

**Extracting detailed metabolic information and connections
from mammalian gut microbiomes via metaproteomics**

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Jose Alfredo Blakeley-Ruiz
December 2020

Copyright © 2020 by Jose Alfredo Blakeley-Ruiz
All rights reserved.

ACKNOWLEDGMENTS

Thank you to the Genome Science and Technology Program for taking me on, especially my adviser Robert L. Hettich. Without him, I may not have discovered the wonderful world of community microbiology. I would like to thank my committee members—Gladys Alexandre for her friendship and mentorship. Mircea Podar, for introducing me to the wonderful world of metagenomics. Steve Abel for teaching me about ODEs, back-of-the-envelope calculations, Markov chains, and Markov models. Richard Giannone for teaching me the nuts and bolts of proteomics and always setting me on the right path.

I would also like to thank my Genome Science and Technology cohort. It's nice to find a group of like-minded people. Thank you to the members of the Hettich lab past and present, especially Weili Xiong, Suresh Poudel, Richard Giannone, and Paul Abraham for their mentorship, Alex Cope for his friendship, and Him Shrestha for tolerating sharing an office with me.

Thank you to the CADES computing cluster facility for making it possible for me to assemble 590 genomes.

Thank you to my parents Jose and Lucinda Blakeley. Their upbringing and constant support put me on the path to choose science as my life. My wonderful wife Jordan Bush for choosing to be my partner in this life. My kids Freddie and Sophia Blakeley. They do not know this yet, but they were born in the middle of this.

DEDICATION

I dedicate this dissertation to my late father, Dr. Jose A Blakeley Perez; without him, I may never have considered pursuing a Ph.D.

ABSTRACT

A diverse community of bacteria populates the mammalian gastrointestinal tract. These populations exist in a balance with the host assisting with critical functions, particularly the metabolism of intractable fibers and immune modulation. Disruption of this balance can lead to diseases such as infection, inflammatory bowel syndrome, and obesity. Common symptoms include chronic pain, chronic inflammation, and altered metabolism. Several taxonomic classifications of bacteria have been associated with these diseases, but recent studies have indicated that these findings are not always statistically valid. An explanation for this is that microbial communities between individuals and even across time can vary substantially even when the individuals have a similar health status. Microbial function, however, is a promising arena to study disease scenarios. Omics methods, which measure the entire gene content of a community, are a particularly powerful set of techniques with which to analyze the potential and active function of microbiome communities. Metaproteomics detects and quantifies proteins directly from environmental samples and can be used to measure gut microbiome functional activity. This dissertation applied the use of LC-MS/MS based metaproteomics and metagenomic sequencing to study gut microbiome function in adult humans with Crohn's disease, preterm infants with necrotizing enterocolitis, and obese and morphine treated mice. Substantial variability across time and individuals was observed at the discrete protein sequence level; however, specific functions such as reactions and metabolic modules were shown to be more conserved. Fully connected metabolic networks and pathways were reconstructed from these metaproteomes, and specific metabolic reactions are shown to be affected by necrotizing colitis, diet-induced obesity, and morphine. This dissertation makes a major step forward by showing discrete metabolic reactions can be effectively analyzed using metaproteomic data.

TABLE OF CONTENTS

1	Overview of the microbiome with an emphasis on gut microbiome function.....	1
1.1	Overview of the human microbiome	1
1.2	Composition of the gut microbiome	2
1.3	Interface between the microbiome and the host	5
1.4	Methods for looking at gut microbiome function	7
1.5	Overview of dissertation	9
2	Technical considerations for using metaproteomics to explore gut microbiome function ...	11
2.1	Introduction.....	11
2.2	Preparation of samples for detection by mass spectrometry	11
2.2.1	Considerations for samples selection.....	12
2.2.2	Isolation of protein from environmental samples	13
2.3	Detection of proteins by mass spectrometry in environmental samples.....	14
2.3.1	Instrument considerations for detection of proteins using mass spectrometry	16
2.3.2	Algorithms for confident detection of peptides by tandem mass spectra	18
2.3.3	Creation of a protein database for database searching.....	20
2.4	Assigning function and taxonomy of origin to proteins detected in environmental samples.....	26
2.4.1	Assignment of taxonomy based on database	26
2.4.2	Assignment of function to detected proteins	28
2.4.3	Analyzing function in environmental samples:	29
2.5	Conclusions.....	30
3	Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes	31
3.1	Introduction.....	31
3.2	Methods.....	32
3.2.1	Patient Cohort	32
	Community DNA preparation.....	33
3.2.2	Shotgun metagenomic sequencing and assembly	33
3.2.3	Metaproteomics sample collection	35
3.2.4	LC-MS/MS	36
3.2.5	Informatics and quantification	36
3.2.6	Data analysis	36
3.3	Results.....	37
3.3.1	Sequence-guided sample selection and metagenome assembly	37
3.3.2	Individual gut metaproteomes reveal substantial proteome variability	38
3.3.3	Description of the Seven Microbial Proteins Identified in All Samples.....	42
3.3.4	Assignment of KEGG Orthologous groups reveals remarkable functional redundancy among protein groups.....	45
3.3.5	Inferred metabolic modules demonstrate persistence of metabolic function in a dynamic microbiome	46
3.3.6	Observation of acetylglucosamine and acetylneuraminate indicate the microbiome can feed itself at the expense of the host.....	50
3.3.7	Observation of short-chain fatty acid associated modules indicate that the host is fed by a changing set of pathways and bacteria.....	53

