Implementation of a bioanalytical metaproteomics assay and design of bioinformatics algorithms to investigate microbiome-modulating effects of resistant starches

James Ryan

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements for the
Master of Science degree in Biochemistry (spec. Bioinformatics)

Ottawa Institute of Systems Biology
Department of Biochemistry, Microbiology, & Immunology
Faculty of Medicine
University of Ottawa

ABSTRACT

The human gut microbiome exists as a community of microorganisms in symbiosis with its host. Prebiotics are functional compounds that modulate this microbial community, promoting the growth and activity of bacteria that are beneficial to human health. Resistant starches (RS), a subclass of prebiotics, are compounds linked to a number of host-beneficial effects when included in human diets. Understanding how RS shapes gut flora composition and function is crucial to understanding these effects; however, these effects are clouded by the complexity of the microbiome's interactions. Comprehensively characterizing microbiome shifts as the result of prebiotics is an intriguing bioanalytical problem. In the thesis project, I hypothesize that: RS changes microbiome biochemical pathway expression community-wide and at different taxonomic levels; that RS forms will affect microbiome bacterial taxonomic distribution; and that a linear programming optimization approach can parsimoniously distribute ambiguous peptide abundances amongst their constituent species, leading to different interpretations of functional and structural characteristics in microbiome metaproteomics data. To address these hypotheses, the thesis project utilizes a combined metaproteomics and bioinformatics approach. The Figeys lab-developed RapidAIM bioanalytical assay is deployed to generate label-free mass spectrometry metaproteomics data, testing for these effects experimentally. Further, Cerberus, a bioinformatics platform for microbiome metaproteomics analyses, was developed to integrate workflows from different software sources into a unified pipeline. Cerberus also implements a novel linear optimization approach addressing the shared-peptide problem. Through experimental data analyses using Cerberus, microbiomes encountering RS showed concerted taxonomic shifts, general and specific functional modulations linked to these taxonomic changes, and a significantly variable pathway expression profile for host-beneficial microbiome processes.

The peptide-species linear optimization procedure demonstrates how naïve approaches to the shared-peptide problem greatly skew downstream taxonomic and functional analyses in metaproteomics experiments, marking an important consideration for microbiome studies seeking to resolve taxon-specific alterations.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Daniel Figeys, for the incredible opportunities I have had throughout my Master's degree. Dr. Figeys allowed me to pursue a thesis project combining elements of both undergraduate degrees I received from Dalhousie University (Biochemistry/Molecular Biology and Computer Science) in the exciting and fast-growing field of microbiome metaproteomics, exactly what I was looking to do for my Master's degree. His mentorship and direction focused my thesis during many stages where I wanted to explore concepts that would have not fit with the project's aims, and my time in the Figeys lab has certainly set the stage for my future career in industry.

I would also like to thank my co-supervisor, Dr. Mathieu Lavallée-Adam, for his guidance, support, and encouragement throughout my degree. Dr. Lavallée-Adam and I both started at the University of Ottawa together in 2016 with new surroundings and his enthusiasm for the potential of my project was apparent at every stage. I greatly appreciated his patience as he and I would debate the intricacies of our bioinformatics algorithms, and ultimately our work together shaped a large portion of this thesis. *Go Bruins*.

Many members of the Figeys and Lavallée-Adam labs had a major impact on my thesis, without which this project would not be what it is.

In the Figeys lab, Dr. Leyuan Li was a constant source of guidance during all stages of the project; she taught me the ins-and-outs of the RapidAIM procedure and assisted at every step, patiently answered my questions, and was always open to discussing problems I was having with protocols or concepts. Considering the amount of projects Dr. Li was involved in during my time at uOttawa, I am continually amazed that she managed to not only have time to help with my experiments and answer my questions but went as far as she did to ensure my experiments succeeded. Dr. Elias Abou Samra and I collaboratively

developed much of Cerberus's functionality since May of 2018, with him giving me direction and myself doing the programming. We spent a lot of time together working out the wrinkles and defining what would be useful for a software that streamlines microbiome metaproteomics analyses, and I greatly appreciate his guidance in this regard. Dr. Janice Mayne really got my project off the ground, helping me with experimental design, defining objectives, contacting companies for samples, editing reports and this thesis, and building a picture of what I could achieve in my time at uOttawa. She was always a friendly face at RGN, and her support shaped my project in many positive ways. Drs. Zhibin Ning, Xu Zhang, and Kai Cheng always had an open door and friendly ears for listening to me work through issues throughout the project, explaining experimental and bioinformatics concepts in ways that gave me a sense of direction during uncertain times. Drs. Amanda Starr, Lioudmila Tepliakova, Shelley Deeke, and Krystal Walker all made a huge difference during my time in the wet-lab, helping me navigate unfamiliar instruments and protocols.

My desk was in the Lavallée-Adam lab, and the people in that office made my degree experience very memorable. I thank Francesca Barry, Alexander Pelletier, Linh Nguyen, Rachel Nadeau, Dallas Nygard, Soroush Shahryari Fard, Emily Roth, and many others for listening to my lab meeting presentations and their patience with my bad jokes. All of them being around my age, it's nice to be confident that we will all find success through our time with Mathieu.

I would also like to thank my family for their encouragement and support during my degree. Moving away from Nova Scotia was a big decision for me but their positive outlook and understanding lead me to make the decision to come to Ottawa and start my Master's. I'm writing this acknowledgements section from my couch at home in Dartmouth, reminding me that everything comes full circle in one way or another. My girlfriend Chantal has been a

major source of support through the last year and a half, and I can't begin to describe how much I have appreciated her patience and understanding while I spent 12-hour days and weekends hunched in front of my computer debugging code. Her sacrifices truly allowed me to finish this thesis.

TABLE OF FIGURES

TABLE OF CONTENTS

1 INTRODUCTION

The human gut microbiome is a community of microorganisms living in symbiosis with its human host [1]. Microbiomes are composed of many organisms, including communities of bacteria, viruses, fungi, protists, and archaea [1]. Through interactions between its constituent species and its host, the gut microbiome affects human health and disorder [2]. Prebiotics are substances that positively modulate microbial communities in a way that benefits the community's host [3]. Positive modulations of the human gut microbiome by prebiotic substances provide a range of host benefits, including promoting immune-system function and protecting from disease [4, 5]. For instance, prebiotics play a role in protecting the host against colorectal cancer (rev. in [6, 7]) and inflammatory bowel diseases [8].

Resistant starches (RS) are prebiotic polysaccharides that resist digestion by pancreatic amylase in the small intestine and are not hydrolyzed to D-glucose. RS can reach the colon, where they can be fermented by the constituents of the gut microbiome [9, 10]. The microbiome uses RS as an energy source, and produces RS fermentation by-products such as short-chain fatty acids (SCFAs) [9, 11], compounds linked to a range of intestinal benefits [11].

Although the host benefits from RS have been demonstrated (rev. in [12]), whether resistant starches have all similar effects on the microbiome and whether individual microbiomes respond differently to resistant starches remain unclear. This is further complicated by the scale and complexity of the immense gut bacterial network: around four trillion bacteria [13] split amongst 500-1000 species [14] comprise this microbiome. Understanding the molecular pathway responses of microbiota upon prebiotic treatment and

the effect of prebiotics on microbiome composition is critical to grasping microbiome dynamics and how interventions can modulate this system.

Omics based approaches have been extensively used to investigate changes in microbiome compositions and functions. In particular, metagenomics can provide gene content and encoded functional attributes from gut microbiome samples [15] and incorporates 16S ribosomal sequencing [16–18]. Metatranscriptomics delves into primary gene expression products to resolve activity profiles of bacteria above the underlying microbial abundance and structure in the microbiome [19, 20]. Quantitative metaproteomics provides a functional view of a microbiome that differs from metatranscriptomics and metagenomics in that it moves past functional precursors (DNA and RNA) to resolve proteins, which affect the vast majority of cellular interactions, processes, and phenotypes. Label-free mass spectrometry (MS)-based metaproteomics analyses provide powerful means to identify and quantify proteins, which constitute biochemical pathway effectors in microbiomes' bacteria, such as enzymes and signalling proteins [21]. Modern study of the human gut microbiome benefits from metaproteomics in that it provides protein abundance data that can elucidate actual microbial function within a microbiome [22].

Recent investigations in the field of gut microbiome metaproteomics have assessed *in vivo* responses to probiotic intervention [23], examining mutualistic interactions between microbiomes and their human hosts [24], and determining microbial signatures of Crohn's disease [25]. Metaproteomics approaches elucidate the ways that microbiomes interact internally in and externally to their communities and lend themselves well to further study of prebiotics' effects on human gut microbiota.

However, significant challenges do exist in metaproteomics-directed microbiome analyses. These include: the complexity and scale of microbiome communities [26] leading

to issues identifying low-abundance proteins; the large amount of homologous proteins present between different bacterial species, an issue compounded by horizontal gene transfer between microbiome bacteria [27]; and the proper distribution of abundance values for 'shared' peptides (peptides whose sequences match more than one protein subsequence entry in a sequence database) between taxa in a microbiome. Therefore, there is a need to develop methods to address limitations of metaproteomic approaches. This thesis tackles the problem of shared peptide distribution in metaproteomics experiments, critically important to human gut microbiome investigations as accurate characterization of this community's biochemical mechanisms can elucidate disease-modulating and host-protective effects.

2 BACKGROUND

2.1 RESISTANT STARCH

2.1.1 Starch structure

Starch is a polysaccharide molecule that acts as the main source of carbohydrates in a human diet. It functions as the main reserve polysaccharide in higher plants [28], who synthesize starches inside plastids to be deposited in other plant storage organs such as roots, tubers, fruits, grains, seeds, leaves, and stems [29]. There are two polysaccharide homopolymer forms that are found in starch: amylose, a linear polysaccharide with α–(1→4) glycosidic bonds between D-glucose moieties, and amylopectin, a branching polysaccharide consisting of "short" (< 36 moiety) and "long" (> 36 moiety) saccharide chains stemming from α–(1→6) branch points that exist on a linear α–(1→4) backbone [28, 30]. The ratio of amylose and amylopectin varies in naturally-found starches. The definition of starch encompasses many different mixtures of these molecules, and also covers variability in starch granule size and shape, amylopectin structure branching and chaining, and the

arrangement of amylose and amylopectin into crystalline and amorphous regions inside the insoluble starch granules [28].

The human body digests starch in a multi-stage process as it passes through the gastrointestinal tract, beginning with salivary amylase hydrolyzing starch bonds in the mouth, then further hydrolysis by pancreatic amylase in the small intestine [9, 10]. As starch is broken down in these areas of the body, its individual sugar moieties become unlinked, producing free glucose, fructose, and galactose [31]. The fraction of starch that successfully undergoes this breakdown process and is absorbed by the gastrointestinal tract is known as "available carbohydrates", a term that encompasses these broken-down starches as well as lactose and fructose [31]. This necessitates the classification of the remaining starch fraction, known as "resistant carbohydrates". Resistant carbohydrates include: RS; non-starch polysaccharides (NSP), a term describing dietary fibre (intrinsic plant cell wall polysaccharides); resistant short-chain carbohydrates (RSCC); and sugar alcohols [31].

2.1.2 Resistant starch classification

The term "resistant starch" was first used by Englyst *et al.* in 1982 to describe "retrograded, enzyme-resistant starch … [that] can be solubilised with 2M potassium hydroxide solution [and] hydrolyzed to glucose with amyloglucosidase" [32]. Currently, relevant literature differs on the definitions and number of RS structural types [9, 10, 12, 33–35]. Resistant starch occurs naturally in three different forms (RS1/RS2/RS3), and synthetically as a fourth (RS4) [10, 12, 33–35] and fifth (RS5) form [9, 12, 31, 34].

The RS1 exists as a physically-inaccessible starch because of its protection by a cell wall in partially-milled grains [9, 10, 12, 33–35]. The protein matrix and cell wall material that constitutes cereal grain and seed endosperm surrounds starch, hindering its digestibility and reducing host glycaemic response to its presence [36]. There exists an inverse

relationship between RS1 content and cell wall integrity, as increased milling of whole seeds leads to cell wall breakdown [37]. Cooked whole legume seeds or kernels maintain an integral cell wall and protein matrix, causing decreased water penetration into their internal starch. The internal starch therefore cannot gelatinize and swell, which limits is availability to enzymes that would hydrolyze it [12]. Foods containing RS1 include extruded durum wheat pasta (semolina) with its high protein content and hard texture [12, 38], and breads made with whole or coarsely ground kernels or grains [12, 39].

The RS2 is described as an enzymatically-resistant starch [9, 12, 33–35] or starch that has poorly-gelatinized [10]. Cooking starches that could be classified as RS2 may cause loss of their enzymatic resistance because cooking causes proper gelatinization [12]. Jane *et al.* [40] list green banana starch, uncooked potato starch, gingko starch, and high-amylose maize starch with B- and C-type polymorphs as examples of RS2 [40]. However, some RS2 starches do exist that exhibit a higher gelatinization temperature than the boiling point of water (100℃), such as high-amylose starch produced by mutations of the *amylose-extender* (*ae*) and branching-enzyme I genes. In these mutated starches, branch chains of intermediate components have much longer lengths and the amylose proportion is higher, factors that increase the gelatinization temperature [12, 41–43].

The RS3 is 'retrograded' amylose and starch – starch that has been cooked and then cooled – forming long-branched chains whose double-helical structure prevents hydrolyzation by digestive enzymes because it cannot fit inside the binding site of amylase [9, 10, 12, 33–35]. This double-helical structure tends to form near refrigeration temperatures (4-5℃) when there is suitable moisture [12]. Once retrogradation has taken place, retrograded amylose cannot be dissociated by cooking, and its gelatinization only occurs at

high temperatures up to 170℃; however, this gelatinization temperature decreases proportionally inverse to its chain length [12].

The RS4 is broadly defined as chemically-modified [9, 10, 33–35]; this chemical modification can be the result of esterification, etherification, or cross-bonding with a number of other chemicals and functional groups to decrease enzyme digestibility (rev in [44]), or by inducing cross-linking, conversion, and substitution that creates atypical moiety linkages [45]. Highly cross-linked starch loses its ability to swell during cooking, causing it to remain in an enzymatically-inaccessible granular form after cooking [12]. Examples of chemical derivatives that provide enzymatic resistance to starch are octenyl succinic [46] and acyl [47] groups: the structural changes that a starch exhibits after their addition partially restricts its hydrolysis, but, intriguingly, bacterial enzymes can still hydrolyze unmodified regions of the starch, making it available as a fermentable sugar that gut bacteria can ferment to produce short-chain fatty acids.

The RS5 has been described in a few ways: as a complexed form of amylose and polar lipids added artificially to foods or because of processing and cooking [9, 12, 31, 34, 37]; and as an amylase-resistant polysaccharide consisting of a water-insoluble linear poly-α-D-1,4-glucan [48]. Specific to polar lipid complexed-RS5, its resistance is a result of single-helical complex between starch amylose and lipids, which prevents starch binding and cleavage [12]. Starch amylopectin also becomes entangled with the amylose-lipid complex, reducing its swelling capacity and preventing hydrolysis [49, 50].

2.1.3 Factors affecting starch digestibility

There are many factors that affect the digestive resistance of a starch, and, by extension, its RS content. These factors can be broken down into environmental, mechanical, and biological factors. Environmentally, the physical form of grains and seeds and its crop-

growing location play a role in digestibility [37]. Biologically, the size and type of starch granules [37], starch association with other food components (lipids, proteins, sugars, gums, other fibre types and plant bioactives that may inhibit human enzymatic activity [51]), the genotype or cultivar of the plant [52–54] and any mutations that it may have as a result of RS-altering genetic engineering [52, 53] play a role in digestibility. Further, the amylose content of a starch defines its digestibility to the extent that amylose is digested slower than amylopectin; because of this, the RS content of a starch is positively associated with the amylose content of that starch [37]. Birt *et al.* [12] assert that "… the single structural aspect with the greatest influence on [starch] digestibility is the degree and type of crystallinity within the [starch] granule … [and that] starch with long, linear chains has a greater tendency to form long crystalline structures than starch with short, highly branched chains". Mechanically, food processing methods such as cooking, milling, high-pressure processing, autoclaving, ultraviolet type-C irradiation, extrusion (forcing material through a die), storage time, and annealing (physical modification of starch in water at temperatures below gelatinization) all affect digestibility [50, 55–63]. In addition to these mechanical methods, grinding of grains reduces RS1 content because of cell wall breakdown [37] and cooking starchy foods in water causes starch gelatinization leading to increased digestibility [37]; however, cooking and then cooling starch (retrogradation) can produce RS3 [37, 64].

2.2 RESISTANT STARCH AS A MICROBIOME MODULATOR

On average, 10-60 g of resistant carbohydrates are available to the large intestinal microbiota each day [65, 66]. These RS forms escape digestion and are present intact in the large intestine, making them available for the human gut microbiome to use as an energy source. Microbiome bacteria's fermentation of resistant starch produces a number of compounds, predominantly gases (carbon dioxide, hydrogen, and methane) and SCFAs [12].

Resistant starch polymers are degraded into glucose, then bacterial glycolysis produces SCFA and (to a lesser extent) other organic acids such as lactate, succinate, and formate. Those bacterial species that are able to metabolize RS as it becomes available in the gut have a growth advantage over other bacteria, as they can utilize RS as an energy source [67].

2.2.1 Bacterial RS import from the gut environment

Bacteria must first bind to starch granules so that they can bring them inside their cells to be processed. A few different complexes have been described that function as complex glycan importers.

The *sus* (starch-utilization-system) gene cluster was prototypically found in *Bacteroidetes thetaiotaomicron* [68] with derivatives later found in other *Bacteriodetes* species (termed '*sus*-like systems' and the '*sus* paradigm') [69, 70]. The left side of figure 2.1 shows an overview of its function. *sus*-like systems express a cell envelope-associated multi-protein group SusRABCDEFG that binds and imports complex glycans. The proposed mechanism for starch import and degradation through the *sus* system is as follows [69, 70]: SusD interacts with starch molecules, anchoring them close to SusG (an α-amylase) which hydrolyzes starch bonds. The resulting maltooligosaccharides are presented to SusC, a β-barrel integral protein that releases these sugars into the periplasm where they are hydrolyzed by SusA (an α-amylase) and SusB (a pullulanase) into smaller sugar chains. SusR transcriptionally regulates the activity of the *sus* cluster, inhibiting its function in response to cellular availability of maltooligosaccharides, amylose, amylopectin, and pullulan. The components SusE and SusF of *sus*-like systems have been described as facilitators of insoluble, granular starch import [69] – starches such as RS – because they increase the *sus* complex's overall affinity for starch [69].

The ‘cellulosome’ also serves as a system for starch import [71]. The right side of figure 2.1 shows an overview of its function. It is a multi-component enzyme complex localized to the extracellular space, found in anaerobic bacteria. In *Ruminococcus*

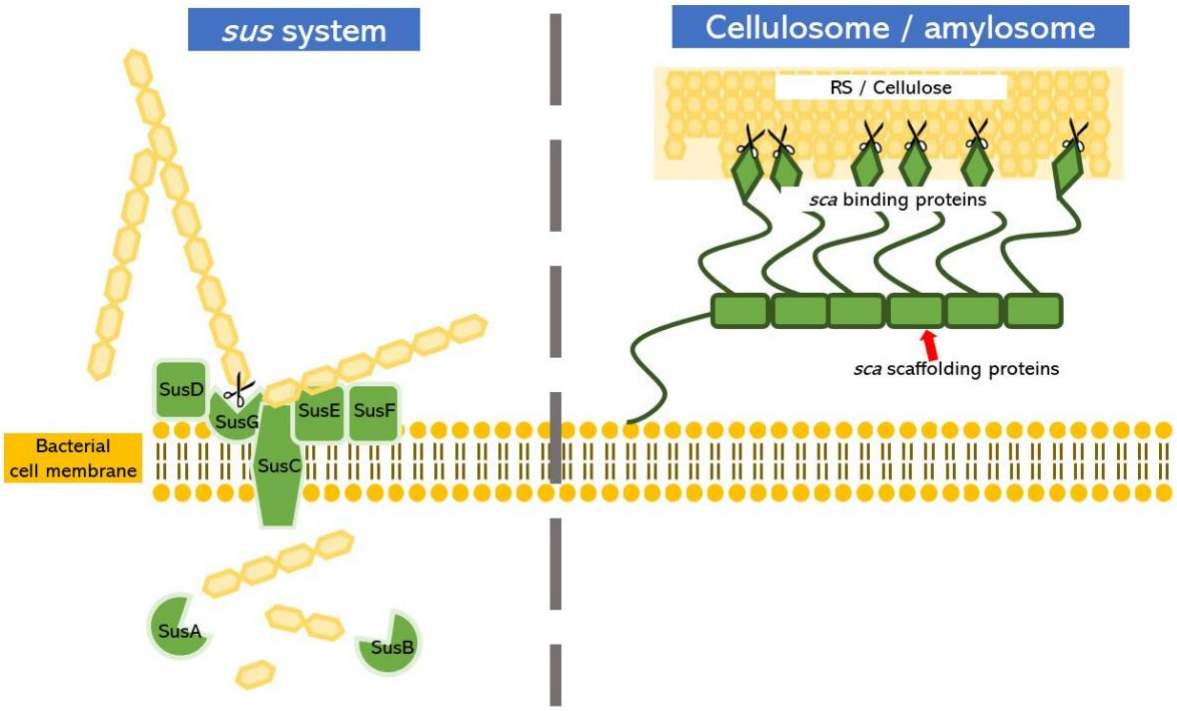


Figure 2.1. Graphical representations of the *sus*-like and cellulosome/amylosome bacterial RS/starch import systems [68].

flavefaciens, the cellulosome is encoded by the *sca* gene cluster, producing scaffolding and binding proteins that import insoluble starches into the cell [71]. The 'amylosome', described in the bacterium *Ruminococcus bromii* [72], is another highly-specialized group of enzymes whose activity closely relates to the cellulosome found in *Ruminococcus flavefaciens*, promoting the concept of *Ruminococcus Bromii* as a keystone species for RS degradation in the human gut. This amylosome acts on the cell surface or extracellularly, binding extracellular starches through dockerin and cohesin protein interactions [72].

Butyrate-producing bacteria may bind RS through a cell-envelope-associated α-amylase: these include *Butyrovibrio fibrisolvens*, *Roseburia intestinalis*, and *Roseburia inulinviorans* [73].

Cooperative RS intake also appears to exist in the human gut microbiome. Here, a system is built between multiple bacteria relying on one or more species to initiate RS degradation, releasing easier-to-process sugars that the remaining bacteria then intake. Milani *et al.* [74] propose that *Bifidobacteria* species exhibit trophic relationships in their carbohydrate resource sharing as they catabolize complex glycans. Mahowald *et al.* [75] describe how, in the mouse intestine, *Bacteroides thetaiotaomicron* responds to the presence of *Eubacterium rectale* by expressing glycoside hydrolases in a *sus*-like system to degrade complex glycans. In response, *Eubacterium rectale* downregulates production of glycan-degrading enzymes and upregulates production of oligosaccharide transporters, suggesting that *Eubacterium rectale* utilizes glycan fragments released by *Bacteroides thetaiotaomicron*'s complex glycan degradation. This 'cross-feeding' may be mediated by *Bacteroides* species' ability to produce extracellular vesicles containing glycoside hydrolases, shifting complex glycan processing to the extracellular milieu and making the resulting cleavage products more available to species such as *Eubacterium rectale* [76]. The

donor organism may receive some kind of benefit; however, the exact nature of the trade-off is unclear in the *Bacteroides thetaiotaomicron*/*Eubacterium rectale* relationship. A different commensal trade-off relationship has been described by Rakoff-Nahoum *et al*. [77], where two species of *Bacteroides* (*ovatus* and *vulgatus*) exhibit reciprocal benefit: after initial inulin (a complex glycan) digestion by *ovatus, vulgatus* utilizes the remaining sugars, leading to an increase in *ovatus* fitness through an unknown mechanism.

2.2.2 Biochemical mechanisms of SCFA production

Starch is broken down into glucose through the actions of bacterial enzymes such as α-amylase (α–(1→4) bond hydrolysis), type-I pullulanase (α–(1→6) bond hydrolysis), and amylopullulanases (type-II pullulanases) that hydrolyzes both types of bonds [12, 78]. Once glucose is freed from RS through these enzymes and the starch intake systems, SCFAs can be synthesized. The main SCFAs produced are formate, acetate, propionate (or propanoate), and butyrate (or butanoate) [79].

A number of pathway mechanisms for SCFA synthesis exist through alternative fermentation of hexose sugars (which comprise RS) [80, 81]. Figure 2.2 gives an overview of these mechanisms. Hexose sugars enter glycolysis, forming phosphoenolpyruvate (PEP), which can then enter four different pathway branches to form acetate, butyrate, or propionate. Two of these pathways involve PEP conversion to pyruvate through the enzyme pyruvate kinase as a first step, with pyruvate then being converted to either acetyl-CoA or oxaloacetate. In the case of pyruvate-to-acetyl-CoA conversion, butyrate can be formed through acetyl-CoA's entry into the butyrate kinase pathway or the butyryl-CoA CoA-transferase pathway. In the case of pyruvate-to-oxaloacetate conversion, propionate can be formed through the acrylate pathway or succinate pathway. To form acetate, pyruvate can be converted to acetyl-CoA by the Wood-Ljungdahl pathway or by direct conversion through

the enzyme pyruvate decarboxylase. Acetyl-CoA can then be converted to acetate through the enzyme acetyl-CoA synthetase. Further use of acetate as a butyrate precursor is described by Duncan *et al*. whereby *Roseburia intestinalis* uses butyryl-CoA:acetyl-CoA transferase as a CoA acceptor to convert butyryl-CoA, the final precursor to butyrate in the butyrate kinase pathway, into butyrate (and acetate into acetyl-CoA) [82]. Lastly, the propanediol pathway can produce propionate in bacteria through conversion of fucose and rhamnose (sugars commonly localized to intestinal mucin) to dihydroxyacetone phosphate (DHAP) that is then processed to propionate [80, 81, 83]. Further, amino acid analog pathways exist to convert glutarate, lysine, and γ-aminobutyrate (GABA) to butyrate (through conversion to the butyrate precursor crotonyl-CoA) [84], suggesting that microbes are resilient to switches in nutritional availability of hexose sugars and consequently highlighting the importance of SCFAs as bacterial metabolites.

An analysis by Louis *et al.* noted the relative rarity of the *buk* (butyrate kinase) gene in butyrate-producing bacterial isolates when compared to the *ptb* (butyryl-CoA:acetate-CoA transferase) gene, suggesting that in the majority of butyrate-producing bacteria, butyrate is formed through acetate usage rather than butyrate kinase pathway-mediated pyruvate-to-butyrate conversion [85]. However, a metagenomic analysis of fecal bacteria found *buk*-related genes being an enriched Clusters of Orthogonal Groups (COG), while *ptb*-related genes were not enriched [15]; taken together with Louis *et al.*'s analysis, it can be said that the individual contributions of *buk* and *ptb* to overall butyrate production is undefined.

