA
1	6.9724	6.6027
2	6.7362	6.3368
4	6.3861	6.3688
8	6.5338	6.3144
12	6.7486	6.4897
19	6.8771	6.5395
29	6.7252	6.4944
34	6.7399	6.463
43	6.7816	6.501

Table 3.1: Average Intensity of Peptides Found at Time Zero and Average Intensity of Peptides Found at Later Time Points. INT represents the average intensity of peptides found at time zero and peptides found at later time points

Day	Number of time zero peptides	Number of non-time zero peptides
0	12094	0
1	7410	1066
2	5771	1160
4	5608	901
8	5581	842
12	5199	657
19	4658	444
29	5223	576
34	5208	610
43	5019	591

Table 3.2: Number of Peptides Found at Time Zero and Number of Peptides Found at Later Time Points

Day	Average RIA for time-zero peptides (%)	Average RIA for non-time-zero peptides (%)	p-value
1	46	50.9	1
2	47.6	48.5	1
4	67.2	64.9	2×10^{-18}
8	73.8	70.9	2×10^{-33}
12	77.9	75.2	7×10^{-32}
19	81	76.6	7×10^{-41}
29	82.6	79	2×10^{-43}
34	84	80.9	2×10^{-42}
43	85.3	81.7	2×10^{-57}

Table 3.3: The Average RIA of Peptides Found at Time Zero and Average RIA of Peptides Found at Later Time Points. The p-value is from a Wilcoxon test Correction was performed using BH procedure.

3.1 Taxonomic and functional characterization of the heavy labeled peptides in mouse gut microbiome

To characterize the functions and taxonomic origin of heavy labeled proteins in microbiome samples, I used the confidently quantified heavy peptides at day 29 to 43. These days were chosen by determining which day the RITZ starts to plateau. Plateau detection was done using a rolling standard deviation, where a window of fixed length moves over the RITZ of each peptide and then computes the standard deviation of the RITZ within the window. The fixed length chosen for the windows was nine days. The average standard deviation across all peptides at each permutation are 9.36% (day: 1, 2, 4, 8), 7.79% (day: 4, 8, 12), 6.84%

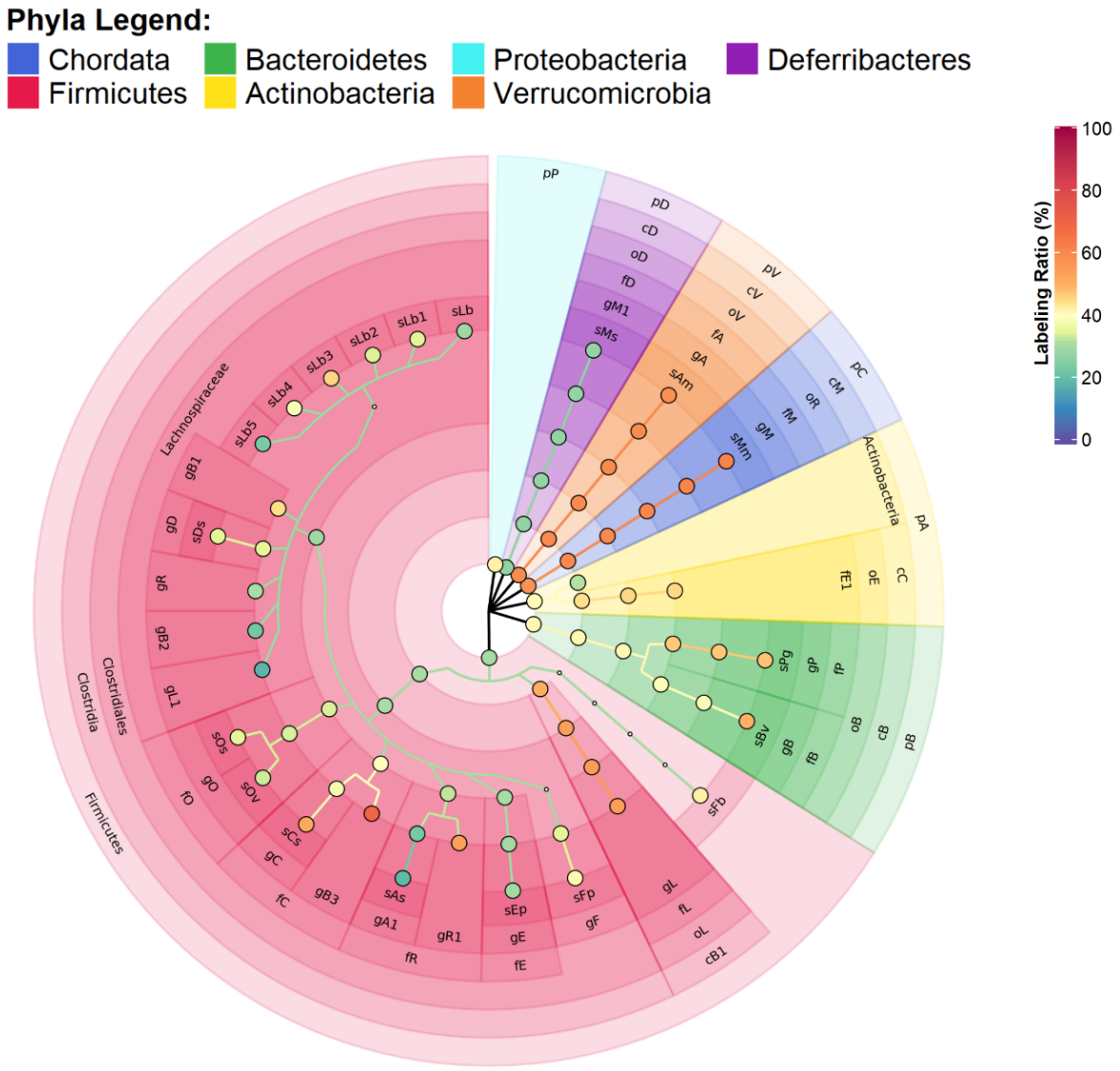
(day: 12, 19), 5.99% (day: 29, 34), and 6.23% (day: 34, 43). Wilcoxon tests showed that the standard deviation in the last window with day 34 and 43 is significantly less than previous windows, except for the second to last one (p-value = 5×10^{-124}, 2×10^{-49}, 1×10^{-4}, and 0.3, respectively; BH procedure). Thus, this suggests that the plateau starts at day 29 and the RITZs past this day reflects the maximum proportion of the labeled proteins that the taxon can attain when using hydrolysate as a nitrogen source.

To assess the contribution of the ^{15}N isotope effect, I compared our dataset against the dataset by Webhofer et al. (2013), where they analysed blood plasma proteins from mice using the same diet as this study. I compared the $\log_2$(^{15}N peptide intensity/^{14}N peptide intensity) values of peptides that were in at least three samples and had a p-value from a one-sample t-test less than 0.05 after FDR correction. This is to ensure that the fold

change is not due to technical variability from the mass spectrometry instrument. The absolute mean $\log_2$ value for the Webhofer dataset(Webhofer et al., 2013) is 0.90 at day 56, while this dataset is 2.19 at day 43, which is significantly higher (p-value $= 1 \times 10^{-122}$).

The taxonomic RITZ was first mapped to a phylogenetic tree using GraPhlAn (Figure 3.6). Taxonomic RITZ was taken as the average across all its distinct peptides. Proteins from mice (Chordata) were within the highest peptide RITZ ratios (RITZ $= 0.584 \pm 0.020$ at 95% Confidence Interval (CI)). Similarly, Verrucomicrobia, a group of mucin degraders who rely on the host for nutrients, had a similar RITZ level (0.606 ± 0.078 at 95% CI). Three of the four dominant phyla in the gut have mid-range RITZ levels including Proteobacteria (0.426 ± 0.079 at 95% CI), Bacteroidetes (0.415 ± 0.010 at 95% CI), and Actinobacteria (0.304 ± 0.196 at 95% CI). Interestingly, the forth and most dominant phyla, Firmicutes, had significantly lower RITZ (0.306 ± 0.008 at 95% CI), than the host cells (p-value$=2 \times 10^{-104}$; two sample t-test; BH procedure) and Bacteroidetes (p-value$=3 \times 10^{-58}$; two sample t-test; BH procedure). While most of the taxa belonging to Firmicutes had low RITZ, the RITZ of the lactic acid bacteria group (e.g. *Lactobacillus*; 0.548 ± 0.077 at 95% CI) was markedly higher. The mean RITZ of each taxon are available in Table A.1.

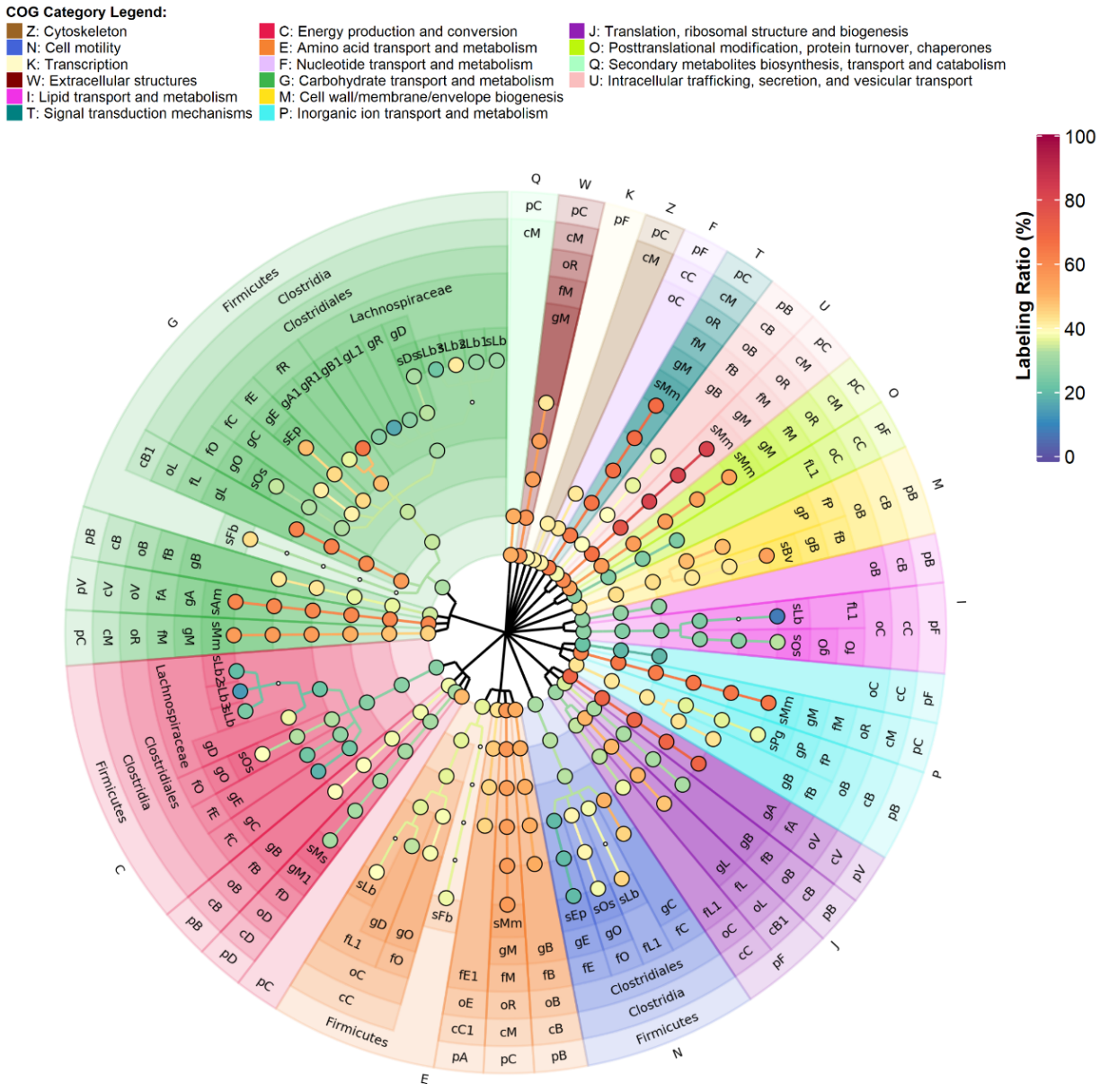
To further examine the functional distributions of ^{15}N labeled proteins, taxa were assigned to categories of COGs (Figure 3.7) using the leading razor proteins associated with their distinct peptides, as reported from MaxQuant. Functions of "carbohydrate transport and metabolism" and "amino acid transport and metabolism", "energy production and conversion", as well as "translation, ribosomal structure, and biogenesis" were among the



Annotation Legend:

pA: Actinobacteria
pB: Bacteroidetes
pC: Chordata
pD: Deferribacteres
pP: Proteobacteria
pV: Verrucomicrobia
cB: Bacteroidia
cB1: Bacilli
cC: Coriobacteriia
cD: Deferribacteres
cM: Mammalia
cV: Verrucomicrobiae
oB: Bacteroidales
oD: Deferribacterales
oE: Eggerthellales
oL: Lactobacillales
oR: Rodentia
oV: Verrucomicrobiales
fA: Akkermansiaceae
fB: Bacteroidaceae
fC: Clostridiaceae
fD: Deferribacteraceae
fE: Eubacteriaceae
fE1: Eggerthellaceae
fL: Lactobacillaceae
fM: Muridae
fO: Oscillospiraceae
fP: Porphyromonadaceae
fR: Ruminococcaceae
gA: Akkermansia
gA1: Anaerotruncus
gB: Bacteroides
gB1: Blautia
gB2: Butyrivibrio
gB3: Butyricicoccus
gC: Clostridium
gD: Dorea
gE: Eubacterium
gF: Flavonifractor
gL: Lactobacillus
gL1: Lachnoclostridium
gM: Mus
gM1: Mucispirillum
gO: Oscillibacter
gP: Parabacteroides
gR: Roseburia
gR1: Ruminococcus
sAm: Akkermansia muciniphila
sAs: Anaerotruncus sp. G3(2012)
sBv: Bacteroides vulgatus
sCs: Clostridium sp. ASF502
sDs: Dorea sp. 5-2
sEp: Eubacterium plexicaudatum
sFb: Firmicutes bacterium ASF500
sFp: Flavonifractor plautii
sLb: Lachnospiraceae bacterium A4
sLb1: Lachnospiraceae bacterium 28-4
sLb2: Lachnospiraceae bacterium COE1
sLb3: Lachnospiraceae bacterium M18-1
sLb4: Lachnospiraceae bacterium A2
sLb5: Lachnospiraceae bacterium 10-1
sMm: Mus musculus
sMs: Mucispirillum schaedleri
sOs: Oscillibacter sp. 1-3
sOv: Oscillibacter valericigenes
sPg: Parabacteroides goldsteinii

Figure 3.6: Relative Intensity from Time Zero of Proteins in Mouse Gut Microbiome from Day 29 to 43. Phylogenetic tree where size and color of the nodes relates to the RITZ of the taxon; color of the cells corresponds to the phylum lineage; and the transparency denotes the rank. For nodes with annotations, the key follows the convention: [first letter of phylogenetic rank][initials][unique identifier]. Nodes are reported when the number of distinct peptides is greater than 3 and the adjusted p-value from a one-sample t-test is below 0.05.



Annotation Legend:

pA: Actinobacteria	oD: Deferribacterales	fP: Porphyromonadaceae	gR: Roseburia
pB: Bacteroidetes	oE: Eggerthellales	fR: Ruminococcaceae	gR1: Ruminococcus
pC: Chordata	oL: Lactobacillales	gA: Akkermansia	sAm: Akkermansia muciniphila
pD: Deferribacteres	oR: Rodentia	gA1: Anaerotruncus	sBv: Bacteroides vulgatus
pF: Firmicutes	oV: Verrucomicrobiales	gB: Bacteroides	sDs: Dorea sp. 5-2
pV: Verrucomicrobia	fA: Akkermansiaceae	gB1: Butyrivibrio	sEp: Eubacterium plexicaudatum
cB: Bacteroidia	fB: Bacteroidaceae	gC: Clostridium	sFb: Firmicutes bacterium ASF500
cB1: Bacilli	fC: Clostridiaceae	gD: Dorea	sLb: Lachnospiraceae bacterium A4
cC: Clostridia	fD: Deferribacteraceae	gE: Eubacterium	sLb1: Lachnospiraceae bacterium COE1
cC1: Coriobacteriia	fE: Eubacteriaceae	gL: Lactobacillus	sLb2: Lachnospiraceae bacterium 28-4
cD: Deferribacteres	fE1: Eggerthellaceae	gL1: Lachnoclostridium	sLb3: Lachnospiraceae bacterium 10-1
cM: Mammalia	fL: Lactobacillaceae	gM: Mus	sMm: Mus musculus
cV: Verrucomicrobiae	fL1: Lachnospiraceae	gM1: Mucispirillum	sMs: Mucispirillum schaedleri
oB: Bacteroidales	fM: Muridae	gO: Oscillibacter	sOs: Oscillibacter sp. 1-3
oC: Clostridiales	fO: Oscillospiraceae	gP: Parabacteroides	sPg: Parabacteroides goldsteinii

Figure 3.7: Functional Distribution of Heavy Labeled Proteins in Mouse Gut Microbiome from day 29 to 43. The tree describes the functional distribution of each taxon. The size and color of the nodes relates to the RITZ of the taxon; color of the cells corresponds to the functional COG category; and the transparency denotes the rank. Nodes were annotated and reported in a similar fashion to figure 3.

categories with the highest number of heavy peptide associated to them. The mean RITZ of each taxon with their associated COG annotations are available in Table A.2.

The host showed high levels of heavy peptides in most categories. In particular, peptides from "intracellular trafficking, secretion, and vesicular transport"(0.612 ± 0.056 at 95% CI), which mostly includes calcium-dependent phospholipid-binding proteins, were among the highest incorporators of ^{15}N indicative of newly synthesized proteins. In contrast, peptides from "secondary metabolites biosynthesis, transport, and catabolism" (0.506 ± 0.293 at 95% CI), which mostly includes multidrug resistant proteins, were among the lowest incorporators of ^{15}N. This was reasonable as these proteins are only expressed in response to drugs(Klaassen and Aleksunes, 2010; Mottino et al., 2000).

I also investigated the light peptide intensities of peptides not identified at time zero and noticed that peptides not identified at time zero are significantly lower in intensity than the ones that were. As seen from Table 3.1, the average log10 intensity, log10(INT), is higher on day 1 for time zero peptides.

3.2 Dynamics of ^{15}N Incorporation into Peptides of the Mouse Gut Microbiota

I next explored the ^{15}N incorporation profiles over time. In particular, I focused on the subset of 2,665 peptides that had sufficient data points to perform time-series modeling (observed in at least 5 time points and in at least 3 samples). I assessed whether the datasets could be modeled using a the three-exponential equation derived by Guan et al.

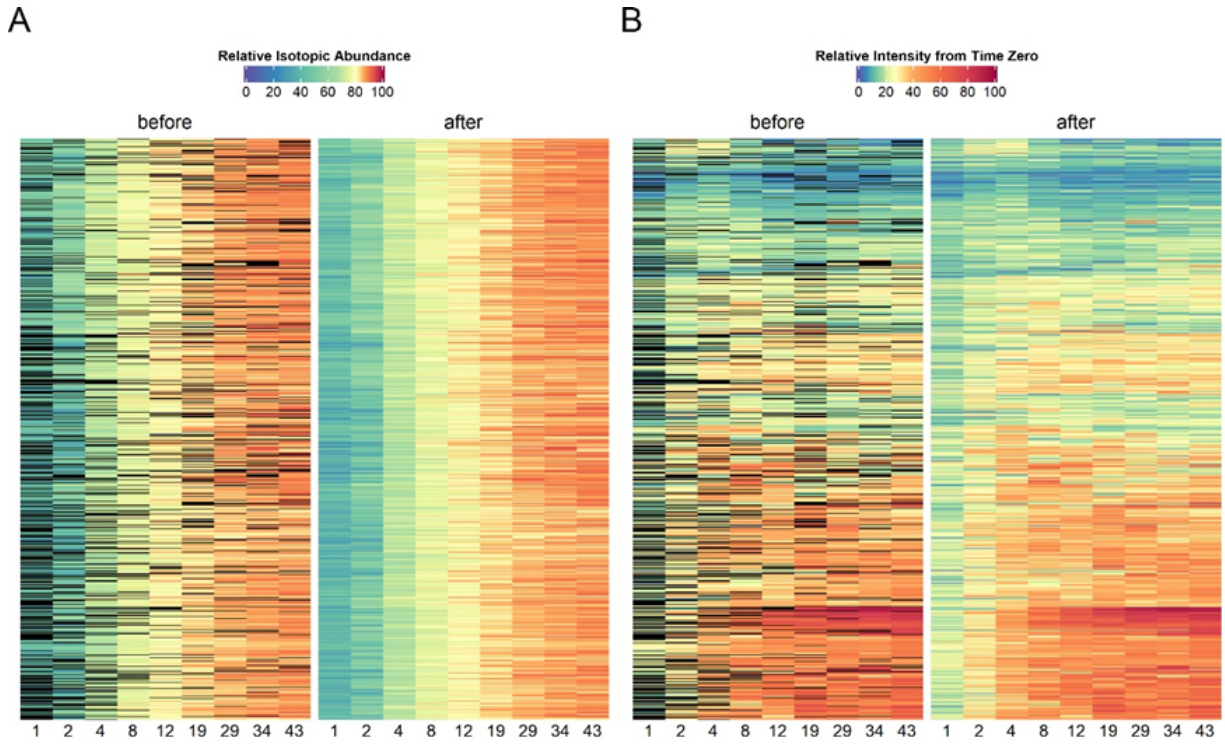
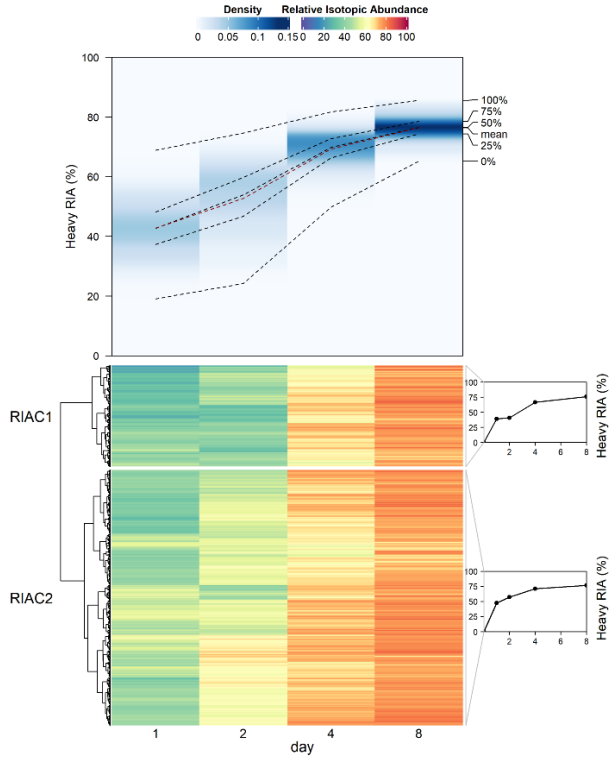

A
Relative Isotopic Abundance
0 20 40 60 80 100
before
after
1 2 4 8 12 19 29 34 43
1 2 4 8 12 19 29 34 43
B
Relative Intensity from Time Zero
0 20 40 60 80 100
before
after
1 2 4 8 12 19 29 34 43
1 2 4 8 12 19 29 34 43

Figure 3.8: Heatmap before and after imputation. A) A heatmap denoting the distribution of relative isotopic abundance (RIA) before and after imputing using the three exponential equation from Guan et al. (2012). The columns are the timepoints and the rows are the averaged RIA of the non-redundant peptides. The color of each cell corresponds to their RIA value, as illustrated by the color key above. B) Similar to A), but with relative intensity (RITZ) instead.