3.4	Discussion	56
4	Metabolic reactions inferred by metaproteomics show connected metabolic networks and functional differences between infants that develop necrotizing enterocolitis and those that do not	58
4.1	Introduction.....	58
4.2	Methods.....	60
4.2.1	Data acquisition	60
4.2.2	Protein database construction and annotation.....	60
4.2.3	Protein inference	61
4.2.4	Construction of graph database.....	61
4.2.5	Data Analysis	62
4.3	Results.....	63
4.3.1	Reactions are substantially more conserved between samples than protein sequences	63
4.3.2	KEGG annotation of detected proteins enables the generation of connected metabolic networks	66
4.3.3	Network analysis reveals numerous host-microbe connections	66
4.3.4	Linear mixed-effects models reveal reactions associated with the development of Necrotizing Enterocolitis	70
4.3.5	Pathway reconstruction reveals differences between NEC and non-NEC	73
4.4	Discussion	75
5	Morphine administration differentially alters gut microbiome composition and function in lean and obese mice	79
5.1	Introduction.....	79
5.2	Methods.....	80
5.2.1	Study Design and Sample Collection	80
5.2.2	Protein Extraction	81
5.2.3	DNA Extraction	81
5.2.4	16S rRNA Amplicon Sequencing and Analysis	82
5.2.5	Whole Genome Metagenomic Sequencing and Analysis.....	83
5.2.6	Metaproteomic Measurement	84
5.2.7	Protein identification.....	84
5.2.8	Richness and principle component comparison between conditions and omics methods	85
5.2.9	Evaluating newness of representative genome's gene calls	85
5.2.10	Differential protein analysis.....	86
5.2.11	Functional analysis.....	86
5.2.12	Cytokine Assays and Analysis.....	87
5.2.13	Urinary Metabolites	87
5.3	Results.....	88
5.3.1	Comprehensive microbiome characterization reveals a mouse gut microbiome membership that is underrepresented in public repositories.....	90
5.3.2	Morphine alters the gut microbiome composition but is eclipsed by obesity effects	94
5.3.3	Differential metaproteomics reveals that taxa-level protein abundances were specific to sample conditions.	94

5.3.4	Both diet and morphine alter microbiome metabolic functions in specific ways.	98
5.4	Discussion	107
6	Conclusions and perspectives on metaproteomics research	112
6.1	Getting a deeper metaproteome	114
6.2	Modeling gut microbiome metabolism.....	116
6.3	Final thoughts.....	117
7	References.....	119
Vita.....		164

LIST OF FIGURES

Figure 2.1: Workflow for taking DNA and protein from a fecal sample and processing it in the mass spectrometer or sequencer.....	12
Figure 2.2: Fragmentation by CID yields b and y ions, which can be used to sequence a peptide. (A) Locations at which b or y ions can be fragmented and their numbering scheme from the N to C terminus. (B) An example MS2 scan where amino acids are inferred based on the difference in mass between y or b ions respectively.....	15
Figure 2.3: Workflow for the inference of proteins from peptides based on database searching. Proteins can then be used to infer metabolic function based on their annotations.	20
Figure 2.4: Percent amino acid identity between protein databases and their top hits in UniParc. Protein databases where derived from metagenomes that where part of a mouse, switchgrass, and human preterm infant study.....	22
Figure 2.5: Typical workflow for the construction of protein databases from genome resolved metagenomics	23
Figure 2.6: Example of refining a metagenome assembeled genome in Anvi'o. Contigs are clusered by tetranucleotide frequency and outlier contigs are removed. This genome was adjusted from a completion score 97.5% and a contamination score of 18.38% to a completion score of 96.79% and a contamination of 0%.	24
Figure 2.7: Number of distinct peptides detected using different protein database formulations. All the metaproteomes from a single preterm infant were searched using different formulations of the infants metagenome: just the sample specfic metagenome, all the metagenomes from the infant, all the metagenomes that had an associated proteome, and all the metagenomes clustered at 100% identity. Sample specific had a signifcantly lower number of peptides detected ($P < 0.01$).	25
Figure 3.1: Number of proteins detected in Chron's subjects. Box plots of (A) the number of microbial (blue), human (orange), and total (grey) protein groups per sample and (B) the number of spectral counts assigned to each of those groups per sample.....	38
Figure 3.2: Technical reproducibility of Chron's samples from several perspectives. (A) Distribution of Spearman Correlation Index of raw spectra, Pearson Correlation Index of Log2 transformed spectra, and Pearson Correlation Index of raw spectra for each sample. (B) Percentage of the total protein groups identified in each sample that were found in both technical duplicates. (C) Linear fit of raw spectra identified in both technical replicates of sample P92_d177. Sample P92_d177 had the highest technical reproducibility per Spearman correlation index. (D) Linear fit of raw spectra identified in both technical replicates of P104_d259. P104_d259 had the lowest spearman correlation.....	39
Figure 3.3: Variation in the protein groups identified across time and individuals, and how that variation decreases with functional groups. (A) Shows the number of microbial protein groups that were observed across all time points, more than one time point, and only one time point for each individual. (B and C) Show the same thing but for KEGG orthologous groups and GoMixer modules respectively. (D) Highlights the percentage of microbial protein groups, KEGG orthologs, and GoMixer modules that are found across all individuals, more than one individual, and only one individual.....	41
Figure 3.4: Heatmap illustrating the hierarchical clustering of the spearman correlation coefficient between mass spectrometry runs.	42
Figure 3.5: Abundance across time of the seven protein groups that were consistently observed across all samples.....	43