2.2.3 Taxonomic distribution of SCFA-producing bacteria

Starch fermentation is carried out by three major phyla, *Firmicutes*, *Bacteroidetes*, and *Actinobacteria*, which account for 95% of total gut bacteria in mammals [12]. The distribution of SCFA-producing biochemical pathways expressed in gut bacteria varies:

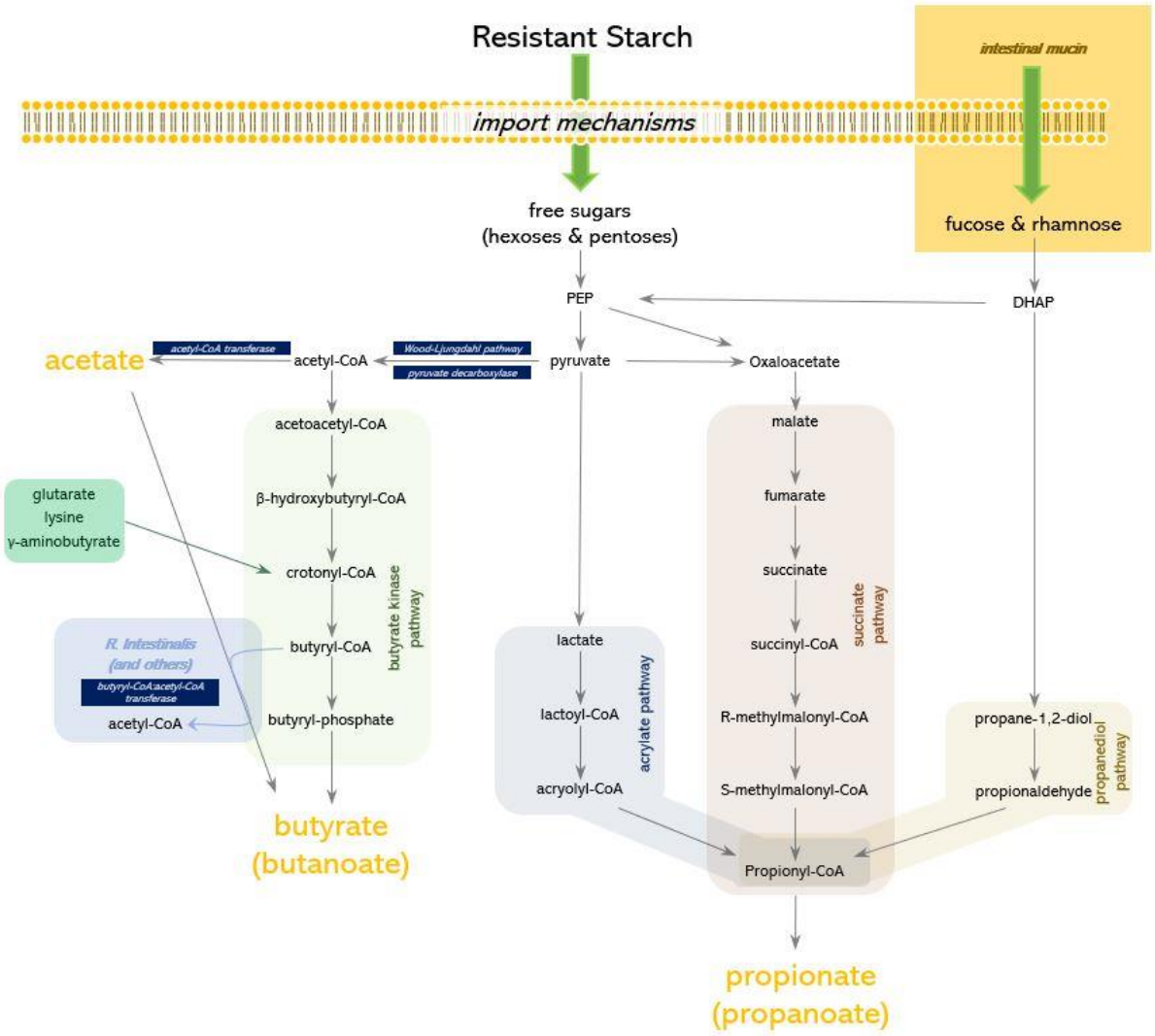

Resistant Starch
intestinal mucin
import mechanisms
free sugars
(hexoses & pentoses)
fucose & rhamnose
PEP
DHAP
acetate
acetyl-CoA transferase
acetyl-CoA
Wood-Ljungdahl pathway
pyruvate decarboxylase
pyruvate
Oxaloacetate
acetoacetyl-CoA
β-hydroxybutyryl-CoA
glutarate
lysine
γ-aminobutyrate
crotonyl-CoA
butyrate kinase pathway
R. Intestinalis (and others)
butyryl-CoA:acetyl-CoA transferase
butyryl-CoA
acetyl-CoA
butyryl-phosphate
butyrate (butanoate)
lactate
lactoyl-CoA
acrylate pathway
acryolyl-CoA
malate
fumarate
succinate
succinate pathway
succinyl-CoA
R-methylmalonyl-CoA
S-methylmalonyl-CoA
propane-1,2-diol
propanediol pathway
propionaldehyde
Propionyl-CoA
propionate (propanoate)

Figure 2.2. Overview of SCFA-producing biochemical pathways in gut microbiome bacteria [80,81,84,85].

pathways producing acetate are widely distributed across phyla, while propanoate and butyrate pathways have so far been described in relatively few genera [79, 83]. Specific to RS fermentation, a small number of organisms seem to be responsible for the majority of butyrate production. These butyrate-forming bacteria are dominated by *Faecalibacterium prausnitzii*, *Eubacterium rectale*, *Eubacterium hallii*, and *Ruminococcus bromii* [86]. *Ruminococcus bromii* acts as a keystone species for RS metabolism: Walker *et al.* show that *Ruminococcus bromii* and its relatives increased from means of 3.8% to 17% of relative abundance in participants' RS-supplemented diets [67], and Ze *et al.* show that its absence from the gut microbiome associates with major decreases in SCFA production [87].

Reichardt *et al*. [83] assessed the phylogenetic associations of human gut bacteria to propionate- and butyrate-producing pathways. The succinate pathway appeared to be the most abundant route for propionate production, with *Bacteroidetes* species heavily associating with succinate pathway enzymes. The acrylate pathway for propionate production was found to be specific for *Firmicutes* species, mainly *Clostridium propionicum* and a few other species to a much lesser degree. Reichardt *et al*. note the low abundance of these acrylate pathway-associated bacteria in their metagenomic analyses, consequently asserting that the succinate pathway is utilized more often than acrylate pathway by the microbiome. Butyrate-producing pathways were associated with *Firmicutes* bacteria, with the butyrate kinase pathway associated to *Coprococcus* species and the butyryl-CoA:acetyl-CoA transferase pathway associated to a wide range of *Firmicutes* species from various genera. The researchers note that most human gut bacteria possessing diagnostic genes for butyrate synthesis (*buk* or *ptb*) lack diagnostic genes for propionate synthesis, suggesting that butyrate and propionate pathways serve as alternative hydrogen sinks whose pathways diverged evolutionarily into different bacterial groups in the gut. However, they did note that

two *Firmicutes* phylum species (*Clostridium catus* and *Roseburia inulinivorans*) were able to produce butyrate and propionate when cultured on different substrates.

2.3 RESISTANT STARCH AS A PROMOTER OF HOST HEALTH

The gut microbiome has been shown to contribute to host immune system development, infection prevention, and absorption of nutrients [88]. The mechanisms of these effects vary, from the microbiome releasing helpful metabolites to its bacteria promoting the growth of host-beneficial strains or preventing colonization of pathogenic strains. Prebiotics (which include RS) act as ecological and functional effectors of the microbiome, whose influence on host health stems more from systemic changes to metabolite usage, protection of the mucus layer, or physiological changes (such as lower gut pH) than from selective induction of taxonomic shifts [89]. Linking RS fermentation into SCFAs to host health outcomes appears to be a central research objective based on available literature. This may be because SCFA production is an easily-measurable condition compared to characterizing microbiome taxonomic or functional shifts in the context of host health.

Availability of SCFAs to host tissues varies: butyrate becomes much less available than propionate or acetate the intestinal epithelium, and propionate is not as available to the liver as acetate [90]. It is therefore useful to consider the effect of each individual SCFA across various biological gradients along with systemic effects in the host. Mechanisms have been uncovered that describe how RS mediates host health outcomes through its fermentation by the microbiome into SCFAs.

2.3.1 Intestinal barrier integrity

Butyrate has been shown to be a crucial metabolite for mediation of intestinal barrier integrity and health. At the intestinal epithelium, it is actively transported into colonocytes

through a sodium (Na^+)-dependent cotransporter [91] and is preferentially used as fuel by colonocytes [92], promoting their proliferation. Butyrate is also an important regulator of tight-junction proteins, which regulate intracellular transport between the lumen and the hepatic portal system [93]. As there exists an association between increased barrier permeability and bacterial translocation (which can trigger inflammatory cascades) [94], butyrate's mediation of barrier integrity provides an anti-inflammatory benefit for the host. Expression of the protein ZO-1 – important to tight-junction function in colonocytes – is increased by colonocyte butyrate intake, and it has also been shown that butyrate reverses aberrant ZO-1 expression in rats, mediating barrier integrity that prevents translocative bacteria-induced inflammatory cascades and reduces hepatic liver injury [95].

2.3.2 Cancer therapeutics

The SCFAs have also been implicated in cancer therapeutic strategies. Butyrate and (to a lesser extent) propionate have been shown to act as histone deacetylase (HDAC) inhibitors [96, 97], compounds that have been widely used as cancer therapies [81]. The HDAC inhibitors prevent deacetylation of histones by HDACs and consequently promote the recruitment of transcriptional machinery to permissive chromatin regions. As mentioned earlier, butyrate induces proliferation of healthy colonocytes, however it also paradoxically induces terminal differentiation and apoptosis in oncolytic cells. This "butyrate paradox" [98, 99] may be explained partly by butyrate's HDAC inhibitive effects causing altered expression of cell differentiation and apoptytic pathways, or that oncolytic colonocytes prefer glucose to SCFAs as an energy substrate (the Warburg effect) and accumulate much more unused butyrate than healthy colonocytes (which oxidize butyrate for energy) leading to HDAC inhibition [98]. Colon cancerous cells may develop from dysregulation of intestinal epithelial cell differentiation, maturation, and apoptosis, all processes that take place rapidly

in intestinal epithelial cells. Butyrate has been shown to induce expression of genes implicated in colonic cell DNA repair in animal studies [100] important because the rapidly-dividing and proliferating intestinal epithelial cells are prone to DNA damage that could influence the aforementioned cellular processes [12]. It has also been shown to modulate signal transduction pathways [101] and increase cancerous cell death by upregulating expression of the apoptytic *Wnt* pathway [102]. Further, a dose-dependent decrease in survival of human colon adenocarcinoma cells *in vitro* has been described [103]. In summary, butyrate may act as a histone acetylation *promoter* in healthy colonocytes and a histone deacetylation *inhibitor* in cancerous colonocytes, provide DNA repair induction properties, and upregulate apoptosis in cancerous colonocytes.

2.3.3 Metabolic disease

Prebiotics have been investigated as therapeutics for metabolic diseases, including diabetes and obesity. Lifestyle changes, including consuming more dietary prebiotics (such as RS), have been associated with improved outcomes for metabolic disease. In diabetes (both types I and II), hyperglycaemia, or increased blood sugar levels, is a daily issue for those affected by the disease, and can lead to systemic tissue toxicity [12]. Factors that influence glycaemic levels include decreased insulin sensitivity, obesity, and attenuated postprandial glucose response [12].

Resistant starch can provide host-protective effects against diabetic symptoms because of its low glycaemic index, and it has been shown that consuming foods with high RS content results in lower postprandial glucose concentrations and insulin response compared to eating foods with normal starches [104–107]. Compared even to regular dietary fibre, RS has been shown to have a more pronounced effect on glucoregulation for obese prediabetic adults [108]. The RSs have also been shown to lower blood glucose, total

cholesterol, and triglyceride levels in diabetic rats, with lower blood glucose levels the result of genetic upregulation of glycogen synthesis and downregulation of gluconeogenesis, and lower total cholesterol and triglyceride levels the result of genetic upregulation of lipid oxidation and cholesterol homeostasis [109]. Conflicting results do exist, however: a 12-week randomized controlled trial on RS supplementation to prediabetic adults showed no significant improvement to glycaemic control [110].

Weight control is also an area of focus for researchers looking at host-protective effects of RS. Increased dietary fibre intake is associated with increased satiety and lowered body mass index (BMI) [111], factors that can influence food consumption. Because RS shares many similarities with dietary fibre, it may aid in weight management in similar ways. Replacement of easily digestible starch with RS lowers the caloric density of foods, a factor associated with increased satiety and weight loss [112]. In humans, increased subjective satiety through dietary RS enrichment has been shown [113–116]; however, other studies describe no change in subjective satiety but quantitative decreases in energy intake [107] and glucagon-like peptide (GLP)-1, a satiety-influencing hormone [117]. A quantitative assessment found that RS-enriched diets associate with increased secretion of GLP-1 and peptide tyrosine (PYY; another satiety-influencing hormone) in rats [112]. In mice, RS effectively ameliorated weight regain in rodents at a rate comparable to exercise, with the regained weight being composed of relatively leaner mass [118].

2.3.4 Differential responses to RS forms

Investigation of RS effects on the gut microbiome *in vivo* describe variable responses to different RS forms. In pig models, RS2 was shown to increase nutrient flow and cause increased levels of *Bifidobacterium* [119] and cause selective proliferation of *Bacillus*, a bacteria genus containing species associated with positive immune-stimulating effects [120].

However, experimental data in humans has shown that SCFA production after RS2 supplementation varies wildly between people [121]. This may be explained by *in vitro* data showing that *Ruminococcus bromii* is a prominent metaboliser of RS2 in single cultures and stimulates RS2 and RS3 metabolism in co-cultures with other species surrounding it [87]. Different sources of RS3 have also been shown to cause differential SCFA production in single cultures [122], underscoring the importance of knowing exactly what the composition of an RS source is. It is important to note, however, that these latter conclusions on single and co-cultures have not been examined in a microbiome capacity, to which *ex vivo* experimentation provides a proxy for *in vivo* microbiome conditions.

2.4 MICROBIOME METAPROTEOMICS

Biochemical processes taking place in the human gut microbiome are difficult to analyse because of the multi-species complexity of organisms within this community. Before the advent of meta-omics approaches using massive parallel sequencing capabilities of recent technologies, researchers relied on culture-based studies to investigate the "post-genomic" content of microbiomes [22, 123]. Culture-based techniques suffer from analytical pitfalls when assessing microbiomes, such as enrichment bias, where standard techniques select for easily-culturable organisms that may not accurately represent human gut microbiome taxonomic structure [123].

Metaproteomics, a term first coined by Rodríguez-Valera to describe abundantly-expressed proteins in environmental samples of microbiological communities [124], is a set of techniques that can capture protein expression in a microbial community through mass spectrometry and bioinformatics. Bottom-up metaproteomics, where protein isolates are proteolytically digested and fractionated into peptides before mass spectrometry analysis, is an established method for complex microbial community analysis. "Shotgun"

metaproteomics couples high-performance liquid chromatography (HPLC) – a method that fractionates peptides by physical and chemical characteristics – to mass spectrometry in a high-throughput manner [21, 125]. Its methodologies can be separated into distinct workflows of sample preparation, peptide fractionation, mass spectrometry, and bioinformatic data analysis [22].

Preparing microbiome samples for bottom-up metaproteomic analysis first involves lysis of microbes to isolate their protein content. While single microbe cultures can be lysed relatively easily, differences between bacterial strains, such as gram-positive and gram-negative cell membranes, confound pan-microbiome cell lysis [126]. Recent refinements to lysis techniques for microbiomes have increased protein yield and identification, such as mechanical disruption methods (bead-beating, ultrasonication) [127] and differential centrifugation of cell components for protein isolation post-lysis [128]. Isolated proteins from microbiome samples must then undergo digestion to produce peptides, often done through tryptic enzymatic cleavage of peptide bonds. This digestion differentiates bottom-up metaproteomics from "top-down" metaproteomics, which maintains whole undigested proteins until their fragmentation into ions during mass spectrometry analysis [129].

Regardless of bottom-up or top-down workflows, prepared peptide (or protein) samples undergo a fractionation procedure to separate the sample mixture by hydrophobicity and charge. As mentioned before, in the case of shotgun techniques, this fractionation takes place through HPLC, although other methods exist to fractionate peptides or proteins based on physical or chemical characteristics. Fractionated peptide (or protein) mixtures then enter a mass spectrometer. In tandem mass spectrometry, proteins/peptides are ionized using an ion source such as an electrospray ionizer (ESI) or matrix-assisted laser desorption/ionization (MALDI) machine. In metaproteomics experiments, ESI is commonly used as it produces a

plurality of ion charge states that can aid in high-mass protein measurement and produce ion fragments that are more easily identifiable than MALDI ions [130]. Fragmented ions are then subjected to two rounds of mass spectrometry selection. In the first round (termed MS1), ions originating from the ion source (termed precursor ions) are separated by their mass-to-charge ratio, or *m/z,* using an electric field in a mass analyzer, such as an Orbitrap [131]. Once selected, a precursor ion undergoes fragmentation into peptide fragments: in bottom-up approaches, collision-induced dissociation (CID) or higher-energy collisional dissociation (HCD) are commonly used for this purpose [132], with electron-transfer dissociation (ETD) used for top-down proteomics because of its improved performance for longer peptide chains compared to CID [133]. These MS1-selected, fragmented peptide ions are then subjected to another round of analysis (MS2), where their m/z values are detected in real-time as spectra and stored in a computer system.

Bioinformatics approaches are used to identify and quantify proteins from mass spectrometry-derived spectral data. In bottom-up approaches, peptide identification takes place with either database search or *de novo* techniques. Database search for peptide sequences requires a protein sequence database that covers expected proteins from a biological sample. In the case of microbiome metaproteomics, the choice of protein sequence database can drastically influence peptide identification performance; this is because of the sequence complexity of a microbiome sample spanning many different microbes, the fact that many of the species present in a microbiome have poor or no sequence coverage in public databases, and that sequences in databases are biased toward easily-culturable, and, therefore, sequenceable species [134, 135].

One method of protein sequence database generation is through metagenomic sequencing of the original microbiome samples, producing a genetic catalog that can be

extrapolated *in silico* to generate expected protein sequences. While this method is in theory ideal for protein database construction as sample-specific data is used, genomic sequencing and assembly errors may lower such a database's coverage of experimentally-derived peptide or protein sequences [126]. Another method involves using publicly-available sequence databases as a basis for *in silico* protein sequence construction, such as the integrated gene catalog (IGC) [136] and UniProt [137]; however, the size of public databases such as these makes false-positive peptide-sequence matches (PSMs) more likely because of its large, non-specific search space [126]. To address this false-positive identification problem, Zhang *et al*. developed MetaPro-IQ [138], an *iterative* database search method for metaproteomics. MetaPro-IQ uses a multi-step process to reduce the search space of a reference database through false discovery rate (FDR) filtering, providing greatly increased search performance for metaproteomic peptide identification and quantification.

2.5 LINEAR PROGRAMMING

In metaproteomics of the human gut microbiome (or any other complex multi-species microbiological sample), the clear majority of peptide sequences found through a label-free tandem-MS experiment will match to multiple protein sequences within a sequence database. In turn, when attempting phylogenetic analyses using these metaproteomics peptide data, these "shared" peptides present a major problem because their non-specific phylogenetic assignment introduces taxonomic ambiguity – if a single peptide sequence can be matched to more than one possible protein (which themselves can map to more than one bacterial species) then a phylogenetic assignment of that peptide becomes unclear at some taxonomic level. An unhelpful pattern develops from this: as one traverses the levels of a sample's bacterial phylogenetic tree from phyla to species, there will be less and less definite taxonomic identifications coming from peptide identifications. Metaproteomics relies on

these peptide identifications to build as complete a picture of the microbiome as possible, so this "shared peptide problem" acts as a compounding limitation to both functional (protein-level) and structural (taxonomic-level) analyses of microbiome metaproteomics data.

Identification of non-redundant or "unique" peptides can serve as a proxy for relative abundances of the whole sample's taxa, as the proportionality of real present taxa should follow the distribution of unique peptides amongst those taxa. MetaPro-IQ [138], a bioinformatics tool developed by members of the Figeys Lab, assigns phylogenetic relationships to positive peptide identifications, and is used in this project. The MetaPro-IQ results assign a phylogenetic resolution for a given peptide, and, for shared peptides, this resolution will only go as far "down" the phylogenetic tree as the lowest common ancestor (LCA) of its assigned species. This presents ambiguity when attempting to distribute shared peptide intensity (abundance) values amongst bacterial species.

Naïve solutions to this peptide distribution problem do exist. One could remove shared peptides from consideration in the experiment, although this reduces the total amount of peptides significantly. One could duplicate the intensity value across all species that match to this peptide, but this increases the total peptide intensity relative to the number of duplication candidate species. A more nuanced method is required: in this thesis, a method is developed for linear optimization of shared peptide intensity values amongst the putative species in a microbiome sample.

2.6 SUMMARY

A metaproteomic lens provides significant analytical power to evaluate the potential modulating effects of three forms of resistant starch (RS2/RS3/RS4) on gut microbiome communities. Implementation of a bioanalytical assay incorporating metaproteomic analyses will provides a base on which further bioinformatics investigations can be performed.

This thesis project uses a metaproteomics and bioinformatics lens to investigate the microbiome-modulating effects of resistant starches and address the shared peptide problem through algorithmic and statistical methods. In relation to resistant starch, I hypothesize that: (1) microbiome biochemical pathway expression will change between different bacterial operational taxonomic units (OTUs) and under different prebiotic conditions, and that (2) RS forms will affect microbiome bacterial taxonomic distribution. In relation to the shared peptide problem, I hypothesize that (3) a linear programming and statistical approach can provide a more accurate distribution for ambiguous peptide abundances, which can lead to changes in functional and structural interpretations of microbiome metaproteomics data.

To address these hypotheses, I have completed an *ex vivo* screening protocol of human gut microbiome samples using RapidAIM (Rapid Analysis of Individual Microbiome), a high-throughput bioanalytical assay with metaproteomic MS data as its result. The goals of this project are to identify and quantify bacterial proteins from gut microbiome samples under different prebiotic conditions, and to design and implement algorithms that identify biochemical pathways a given prebiotic treatment affects and distinguish those bacteria that are driving metabolism at differential levels. The findings of this research show modulating effects of prebiotics – in particular, RS – on gut microbial communities: microbiomes encountering RS display concerted taxonomic shifts, general and specific functional modulations linked to these taxonomic changes, and a significantly-variable pathway expression profile for host-beneficial microbiome processes. Further, the peptide-species linear optimization procedure demonstrates how naïve approaches to the shared-peptide problem greatly skew downstream taxonomic and functional analyses in

metaproteomics experiments, marking an important consideration for microbiome studies seeking to resolve taxon-specific alterations.

3 METHODS

3.1 PREBIOTICS SAMPLE SELECTION

The prebiotics used in this project will be RS2/RS3/RS4, with the inulin-type prebiotic fructooligosaccharide (FOS) acting as a positive control. The FOS is considered to be the preeminent prebiotic, with the weight of evidence supporting its positive function [139–141]. Negative controls will be samples not encountering any prebiotic compound.

Four commercial samples of RS compounds were obtained:

RS2 - Hi Maize 260 [Ingredion, Inc., Westchester, IL, USA] – 60% RS2

RS3 - Novelose 330 [Ingredion, Inc., Westchester, IL, USA] – 28-38% RS3

RS4 - Fibersym RW [MGP Ingredients, Atchison, KS, USA] – 85% RS4

FOS - Orafti P95 [BENEO, Inc., Parsippany, NJ, USA] – 93.2-97.5% FOS

These specific compounds have been cited in previous literature on microbiome responses (RS2 and RS3 in [87, 142, 143]; RS4 in [143]; FOS in [144–146]). The purity of all of these samples was determined by the manufacturers using AOAC method 991.43 [147]. The RS1 and RS5 samples were not obtained. Literature examining RS1/RS5 in an experimental context was also not found during literature review.

3.2 CULTURING, PREBIOTIC TREATMENT, AND SAMPLE PREPARATION

Human stool samples provide a close characterization of the *in vivo* gut microbiome. Human fecal samples were treated with the RS2/RS3/RS4/FOS samples through *ex vivo* culturing of their bacteria. This was done through the Figeys lab-developed RapidAIM procedure (Genome Canada/Ontario Genomics Grant #9408), a high-throughput bioanalytical assay testing for the effects of compounds on a microbiome with

metaproteomics data as its result. Figure 1 gives a general overview of the procedure used in the thesis project to generate mass spectrometry metaproteomics samples.

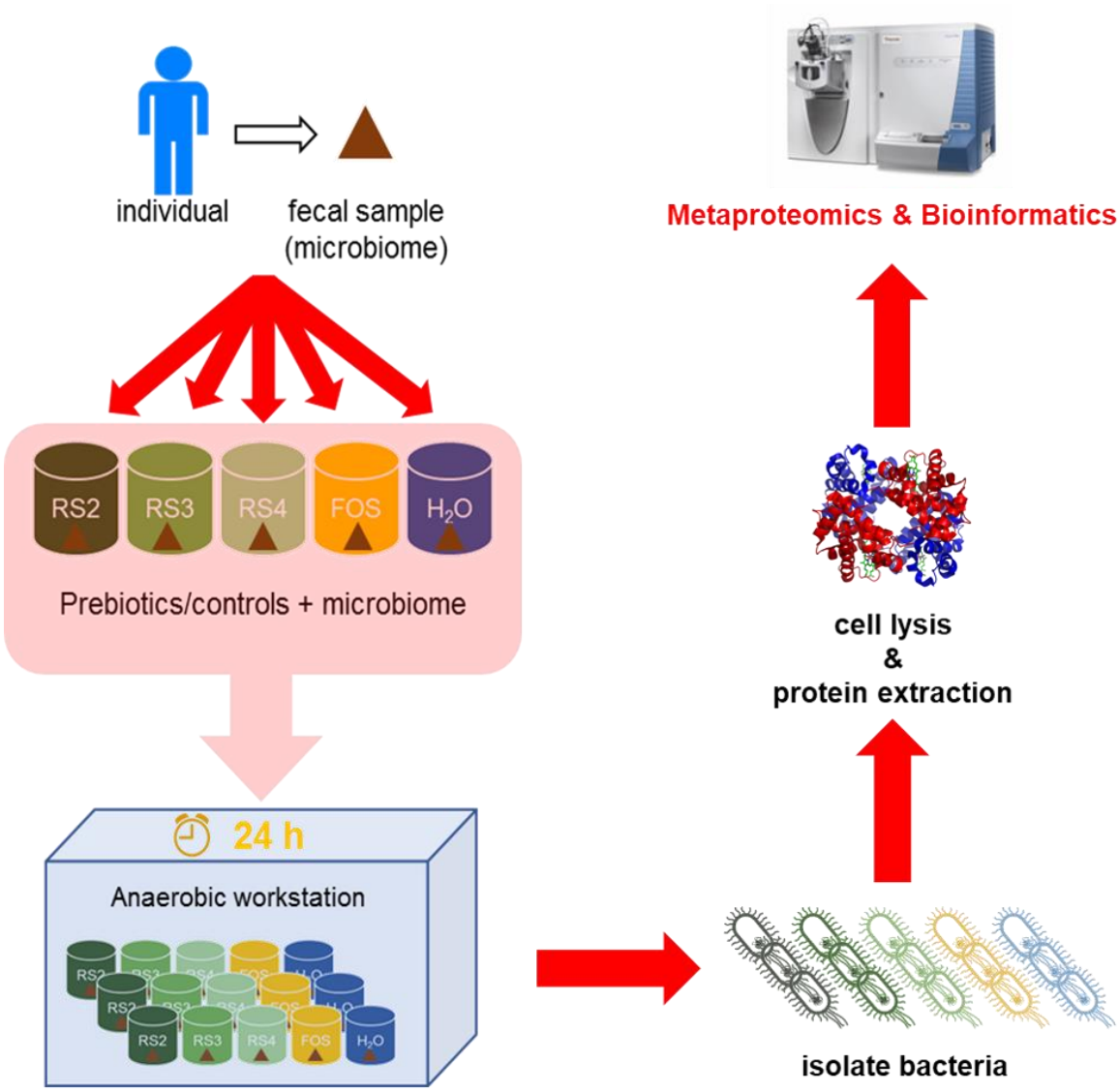


Figure 3.1. Overview of procedure to generate metaproteomics mass spectrometry samples of gut microbiomes from human fecal samples.

3.2.1 Microbiome sample selection and collection

In total, ten stool samples were collected, and six were chosen as final samples for the thesis project (four were excluded because of sample discontinuity or changes in protocol). The Figeys lab has a human feces collection protocol in place (#2016-0585-01H) that is authorised through the University of Ottawa's Ottawa Health Science Network Research Ethics Board. These samples serve as the source of microbiome material for the thesis project.

The collection protocol is as follows: First, potential participants contact the research coordinator to volunteer, and a meeting is set up where the details of the study is discussed. Each potential participant is designated a number to blind researchers from sample identity. On the day of collection, the participant is given a kit containing a 50 mL BD Falcon™ tube containing 12.5 mL of pre-reduced PBS [311-010-CL; Fisher Scientific, Fair Lawn, NJ, USA] (had been placed in an anaerobic chamber 24 hours before collection) and 0.0125 g of L-cysteine [C7352; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (0.1% w/v; added just before sample collection), a sheet of plastic wrap that will be placed over a toilet seat, and a sterile spoon with which to collect the stool sample and place in the 50 mL tube. Once a participant has collected their stool sample, the kit is returned to researchers.

3.2.2 Anaerobic culturing

Each participant's stool sample was cultured in 20 separate tubes: four technical replicates for each of the five different treatment conditions (RS2, RS3, RS4, FOS, and water), with an in-solution stool concentration of 2% w/v. Once collected (as described above), each human fecal sample was weighed to determine if enough fecal matter was collected to reach the 2% w/v inoculation concentration.

In each culture tube, 4 mL of a basic nutrient and salt medium was added prior to stool inoculation. This medium consisted of peptone water [70179; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (0.2 g; 0.2% w/v), yeast extract [212750; BD Biosciences, Sparks, MD, USA] (0.2 g; 0.2% w/v), monopotassium phosphate [P5655; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (0.045 g; 0.045% w/v), dipotassium phosphate [PX1570-1; EMD Millipore Canada Ltd., Etobicoke, ON, Canada] (0.045 g; 0.045% w/v), sodium hydroxide [S8045; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (0.09 g; 0.09% w/v), sodium bicarbonate [SX0320-3; EMD Millipore Canada Ltd., Etobicoke, ON, Canada] (0.4 g; 0.4% w/v), magnesium sulphate heptahydrate [230391; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (0.009 g; 0.009% w/v), calcium chloride [CX0156-1; EMD Millipore Canada Ltd., Etobicoke, ON, Canada] (0.009 g; 0.009% w/v), bile salts [48305; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (0.05 g; 0.05% w/v), Tween 80 [P1754; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (200 µL; 0.2% v/v), and distilled water (~90 mL). This ~100 mL solution was scaled to account for the total number of samples, autoclaved, corrected to pH ~7, and placed in an anaerobic chamber for 24 hours.

Previous literature involving microbiome responses to prebiotics using fecal slurries state sample inoculations with a 1% w/v prebiotic concentration [144, 148]; this concentration was used for this project's culture samples. 0.04 g (1% w/v) of RS/FOS was added to each corresponding culture tube. Fecal samples (in the 50 mL BD Falcon tube, with 12.5 mL PBS and 0.0125 g L-cysteine) were vortex mixed and strained through a cloth gauze to remove solids. The resulting fecal slurry was then added to each culture tube at an amount necessary to reach 2% w/v inoculation concentration. Culture tubes were then placed on a shaking platform (300 rpm) inside the anaerobic chamber at room temperature for 24 hours.

3.2.3 Cell washing

A cell washing procedure was then followed. Culture sample tubes were taken out of the anaerobic chamber and centrifuged for five minutes (300*g*/4°C), their supernatants decanted to 15 mL BD Falcon™ tubes, then centrifuged again at the same settings. This was repeated three times to collect all supernatant.