(2012) for computing protein turnover. This regression model predicted missing values with an average root mean square error (RMSE) of 7.27% for RITZ and 5.99% for RIA. The results before and after imputation are available in Figure 3.8.

The overall nitrogen flux was then evaluated in mice gut microbiome (Figure 3.9A). Nitrogen incorporation appeared to be rapid in the first few days but then slowed down (Figure 3.10B). Additionally, the vast majority of peptides did not reach an RIA of 100%, where the interquartile range was between 88.9% and 91.8% at day 43. To further examine whether there were differences in nitrogen incorporation rate of different peptides, I performed hierarchical clustering using the first four time points (day 1, 2, 4, and 8), where higher RIA variations were observed. Three distinct clusters (RIAC1, RIAC2, and RIAC3) were identified with peptides in cluster RIAC3 showing the fastest rate of isotope incorporation. Interestingly, a noticeable lag in incorporation was present in cluster RIAC1 from day 1 to day 2, and more subtlety in RIAC2. Hypergeometric tests (Table A.3) showed the species that are significantly over-represented in cluster RIAC1 are *Bacteroides vulgatus* ($p = 2 \times 10^{-3}$; BH procedure), host cells (*Mus musculus*; $p = 3 \times 10^{-4}$; BH procedure), and *Parabacteroides goldsteinii* ($p = 8 \times 10^{-4}$; BH procedure). To investigate this further, a pairwise t-test between taxa using their distinct heavy peptide RIA accompanied by BH procedure was performed to identify which species had significantly less nitrogen incorporation at day 2. From the tests, mouse cells had significantly less incorporation in day 2 than *Dorea sp. 5-2* (p-value=2×10^{-3}), and *Lachnospiraceae bacterium A4* (p-value=5×10^{-4}), and *Parabacteroides goldsteinii* had significantly less incorporation than *Dorea sp. 5-2* (p-value=3×10^{-4}), *Eubacterium plexicaudatum* (p-value=1×10^{-2}),

A


Density
Relative Isotopic Abundance
0 0.05 0.1 0.15
0 20 40 60 80 100
Heavy RIA (%)
100%
75%
50%
mean
25%
0%
RIAC1
RIAC2
Heavy RIA (%)
day
1 2 4 8

B

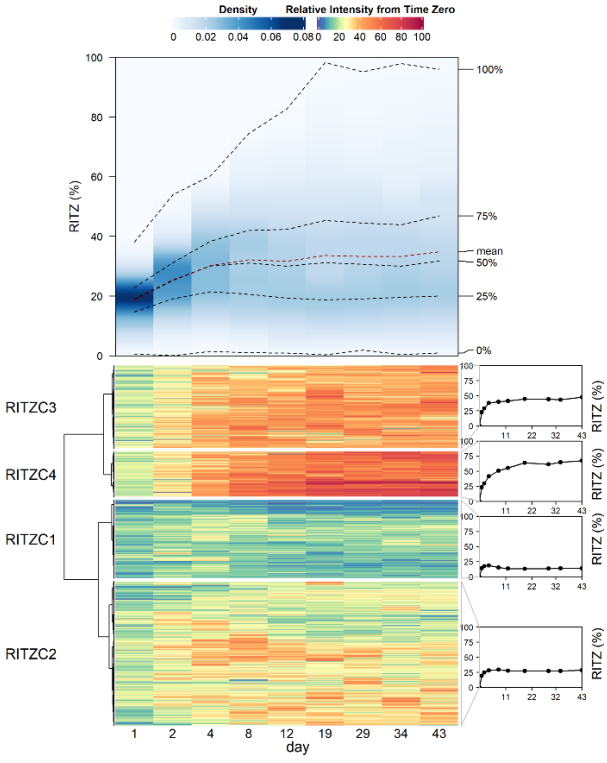

Density
Relative Intensity from Time Zero
0 0.02 0.04 0.06 0.08
0 20 40 60 80 100
RITZ (%)
100%
75%
mean
50%
25%
0%
RITZC3
RITZC4
RITZC1
RITZC2
RITZ (%)
day
1 2 4 8 12 19 29 34 43

Figure 3.9: Isotope Incorporation Profile. A) The top graph shows the distribution of RIA at each time point. Dark blue denotes a high density and light blue denotes a low density. At the bottom is a heatmap where the rows are peptides and the columns are the days at which the RIA was recorded. B) Similar to A except that it denotes relative intensity over time. Both heatmaps were clustered using Ward's minimum variance, as implemented in,74 using the Euclidean distance.

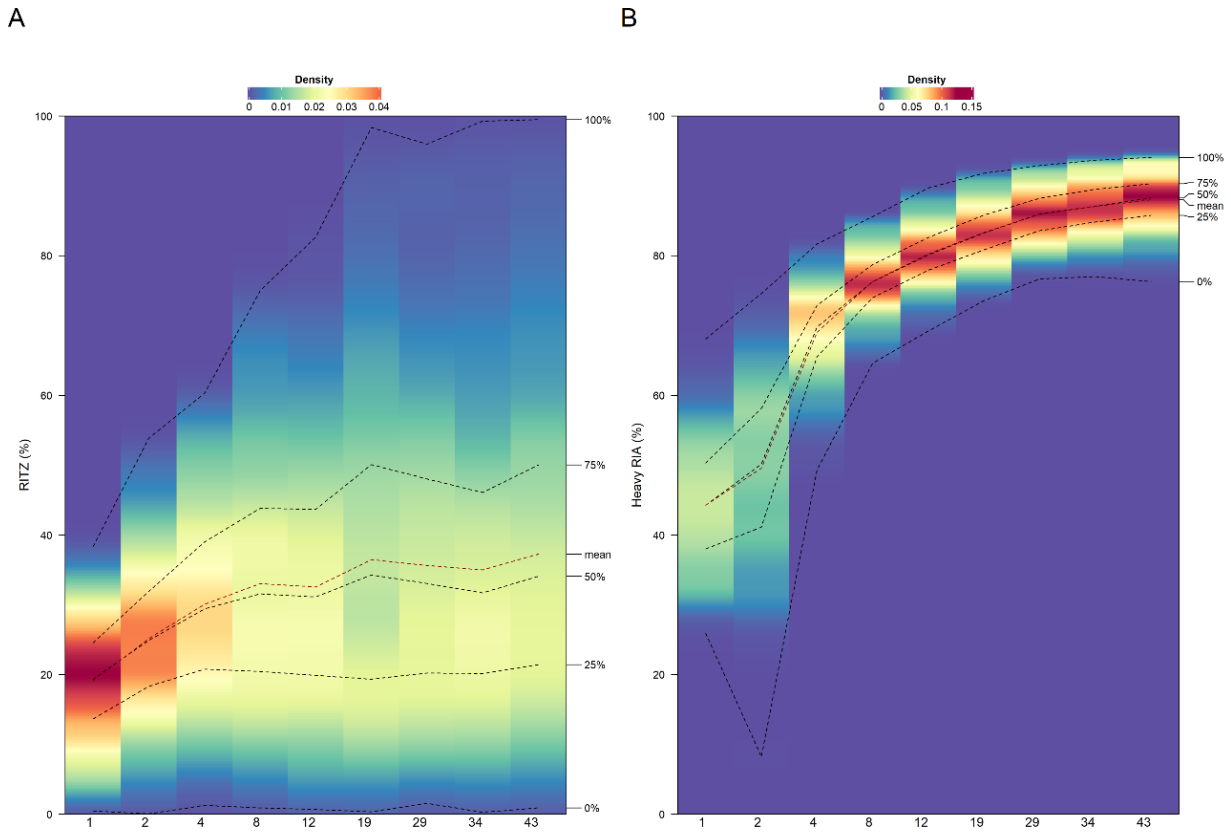

A
B
Density
0 0.01 0.02 0.03 0.04
Density
0 0.05 0.1 0.15
RITZ (%)
Heavy RIA (%)
100%
75%
mean
50%
25%
0%
1 2 4 8 12 19 29 34 43

Figure 3.10: RIA and RITZ Distribution of Heavy Peptides Below 10% LFDR Over Time. The density distribution of A) relative intensity (RITZ) and B) relative isotopic distribution (RIA) at day 1, 2, 4, 8, 12, 19, 29, 34, and 43. The black line represents the quantile of RIA and RITZ at 0%, 25%, 50%, 75%, and 100% and the red line represents the mean RIA and RITZ over time.

Lachnospiraceae bacterium 10-1 (p-value=9×10^{-3}), *Lachnospiraceae bacterium 28-4* (p-value=3×10^{-2}), *Lachnospiraceae bacterium A4* (p-value=2×10^{-5}), and *Oscillibacter sp. 1-3* (p-value=2×10^{-3}).

In order to group the different rates of newly synthesized proteins, hierarchical clustering of the RITZ over time was performed (Figure 3.9B). A total of four clusters were identified (LRC1, LRC2, LRC3, and LRC4). Peptides from cluster LRC1 showed the lowest RITZ (day 43 = 0.137 ± 0.005 at 95% CI), which became constant after day 4. Hypergeometric tests (Table A.4) showed that Firmicutes is significantly over-represented in this cluster (p-value = 4×10^{-5}; BH procedure). Cluster LRC2, LRC3, and LRC4 also remained stable until day 43 after the 4^{th} day of ^{15}N diet feeding, indicating an equilibrium state was reached by these peptides. Peptides from cluster LRC4 reached the highest RITZ ratio at day 43 (0.693 ± 0.015 at 95% CI). Hypergeometric tests showed that the taxa over-represented in cluster LRC4 are mouse cells (p-value = 6×10^{-49}; BH procedure) and *Parabacteroides goldsteinii* (p-value = 2×10^{-3}; BH procedure). This is consistent with Figure 3.6, where the majority of the mouse cells had high RITZ. Altogether, these findings suggest that the rates of newly synthesized proteins varied among different taxa.

3.3 ^{15}N incorporation rate differed according to microbial phylogeny

To further investigate whether the incorporation rate of nitrogen differed between microorganisms in mouse gut microbiome, I mapped all the above peptides to taxa and identified

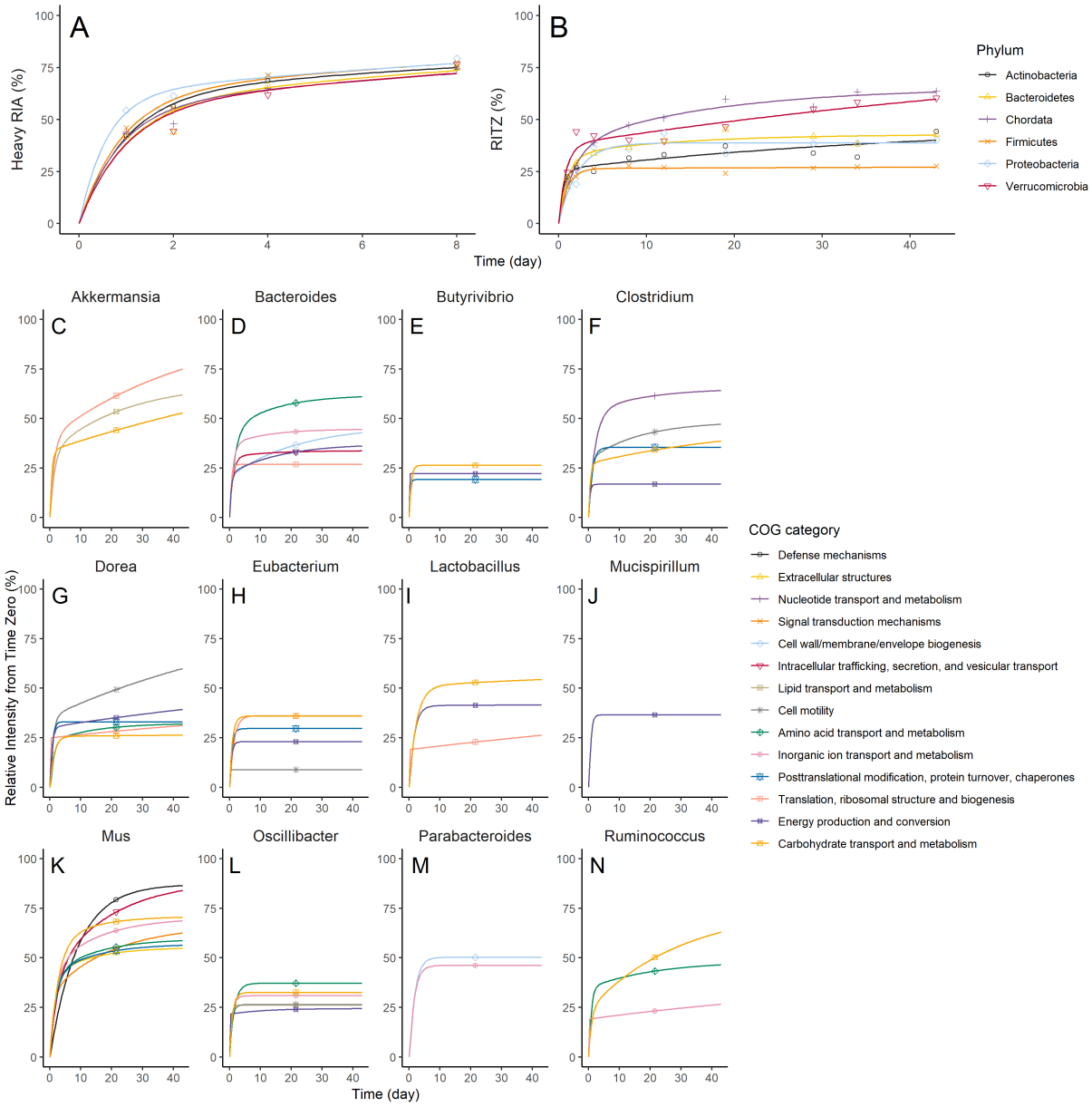

A
B
Heavy RIA (%)
RITZ (%)
Time (day)
Phylum
Actinobacteria
Bacteroidetes
Chordata
Firmicutes
Proteobacteria
Verrucomicrobia
Akkermansia
Bacteroides
Butyrivibrio
Clostridium
C
D
E
F
Dorea
Eubacterium
Lactobacillus
Mucispirillum
G
H
I
J
Relative Intensity from Time Zero (%)
Mus
Oscillibacter
Parabacteroides
Ruminococcus
K
L
M
N
Time (day)
COG category
Defense mechanisms
Extracellular structures
Nucleotide transport and metabolism
Signal transduction mechanisms
Cell wall/membrane/envelope biogenesis
Intracellular trafficking, secretion, and vesicular transport
Lipid transport and metabolism
Cell motility
Amino acid transport and metabolism
Inorganic ion transport and metabolism
Posttranslational modification, protein turnover, chaperones
Translation, ribosomal structure and biogenesis
Energy production and conversion
Carbohydrate transport and metabolism

Figure 3.11: Taxon-Specific RIA and RITZ Profiles Over Time. (A) phylum level RIA and (B) RITZ profiles over time; (C-N) Average COG RITZ profiles over time of the 12 most abundant genera.

3 superkingdom, 2 kingdom, 9 phylum, 12 class, 12 order, 15 family, 17 genus, and 21 species that have more than 3 distinctive peptides. RIA and RITZ of each taxon at each time point were then calculated using the average of all distinctive peptides (see Appendix A.1.1 for average RIA and RITZ over time for all taxa). At the phylum level, the RIA profiles over time were similar among all the identified seven phyla (Figure 3.11A), with the exception of host cells (Chordata) and Bacteroidetes between day 1 and 2. This further supports the hypergeometric tests performed in the clusters presented in Figure 3.9A. Obvious different RITZ profiles were also observed. The synthesis rate for Firmicutes clearly exhibited a similar pattern found in cluster LRC2 in Figure 3.9B, and reached a lower plateau than other abundant phyla, such as Bacteroidetes, Proteobacteria and the mice host. In agreement with the results obtained above, the mice proteins showed the highest RITZ after 4 days of ^{15}N diet. These findings suggest that the incorporation of ^{15}N to the host proteins is faster than the microbiome. Figure 3.11C-N shows the average RITZ overtime across all their COG categories for the most abundant genera.

3.4 Proteome Dynamics of *Arabidopsis thaliana* Seedling Roots

MetaProfiler can also be applied to datasets other than microbiome data and without protein spike-ins. Here, the dataset from Gajer et al. (2012) was applied to the same workflow and data analysis. The dataset includes soluble, microsomal, and organellar fractions obtained from *Arabidopsis thaliana* roots, where the samples were collected after

0, 4, 8, 16, 24, 32, 40, and 48 hours of feeding with ^{15}N labeled diet. It is important to note that two files are missing in the Pride repository (hour 16 for the soluble fraction and hour 48 for the microsomal fraction). The workflow generated 4,629 non-redundant heavy peptides from 5,086 peptides identified at time zero. In the original paper, they reported the incorporations 1845, 1998, and 6225 peptides in all the time points from the enriched soluble, organellar and microsomal fractions of Arabidopsis seedling roots, respectively. However, after removing peptides not present at time zero, not present in at least 3 time points, and not belonging to a protein with at least 2 unique peptides associated to them, they were left with 345, 321, and 798 peptides, respectively. By applying the same filtering strategy, as well as applying a correlation threshold of 0.6 and LFDR filtering at 10%, I obtained 662, 660, and 1,433 peptides.

The median LFDR for each time point collected were 99.5%, 72.3%, 10.5%, 2.9%, 1.3%, 1.6%, and 1.0% and the fitted joint distributions of RIA and RITZ are shown in Figure 3.12 for the soluble fraction, Figure 3.13 for the microsomal fraction, and Figure 3.14 for the organellar fraction. As seen from the figures, false discoveries were high in the first two timepoints, causing all the data from hour 4 to be lost and for most to be lost in hour 8. In the paper by Gajer et al. (2012), they mention that some peptides varied in LR in at hour 4, however they mention that they are no obvious outliers in their datasets. This emphasizes the importance of doing confidence assessments to ensure the quality of the biological interpretations of the results.

In order to group the different rates of protein turnover, hierarchical clustering of the LR over time was performed (Figure 3.15). A total of three clusters were identified (LRC1,

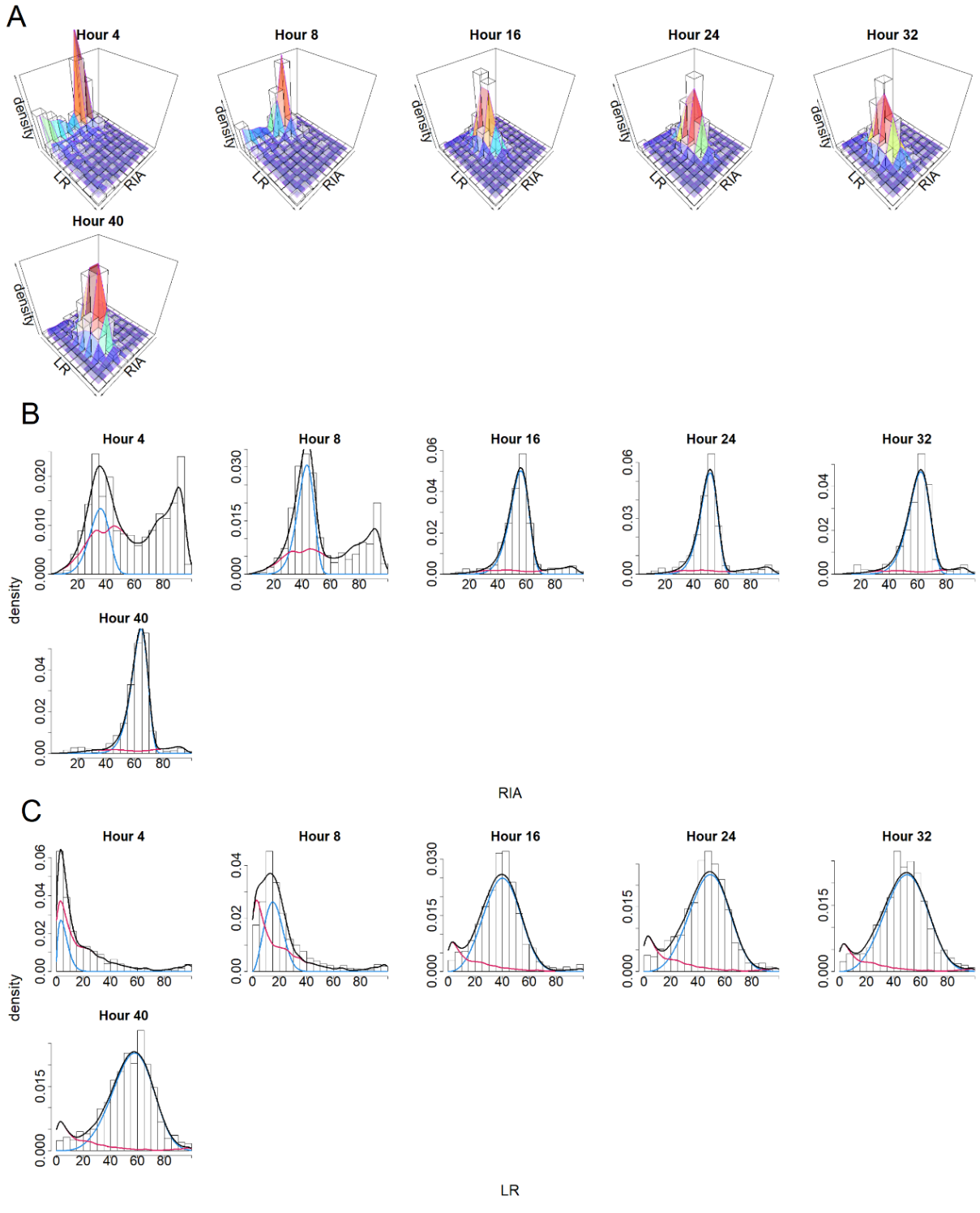

A
Hour 4
Hour 8
Hour 16
Hour 24
Hour 32
Hour 40
density
LR
RIA
B
Hour 4
Hour 8
Hour 16
Hour 24
Hour 32
Hour 40
density
RIA
C
Hour 4
Hour 8
Hour 16
Hour 24
Hour 32
Hour 40
density
LR

Figure 3.12: LR and RIA Distribution of Mixture, False, and True discoveries for the Soluble Fraction Sample. A) A bar plot representing the histogram of RIA and LR at each time point. The surface plot is the estimated joint density distribution calculated from the maximum likelihood estimate. B) and C) are the marginals from the of joint distribution where false discoveries is in red, mixture is in black, and true discoveries is in blue.