Figure 3.6: Example InterPro Scan result for 7 protein groups found across all Crohn's disease study samples. Specifically, InterPro Scan Result for ProteinGroup_104293 representative sequence. Sequence submitted 2/27/2018. Revealed domains IPR020828 and IPR020829 which are the NAD(P) binding and catalytic domains respectively for Glyceraldehyde 3-phosphate dehydrogenase.	44
Figure 3.7: Metabolic map of all the GOMixer modules rendered in the GOMixer web application. Highlighted lines represent all the modules that were inferred with peptide evidence in at least one sample. Line thickness represents how many samples the module was observed in with the thickest lines representing modules that were represented across all samples, the medium thickest lines being represented by modules inferred in all individuals, and the least thick lines representing modules not seen in all individuals. Red colored lines represent modules that were observed in Firmicutes, Actinobacteria, Bacteroidetes, and Proteobacteria. Green lines represent modules that were observed in at least two of the phyla. Black lines represent modules only represented in Firmicutes. Yellow lines represent modules only observed in Bacteroidetes. Purple lines represent modules only represented in Proteobacteria. Two modules that were unique to other phyla were MF0027 inferred in Euryarcheota and MF0010 inferred in Fusobacteria. These were colored beige and orange, respectively.	47
Figure 3.8: Trends by phylum in modules pertaining to degradation of food products that were consistently observed across all samples. MF0029 and MF0030 represent different pathways that degrade Threonine the only amino acid degradation pathway that was seen across all samples. MF0048 represents Lactose degradation. MF0062 represents starch degradation a critical polysaccharide food source that comes from plants. MF0108 represents glycerol degradation, the only lipid degradation pathway that is observed across all samples. MF0124 represent Fucose degradation which is a mucosal glycan that the host produces and that the bacteria in the gut microbiome consume. were observed in two or more phyla, but not always in the same sample.	49
Figure 3.9: Detailed pathway describing modules MF0003 and MF0005, which together degrade acetylglucosamine and acetylneuraminate into pyruvate, fructose-6-phosphate, and acetate. All the KEGG ortholog terms in this figure had peptide evidence in our data set. K00656, while not a part of these modules, had proteomic evidence in every sample of this database, and was the method for converting pyruvate to Acetyl-CoA that there was the most protein evidence for.	51
Figure 3.10: Representation of direct routes to short chain fatty acids and acetyl-CoA with proteomic evidence. All KEGG Ortholog terms mentioned in this figure had proteomic evidence in at least one sample, although many had evidence across all individuals and most samples. For example, K00656 had proteomic evidence in all samples and in Firmicutes, Actinobacteria, Proteobacteria, and Bacteroidetes, but not always at the same time. These enzymes form a sizable representation of the pathways to these metabolites with proteomic evidence; however, not all enzymes that can produce these molecules are included.	55
Figure 4.1: Structure of baby-sample-protein-reaction-compound graph database.	62
Figure 4.2: Observed genes are radically different across samples, but reactions are more conserved.	64
Figure 4.3: Reaction to compound metabolic networks reveal primarily a single connected component. (a) Representative compound to reaction graph for a single sample, S_3_25.1. (b) Reaction to reaction graph of the same sample after a bipartite projection. (c) Compound	

to compound graph of same sample after a bipartite projection. (d) Categorical dot plot depicting the number of nodes in the largest component divided by the total number of nodes for both the compound and reaction graphs. Red nodes represent reactions and blue nodes represent compounds.	68
Figure 4.4: A substantial number of reactions are shared between host and microbiome. (a) Reaction to reaction graph for S_3.25.1 showing microbial (purple), shared (orange), and host (green) reactions. (b) Venn diagram showing number of reactions that are microbial, shared, or host across all samples.	69
Figure 4.5: Ratio of host to microbial reactions changes across time. Regression plots showing a variety of metrics across time shaded region is 95% confidence interval as predicted by jtools. (a) Ratio of host reactions relative to total reactions across time. (b) Ratio of microbial reactions relative to total reactions across time. (c) Number of reactions unique to a genera/relative to total reactions across time. (d) Number of reactions unique to a genera, when only considering microbial or shared reactions. Using all the reaction to reaction networks it was calculated how many edges were microbe to microbe or host to host. (e) Number of host to host edges relative to total number of edges with a host reaction. (f) Number of host to microbe edges relative to the total number of edges with a host reaction. (g) Number of microbe to microbe edges relative to total edges with a microbial reactions. (h) Number of microbe to host edges relative to total number of edges with microbial reactions.	71
Figure 4.6: Reactions significantly different between NEC and non-NEC. (a) $-\log_{10}(\text{p-values})$ versus $\log_2(\text{fold-change})$ for reaction abundance calculated via summed NSAF. (b) $-\log_{10}(\text{p-values})$ versus $\log_2(\text{fold-change})$ for reaction abundance calculated via relative number of proteins.	72
Figure 4.7: Reaction to compound network made up of all the reactions that are significantly different between NEC and non-NEC using summed NSAF abundance. Green nodes are up in infants that do not develop NEC and orange nodes are up in infants that develop NEC. Square blue nodes are compounds, circle nodes are microbial reactions, diamond nodes are shared reactions, and triangle nodes are host reactions.	73
Figure 4.8: Compound to reaction network of reactions that are significantly different in abundance using relative numbers of proteins. Green nodes are up in infants that do not develop NEC and orange nodes are up in infants that develop NEC. Square blue nodes are compounds, circle nodes are microbial reactions, diamond nodes are shared reactions, and triangle nodes are host reactions. In addition, pathway from chorismate to enterochelin, a siderophore, is depicted.	74
Figure 5.1: Experimental design for mouse study. (A) Flow chart of starting hypotheses of bidirectional interaction between pain, obesity, inflammation, opioid use and the gut microbiome. (B) Side by side comparison of db mice vs B6 Mice. (C) Table describing experimental setup. (D) Boxplot showing weight in grams of mice. P-Values calculated by TukeyHSD in R. (E) Boxplot depicting morphine quantities in mouse urine Analyzed by triple quadrupole mass spectrometer. Images are from the Jackson Laboratory website.	89
Figure 5.2: Cytokine analysis reveals different markers of inflammation for different types of obesity and morphine.	90
Figure 5.3: Alpha diversity metrics for samples collected directly from colon versus in a metabolic cage.	91