For each sample, 2.0 mL of supernatant was then added to a 2.0 mL Axygen® tube, then all tubes were centrifuged for 20 minutes (14,000*g*/4°C). This produced a bacterial pellet from bacteria in the supernatant. Each tube's supernatant was then removed so that only the bacterial pellet remained. The bacterial pellet was then washed three times (wash: added 1.5 mL PBS, vortexed, then centrifuged for 20 minutes (14,000*g*/4°C)).

3.2.4 Cell lysis

A cell lysis procedure was then followed. 200 µL of a lysis buffer (10 mL recipe: urea [U5128; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (3.8 g, 38% w/v); 20% SDS [L3771; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (2 mL, 20% v/v); 1M Tris-HCl [C4706; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (0.5 mL, 5% v/v); ddH_2O (4 mL, 40% v/v); PhosSTOP™ tablet [4906837001; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (one tablet); cOmplete mini™ [4693124001; ; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (one tablet)) was added to each sample culture tube. Each sample was then ultrasonicated (25% amplitude) for 30 seconds, placed on ice for 30 seconds, then ultrasonicated again (same setting) for 30 seconds. Sample tubes were then centrifuged for 10 minutes (16,000*g*/8°C). Supernatants were transferred to new 2.0 mL Axygen® tubes, then 1 mL of a -20°C pre-chilled precipitation solution was added (50% v/v acetone [A949-4; Fisher Scientific, Fair Lawn, NJ, USA]; 50% v/v ethanol [1009; Commercial Alcohols,

Tiverton, ON, Canada]; 0.1% v/v acetic acid [A38-212; Fisher Scientific, Fair Lawn, NJ, USA]). Proteins in samples were then precipitated overnight at -20°C.

3.2.5 Isolated protein sample preparation

A protein washing procedure was then followed. Samples were centrifuged for 25 minutes (16,000*g*/4°C), then supernatants were removed. The following was repeated three times: 1 mL of cold (-20°C) acetone was added to each tube, then tubes were placed in an ultrasonic bath to resuspend the proteins in solution, samples were centrifuged for 10 minutes (16,000*g*/4°C), then supernatant removed through a vacuum suction tip in a fume hood. This resolved a washed protein pellet in each tube.

An in-solution protein digestion procedure was then followed. Samples were re-suspended in 100 µL of ABC buffer (10 mL recipe: urea, 3.6 g, 36% w/v; 1M ABC [113164; EMD Millipore Canada Ltd., Etobicoke, ON, Canada] (Stock soln., 0.5 mL, 5% v/v); dH_2O (7 mL, 70% v/v)). Then, a DC™ assay [Bio-Rad Laboratories, Mississauga, ON, Canada] was performed to determine protein content in each sample tube. Once complete, 50 µg of protein from each sample was aliquoted into new 2.0 mL Axygen® tubes for in-solution digestion, with the remainder stored at -20°C. 2 µL of DTT (Cleland's reagent) [43815; Sigma-Aldrich Canada Co., Oakville, ON, Canada] was added to each tube, then tubes were mixed on for 30 minutes on an Eppendorf Thermomixer C (800 rpm/56°C). 2 µL of IAA (iodoacetamide) [I1149; Sigma-Aldrich Canada Co., Oakville, ON, Canada] was then added to each sample, and tubes were placed in darkness for 40 minutes. 4 µL of a trypsin solution containing 1 µg trypsin [T1426; Sigma-Aldrich Canada Co., Oakville, ON, Canada] (lyophilized trypsin, 50 µg, 25% w/v; 200 µL 50 mM ABC) was added to each tube, then tubes were incubated on an Eppendorf Thermomixer C (800 rpm/37°C) for 24 hours.

A desalting procedure was then followed. Elution columns were made from 20 µL pipette tips with their tip end sliced off using a razor blade. 50 µL of a C18 bead [ReproSil120 C18-AQ; Dr. Maisch GmbH, Ammerbuch-Entringen, DE] solution (500 mg C18 beads, 5 mL acetonitrile [34851; Sigma-Aldrich Canada Co., Oakville, ON, Canada]) was added to each tip, acting as the stationary phase for elution of proteins. C18 beads in the elution tips were activated (200 µL of acetonitrile added to each tip, then centrifuged for 2 minutes (200*g*, room temp.)) then equilibrated (200 µL 0.1% v/v formic acid [F0507; Sigma-Aldrich Canada Co., Oakville, ON, Canada] added to each tip, then centrifuged for 2 minutes (200*g*, room temp.)). Samples were then acidified to pH ~2-3 using 5% v/v formic acid and loaded into elution columns (centrifuged as before) until all of sample was added. Elution columns were then washed by adding 200 µL 0.1% v/v formic acid and centrifuging as before until all wash solution had eluted through column. Finally, an elution solution (260 µL of 80% v/v acetonitrile and 20% v/v of a 0.1% v/v formic acid solution) was added to each elution column whose elutant was collected in new sample tubes after centrifugation for two minutes (100*g*/room temp.).

3.2.6 Mass spectrometry

All samples were then placed in a Speed-Vac [Savant™ SPD131DDA SpeedVac™ Concentrator; Thermo-Fisher Scientific, Waltham, MA, USA] to evaporate supernatant. After solubilizing samples with 50 µL 0.1% v/v formic acid, samples were vortex mixed and 2 µL of each was placed in a 96-well MS plate. All samples were run on an Orbitrap Elite™ Hybrid Ion Trap Mass Spectrometer [Thermo-Fisher Scientific, Waltham, MA, USA] with a four-hour gradient, and spectral data was collected as .RAW files.

Metaproteomics MS analysis allows for protein identification and quantification and serves as a basis for identifying biochemical pathways that are differentially expressed.

The following computational tools will distinguish bacteria that have increased or decreased activity relative to the community and will illuminate differential metabolic profiles between bacterial species microbiome compositions across prebiotic treatments.

4 RESULTS

4.1 CERBERUS

4.1.1 Outline

Gleaning important information from mass spectrometry data involves a massive amount of data analysis and organization. In metaproteomics experiments, this process requires a deep understanding and application of biological, statistical, and bioinformatics concepts. Researchers in the -omics space look to software implementations that streamline the process of data organization and visualization to save time and for ease of use. Examples of software packages for -omics workflows are iMetaLab [149], MetaboAnalyst [150], Perseus [151], and Pathview [152]. While each software package contains several exclusive features, an -omics project may require synthesis of results from one or more software packages or workflows, and even involve outside programming or scripting to streamline and automate data analysis. For example, a limitation of iMetaLab is that it cannot process MaxQuant output created through a different program as its design as a pipeline requires .RAW files as inputs. Perseus, MetaboAnalyst, and Pathview are built to be broad analysis pipelines with varying functions, and their designs necessitate that they can handle input files that may not originate from a microbiome metaproteomics experiment. Further, these software (excluding iMetaLab) do not perform the initial database search procedure for .RAW files, requiring this processing be done by some other software package.

A metaproteomics experiment using tandem-MS/MS workflows produces mass spectrometry data in the form of .RAW spectra files. In the thesis project, and others like it,

one .RAW file is produced for each replicate or sample in the experiment. For large projects involving many samples, the size and complexity of MS-generated data becomes very large: for example, the total size of the .RAW files generated in the thesis project amounts to over 130 GB of data. Needless to say, gleaning functional and taxonomic information from these data would take an incredible amount of time if it was not for the utility of computer programs and algorithms to automate data cleaning, data sorting, statistics, and visualizations.

To this end, Cerberus, a software package, was developed to analyze data generated in the thesis project and can be extended to other metaproteomics projects if given proper databases. Cerberus is a Python 3.6 [153] and R >3.5.0 [154] software package that sits at the interface between iMetaLab database search outputs from .RAW files and downstream analyses. For large metaproteomics datasets that hold a large amount of sample data, Cerberus allows researchers to do initial database search processing using iMetaLab, take the iMetaLab output files, and then perform a range of statistical analyses and visualizations – along with linear programming analyses. Cerberus is so named because it integrates processes from three established software packages – iMetaLab, PathviewR, and MetaboAnalystR – and Cerberus is a mythical three-headed dog (and the name of a Figeys lab member's own dog as well).

The focuses of the thesis project are to describe resistant starch-mediated functional and taxonomic shifts experienced by human gut microbiomes, as well as addressing the shared-peptide problem that influences taxonomic and functional analyses of metaproteomics data. Parlaying these focuses into Cerberus lead to its development as an integrative workflow software package, taking out the legwork and tediousness of organizing mass

spectrometry data for statistics and visualization generation. Cerberus has been deposited on GitHub (https://github.com/jackrrryan/uolab).

4.1.2 Data import, processing, and annotation

Figure 4.1 outlines a general workflow utilizing Cerberus for metaproteomics data analysis.

The Figeys lab-developed software suite iMetaLab [149] performs a crucial first step as it produces outputs that serve as inputs for Cerberus. iMetaLab imports .RAW files, constructs sample-specific peptide spectra search databases, and outputs a set of result files including peptide, protein, and taxonomic abundance assessments.

Figure 4.2 outlines the data inputs and outputs of Cerberus's functional and taxonomic analyses. Cerberus inputs two files produced by iMetaLab: proteinGroups.txt and Metalab_taxonomy.xlsx. proteinGroups.txt organizes peptide-protein assignment information into table rows or "proteinGroups", with each proteinGroup containing one or more "proteinIDs" (UniProt [137] protein identifier codes), based on the peptides that were identified in the experiment. This file also contains all sample-specific proteinGroup intensities (proxy for abundance) for each sample in the experiment. Metalab_taxonomy.xlsx is an Excel spreadsheet file whose two sheets can be converted into .csv files, making them easier to read by software programs. Both sheets contain taxonomic abundance assignments for every taxonomic level that was found in any sample (based on mappings done by MetaLab between peptide sequences and specific taxa), with one sheet containing taxa with at least one peptide assignment and the other containing taxa with at least three peptide assignments. Taken together, proteinGroups.txt and Metalab_taxonomy.xlsx represent a basis for deeper taxonomic and functional analyses using Cerberus.

The first step for Cerberus is pre-processing of proteinGroups and taxonomic

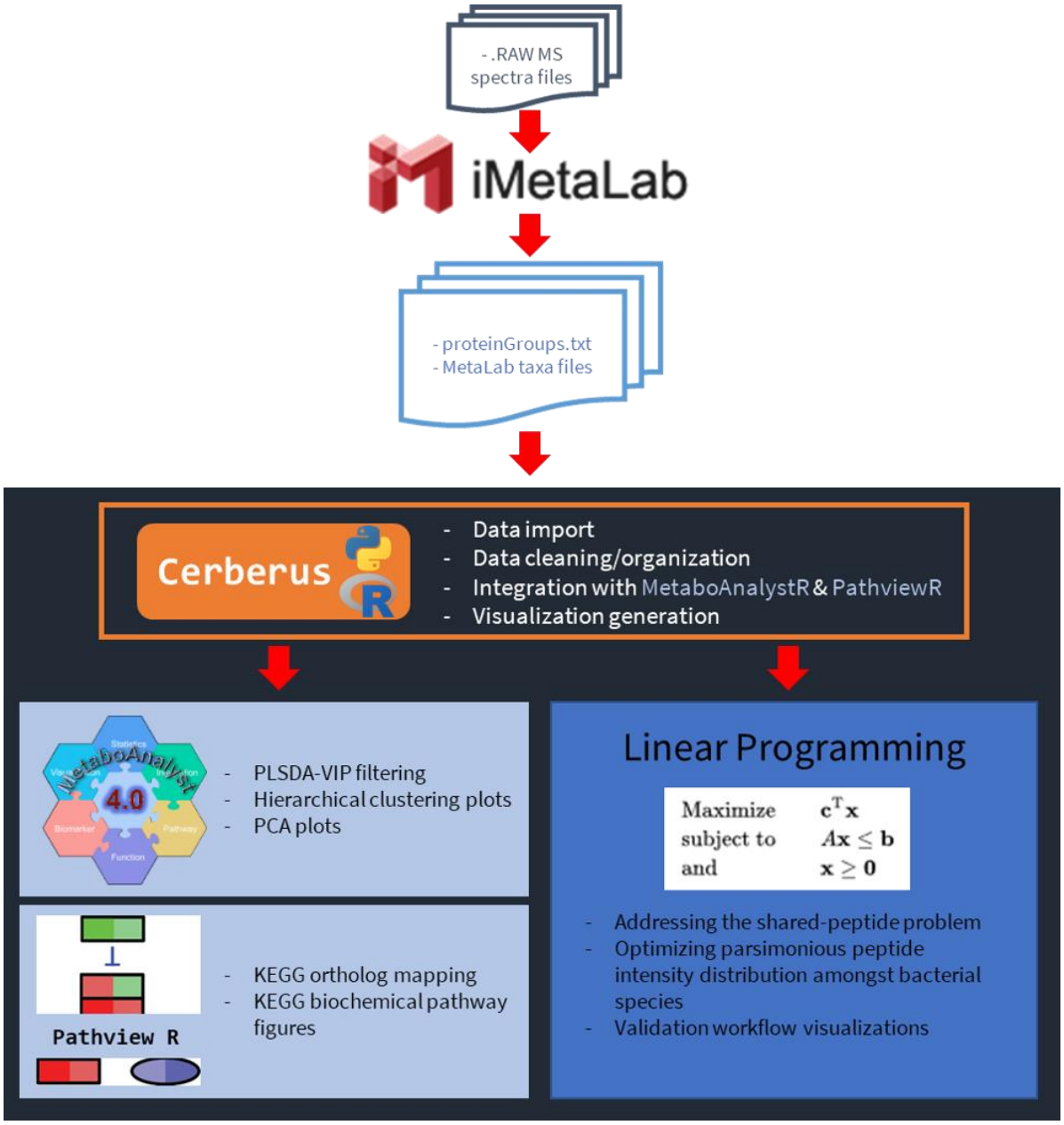

- .RAW MS spectra files
iMetaLab
- proteinGroups.txt
- MetaLab taxa files
Cerberus
- Data import
- Data cleaning/organization
- Integration with MetaboAnalystR & PathviewR
- Visualization generation
MetaboAnalyst
4.0
- PLSDA-VIP filtering
- Hierarchical clustering plots
- PCA plots
Pathview R
- KEGG ortholog mapping
- KEGG biochemical pathway figures
Linear Programming
Maximize $\mathbf{c}^T\mathbf{x}$
subject to $A\mathbf{x} \leq \mathbf{b}$
and $\mathbf{x} \geq \mathbf{0}$
- Addressing the shared-peptide problem
- Optimizing parsimonious peptide intensity distribution amongst bacterial species
- Validation workflow visualizations

Figure 4.1. Overview of Cerberus, and of the bioinformatics workflow that Cerberus inhabits.

abundance data into internal data structures. Comparisons between treatment conditions make up the bulk of Cerberus's functions, and data pre-processing involves organizing proteinGroups.txt and Metalab_taxonomy.xlsx into pandas [155] data structures holding all groups' intensity and taxonomy data for experimental samples present in these files.

The KEGG ortholog functional annotation database [156] is used in Cerberus for functional annotations. Its data is organized into a hierarchical directed graph with nodes at four levels (defined as L1, L2, L3, and L4 in this thesis). This structure provides a hierarchical and logical organization of functional divisions: L4 nodes (or leaves) represent specific enzymes and their orthologs such as glucose-6-phosphate isomerase (KEGG L4 ID: K01810); L3 nodes (internal) represent biochemical pathways such as Glycolysis/Gluconeogenesis (KEGG L3 ID: ko00010); L2 nodes represent groupings of biochemical pathways, such as Carbohydrate Metabolism (KEGG L2 ID: 09101); and L1 nodes represent general functional groupings, such as Metabolism (KEGG L1 ID: 09100; the parent node of Carbohydrate Metabolism). L1, L2, and L3 can be considered a traditional tree with descending nodes representing increasing specificity of functional groupings. L4 nodes can have more than one L3 parent node (inbound edges), as an enzyme may be involved in multiple biochemical pathways. A webpage-based KEGG ortholog ID hierarchy viewer is provided at https://www.genome.jp/kegg-bin/get_htext#A1.

Taxonomic annotations are done through MetaLab-sourced proteinID-taxa mapping database files (pro2taxa_homo.csv or pro2taxa_mouse.csv, depending on the experiment's source organism). Functional annotations are done through KEGG GhostKOALA-generated database files containing proteinID-functional mappings [157] (KEGG-proteinID-human.csv or KEGG-proteinID-mouse.csv, depending on the experiment's source organism). GhostKOALA maps UniProt proteinIDs to L4 KEGG IDs, completing the final annotation

link to be able to functionally describe proteinIDs found through peptide database searches in a metaproteomics experiment.

Each proteinGroup has a collection of proteinIDs, which themselves share a set of peptide assignments. To annotate a proteinGroup with a KEGG functional mapping, each proteinID's KEGG L4 annotations are retrieved, and the set union of all proteinID's KEGG L4s is calculated. This KEGG L4 set is then assigned to the proteinGroup and represents its functional mapping. For taxonomic annotation, a proteinGroup's proteinIDs are assessed from most-confident to least confident, and if a proteinID has a taxonomic annotation, it is assigned to the whole proteinGroup.

Once annotation is complete, Cerberus builds a Q-filtered association dataset, where each treatment group's proteinGroups are compared with every other treatment group to determine which proteinGroups are common between and exclusive to treatment groups. Consider an experiment with five treatment groups (e.g. control and 12-, 24-, 36-, and 48-hour incubations), and each treatment group contains five technical replicates (in total, 25 samples between all treatment groups). Cerberus finds the union set and exclusive sets of proteinGroups for each combination of treatment groups (in this example, ${}_5C_2 = 10$ possible treatment group combinations).

Then, Q-filtering is performed on these associated proteinGroups. Since metaproteomic datasets commonly consist of a range of samples originating from multiple individual organisms and multiple treatment conditions, there exists a hierarchical sample organization that becomes useful in a filtering procedure. As such, Cerberus takes hierarchical sample organizations into account. There are two modes for Q-filtering: one for cases where treatment groups have more than four replicates and one where treatment groups have less than four replicates.

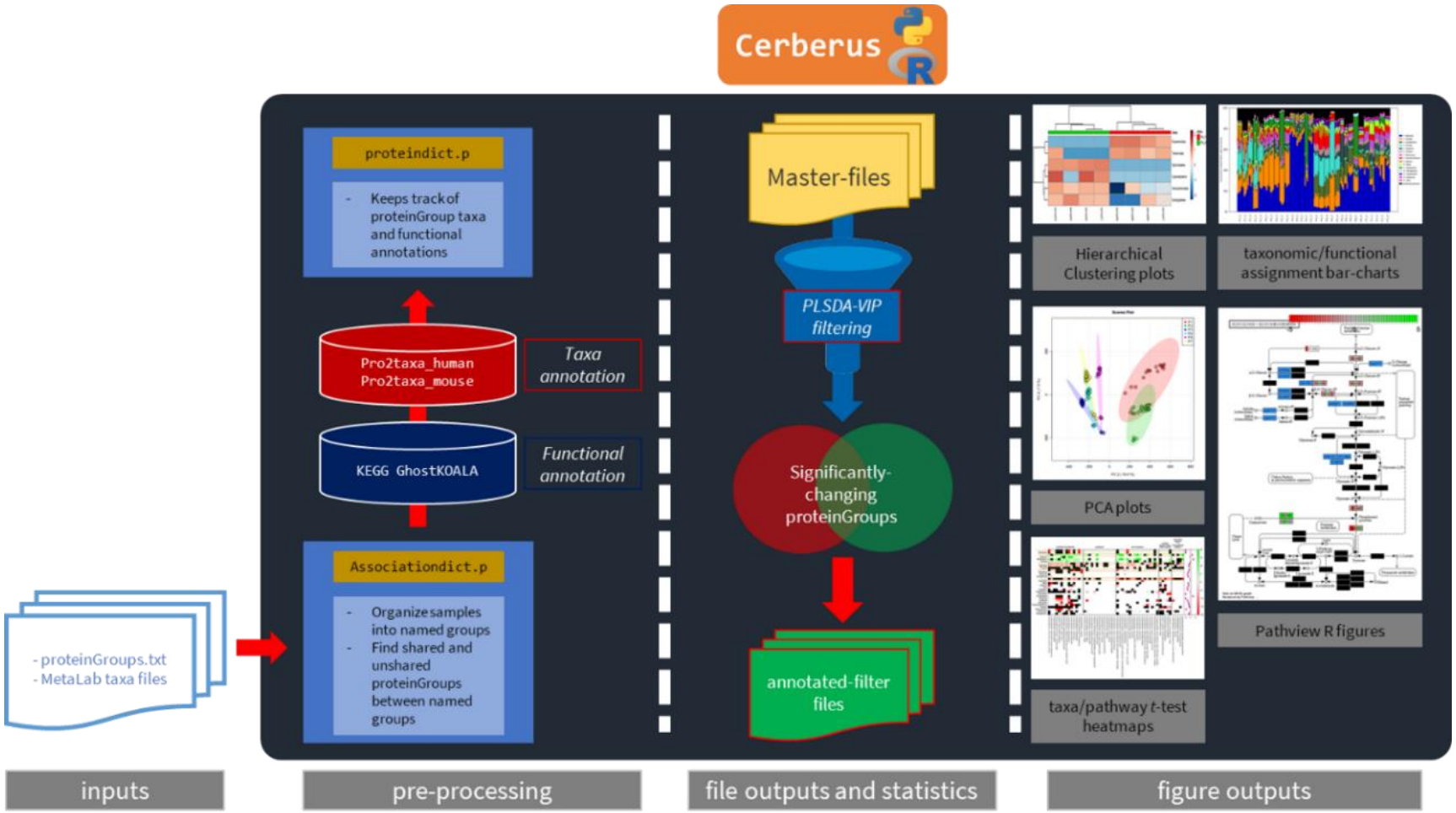

Cerberus
proteindict.p
- Keeps track of proteinGroup taxa and functional annotations
Pro2taxa_human
Pro2taxa_mouse
Taxa annotation
KEGG GhostKOALA
Functional annotation
Associationdict.p
- Organize samples into named groups
- Find shared and unshared proteinGroups between named groups
- proteinGroups.txt
- MetaLab taxa files
Master-files
PLSDA-VIP filtering
Significantly-changing proteinGroups
annotated-filter files
Hierarchical Clustering plots
taxonomic/functional assignment bar-charts
PCA plots
Pathview R figures
taxa/pathway t-test heatmaps
inputs
pre-processing
file outputs and statistics
figure outputs

Figure 4.2. A closer look at Cerberus's functional and taxonomic analysis workflow.

In the former case, the process is as follows:

For a given treatment group comparison between groups *A* and *B*, proteinGroups belonging to the set $A \cup B$ with a non-zero intensity in at least 50% of technical replicates are kept (Q50 filtering), and for the *A*- and *B*-exclusive sets, only proteinGroups that have a non-zero intensity in at least 80% of technical replicates are kept (Q80 filtering).

For the latter case, the process takes place on individual treatment groups, and is as follows:

Let *T* be the set of all treatment groups from all organisms. For a given proteinGroup *g* in the set of all proteinGroups *G*, find the subset T_g of *T* composed of treatment groups that have a non-zero intensity for *g* in >50% of their technical replicates. Next, let T_n be treatment groups in *T* that are *not* in T_g but have at least one technical replicate with a non-zero intensity for *G*. For each treatment group *r* in T_n, check how many of *r*'s organism's other treatment groups have at least one replicate with a non-zero intensity for *G*. If this count is >50%, then add *r* to the final set T_P (T_P also contains T_g). Next, go through each treatment group association; for a given treatment group comparison between groups *A* and *B*, find all proteinGroups that are in *A* and *B* and that are in T_P.

Two experiment-specific compressed database files are the result of this pre-processing step: *associationdict.p*, which holds all proteinGroup associations between treatment groups, and *proteindict.p*, which holds taxonomic and functional annotation information for all proteinGroups. Further, a set of files known as master-files hold this filtering and association information and are saved in .csv format for easy viewing by the user.

Cerberus then performs data transformation and imputation using a few different protocols. Users can select no data transformation or imputation, or a combination of one transformation technique and one imputation technique. For transformation, data can be scaled by sum or mean. These transform methods are as follows:

- Sample vector $S = \{s_1, \ldots, s_k, \ldots s_n\}$
- find sum(k) for each sample k in S, assign to vector S_{sum}
- find max(S_{sum})
- calculate scale factor vector $S_c = \{\frac{sum(s1)}{\max(S\text{sum})}, \ldots \frac{sum(sk)}{\max(S\text{sum})}, \ldots, \frac{sum(sn)}{\max(S\text{sum})}\}$
- multiply values in each sample in the proteinGroups table by each sample's scaled factor in S_c
- replace 'sum' with 'mean' for mean transformation.

Imputation can take place through several methods:

- apply the minimum non-zero value to all zero values in the proteinGroup table.
- apply the minimum non-zero value multiplied by a random number in the range [0.99,1.01] to all zero values in the proteinGroups table.
- Do a sequential KNN (*k*-nearest neighbours) imputation to the proteinGroups table.

This transformed and imputed proteinGroups table is used for further analyses in Cerberus.

The next step for Cerberus is a proteinGroup filtering procedure to find and retain statistically-significantly-changing proteinGroups between treatment conditions. Another set of files are produced for each treatment group combination, where the shared set of proteins between two treatment groups undergo a further PLSDA (partial least squares discriminant analysis) -VIP (variable importance in projection) filtering step. Cerberus interfaces with MetaboAnalystR [150], an R package that provides statistical and visualization outputs, to perform this step. PLSDA is a feature selection procedure that finds a linear regression model and features of statistical importance for categorical data [158]. In these experiments,

features can be considered as proteinGroups. A VIP coefficient is calculated for each feature, with large VIP coefficients (greater than one) assigned to features that provide the most explanatory power for the predictive PLS model; in metaproteomics analyses, these features are proteinGroups that explain the most variability between treatment groups. VIP coefficients provide a valuable means of filtering group comparison data so that only the most "important" (VIP > 1) proteinGroups remain and limits a treatment group comparison's shared proteinGroup dataset to those proteinGroups that exhibit a great degree of variability between treatment groups in the context of those treatment groups' data.

Once PLSDA-VIP filtering is complete, the proteinGroup data can be considered split into two meta-groups: *unfiltered*, where PLSDA-VIP filtering does not take place, and *PLSDA-log2*, where proteinGroup intensity data is $\log_2$-transformed before running PLSDA-VIP filtering. Each meta-group contains a full output of taxa-/function-annotated treatment group association files, and Cerberus maintains internal data structures holding all this information as well. Along with these proteinGroup datasets, the MetaLab taxa file data also persist in Cerberus's data structures, and parallel PLSDA-VIP-filtered association datasets are also produced for use in figure generation.

4.1.3 Figure generation

Cerberus uses these datasets to generate figures. On runtime, the user supplies a text file containing specifications for figure generation. As every experiment is different, it is impossible for Cerberus to infer what treatment groups are controls (or samples) from experiment sample names, and what treatment group comparisons the user wants to include in various figures. To address this issue, the options file allows the user to define exactly what treatment groups and comparisons to analyze, removing this responsibility from Cerberus. The options file contains a list of headers corresponding to different figures, sub-

headers corresponding to different “runs”, and line-separated strings corresponding to individual group comparisons. A correctly formatted options file gives Cerberus necessary information to be able to generate tailored figures from the following five categories.

4.1.3.1 Taxa PLSDA-VIP Hierarchical clustering plots

Taxonomic data from the 1-peptide and 3-peptide MetaLab taxa files are sample-scaled at the phylum, genus, and species levels. Scaling is done individually on samples: a sample’s sum of all present taxa at phylum, genus, and species levels are calculated, all values at that level are divided by this sum, and this value is then multiplied by the number of different taxa for that level. This data table is then broken into treatment group comparisons (as described before), undergoes PLSDA-VIP filtering (taxa with a VIP coefficient < 1 are filtered out), and is submitted to MetaboAnalystR through an R script. MetaboAnalystR outputs a hierarchical clustering plot displaying treatment sample and taxonomic clusters using the Ward clustering algorithm [159] and a Euclidean distance measure.

4.1.3.2 Taxonomic and functional (KEGG) assignment bar-charts

Once annotation is complete, a bar-chart is generated for each of phyla-, genus-, and species-level taxonomic assignments. The Python Matplotlib software package [160] is used to plot these data, and the unfiltered MetaLab taxa file data is used for these figures. After proteinGroup annotation has taken place, KEGG L2 and L3 functional annotations for all proteinGroups are displayed as bar-charts. For each sample in the experiment, Cerberus takes all proteinGroup data and sums the intensities of those proteinGroups that are annotated with a given L3 or L2 KEGG (depending on the bar-chart being generated). This creates a summary data structure holding the summed intensity of each L3 and L2 KEGG, then used to plot the bar-charts.

4.1.3.3 Principal components analysis (PCA) plots

MetaboAnalystR includes a function that takes in a .csv file of features and outputs a PCA plot given those features' treatment groupings. Users can specify "meta-groups", combining multiple treatment groups into one group containing all pertaining samples. Cerberus $\log_2$-transforms all proteinGroup intensity data and runs this function, producing PCA plots as .png image files.

4.1.3.4 Taxa/Pathway t-test heatmaps

These heatmaps display an overview of relationships existing between taxonomically- and functionally-annotated proteinGroup data. Figure 4 outlines the process of assigning data and creating these figures.

Taxa annotations at the phylum- and genus-levels are considered. As described in Figure 4, the filtering and calculation process is done for every taxa/L3 KEGG pathway combination, but Cerberus selects subsets of KEGG L3 pathways to display on the heatmap figures together. This is done for clarity and for focusing the figures on specific functions. The subsets of KEGG L3 pathways are called "pathway groups", and their member KEGG L3 pathways share a similar over-arching function, such as "Amino Acid Metabolism" or "Lipid Metabolism".