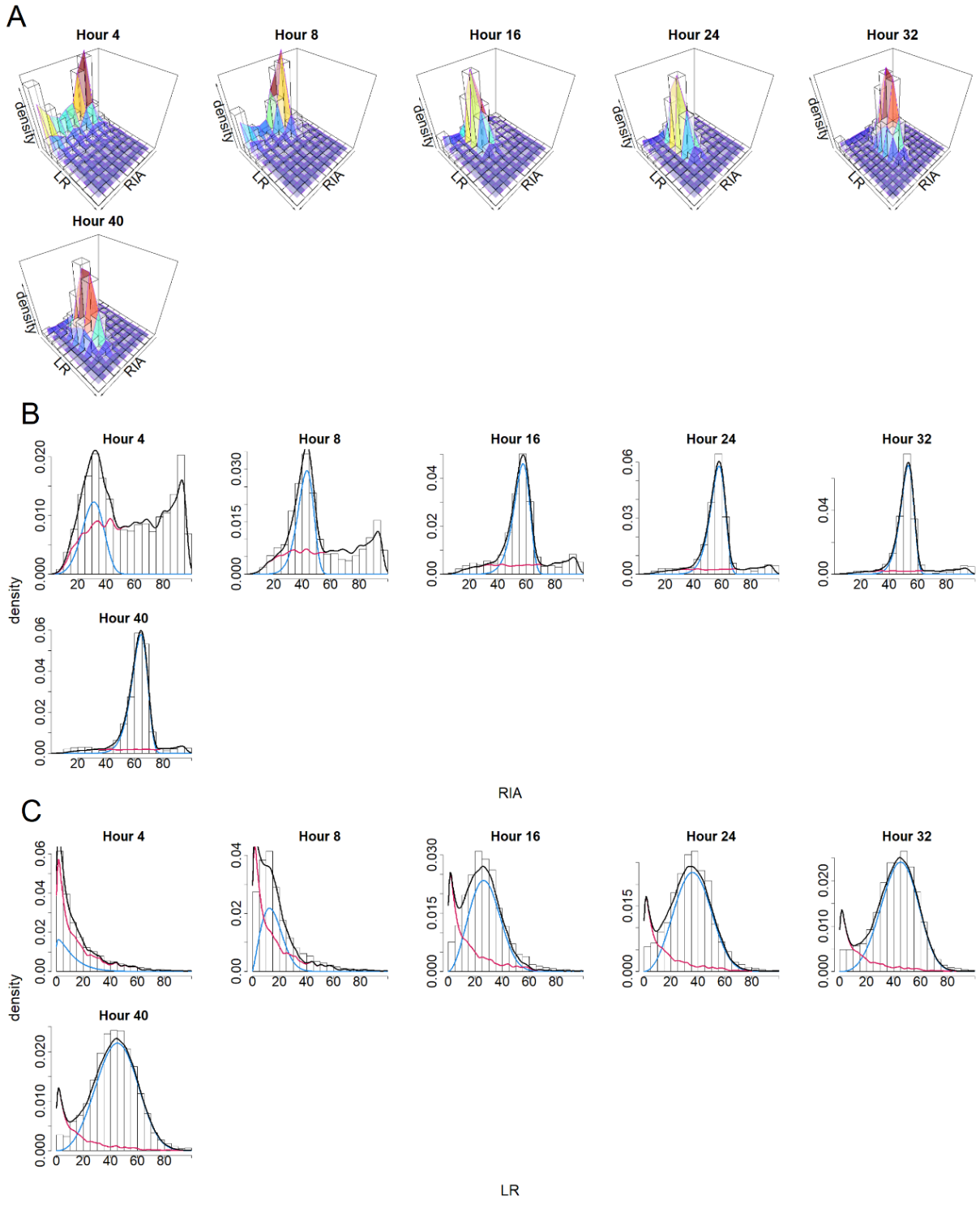

A
Hour 4
Hour 8
Hour 16
Hour 24
Hour 32
Hour 40
density
LR
RIA
B
Hour 4
Hour 8
Hour 16
Hour 24
Hour 32
Hour 40
density
RIA
C
Hour 4
Hour 8
Hour 16
Hour 24
Hour 32
Hour 40
density
LR

Figure 3.13: LR and RIA Distribution of Mixture, False, and True discoveries for the Microsomal Fraction Sample. A) A bar plot representing the histogram of RIA and LR at each time point. The surface plot is the estimated joint density distribution calculated from the maximum likelihood estimate. B) and C) are the marginals from the of joint distribution of RIA and LR where false discoveries is in red, mixture is in black, and true discoveries is in blue.

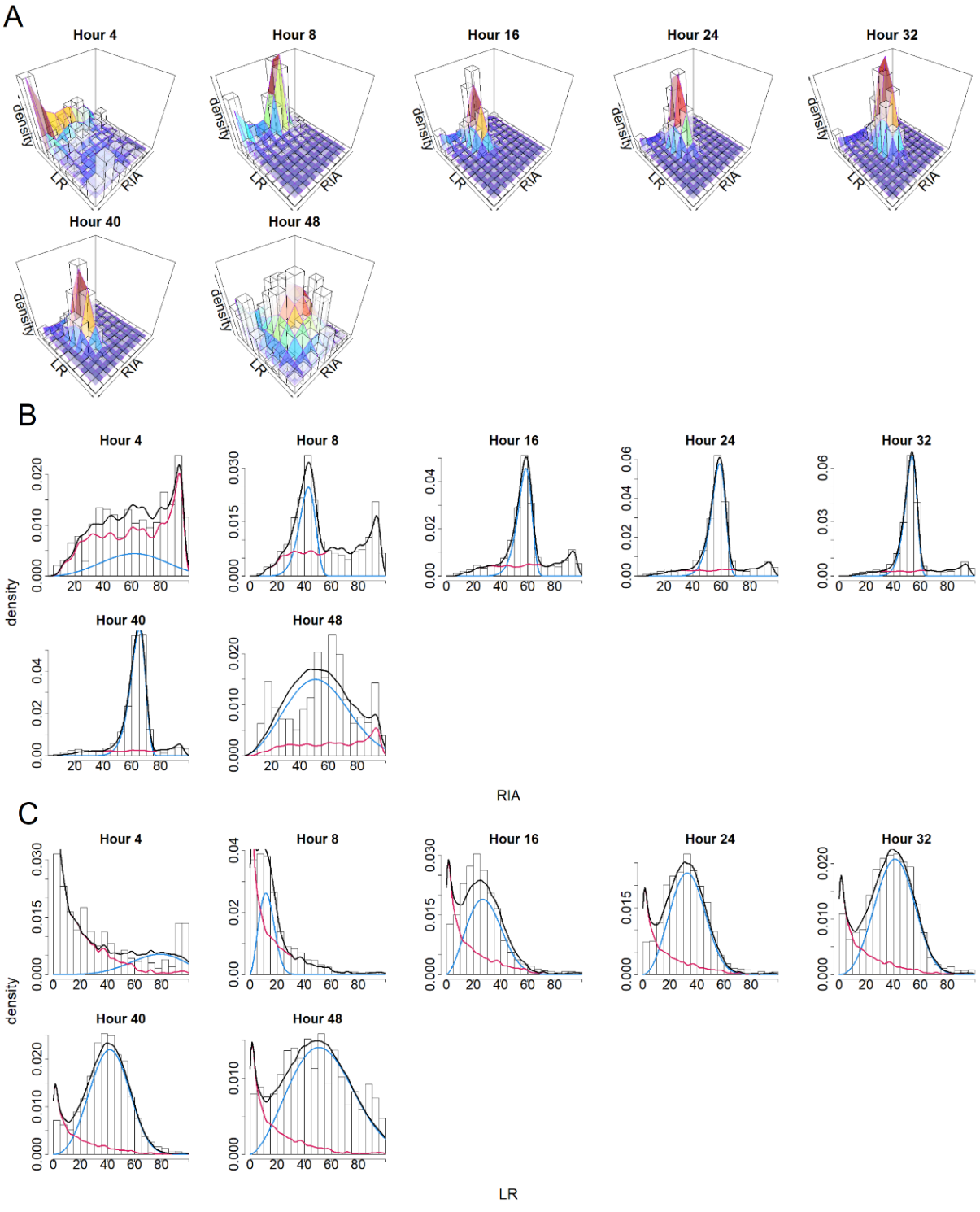

A
Hour 4
Hour 8
Hour 16
Hour 24
Hour 32
Hour 40
Hour 48
density
LR
RIA
B
Hour 4
Hour 8
Hour 16
Hour 24
Hour 32
Hour 40
Hour 48
density
RIA
C
Hour 4
Hour 8
Hour 16
Hour 24
Hour 32
Hour 40
Hour 48
density
LR

Figure 3.14: LR and RIA Distribution of Mixture, False, and True discoveries for the Organellar Fraction Sample. A) A bar plot representing the histogram of RIA and LR at each time point. The surface plot is the estimated joint density distribution calculated from the maximum likelihood estimate. B) and C) are the marginals from the of joint distribution where false discoveries is in red, mixture is in black, and true discoveries is in blue.

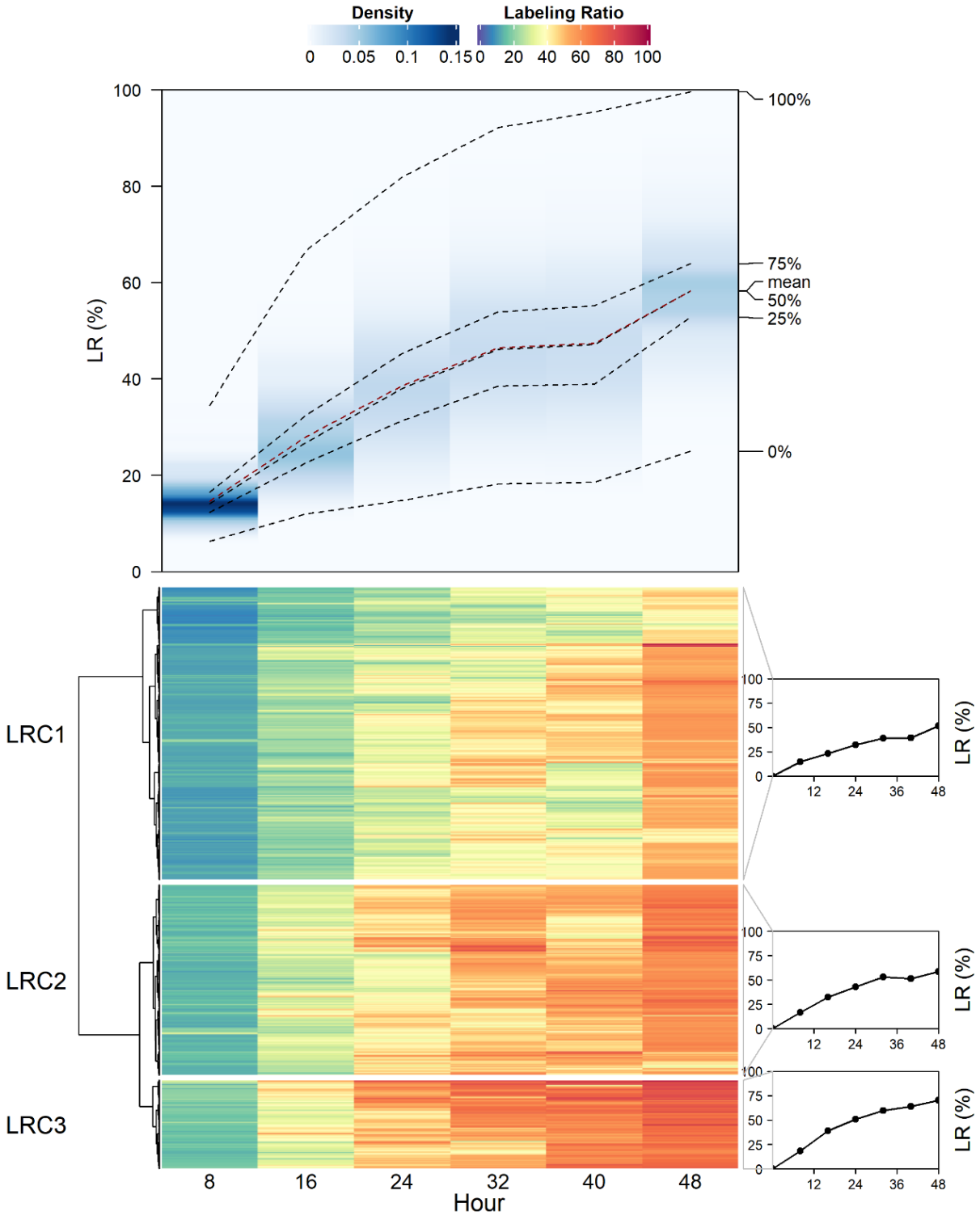

Density
0 0.05 0.1 0.15
Labeling Ratio
0 20 40 60 80 100
LR (%)
100
80
60
40
20
0
100%
75%
mean
50%
25%
0%
LRC1
LRC2
LRC3
8 16 24 32 40 48
Hour
LR (%)
12 24 36 48

Figure 3.15: Isotope Incorporation Profile of Arabidopsis thaliana Seedling Roots. A) The top graph shows the distribution of RIA at each time point. Dark blue denotes a high density and light blue denotes a low density. At the bottom is a heatmap where the rows are peptides and the columns are the days at which the RIA was recorded. B) Similar to A except that it denotes relative intensity over time. Both heatmaps were clustered using Ward's minimum variance, as implemented in,74 and using the Euclidean distance.

LRC2, and LRC3), where cluster LRC1 had the lowest rate. Hypergeometric tests showed that the organellar fraction is significantly over-represented in this cluster (p-value = 1×10^{-7}; BH procedure). Peptides from cluster LRC3 displayed the highest rate from the three clusters and hypergeometric tests showed that the soluble fraction sample is over-represented in LRC3 (p-value = 4×10^{-9}; BH procedure). As for LRC2, the microsomal fraction is significantly over-represented in this cluster (p-value = 1×10^{-7}; BH procedure). These results are all in agreement with Fan et al. (2016).

Of the 251 proteins analyzed in Fan et al. (2016) 248 proteins were observed from the results generated by our workflow. The coefficient of variation of the RMSD between the log2(k) values of the two sets of results is 7% and the Pearson correlation coefficient is 0.84. In addition to these proteins, the incorporation of 743 proteins were also analyzed, making a total of 991 unique proteins, The fitted curves for the three samples is shown in Figure 3.16A, as well as the fitted curves for the 12 proteins with the most peptides associated to them (Figure 3.16B-M).

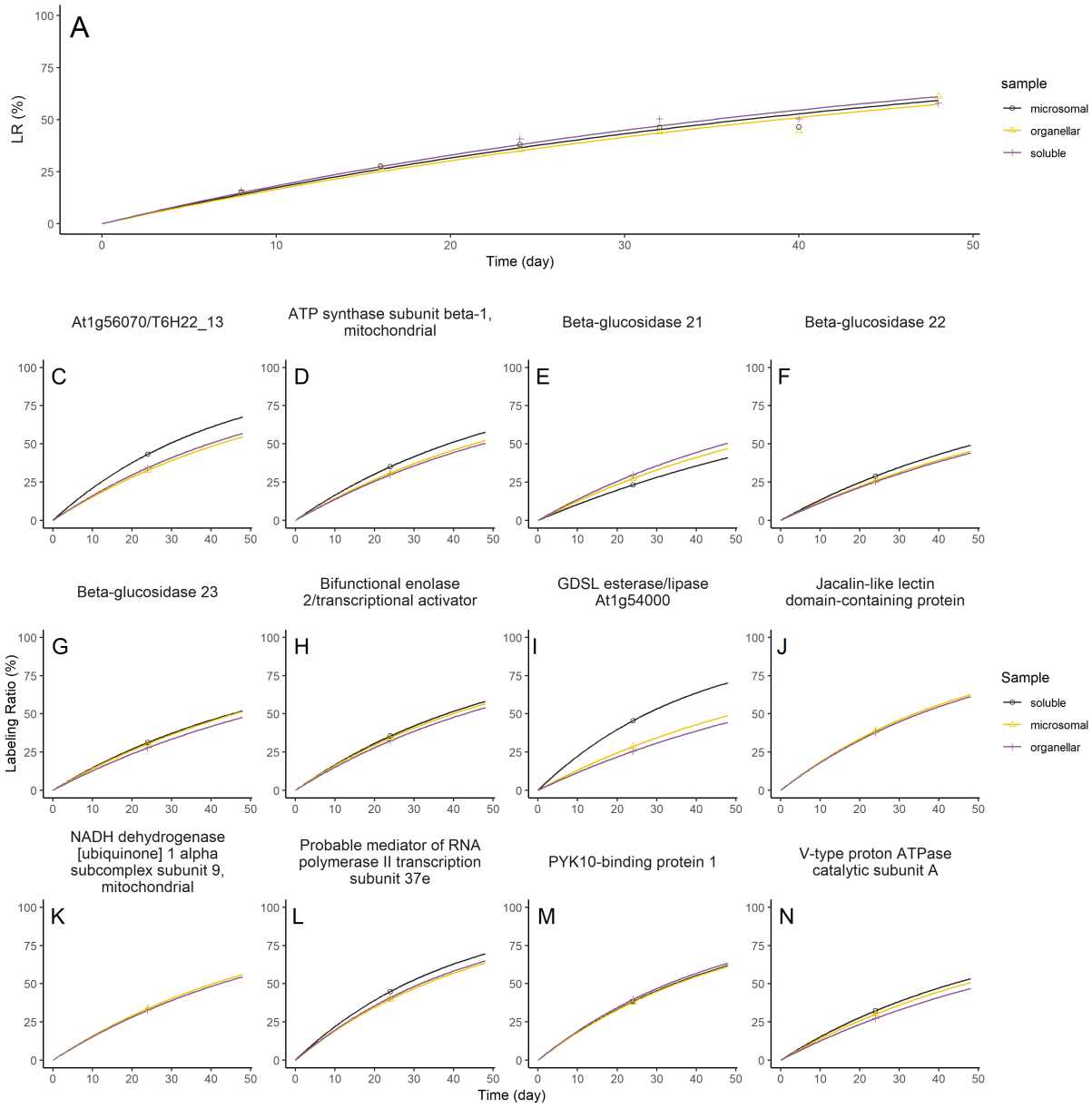

A
LR (%)
Time (day)
sample
microsomal
organellar
soluble
At1g56070/T6H22_13
ATP synthase subunit beta-1, mitochondrial
Beta-glucosidase 21
Beta-glucosidase 22
C
D
E
F
Beta-glucosidase 23
Bifunctional enolase 2/transcriptional activator
GDSL esterase/lipase At1g54000
Jacalin-like lectin domain-containing protein
G
H
I
J
Labeling Ratio (%)
Sample
soluble
microsomal
organellar
NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 9, mitochondrial
Probable mediator of RNA polymerase II transcription subunit 37e
PYK10-binding protein 1
V-type proton ATPase catalytic subunit A
K
L
M
N
Time (day)

Figure 3.16: LR Profiles Over Time. (A) the LR profiles for each sample; (B-M) Average sample LR profiles over time for the 12 proteins with the most peptides associated to them

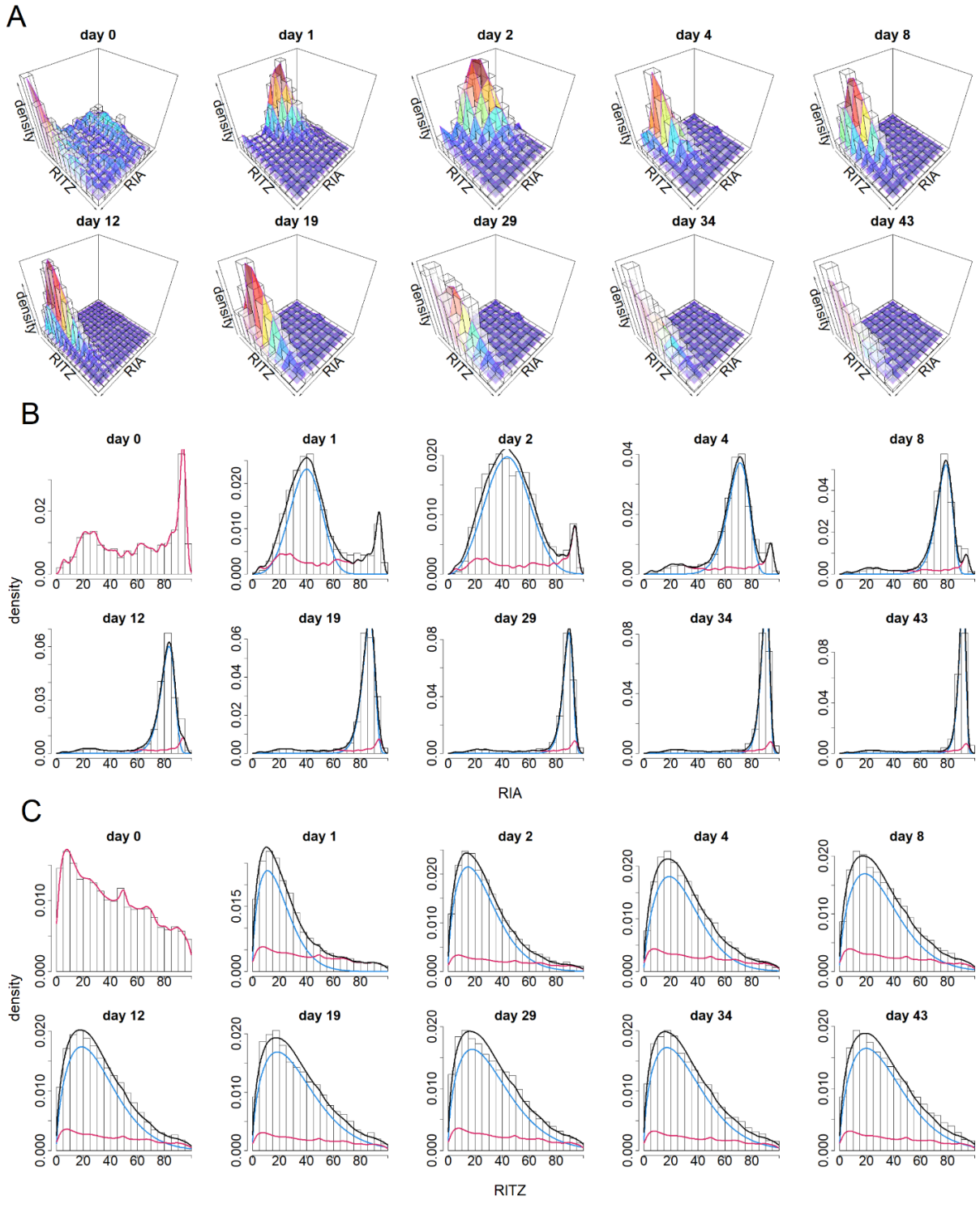

A
day 0
day 1
day 2
day 4
day 8
day 12
day 19
day 29
day 34
day 43
density
RITZ
RIA
B
density
RIA
C
density
RITZ

Figure 3.17: RITZ and RIA Distribution of Mixture, False, and True discoveries. A) A bar plot representing the histogram of RIA and RITZ at each time point. The surface plot is the estimated joint density distribution calculated from the maximum likelihood estimate. B) and C) are the marginals from the of joint distribution where false discoveries is in red, mixture is in black, and true discoveries is in blue.