Figure 5.4: Comprehensive profiling of the mouse gut microbiome reveals a microbiome underrepresented in public repositories. (A-C) Boxplots depicting the number of ASV's, quantifiable genomes, and proteins detected in each sample respectively. (D) pie chart depicting the percentage of ASV's classified using the Silva database down to the species level, unclassified at the species level, and then unclassified at the genera level. (E) pie chart depicting the percentage of genomes classified down to the species, genus, family, class, or order level. All genomes were classified at the phylum level. (F) boxplot depicting amino acid percent identify match of each protein to environmental NR, mouse catalog, NR, and UniParc databases. (G) boxplot of nucleotide percent identity match of each ASV to the NCBI 16S rRNA database.	93
Figure 5.5: Diet induced obesity alters microbiome composition more than genetic obesity or morphine treatment. Principle component analysis of log transformed mean centered counts derived from 16S rRNA read coverage, genome coverage, and metaproteome derived mass spectra assigned to genomes.	95
Figure 5.6: Medium quality genomes cluster by experimental group using hierarchical clustering. (A) Scatter plot of % contamination versus % completion of each genomes as calculated by checkM. Redlines depict a 75% completion and 5% contamination score. 83% of genomes are above and below that threshold respectively. (B) Heatmap of normalized inner-quantile of mean coverage per genomes as calculated by Anvi'o. Dendrograms calculated by ward clustering.....	96
Figure 5.7: Beta-diversity analysis of mouse 16S data. (A-D) PCoA plots of bray-curtis, unweighted UniFrac, weighted UniFrac, and Jaccard respectively. (E) Ward clustered heatmap of all the ASV's collected from colon feces samples.....	97
Figure 5.8: Differential protein abundance analysis reveals specific Lachnospiraceae species that favor B6 saline over B6 morphine and different kinds of obesity. (A) Bubble plot of species with at least 15 or more significant proteins in a comparisons between B6 saline and other experimental conditions. (B-E) Volcano plots of each binary comparison between B6 saline and another experimental condition. Dotted black lines depict a $\log_2$ fold change of 0.5 or greater, and Benjamini Hochberg procedure $Q < 0.1$ threshold. P-values were calculated by student T-test only comparing proteins that had at least 3 in one condition. Colors are species specific.....	99
Figure 5.9: Species enriched for up-regulated proteins align with number proteins and genome sequence coverage of representative genomes. Boxplots depicting the number of proteins and inner-quantile mean coverage for representative species from Figure 4.	100
Figure 5.10: Random forest classifaction of 16S data show high abundance of Lachnospiraceae to be the best predictor of B6 saline treated mice.	101
Figure 5.11: Counting proteins into functional categories reveals functional shifts between morphine versus saline and B6/db versus DIO.Boxplots depicting representative broad functional categories, KEGG pathways and modules, and specific KEGG terms that are significantly different between experimental conditions after the first filtration. At the top the broad COG category G represents the relative proportion of proteins that belong to that category. For the KEGG orthologs K00965 and K02382 these are the centered-log-ratio of tabulated proteins after filtering out all KO terms that did not have 4 proteins in at least one samples. The remaining boxplots represent KEGG pathways or modules, which also had their proteins tabulated and normalized by centered-log-ratio after filtering any pathways or	

modules that did not have at least 4 proteins in at least one sample. All p-values are from TukeyHSD after two-way ANOVA.....	103
Figure 5.12: Counting proteins into KEGG orthologs reveals diet fiber degrading enzymes are more present in db and B6 than DIO. Number of quantified proteins per sampler were counted for each KEGG ortholog group. Kegg orthologs that had less 7 proteins max in a sample were filtered out. The remaining KEGG orthologs were normalized by center log ratio. All P-values are calculated by TukeyHSD after two-way ANOVA. All enzymes pass a Benjamini Hochberg procedure $Q < 0.25$. Starting at the top principle component analysis reveals that DIO has a distinct functional profile from db and B6, while B6 and db are overall indistinguishable. Boxplots depict specific KEGG Orthologs that degrade starch, xylan, arabinogalactan, or pectin and have significantly more normalized proteins in db and B6 than DIO.	104
Figure 5.13: Counting number of proteins in KEGG orthologs reveals a complete pathway from mucus to D-Fructose 6-phosphate that is more abundant in DIO than B6 or db. (A) Boxplots supporting the notion that high fat drives the microbiome towards starvation stress and degradation of mucin. (B) Proposed pathway from mucin to fructose-6-phosphate. (C) Specific KEGG orthologs supporting the pathway. Here KEGG orthologs were filtered out if they did not have 7 proteins in at least one sample. Then they were normalized by centered-log-ration. Finally p-values were calculated by TukeyHSD following two-way ANOVA.	106