The *t*-statistics and p-values for all taxa/pathway comparisons are saved to Cerberus's data structure, and p-value ≤ 0.05 values are plotted as coloured squares on a heatmap. There are two modes for this plotting function: *multi-comparison* displays a set of results for a list of user-defined treatment group comparisons, useful for comparing multiple samples versus some control; and *pathwaygroups* generates one figure for a single treatment group comparison and contains all "pathway groups" defined in Cerberus, useful for providing a broad biochemical pathway and taxonomic summary of a single treatment group's response

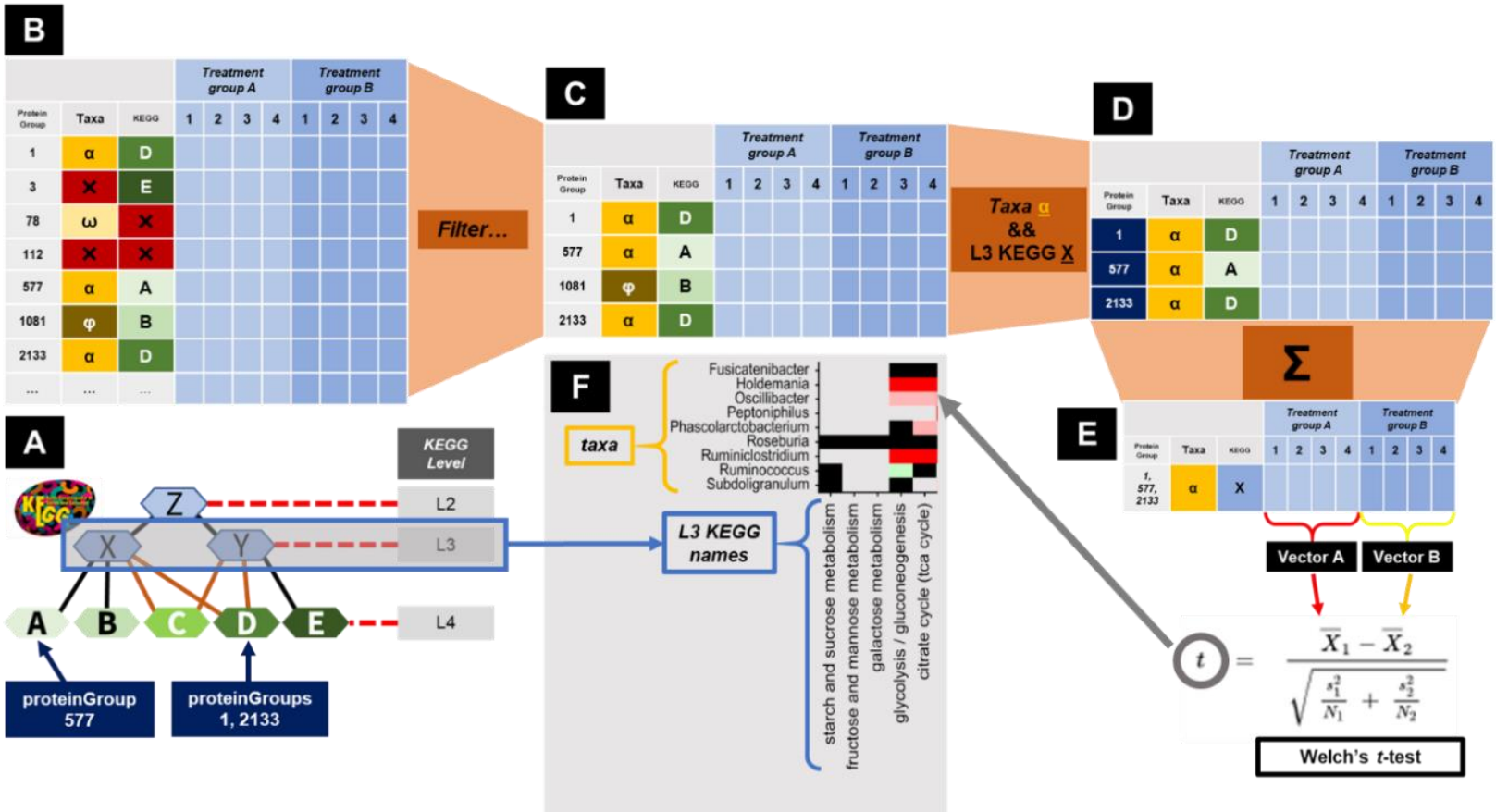

B
Treatment group A
Treatment group B
Protein Group
Taxa
KEGG
1 2 3 4 1 2 3 4
1 α D
3 X E
78 ω X
112 X X
577 α A
1081 φ B
2133 α D
... ...
Filter...
C
Taxa α
&&
L3 KEGG X
D
Σ
E
1, 577, 2133 α X
Vector A
Vector B
Welch's t-test
A
KEGG Level
L2
L3
L4
Z
X
Y
A B C D E
proteinGroup 577
proteinGroups 1, 2133
F
taxa
Fusicatenibacter
Holdemania
Oscillibacter
Peptoniphilus
Phascolarctobacterium
Roseburia
Ruminiclostridium
Ruminococcus
Subdoligranulum
L3 KEGG names
starch and sucrose metabolism
fructose and mannose metabolism
galactose metabolism
glycolysis / gluconeogenesis
citrate cycle (tca cycle)

Figure 4.3. Overview of t-test heatmap figure generation.
This shows the process of isolating vectors describing some treatment groups A and B's proteinGroups that have been annotated with taxa α and L3 KEGG node X. Panel A shows a condensed example of the KEGG ortholog database's quasi-tree structure. KEGG levels are shown on the right, along with an example of how L4 KEGG nodes can be connected through inbound edges to L3 KEGG nodes (see L4 KEGG nodes C and D). Panel B shows a toy data table with taxa and L4 KEGG annotations. Panel C shows the same data table after it has had proteinGroups (rows) removed that do not have both a taxon and KEGG annotation. Panel D shows the same table after it has had all proteinGroups removed that do not match taxa α. Panel E shows the process of summing the three remaining protein groups together: they all have been annotated with L4 KEGGs (nodes A and D) whose parent node is L3 KEGG node X, as well as all being annotated with taxa α. A Welch's t-test is performed on group A versus group B's vector values using the SciPy ttest_ind function [161], and its t-statistic is displayed as a coloured box corresponding to taxa α and L3 KEGG node X in the heatmap figure. Panel F shows a subsection of an example heatmap figure. This process is done for every taxa/L3 KEGG pathway combination to produce a matrix heatmap.

versus a single control. For each mode, two figures are generated, one where p-value > 0.05 values are displayed as black squares, and one where p-value > 0.05 values are displayed as white squares. This is done so that a user can determine which taxa-pathway comparisons *did* have proteinGroups that matched to them but *did not* have a significant *t*-test result.

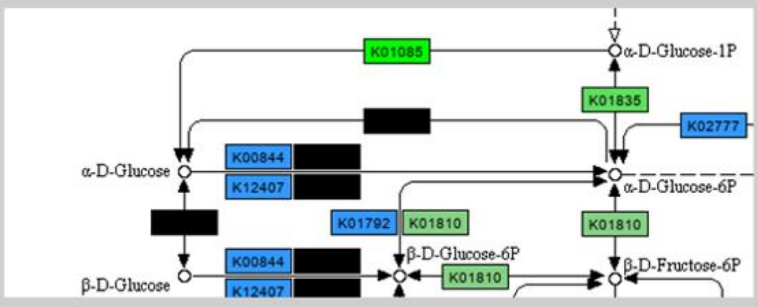
4.1.3.5 Pathview figures

Pathview [152] is an R software package that integrates KEGG functional mappings into KEGG L3 biochemical pathway visualizations, allowing for intuitive overviews of KEGG functional expression levels between treatment groups. Cerberus interfaces with Pathview through an R script, handing it prepared data files for treatment group comparisons. Pathview imports a pathway-specific .png file and .xml file to run its figure generation. The .png file acts as a template for Pathview to colour, and the .xml file holds L4 KEGG mappings and colour locations. The .png and .xml files originates from the KEGG database online portal and Pathview either downloads it at runtime through a file transfer protocol (FTP) request or imported directly to Pathview from a local directory. Cerberus holds all pathway .png and .xml files in a local directory, eliminating FTP running time that would otherwise slow the automated Pathview figure generation to a crawl.

The data files Cerberus generates for Pathview are composed in a similar fashion to the taxa-pathway *t*-test data, except taxa annotations do not factor into the filtering procedure. This results in a table where rows are L4 KEGG IDs, columns are the two treatment groups for a given comparison, and bins are the *mean* of the *sum* of proteinGroup intensities that are annotated for that row's L4 KEGG ID.

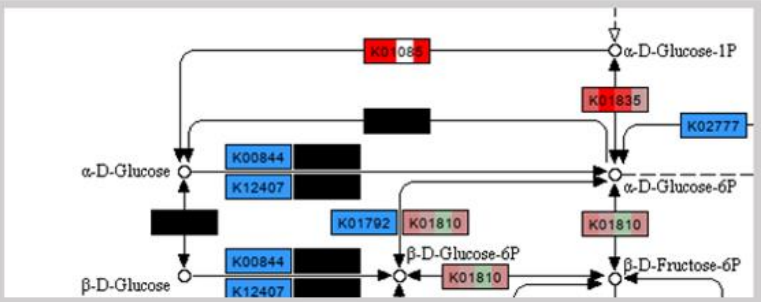
Figure 4.4 shows an example of the differences between single-comparison and multi-comparison Pathview figures. Rectangular boxes represent enzymes mapped to the L4 KEGG IDs shown inside them. Cerberus uses a database file with a list of L4 KEGG IDs

A



	12h	control	Log_2-FC
K01085	500000	15000	5.0589
K01035	1900000	65000	4.8694
K01810	300000	55000	2.4475

B



	24h	control	Log_2-FC	36h	control	Log_2-FC	48h	control	Log_2-FC	72h	control	Log_2-FC
K01085	400	15000	-5.2288	200	15000	-6.2288	0	15000	NA	200	15000	-5.2288
K01035	8000	65000	-3.0224	2000	65000	-5.0224	2500	65000	-4.7004	29000	65000	-1.1644
K01810	15000	55000	-1.8745	25000	55000	-1.1375	180000	55000	1.7105	15000	55000	-1.8745

Figure 4.4. Examples of Cerberus's PathviewR integration for KEGG pathway maps. Panel A displays a dummy data table and its Pathview figure for a single treatment group comparison (e.g. 12-hour incubation vs. control). Panel B shows the same pathway subset as Panel A with an example output for a multi-comparison Pathview figure using the dummy data in the lower table.

deemed diagnostic for each pathway's activity, and these specific L4 KEGG IDs can be considered a subset of all L4 KEGG IDs displayed on an empty Pathview figure. Green/grey/red boxes denote the colourbar-mapped value of a fold-change calculation between the two treatment groups in the comparison, blue boxes represent enzymes that map to KEGG L4 IDs without any proteinGroup annotations in that treatment group comparison, and black boxes represent enzymes that either are not present in the KEGG ortholog database or are not diagnostic for activity in the displayed pathway. In multi-comparison Pathview figures, each KEGG rectangle box is "sliced" into even-width segments corresponding to the value for a column in the input data file. This allows for the fold-change values from multiple comparisons (e.g. control vs. 12-hour incubation, control vs. 24-hour incubation, control vs. 36-hour incubation, etc.) to be displayed side-by-side in the same KEGG rectangle box, making it easier for viewers to discern variable expression patterns between treatment group comparisons.

Cerberus uses the *unfiltered* and *log2-PLSDA-VIP* datasets as the data source for these Pathview figures.

4.1.4 Linear programming

4.1.4.1 Formulation

Cerberus includes a collection of functions to execute a linear programming (LP) approach, addressing the shared-peptide problem present in metaproteomics mass spectrometry experiments. Inspired by the work of He *et al*. [162], its goal is assigning a *parsimonious* shared peptide intensity distribution amongst constituent bacterial species with respect to all other peptide measurements obtained in the experiment.

The LP algorithm's objective function is as follows:

$$\min_{D} \sum t_k \qquad \textbf{4.1}$$

Where *D* is the matrix holding all intensity values for all *n* peptides vs. all possible bacterial species given shared peptides' phylogenetic resolution, and t_k is the **maximum** peptide intensity value for the *k*th bacterial species' vector d_k. The goal of the objective function is to minimize the sum of the maximum values for all individual bacterial species and optimizes for the fact that bacteria with the greatest intensities for unique peptides will gather most of the shared peptide abundances.

Constraints for this LP algorithm are as follows:

$$\forall j: b_j - \sum_{k|(y_j,z_k)\in E} = 0 \qquad \textbf{4.2}$$

Where y_j is the *j*th peptide, z_k is the *k*th bacterial species, and *E* is the set of edges from a graph $G=\{V,E\}$ with *V* representing vertexes assigned to all peptides and all species. In this way, *D* can be considered the adjacency matrix of a bipartite graph *G* which connects peptide "vertices" (a subset of *V*) to bacterial species vertices (another subset of *V*) through edges *E*. This constraint forces the sum of intensity values for a given peptide's species to equal the total intensity value of the peptide measured by mass spectrometry in the experiment.

Further,

$$\forall j,k : d_{jk} \leq t_k \qquad \textbf{4.3}$$

constrains any matrix bin d_{jk}'s intensity value to be less than its species vector d_k's maximum intensity value (t_k).

Finally,

$$\forall j,k:\ d_{jk} \begin{cases} = 0 & if\ (y_j, z_k) \notin E \\ \geq 0 & else \end{cases} \qquad \textbf{4.4}$$

asserts that matrix bins representing unmatched species and peptides must be 0, and greater than or equal to 0 otherwise (matched peptides and species). In other words, peptide sequences that are not known to be expressed by a given bacterial species, will have a 0 entry in *D* for that bacterial species.

4.1.4.2 Rationale

This LP formulation carries the rationale that bacterial species with large unique peptide intensities are likely to receive large distributions of shared peptide intensities. In other words, "the rich get richer" – the more peptides a bacterial species is associated with, the more it is likely to have shared peptides associated with it versus other species associated with few peptides.

For instance, consider the following toy example matrix:

$$D_{j,k} = \begin{array}{c} \\ j_0 \\ j_1 \\ j_2 \\ j_3 \end{array} \begin{array}{c} \begin{array}{ccccc} k_0 & k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} ? & - & - & - & - \\ ? & - & ? & - & ? \\ - & - & - & ? & - \\ - & ? & ? & ? & - \end{bmatrix} \end{array} \qquad \textbf{4.5}$$

In **4.5**, four peptides (rows *j*) and five bacterial species (columns *k*) are organized, where ones represent possible peptide assignments and zeros represent where there can be no match between a peptide and a species. Consider species column k_0: peptides j_0 *and* j_1 can possibly assign a share of their individual intensities to this species. In this toy example, let peptide j_0's intensity be 100, and peptide j_1's intensity be 110. Note that peptide j_0 is a unique peptide – it has a single specific species assignment – so the linear objective will place all its intensity value (100) into the $D_{0,0}$ matrix bin:

$$D_{j,k} = \begin{array}{c} \\ j_0 \\ j_1 \\ j_2 \\ j_3 \\ \textbf{max} \end{array} \begin{array}{c} \begin{array}{ccccc} k_0 & k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} \mathbf{100} & - & - & - & - \\ ? & - & ? & - & ? \\ - & - & - & ? & - \\ - & ? & ? & ? & - \end{bmatrix} \\ \begin{array}{ccccc} \mathbf{100} & ? & ? & ? & ? \end{array} \end{array} \quad \textbf{4.6}$$

If we were only to consider this information, then t_0 (species k_0's maximum peptide intensity value) will simply be the intensity value of peptide j_0. Because the objective function is attempting to minimize the sum of each column's maximum value, the linear formulation incentivizes giving the majority of peptide j_1's intensity (which is a shared peptide between species k_0, k_2, and k_4) to bin $D_{1,0}$ in such a way that bin $D_{1,0}$'s value does not exceed bin $D_{0,0}$'s value. A possible state of this matrix given this incentivization could be this:

$$D_{j,k} = \begin{array}{c} \\ j_0 \\ j_1 \\ j_2 \\ j_3 \\ \textbf{max} \end{array} \begin{array}{c} \begin{array}{ccccc} k_0 & k_1 & k_2 & k_3 & k_4 \end{array} \\ \begin{bmatrix} \mathbf{100} & - & - & - & - \\ \mathbf{90} & - & \mathbf{10} & - & \mathbf{10} \\ - & - & - & ? & - \\ - & ? & ? & ? & - \end{bmatrix} \\ \begin{array}{ccccc} \mathbf{100} & ? & \mathbf{10} & ? & \mathbf{10} \end{array} \end{array} \quad \textbf{4.7}$$

$D_{1,0}$ gets 90/110 out of peptide j_1's intensity, and the maximum value of column k_0 remains 100. Ultimately, this is what the linear solver wants to achieve – a minimization of each column's maximum value. Consider extrapolating these concepts to the rest of the matrix, and this majority-share pattern makes sense for other peptides in this situation: for example, a similar association exists between peptides j_2 (unique) and j_3 (shared). In this sense, bacterial species with evidence of their existence (unique peptides) will garner a larger share of shared peptides' abundance.

4.1.4.3 Validation

The LP model uses peptide/taxonomy information supplied through iMetaLab and MaxQuant to parsimoniously distribute peptide abundances amongst bacterial species that a given sample may possess. However, there is no "gold standard" for this model in a metaproteomics context since we do not actually know the proportion of each bacterial species in the analyzed samples and the amount each contribute to the measurement of a peptide. The lack of a gold standard for this LP model presents a pitfall for its analysis and usefulness. To address this, a focus of this facet of the thesis project has been to develop validation procedures to determine the LP procedure's performance against simulated conditions.

The LP validation procedure starts with a representative peptide intensity/taxonomic identification dataset. These data are organized into a table like that shown in equation **4.5**, called D_V. Along with this table, a vector I contains all the peptide intensity values measured during the representative experiment, a vector P contains the list of peptides in D_V (rows in the table), and a vector S contains "species abundance" values, which can be interpreted as the sum of all peptide intensities assigned to that species.

Next, a set of rules is applied to the dataset so that the validation procedure mimics biological realities of running metaproteomics tandem-MS/MS experiments. There are two problems to consider here:

Subset-species problem: Consider all of the possible mappings between a sample's experimentally detected peptides and their possible assignments to protein sequences extrapolated from a bacterial genomic database. Any microbiome sample that undergoes mass spectrometry will only contain a subset of these mappings – many of mapped

species do not actually exist in the sample, yet non-existent species are still considered because sub-sequences of their proteins map to experimentally-determined peptides. This is a by-product of spectral database search performance and the species/protein sequence coverage of sequence databases, and the protein-centric nature of metaproteomics experiments. Even in the case of very well optimized peptide/species matrices, there will be species that could receive a share of a peptide's abundance after LP optimization, even though that species does not exist in the sample.

Peptide-error problem: a tandem-MS/MS instrument has some element of uncertainty in its measurements. This results in experimental peptide intensity values deviating from their actual abundance in the sample by some factor. Accounting for this deviation is critical when using these raw peptide intensity values as the basis for their expression distributions across their putatively constituent species.

The LP validation procedure must account for these two problems.

The validation procedure supplies some information to D_V, P, and S before running, as described here:

To address the *subset-species problem*, randomly sampled peptide intensity values from vector I are assigned to vector S, giving putative species a proxy abundance value – however, only a subset of species in S are given randomly sampled values. Determining which species receive a randomly-sampled value is done by calculating the amount r of "definite" species from S (species that have at least one unique peptide assignment), then randomly choosing r species from S (called S_r). For example, in equation **4.5**, $r = 2$ (peptides j_0 and j_2). All species which belonging to the set $(S - S_r)$ receive a proxy abundance value of 0.

Then, based on the proxy abundance values assigned to *S*, pre-LP peptide intensities are calculated for *P*. Every peptide *j* has a set of species in *S* that it is assigned to (j_S). For any peptide *j* in *P*, the sum of proxy abundance values in *S* that *j* matches to is calculated. This gives a pre-LP peptide intensity to each peptide represented in D_V.

To address the *peptide-error problem*, each peptide's intensity value is perturbed by a factor ε sampled from a $\mathcal{N}(1, \sigma)$ distribution. The value of σ can change depending on the validation procedure's parameters. Each peptide intensity j_i in *P* is multiplied by an individually sampled ε to generate j_i's perturbed peptide intensity j_ε.

An LP iteration is run given the completed *P* vector with its perturbed peptide intensity values.

After the LP iteration is completed, a Euclidean distance is calculated between each peptide's newly-distributed intensity values and values in *S* that each peptide corresponds to:

$$\forall j \in P : j_d = \frac{\sqrt{\sum_{i=1}^{m} \left(D_{j,i}^{V} - S_i^j\right)^2}}{m} \qquad \textbf{4.8}$$

With $S^j = S_i^j, \ldots, S_m^j$ being the subset of species corresponding to some peptide *j*.

Steps 1-5 are then repeated using the same D_V for *x* iterations.

Measuring LP validation performance involves examining all j_d to see how well the LP procedure recapitulates the distribution of peptide intensity values based on the pre-determined bacterial species abundances. A j_d of 0 indicates a perfect match between expected and observed LP values, with increasing j_d indicating increasing magnitude of match error. The thesis project benchmarks the two naïve methods previously mentioned above: *naïve-split* and *naïve-duplicate*. The n*aïve-split* dataset is composed by evenly

splitting all individual peptide intensities from *P* amongst their constituent species in each peptide *j*'s S_j, and the *naïve-duplicate* dataset is composed by duplicating all individual peptide intensities in the same fashion. It is worth nothing that while the LP algorithm is described in the thesis project as distributing peptide intensity values across bacterial species, it can be used to distribute values across any OTU level in a community's taxonomy.

This algorithm has been implemented in Python 3.6 code as part of Cerberus and contains major contributions from the PuLP package [163].

4.2 GENERAL EXPERIMENTAL RESULTS

4.2.1 Mass spectrometry quantification

The thesis project relied on a metaproteomic approach to characterize the functional and taxonomic shifts that an *ex vivo* microbiome sample exhibits when exposed to different types of resistant starch. Samples from six participants were co-cultured as described in section 3.2. 116 .RAW files corresponding to technical replicates for treatment conditions were produced through tandem-MS/MS and analyzed using MetaLab. 5,119,332 MS/MS spectra were identified from these files, and a total of 80,297 peptides and 21,240 proteinGroups were identified using a false discovery rate (FDR) threshold of 1%. MetaLab-generated taxa files showed that 142 species with >= 3 peptide assignments (corresponding to 79 genera and 16 phyla) were identified throughout all 116 evaluated samples.

Figure 4.5 shows an overview of the thesis project's experimental results. Clustering and component/coordinate decomposition approaches are used to describe high-level relationships and differences between experimental samples. Cerberus (interfacing with MetaboAnalyst) was used to generate principal component analysis (PCA) plots, and The R package *vegan* [164] was used to calculate and plot a principal coordinate analysis (PCoA)

plot of Bray-Curtis dissimilarities as well as run an analysis of similarities (ANOSIM) of species abundances between all samples in the dataset.

The experimental data shows close clustering amongst samples originating from the same participant, as seen in Figure 4.5 panels C (PCA plot), D (Bray-Curtis PCoA plot), and E (hierarchical clustering plot). Panels C and D display clear clustering of each participant's treatment samples from proteinGroup intensity and Bray-Curtis dissimilarity scores, respectively, regardless of the treatment each sample was subjected to. These data were then run through a principal coordinate analysis (PCoA) in R. The ANOSIM *R* value was 0.9262

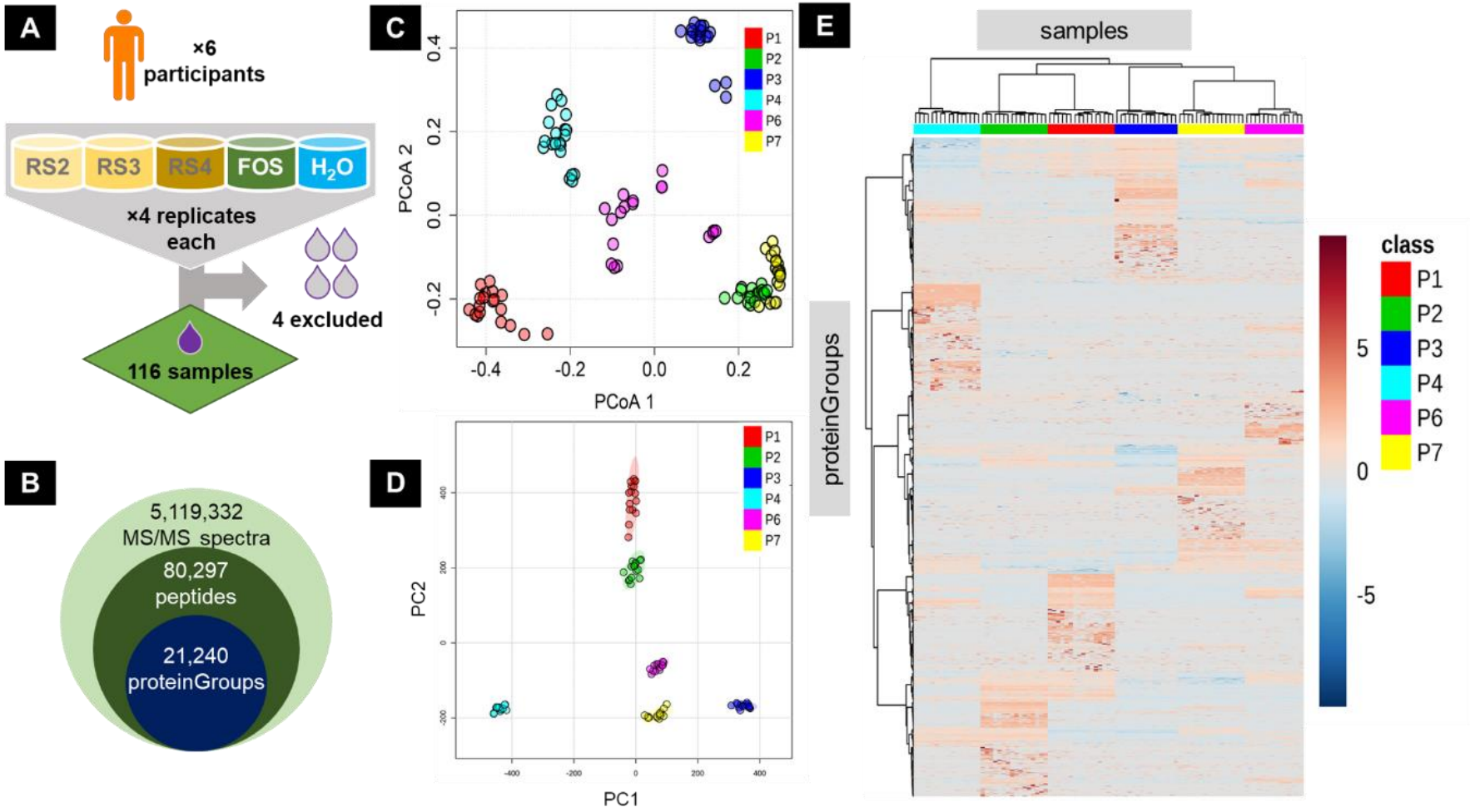

A
×6
participants
RS2
RS3
RS4
FOS
H2O
×4 replicates
each
4 excluded
116 samples
B
5,119,332
MS/MS spectra
80,297
peptides
21,240
proteinGroups
C
PCoA 2
PCoA 1
P1
P2
P3
P4
P6
P7
D
PC2
PC1
E
samples
proteinGroups
class
P1
P2
P3
P4
P6
P7
5
0
-5

Figure 4.5. Overview of the thesis project's experimental data.
Panel A describes the experimental sample division and organization. Panel B displays the quantities of MS/MS spectra, identified peptides, and identified proteinGroups in the experiment. Panel C shows a principal coordinate analysis (PCoA) plot of Bray-Curtis dissimilarities between samples, coloured by participant. Panel D shows a principal component analysis (PCA) plot of the first two components for the experimental dataset (FOS [positive control] samples are not shown as their components lie far outside of the rest of the samples' components, making the plot hard to read). Panel E shows a hierarchical clustering plot (Euclidean distance measure, Ward clustering algorithm) of all samples' proteinGroup intensities (FOS samples removed).

($p = 0.001$), suggesting a high degree of similarity between samples originating from the same individual (as opposed to similarity between samples originating from different individuals). Panel E reinforces this pattern, with visible "blocks" of proteinGroup intensities appearing in corresponding columns for each participant. When these analyses are done on treatment group divisions as opposed to participant group divisions, there does not appear to be any distinct clustering pattern in PCA or hierarchical clustering plots (see figure 4.6 below). Taken together, these results suggest the existence of expression "enterotypes" specific to individual participants, where differences between individuals' microbiome composition/expression account for far more variation than *ex vivo* co-culturing with RS types.

4.2.2 Resistant starch-mediated microbiome shifts

The hypotheses of the thesis project regarding RS-microbiome interactions are that (1) microbiome biochemical pathway expression will change between different bacterial operational taxonomic units (OTUs) and under different prebiotic conditions, and that (2) RS forms will affect microbiome bacterial taxonomic distribution. To address these, several statistical analyses and visualizations were performed using Cerberus.

Figure 4.6 displays PCA plots of log2-transformed proteinGroup intensity data. As previously mentioned, the figures as a whole show a clear clustering of samples amongst their participant label, demarcating "enterotypes" that, in this principal component decomposition, show that inter-sample variability is much more dependent on the participant that the sample originated from rather than the treatment the sample was exposed to. When the same PCA results are coloured according to the treatment condition that each sample was exposed to (top-right), the plot shows no discernable clustering. In the middle- and bottom-row plots, variability between samples corresponding to treatment conditions can be seen.