Chapter 4

Discussion

I investigated the *in vivo* dynamics of the microbiome using metaproteomics coupled with protein-SIP. Our experiment differs from previous protein-sip studies(Oberbach et al., 2017; Fan et al., 2016; Arike et al., 2020) by the addition of a protein spike-in to our labeled sample. One disadvantage of a protein spike-in is that it would mask the rate of disappearance of the light peptide(Sachsenberg et al., 2015), thus making protein turnover estimation difficult. However, it is arguable that this approach can be used to measure protein turnover in the gut microbiome. Many factors such as elimination, de novo synthesis, protein recycling, etc will influence the observed abundance of the light and heavy peptide. Also, without a spike-in, very few peptides are observed as demonstrated by Oberbach et al. (2017). Thus, since our experiment mostly concerns with the general activity of the microbiome, I instead estimated the rate of newly synthesized (i.e. heavy) peptides by using the light peptide as a form of internal standard.

One challenge when dealing with ^{15}N labeled proteomic data, from the microbiome in

particular, is the difficulty in confidently identifying peptides from MS/MS spectra due to the partial incorporation of ^{15}N(Gajer et al., 2012; Egert et al., 2007). To address this issue, unlabeled microbiome samples (time = 0) were spiked in every partially labeled sample to ensure the presence of the corresponding reference light peptide. This allowed to achieve a greater number of peptides identified than the previous research by Oberbach et al. (2017).

Sachsenberg et al. (2015) were the first to introduce confidence assessment in protein-SIP experiments. In their paper, they estimated the quality of the results by calculating false positive rate (FPR) using contaminants or unlabeled reference samples. FPR is the expected proportion of all the true false positives in a study population selected as hits, whereas FDR is the expected proportion of true false positives among declared hits. The number of true false positives in a study population is typically unknown, whereas the number of declared as hits is known. Therefore, FPR is not typically verifiable and of less interest relative to FDR, which is verifiable. One can similarly use the scores from the contaminants and the scores from the observed values to estimate FDR. Unfortunately, when testing for hundreds or thousands of hypotheses, the distinction between true and false discoveries become blurred in an overall false-discovery rate(Verheggen et al., 2020; Zhang et al., 2017). An empirical Bayes model, on the other hand, allows for an error rate to be attributed for each selected feature. Thus, the strength of the evidence between each test is not lost in local false discovery rate. Another advantage is that LFDR can use multiple parameters, which increases power(Korthauer et al., 2019).

Since the median LFDR for day 1, 2, 4, 8, 12, 19, 29, 34, and 43 are 11.81%, 8.7%, 5.64%, 4.71%, 4.67%, 4.37%, 5.48%, 4.8%, and 4.61%, respectively, the threshold was set

at 10% so that earlier timepoints would not be filtered out completely. The reason for LFDR being higher in earlier timepoints is because low RIAs are more likely to be false as MetaProSIP may sometimes mistake the light peptide feature as being overlapped with the heavy peptide feature at time zero. This is why a higher density of RIA is observed for false discoveries in earlier timepoints in Figure 3.17B. In addition, the quantified features that are the result of experimental noise will often be low in RITZ as seen in Figure 3.17B.

Plateau detection was done using a rolling standard deviation, where a window of fixed length moves over the RITZ of each peptide and then computes the standard deviation of the RITZ within the window. Since windows with larger start to end gaps will tend to have smaller deviations and at least two observations are needed to compute standard deviation, nine days were chosen as the length. This span has the smallest root mean square deviation between the width of each window, compared to other spans. The Wilcoxon tests suggest that the plateau starts at day 29 and the RITZs past this day reflects the maximum proportion of the labeled proteins that the taxon/function can attain when using hydrolysate as a nitrogen source. It is important to note that several studies have reported that ^{15}N metabolic labeling can effect certain cells, causing certain genes to be under or over expressed.(Filiou et al., 2012; Webhofer et al., 2013; Millard et al., 2015) To assess the contribution of the ^{15}N isotope effect, I compared our dataset against the dataset by Webhofer et al.,(Webhofer et al., 2013) where they analysed blood plasma proteins from mice using the same diet as this study. However, the mice were fed *in utero* and for 56 days after birth, instead. The results from the comparison suggests that diet-induced compositional changes in microbiota may affect RITZ. In addition to the isotope effect

from ^{15}N, factors such as nitrogen uptake, de novo synthesis, protein recycling, and internal storage will also affect heavy protein abundance in this study.

As shown in Figure 3.2, proteins from mice (Chordata) were within the highest peptide relative intensities (0.572 ± 0.015 at 95% CI). This is not particularly surprising as dietary proteins are the most common source of nitrogen for mammals. Firmicutes, on the other hand, consistently showed lower relative intensities. A possible explanation is that they do not process hydrolysate efficiently and may obtain their nitrogen source elsewhere, such as from fiber(Zhang et al., 2018a) or metabolites from other microorganism in the microbial community. Nitrogen fixation could also provide another source of unlabeled nitrogen; however, very few mammalian gut microbes are capable of nitrogen fixation(Berthold et al., 2009; Alka et al., 2019). It is more likely that the microbiota cells obtain light amino acids via host protein degradation, and *de novo* synthesis(Gajer et al., 2012). Overtime, the nitrogen pool will gradually grow towards heavy isotopes as the host increasingly incorporates the labeled heavy nitrogen. Lactic acid bacteria were notably the exception from the phyla Firmicutes as they are among the highest in RITZ. This suggests potential differences in nutritional mechanisms of lactic acid bacteria with other Firmicutes species. Functions of "carbohydrate transport and metabolism" and "amino acid transport and metabolism", "energy production and conversion", as well as "translation, ribosomal structure, and biogenesis" were among the categories with the highest number of heavy peptide associated to them. The composition of functions was similar to previous metaproteomic studies on the gut microbiome(Weisser and Choudhary, 2017), indicating that the labeling might not be biased to specific functional categories.

The RITZ of proteins belonging to the functional group "amino acid transport and metabolism" corresponded well with the overall RITZ of heavy peptides in each taxon (Pearson's correlation coefficient = 0.80). This makes sense as those which efficiently degrade hydrolysate will generally see a higher level of heavy peptides. However, there were a few taxa that displayed high overall levels of RITZ without having peptides belonging to this category. In particular, *Akkermansia muciniphila*, showed high RITZ in the functional group "carbohydrate transport and metabolism" (0.558 ± 0.108 at 95% CI; Figure 3.11C), which was consistent with its mucin degrading ability. Another taxon which showed this pattern is *Lactobacillus* (0.595 ± 0.048 at 95% CI; Figure 3.11J). Although not as well characterized, several species of this lineage have been shown to possess proteins that degrade mucin(Chacón and Duong, 2018). In addition, both *Akkermansia* and *Lactobacillus* were high in "translation, ribosomal structure and biogenesis" (0.667 ± 0.081 at 95% CI) and 0.462 ± 0.150 at 95% CI), respectively), which indicates that both are rapid growers.

By studying RIA over time, I analyzed the elemental flux of nitrogen in the gut and the results shows that its availability is consistent to most species. However, there were a few species and strains that showed a delay in nitrogen availability. In particular, certain strains of *Parabacteroides goldsteinii* had no increase in heavy nitrogen availability after day 1. Cells from the host also exhibited this pattern. However, once heavy nitrogen became readily available to them, these strains started to catch up with the earlier incorporators. When analyzing the RITZ over time, on the other hand, it was shown that Verrucomicrobia had a characteristic increasing trend and never fully reached a plateau over 43 days. This is also observed in the genus, *Akkermansia* (Figure 3.11C), for cate-

gory "translation, ribosomal structure and biogenesis" and "carbohydrate transport and metabolism". A similar pattern was observed for *Lactobacillus*(Figure 3.11J) in these categories as well, which further provided support of their mucin degrading capabilities. It is possible that these taxa may play important roles in foraging nitrogen from the host to the microbiome. Thus, our findings may suggest that these taxa, especially *A. muciniphila*, are keystone species for the metabolic flux of nitrogen in the microbiome.

For this study I retained peptides that were not identified at time zero. It is possible that peptides identified at time zero were the result of MS/MS from heavy peptides being wrongfully assigned to a light peptide. One way that one can test the likelihood of these types of misassignments is by searching light peptides against a fully ^{15}N labeled database. Since there were nine times more ^{15}N peptides, I took the sample specific database generated from MetaLab and added randomly selected proteins from the large metaproteome database until the size became nine times larger than its original size. I then added ^{15}N labeled proteins via fixed modifications and searched this database with the day zero samples. The resulting number of peptide identifications were seven at an FDR threshold of 1% (PSM level).

In the case that a heavy peptide would be assigned with a light peptide from MS/MS, they will likely get filtered out: 1) when removing results that do not have two RIAs: one below and above $^{15}N_{\text{max}}{}^{-1}$, 2) when running MetaProSIP as peaks with a Pearson correlation lower than 0.6 are removed, and 3) when filtering via LFDR as they will likely have high local false from the distorted RIA and LR values.

I also investigated as to why some light peptides were not identified at time zero but

identified in subsequent days. One possibility is that MS/MS were not generated at time zero for these peptides, which would be consistent with the results from Table 3.1. Namely, peptides identified in the subsequent days were too low to be selected for MS2 fragmentation at time zero, but due to having lower protein turnover, their peak intensities did not reduce as much as the other peptides and were able to be selected for MS2 fragmentation. This would also mean that, eventually, their intensities will become too low again to be selected for MS2 fragmentation. One can see this pattern in the number of peptide identifications as well (Table 3.2). where from day 1 to day 2, you see an increase in the number of ids for non-time zero peptides, while you get ˜2000 less ids for time zero peptides. Eventually, however, we get ˜50% less non-time zero peptide ids on day 43 than on day 1, while for time zero, we get 30% less. This also explains why they were mostly found in RITZC1 due to their slow protein turnover. RIA was significantly lower as well for these peptides from day 4 to day 43 compared to peptides found at time zero (Table 3.3).

Chapter 5

Conclusions

We investigated the *in vivo* dynamics of the microbiome using metaproteomics coupled with protein-SIP. Using dataset with light protein spike-in, I developped a new bioinformatics pipeline and a new bioinformatic tool called MetaProfiler which greatly enhanced the number of peptides identified (15,297), and quantified (10,173; 7,108 with LFDR $< 10\%$) with ^{15}N incorporations. I used a novel approach to assess confidence of protein-SIP and I used machine learning algorithms to conduct the most in-depth analysis of mice gut microbiome dynamic study using protein-SIP, at the time of this study. I believe that the outcome of this study will allow more thorough research in microbiome dynamics using MetaProfiler and provide deeper insight into the functionality of the gut microbiome.

Chapter 6

Future Work

One of the main limitations of using mouse models is that they cannot replicate human systems. Genetic variations, medical history, lifestyle choices, and so on, are important factors to consider in microbiome dynamics which are not taken into consideration when using mouse models. Therefore, future work in microbiome experiments using protein-SIP would need to incorporate *in vitro* human models. In addition, culturomics would eliminate factors that would affect light and heavy protein intensities over time such as elimination, de novo synthesis from neighboring species, host internal storage, and so on. Thus, we would be able to better understand the mechanisms affecting the proteome dynamics of the gut microbiome. While several studies are starting to look at the effects of therapeutic treatments on gut microbiome(Maier et al., 2018; Chen et al., 2017; Le Bastard et al., 2018; Li et al., 2020), no study so far has investigated the effects of these treatments on microbiome dynamics using protein-SIP. With the ever-increasing number of treatments, there is a pressing need in developing high-throughput methods to assess drug-microbiome

dynamics interactions

References

Aggarwal, K., Choe, L. H. and Lee, K. H. (2006). Shotgun proteomics using the iTRAQ isobaric tags. Briefings in Functional Genomics *5*, 112–120.

Alka, O., Sachsenberg, T., Bichmann, L., Pfeuffer, J., Weisser, H., Wein, S., Netz, E., Rurik, M., Kohlbacher, O. and Rost, H. (2019). OpenMS for open source analysis of mass spectrometric data. PeerJ Preprints *7*, e27766v1.

Arike, L., Seiman, A., van der Post, S., Piñeiro, A. M. R., Ermund, A., Schütte, A., Bäckhed, F., Johansson, M. E. and Hansson, G. C. (2020). Protein turnover in epithelial cells and mucus along the gastrointestinal tract is coordinated by the spatial location and microbiota. Cell reports *30*, 1077–1087.

Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ *3*, e1029.

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H. et al. (2020). Microbiome definition re-visited: old concepts and new challenges. Microbiome *8*, 1–22.

Berry, D. and Loy, A. (2018). Stable-isotope probing of human and animal microbiome function. Trends in microbiology *26*, 999–1007.

Berry, D., Stecher, B., Schintlmeister, A., Reichert, J., Brugiroux, S., Wild, B., Wanek, W., Richter, A., Rauch, I., Decker, T. et al. (2013). Host-compound foraging by intestinal microbiota revealed by single-cell stable isotope probing. Proceedings of the National Academy of Sciences *110*, 4720–4725.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K. and Wiswedel, B. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter *11*, 26–31.

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N. et al. (2011). Moving pictures of the human microbiome. Genome biology *12*, 1–8.

Chacón, J. E. and Duong, T. (2018). Multivariate kernel smoothing and its applications. CRC Press.

Charrad, M. and Ghazzali, N. (2014). Package 'nbclust'. Journal of statistical software *61*, 1–36.

Chen, T., Li, J., Chen, T., Sun, C. C. and Zheng, Y. (2017). Tablets of multi-unit pellet system for controlled drug delivery. Journal of Controlled Release *262*, 222–231.

Cheng, K., Ning, Z., Zhang, X., Li, L., Liao, B., Mayne, J., Stintzi, A. and Figeys, D. (2017). MetaLab: an automated pipeline for metaproteomic data analysis. Microbiome *5*, 1–10.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X. et al. (2016). A survey of best practices for RNA-seq data analysis. Genome biology *17*, 13.

Cox, J. and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol *26*, 1367–72.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V. and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. Journal of proteome research *10*, 1794–1805.

Eddelbuettel, D. (2010). RcppDE: Global optimization by differential evolution in C++. R package version 0.1. 0, URL http://CRAN.R-project.org/package=RcppDE .

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. Journal of the American statistical association *96*, 1151–1160.

Egert, M., De Graaf, A. A., Maathuis, A., De Waard, P., Plugge, C. M., Smidt, H., Deutz, N. E., Dijkema, C., De Vos, W. M. and Venema, K. (2007). Identification of glucose-fermenting bacteria present in an in vitro model of the human intestine by RNA-stable isotope probing. FEMS microbiology ecology *60*, 126–135.

Eng, J. K., McCormack, A. L. and Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the american society for mass spectrometry *5*, 976–989.

Fan, K.-T., Rendahl, A. K., Chen, W.-P., Freund, D. M., Gray, W. M., Cohen, J. D. and Hegeman, A. D. (2016). Proteome scale-protein turnover analysis using high resolution mass spectrometric data from stable-isotope labeled plants. Journal of proteome research *15*, 851–867.

Filiou, M. D., Varadarajulu, J., Teplytska, L., Reckow, S., Maccarrone, G. and Turck, C. W. (2012). The 15N isotope effect in Escherichia coli: A neutron can make the difference. Proteomics *12*, 3121–3128.

Flores, G. E., Caporaso, J. G., Henley, J. B., Rideout, J. R., Domogala, D., Chase, J., Leff, J. W., Vázquez-Baeza, Y., Gonzalez, A., Knight, R. et al. (2014). Temporal variability is a personalized feature of the human microbiome. Genome biology *15*, 531.

Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., Vatanen, T., Hall, A. B., Mallick, H., McIver, L. J., Sauk, J. S., Wilson, R. G., Stevens, B. W., Scott, J. M., Pierce, K., Deik, A. A., Bullock, K., Imhann, F., Porter, J. A., Zhernakova, A., Fu, J., Weersma, R. K., Wijmenga, C., Clish, C. B., Vlamakis, H., Huttenhower, C. and Xavier, R. J. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol *4*, 293–305.

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M., Zhong, X., Koenig, S. S., Fu, L., Ma, Z. S., Zhou, X. et al. (2012). Temporal dynamics of the human vaginal microbiota. Science translational medicine *4*, 132ra52–132ra52.

Gu, Z., Eils, R. and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849.

Guan, S., Price, J. C., Ghaemmaghami, S., Prusiner, S. B. and Burlingame, A. L. (2012). Compartment modeling for mammalian protein turnover studies by stable isotope metabolic labeling. Analytical chemistry *84*, 4014–4021.

Heintz-Buschart, A., May, P., Laczny, C. C., Lebrun, L. A., Bellora, C., Krishna, A., Wampach, L., Schneider, J. G., Hogan, A., De Beaufort, C. et al. (2016). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nature microbiology *2*, 1–13.

Heintz-Buschart, A. and Wilmes, P. (2018). Human gut microbiome: function matters. Trends in microbiology *26*, 563–574.

Helmink, B. A., Khan, M. A. W., Hermann, A., Gopalakrishnan, V. and Wargo, J. A. (2019). The microbiome, cancer, and cancer therapy. Nat Med *25*, 377–388.

Herrmann, E., Young, W., Rosendale, D., Conrad, R., Riedel, C. U. and Egert, M. (2017a). Determination of resistant starch assimilating bacteria in fecal samples of mice by in vitro RNA-based stable isotope probing. Frontiers in microbiology *8*, 1331.

Herrmann, E., Young, W., Rosendale, D., Reichert-Grimm, V., Riedel, C. U., Conrad, R. and Egert, M. (2017b). RNA-based stable isotope probing suggests Allobaculum spp. as particularly active glucose assimilators in a complex murine microbiota cultured in vitro. BioMed Research International *2017*.

Jagtap, P. D., Blakely, A., Murray, K., Stewart, S., Kooren, J., Johnson, J. E., Rhodus, N. L., Rudney, J. and Griffin, T. J. (2015). Metaproteomic analysis using the Galaxy framework. Proteomics *15*, 3553–3565.

Jehmlich, N., Vogt, C., Lunsmann, V., Richnow, H. H. and von Bergen, M. (2016). Protein-SIP in environmental studies. Curr Opin Biotechnol *41*, 26–33.

Kim, H. J., Li, H., Collins, J. J. and Ingber, D. E. (2016). Contributions of microbiome and mechanical deformation to intestinal bacterial overgrowth and inflammation in a human gut-on-a-chip. Proceedings of the National Academy of Sciences *113*, E7–E15.

Klaassen, C. D. and Aleksunes, L. M. (2010). Xenobiotic, bile acid, and cholesterol transporters: function and regulation. Pharmacol Rev *62*, 18 – 23.

Komaroff, A. L. (2017). The Microbiome and Risk for Obesity and Diabetes. JAMA *317*, 355–356.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J. and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. Genome biology *20*, 1–21.

Kovatcheva-Datchary, P. and Egert, M. (2009). M., M. Rajilić-Stojanović, A. de Graaf, H. Smidt, W. de Vos and K. Venema, Linking phylogenetic identities of bacteria to starch fermentation in an in vitro model of the large intestine by RNA-based stable isotope probing. Environ. Microbiol *11*, 914–926.

Lagier, J.-C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., Levasseur, A., Rolain, J.-M., Fournier, P.-E. and Raoult, D. (2018). Culturing the human microbiota and culturomics. Nature Reviews Microbiology *16*, 540–550.

Lavelle, A. and Sokol, H. (2018). Gut microbiota: beyond metagenomics, metatranscrip-

tomics illuminates microbiome functionality in IBD. Nature Reviews Gastroenterology & Hepatology *15*, 193–194.

Le Bastard, Q., Al-Ghalith, G., Grégoire, M., Chapelet, G., Javaudin, F., Dailly, E., Batard, E., Knights, D. and Montassier, E. (2018). Systematic review: human gut dysbiosis induced by non-antibiotic prescription medications. Alimentary Pharmacology & Therapeutics *47*, 332–345.

Li, L., Ning, Z., Zhang, X., Mayne, J., Cheng, K., Stintzi, A. and Figeys, D. (2020). RapidAIM: A culture-and metaproteomics-based Rapid Assay of Individual Microbiome responses to drugs. Microbiome *8*, 1–16.

Liu, C., Song, C.-Q., Yuan, Z.-F., Fu, Y., Chi, H., Wang, L.-H., Fan, S.-B., Zhang, K., Zeng, W.-F., He, S.-M. et al. (2014). pQuant improves quantitation by keeping out interfering signals and evaluating the accuracy of calculated ratios. Analytical chemistry *86*, 5286–5294.

Livanos, A., Greiner, T., Vangay, P., Pathmasiri, W., Stewart, D., McRitchie, S., Li, H., Chung, J., Sohn, J., Kim, S. et al. (2016). Antibiotic-mediated gut microbiome perturbation accelerates development of type 1 diabetes in mice. Nat Microbiol 1: 16140.

Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E. E., Brochado, A. R., Fernandez, K. C., Dose, H., Mori, H. et al. (2018). Extensive impact of non-antibiotic drugs on human gut bacteria. Nature *555*, 623–628.