1 OVERVIEW OF THE MICROBIOME WITH AN EMPHASIS ON GUT MICROBIOME FUNCTION

Text adapted from:

Blakeley-Ruiz, J.A., Erickson, A.R., Cantarel, B.L., Xiong, W., Adams, R., Jansson, J.K., Fraser, C.M., Hettich, R.L. Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes. *Microbiome* 7, 18 (2019). <https://doi.org/10.1186/s40168-019-0631-8>

Blakeley-Ruiz, J. A., McClintock, C. S., Lydic, R., Baghdoyan, H. A., Choo, J. J., & Hettich, R. L. Combining integrated systems-biology approaches with intervention-based experimental design provides a higher-resolution path forward for microbiome research. *Behavioral and Brain Sciences*, 42 (2019). DOI:10.1017/S0140525X18002911

J. Alfredo Blakeley-Ruiz's contributions include data analysis, literature review, writing of the original manuscript, revision, and response to reviewers.

1.1 Overview of the human microbiome

Text adapted from Blakeley-Ruiz et al. 2019. Microbiome 7, 18.

The human body hosts a dynamic ecosystem of microbial organisms that form an integral part of the overall health maintenance of the host (Clemente et al. 2012). This microbiota comprises several similar but distinct niches spread out across every surface and cavity of the body, including the skin, nose, mouth, genital tract, and gut, where they perform similar but distinct functions (The Human Microbiome Project 2012). Initial estimates suggested an approximate 10 to 1 ratio of bacteria to human cells in the human body using back-of-the-envelope calculations; however, since then, a more conservative estimate of 1.3 bacterial cells for every human somatic cell has been calculated (Sender et al. 2016, Gilbert et al. 2018). The microbial population of any one person can contain more than 1000 distinct bacterial species, expanding the genetic repertoire of the human body from approximately 20,000 genes to well over a million (Locey and Lennon 2016). Many of these genes, however, represent redundant functions. A microbiome that has a healthy interaction with the host, termed “eubiosis,” has been credited with assisting the host with critical functional roles, including expanded metabolism, pathogen defense, immune development, and immune modulation (Koropatkin et al. 2012, den Besten et al. 2013, Smith et al. 2013). A breakdown of this healthy interaction, termed “dysbiosis,” correlates with unhealthy phenotypes such as obesity, allergies, chronic pain, and inflammation (Bisgaard et al. 2011, Ridaura et al. 2013, Rea et al. 2017).

The largest microbial population in terms of biomass is found in the gastrointestinal tract (Rosner 2014). The gut microbiota plays a critical role in assisting the host with the metabolism of indigestible food products and immune modulation (den Besten et al. 2013, Sun et al. 2017). Fermentation products produced by the gut microbiota, including short-chain fatty acids, lie at the intersection of human host-microbiota interactions. These fermentation products, particularly butyrate, propionate, and acetate, represent major sources of energy for the host especially in colonocytes (McNeil 1984, Donohoe et al. 2011), and play a role in host immune modulation, health, and disease (Vinolo et al. 2011, Lin et al. 2012, Smith et al. 2013, Iraporda et al. 2015).

Distinct microbial niches populate the different regions of the gastrointestinal tract (Suchodolski et al. 2005). Starting in the small intestine, the diversity of the microbiome increases as it progresses down the intestinal tract with the most diverse population present in the distal colon. Although distinct microbes populate different regions of the gastrointestinal tract, fecal samples capture a representative sampling of the entire gastrointestinal tract (Eckburg et al. 2005).

1.2 Composition of the gut microbiome

In total, the species that make up the human microbiome represent phyla from all domains of life, including archaea, fungi, bacteria, and viruses. Bacteria remain the most well studied because 16S rRNA and whole-genome sequencing methods favor bacteria (Goodrich et al. 2014); however, use of the internal transcribed spacer region (ITS), as well as advances in metagenomic analysis methods (Pride et al. 2012, Schoch et al. 2012, West et al. 2018), promises more representation of fungi and viruses in future studies. Among bacteria, the majority of species in the microbiome come from the phyla Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria, with Firmicutes and Bacteroidetes making up the highest proportion. Early and more recent studies have suggested that a higher proportion of Firmicutes relative to Bacteroidetes may be indicative of several microbiome-related diseases; however, more recent meta-analyses challenge these findings (Sze and Schloss 2016). A variety of host and environmental factors affect the composition of an individual's gut microbiome and its similarity to others.