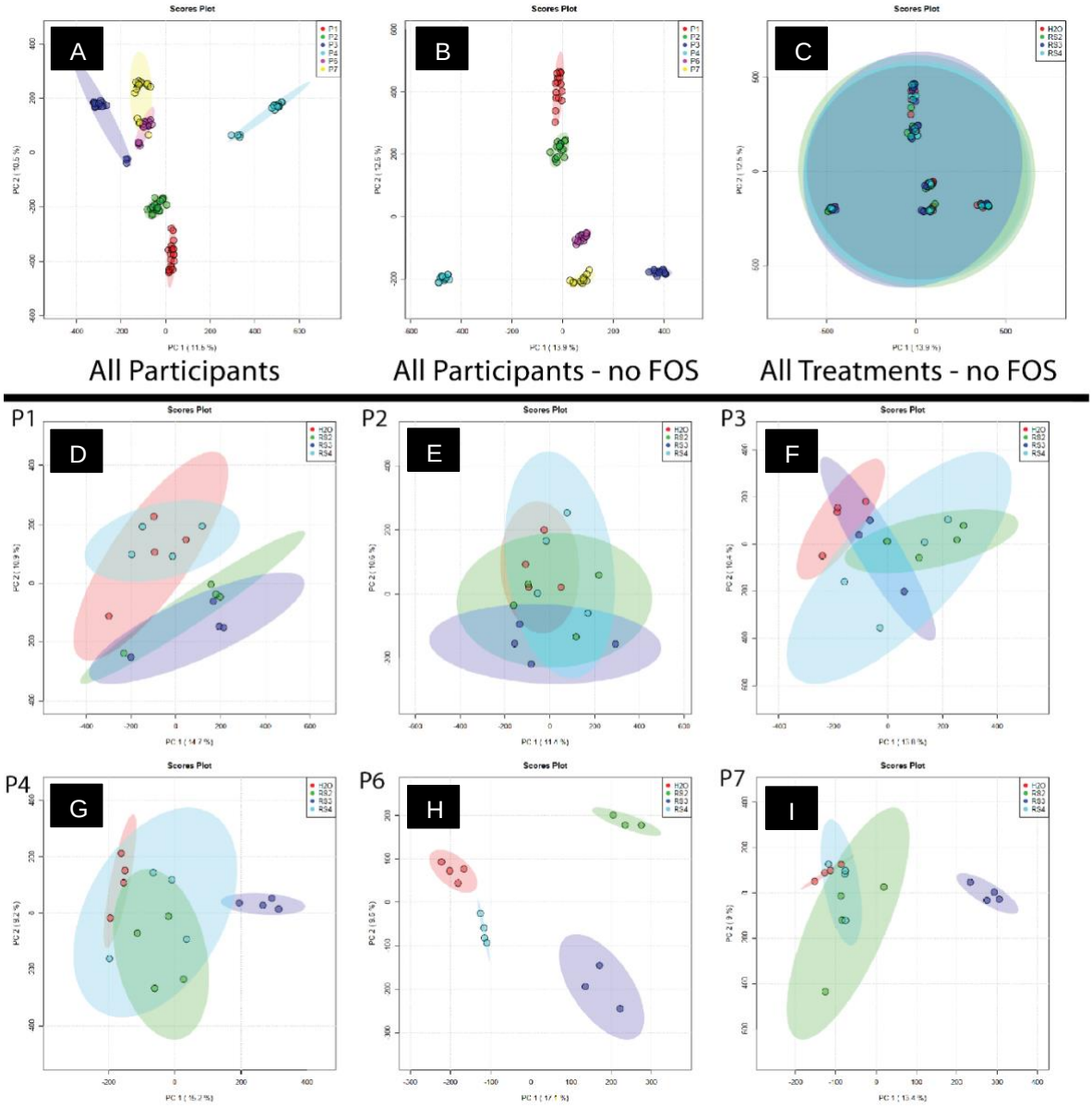

A
B
C
All Participants
All Participants - no FOS
All Treatments - no FOS
P1
D
P2
E
P3
F
P4
G
P6
H
P7
I

Figure 4.6. PCA plots of log2-transformed and Q-filtered proteinGroup intensity data. Each point represents a single sample from the overall experiment. In the top row, points are coloured by membership to participants (panels A and B) and treatment conditions (panel C). Plots in the middle and bottom rows (panels D-I) contain only samples from their labelled participant (top-left corner of each plot). Points are coloured by membership to treatment conditions. In all plots, coloured ellipses correspond to two-sigma confidence areas for all points belonging to a participant. FOS samples have been removed from each plot in this figure as their samples lie far away from the other treatment groups' samples in component space and obscure differences between the other treatment groups.

For participants 1,2,3, and 4, there appears to be little discernable difference between treatment groups, save for some defined separation between RS3's samples (dark blue) and control (H_2O; red). This suggests that for these participants, RS treatments did not lead to large differences in overall proteinGroup expression patterns. In participants 6's and 7's data, treatment groups resolve to defined clusters in principal component space, suggesting a participant-specific RS effect on proteinGroup expression in these two individuals.

While PCAs provide a broad overview of sample variation through a decomposition procedure and broadcasting to a component space, specific functional differences between treatment conditions in participant sample groups are not picked up by the PCA procedure. These differences include biochemical pathway enrichment variations at the treatment condition and participant levels and require further analytic and visualization techniques.

4.2.2.1 Functional shifts

Functional changes between treatment groups were assessed using Cerberus-generated visualizations. Figure 4.7 gives an overview of functional mappings for each sample in the total experiment. In this figure, FOS samples have a visible functional output difference when compared to other treatment groups from the same participant, regardless of which participant the samples originated from. The majority of FOS samples' functional variation relative to other treatment groups stems from enhanced expression of proteins annotated for pyruvate metabolism pathways (bottom panel of figure 4.7, dark blue bar sections at bottom). Since FOS samples were used as positive controls in the RapidAIM experiment and their functional profile differs significantly enough from negative control (H_2O), these results confirm it as a working positive control for microbiome shifts.

Functional information for the entire microbiome's protein expression data gives a broad view of differences between treatment groups. To establish more specific functional

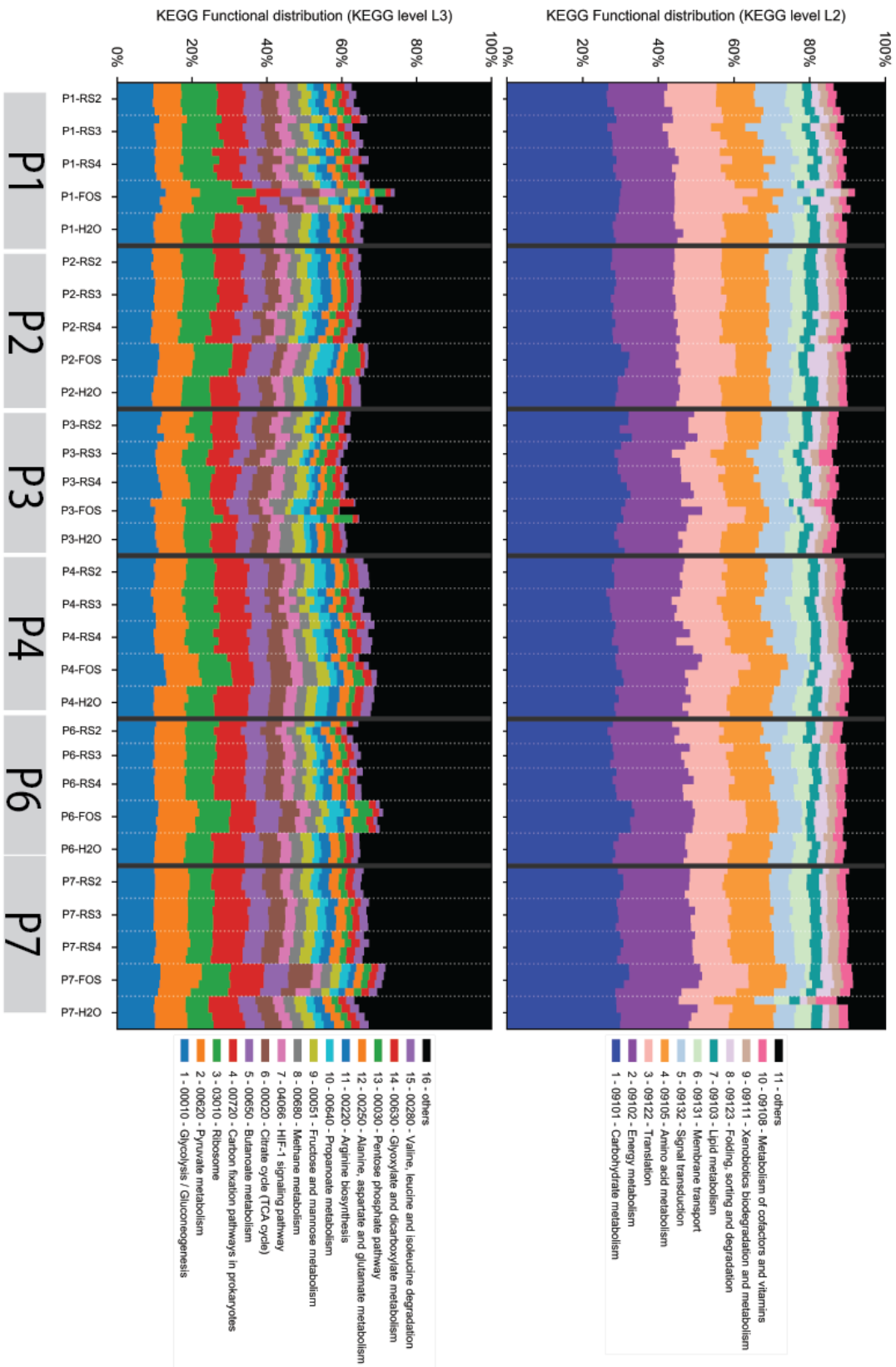

KEGG Functional distribution (KEGG level L2)
KEGG Functional distribution (KEGG level L3)
0%
20%
40%
60%
80%
100%
P1
P2
P3
P4
P6
P7
P1-RS2
P1-RS3
P1-RS4
P1-FOS
P1-H2O
P2-RS2
P2-RS3
P2-RS4
P2-FOS
P2-H2O
P3-RS2
P3-RS3
P3-RS4
P3-FOS
P3-H2O
P4-RS2
P4-RS3
P4-RS4
P4-FOS
P4-H2O
P6-RS2
P6-RS3
P6-RS4
P6-FOS
P6-H2O
P7-RS2
P7-RS3
P7-RS4
P7-FOS
P7-H2O
11 - others
10 - 09108 - Metabolism of cofactors and vitamins
9 - 09111 - Xenobiotics biodegradation and metabolism
8 - 09123 - Folding, sorting and degradation
7 - 09103 - Lipid metabolism
6 - 09131 - Membrane transport
5 - 09132 - Signal transduction
4 - 09105 - Amino acid metabolism
3 - 09122 - Translation
2 - 09102 - Energy metabolism
1 - 09101 - Carbohydrate metabolism
16 - others
15 - 00280 - Valine, leucine and isoleucine degradation
14 - 00630 - Glyoxylate and dicarboxylate metabolism
13 - 00030 - Pentose phosphate pathway
12 - 00250 - Alanine, aspartate and glutamate metabolism
11 - 00220 - Arginine biosynthesis
10 - 00640 - Propanoate metabolism
9 - 00051 - Fructose and mannose metabolism
8 - 00680 - Methane metabolism
7 - 04066 - HIF-1 signaling pathway
6 - 00020 - Citrate cycle (TCA cycle)
5 - 00650 - Butanoate metabolism
4 - 00720 - Carbon fixation pathways in prokaryotes
3 - 03010 - Ribosome
2 - 00620 - Pyruvate metabolism
1 - 00010 - Glycolysis / Gluconeogenesis

Figure 4.7. Overview of KEGG functional mappings for each participants' treatment groups' samples.
Top panel shows **KEGG L2** functional mappings, and bottom panel shows **KEGG L3** functional mappings. Functional abundance data was scaled across all samples for comparison purposes.

changes at the pathway and enzyme level, PathviewR figures were produced using Cerberus. These PathviewR figures were used to build the major portion (panel (c)) of figure 4.8, which displays functional enrichment differences for SCFA producing pathways, related reactions, and the butyrate pathway specifically. These data show a few significant trends related to SCFA production and biochemical pathways that feed into them. First, different RS have different taxa-function profiles related to butyrate production (panels (a)). In all RS, bacteria belonging to the *Firmicutes* phylum show the most variability regarding this set of KEGG L3 pathways, with proteinGroup activity appearing balanced relative to negative control (H_2O) as shown by the same-size green and red semicircles in the *Firmicutes* row. However, there does appear to be an RS-specific shift in taxa-function linked expression for these SCFA pathways, shown clearly in the *Bacteroidetes* phylum rows. RS2-related *Bacteroidetes* functional profiles parallel the balance shown in RS2-related *Firmicutes* function, but RS3 and RS4 results show that *Bacteroidetes* bacteria are much more up-regulated and down-regulated (respectively) for SCFA-related function relative to control.

The data presented in figure 4.8C consists of an aggregation of enzyme-level outputs from five PathviewR figures (Butanoate (Butyrate) Metabolism; Arginine and Proline Metabolism; Alanine, Aspartate, and Glutamate Metabolism; Glycolysis / Gluconeogenesis; and Pyruvate Metabolism), generated through Cerberus. All RS sources show an upregulation trend for major portions of pathways integral to butyrate production. This is apparent in the metabolism of early butyrate precursors, shown in reactions converting pyruvate through precursors acetyl-CoA, acetoacetyl-CoA, β-hydroxybutyryl-CoA, and crotonyl-CoA. These upregulated enzymes are an expected result as RS-encountering microbiome bacteria will have readily available 5- and 6-carbon sugars to use for this metabolic process (among others). Interestingly, the final direct conversion step of butyryl-

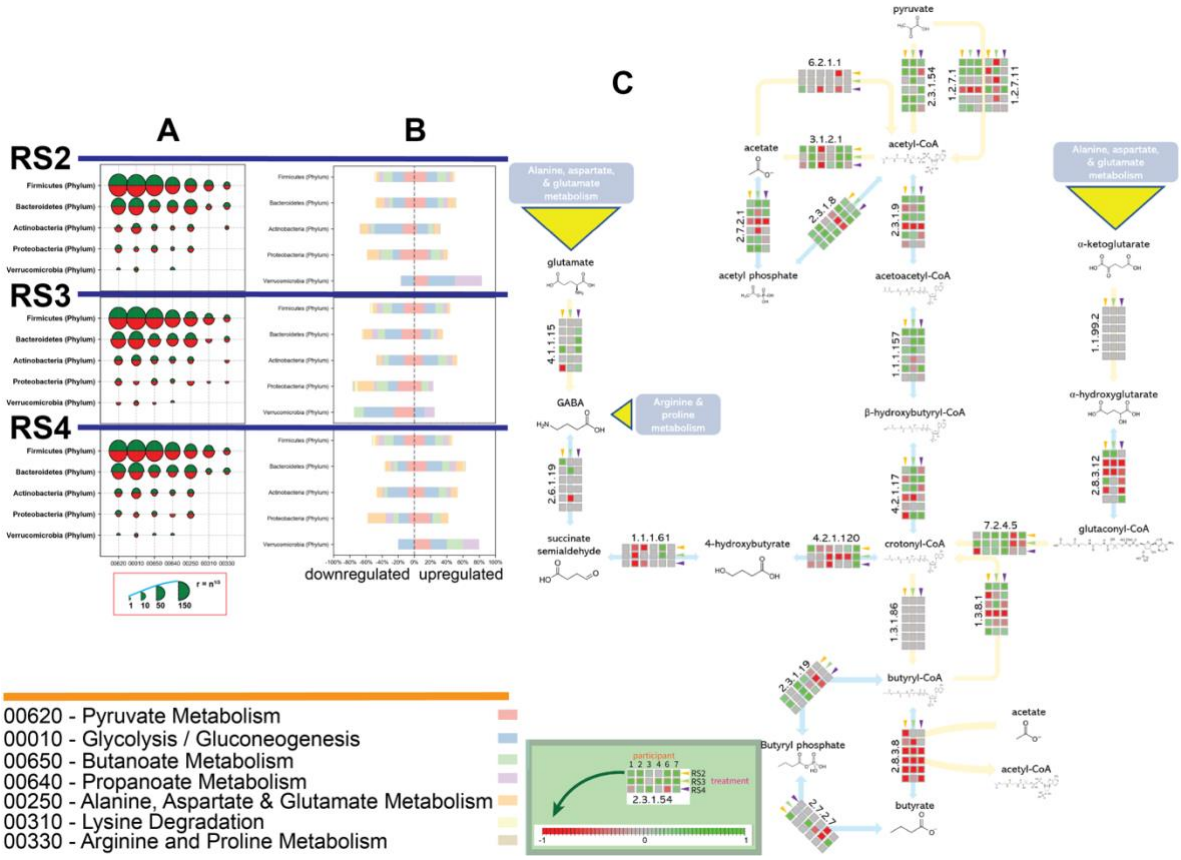

A
B
C
RS2
RS3
RS4
Firmicutes (Phylum)
Bacteroidetes (Phylum)
Actinobacteria (Phylum)
Proteobacteria (Phylum)
Verrucomicrobia (Phylum)
downregulated
upregulated
00620 - Pyruvate Metabolism
00010 - Glycolysis / Gluconeogenesis
00650 - Butanoate Metabolism
00640 - Propanoate Metabolism
00250 - Alanine, Aspartate & Glutamate Metabolism
00310 - Lysine Degradation
00330 - Arginine and Proline Metabolism
pyruvate
acetate
acetyl-CoA
acetyl phosphate
acetoacetyl-CoA
β-hydroxybutyryl-CoA
crotonyl-CoA
butyryl-CoA
Butyryl phosphate
butyrate
glutamate
GABA
succinate semialdehyde
4-hydroxybutyrate
α-ketoglutarate
α-hydroxyglutarate
glutaconyl-CoA
Alanine, aspartate, & glutamate metabolism
Arginine & proline metabolism
6.2.1.1
2.3.1.54
1.2.7.1
1.2.7.11
3.1.2.1
2.7.2.1
2.3.1.8
2.3.1.9
4.1.1.15
1.1.1.157
2.6.1.19
1.1.1.61
4.2.1.120
4.2.1.17
7.2.4.5
1.3.1.86
1.3.8.1
2.3.1.19
2.8.3.8
2.7.2.7
1.1.99.2
2.8.3.12
participant
treatment
RS2
RS3
RS4

Figure 4.8. Analysis of SCFA-linked biochemical pathways in the context of the whole microbiome and specific taxa.
(a) Taxon-function distribution of proteinGroups responding to RS2, RS3, and RS4 through their assignment to KEGG L3 pathway IDs related to SCFA production. Responding proteinGroups were selected using a partial least squares discriminant analysis (PLS-DA) of treatment group comparisons for all participant datasets, with only those proteinGroups having a VIP score ≥ 1 (significantly different relative to control) being considered. Data is relative to negative control (H_2O). The radius of each semi-circle corresponds to the number of significantly changing proteinGroups with a positive (green) or negative (red) log_2-fold-change relative to control. **(b)** Phyla-level shifts in SCFA-related functional activities (KEGG L3 pathways). The origin of these data is the same as the data used in A. The percentages of the total numbers of up- or down-regulated proteinGroups corresponding to each KEGG L3 pathway are shown as coloured blocks on either side of the median line (0%), and a color legend is displayed at the bottom-right of the figure. **(c)** Effects of RS2, RS3, and RS4 treatment on enzyme expression levels relative to negative control (H_2O) for enzymes involved in butyrate (butanoate) metabolism. Arrow color represents bidirectional or unidirectional enzyme function. In each matrix, rows of boxes represent treatment condition (RS2, RS3, or RS4), and columns of boxes represent participants. Box color represents log_2-fold change of the sum of PLSDA-VIP ≥ 1 proteinGroups annotated for a KEGG L4 ortholog of a given enzyme (E.C. numbers shown).

CoA to butyrate (E.C. 2.8.3.8, butyryl CoA:acetate CoA transferase) shows a clear down-regulation of activity for all RS compounds relative to control. Instead, butyryl-CoA conversion to butyrate through the intermediate butyryl phosphate (catalyzed by E.C.#s 2.3.1.19 and 2.7.2.7) shows varying up- and down-regulation relative to control across different RS treatments and different participants. Despite these two enzymes not showing a clear upregulation trend in most participant/treatment pairs, these data strongly suggest that these RS-responding microbiomes are significantly shifting expression of enzymes involved in the last step of butyrate production.

4.2.2.2 Taxonomic shifts

Taxonomic variation between samples was captured from iMetaLab and Cerberus through visualization procedures. Figure 4.9 outlines the overall taxonomic variability present between samples, participants, and treatment conditions in the thesis experiment. This visualization uses taxonomic abundance data that has been mean-transformed and scaled to fit in the [0,1] range (0% to 100% of total abundance). In this figure, there is a great deal of taxonomic variability between participants and less variability between treatment conditions, similar to the pattern shown by functional changes in figure 4.8. Each participant's FOS samples also appear to be more taxonomically variable relative to each participant's other treatment groups, further confirming it as a positive control in this experiment.

While RS treatment group replicates do display differences relative to H_2O in figure 4.9, the degree to which RS alters taxonomic profile is overshadowed by phylum, genus, and species profiles intrinsic to each participant's microbiome. This trend is further displayed in figure 4.10. Bray-Curtis and Jaccard dissimilarity matrices were constructed using the Python 3.6 package skbio [165], some of whose functions are integrated into Cerberus. These

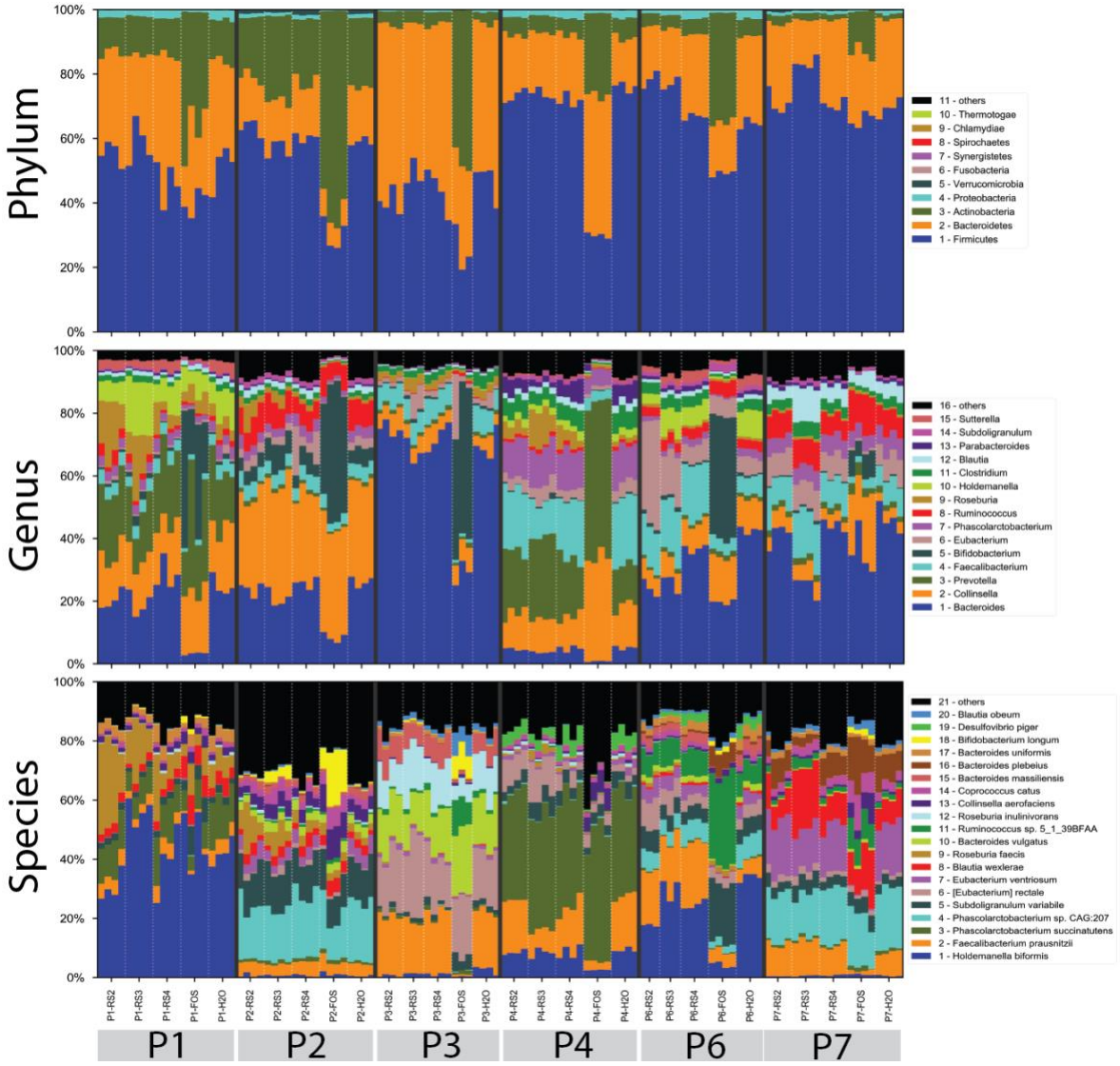
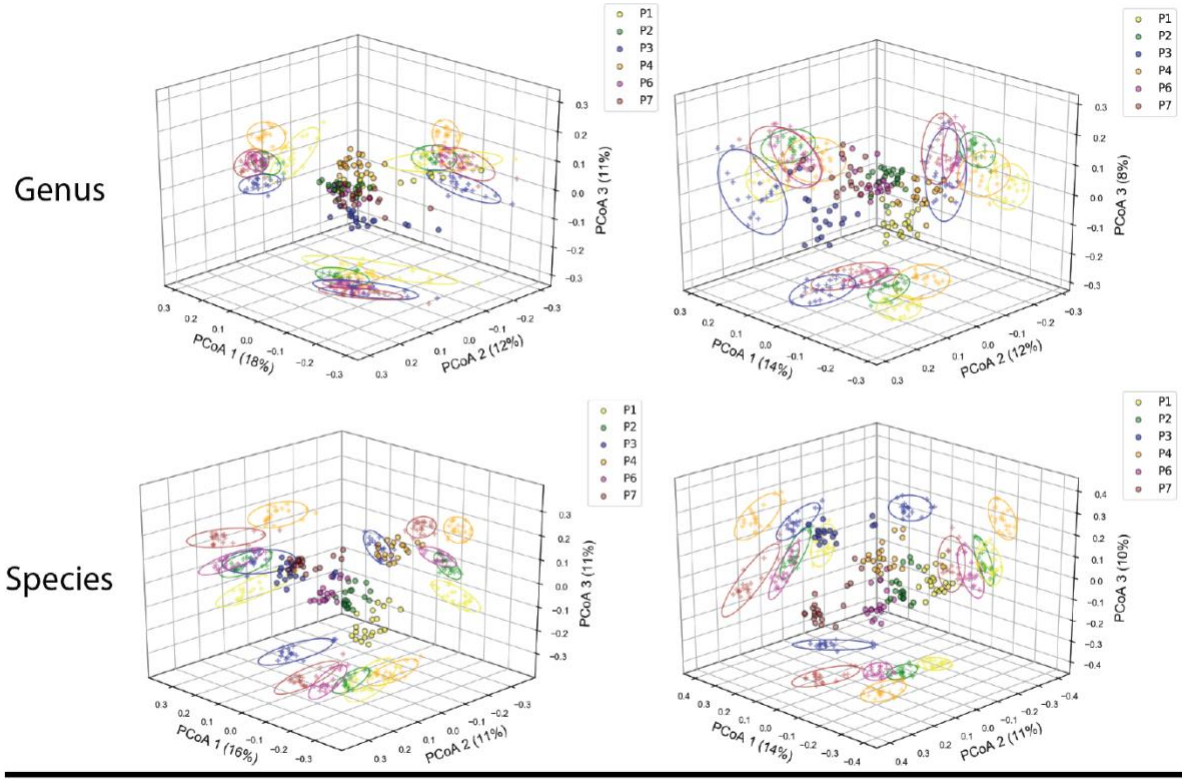
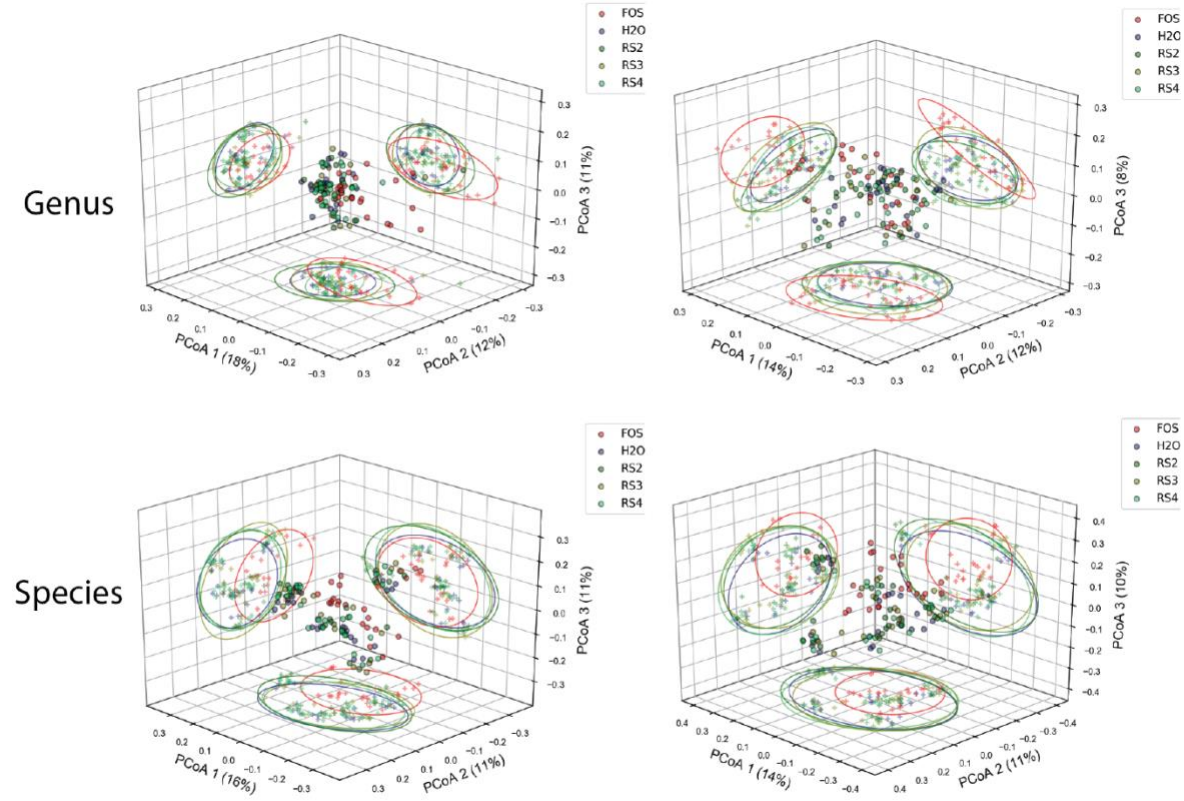

Phylum
100%
80%
60%
40%
20%
0%
11 - others
10 - Thermotogae
9 - Chlamydiae
8 - Spirochaetes
7 - Synergistetes
6 - Fusobacteria
5 - Verrucomicrobia
4 - Proteobacteria
3 - Actinobacteria
2 - Bacteroidetes
1 - Firmicutes
Genus
16 - others
15 - Sutterella
14 - Subdoligranulum
13 - Parabacteroides
12 - Blautia
11 - Clostridium
10 - Holdemanella
9 - Roseburia
8 - Ruminococcus
7 - Phascolarctobacterium
6 - Eubacterium
5 - Bifidobacterium
4 - Faecalibacterium
3 - Prevotella
2 - Collinsella
1 - Bacteroides
Species
21 - others
20 - Blautia obeum
19 - Desulfovibrio piger
18 - Bifidobacterium longum
17 - Bacteroides uniformis
16 - Bacteroides plebeius
15 - Bacteroides massiliensis
14 - Coprococcus catus
13 - Collinsella aerofaciens
12 - Roseburia inulinivorans
11 - Ruminococcus sp. 5_1_39BFAA
10 - Bacteroides vulgatus
9 - Roseburia faecis
8 - Blautia wexlerae
7 - Eubacterium ventriosum
6 - [Eubacterium] rectale
5 - Subdoligranulum variabile
4 - Phascolarctobacterium sp. CAG:207
3 - Phascolarctobacterium succinatutens
2 - Faecalibacterium prausnitzii
1 - Holdemanella biformis
P1-RS2
P1-RS3
P1-RS4
P1-FOS
P1-H2O
P2-RS2
P2-RS3
P2-RS4
P2-FOS
P2-H2O
P3-RS2
P3-RS3
P3-RS4
P3-FOS
P3-H2O
P4-RS2
P4-RS3
P4-RS4
P4-FOS
P4-H2O
P6-RS2
P6-RS3
P6-RS4
P6-FOS
P6-H2O
P7-RS2
P7-RS3
P7-RS4
P7-FOS
P7-H2O
P1
P2
P3
P4
P6
P7

Figure 4.9. Overview of taxonomic mappings for each participants' treatment groups' samples.
Taxonomic abundance data was scaled across all samples for comparison purposes.