Maruvada, P., Leone, V., Kaplan, L. M. and Chang, E. B. (2017). The Human Microbiome and Obesity: Moving beyond Associations. Cell Host Microbe *22*, 589–599.

McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C. A., Vanderpool, C. K. and Tjaden, B. (2013). Computational analysis of bacterial RNA-Seq data. Nucleic acids research *41*, e140–e140.

McDonald, J. A., Fuentes, S., Schroeter, K., Heikamp-deJong, I., Khursigara, C. M., de Vos, W. M. and Allen-Vercoe, E. (2015). Simulating distal gut mucosal and luminal communities using packed-column biofilm reactors and an in vitro chemostat model. Journal of microbiological methods *108*, 36–44.

Millard, P., Portais, J.-C. and Mendes, P. (2015). Impact of kinetic isotope effects in isotopic studies of metabolic systems. BMC systems biology *9*, 1–13.

Mottino, A. D., Hoffman, T., Jennes, L. and Vore, M. (2000). Expression and localization of multidrug resistant protein mrp2 in rat small intestine. J Pharmacol Exp Ther *293*, 717–23.

Muth, T., Kohrs, F., Heyer, R., Benndorf, D., Rapp, E., Reichl, U., Martens, L. and Renard, B. Y. (2018). MPA portable: a stand-alone software package for analyzing metaproteome samples on the go. Analytical chemistry *90*, 685–689.

Oberbach, A., Haange, S. B., Schlichting, N., Heinrich, M., Lehmann, S., Till, H., Hugenholtz, F., Kullnick, Y., Smidt, H., Frank, K., Seifert, J., Jehmlich, N. and von Bergen, M. (2017). Metabolic in Vivo Labeling Highlights Differences of Metabolically Active Microbes from the Mucosal Gastrointestinal Microbiome between High-Fat and Normal Chow Diet. J Proteome Res *16*, 1593–1604.

Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S. and Mann, M. (2007).

Higher-energy C-trap dissociation for peptide modification analysis. Nature methods *4*, 709–712.

Park, S. K. R., Aslanian, A., McClatchy, D. B., Han, X., Shah, H., Singh, M., Rauniyar, N., Moresco, J. J., Pinto, A. F., Diedrich, J. K. et al. (2014). Census 2: isobaric labeling data analysis. Bioinformatics *30*, 2208–2209.

Pedrioli, P. G. (2010). Trans-proteomic pipeline: a pipeline for proteomic analysis. In Proteome Bioinformatics pp. 213–238. Springer.

Perkins, D. N., Pappin, D. J., Creasy, D. M. and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. ELECTROPHORESIS: An International Journal *20*, 3551–3567.

Price, J. C., Guan, S., Burlingame, A., Prusiner, S. B. and Ghaemmaghami, S. (2010). Analysis of proteome dynamics in the mouse brain. Proceedings of the National Academy of Sciences *107*, 14508–14513.

Ralls, M. W., Demehri, F. R., Feng, Y., Raskind, S., Ruan, C., Schintlmeister, A., Loy, A., Hanson, B., Berry, D., Burant, C. F. et al. (2016). Bacterial nutrient foraging in a mouse model of enteral nutrient deprivation: insight into the gut origin of sepsis. American Journal of Physiology-Gastrointestinal and Liver Physiology *311*, G734–G743.

Rauniyar, N., McClatchy, D. B. and Yates III, J. R. (2013). Stable isotope labeling of mammals (SILAM) for in vivo quantitative proteomic analysis. Methods *61*, 260–268.

Röst, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E. et al. (2016). OpenMS: a flexible open-

source software platform for mass spectrometry data analysis. Nature methods *13*, 741.

Rowland, I., Gibson, G., Heinken, A., Scott, K., Swann, J., Thiele, I. and Tuohy, K. (2018). Gut microbiota functions: metabolism of nutrients and other food components. European journal of nutrition *57*, 1–24.

Sachsenberg, T., Herbst, F.-A., Taubert, M., Kermer, R., Jehmlich, N., von Bergen, M., Seifert, J. and Kohlbacher, O. (2015). MetaProSIP: automated inference of stable isotope incorporation rates in proteins for functional metaproteomics. Journal of proteome research *14*, 619–627.

Sato, T. and Clevers, H. (2013). Growing self-organizing mini-guts from a single intestinal stem cell: mechanism and applications. Science *340*, 1190–1194.

Senko, M. W., Beu, S. C. and McLaffertycor, F. W. (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. Journal of the American Society for Mass Spectrometry *6*, 229–233.

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S. et al. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nature Biotechnology ., 1–10.

Sleno, L. and Volmer, D. A. (2004). Ion activation methods for tandem mass spectrometry. Journal of mass spectrometry *39*, 1091–1112.

Slysz, G. W., Steinke, L., Ward, D. M., Klatt, C. G., Clauss, T. R., Purvine, S. O., Payne, S. H., Anderson, G. A., Smith, R. D. and Lipton, M. S. (2014). Automated data extraction from in situ protein-stable isotope probing studies. Journal of proteome research *13*, 1200–1210.

Starr, A. E., Deeke, S. A., Li, L., Zhang, X., Daoud, R., Ryan, J., Ning, Z., Cheng, K., Nguyen, L. V., Abou-Samra, E. et al. (2018). Proteomic and metaproteomic approaches to understand host–microbe interactions. Analytical chemistry *90*, 86–109.

Tang, W. H. and Hazen, S. L. (2017). The Gut Microbiome and Its Role in Cardiovascular Diseases. Circulation *135*, 1008–1010.

Taubert, M., Vogt, C., Wubet, T., Kleinsteuber, S., Tarkka, M. T., Harms, H., Buscot, F., Richnow, H.-H., Von Bergen, M. and Seifert, J. (2012). Protein-SIP enables time-resolved analysis of the carbon flux in a sulfate-reducing, benzene-degrading microbial consortium. The ISME journal *6*, 2291–2301.

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T. and Hamon, C. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Analytical chemistry *75*, 1895–1904.

Tremlett, H., Bauer, K. C., Appel-Cresswell, S., Finlay, B. B. and Waubant, E. (2017). The gut microbiome in human neurological disease: A review. Ann Neurol *81*, 369–382.

Verheggen, K., Ræder, H., Berven, F. S., Martens, L., Barsnes, H. and Vaudel, M. (2020).

Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows. Mass spectrometry reviews *39*, 292–306.

Wang, Y., Ahn, T.-H., Li, Z. and Pan, C. (2013). Sipros/ProRata: a versatile informatics system for quantitative community proteomics. Bioinformatics *29*, 2064–2065.

Webhofer, C., Zhang, Y., Brusis, J., Reckow, S., Landgraf, R., Maccarrone, G., Turck, C. W. and Filiou, M. D. (2013). 15N metabolic labeling: Evidence for a stable isotope effect on plasma protein levels and peptide chromatographic retention times. Journal of proteomics *88*, 27–33.

Weisser, H. and Choudhary, J. S. (2017). Targeted feature detection for data-dependent shotgun proteomics. Journal of proteome research *16*, 2964–2974.

Whipps, J., Lewis, K. and Cooke, R. (1988). Mycoparasitism and plant disease control. Fungi in biological control systems ., 161–187.

Wickham, H. (2016). ggplot2: elegant graphics for data analysis. Springer.

Xiao, L., Feng, Q., Liang, S., Sonne, S. B., Xia, Z., Qiu, X., Li, X., Long, H., Zhang, J., Zhang, D. et al. (2015). A catalog of the mouse gut metagenome. Nature biotechnology *33*, 1103.

Young, W., Egert, M., Bassett, S. A. and Bibiloni, R. (2015). Detection of sialic acid-utilising bacteria in a caecal community batch culture using RNA-based stable isotope probing. Nutrients *7*, 2109–2124.

Zhang, X., Chen, W., Ning, Z., Mayne, J., Mack, D., Stintzi, A., Tian, R. and Figeys, D.

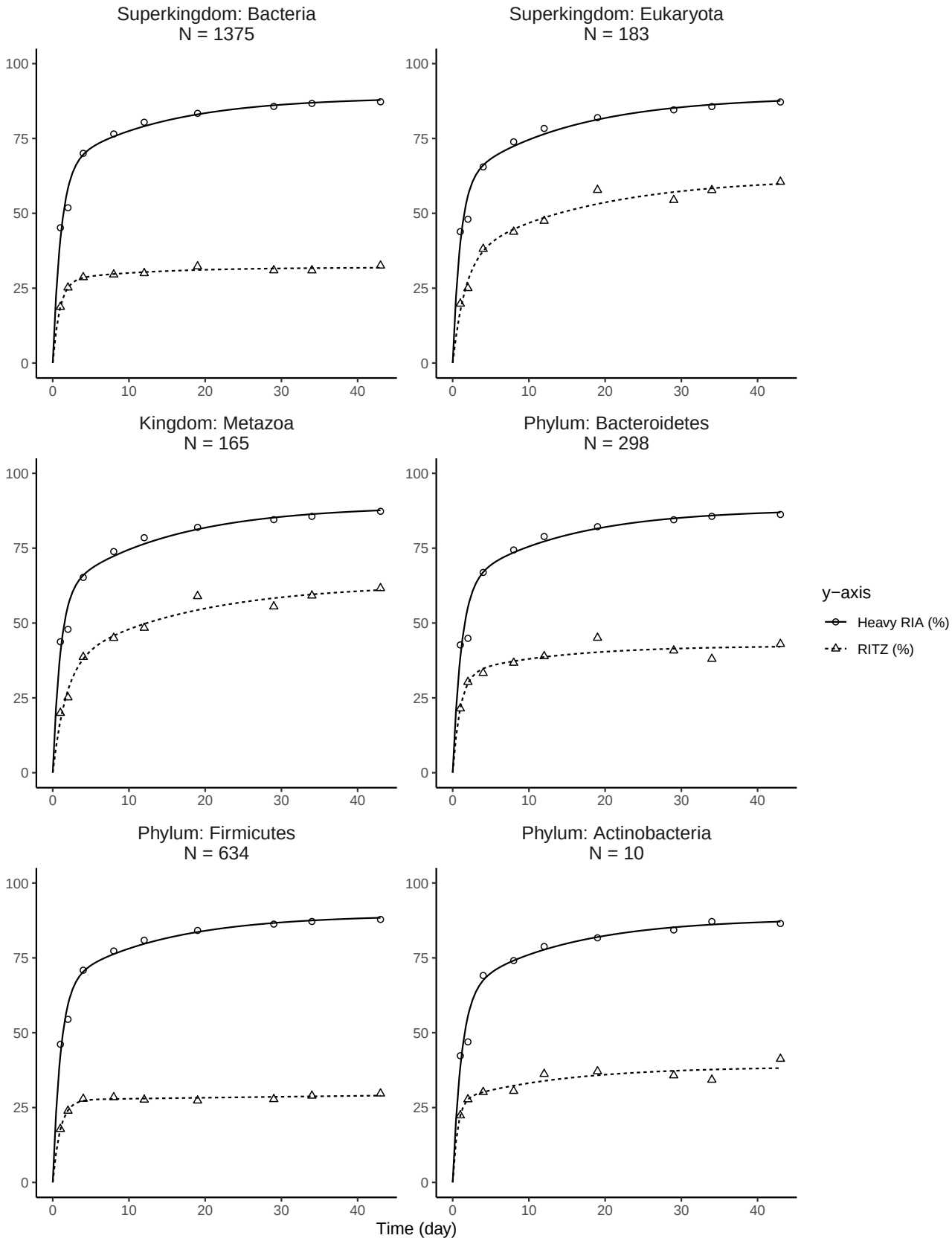
(2017). Deep Metaproteomics Approach for the Study of Human Microbiomes. Anal Chem *89*, 9407–9415.

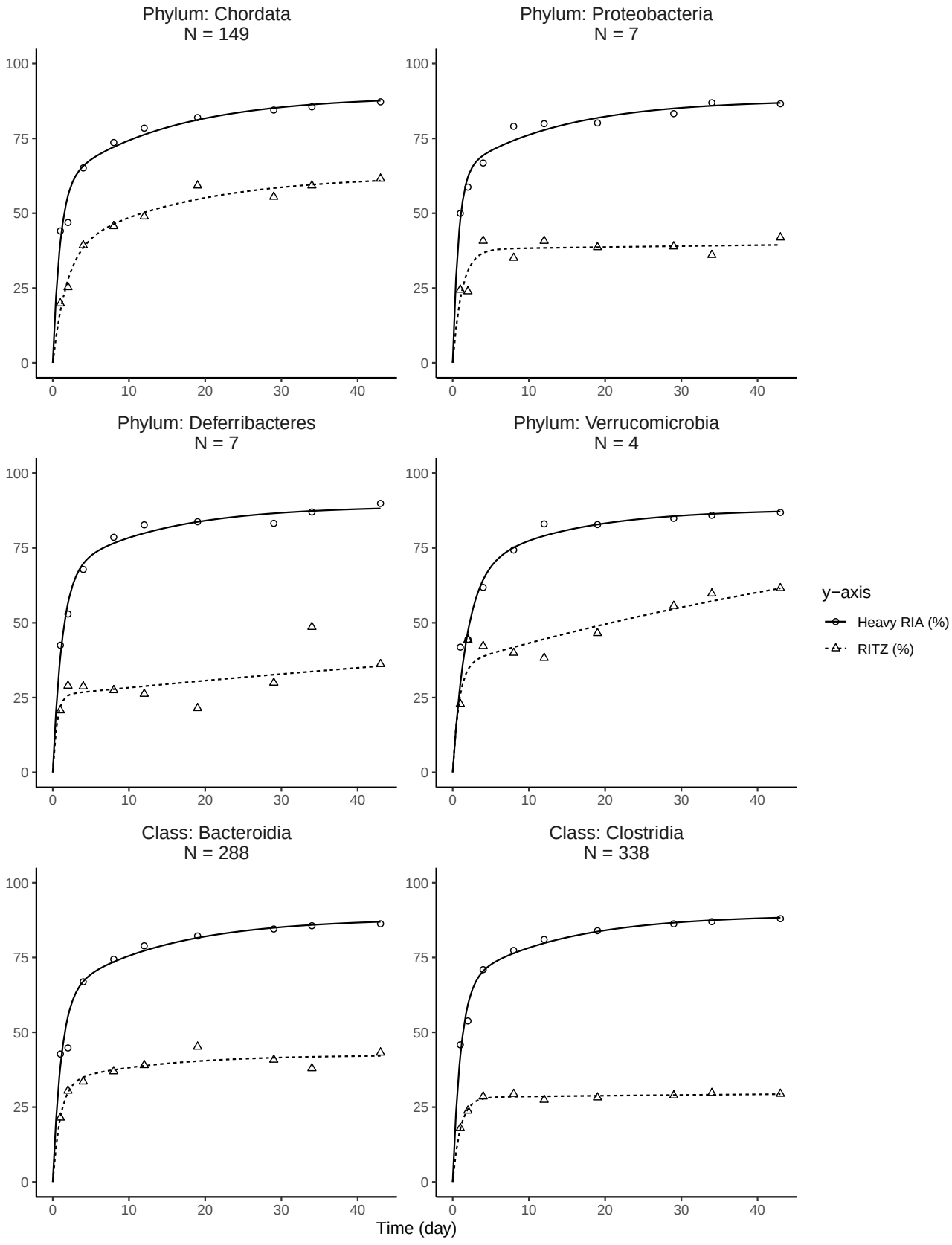
Zhang, X., Deeke, S. A., Ning, Z., Starr, A. E., Butcher, J., Li, J., Mayne, J., Cheng, K., Liao, B., Li, L. et al. (2018a). Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. Nature communications *9*, 1–14.

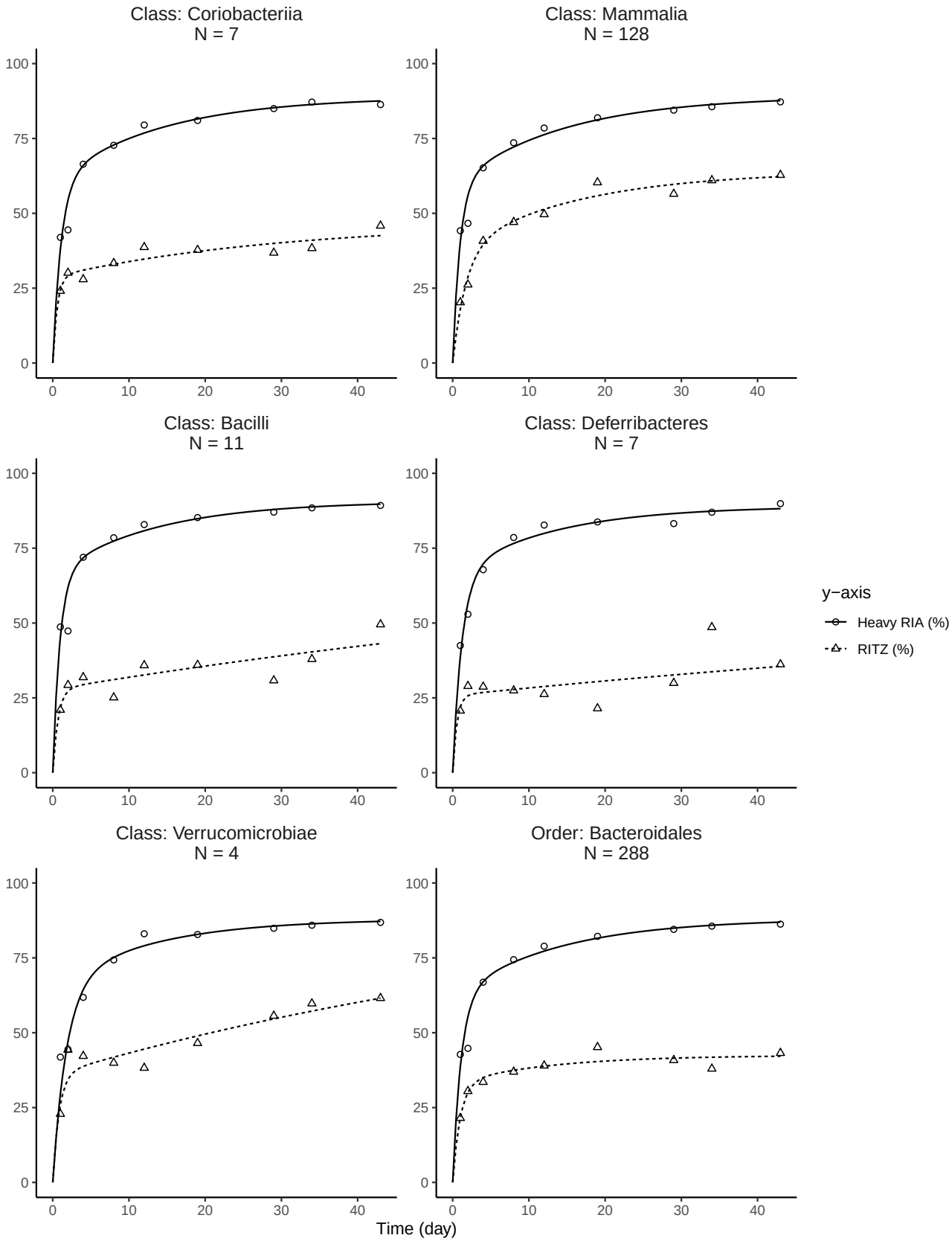
Zhang, X., Li, L., Mayne, J., Ning, Z., Stintzi, A. and Figeys, D. (2018b). Assessing the impact of protein extraction methods for human gut metaproteomics. J Proteomics *180*, 120–127.