At birth, microorganisms rapidly colonize the gut within the first few days of life. Early in life, the microbiome can change substantially in composition over time, sometimes having a

more different composition across time than between individuals (Brown et al. 2018). Over time, the microbiome proceeds through several successions increasing in diversity until it plateaus at around 2-3 years of life, becoming more similar to an adult gut microbiome (Yatsunenko et al. 2012, Yassour et al. 2016). During this colonization period, the microbiome modulates the immune system, making it accustomed to commensal bacteria. For example, the diversity of the microbiome during this colonization period modulates the levels of inflammatory cytokines (Cahenzli et al. 2013). A variety of factors can affect this development, including birth mode, antibiotic use, and diet. A vaginal birth microbiome tends to end up similar to the mother's microbiome, while the caesarian tends to start more similar to the skin microbiome and ends up distinct from a vaginal delivery microbiome (Dunn et al. 2017, Wampach et al. 2017). Breast milk tends to favor *Bifidobacteria* and *Lactobacillus*. In contrast, formula tends to favor *Bacteroides* and *Clostridium* (Arrieta et al. 2014), and transitions between colonization successions appear to be associated with the introduction of solids (Fallani et al. 2011). Antibiotic use, in particular, delays the colonization of certain bacteria like *Bacteroidetes* and correlates with asthma and autoimmune diseases later in life (Arrieta et al. 2014).

After the colonization period, the gut microbiome becomes more personalized, with individual gut microbiota being more similar across time than between individuals. It appears, however, that some individual gut microbiomes change more over time in composition than others (Flores et al. 2014). How much the microbiome changes may also be disease-related (Halfvarson et al. 2017). The diversity of the gut microbiome decreases in old age, while the proportion of opportunistic bacteria increases (Nagpal et al. 2018). After the colonization period, the gut microbiome also becomes similar to the mother's microbiome, suggesting a genetic component to the composition of the microbiota. Monozygotic twins have a significantly more similar microbiome (Goodrich et al. 2014). Several studies used twins to account for this genetic variation; however, disease state has been shown to have a greater effect on the microbiome (Erickson et al. 2012, Ridaura et al. 2013). Much of this genetic effect could be driven by co-habitation. Non-related partners have significantly similar microbiomes to each other (Dill-McFarland et al. 2019), and the addition of animals appears to increase this similarity. Families with dogs have more similar microbiomes than those that do not (Song et al. 2013). Geography appears to have an impact on the composition of the gut microbiome. People from South America have distinct gut microbiomes from those in North America (Yatsunenko et al. 2012),

and rural Indians have distinct microbiomes from urban Indians (Das et al. 2018). Differences in diet between the communities may drive much of this geography-based difference. Diet plays a significant role in the composition of the gut microbiome. Westernized diets substantially alter the gut microbiome (Bisanz et al. 2019), and a transition from an animal-based diet to a vegetarian diet can alter the gut microbiome within a single day (David et al. 2013).

Many diseases interrelate with the composition of the gut microbiome. A gnotobiotic mouse model showed that an obese microbiome introduced to a lean mouse could trigger an obese phenotype; however, obesity is a bit more complex than a singular entity (Ridaura et al. 2013). Diet-induced obesity alters the gut microbiome distinctly and more significantly than genetically induced obesity, implying that though an obese microbiome can cause an obese phenotype, the underlying factors that cause obesity can also alter the microbiome (Pfalzer et al. 2015). Chronic gastrointestinal diseases such as Crohn's disease, irritable bowel syndrome, and inflammatory bowel syndrome correlate with distinctly altered microbiomes as well (Gilbert et al. 2018). Twins discordant for these diseases have been shown to have as different microbiomes from each other as non-twins, indicating that the effect of these diseases on the microbiome is more potent than any genetic effect and that factors other than genetics also play a role in the development of these diseases (Erickson et al. 2012). Crohn's disease in particular is associated with intolerance of vegetables (Ballegaard et al. 1997). This intolerance implies a disfunction in microbial metabolic function since vegetables are typically associated with a healthy microbiota.

Though the most abundant phyla appear to be consistent at lower taxa, the variation across time and individuals can be quite substantial. The set of species present in the gut microbiome can be completely different between two individuals, and core species within an individual can be as low as 5% (Gilbert et al. 2018). In one of the early microbiome studies, 70% of the species detected were unique to a single individual (Ley et al. 2006). Microbial composition tends to be more consistent within an individual over time than across individuals, suggesting a personalized microbiome. Despite this personalization, the microbiome can be more subject to variation over time in some individuals versus others, and at least one study found this increased variance to be associated with disease (Halfvarson et al. 2017). Perturbations to the gut microbiome can lead to significant changes in composition within a single day, and more prolonged perturbations can have long term effects. With this much variation, it can be challenging to trace diseases associated with the microbiome to a specific set of taxa. The

functionality of the microbiome, however, is more consistent and may be a more tractable space in which to study microbiome associated effects.