Bray-Curtis

Jaccard

Figure 4.10. Principal Coordinate Analysis (PCoA) plots of Bray-Curtis and Jaccard dissimilarity matrices generated from comparisons between genus- and species-level abundances for all samples in the total experiment.
Plots display PC 1, 2, and 3 decompositions, with comparisons between PCs shown on the faces of the three-dimensional plots (see the axis labels on each plot for the specific PC comparison displayed on that face). Points (samples) are coloured according to the participant (above the black line) or treatment (below the black line) they belong to. Coloured ellipses on faces show the two-sigma confidence interval for each participants' or treatments' samples in that PC comparison. Points with black outlines demarcate the position of each sample in the three-dimensional coordinate space.

dissimilarity matrices were then transformed through a principal coordinate analysis (PCoA) procedure, and the three resulting latent features were projected onto axes displayed in this figure. PCoA analyses using these distance metrics resolve similar structures from the iMetaLab taxonomic dataset. When considering the participant that each sample originates from, both Bray-Curtis and Jaccard distances show that samples cluster among their shared participant in the first three PCoA features at both the genus- and species-level. When samples are grouped by their shared treatment, this clustering largely disappears, except for some discernable differences shown in the FOS-treated samples (dark red color in the bottom four plots). Taken together, these taxonomic results affirm that inter-microbiome variability plays a much larger role than treatment condition in shaping the bacterial profile of each sample.

A number of bacterial species have been noted in literature as being involved in RS degradation and SCFA production [71–73, 75, 82, 83, 86, 166]. The iMetaLab taxonomic output dataset was used as the source of overall abundance data, which was then scaled (x/sum) and the $\log_2$-fold change in RS2, RS3, and RS4 samples vs. H_2O samples was calculated for a collection of species. Figure 4.11 displays $\log_2$-fold change values for different divisions of data at the species level.

These data show a few trends. In particular, the species *Eubacterium rectale*, *Roseburia faecis*, *Butyrivibrio crossotus*, and *Butyricicoccus pullicaecorum* have a $\log_2$-fold change value of at least 1 (approximately double the abundance versus control) for at least one RS compound when the species values of all samples from all participants are averaged. These species also appear to follow a trend of upregulation at the participant level regardless of RS treatment or participant (see heatmap panels at bottom of figure 4.11). *Bacteroides thetaiotaomicron* shows a $\log_2$-fold change below -1 for RS3 and overall is negative for the

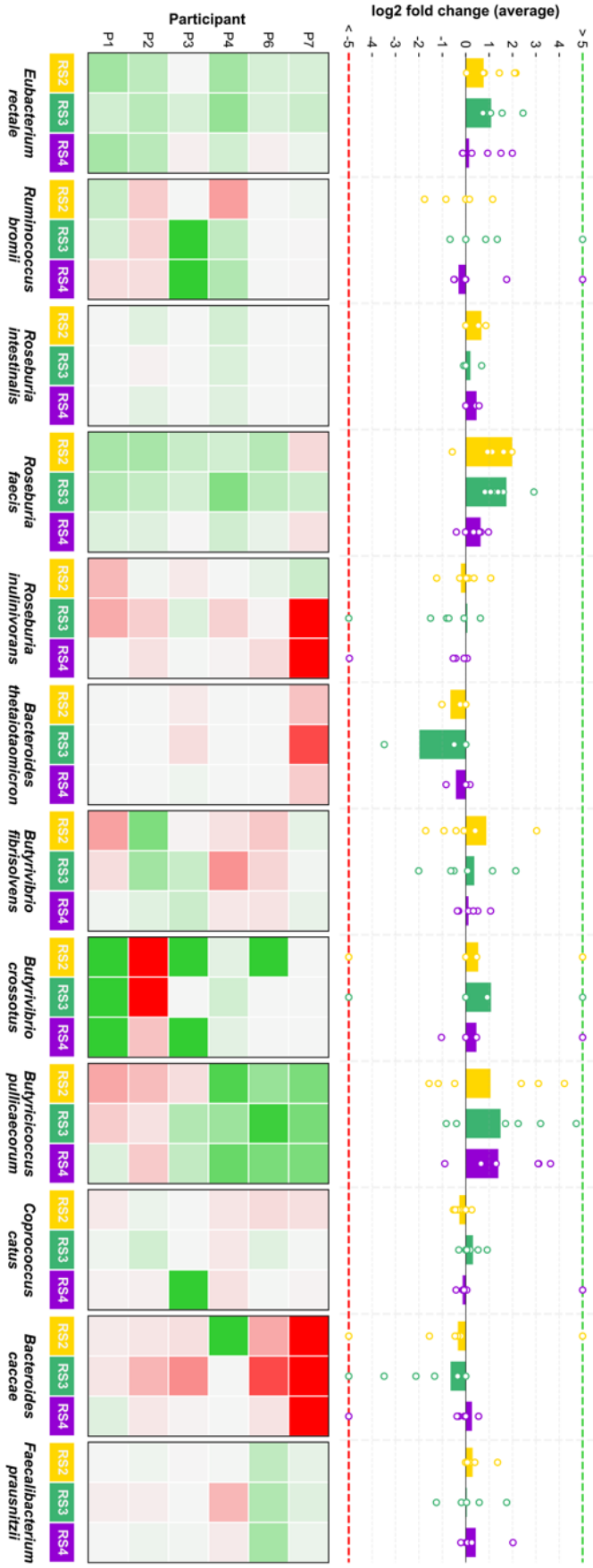

log2 fold change (average)
> 5
4
3
2
1
0
-1
-2
-3
-4
< -5
Participant
P1
P2
P3
P4
P6
P7
RS2
RS3
RS4
Eubacterium rectale
Ruminococcus bromii
Roseburia intestinalis
Roseburia faecis
Roseburia inulinivorans
Bacteroides thetaiotaomicron
Butyrivibrio fibrisolvens
Butyrivibrio crossotus
Butyricicoccus pullicaecorum
Coprococcus catus
Bacteroides caccae
Faecalibacterium prausnitzii

Figure 4.11. log2-fold change values for scaled species-level abundances of SCFA-producing species.
The bottom row of heatmaps shows log2-fold change of RS2/RS3/RS4 vs. negative control (H2O) for the mean of each participant's treatment samples in each box. The top panel shows log2-fold change for the mean of all participants' treatment samples as bars and shows the same values as the heatmap as coloured dots. Heatmap color range corresponds to the y-axis of the top panel (green is positive, red is negative), maximum is +5 and minimum is -5. In the top panel, log2-fold change values above 5 are displayed as dots at the '>5' y-axis tick, and log2-fold change values below -5 are displayed as dots at the '<-5' y-axis tick. In the bottom panel, the same logic follows, with values outside of the [-5,5] range shown as deep red or deep green, respectively. Log2-fold changes of 1 or -1 correspond to a doubling (or halving) of the abundance of that species relative to control.

other RS treatments, although this appears to be due to major negative values from two participants' (P3 and P7) samples skewing the overall fold change profile in the negative direction. Other species such as *Ruminococcus bromii, Roseburia inulinivorans, Butyrivirbio fibrisolvens, Coprococcus catus, Bacteroides caccae*, and *Faecalibacterium prausnitzii* show a highly-variable abundance profile relative to control. Taken together, these data show that the RapidAIM procedure captures species-level taxonomic shifts taking place in RS-treated microbiome samples relative to control, with a number of species integral to SCFA production displaying significant abundance changes when data from all participants and each individual participant are considered.

4.3 LINEAR PROGRAMMING RESULTS

Overall, the linear programming strategy more accurately recovers the bacterial species-shared peptide associations and their abundances than the naïve approaches. For the identified peptides, a file was generated to show all possible bacterial species that contain proteins matching sub-sequences to the identified peptides. In effect, this file represents a dataset of all possible peptide-to-species mappings based purely on sequence matching and was used as the basis for peptide intensity redistribution through linear programming.

4.3.1 Post-optimization peptide redistribution

To calculate taxonomic and functional shifts, the dataset was split into three parallel organizations: a dataset holding the post-LP-optimized peptide-species abundance assignments; a dataset holding peptide-species abundance assignments that arise from the *naïve-split* method (a given peptide's intensity is split evenly amongst the species that it can possibly be assigned to); and a dataset holding peptide-species abundance assignments that arise from the *naïve-duplicate* method (a given peptide's intensity is duplicated for the species that it can possibly be assigned to).

4.3.1.1 Taxonomic changes

Linear programming optimization of the peptide-species abundance assignments produces a defined shift in the taxonomic characteristics of the dataset. Figure 4.12 shows an overview of general taxonomic changes compared to naïve peptide-species distribution approaches. The naïve approaches will necessarily assign some positive intensity value to *all* of a peptide's constituent species, dependent on the coverage of the sequence database being used to generate peptide-species mappings. The subset-species problem is shown plainly through these naïve distribution results. Linear programming addresses the subset-species problem by influencing the distribution of peptide intensities amongst species to be more parsimonious. A division of peptide-species mappings are left with no value after linear programming, making the summed intensities of some taxa equal to zero. This effect is shown clearly in the right-side bar chart of figure 4.12, where both naïve approaches have many more species (bars) present than the linear programming approach does.

4.3.1.2 Functional changes

To determine functional shifts described by peptide abundances that arise from peptide-species distribution through the LP procedure, an annotation dataset needed to be constructed. Here, the protein sequence database used for peptide identification was filtered to contain only protein sequences that an experiment's peptides matched to. Peptides were mapped to this subset of proteins. These proteins themselves were mapped to their taxonomic LCA according to the same sequence database – in effect, the same as the linear programming input dataset.

There are a few things to consider when performing functional analyses on these data:

- A peptide can be mapped to multiple species.

- Of these species, one or many will have received a share of the peptide's intensity after linear programming optimization.

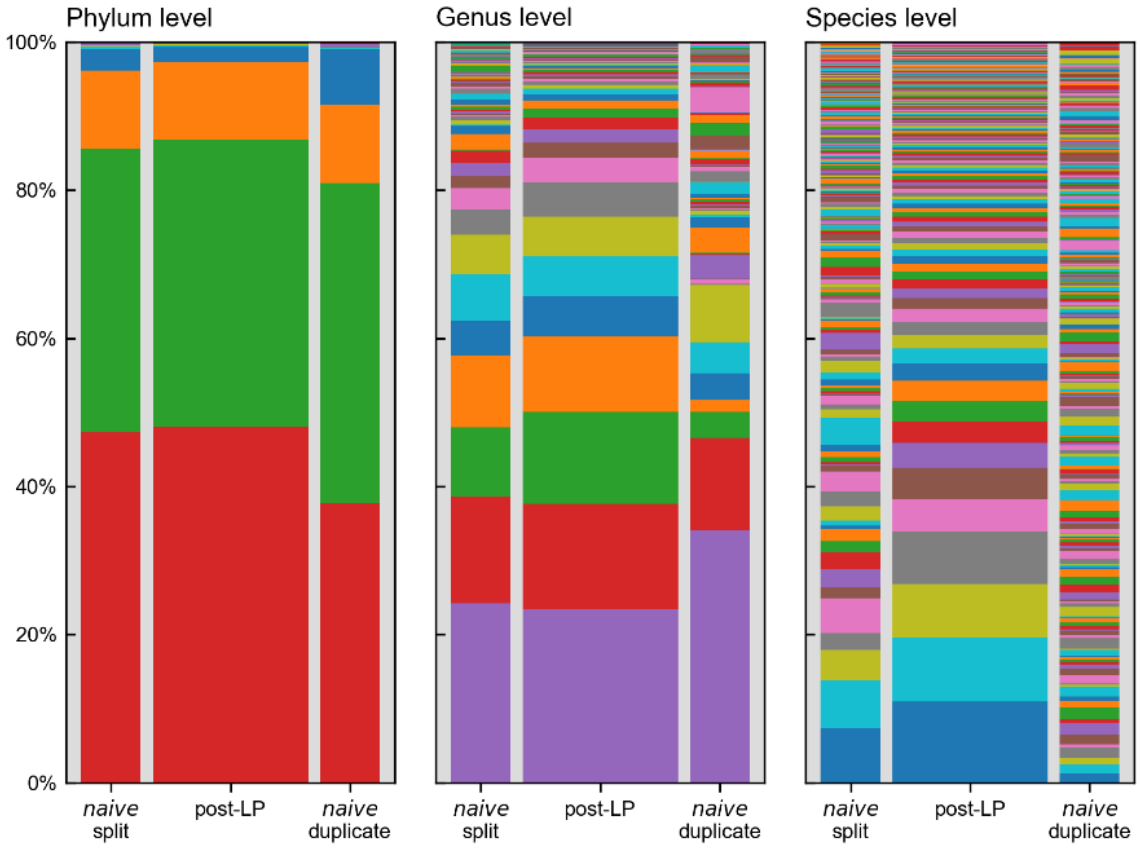


Figure 4.12. Taxonomic shifts as a result of linear programming optimization at three different taxonomic levels (phylum-, genus-, and species-level).
Coloured bar blocks represent the relative abundance of an OTU at that taxonomic level.

Further:

- A peptide can be mapped to multiple proteins.
- Proteins can be mapped to one, multiple, or no KEGG L4 IDs.
- Proteins can be mapped to one, multiple, or no taxa.

To form the taxa-function annotation dataset, these data must be constructed such that only the KEGG L4 IDs corresponding to proteins that *themselves* map to species with an above-zero intensity for a given peptide are considered.

Figure 4.13 displays the result of this taxa-functional mapping for the linear programming output dataset. The results show that linear programming distribution of peptide intensities amongst constituent species causes a defined shift in the functional character of the metaproteomics dataset when compared to naïve methods of peptide-species distribution. This shift is most clearly seen in how the $\log_2$-fold change between linear programming results and naïve-split results (left panel of figure 4.13) shows relative decreases (red) or complete absence of values (black) in most of the found taxa-function mappings. It is important to note, however, that the linear programming procedure also managed to assign peptide intensities to species that, when abstracted to taxa-functional mappings as shown in this figure, produced a relative *increase* in taxa-linked functional intensity, denoted by the sparse light-blue boxes in the top-left corner of figure 4.13's far-left panel.

4.3.2 Validation results

Assessing the performance of the LP optimization procedure falls on the validation protocols described in the methods section (specifically *4.1.4.3, Validation*).

Figure 4.14 shows an overview of the validation results for one sample. The validation procedure's ε factor adds a level of randomness to each peptide intensity value. It

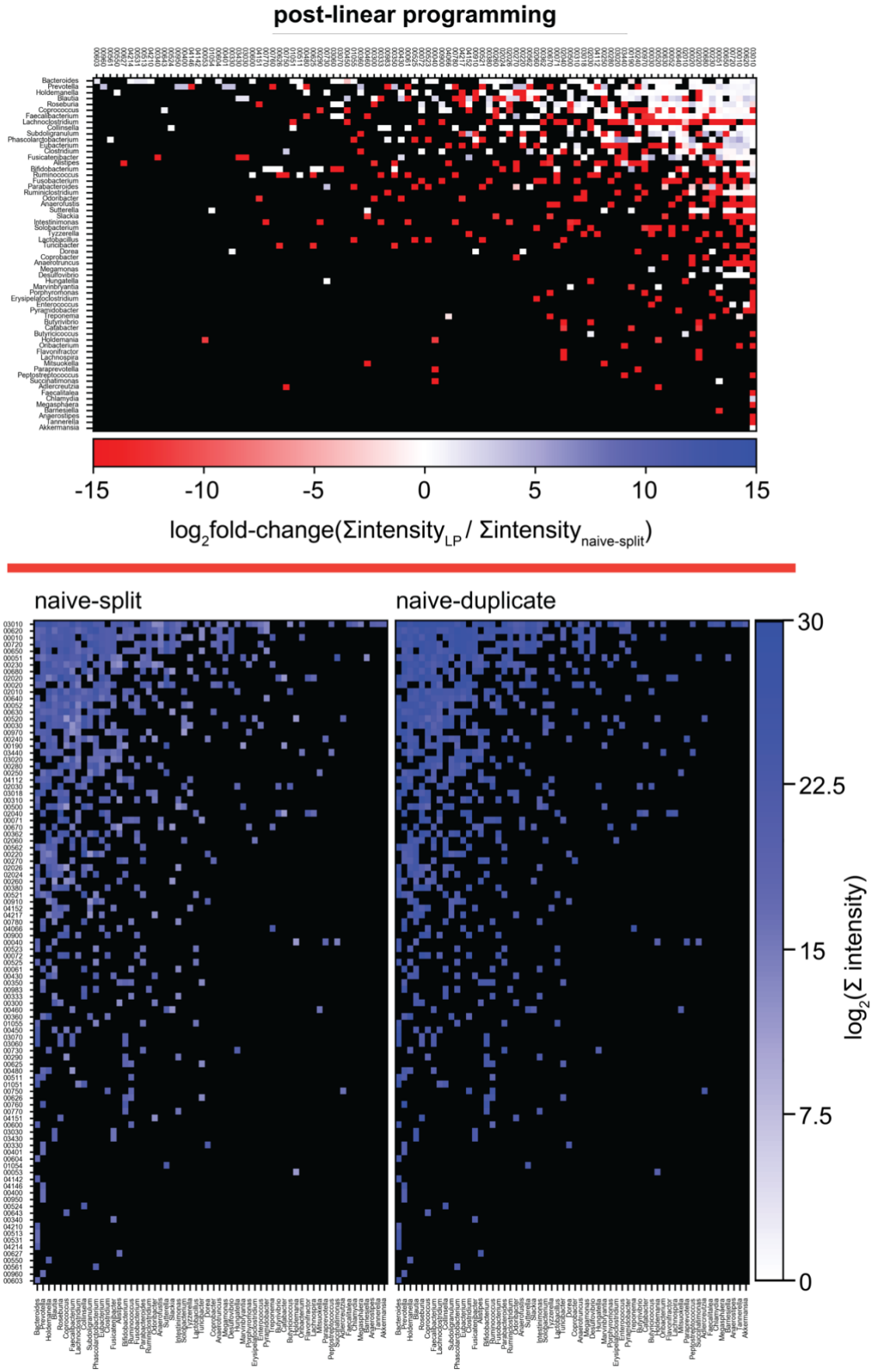

post-linear programming
-15
-10
-5
0
5
10
15
log2fold-change(Σintensity LP / Σintensity naive-split)
naive-split
naive-duplicate
30
22.5
15
7.5
0
log2(Σ intensity)

Figure 4.13. Post-linear programming peptide distribution results mapped onto genus-level taxonomy and KEGG L3 pathways IDs.
Top box shows $\log_2$-fold change values comparing post-linear programming summed intensities vs. naïve-split summed intensities. The value in each box represents the $\log_2$-fold change of the sum of all peptides that map to that KEGG L3 ID/genus for post-LP vs. naïve-split. Bottom boxes show the $\log_2$-summed intensity for peptides corresponding to genera mapped to KEGG L3 IDs. Left-side are results for the naïve-split dataset, and right side are results for the naïve-duplicate dataset.

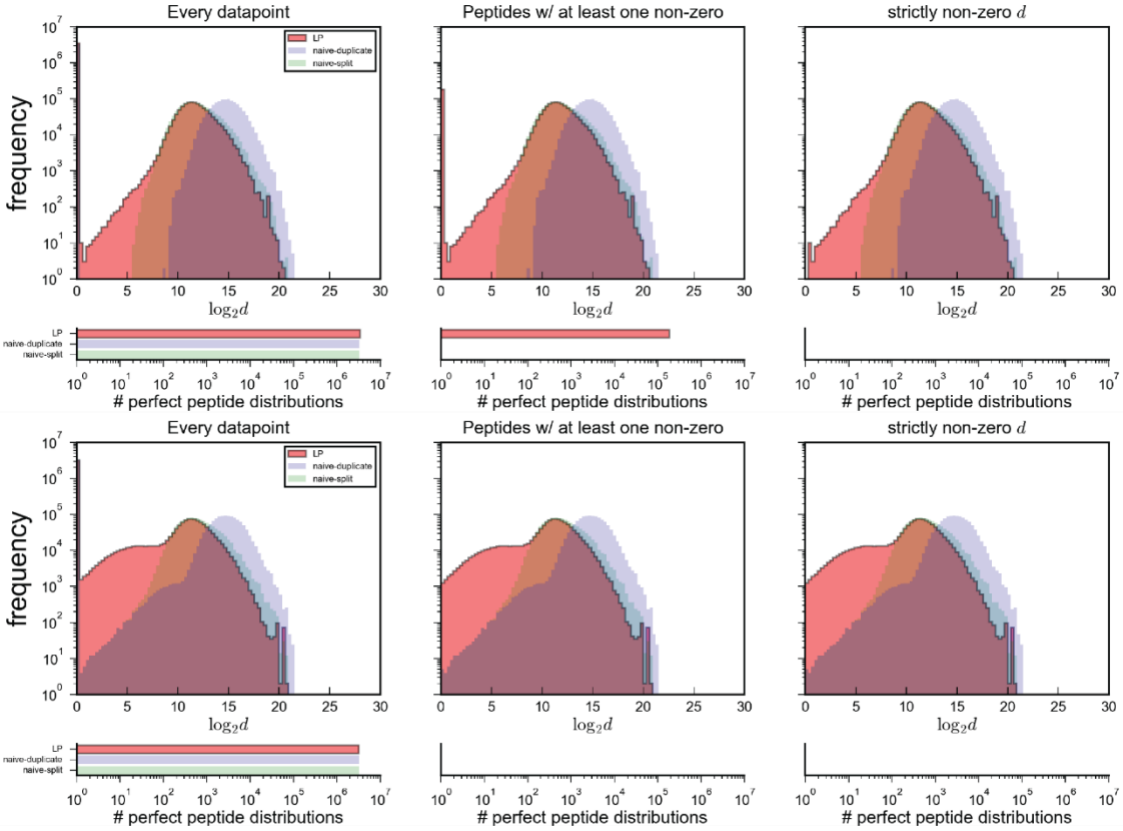

Every datapoint
Peptides w/ at least one non-zero
strictly non-zero d
frequency
LP
naive-duplicate
naive-split
$\log_2 d$
perfect peptide distributions

Figure 4.14. Histogram of Euclidean distances calculated from linear programming validation results of one sample, run over 1000 iterations, using σ factors of 0 (top row) and 0.01 (bottom row).

is selected from a $\mathcal{N}(1, \sigma)$ distribution whose σ value remains the same for all iterations in a validation run, meaning that the sampling range for ε does not change, however ε still is random for each peptide intensity value in each iteration. ε accounts for stochasticity in the experiment that produced the peptide intensity values, and its different values' effect on the validation results are used as a benchmark for how the validation procedure reacts to initial parameters, and, further, how those parameters fit the pre-assigned species abundances after all validation iterations are complete. The D_V matrix holds all possible peptide-species pairs, and a peptide's set of possible species is based purely on the underlying database's coverage of taxonomically-linked protein sequences, meaning that for any peptide in D_V, there may (and most likely will) be many mapped species that are not present in the sample but nonetheless can receive some share of intensity values after optimization has completed. The S vector of pre-assigned species abundances will not change throughout all iterations of the validation procedure, and the actual optimization procedure is agnostic of S during its run. Therefore, increasing values of σ can be interpreted as giving the validation procedure an increasing ability to produce an optimized D_V matrix holding peptide-species vectors that fit S's species abundance values.

Given this interpretation, figure 4.13 shows two results.

First, a validation procedure where σ is 0 means that all ε values will be 1 and no peptide intensities will change from their original values in any iteration, yet regardless of this model rigidity, LP resolves more perfect peptide intensity distributions than either of the naïve-split or naïve-duplicate approaches. Further, when LP arrives at a solution for a peptide's intensity distribution that is *not* perfect – there is a Euclidean distance above 0 – the degree of that error is less that the degree of error in either of the naïve approaches (see the

left-side tail of each distribution in the histograms). When the σ value is 0.01 (results shown in second row of histograms in figure 4.8), an element of randomness is introduced into the validation procedure's initial conditions as described before; the results show that the LP procedure resolves more perfect peptide distributions *and* its degree of error is less than either of the naïve approaches. The LP degree of error also exhibits a somewhat bimodal distribution, with a skewed clustering of incorrectly-distributed peptide intensities in the lower range of histogram bins ($d < 10$). Taken together, these validation results show that LP resolves more perfect peptide distributions amongst constituent species than both naïve approaches, and, further, when LP produces incorrect peptide distributions, it is wrong *to a lesser degree* than either of the naïve approaches.

5 DISCUSSION

In the thesis project, the Figeys lab-developed RapidAIM metaproteomics assay was implemented as a procedure to describe functional and taxonomic changes present in microbiomes encountering RS 2, 3, and 4. Further, the need for comprehensive and efficient data analysis to investigate functional and taxonomic changes lead to the development of the Cerberus software package. Lastly, a LP approach was developed, deployed, and validated to address the shared-peptide problem in metaproteomics studies.

The thesis project looked at taxonomic and functional characteristics of microbiomes exposed to RS 2/3/4 versus a negative control (H_2O). Through the RapidAIM procedure, a massive dataset representing microbiomes from six individuals was collected, generated, and analyzed.

When comparing $\log_2$-fold change values of SCFA-involved species in RS-treated samples vs. control (Figure 4.11), many species exhibited a significant (>=1) relative increase in abundance, suggesting their movement into new microbiome taxonomic niches as

a result of RS treatment. These significantly-increasing species include: *Butyricicoccus pullicaecorum* (a species tolerant to environmental variability and suggested as a probiotic for IBD patients [167]), which showed a $\log_2$-fold change value above 2 for all three RS types and showed the same profile in the majority of participants at the participant level (heatmap in figure 4.11); *Roseburia faecis* and *Eubacterium rectale,* showing $\log_2$-fold change values above 1 for RS2/RS3 (*R. faecis*) and RS3 (*E. rectale*) and the same positive profile in the vast majority of participant-specific comparisons; and *Butyrivibrio crossotus,* showing a $\log_2$-fold change above 1 for RS3-treated samples but highly-variable participant-level responses (very high values for P1, P3, and P6, but very low values for P2). Two *Butyrivibrio* species (*crossotus and fibrisolvens*) exhibited relative increases in abundance (to a lesser degree – between 0 and 1) for all three RS types when all participants' data are considered, however there doesn't appear to be a symmetrical participant-level up- or down-trend for these species. Further, the species *Ruminococcus bromii, Roseburia inulinivorans, Coprococcus catus, Bacteroides caccae,* and *Faecalibacterium prausnitzii* showed highly variable responses to RS treatments in each participant that results in offsetting $\log_2$-fold change values when all participant's values are considered.