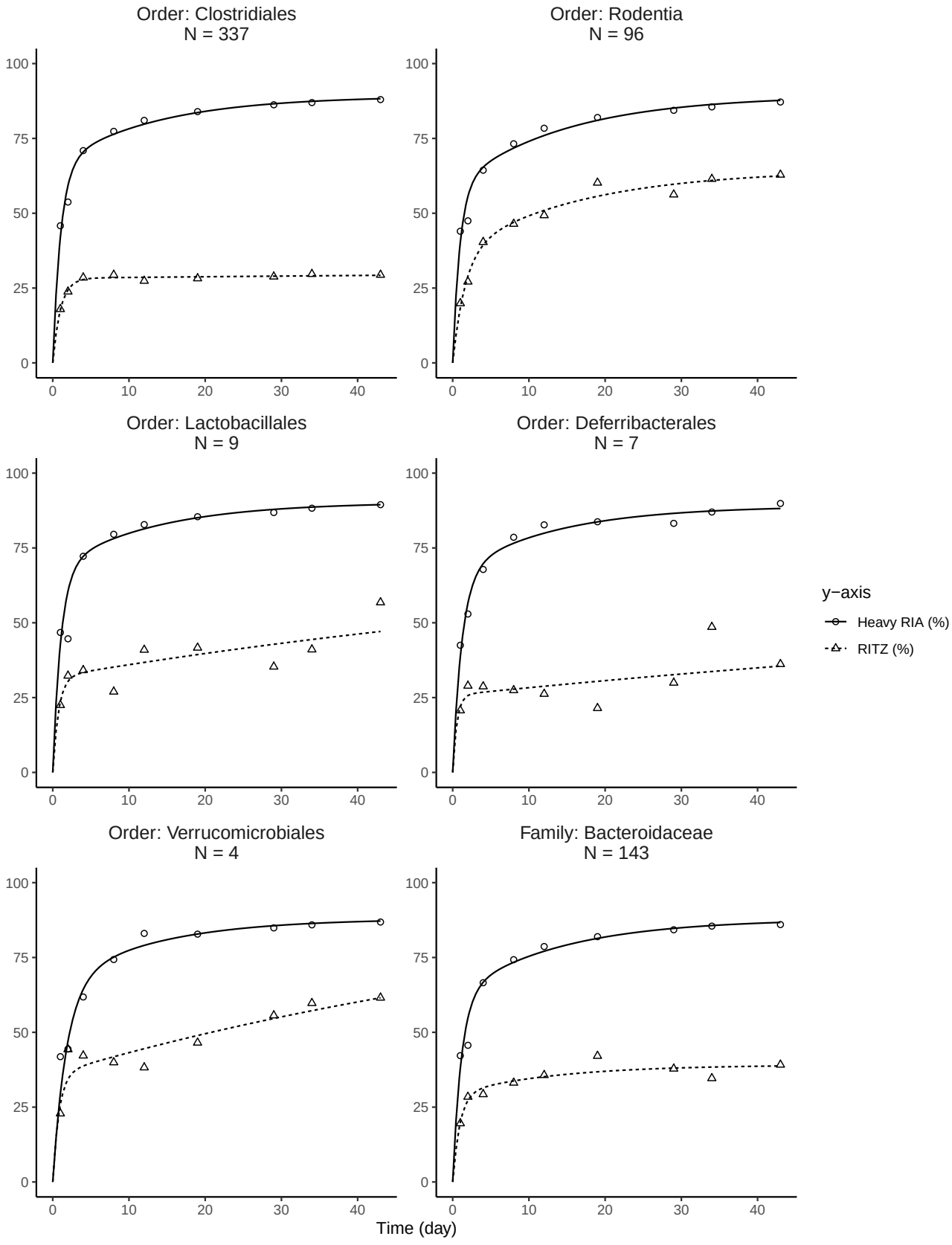
APPENDICES

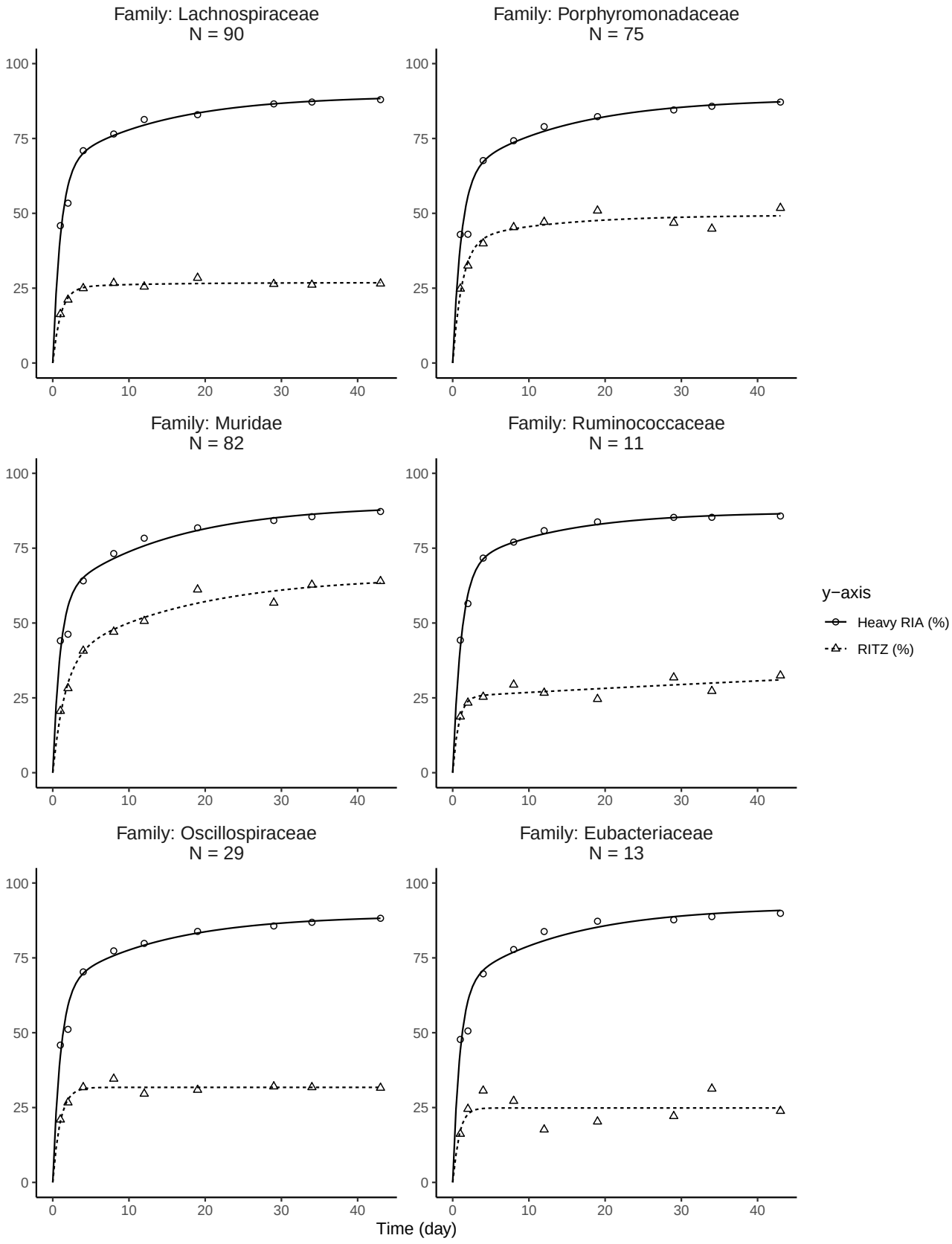
A.1 Supplemental Figures

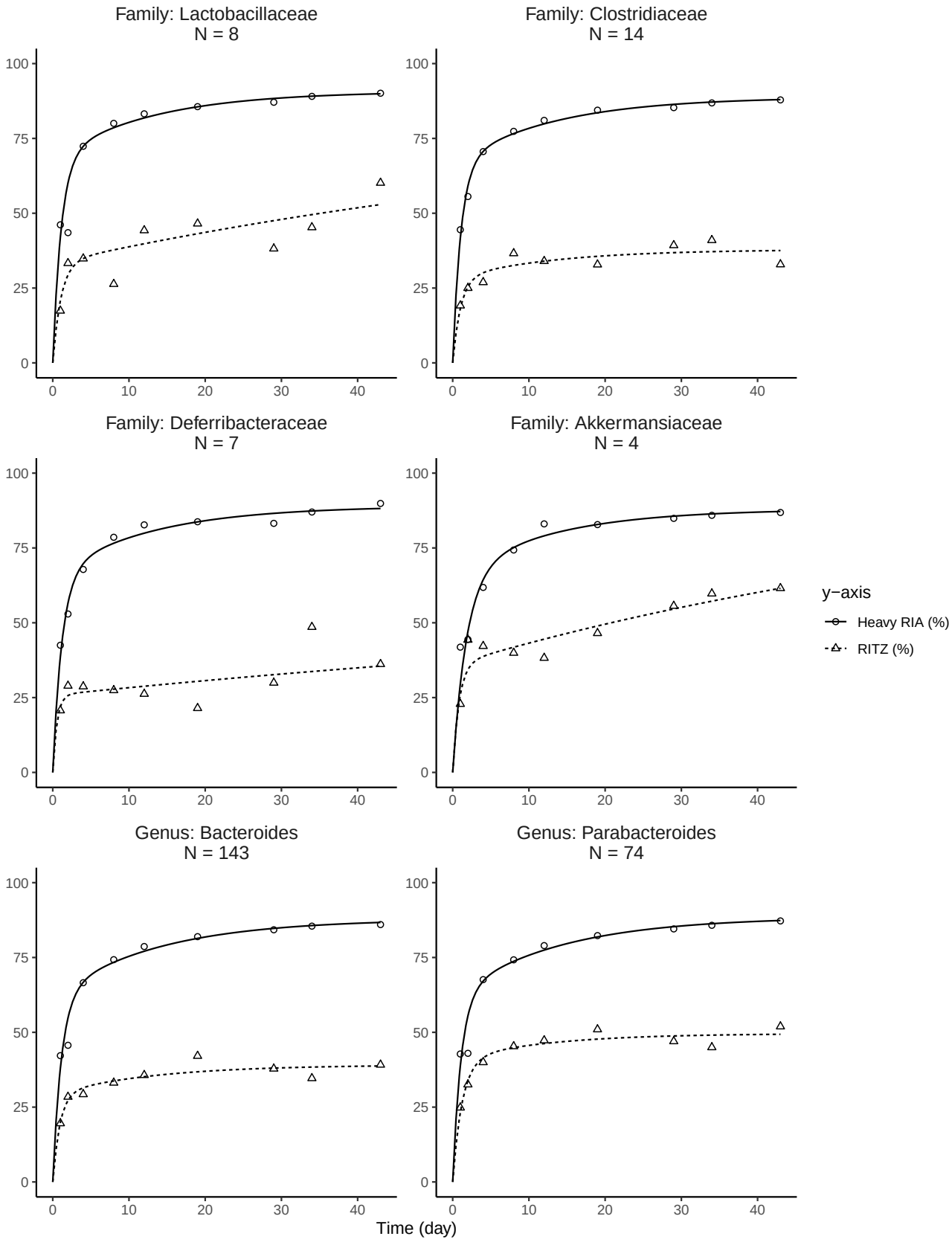

Superkingdom: Bacteria
N = 1375
Superkingdom: Eukaryota
N = 183
Kingdom: Metazoa
N = 165
Phylum: Bacteroidetes
N = 298
Phylum: Firmicutes
N = 634
Phylum: Actinobacteria
N = 10
100
75
50
25
0
0
10
20
30
40
Time (day)
y−axis
Heavy RIA (%)
RITZ (%)

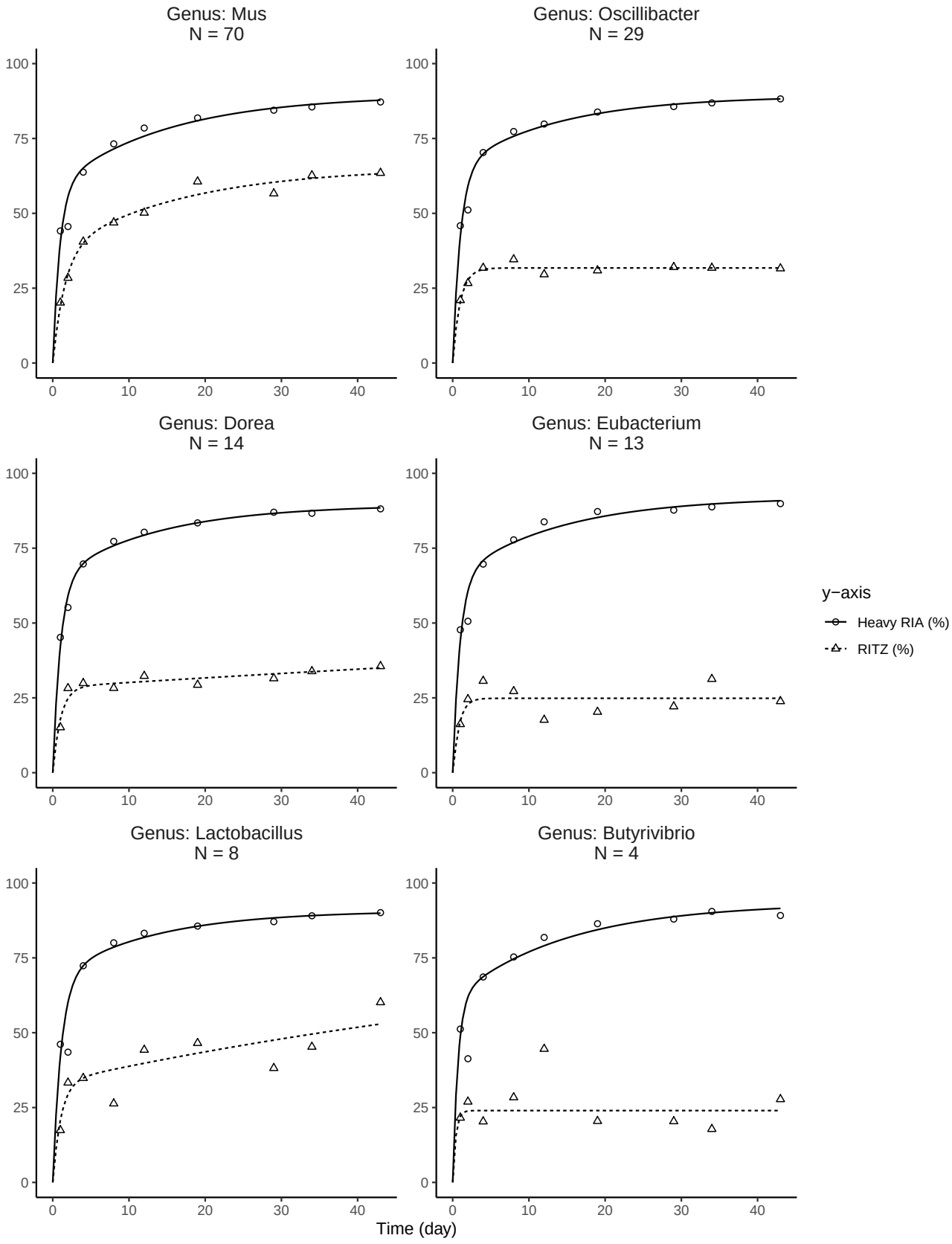

Phylum: Chordata
N = 149
Phylum: Proteobacteria
N = 7
Phylum: Deferribacteres
N = 7
Phylum: Verrucomicrobia
N = 4
Class: Bacteroidia
N = 288
Class: Clostridia
N = 338
100
75
50
25
0
0
10
20
30
40
Time (day)
y−axis
Heavy RIA (%)
RITZ (%)

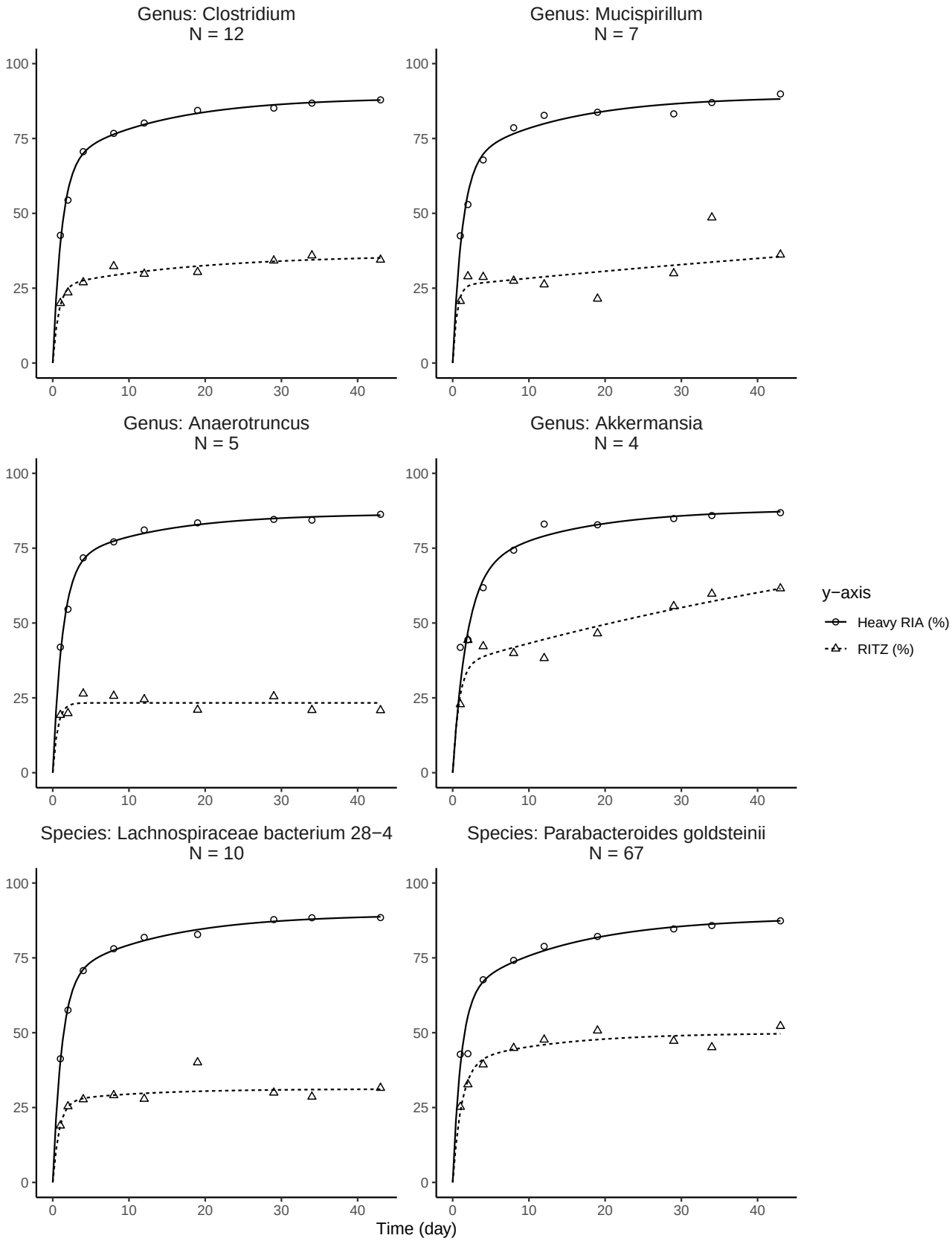

Class: Coriobacteriia
N = 7
Class: Mammalia
N = 128
Class: Bacilli
N = 11
Class: Deferribacteres
N = 7
Class: Verrucomicrobiae
N = 4
Order: Bacteroidales
N = 288
100
75
50
25
0
0
10
20
30
40
Time (day)
y−axis
Heavy RIA (%)
RITZ (%)

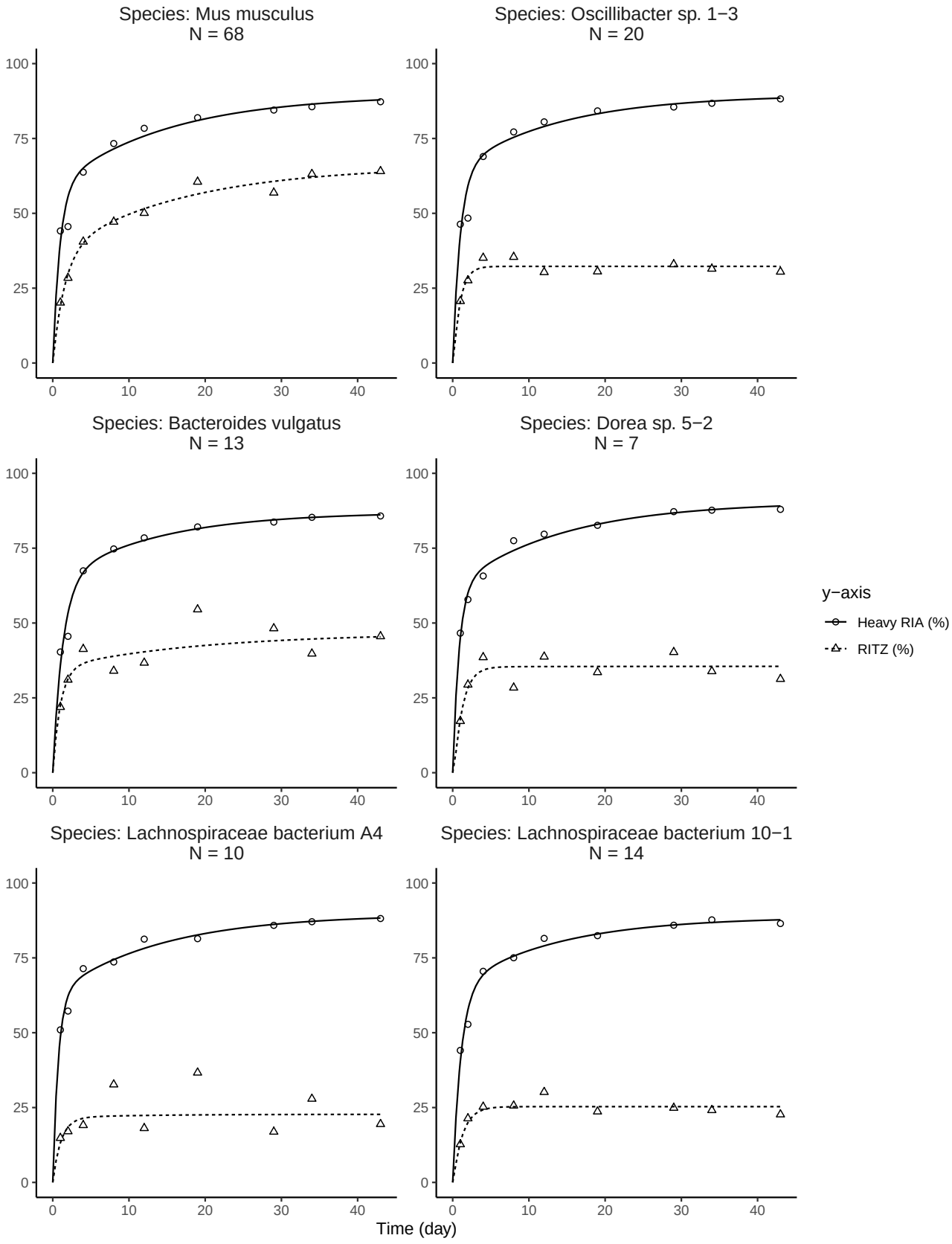

Order: Clostridiales
N = 337
Order: Rodentia
N = 96
Order: Lactobacillales
N = 9
Order: Deferribacterales
N = 7
Order: Verrucomicrobiales
N = 4
Family: Bacteroidaceae
N = 143
100
75
50
25
0
0
10
20
30
40
Time (day)
y−axis
Heavy RIA (%)
RITZ (%)