1.3 Interface between the microbiome and the host

The gut microbiome plays critical roles in metabolism, immune modulation, and pathogen defense. Metabolically its primary role involves degrading dietary products into products that can be used by the host for energy. Typically the human genome has the enzymes needed to degrade starch, lactose, and sucrose but cannot degrade more complex plant fibers like pectins, xylans, and hemicelluloses (Koropatkin et al. 2012). In the gut, particularly in the colon, the microbiome degrades these glycans into sugars and ferments them via the typical anaerobic pathways. These pathways yield a variety of products, with short-chain fatty acids being the most well studied (den Besten et al. 2013). The most abundant short-chain fatty acids in the gut are acetate, propionate, and butyrate, with acetate being the most abundant. Approximately 95% of short-chain fatty acids produced by bacteria are rapidly taken up by the host (Topping and Clifton 2001). These short-chain fatty acids account for the majority of the energy needed by colonocytes (Donohoe et al. 2011). Overall these short-chain fatty acids have been suggested to account for 10% of the host's energy in western diets, with the hypothesis that this could be higher in more plant-based diets (McNeil 1984). Diet can impact this process. Westernized diets that are high in fat at the expense of fiber decrease the amount of short-chain fatty acids produced (Jakobsdottir et al. 2013). Interestingly, low in fiber diets diet can cause certain bacteria, like *Bacteroides thetaiotaomicron*, to transition away from digesting diet fibers and towards digesting the intestinal mucosal interface, which may be a significant factor in the disruption of the mucosal layer leading to disease (Desai et al. 2016). Production of short-chain fatty acids is mutliphyletic, and the extent to which different bacteria perform this function remains to be fully understood (Parada Venegas et al. 2019).

A mucosal layer separates the epithelial lining of the intestine from the gut microbiome and other potentially harmful substances. In the colon, there are two mucosal layers, an impenetrable layer that protects the epithelial cells and a second layer that is populated by specific bacteria (Johansson et al. 2013). The gut microbiome interacts with the host at this interface. Glycoproteins called mucins, produced by goblet cells, make up the bulk of the mucosal layer (Johansson 2012). Mucins form a gel-like layer that permits beneficial molecules

like sugars to be taken up by the host but prevents harmful substances, such as antigens and microbes, from interfacing with the epithelial cells.

Glycans made up of several different sugars, including acetylgalactosamine, acetylglucosamine, acetylneuraminate, fucose, and galactose, glycosylate mucins at serine and threonine residues (Holmen Larsson et al. 2013, Marcobal et al. 2013). Several bacteria can degrade these glycans for energy (Berry et al. 2013). Specifically, *Bacteroides thetaiaetamicon* and *Akkermansia muciniphila* have been well-documented to degrade mucin; however, it seems *Akkermansia muciniphila* uses mucin as its primary form of energy (Van Herreweghen et al. 2017), while *Bacteroides thetaiaetamicon* is more opportunistic with the ability to degrade both dietary products and mucin (Sonnenburg et al. 2005). Other bacteria have been shown to forage mucin as well (Berry et al. 2013); however, many of the mechanisms by which the microbiome harvests mucin and the extent to which different microbes do this remains poorly understood. Proper formation of the mucosal layer requires the presence of the microbiome. Germ-free mice have been shown to have a more permeable mucosal layer, which can be recovered after long term colonization by the microbiome (Johansson et al. 2015). Specific bacteria have been shown to be associated with the production of mucin. *Akkermansia muciniphila* correlates with an increase in sulfated mucins and negatively correlates with ulcerative colitis (Earley et al. 2019). Short-chain fatty acids produced by bacteria, particularly butyrate and propionate, have been shown to increase mucin production (Burger-van Paassen et al. 2009). Thus, it may be beneficial for certain bacteria to feed on the mucosal layer but harmful for others.

Many of the primary gut microbiome diseases exhibit disruption of the mucosal layer as a symptom, including obesity, ulcerative colitis, Crohn's disease, and necrotizing enterocolitis (NEC) (Dorofeyev et al. 2013, Jiang et al. 2013, Gulhane et al. 2016). Disruption of the mucosal layer correlates with increased gut permeability and inflammation, which represent the hallmarks of these diseases (Poret et al. 2018, Ravisankar et al. 2018). Short-chain fatty acids and likely other bacteria-derived metabolites are associated with inflammation suppression, thus creating a link between microbial metabolism, the formation of the mucosal layer, and chronic inflammation (Parada Venegas et al. 2019). Increasing understanding of the details behind microbial metabolism is critical for ameliorating these diseases.

1.4 Methods for looking at gut microbiome function

Text adapted from Blakeley-Ruiz et al. 2019 Behavioral and Brain Sciences, 42.

Deriving a more detailed understanding of how the metabolic activities of microbial communities affect host physiology requires an expanded experimental design. Many studies have demonstrated that microbial community composition varies across individuals, but clusters as a function of age, geography, diet, cohabitation, and health (The Human Microbiome Project 2012, Yatsunenکو et al. 2012, David et al. 2013, Song et al. 2013, Halfvarson et al. 2017). This implies that all these factors have an impact on the host's microbial composition, and that replicating the composition of the microbiome without exact replication of experimental conditions (e.g., individuals, laboratories, diet) is complicated. This is a formidable problem even with large sample sizes. This limitation suggests that there is less of a reproducibility crisis (Sze and Schloss 2016), and more of a need to design more elaborate and appropriate experiments (Knight et al. 2018).

There is evidence that despite significant variation in microbial composition, broad biological function appears to remain similar (The Human Microbiome Project 2012). Thus, microbiome research will be greatly propelled by experimental designs that combine intervention studies with systems-biology techniques that seek to characterize overall microbiome function. Currently, four “omics” techniques capture “detailed” molecular-level information about a microbial community: whole community genome (metagenome) sequencing (The Human Microbiome Project 2012), metatranscriptome sequencing (Jorth et al. 2014), metaproteome identification (Erickson et al. 2012), and metabolome characterization (Karu et al. 2018). Each of these techniques has advantages and limitations. However, a combination of these techniques should lead to a high-resolution portrayal of microbial function and potentially microbial metabolic reconstruction at the community level.