Taken together, these data show that many species, with literature supporting their function as SCFA producers, expand into new microbiome taxonomic niches when exposed to RS. However, there are some notable exceptions in these data: *Ruminococcus bromii*, a species described as integral to SCFA production in the human gut microbiome [67, 72, 86, 87], showed very little change relative to control in all three RS types, and *Bacteroides thetaiotamicron*, containing a *sus*-like glycan import system [68] and involved in cross-feeding with *Eubacterium rectale* to produce SCFA cooperatively [75], displayed a large negative change in abundance relative to control for all three RS types (opposite to

Eubacterium rectale with increases in relative abundance). Based on these data, it is possible that *Eubacterium rectale* is already exposed to enough RS cleavage products from overall microbiome metabolism that *Bacteroides thetaiotamicron* may not need to participate in a cross-feeding system with *rectale*, or the activity of *thetaiotamicron* as a primary RS degrader for *rectale* may be independent of changes in its relative abundance within the microbiome.

The common theme amongst all of these species-level abundance change profiles is that inter-individual variability overwhelms treatment-specific responses for RS compounds. This theme extends to other taxonomic- and functional-level analyses in this thesis project; the phylum-, genus-, and species-level taxonomic profiles and PCoA projection plots in figures 4.9 and 4.10 serve as evidence of this. At the phylum level, *Firmicutes* and *Bacteroidetes* dominate in all samples, but *Actinobacteria* varies in its relative abundance, with much higher shares in P1 and P2 compared to other participants (top panel, figure 4.9). Further, P3 has a more even *Firmicutes*/*Bacteroidetes* split as a share of total abundance compared to all other participants. Importantly, dramatic shifts do take place when samples from any participant are subject to the FOS positive control, with a marked relative increase in *Actinobacteria* abundance compared to other phyla. In all samples, phyla-level inter-individual taxonomic variability translates to variability at lower taxonomic levels, such as genus and species, with entire taxa appearing and disappearing based on each sample's originating individual. FOS's variability relative to negative control in these microbiomes is captured through the RapidAIM procedure regardless of sample source, and evidence showing FOS-treated microbiome samples changing drastically gives a point of reference for RS-treated samples, whose variability exists on a much smaller scale. The Bray-Curtis and Jaccard distance PCoA projections shown in figure 4.10 support this participant-level

taxonomic variability at the genus- and species-level, and the high overlap present in PCoAs for treatment conditions reinforces the finding that inter-individual variability overpowers inter-treatment variability in these microbiomes.

Considering inter-individual variability is therefore crucial when interpreting results from multi-individual microbiome experiments where treatment conditions confer smaller-scale changes. Human gut microbiome enterotypes have been suggested to be geographically-, age-, diet-, and host-genetic-dependent [168], and bacterial species associated with the gut microbiome undergo genome recombination at a rate 25 times higher than other non-human ecological community isolates [169]; these features of microbiomes further capture the scope of their inter-individual variability. While average values for *all* of a participant's data can be resolved and are useful in determining experiment-wide trends, context is critical; multi-individual data pools serve as pieces that contribute to a snapshot of microbiome complexity and add complexity to the process of drawing conclusions from microbiome metaproteomics datasets.

Taxonomy informs function in the microbiome, but bacterial communities in symbiosis must maintain their functional flux in response to changing environmental conditions, such as RS treatment. Results from the thesis project show how, regardless of participant and taxonomy, functional profiles remain similar when subjected to different conditions (Figure 4.7). This functional homeostasis arises from taxonomic shifts as well as horizontal gene transfer between bacteria [170]. Despite this, results from this project show how defined responses to RS treatment can be gleaned from overall function as well as functions linked to taxonomic shifts. Figure 4.8 exhibits how analyses focused on overall microbiome protein output resolve enzyme- and pathway-level differences, and, further, how the interplay between taxonomy and function modulates microbiome response.

Different RS forms cause fluctuations in proteins associated with KEGG L3 pathways, and this variability is captured in panels A and B in figure 4.8. PLSDA VIP ≥ 1 ProteinGroup intensities are assigned KEGG L4 annotations through Cerberus and database search, and the summed KEGG L4 intensity is derived. KEGG L4s are then aggregated to KEGG L3 pathways (Figure 4.8A). These data show that the majority of variable functional activity emerges from *Firmicutes* and *Bacteroidetes* bacteria, and that different RS forms produce different taxa-function-linked responses. RS3-treated samples have more downregulated KEGG L4 IDs (red semi-circles) compared to upregulated KEGG L4 IDs (green semi-circles) for *Bacteroidetes* proteins across KEGG L3 pathways involved in SCFA production, while RS2- and RS4-responding *Bacteroidetes* display more up-regulation for KEGG L4 IDs involved in SCFA production. This pattern can be seen in Figure 4.8B, where KEGG L4 variability in Bacteroidetes trends down in RS3 and slightly up in RS2 and RS4. Regardless of RS treatment, *Firmicutes* appears to stay balanced between upregulated and downregulated KEGG L4s, suggesting that *Bacteroidetes* shifts its functional profile relative to itself to a higher degree in response to RS forms than *Firmicutes*. *Bacteroidetes* species have been described as major drivers of propanoate production in human gut microbiomes, and *Firmicutes* species are associated with butyrate production overall [171]; in the context of these data, there doesn't appear to be a significant phylum-level trend towards either of these relationships in RS-treated microbiome samples.

When considering treatment- and participant-specific functional enrichment, building pathway maps like the one in panel C of figure 4.8 provides an intuitive and information-dense visualization of enzyme-specific variability in specific biochemical pathways. Analyzing biochemical pathways, especially in microbiome studies where functional data is derived from thousands of different bacterial species in a community, digging into enzyme-

level variability in response to compounds presents a high-resolution look at low-level microbiome metabolism and is crucial for defining precise modulations. Results from the thesis project show that enzymes involved in butyrate production are upregulated at many precursor steps, and, interestingly, suggest that these RS-treated microbiomes preferentially utilize a butyryl phosphate intermediate as a final precursor to butyrate instead of direct conversion from butyryl-CoA to butyrate through the enzyme butyryl CoA:acetate CoA transferase (E.C.# 2.8.3.8). Louis *et al*. (2017) describe how, in the microbiome, this last step to butyrate production produces ATP through the butyryl-phosphate mechanism, but can also conserve the energy of the CoA group on butyryl-CoA by transferring it to acetate, itself an SCFA and a common substrate in the microbiome [171]. The authors also point to the prevalence of CoA-transferase enzymes in the microbiome as a factor in its function, noting that one microbiome species has 14 such enzymes encoded in its genome [172], and that CoA transferases have been shown to be accepting of many different substrates. Data from this project show a notable downregulation of CoA transferase activity at this final conversion step, suggesting that these microbiomes prefer ATP production through butyryl-phosphate conversion to CoA maintenance, possibly because of an abundance of CoA produced through other metabolism, an increase in CoA transferase activity at other metabolic reactions, or a forced downregulation of this specific CoA transferase activity in order to produce ATP.

Interpretations of enzyme-level shifts in response to RS treatment must take into account the experimentally-derived information on each enzyme and are dependent on a few different factors. The KEGG database provides enzyme orthologues to KEGG L4 IDs through KEGG ENZYME [173], an annotation database holding mappings between KO (KEGG Ortholog) identifiers and Enzyme Commission (E.C.) numbers [174]. KO identifiers

represent “functional orthologs in the context of [KEGG biochemical] pathway maps [and] are also defined as [protein] sequence similarity groups”[173], all of which have been experimentally characterized in literature. Their relationship is many-to-many: one KO can correspond to multiple E.C.s, and multiple KOs can correspond to one E.C. The KEGG PATHWAY database uses these KO-to-E.C. mappings to construct biochemical relationships as reactions, and mappings from the KEGG PATHWAY database are what PathviewR builds its pathway figures upon.

Since KO identifiers are by definition built on functional *or* sequence-based orthologs, there exists three levels of abstraction from a MS-derived peptide identification to an ultimate value assignment for a specific E.C. number. After proteinGroups are built through MaxQuant (level 1), Cerberus’s annotation procedure assigns one or many KEGG L4 IDs (KO identifiers) to each proteinGroup (level 2), and then, through the KEGG ENZYME database, the proteinGroup’s KEGG L4 IDs are mapped to E.C. codes (level 3). In this last level of abstraction, it cannot be determined whether a KEGG L4 ID-to-E.C. mapping exists because of protein sequence similarity or functional similarity; nonetheless, PathviewR performs this last step to produce a figure output. All of these steps rely on accurate and robust databases that limit generalizations as much as possible. The problem with a mapping and aggregation method such as this is that database limitations may force mapping generalizations that are ambiguous in the context of the exact pathway, function, and enzyme that catalyzes a reaction. An example of a generalization could be mapping a KEGG L4 ID to an E.C. whose mapping set consists of up to hundreds of KEGG L4 IDs, as many E.C.s describe “non-specific” enzyme functions that could be part of many different pathways. In effect, this means that a peptide found in an experiment originating from a specific bacterial protein expressed in a microbiome sample could ultimately have its

intensity mapped to a vast array of proteinGroups, KEGG L4 IDs, and E.C.s, contributing its intensity value to all of these levels of abstraction, and, in the context of PathviewR figures, contribute its intensity value to enzymes in multiple KEGG pathways. Taken together, these factors underscore the importance of good, well-annotated, and experimentally-validated databases that limit generalization as much as possible, all of which are goals of the KEGG project. As more experimental data is gathered and validated by the KEGG consortium, improved generalization accuracy will be achieved and lead to increased resolution of orthologous pathway mapping and functional enrichment analyses, especially for microbiome studies.

In the thesis project, a linear optimization approach was developed to address the shared-peptide problem present in metaproteomics studies. This problem manifests itself in the data analysis stage of a metaproteomics experiment, where knowing the abundance of peptides and proteins in a sample allows for derivation of bacterial abundance at different taxa.

The peptide-/protein-to-taxa derivation step is often performed through aggregation and mapping of the peptide sequences to a protein sequence database, whose data was either gathered through a metagenomic experiment on the same microbiome sample or composed of an amalgamation of many different metagenomic experiments to generate the most taxonomic coverage possible using current techniques. An example of the latter is the Integrated Gene Catalog (IGC) [136], the protein sequence database used by iMetaLab analyses. The IGC is a combined database of 1,267 metagenomic sequenced microbiome samples, with protein sequences extrapolated *in silico* from gene sequences. This database provides both protein sequence information and taxonomic mapping to these proteins. When bottom-up metaproteomic datasets are checked against the IGC protein database, each

peptide is assessed for sequence matches to each database protein, forming an experiment-specific peptide-protein map.

This procedure streamlines proteomic abundance measurements through an implicit assumption: that any species present in the experimental microbiome sample *could* be present in the IGC, as the originating organism for both experimental microbiome samples and IGC samples is the same (human) and taxon definitions from all 1,267 IGC datasets combined together should cover a very broad range of taxa. However, pitfalls with using this approach exist. Importantly, the IGC's large database coverage means that there is an increased possibility of false mappings between any experimental peptide sequence and any IGC database protein sequence. This potentially causes situations where an experimental peptide ends up being mapped to a set of proteins, some of which are not present in the experimental sample, inserting more ambiguity into defining accurate taxonomic mappings based on peptide measurements alone. The opposite problem may arise as well: despite IGC taxonomic coverage, there may be bacterial taxa present in the experimental sample that are not present in IGC, especially when doing species- or genus-level taxonomic assignments, because of the required sequence resolution needed to confidently distinguish sequences belonging to a given species or genus from another.

These pitfalls all stem from a common issue: the sequence database does not originate from the same sample as the experimental peptide measurements. Addressing these pitfalls during experimental analysis may involve using filtering procedures with arbitrarily defined parameters. An example of this is when the taxonomic abundance profile of a microbiome sample is inferred from its experimental peptide measurements. A table of taxonomic abundances can be constructed by looking at each peptide measurement, mapping its sequence to a database such as IGC, and summing the intensities of all peptides that have

been mapped to any given species' proteins. To calculate the overall abundance of higher-level taxa, such as genera and phyla, all the child node intensities of each taxa are summed together. This creates a peptide intensity-based taxonomic tree structure for the experimental sample. However, a researcher should be wary of having confidence in an aggregation procedure such as this. Errors here can be broken into two categories: database-dependent and microbiome-dependent. These error modes obscure real taxonomic abundance measurements through metaproteomics procedures and will likely cause an over-assignment of abundances to all taxa.

Database-dependent errors arise when a peptide is mapped to a set of species in a sequence database, and a subset of those species are either not actually present in the microbiome sample or do not express the peptide in the sample. If the procedure to define taxonomic abundances described in the last paragraph is followed, it will result in a taxonomic tree containing entries for taxa that do not exist, entries for taxa that do exist but not at the abundance ascribed to it by the summing procedure, or, more likely, a mixture of both scenarios. Microbiome-dependent errors arise when a peptide is mapped to a set of species in a sequence database, but the subset of species that are *actually* responsible for expressing the protein containing that peptide is undefined.

There a few naïve methods to address peptide-species mapping ambiguity. One could simply remove all shared peptides from consideration, leaving a dataset containing only peptides unique to one species in the sequence database; however, this method removes a majority of peptides found through mass spectrometry from consideration. In the thesis project, two methods were applied: splitting peptide intensity measurements evenly amongst each mapped species (*naïve-split* datasets), or duplicating intensity measurements to each mapped species (*naïve-duplicate* datasets). On top of these two naïve methods, a further

filtering step can be performed, where the taxonomic tree is constructed to only contain species that have *n* peptide mappings. Defining a value for *n* does present more problems, however: low *n* values will be too permissive of multiplexed peptide-species mappings, and high *n* values will be too restrictive and lead to true peptide-species mappings and intensities being filtered out. It is therefore up to the researcher to define an *n* value that allows for accurate taxonomic mappings in the context of the error that a given *n* value may introduce into their dataset. For instance, iMetaLab uses $n=1$ and $n=3$ as its cut-offs to reject a species from the final calculated taxonomic tree.

These strategies all inherit one major underlying flaw: there is no gold-standard dataset with which to compare results. Metaproteomics experiments can only work with the observed data that they output, composed of peptide intensity measurements, which then must be combined, altered, and transformed to produce protein- and taxonomic-level abundance measurements. A standard dataset could be constructed using a targeted whole-genome metagenomic experiment on a microbiome sample, producing gene-level abundance information on its species, and then using the results to build a parallel taxonomic abundance dataset that can be used as a reference for metaproteomic peptide-to-taxonomy assignment estimation; however, this method also relies on the assumption that a bacterial species' total protein expression will be closely linked to its real abundance in the microbiome. In a broad sense, this is a somewhat safe assumption to make, as there are so many bacteria in a microbiome community and the contribution of a low-abundance bacteria to the expression of a certain protein may be insignificant in a relative sense. However, the microbiome is a dynamic system with bacteria cooperating, competing, and vying for a place in its equilibrium, and a low-abundance species may become higher-abundance as a result of producing a specific antimicrobial compound that kills off a competitor, allowing it to gain a

greater foothold in the community. This dynamism suggests that there are cases where protein expression is not closely linked to bacterial abundance, but rather linked to several complex factors, such as opportunistic bacterial protein expression, environmental changes, and susceptibility of the microbiome to minor and major alterations. This is an important concept to consider for all researchers, and specifically those studying the effects of environmental changes such as introduction of drugs and other compounds to the microbiome, where defining bacteria that are significantly over-expressing a certain protein or group of proteins is important to identifying targeted effects.

Sample-specific sequence databases could be constructed using a whole-genome sequencing procedure, and protein sequences would then be derived using *in silico* methods, much like IGC's protein sequence database. Using sample-specific sequence databases for metaproteomics data analysis addresses some of the challenges present with massive public databases. The benefit of restricting the sequence database to only be composed of information that was derived from the same microbiome sample is that the resulting protein sequence database will only contain sequences that could have possibly come from a species in the sample. This increases specificity in database search procedures, but sensitivity to unknown protein sequences will be a function of the initial metagenomic procedure's accuracy and coverage. One can foresee a situation where there is a low-abundance bacterium present in the microbiome sample, it is not picked up during metagenomic sequencing, and its protein expression is either mistaken for a different bacterial species (through sequence alignment and matching protocols) or missed altogether.

Consideration of these factors inspired the development of the LP protocol. Further, defining a performance metric for the LP protocol in the context of these factors lead to development of its validation procedure. Since the only data to work with was the peptide

intensity information produced by the mass spectrometer, proxies for a few features of dynamic microbiomes needed to be characterized. These include the assignment of pre-optimization bacterial abundance measures in an iterative protocol, the introduction of randomness through the ε-factor and $\mathcal{N}(1, \sigma)$ sampling, and comparison of linear programming results to naïve approaches of peptide-species abundance assignment. Results from the validation procedure showed a distinct distribution shift of LP peptide-species distance measures towards $d = 0$ compared to both the naïve-split and naïve-duplicate methods, suggesting that when LP produces an incorrect peptide-species redistribution, it is incorrect *to a lesser degree* than either naïve method. Euclidean distance was used as the distance function for this investigation because it takes into account scalar values of each element in a peptide-species vector, allowing for the incorporation of cases where linear programming has redistributed a peptide's abundance to a species but not to the exact amount specified in the *S* species abundance vector (as opposed to using a binary distance function such as the Hamming distance, which would only take into account the presence or absence of an intensity value for a peptide-species pair).

A workflow that was explored during the earlier parts of the LP protocol's development was using metagenomic data to effectively filter out bacterial species from the overall pre-LP data table, composed of peptides (rows) and species (columns). The pre-LP table was derived through iMetaLab, where it matched each peptide sequence to the set of species that have some protein matching a peptide's subsequence. For non-unique peptides, it must be assumed that many of the peptide-species matches are incorrect: a database sequence match does not mean that the bacterial species is present in the sample itself. If these peptide-species database matches were then combined with species-level metagenomic information about the presence or absence of a specific species in the microbiome sample, another level

of analytical validation supports the existence of that species in the sample. Linear programming could then be performed on a filtered dataset where peptide-species matches are only considered if they also can be backed up by the species being in the metagenomic dataset. This would undoubtedly speed up LP performance and increase sample-specific peptide distribution. Further, metagenomics-informed intensity optimization as described would also allow for more accurate estimation of bacterial species abundance, in the sense that the optimized taxonomic profile (metaproteomics data informed by metagenomics data) could be compared directly against taxonomic abundance measurements from the metagenomics dataset. Combining these two experimental datasets would ultimately increase confidence in calculated taxonomic distributions using bioinformatics.

In 16S sequencing – a metagenomic method that amplifies the 16S rRNA gene using PCR to define sequence similarity and abundance of bacterial taxa in a sample – taxonomic resolution is limited because the PCR primers used must be able to amplify as many bacterial 16S hypervariable regions as possible, while remaining specific enough to capture taxonomic differences [175]. This method was attempted during the thesis project in an effort to obtain species-level taxonomic profiles for microbiome samples, but the data was ultimately deemed unsuitable because there wasn't enough taxonomic resolution for it to be effectively applied as a filter to the pre-LP dataset. Recently, procedures have been developed to address 16S sequencing's taxonomic resolution limitation that could be viable as a parallel species-level definition methods [175–177]. Future analyses using LP approaches should consider parallel metagenomics sequencing of this kind to increase the performance of optimization algorithms addressing the shared-peptide problem.

6 FUTURE DIRECTION

The thesis project used six human gut microbiome samples to investigate the effects of three RS compounds on their function and taxonomy. Through the RapidAIM procedure, a wealth of metaproteomics mass spectrometry data were generated and assessed using a variety of software tools, most notably Cerberus and iMetaLab. While features of functional and taxonomic variability were established using this workflow, statistical significance at the participant and treatment level was hard to achieve in many circumstances, owed to the fact that peptide, protein, and taxonomic data across different microbiomes was so variable and participant specific ("enterotypes"). Further, technical replicate counts for each participants' treatment conditions (e.g. P1's RS2-treated samples) were either three or four; this made comparing treatments against each other much more susceptible to exploding p-values for *t*-tests and other statistical procedures. For further analyses of microbiome responses to RS treatment, it may be prudent to increase technical replicate counts and increase the number of participants and microbiome samples gathered. Compiling a larger metaproteomics dataset will aid not only in statistical power but also may allow for detection of trends at the participant level that can be taken into consideration when all participant data is pooled and analyzed.

Although the same amount of each prebiotic compound (RS and FOS) was exposed to each microbiome, these data show that RS does not appear to influence microbiome taxonomy or function to the same scale as the FOS positive control. Because these data support FOS as a functioning positive control in a microbiome metaproteomics context, increasing the culture concentration of RS compounds may be necessary to resolve signal from noise and elucidate defined functional and taxonomic characteristics of microbiomes responding to these compounds. The FOS compound used was ~98% pure FOS, while the

RS compounds had actual RS contents as low as 23%; increasing the concentration of RS compounds relative to FOS should be considered to ensure treatment consistency.

For the LP procedure, building a parallel metagenomics dataset that has species-level resolution would greatly aid in performing peptide-species intensity redistribution, as removing species that are not present in a parallel metagenomics dataset ensures that only those species that have been verified to be in the sample through a non-protein-dependent protocol are considered. This would increase confidence in LP redistribution results, as the base-level peptide-species relationship table would contain much less spurious associations for the LP procedure to take into account.

7 REFERENCES

1. Mai V, Draganov P V. 2009. Recent advances and remaining gaps in our knowledge of associations between gut microbiota and human health. World J Gastroenterol 15:81–85.

2. Shreiner AB, Kao JY, Young VB. 2015. The gut microbiome in health and in disease. Curr Opin Gastroenterol 31:69–75.

3. Hutkins RW, Krumbeck JA, Bindels LB, Cani PD, Fahey G, Goh YJ, Hamaker B, Martens EC, Mills DA, Rastal RA, Vaughan E, Sanders ME. 2016. Prebiotics: Why definitions matter. Curr Opin Biotechnol 37:1–7.

4. Schley PD, Field CJ. 2002. The immune-enhancing effects of dietary fibres and prebiotics. Br J Nutr 87 Suppl 2:S221–S230.

5. Lohner S, Küllenberg D, Antes G, Decsi T, Meerpohl JJ. 2014. Prebiotics in healthy infants and children for prevention of acute infectious diseases: A systematic review and meta-analysis. Nutr Rev 72:523–531.

6. Liong MT. 2008. Roles of probiotics and prebiotics in colon cancer prevention: Postulated mechanisms and in-vivo evidence. Int J Mol Sci 9:854–863.

7. Raman M, Ambalam P, Kondepudi KK, Pithva S, Kothari C, Patel AT, Purama RK, Dave JM, Vyas BRM. 2013. Potential of probiotics, prebiotics and synbiotics for management of colorectal cancer. Gut Microbes 4:181–192.

8. Hedin C, Whelan K, Lindsay JO. 2007. Evidence for the use of probiotics and prebiotics in inflammatory bowel disease: a review of clinical trials. Proc Nutr Soc 66:307–15.

9. Fuentes-Zaragoza E, Sanchez-Zapata E, Sendra E, Sayas E, Navarro C, Fernandez-Lopez J, Perez-Alvarez JA. 2011. Resistant starch as prebiotic: A review. Starch 63:406–415.

10. Yao N, Paez A V, White PJ. 2009. Structure and function of starch and resistant starch from corn with different doses of mutant Amylose-Extender and Floury-1 alleles. J Agric Food Chem 57:2040–2048.

11. Canani RB, Costanzo M Di, Leone L, Pedata M, Meli R, Calignano A, Canani RB, Costanzo M Di. 2011. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol 17:1519–1528.

12. Birt DF, Boylston T, Hendrich S, Jane J-L, Hollis J, Li L, McClelland J, Moore S, Phillips GJ, Rowling M, Schalinske K, Scott MP, Whitley EM. 2013. Resistant Starch: Promise for Improving Human Health. Adv Nutr An Int Rev J 4:587–601.

13. Sender R, Fuchs S, Milo R. 2016. Are we really vastly outnumbered? Revisiting the

ratio of bacterial to host cells in humans. Cell 164:337–40.

14. Sommer F, Bäckhed F. 2013. The gut microbiota--masters of host development and physiology. Nat Rev Microbiol 11:227–38.

15. Gill SRS, Pop M, DeBoy RTR, Eckburg PBP, Peter J, Samuel BS, Gordon JI, Relman DA, Fraser- CM, Nelson KE. 2006. Metagenomic Analysis of the Human Distal Gut Microbiome. Science (80-) 312:1355–1359.

16. Kim M, Morrison M, Yu Z. 2011. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. J Microbiol Methods 84:81–87.

17. Huse SM, Ye Y, Zhou Y, Fodor AA. 2012. A core human microbiome as viewed through 16S rRNA sequence clusters. PLoS One 7:1–12.

18. Ward D V., Gevers D, Giannoukos G, Earl AM, Meth?? BA, Sodergren E, Feldgarden M, Ciulla DM, Tabbaa D, Arze C, Appelbaum E, Aird L, Anderson S, Ayvaz T, Belter E, Bihan M, Bloom T, Crabtree J, Courtney L, Carmichael L, Dooling D, Erlich RL, Farmer C, Fulton L, Fulton R, Gao H, Gill JA, Haas BJ, Hemphill L, Hall O, Hamilton SG, Hepburn TA, Lennon NJ, Joshi V, Kells C, Kovar CL, Kalra D, Li K, Lewis L, Leonard S, Muzny DM, Mardis E, Mihindukulasuriya K, Magrini V, O'Laughlin M, Pohl C, Qin X, Ross K, Ross MC, Rogers YHA, Singh N, Shang Y, Wilczek-Boney K, Wortman JR, Worley KC, Youmans BP, Yooseph S, Zhou Y, Schloss PD, Wilson R, Gibbs RA, Nelson KE, Weinstock G, DeSantis TZ, Petrosino JF, Highlander SK, Birren BW. 2012. Evaluation of 16s rDNA-based community profiling for human microbiome research. PLoS One 7.

19. Ursell LK, Knight R. 2013. Xenobiotics and the human gut microbiome: Metatranscriptomics reveal the active players. Cell Metab 17:317–318.

20. Jorth P, Turner KH, Gumus P. 2014. Metatranscriptomics of the Human Oral Microbiome during Health. MBio 5:1–10.

21. Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, Lefsrud MG, Apajalahti J, Tysk C, Hettich RL, Jansson JK. 2009. Shotgun metaproteomics of the human distal gut microbiota. ISME J 3:179–89.

22. Kolmeder CA, de Vos WM. 2014. Metaproteomics of our microbiome - Developing insight in function and activity in man and model systems. J Proteomics 97:3–16.

23. Kolmeder CA, Salojärvi J, Ritari J, Been M De, Raes J, Falony G, Vieira-silva S, Kekkonen RA, Corthals GL, Palva A, Salonen A, Vos WM De. 2016. Faecal metaproteomic analysis reveals a personalized and stable functional microbiome and limited effects of a probiotic intervention in adults. PLoS One 1–23.

24. Kolmeder CA, de Been M, Nikkila J, Ritamo I, Matto J, Valmu L, Salojärvi J, Palva A, Salonen A, de Vos WM. 2012. Comparative metaproteomics and diversity analysis

of human intestinal microbiota testifies for its temporal stability and expression of core functions. PLoS One 7.

25. Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, Shah M, Halfvarson J, Tysk C, Henrissat B, Raes J, Verberkmoes NC, Fraser CM, Hettich RL, Jansson JK. 2012. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's Disease. PLoS One 7.

26. Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. 2012. Host-bacterial mutualism in the human intestine. Science (80-) 307:1915–1920.

27. Muth T, Benndorf D, Reichl U, Rapp E, Martens L. 2013. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. Mol Biosyst 9:578–585.

28. Wang S, Copeland L. 2015. Effect of acid hydrolysis on starch structure and functionality: a review. Crit Rev Food Sci Nutr 55:1081–1097.

29. Tetlow IJ. 2011. Starch biosynthesis in developing seeds. Seed Sci Res 21:5–32.

30. Bertoft E. 2013. On the building block and backbone concepts of amylopectin structure. Cereal Chem 90:294–311.

31. Englyst KN, Liu S, Englyst HN. 2007. Nutritional characterization and measurement of dietary carbohydrates. Eur J Clin Nutr 61:19–39.

32. Englyst H, Wiggins HS, Cummings JH. 1982. Determination of the non-starch polysaccharides in plant foods by gas-liquid chromatography of constituent sugars as alditol acetates. Analyst 107:307.

33. Sajilata MG, Singhal RS, Kulkarni PR. 2006. Resistant starch - A review. Compr Rev Food Sci Food Saf 5:1–17.

34. Jiang H, Jane J-L, Acevedo D, Green A, Shinn G, Schrenker D, Srichuwong S, Campbell M, Wu Y. 2010. Variations in starch physicochemical properties from a generation-means analysis study using Amylomaize V and VII parents. J Agric Food Chem 58:5633–5639.

35. Higgins JA, Brown IL. 2013. Resistant starch: a promising dietary agent for the prevention / treatment of inflammatory bowel disease and bowel cancer. Curr Opin Gastroenterol 29:190–194.

36. O'Dea K, Nestel PJ, Antonoff L. 1980. Physical factors influencing postprandial glucose and insulin responses to starch. Am J Clin Nutr 33:760–765.

37. Lockyer S, Nugent AP. 2017. Health effects of resistant starch. Nutr Bull.

38. Granfeldt Y, Björck I. 1991. Glycemic response to starch in pasta: a study of mechanisms of limited enzyme availability. J Cereal Sci 14:47–61.