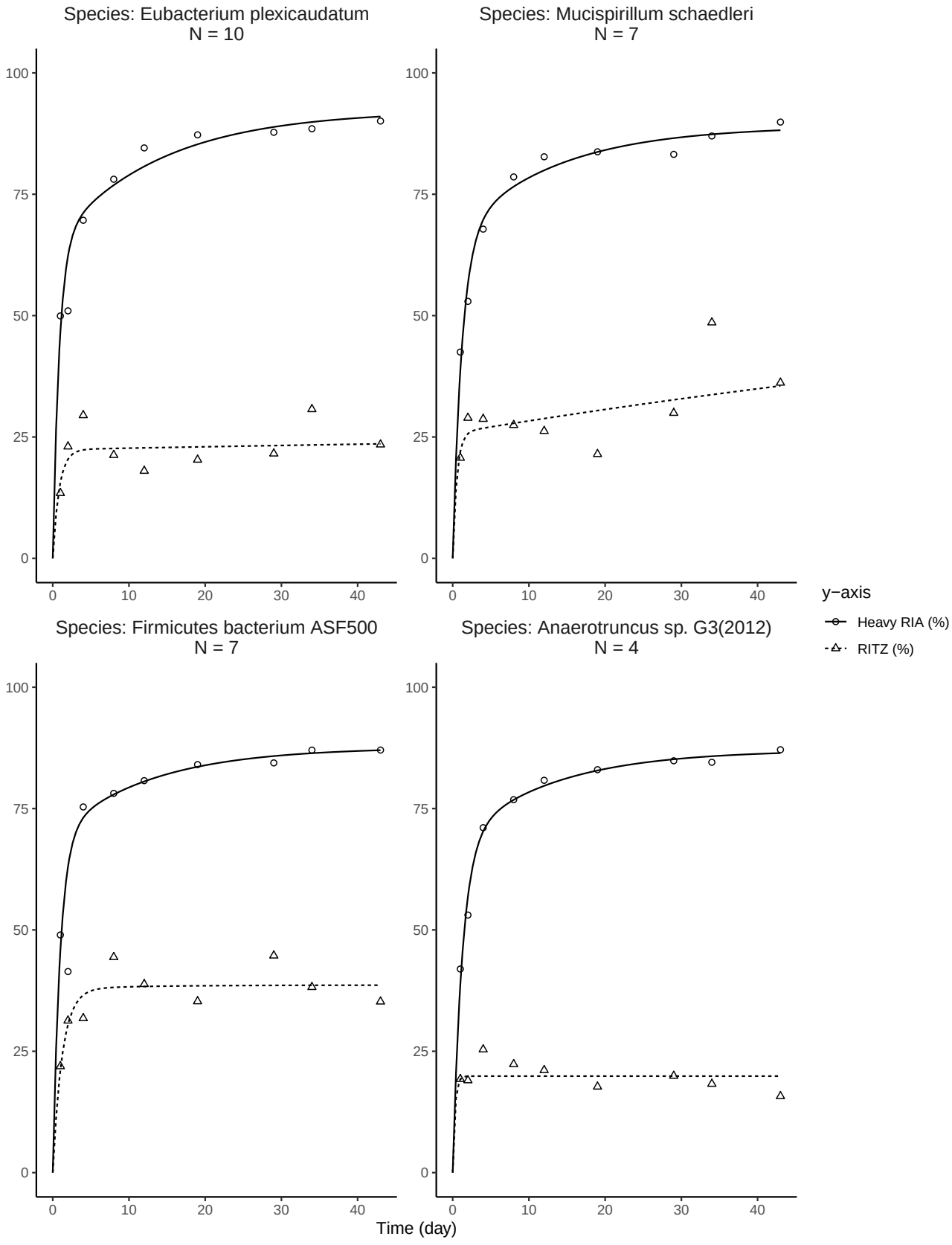

Family: Lachnospiraceae
N = 90
Family: Porphyromonadaceae
N = 75
Family: Muridae
N = 82
Family: Ruminococcaceae
N = 11
Family: Oscillospiraceae
N = 29
Family: Eubacteriaceae
N = 13
100
75
50
25
0
0
10
20
30
40
Time (day)
y−axis
Heavy RIA (%)
RITZ (%)


Family: Lactobacillaceae
N = 8
Family: Clostridiaceae
N = 14
Family: Deferribacteraceae
N = 7
Family: Akkermansiaceae
N = 4
Genus: Bacteroides
N = 143
Genus: Parabacteroides
N = 74
100
75
50
25
0
0
10
20
30
40
Time (day)
y−axis
Heavy RIA (%)
RITZ (%)


Genus: Mus
N = 70
Genus: Oscillibacter
N = 29
Genus: Dorea
N = 14
Genus: Eubacterium
N = 13
Genus: Lactobacillus
N = 8
Genus: Butyrivibrio
N = 4
100
75
50
25
0
0
10
20
30
40
Time (day)
y−axis
Heavy RIA (%)
RITZ (%)


Genus: Clostridium
N = 12
Genus: Mucispirillum
N = 7
Genus: Anaerotruncus
N = 5
Genus: Akkermansia
N = 4
Species: Lachnospiraceae bacterium 28−4
N = 10
Species: Parabacteroides goldsteinii
N = 67
100
75
50
25
0
0
10
20
30
40
Time (day)
y−axis
Heavy RIA (%)
RITZ (%)


Species: Mus musculus
N = 68
Species: Oscillibacter sp. 1−3
N = 20
Species: Bacteroides vulgatus
N = 13
Species: Dorea sp. 5−2
N = 7
Species: Lachnospiraceae bacterium A4
N = 10
Species: Lachnospiraceae bacterium 10−1
N = 14
100
75
50
25
0
0
10
20
30
40
Time (day)
y−axis
Heavy RIA (%)
RITZ (%)


Species: Eubacterium plexicaudatum
N = 10
Species: Mucispirillum schaedleri
N = 7
Species: Firmicutes bacterium ASF500
N = 7
Species: Anaerotruncus sp. G3(2012)
N = 4
100
75
50
25
0
0
10
20
30
40
Time (day)
y−axis
Heavy RIA (%)
RITZ (%)

A.2 Supplemental Tables

Table A.1: Average taxonomic RITZ from day 29 to 43

Rank	LCA	RITZ mean (%)	p.value
Phylum	Firmicutes	31.4	0
Phylum	Proteobacteria	40.2	2.13×10^{-18}
Phylum	Bacteroidetes	40.6	0
Phylum	Chordata	60.2	5.38×10^{-307}
Phylum	Actinobacteria	38.3	3.11×10^{-17}
Phylum	Euryarchaeota	24.3	0.0183
Phylum	Verrucomicrobia	58.2	5.01×10^{-14}
Phylum	Fusobacteria	8.25	0.0379
Phylum	Deferribacteres	27.4	4.17×10^{-11}
Phylum	Ascomycota	27.7	6.96×10^{-6}
Phylum	Platyhelminthes	30.6	0.0243
Phylum	Spirochaetes	34.4	0.0407
Phylum	Cyanobacteria	33.2	0.0167
Class	Clostridia	31.9	0
Class	Bacteroidia	40.6	0
Class	Mammalia	62.7	1.87×10^{-279}
Class	Negativicutes	10.3	0.003 13
Class	Alphaproteobacteria	41.9	1.43×10^{-6}
Class	Coriobacteriia	40.4	2.75×10^{-15}
Class	Deltaproteobacteria	60.5	0.008 48
Class	Bacilli	49.3	1.35×10^{-19}
Class	Thermococci	24.3	0.0183
Class	Verrucomicrobiae	58.2	5.01×10^{-14}
Class	Fusobacteriia	8.25	0.0379
Class	Deferribacteres	27.4	4.17×10^{-11}
Class	Actinobacteria	33.6	0.000 458
Class	Gammaproteobacteria	33.9	7.68×10^{-10}
Class	Spirochaetia	34.4	0.0407
Class	Flavobacteriia	35.5	0.006 21
Order	Clostridiales	31.8	0
Order	Bacteroidales	40.6	0
Order	Rodentia	62.9	3.15×10^{-228}
Order	Selenomonadales	10.3	0.003 13
Order	Rhodobacterales	34.8	1.01×10^{-5}
Order	Eggerthellales	45.5	1.84×10^{-7}
Order	Lactobacillales	53.5	2.3×10^{-18}
Order	Thermococcales	24.3	0.0183
Order	Verrucomicrobiales	58.2	5.01×10^{-14}
Order	Fusobacteriales	8.25	0.0379
Order	Rhizobiales	56.1	0.006 21
Order	Deferribacterales	27.4	4.17×10^{-11}

Continued on next page

Table A.1 – continued from previous page

Rank	LCA	RITZ mean (%)	p.value
Order	Desulfovibrionales	88.2	0.002 86
Order	Xanthomonadales	20.2	0.0355
Order	Spirochaetales	34.4	0.0407
Order	Alteromonadales	36.6	0.000 228
Order	Flavobacteriales	35.5	0.006 21
Family	Lachnospiraceae	30.7	3.73×10^{-111}
Family	Bacteroidaceae	38	1.47×10^{-208}
Family	Rhodobacteraceae	34.8	1.01×10^{-5}
Family	Oscillospiraceae	33	5.18×10^{-35}
Family	Eggerthellaceae	45.5	1.84×10^{-7}
Family	Porphyromonadaceae	46.7	1.44×10^{-134}
Family	Muridae	63.4	3.22×10^{-202}
Family	Clostridiaceae	42.6	2.49×10^{-26}
Family	Lactobacillaceae	54	6.03×10^{-17}
Family	Thermococcaceae	24.3	0.0183
Family	Ruminococcaceae	34	7.62×10^{-15}
Family	Eubacteriaceae	30	1.54×10^{-12}
Family	Akkermansiaceae	58.2	5.01×10^{-14}
Family	Fusobacteriaceae	8.25	0.0379
Family	Bradyrhizobiaceae	56.1	0.006 21
Family	Deferribacteraceae	27.4	4.17×10^{-11}
Family	Desulfovibrionaceae	88.2	0.002 86
Family	Xanthomonadaceae	20.2	0.0355
Family	Spirochaetaceae	34.4	0.0407
Family	Flavobacteriaceae	36.9	0.005 89
Genus	Bacteroides	38	1.47×10^{-208}
Genus	Oscillibacter	33	5.18×10^{-35}
Genus	Blautia	44.5	0.001 13
Genus	Parabacteroides	46.9	7.02×10^{-133}
Genus	Clostridium	40.9	3.48×10^{-20}
Genus	Mus	63.5	4.16×10^{-171}
Genus	Lactobacillus	53.5	1.21×10^{-15}
Genus	Ruminiclostridium	32.2	0.000 126
Genus	Roseburia	36.6	0.000 209
Genus	Eubacterium	29.4	3.57×10^{-12}
Genus	Akkermansia	58.2	5.01×10^{-14}
Genus	Bradyrhizobium	56.1	0.006 21
Genus	Herbinix	23.8	0.0431
Genus	Mucispirillum	27.4	4.17×10^{-11}
Genus	Lachnoclostridium	19.9	0.000 259
Genus	Dorea	34.7	2.24×10^{-17}
Genus	Coprococcus	35	0.0306

Continued on next page

Table A.1 – continued from previous page

Rank	LCA	RITZ mean (%)	p.value
Genus	Anaerotruncus	24.4	5.11×10^{-7}
Genus	Butyrivibrio	22.8	0.000 168
Genus	Flavonifractor	37.3	0.000 45
Genus	Subdoligranulum	41	0.006 62
Genus	Butyricicoccus	71.8	0.006 62
Genus	Ruminococcus	51	0.000 13
Genus	Enterorhabdus	51.2	0.0493
Genus	Treponema	34.4	0.0407
Species	Lachnospiraceae bacterium 10-1	27.4	1.22×10^{-12}
Species	Lachnospiraceae bacterium A4	31.1	3.62×10^{-17}
Species	Lachnospiraceae bacterium 28-4	35.2	7.62×10^{-15}
Species	Lachnospiraceae bacterium COE1	38.9	0.000 236
Species	Parabacteroides goldsteinii	47.3	7.54×10^{-120}
Species	Clostridium sp. ASF502	55.2	1.3×10^{-12}
Species	Mus musculus	64	7.41×10^{-166}
Species	Lachnospiraceae bacterium 3-2	34.9	0.000 187
Species	Lachnospiraceae bacterium JC7	21	0.000 461
Species	Ruminiclostridium sp. KB18	28.6	0.001 16
Species	Eubacterium plexicaudatum	29.2	1.08×10^{-8}
Species	Oscillibacter sp. 1-3	34.7	8.04×10^{-30}
Species	Akkermansia muciniphila	57	2.46×10^{-8}
Species	Lachnospiraceae bacterium M18-1	45	2.31×10^{-7}
Species	Mucispirillum schaedleri	27.4	4.17×10^{-11}
Species	Dorea sp. 5-2	35.5	1.84×10^{-9}
Species	Bacteroides vulgatus	48.9	2.98×10^{-26}
Species	Lachnospiraceae bacterium A2	41.5	9.46×10^{-5}
Species	Firmicutes bacterium ASF500	43.3	2.98×10^{-14}
Species	Anaerotruncus sp. G3(2012)	21.9	9.51×10^{-6}
Species	Flavonifractor plautii	40.3	0.000 455
Species	Subdoligranulum variabile	41	0.006 62
Species	Oscillibacter sp. KLE 1745	59.3	0.0209
Species	Enterorhabdus caecimuris	51.2	0.0493
Species	Oscillibacter valericigenes	29.1	0.002 67
Species	Butyrivibrio crossotus	17.3	0.001 28
Species	Eubacterium sp. 14-2	32.3	0.000 461
Species	Coprococcus comes	35	0.0306

Table A.2: Average taxonomic RITZ from day 29 to 43

Rank	LCA	NOG category	RITZ mean (%)	p-value
Phylum	Firmicutes	Lipid transport and metabolism	28.4	8.23×10^{-50}
Phylum	Proteobacteria	Carbohydrate transport and metabolism	35.7	6.85×10^{-7}
Phylum	Bacteroidetes	Carbohydrate transport and metabolism	35	2.95×10^{-6}
Phylum	Chordata	Intracellular trafficking, secretion, and vesicular transport	62.3	6.72×10^{-37}
Phylum	Firmicutes	Energy production and conversion	27	2.04×10^{-156}
Phylum	Firmicutes	Carbohydrate transport and metabolism	33.5	7.33×10^{-222}
Phylum	Bacteroidetes	Cell wall/membrane/envelope biogenesis	42.6	1.54×10^{-51}
Phylum	Firmicutes	Cell motility	32.7	3.04×10^{-45}
Phylum	Proteobacteria	Amino acid transport and metabolism	42.8	3.26×10^{-5}
Phylum	Firmicutes	Amino acid transport and metabolism	36.2	8.73×10^{-83}
Phylum	Firmicutes	Translation, ribosomal structure and biogenesis	31.1	1.07×10^{-107}
Phylum	Bacteroidetes	Inorganic ion transport and metabolism	42.9	3.45×10^{-84}
Phylum	Firmicutes	Posttranslational modification, protein turnover, chaperones	27.6	3×10^{-34}
Phylum	Actinobacteria	Amino acid transport and metabolism	41.6	1.16×10^{-11}
Phylum	Firmicutes	Secondary metabolites biosynthesis, transport and catabolism	35.3	0.0155
Phylum	Chordata	Inorganic ion transport and metabolism	62.3	9.16×10^{-36}
Phylum	Firmicutes	Nucleotide transport and metabolism	43	6.45×10^{-22}
Phylum	Chordata	Amino acid transport and metabolism	56.5	1.38×10^{-69}
Phylum	Chordata	Posttranslational modification, protein turnover, chaperones	56	4.63×10^{-89}
Phylum	Firmicutes	Cell wall/membrane/envelope biogenesis	29.2	0.0386
Phylum	Chordata	Signal transduction mechanisms	65.9	9.49×10^{-44}
Phylum	Firmicutes	Replication, recombination and repair	31.6	0.006 61
Phylum	Bacteroidetes	Energy production and conversion	37.2	8.75×10^{-19}
Phylum	Verrucomicrobia	Translation, ribosomal structure and biogenesis	67.2	3.14×10^{-6}
Phylum	Fusobacteria	Lipid transport and metabolism	8.25	0.0355
Phylum	Verrucomicrobia	Carbohydrate transport and metabolism	59.1	6.67×10^{-6}
Phylum	Chordata	Cytoskeleton	41.6	3.46×10^{-5}
Phylum	Bacteroidetes	Intracellular trafficking, secretion, and vesicular transport	37.8	3.61×10^{-23}
Phylum	Chordata	Carbohydrate transport and metabolism	53.8	1.92×10^{-25}
Phylum	Bacteroidetes	Lipid transport and metabolism	28.3	0.0466
Phylum	Bacteroidetes	Posttranslational modification, protein turnover, chaperones	37.5	0.027
Phylum	Bacteroidetes	Amino acid transport and metabolism	50.6	1.71×10^{-8}
Phylum	Actinobacteria	Energy production and conversion	35.7	4.37×10^{-6}
Phylum	Chordata	Defense mechanisms	87.3	1.56×10^{-20}
Phylum	Bacteroidetes	Translation, ribosomal structure and biogenesis	35.5	1.52×10^{-6}
Phylum	Firmicutes	Transcription	36	0.000 898
Phylum	Proteobacteria	Energy production and conversion	52.6	0.000 205
Phylum	Firmicutes	Inorganic ion transport and metabolism	22.4	1.24×10^{-7}
Phylum	Deferribacteres	Energy production and conversion	31.9	1.29×10^{-6}
Phylum	Chordata	Extracellular structures	61.5	1.33×10^{-16}
Phylum	Chordata	Secondary metabolites biosynthesis, transport and catabolism	50.9	0.006 64
Phylum	Chordata	Energy production and conversion	51.3	1.63×10^{-10}
Phylum	Actinobacteria	Translation, ribosomal structure and biogenesis	20	0.000 66
Phylum	Ascomycota	Carbohydrate transport and metabolism	27.3	0.000 707
Phylum	Firmicutes	Intracellular trafficking, secretion, and vesicular transport	47	0.003 33
Phylum	Proteobacteria	Lipid transport and metabolism	30.7	0.003 67
Class	Clostridia	Lipid transport and metabolism	26.3	1.41×10^{-22}
Class	Bacteroidia	Carbohydrate transport and metabolism	37.3	7.4×10^{-6}
Class	Mammalia	Intracellular trafficking, secretion, and vesicular transport	69.3	1.42×10^{-28}
Class	Clostridia	Energy production and conversion	26.9	2.12×10^{-94}
Class	Clostridia	Carbohydrate transport and metabolism	35.1	1.6×10^{-144}
Class	Bacteroidia	Cell wall/membrane/envelope biogenesis	43.3	3.59×10^{-48}
Class	Clostridia	Cell motility	34	4.94×10^{-34}
Class	Alphaproteobacteria	Amino acid transport and metabolism	37.2	0.001 25
Class	Clostridia	Translation, ribosomal structure and biogenesis	35.2	9.49×10^{-44}
Class	Bacteroidia	Inorganic ion transport and metabolism	42.4	1.21×10^{-73}
Class	Clostridia	Posttranslational modification, protein turnover, chaperones	26.2	1.55×10^{-22}
Class	Coriobacteriia	Amino acid transport and metabolism	45.2	2.38×10^{-12}
Class	Clostridia	Amino acid transport and metabolism	35.4	1.6×10^{-30}
Class	Clostridia	Secondary metabolites biosynthesis, transport and catabolism	35.3	0.0155
Class	Mammalia	Inorganic ion transport and metabolism	64.5	5.16×10^{-34}
Class	Clostridia	Nucleotide transport and metabolism	42.9	8.8×10^{-9}
Class	Mammalia	Amino acid transport and metabolism	56.6	1.96×10^{-58}
Class	Mammalia	Posttranslational modification, protein turnover, chaperones	57.6	5.69×10^{-81}
Class	Bacilli	Translation, ribosomal structure and biogenesis	49.5	2.4×10^{-5}
Class	Mammalia	Signal transduction mechanisms	65.9	9.49×10^{-44}
Class	Bacteroidia	Energy production and conversion	37.3	8.85×10^{-19}
Class	Verrucomicrobiae	Translation, ribosomal structure and biogenesis	67.2	3.14×10^{-6}
Class	Fusobacteriia	Lipid transport and metabolism	8.25	0.0355
Class	Verrucomicrobiae	Carbohydrate transport and metabolism	59.1	6.67×10^{-6}
Class	Mammalia	Cytoskeleton	41.6	3.46×10^{-5}
Class	Bacteroidia	Intracellular trafficking, secretion, and vesicular transport	38.7	3.09×10^{-23}
Class	Mammalia	Carbohydrate transport and metabolism	59.2	3.26×10^{-21}
Class	Bacteroidia	Lipid transport and metabolism	28.3	0.0466
Class	Bacteroidia	Posttranslational modification, protein turnover, chaperones	37.5	0.027
Class	Bacteroidia	Amino acid transport and metabolism	50.6	1.71×10^{-8}
Class	Coriobacteriia	Energy production and conversion	35.7	4.37×10^{-6}
Class	Bacilli	Carbohydrate transport and metabolism	54.2	3.07×10^{-14}
Class	Mammalia	Defense mechanisms	87.3	1.56×10^{-20}
Class	Clostridia	Transcription	38.8	0.004 93

Continued on next page

Table A.2 – continued from previous page

Rank	LCA	NOG category	RITZ mean (%)	p-value
Class	Deltaproteobacteria	Energy production and conversion	42	0.005 31
Class	Clostridia	Inorganic ion transport and metabolism	20.4	2.76×10^{-5}
Class	Deferribacteres	Energy production and conversion	31.9	1.29×10^{-6}
Class	Bacteroidia	Translation, ribosomal structure and biogenesis	32	7.75×10^{-8}
Class	Mammalia	Extracellular structures	61.5	1.33×10^{-16}
Class	Bacilli	Energy production and conversion	41.6	0.0093
Class	Mammalia	Secondary metabolites biosynthesis, transport and catabolism	50.9	0.006 64
Class	Gammaproteobacteria	Carbohydrate transport and metabolism	34.4	0.000 85
Class	Actinobacteria	Translation, ribosomal structure and biogenesis	20	0.000 66
Class	Gammaproteobacteria	Amino acid transport and metabolism	32.2	0.002 19
Class	Clostridia	Cell wall/membrane/envelope biogenesis	29.2	0.0386
Order	Clostridiales	Lipid transport and metabolism	26.3	1.41×10^{-22}
Order	Bacteroidales	Carbohydrate transport and metabolism	37.3	7.4×10^{-6}
Order	Rodentia	Intracellular trafficking, secretion, and vesicular transport	77	6.99×10^{-15}
Order	Clostridiales	Energy production and conversion	26.9	2.12×10^{-94}
Order	Clostridiales	Carbohydrate transport and metabolism	35	4.36×10^{-144}
Order	Bacteroidales	Cell wall/membrane/envelope biogenesis	43.3	3.59×10^{-48}
Order	Clostridiales	Cell motility	34	4.94×10^{-34}
Order	Rhodobacterales	Amino acid transport and metabolism	37.2	0.001 25
Order	Clostridiales	Translation, ribosomal structure and biogenesis	35	3.56×10^{-42}
Order	Bacteroidales	Inorganic ion transport and metabolism	42.4	1.21×10^{-73}
Order	Clostridiales	Posttranslational modification, protein turnover, chaperones	26.2	1.55×10^{-22}
Order	Eggerthellales	Amino acid transport and metabolism	45.5	1.99×10^{-7}
Order	Clostridiales	Amino acid transport and metabolism	35.4	1.6×10^{-30}
Order	Clostridiales	Secondary metabolites biosynthesis, transport and catabolism	35.3	0.0155
Order	Rodentia	Inorganic ion transport and metabolism	63.7	7.4×10^{-27}
Order	Clostridiales	Nucleotide transport and metabolism	42.9	8.8×10^{-9}
Order	Rodentia	Amino acid transport and metabolism	53.8	2.89×10^{-39}
Order	Rodentia	Posttranslational modification, protein turnover, chaperones	56.8	2.63×10^{-71}
Order	Lactobacillales	Translation, ribosomal structure and biogenesis	49.5	2.4×10^{-5}
Order	Rodentia	Signal transduction mechanisms	67.2	2.19×10^{-44}
Order	Bacteroidales	Energy production and conversion	37.3	8.85×10^{-19}
Order	Verrucomicrobiales	Translation, ribosomal structure and biogenesis	67.2	3.14×10^{-6}
Order	Fusobacteriales	Lipid transport and metabolism	8.25	0.0355
Order	Verrucomicrobiales	Carbohydrate transport and metabolism	59.1	6.67×10^{-6}
Order	Bacteroidales	Intracellular trafficking, secretion, and vesicular transport	38.7	3.09×10^{-23}
Order	Rodentia	Carbohydrate transport and metabolism	63.9	4.01×10^{-20}
Order	Bacteroidales	Lipid transport and metabolism	28.3	0.0466
Order	Bacteroidales	Posttranslational modification, protein turnover, chaperones	37.5	0.027
Order	Bacteroidales	Amino acid transport and metabolism	50.6	1.71×10^{-8}
Order	Lactobacillales	Carbohydrate transport and metabolism	58.9	1.32×10^{-13}
Order	Rodentia	Defense mechanisms	87.3	1.56×10^{-20}
Order	Clostridiales	Transcription	38.8	0.004 93
Order	Clostridiales	Inorganic ion transport and metabolism	20.4	2.76×10^{-5}
Order	Deferribacterales	Energy production and conversion	31.9	1.29×10^{-6}
Order	Bacteroidales	Translation, ribosomal structure and biogenesis	32	7.75×10^{-8}
Order	Rodentia	Extracellular structures	58	1.18×10^{-12}
Order	Lactobacillales	Energy production and conversion	41.6	0.0093
Order	Rodentia	Secondary metabolites biosynthesis, transport and catabolism	50.9	0.006 64
Order	Clostridiales	Cell wall/membrane/envelope biogenesis	29.2	0.0386
Family	Lachnospiraceae	Lipid transport and metabolism	25.2	2.17×10^{-5}
Family	Bacteroidaceae	Carbohydrate transport and metabolism	42.2	0.019
Family	Lachnospiraceae	Energy production and conversion	22.5	2.73×10^{-31}
Family	Lachnospiraceae	Carbohydrate transport and metabolism	32.5	1.86×10^{-36}
Family	Bacteroidaceae	Cell wall/membrane/envelope biogenesis	41.6	4.94×10^{-34}
Family	Lachnospiraceae	Cell motility	39.9	2.4×10^{-16}
Family	Rhodobacteraceae	Amino acid transport and metabolism	37.2	0.001 25
Family	Bacteroidaceae	Inorganic ion transport and metabolism	42.6	6.93×10^{-43}
Family	Lachnospiraceae	Posttranslational modification, protein turnover, chaperones	25.9	1.71×10^{-6}
Family	Eggerthellaceae	Amino acid transport and metabolism	45.5	1.99×10^{-7}
Family	Porphyromonadaceae	Inorganic ion transport and metabolism	36.7	1.77×10^{-9}
Family	Clostridiaceae	Carbohydrate transport and metabolism	40.5	8.88×10^{-8}
Family	Muridae	Inorganic ion transport and metabolism	64.9	8.05×10^{-21}
Family	Clostridiaceae	Nucleotide transport and metabolism	54.2	0.000 202
Family	Muridae	Amino acid transport and metabolism	54.3	1.97×10^{-37}
Family	Muridae	Posttranslational modification, protein turnover, chaperones	57	6.38×10^{-68}
Family	Lactobacillaceae	Translation, ribosomal structure and biogenesis	49.5	2.4×10^{-5}
Family	Muridae	Signal transduction mechanisms	69.2	5.68×10^{-39}
Family	Eubacteriaceae	Carbohydrate transport and metabolism	39.2	3.46×10^{-5}
Family	Akkermansiaceae	Translation, ribosomal structure and biogenesis	67.2	3.14×10^{-6}
Family	Fusobacteriaceae	Lipid transport and metabolism	8.25	0.0355
Family	Oscillospiraceae	Lipid transport and metabolism	27.5	4.87×10^{-5}
Family	Akkermansiaceae	Carbohydrate transport and metabolism	59.1	6.67×10^{-6}
Family	Lachnospiraceae	Translation, ribosomal structure and biogenesis	43.5	1.74×10^{-9}
Family	Lachnospiraceae	Amino acid transport and metabolism	36.5	1.35×10^{-8}
Family	Muridae	Carbohydrate transport and metabolism	67.9	1.75×10^{-21}
Family	Clostridiaceae	Translation, ribosomal structure and biogenesis	41.2	0.008 46
Family	Bacteroidaceae	Amino acid transport and metabolism	50.3	3.75×10^{-6}
Family	Oscillospiraceae	Energy production and conversion	28.8	8.16×10^{-6}
Family	Eubacteriaceae	Cell motility	20.1	0.0269
Family	Lactobacillaceae	Carbohydrate transport and metabolism	61.5	9.2×10^{-14}

Continued on next page

Table A.2 – continued from previous page

Rank	LCA	NOG category	RITZ mean (%)	p-value
Family	Muridae	Defense mechanisms	87	3.55×10^{-15}
Family	Oscillospiraceae	Amino acid transport and metabolism	38.8	5.92×10^{-5}
Family	Oscillospiraceae	Carbohydrate transport and metabolism	33.5	2.42×10^{-21}
Family	Bacteroidaceae	Intracellular trafficking, secretion, and vesicular transport	36.4	2.2×10^{-15}
Family	Oscillospiraceae	Cell motility	38.6	0.001 42
Family	Bacteroidaceae	Energy production and conversion	39.1	2.66×10^{-14}
Family	Clostridiaceae	Cell motility	52	1.49×10^{-6}
Family	Clostridiaceae	Energy production and conversion	25.1	0.000 889
Family	Eubacteriaceae	Energy production and conversion	27.5	0.001 88
Family	Ruminococcaceae	Amino acid transport and metabolism	42.7	0.0233
Family	Ruminococcaceae	Carbohydrate transport and metabolism	47.3	1.51×10^{-6}
Family	Eubacteriaceae	Translation, ribosomal structure and biogenesis	33.9	0.002 35
Family	Deferribacteraceae	Energy production and conversion	31.9	1.29×10^{-6}
Family	Clostridiaceae	Posttranslational modification, protein turnover, chaperones	27.1	0.0224
Family	Muridae	Extracellular structures	58	1.18×10^{-12}
Family	Porphyromonadaceae	Cell wall/membrane/envelope biogenesis	48.1	8.97×10^{-12}
Family	Porphyromonadaceae	Energy production and conversion	37.3	0.002 67
Family	Lactobacillaceae	Energy production and conversion	41.6	0.0093
Family	Muridae	Secondary metabolites biosynthesis, transport and catabolism	50.9	0.006 64
Family	Bacteroidaceae	Translation, ribosomal structure and biogenesis	31.8	0.000 105
Family	Oscillospiraceae	Inorganic ion transport and metabolism	25	0.005 67
Family	Muridae	Intracellular trafficking, secretion, and vesicular transport	82.1	1.96×10^{-11}
Genus	Bacteroides	Carbohydrate transport and metabolism	42.2	0.019
Genus	Bacteroides	Cell wall/membrane/envelope biogenesis	41.6	4.94×10^{-34}
Genus	Bacteroides	Inorganic ion transport and metabolism	42.6	6.93×10^{-43}
Genus	Parabacteroides	Inorganic ion transport and metabolism	36.7	1.77×10^{-9}
Genus	Clostridium	Carbohydrate transport and metabolism	40.7	3.63×10^{-7}
Genus	Mus	Inorganic ion transport and metabolism	64.2	2.84×10^{-20}
Genus	Mus	Amino acid transport and metabolism	55.9	1.89×10^{-32}
Genus	Mus	Posttranslational modification, protein turnover, chaperones	55.1	5.92×10^{-56}
Genus	Lactobacillus	Translation, ribosomal structure and biogenesis	47.2	0.000 274
Genus	Mus	Signal transduction mechanisms	70.7	3.52×10^{-33}
Genus	Roseburia	Cell motility	47.2	0.000 777
Genus	Eubacterium	Carbohydrate transport and metabolism	39.2	3.46×10^{-5}
Genus	Akkermansia	Translation, ribosomal structure and biogenesis	67.2	3.14×10^{-6}
Genus	Oscillibacter	Lipid transport and metabolism	27.5	4.87×10^{-5}
Genus	Akkermansia	Carbohydrate transport and metabolism	59.1	6.67×10^{-6}
Genus	Lachnoclostridium	Carbohydrate transport and metabolism	16	0.001 28
Genus	Dorea	Carbohydrate transport and metabolism	32.9	0.000 657
Genus	Dorea	Energy production and conversion	38	2.98×10^{-7}
Genus	Mus	Carbohydrate transport and metabolism	66.6	2.05×10^{-17}
Genus	Bacteroides	Amino acid transport and metabolism	50.3	3.75×10^{-6}
Genus	Oscillibacter	Energy production and conversion	28.8	8.16×10^{-6}
Genus	Eubacterium	Cell motility	20.1	0.0269
Genus	Lactobacillus	Carbohydrate transport and metabolism	61.5	9.2×10^{-14}
Genus	Mus	Defense mechanisms	87	3.55×10^{-15}
Genus	Oscillibacter	Amino acid transport and metabolism	38.8	5.92×10^{-5}
Genus	Oscillibacter	Carbohydrate transport and metabolism	33.5	2.42×10^{-21}
Genus	Bacteroides	Intracellular trafficking, secretion, and vesicular transport	36.4	2.2×10^{-15}
Genus	Oscillibacter	Cell motility	38.6	0.001 42
Genus	Bacteroides	Energy production and conversion	39.1	2.66×10^{-14}
Genus	Clostridium	Cell motility	46.4	3.56×10^{-7}
Genus	Clostridium	Energy production and conversion	21.8	0.001 25
Genus	Eubacterium	Energy production and conversion	22.8	0.000 397
Genus	Flavonifractor	Energy production and conversion	31.8	0.009 38
Genus	Subdoligranulum	Carbohydrate transport and metabolism	41	0.006 29
Genus	Eubacterium	Translation, ribosomal structure and biogenesis	33.9	0.002 35
Genus	Mucispirillum	Energy production and conversion	31.9	1.29×10^{-6}
Genus	Roseburia	Carbohydrate transport and metabolism	27.4	0.003 53
Genus	Dorea	Amino acid transport and metabolism	34	0.005 68
Genus	Parabacteroides	Cell wall/membrane/envelope biogenesis	48.1	8.97×10^{-12}
Genus	Anaerotruncus	Carbohydrate transport and metabolism	36.2	8.74×10^{-5}
Genus	Lactobacillus	Energy production and conversion	41.6	0.0093
Genus	Mus	Secondary metabolites biosynthesis, transport and catabolism	50.9	0.006 64
Genus	Bacteroides	Translation, ribosomal structure and biogenesis	31.8	0.000 105
Genus	Enterorhabdus	Amino acid transport and metabolism	51.2	0.0476
Genus	Ruminococcus	Carbohydrate transport and metabolism	67.1	0.0106
Genus	Oscillibacter	Inorganic ion transport and metabolism	25	0.005 67
Genus	Mus	Intracellular trafficking, secretion, and vesicular transport	82.1	1.96×10^{-11}
Genus	Butyrivibrio	Carbohydrate transport and metabolism	26.4	0.004 23
Genus	Mus	Extracellular structures	50.3	2.55×10^{-7}
Genus	Dorea	Posttranslational modification, protein turnover, chaperones	33.5	0.04
Species	Lachnospiraceae bacterium A4	Carbohydrate transport and metabolism	31.7	5.56×10^{-6}
Species	Lachnospiraceae bacterium 28-4	Cell motility	40.7	0.0036
Species	Lachnospiraceae bacterium COE1	Carbohydrate transport and metabolism	34.8	0.000 99
Species	Parabacteroides goldsteinii	Inorganic ion transport and metabolism	37.7	1.01×10^{-8}
Species	Clostridium sp. ASF502	Carbohydrate transport and metabolism	50.6	0.000 129
Species	Lachnospiraceae bacterium 28-4	Energy production and conversion	20.8	7.7×10^{-10}
Species	Mus musculus	Inorganic ion transport and metabolism	64.2	2.84×10^{-20}
Species	Mus musculus	Amino acid transport and metabolism	55.9	1.89×10^{-32}
Species	Mus musculus	Posttranslational modification, protein turnover, chaperones	55.1	5.92×10^{-56}
Species	Lachnospiraceae bacterium 3-2	Carbohydrate transport and metabolism	35.4	0.006 64
Species	Mus musculus	Signal transduction mechanisms	70.7	3.52×10^{-33}

Continued on next page

Table A.2 – continued from previous page

Rank	LCA	NOG category	RITZ mean (%)	p-value
Species	Oscillibacter sp. 1-3	Lipid transport and metabolism	34.1	0.000 157
Species	Akkermansia muciniphila	Carbohydrate transport and metabolism	59	3.97×10^{-5}
Species	Lachnospiraceae bacterium M18-1	Translation, ribosomal structure and biogenesis	46.6	0.006 64
Species	Dorea sp. 5-2	Carbohydrate transport and metabolism	32.2	0.0246
Species	Lachnospiraceae bacterium A4	Amino acid transport and metabolism	38.8	0.000 523
Species	Mus musculus	Carbohydrate transport and metabolism	66.6	2.05×10^{-17}
Species	Lachnospiraceae bacterium A4	Lipid transport and metabolism	8.05	0.0463
Species	Lachnospiraceae bacterium M18-1	Cell motility	38.3	0.000 453
Species	Eubacterium plexicaudatum	Cell motility	20.1	0.0269
Species	Lachnospiraceae bacterium A2	Carbohydrate transport and metabolism	45.1	0.000 758
Species	Lachnospiraceae bacterium 28-4	Carbohydrate transport and metabolism	41.9	2.14×10^{-7}
Species	Eubacterium plexicaudatum	Carbohydrate transport and metabolism	40.7	0.0014
Species	Mus musculus	Defense mechanisms	85.6	4.35×10^{-12}
Species	Firmicutes bacterium ASF500	Amino acid transport and metabolism	37.9	0.003 62
Species	Lachnospiraceae bacterium A4	Energy production and conversion	21.7	0.001 17
Species	Firmicutes bacterium ASF500	Cell motility	39.7	0.0259
Species	Oscillibacter sp. 1-3	Cell motility	38.6	0.001 42
Species	Bacteroides vulgatus	Cell wall/membrane/envelope biogenesis	53.5	7.01×10^{-22}
Species	Firmicutes bacterium ASF500	Carbohydrate transport and metabolism	44.8	1.58×10^{-9}
Species	Lachnospiraceae bacterium 10-1	Energy production and conversion	14.5	2.41×10^{-6}
Species	Oscillibacter sp. 1-3	Carbohydrate transport and metabolism	34.6	3.99×10^{-19}
Species	Oscillibacter sp. 1-3	Energy production and conversion	34.5	0.000 193
Species	Flavonifractor plautii	Energy production and conversion	31.8	0.009 38
Species	Subdoligranulum variabile	Carbohydrate transport and metabolism	41	0.006 29
Species	Mucispirillum schaedleri	Energy production and conversion	31.9	1.29×10^{-6}
Species	Lachnospiraceae bacterium A4	Cell motility	46.9	0.001 14
Species	Firmicutes bacterium ASF500	Energy production and conversion	56.5	0.0246
Species	Dorea sp. 5-2	Amino acid transport and metabolism	31.8	0.0144
Species	Parabacteroides goldsteinii	Cell wall/membrane/envelope biogenesis	44.8	2.95×10^{-6}
Species	Anaerotruncus sp. G3(2012)	Carbohydrate transport and metabolism	32.8	0.012
Species	Mus musculus	Secondary metabolites biosynthesis, transport and catabolism	50.9	0.006 64
Species	Enterorhabdus caecimuris	Amino acid transport and metabolism	51.2	0.0476
Species	Oscillibacter valericigenes	Amino acid transport and metabolism	32.3	0.001 66
Species	Oscillibacter sp. 1-3	Inorganic ion transport and metabolism	25	0.005 67
Species	Mus musculus	Intracellular trafficking, secretion, and vesicular transport	82.1	1.96×10^{-11}
Species	Lachnospiraceae bacterium 10-1	Cell motility	34.7	0.001 55
Species	Lachnospiraceae bacterium A4	Translation, ribosomal structure and biogenesis	38.1	0.008 13
Species	Eubacterium plexicaudatum	Energy production and conversion	23.4	0.0266
Species	Lachnospiraceae bacterium 10-1	Carbohydrate transport and metabolism	32.7	5.64×10^{-6}
Species	Firmicutes bacterium ASF500	Lipid transport and metabolism	33.6	0.0135
Species	Dorea sp. 5-2	Energy production and conversion	37.5	0.0147
Species	Dorea sp. 5-2	Posttranslational modification, protein turnover, chaperones	33.5	0.04

Table A.3: Hygeometric tests for over-representation of taxa in RIA hierarchical clusters

name	group	RIAC1	RIAC2
Bacteroidaceae	Family	2.29×10^{-11}	
Bacteroidales	Order	2.34×10^{-25}	
Bacteroides	Genus	2.29×10^{-11}	
Bacteroides vulgatus	Species	0.0223	
Bacteroidetes	Phylum	1.63×10^{-26}	
Bacteroidia	Class	2.34×10^{-25}	
Chordata	Phylum	0.006 58	
Clostridia	Class		0.0164
Clostridiales	Order		0.0178
Firmicutes	Phylum		1.17×10^{-7}
Lachnospiraceae	Family		0.039
Mammalia	Class	0.0118	
Muridae	Family	0.006	
Mus	Genus	0.004 65	
Mus musculus	Species	0.002 32	
Parabacteroides	Genus	2.63×10^{-9}	
Parabacteroides goldsteinii	Species	2.46×10^{-9}	
Porphyromonadaceae	Family	4.75×10^{-9}	
Rodentia	Order	0.0356	

Table A.4: Hygeometric tests for over-representation of taxa in RITZ hierarchical clusters

name	group	RITZC1	RITZC2	RITZC3	RITZC4
Akkermansia	Genus			0.0029	
Akkermansia muciniphila	Species			0.0125	
Akkermansiaceae	Family			0.0029	
Bacteroidaceae	Family		0.005 16	0.002 13	
Bacteroidales	Order			2.21×10^{-18}	
Bacteroides	Genus		0.005 16	0.002 13	
Bacteroides vulgatus	Species			0.0012	
Bacteroidetes	Phylum			2.7×10^{-18}	
Bacteroidia	Class			2.21×10^{-18}	
Chordata	Phylum			1.78×10^{-13}	4.32×10^{-49}
Clostridia	Class	0.0188			
Clostridiales	Order	0.0172			
Clostridium	Genus				0.0363
Clostridium sp. ASF502	Species				0.0227
Coriobacteriia	Class		0.0308		
Firmicutes	Phylum	2.62×10^{-5}	0.0398		
Lachnospiraceae	Family	0.000 85			
Lachnospiraceae bacterium 10-1	Species	0.0156			
Mammalia	Class			5.64×10^{-11}	1.22×10^{-48}
Muridae	Family			2.05×10^{-9}	1.76×10^{-32}
Mus	Genus			4.1×10^{-8}	4.25×10^{-27}
Mus musculus	Species			1.31×10^{-7}	1.71×10^{-26}
Oscillibacter	Genus		0.003 27		
Oscillibacter sp. 1-3	Species		2.19×10^{-5}		
Oscillospiraceae	Family		0.003 27		
Parabacteroides	Genus			4.18×10^{-21}	
Parabacteroides goldsteinii	Species			2.13×10^{-19}	
Porphyromonadaceae	Family			1.18×10^{-20}	
Rodentia	Order			5.09×10^{-10}	3×10^{-35}
Verrucomicrobia	Phylum			0.0029	
Verrucomicrobiae	Class			0.0029	
Verrucomicrobiales	Order			0.0029	

Curriculum Vitae

Patrick Smyth

Resume

AREAS OF EXPERTISE

Reporting skills
Academic research
Cell and molecular techniques
Synthesis and separation techniques.
Analytical techniques
Bioinformatics
Statistics
Software Development
C++
R language

PROFESSIONAL

Bachelor's degree in Biochemistry

PERSONAL SKILLS

Initiative
Highly competent
Efficient with computers

PERSONAL SUMMARY

I have excellent research potential and an ability to actively contribute to the research project's goals as well as a good standard of written English and French. I can interact with all researchers in a constructive, creative and professional manner.
I possess expertise in cell and molecular techniques, synthesis and separation techniques, analytical techniques, biostatistics/informatics, software development, and metaproteomics.

PROFESSIONAL EXPERIENCE

University of Ottawa
MASTERS THESIS September 2018 – December 2020

Developed a software to identify and analyze partially labeled peptides for microbiome dynamics. My research led to a publication in the journal of analytical chemistry (DOI: 10.1021/acs.analchem.0c02070). I was supervised by Dr. Daniel Figeys.

Duties:

- Building and testing software using R and C++.
- Performing the most current and effective methods in bioinformatics and biostatistics for analysis of big data.
- Analyzing gut microbiome data.
- Making presentable figures about research findings.
- Frequent participation in conferences and professional workshops.
- Searching and summarizing literature related to research.
- Producing a final thesis report and presenting the research findings to lab colleagues and committee members.

University of Ottawa
TECHNOMISE PROGRAM January 2018 – December 2020

I participated in various professional development workshops and seminars related to gut microbiome research as part of the TECHNOMISE program. Such workshops include an omics bootcamp and classes on project management, team building, and communication skills.

ASMS
Conference presentation July 2020

Presented a 10-minute talk related to my research at the American Society for Mass Spectrometry (ASMS) conference.

ISMB
Poster presentation June 2020

Presented a poster and an 8-minute short talk at ISMB 2020

AREAS OF EXPERTISE

Reporting skills
Academic research
Cell and molecular techniques
Synthesis and separation techniques.
Analytical techniques
Bioinformatics
Statistics
Software Development
C++
R language

PROFESSIONAL

Bachelor's degree

PERSONAL SKILLS

Initiative
Highly competent
Efficient with computers

KEY SKILLS AND COMPETENCIES

analytical chemistry

- Knowledgeable in the study of activity, chemical equilibrium and titration curves on complexation and neutralization systems.
- Treatment of analytical data.
- Applying methods for precipitation and solubility product.

Spectroscopy

- Knowledgeable in the principles of UV-vis-, IR-, NMR-, CD-spectroscopy, and mass spectrometry, as well as their spectral analysis.

Chemistry

- Knowledgeable in various synthesis techniques
- Liquid and gas chromatography, including other separation and purification techniques.

Cell and molecular biology

- Knowledgeable in the amplification and preparation of DNA constructs.
- Expression and purification of desired proteins.
- Separation of protein, DNA, and RNA.
- Cell culture techniques.

Bioinformatics

- Knowledgeable in algorithms for the classification and analysis of big data.
- Machine Learning and AI
- Metaproteomics
- Software Development.
- Biostatistics (Particularly in Bayesian inference)

ACADEMIC QUALIFICATIONS

BSc (Hons) Biochemistry
Concordia University 2013 - 2018