39. Jenkins DJA, Wesson V, Wolever TM, Jenkins AL, Kalmusky J, Guidici S, Csima A, Josse RG, Wong GS. 1988. Wholemeal versus wholegrain breads: proportion of whole or cracked grain and the glycaemic response. BMJ 297:958–60.

40. Jane J-L, Aol Z, Duvick SA, Wiklund M, Yoo S-H, Wong1 K-S, Gardner C. 2003. Structures of Amylopectin and Starch Granules: How Are They Synthesized? Japanese Soc Appl Glycosci 167:167–172.

41. Li L, Jiang H, Campbell M, Blanco M, Jane J lin. 2008. Characterization of maize amylose-extender (ae) mutant starches. Part I: Relationship between resistant starch contents and molecular structures. Carbohydr Polym 74:396–404.

42. Jiang H, Campbell M, Blanco M, Jane JL. 2010. Characterization of maize amylose-extender (ae) mutant starches: Part II. Structures and properties of starch residues remaining after enzymatic hydrolysis at boiling-water temperature. Carbohydr Polym 80:1–12.

43. Regina A, Bird A, Topping D, Bowden S, Freeman J, Barsby T, Kosar-Hashemi B, Li Z, Rahman S, Morell M. 2006. High-amylose wheat generated by RNA interference improves indices of large-bowel health in rats. Proc Natl Acad Sci 103:3546–3551.

44. Nugent AP. 2005. Health properties of resistant starch. Nutr Bull 30:27–54.

45. Jung M, Jun S, Ick S, Ri M, Joo C, Kim Y, Il W, Wha T. 2008. Resistant glutarate starch from adlay: Preparation and properties. Carbohydr Polym 74:787–796.

46. Zhang B, Huang Q, Luo FX, Fu X, Jiang H, Jane JL. 2011. Effects of octenylsuccinylation on the structure and properties of high-amylose maize starch. Carbohydr Polym 84:1276–1281.

47. Annison G, Illman RJ, Topping DL. 2003. Acetylated, propionylated or butyrylated starches raise large bowel short-chain fatty acids preferentially when fed to rats. J Nutr 133:3523–8.

48. Frohberg C, Quanz M. 2009. Use of linear poly-alpha-1,4-glucans as resistant starch. 20080249297. United States, Germany.

49. Hasjim J, Lee S, Hendrich S, Setiawan S, Ai Y, Jane J. 2010. Molecular Diversity and Health Benefits of Carbohydrates from Cereals and Pulses Characterization of a Novel Resistant-Starch and Its Effects on Postprandial Plasma-Glucose and Insulin Responses. Cereal Chem 87:257–262.

50. Seneviratne HD, Biliaderis CG. 1991. Action of α-amylases on amylose-lipid complex superstructures. J Cereal Sci 13:129–143.

51. Lochocka K, Bajerska J, Glapa A, Fidler-Witon E, Nowak JK, Szczapa T, Grebowiec P, Lisowska A, Walkowiak J. 2015. Green tea extract decreases starch digestion and absorption from a test meal in humans: A randomized, placebo-controlled crossover

study. Sci Rep 5:1–5.

52. Wei C, Xu B, Qin F, Yu H, Chen C, Meng X, Zhu L, Wang Y, Gu M, Liu Q. 2010. C-Type Starch from High-Amylose Rice Resistant Starch Granules Modified by Antisense RNA Inhibition of Starch Branching Enzyme. J Agric Food Chem 58:7383–7388.

53. Carciofi M, Blennow A, Jensen SL, Shaik SS, Henriksen A, Buléon A, Holm PB, Hebelstrup KH. 2012. Concerted suppression of all starch branching enzyme genes in barley produces amylose-only starch granules. BMC Plant Biol 12:1–16.

54. Pinhero RG, Waduge RN, Liu Q, Sullivan JA, Tsao R, Bizimungu B, Yada RY. 2016. Evaluation of nutritional profiles of starch and dry matter from early potato varieties and its estimated glycemic impact. Food Chem 203:356–366.

55. Englyst HN, Kingman SM, Cummings JH. 1992. Classification and Measurement of Nutritionally Important Starch Fractions. Eur J Clin Nutr 46:S33–S50.

56. Muir JG, O'Dea K. 1992. Measurement of resistant starch: Factors affecting the amount of starch escaping digestion in vitro. Am J Clin Nutr 56:123–127.

57. Chung HJ, Liu Q, Hoover R. 2009. Impact of annealing and heat-moisture treatment on rapidly digestible, slowly digestible and resistant starch levels in native and gelatinized corn, pea and lentil starches. Carbohydr Polym 75:436–447.

58. Chung H-J, Liu Q, Huang R, Yin Y, Li A. 2010. Physicochemical Properties and In Vitro Starch Digestibility of Cooked Rice from Commercially Available Cultivars in Canada. Cereal Chem 87:297–304.

59. Rohlfing KA, Pollak LM, White PJ. 2010. Exotic corn lines with increased resistant starch and impact on starch thermal characteristics. Cereal Chem 87:190–193.

60. Singh J, Dartois A, Kaur L. 2010. Starch digestibility in food matrix: a review. Trends Food Sci Technol 21:168–180.

61. Pollak LM, Scott MP, Duvick SA. 2011. Resistant Starch and Starch Thermal Characteristics in Exotic Corn Lines Grown in Temperate and Tropical Environments. Cereal Chem 88:435–440.

62. Linsberger-Martin G, Lukasch B, Berghofer E. 2012. Effects of high hydrostatic pressure on the RS content of amaranth, quinoa and wheat starch. Starch/Staerke 64:157–165.

63. Dundar AN, Gocmen D. 2013. Effects of autoclaving temperature and storing time on resistant starch formation and its functional and physicochemical properties. Carbohydr Polym 97:764–771.

64. Raatz SK, Idso L, Johnson LAK, Jackson MI, Combs GF. 2016. Resistant starch

analysis of commonly consumed potatoes: Content varies by cooking method and service temperature but not by variety. Food Chem 208:297–300.

65. Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F. 2013. The Prokaryotes: Human Microbiology.

66. Topping DL, Clifton PM. 2001. Short-Chain Fatty Acids and Human Colonic Function: Roles of Resistant Starch and Nonstarch Polysaccharides. Physiol Rev 81:1031–1064.

67. Walker AW, Ince J, Duncan SH, Webster LM, Holtrop G, Ze X, Brown D, Stares MD, Scott P, Bergerat A, Louis P, McIntosh F, Johnstone AM, Lobley GE, Parkhill J, Flint HJ. 2011. Dominant and diet-responsive groups of bacteria within the human colonic microbiota. ISME J 5:220–230.

68. Shipman JA, Berleman JE, Salyers AA. 2000. Characterization of four outer membrane proteins involved in binding starch to the cell surface of Bacteroides thetaiotaomicron. J Bacteriol 182:5365–5372.

69. Martens EC, Koropatkin NM, Smith TJ, Gordon JI. 2009. Complex glycan catabolism by the human gut microbiota: The bacteroidetes sus-like paradigm. J Biol Chem 284:24673–24677.

70. Flint HJ, Scott KP, Duncan SH, Louis P, Forano E. 2012. Microbial degradation of complex carbohydrates in the gut. Gut Microbes 3.

71. Flint HJ, Bayer EA, Rincon MT, Lamed R, White BA. 2008. Polysaccharide utilization by gut bacteria: Potential for new insights from genomic analysis. Nat Rev Microbiol 6:121–131.

72. Ze X, David Y Ben, Laverde-Gomez JA, Dassa B, Sheridan PO, Duncan SH, Louis P, Henrissat B, Juge N, Koropatkin NM, Bayer EA, Flint HJ. 2015. Unique organization of extracellular amylases into amylosomes in the resistant starch-utilizing human colonic firmicutes bacterium ruminococcus bromii. MBio 6:1–11.

73. Birt DF, Phillips GJ. 2014. Diet, genes, and microbes: Complexities of colon cancer prevention. Toxicol Pathol 42:182–188.

74. Milani C, Andrea Lugli G, Duranti S, Turroni F, Mancabelli L, Ferrario C, Mangifesta M, Hevia A, Viappiani A, Scholz M, Arioli S, Sanchez B, Lane J, Ward D V., Hickey R, Mora D, Segata N, Margolles A, Van Sinderen D, Ventura M. 2015. Bifidobacteria exhibit social behavior through carbohydrate resource sharing in the gut. Sci Rep 5:1–14.

75. Mahowald MA, Rey FE, Seedorf H, Turnbaugh PJ, Fulton RS, Wollam A, Shah N, Wang C, Magrini V, Wilson RK, Cantarel BL, Coutinho PM, Henrissat B, Crock LW, Russell A, Verberkmoes NC, Hettich RL, Gordon JI. 2009. Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. Proc

Natl Acad Sci 106:5859–5864.

76. Ndeh D, Gilbert HJ. 2018. Biochemistry of complex glycan depolymerisation by the human gut microbiota. FEMS Microbiol Rev 42:146–164.

77. Rakoff-Nahoum S, Foster KR, Comstock LE. 2016. The evolution of cooperation within the gut microbiota. Nature 533:255–259.

78. Erra P, Debeire P, O'Donohue MJ. 1999. The Type II Pullulanase of Thermococcus hydrothermalis: Molecular. J Bacteriol 181:3284.

79. Morrison DJ, Preston T. 2016. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. Gut Microbes 7:189–200.

80. Louis P, Scott KP, Duncan SH, Flint HJ. 2007. Understanding the effects of diet on bacterial metabolism in the large intestine. J Appl Microbiol 102:1197–1208.

81. Koh A, De Vadder F, Kovatcheva-Datchary P, Bäckhed F. 2016. From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. Cell 165:1332–1345.

82. Duncan SH, Hold GL, Barcenilla A, Stewart CS, Flint HJ. 2002. Roseburia intestinalis sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces. Int J Syst Evol Microbiol 52:1615–1620.

83. Reichardt N, Duncan SH, Young P, Belenguer A, McWilliam Leitch C, Scott KP, Flint HJ, Louis P. 2014. Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. ISME J 8:1323–1335.

84. Vital M, Howe AC, Tiedje JM. 2014. Revealing the Bacterial Butyrate Synthesis Pathways by Analyzing (Meta)genomic Data. Am Soc Microbiol 5:e00889-14.

85. Louis P, Duncan SH, Mccrae SI, Millar J, Jackson MS, Flint HJ, Al LET. 2004. Restricted Distribution of the Butyrate Kinase Pathway among Butyrate-Producing Bacteria from the Human Colon. J Bacteriol 186:2099–2106.

86. Louis P, Young P, Holtrop G, Flint HJ. 2010. Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-CoA:acetate CoA-transferase gene. Environ Microbiol 12:304–314.

87. Ze X, Duncan SH, Louis P, Flint HJ. 2012. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. ISME J 1–9.

88. Serikov I, Russell SL, Antunes CM, Finlay BB. 2010. Gut microbiota in health and disease. Physiol Rev 90:859–904.

89. Makki K, Deehan EC, Walter J, Bäckhed F. 2018. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. Cell Host Microbe 23:705–715.

90. Macfarlane GT, Gibson GR, Cummings JH. 1992. Comparison of fermentation reactions in different regions of the human colon. J Appl Bacteriol 72:57–64.

91. Gupta N, Martin PM, Prasad PD, Ganapathy V. 2006. SLC5A8 (SMCT1)-mediated transport of butyrate forms the basis for the tumor suppressive function of the transporter. Life Sci 78:2419–2425.

92. van der Beek CM, Bloemen JG, van den Broek MA, Lenaerts K, Venema K, Buurman WA, Dejong CH. 2015. Hepatic Uptake of Rectally Administered Butyrate Prevents an Increase in Systemic Butyrate Concentrations in Humans. J Nutr 145:2019–2024.

93. Wang HB, Wang PY, Wang X, Wan YL, Liu YC. 2012. Butyrate enhances intestinal epithelial barrier function via up-regulation of tight junction protein claudin-1 transcription. Dig Dis Sci 57:3126–3135.

94. Cani PD, Bibiloni R, Knauf C, Neyrinck AM, Delzenne NM, Burcelin R. 2008. Changes in gut microbiota control metabolic diet–induced obesity and diabetes in mice. Diabetes 57:1470–1481.

95. Liu B, Qian J, Wang Q, Wang F, Zhenyu M, Qiao Y. 2014. Butyrate protects rat liver against total hepatic ischemia reperfusion injury with bowel congestion. PLoS One 9:2–9.

96. Johnstone RW. 2002. Histone-deacetylase inhibitors: novel drugs for the treatment of cancer. Nat Rev Drug Discov 1:287–299.

97. Flint HJ, Scott KP, Louis P, Duncan SH. 2012. The role of the gut microbiota in nutrition and health. Nat Rev Gastroenterol Hepatol 9:577–589.

98. Donohoe DR, Collins LB, Wali A, Bigler R, Sun W, Bultman SJ. 2012. The Warburg Effect Dictates the Mechanism of Butyrate-Mediated Histone Acetylation and Cell Proliferation. Mol Cell 48:612–626.

99. Lupton JR. 2004. Diet Induced Changes in the Colonic Environment and Colorectal Cancer Microbial Degradation Products Influence Colon Cancer Risk : J Nutr 134:479–482.

100. Conlon MA, Kerr CA, McSweeney CS, Dunne RA, Shaw JM, Kang S, Bird AR, Morell MK, Lockett TJ, Molloy PL, Regina A, Toden S, Clarke JM, Topping DL. 2012. Resistant Starches Protect against Colonic DNA Damage and Alter Microbiota and Gene Expression in Rats Fed a Western Diet. J Nutr 142:832–840.

101. Bordonaro M, Mariadason JM, Aslam F, Heerdt BG, Augenlicht LH. 1999. Butyrate-induced apoptotic cascade in colonic carcinoma cells: modulation of the beta-catenin-Tcf pathway and concordance with effects of sulindac and trichostatin A but not curcumin. Cell Growth Differ 10:713–20.

102. Lazarova DL, Bordonaro M, Carbone R, Sartorelli AC. 2004. Linear relationship

between Wnt activity levels and apoptosis in colorectal carcinoma cells exposed to butyrate. Int J Cancer 110:523–531.

103. Cruz-Bravo RK, Guevara-Gonzalez M, Garcia-Gasca T, Campos-Vega R, Oomah BD, Loarca-Pina G. 2011. Fermented nondigestible fraction from common bean (Phaseolus vulgaris L.) cultivar Negro 8025 modulates HT-29 cell behavior. J Food Sci 76:T41–T47.

104. Kendall CW, Esfahani A, Sanders LM, Potter SM, Vidgen E. 2010. The effect of a pre-load meal containing resistant starch on spontaneous food intake and glucose and insulin responses. J Food Technol.

105. Stewart ML, Zimmer JP. 2018. Postprandial glucose and insulin response to a high-fiber muffin top containing resistant starch type 4 in healthy adults: a double-blind, randomized, controlled trial. Nutrition 53:59–63.

106. Alfa MJ, Strang D, Tappia PS, Olson N, DeGagne P, Bray D, Murray B-L, Hiebert B. 2018. A Randomized Placebo Controlled Clinical Trial to Determine the Impact of Digestion Resistant Starch MSPrebiotic® on Glucose, Insulin, and Insulin Resistance in Elderly and Mid-Age Adults. Front Med 4:1–12.

107. Al-Mana NM, Robertson MD. 2018. Acute Effect of Resistant Starch on Food Intake, Appetite and Satiety in Overweight/Obese Males. Nutrients 10:1993.

108. Dodevska MS, Sobajic SS, Djordjevic PB, Dimitrijevic-Sreckovic VS, Spasojevic-Kalimanovska V V., Djordjevic BI. 2016. Effects of total fibre or resistant starch-rich diets within lifestyle intervention in obese prediabetic adults. Eur J Nutr 55:127–137.

109. Zhou ZK, Wang F, Ren XC, Wang Y, Blanchard C. 2015. Resistant starch manipulated hyperglycemia/hyperlipidemia and related genes expression in diabetic rats. Int J Biol Macromol 75:316–321.

110. Peterson CM, Beyl RA, Marlatt KL, Martin CK, Aryana KJ, Marco ML, Martin RJ, Keenan MJ, Ravussin E. 2018. Effect of 12 wk of resistant starch supplementation on cardiometabolic risk factors in adults with prediabetes: A randomized controlled trial. Am J Clin Nutr 108:492–501.

111. Wanders AJ, van den Borne JJ, de Graaf C, Hulshof T, Johnathan MC, Kristensen M, Mars M, Schols HA, Feskens EJ. 2011. Effects of dietary fibre on subjective appetite, energy intake and body weight: A systematic review of randomized controlled trials. Obes Rev 12:724–739.

112. Keenan MJ, Zhou J, McCutcheon KL, Raggio AM, Bateman HG, Todd E, Jones CK, Tulley RT, Melton S, Martin RJ, Hegsted M. 2006. Effects of resistant starch, a non-digestible fermentable fiber, on reducing body fat. Obesity 14:1523–1534.

113. Bodinham CL, Frost GS, Robertson MD. 2010. Acute ingestion of resistant starch reduces food intake in healthy adults. Br J Nutr 103:917–922.

114. Willis HJ, Eldridge AL, Beiseigel J, Thomas W, Slavin JL. 2009. Greater satiety response with resistant starch and corn bran in human subjects. Nutr Res 29:100–105.

115. Hoffmann Sardá FA, Giuntini EB, Gomez MLPA, Lui MCY, Negrini JAE, Tadini CC, Lajolo FM, Menezes EW. 2016. Impact of resistant starch from unripe banana flour on hunger, satiety, and glucose homeostasis in healthy volunteers. J Funct Foods 24:63–74.

116. Gentile CL, Ward E, Holst JJ, Astrup A, Ormsbee MJ, Connelly S, Arciero PJ. 2015. Resistant starch and protein intake enhances fat oxidation and feelings of fullness in lean and overweight/obese women. Nutr J 14:1–10.

117. A.S. K, W. T, J.L. S. 2012. Resistant starch and pullulan reduce postprandial glucose, insulin, and GLP-1, but have no effect on satiety in healthy humans. J Agric Food Chem 60:11928–11934.

118. Higgins JA, Jackman MR, Brown IL, Johnson GC, Steig A, Wyatt HR, Hill JO, Maclean PS. 2011. Resistant starch and exercise independently attenuate weight regain on a high fat diet in a rat model of obesity. Nutr Metab (Lond) 8:49.

119. Regmi PR, Metzler-zebeli BU, Ga MG, Kempen TATG Van. 2011. Starch with high amylose content and low In vitro digestibility increases intestinal nutrient flow and microbial fermentation and selectively promotes Bifidobacteria in pigs. J Nutr 39:1273–1280.

120. Xiang G, Han Z, Yu B, Qi D, Chen H. 2012. Effects of different starch sources on Bacillus spp. in intestinal tract and expression of intestinal development related genes of weanling piglets. Mol Biol Rep 39:1869–1876.

121. Mcorist AL, Miller RB, Bird AR, Keogh JB, Noakes M, Topping DL, Conlon MA. 2011. Fecal butyrate levels vary widely among individuals but are usually increased by a diet high in resistant starch. J Nutr 141:883–889.

122. Yuli E, Purwadaria T, Thenawidjaja M. 2012. Anaerobe fermentation RS3 derived from sago and rice starch with Clostridium butyricum BCC B2571 or Eubacterium rectale DSM 17629. Anaerobe 18:55–61.

123. Wilmes P, Bond PL. 2006. Metaproteomics: Studying functional gene expression in microbial ecosystems. Trends Microbiol 14:92–97.

124. Rodríguez-Valera F. 2004. Environmental genomics, the big picture? FEMS Microbiol Lett 231:153–158.

125. Nesvizhskii AI. Protein Identification by Tandem Mass Spectrometry and Sequence Database Searching. Mass Spectrom Data Anal Proteomics 367:87–120.

126. Cheng K, Ning Z, Zhang X, Mayne J, Figeys D. 2018. Separation and characterization of human microbiomes by metaproteomics. TrAC - Trends Anal Chem 108:221–230.

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

128. Tanca A, Palomba A, Pisanu S, Addis MF, Uzzau S. 2015. Enrichment or depletion? The impact of stool pretreatment on metaproteomic characterization of the human gut microbiota. Proteomics 15:3474–3485.

129. Chen B, Brown KA, Lin Z, Ge Y. 2018. Top-Down Proteomics: Ready for Prime Time? Anal Chem 90:110–127.

130. Chait BT. 2011. Mass Spectrometry in the Postgenomic Era. Annu Rev Biochem 80:239–246.

131. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Cooks RG. 2005. The Orbitrap: A new mass spectrometer. J Mass Spectrom 40:430–443.

132. Senko MW, Remes PM, Canterbury JD, Mathur R, Song Q, Eliuk SM, Mullen C, Earley L, Hardman M, Blethrow JD, Bui H, Specht A, Lange O, Denisov E, Makarov A, Horning S, Zabrouskov V. 2013. Novel parallelized quadrupole/linear ion trap/orbitrap tribrid mass spectrometer improving proteome coverage and peptide identification rates. Anal Chem 85:11710–11714.

133. Good DM, Wirtala M, McAlister GC, Coon JJ. 2007. Performance Characteristics of Electron Transfer Dissociation Mass Spectrometry. Mol Cell Proteomics 6:1942–1951.

134. Tanca A, Palomba A, Fraumene C, Pagnozzi D, Manghina V, Deligios M, Muth T, Rapp E, Martens L, Addis MF, Uzzau S. 2016. The impact of sequence database choice on metaproteomic results in gut microbiota studies. Microbiome 4:51.

135. Muth T, Renard BY, Martens L. 2016. Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. Expert Rev Proteomics 13:757–69.

136. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, Juncker AS, Manichanh C, Chen B, Zhang W, Levenez F, Wang JJJ, Xu X, Xiao L, Liang S, Zhang D, Zhang Z, Chen W, Zhao H, Al-Aama JY, Edris S, Yang H, Wang JJJ, Hansen T, Nielsen HB, Brunak S, Kristiansen K, Guarner F, Pedersen O, Doré J, Ehrlich SD, Bork P, Wang JJJ. 2014. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol 32:834–841.

137. Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, Alpi E, Antunes R, Arganiska J, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Chavali G, Cibrian-Uhalte E, Da Silva A, De Giorgi M, Dogan T, Fazzini F, Gane P, Castro LG, Garmiri P, Hatton-Ellis E, Hieta R, Huntley R, Legge D, Liu W, Luo J, Macdougall A, Mutowo P, Nightingale A, Orchard S, Pichler K, Poggioli D, Pundir S, Pureza L, Qi G, Rosanoff S, Saidi R, Sawford T, Shypitsyna A, Turner E, Volynkin V, Wardell T, Watkins X, Zellner H, Cowley A, Figueira L, Li W, McWilliam H, Lopez R, Xenarios

I, Bougueleret L, Bridge A, Poux S, Redaschi N, Aimo L, Argoud-Puy G, Auchincloss A, Axelsen K, Bansal P, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Boutet E, Breuza L, Casal-Casas C, De Castro E, Coudert E, Cuche B, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gasteiger E, Gehant S, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Jungo F, Keller G, Lara V, Lemercier P, Lieberherr D, Lombardot T, Martin X, Masson P, Morgat A, Neto T, Nouspikel N, Paesano S, Pedruzzi I, Pilbout S, Pozzato M, Pruess M, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Wu CH, Arighi CN, Arminski L, Chen C, Chen Y, Garavelli JS, Huang H, Laiho K, McGarvey P, Natale DA, Suzek BE, Vinayaka CR, Wang Q, Wang Y, Yeh LS, Yerramalla MS, Zhang J. 2015. UniProt: A hub for protein information. Nucleic Acids Res 43:D204–D212.

138. Zhang X, Ning Z, Mayne J, Moore JI, Li J, Butcher J, Deeke SA, Chen R, Chiang CK, Wen M, Mack D, Stintzi A, Figeys D. 2016. MetaPro-IQ: A universal metaproteomic approach to studying human and mouse gut microbiota. Microbiome 4:1–12.

139. Italiana A, Fanaro S, Moro G, Minoli I, Mosca M, Fanaro S, Jelinek J, Stahl B, Boehm G. 2002. Dosage-related bifidogenic effects of galacto- and fructooligosaccharides in formula-fed term infants. J Pediatr Gastroenterol Nutr 34:291–295.

140. Rastall RA, Gibson GR. 2015. Recent developments in prebiotics to selectively impact beneficial microbes and promote intestinal health. Curr Opin Biotechnol 32:42–46.

141. Bouhnik Y, Vahedi K, Achour L, Attar A, Pochart P, Marteau P, Flourie B, Bornet F, Rambaud J. 1999. Human nutrition and metabolism increases fecal Bifidobacteria in healthy humans. J Nutr 129:113–117.

142. Le Leu RK, Brown IL, Hu Y, Bird AR, Jackson M, Esterman A, Young GP. 2005. A synbiotic combination of resistant starch and Bifidobacterium lactis facilitates apoptotic deletion of carcinogen-damaged cells in rat colon. J Nutr 135:996–1001.

143. Ratnayake WS, Jackson DS. 2008. Thermal behavior of resistant starches RS 2, RS 3, and RS 4. J Food Sci 73:356–366.

144. Rycroft CE, Jones MR, Gibson GR, Rastall RA. 2001. A comparative in vitro evaluation of the fermentation properties of prebiotic oligosaccharides. J Appl Microbiol 91:878–887.

145. Rada V, Nevoral J, Trojanova I, Tomankova E, Smehilova M, Killer J. 2008. Anaerobe growth of infant faecal bifidobacteria and clostridia on prebiotic oligosaccharides in in vitro conditions. Anaerobe 14:205–208.

146. Liu Y, Gibson GR, Walton GE. 2016. An in vitro approach to study effects of prebiotics and probiotics on the faecal microbiota and selected immune parameters relevant to the elderly. PLoS One 1–18.

147. AOAC Official Methods of Analysis. 1995. AOAC Official Method 991.43: Total,

Soluble, and Insoluble Dietary Fibre in Foods, p. 7–9. *In* Cereal Foods.

148. Gibson GR, Rastall RA. 2002. Comparison of the in vitro bifidogenic properties of pectins and pectic-oligosaccharides. J Appl Microbiol 505–511.

149. Liao B, Ning Z, Cheng K, Zhang X, Li L, Mayne J, Figeys D. 2018. iMetaLab 1.0: A web platform for metaproteomics data analysis. Bioinformatics 34:3954–3956.

150. Chong J, Xia J. 2018. MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. Bioinformatics 34:4313–4314.

151. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. 2016. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods 13:731–740.

152. Luo W, Brouwer C. 2013. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics 29:1830–1831.

153. Foundation PS. Python Language Reference, version 3.6.

154. Team RC. 2018. R: A Language and Environment for Statistical Computing. 3.5.2. Vienna, Austria.

155. Mckinney W, Pydata Development Team. 2019. pandas : powerful Python data analysis toolkit release 0.24.1. Python Packag.

156. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40:109–114.

157. Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol 428:726–731.

158. Wold S, Sjöström M, Eriksson L. 2001. PLS-regression: A basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130.

159. Ward JT. 1963. Heirarchical Grouping to Optimize an Objective Function. J Am Stat Assoc 58:236–244.

160. Hunter JD. 2007. MATPLOTLIB: A 2D Graphics Environment. Sci Program 90–95.

161. Jones E, Oliphant T, Peterson P, others. 2001. SciPy: Open source scientific tools for Python.

162. He Z, Huang T, Liu X, Zhu P, Teng B, Deng S. 2016. Protein inference: A protein quantification perspective. Comput Biol Chem 63:21–29.

163. Mitchell S, O'Sullivan M, Dunning I. 2011. PuLP: A Linear Programming Toolkit for Python. Dep Eng Sci Univ Auckl.

164. Dixon P. 2003. VEGAN, a package of R functions for community ecology. J Veg Sci 14:927–930.

165. Team scikit-bio D. 2014. scikit-bio. v0.5.5.

166. Baxter NT, Schmidt AW, Venkataraman A, Kim KS, Waldron C, Schmidt TM. 2019. Dynamics of Human Gut Microbiota and Short-Chain Fatty Acids in Response to Dietary Interventions with Three Fermentable Fibers. MBio 10:e02566-18.

167. Geirnaert A, Steyaert A, Eeckhaut V, Debruyne B, Arends JBA, Van Immerseel F, Boon N, Van de Wiele T. 2014. Butyricicoccus pullicaecorum, a butyrate producer with probiotic potential, is intrinsically tolerant to stomach and small intestine conditions. Anaerobe 30:70–74.

168. Mao L, Franke J. 2015. Symbiosis, dysbiosis, and rebiosis-The value of metaproteomics in human microbiome monitoring. Proteomics 15:1142–1151.

169. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480:241–244.

170. Reigstad CS, Kashyap PC. 2013. Beyond phylotyping: Understanding the impact of gut microbiota on host biology. Neurogastroenterol Motil 25:358–372.

171. Louis P, Flint HJ. 2017. Formation of propionate and butyrate by the human colonic microbiota. Environ Microbiol 19:29–41.

172. Bui TPN, Ritari J, Boeren S, De Waard P, Plugge CM, De Vos WM. 2015. Production of butyrate from lysine and the Amadori product fructoselysine by a human gut commensal. Nat Commun 6:1–10.

173. Kanehisa M. 2017. Enzyme Annotation and Metabolic Reconstruction using KEGG, p. 135–145. *In* Methods in Molecular Biology.

174. McDonald AG, Tipton KF. 2014. Fifty-five years of enzyme classification: Advances and difficulties. FEBS J 281:583–592.

175. Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, Soen Y, Shental N. 2018. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. Microbiome 6:1–13.

176. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. 2016. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun 469:967–977.

177. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL, Wong GKS. 2016. Characterization of the gut microbiome using 16S or shotgun metagenomics. Front Microbiol 7:1–17.