Whole metagenomic sequencing provides a backbone for investigating microbiome function via the creation of a composite gene-catalog. Initially, reference sequence databases were used to characterize the taxonomy and function of a microbial community (The Human Microbiome Project 2012, Stewart et al. 2018), but were limited by the range of known reference genomes, and functional ortholog databases. Advancements in the understanding of genomic features allow combining tetranucleotide frequency with sequence coverage to bin the results from metagenomic assemblies into metagenome assembled genomes (MAGs) *de-novo* (Kang et

al. 2019). This approach provides a set of low to high-quality draft genomes, which can provide subspecies information about the gene repertoire of a microbial community (Brown et al. 2018). Although metagenomic sequencing allows categorizing a set of genes into their respective taxonomy and predicting their function, metagenomics is limited to only predicting potential function because it contributes no information about the actual translation of genes into proteins.

Although the laboratory technique behind metatranscriptomics is more similar to metagenomic sequencing, both metatranscriptomics and metaproteomics explore a similar question: which predicted genes are being transcribed and subsequently translated into proteins? Transcriptome sequencing identifies genes that are being transcribed into RNA and generally produces more total identifications, while LC-MS/MS based metaproteomics ultimately identifies proteins, thus providing confirmation that a metagenomics predicted gene is translated into protein (Erickson et al. 2012, Jorth et al. 2014). Although the number of identified proteins tends to be fewer than transcripts, it is unclear how many proteins microbes utilize in a community. Strides have been made by integrating MAGs with metaproteomes in lower diversity communities, but much work remains to answer this question (Xiong et al. 2017). Theoretically, metaproteomics should provide a better connection to phenotype since functional products of gene expression are measured.

Gas (GC) and liquid (LC) chromatography paired with mass spectrometry provide an approach for detecting metabolites within microbial communities (Karu et al. 2018). Although these molecules are difficult to connect to their organism of origin without complementary gene information, metabolomics provides a framework to evaluate metabolic hypotheses derived from genetic or protein information. Enzyme-to-compound databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2016) and MetaCyc (Caspi et al. 2018), can provide annotation of recognized genes with an enzyme identifier, thereby connecting specific taxa, genes, and proteins to specific molecular functions at the compound level. Ultimately, these features can create a framework by which metabolomic data can be integrated with microbiome data.

Using constrained experimental conditions combined with specific hypothesis-driven intervention, researchers can connect the power of cause-and-effect experimental design with integrated “omics” to characterize the microbiome under different interventions. Time-course studies would be ideal, as the sample subject can function as its own internal control. Eventually,

as the community of investigators transitions into the next phase of microbiome research, general intervention-based studies and broad characterization of microbial composition at the population level will no longer be sufficient to provide novel insights into the gut microbiome. Holistic metabolic modeling through systems-biology measurements of individual microbiomes combined with intervention-based experimental design provides a path forward.

1.5 Overview of dissertation

This dissertation seeks to advance microbiome research by exploring the use of metaproteomics to look at specific metabolic functions. The most specific function possible is the individual reaction an enzyme catalyzes. Databases such as MetaCyc and KEGG have a wealth of data associating protein sequences to the reactions they catalyze (Kanehisa et al. 2016, Caspi et al. 2018). By annotating genes in a metagenome with this data, the potential metabolic functions of a microbial community can be explored, while retaining a link to the genome of origin. Metagenomics, however, only measures the predicted genes of a community with no knowledge about whether that gene is being transcribed or translated into protein or even if it can be translated into protein. Metaproteomics adds another layer of information by detecting the most abundant genes that have been translated into proteins. In this dissertation, it is shown that although individual proteins can be highly varied between conditions, the specific functions that they code for are much more conserved and can be used to explore shifts in host and microbial metabolic function due to perturbation. This dissertation is divided into six chapters, including this introductory chapter. The chapters include one methodological chapter, three experimental chapters, and a concluding chapter.

Chapter 2 provides an overview of many of the methodological considerations needed to detect proteins in environmental samples. This chapter covers all the steps from preparing samples to detecting proteins using a mass spectrometer. Some specific environmental considerations are included, such as the preparation of protein databases, and some recommendations for assigning taxonomy and function to proteins are mentioned at the end.

Chapter 3 explores the metaproteomes of five Crohn's disease patients after resection surgery. Each patient had five samples collected throughout a single year. The purpose of this study was to explore the stability of the metaproteome across time and individuals. This included determining how similar metaproteomes were across time and individuals, whether the same was

true for the functions inferred from the metaproteomes, and how redundant those functions were across taxa.

Chapter 4 represents a natural progression from chapter 3 by using samples taken from preterm infants. In this study, it was explored whether the specific reactions that annotate the metaproteome were more conserved across time and samples than the proteins themselves and whether these reactions could be connected by the compounds they catalyze to build fully connected metabolic networks. These reactions could be used to identify functional differences between infants that developed necrotizing enterocolitis and those that did not.

The goal of Chapter 5 was to see if morphine differentially alters the microbiome of mice with different kinds of obesity from a taxonomical and functional perspective. In this chapter, the taxonomical composition of the gut microbiome was analyzed by 16S rRNA sequencing and genome resolved metagenomics, and the function was analyzed using metaproteomics to infer detailed metabolic functions as described in Chapter 4.

2 TECHNICAL CONSIDERATIONS FOR USING METAPROTEOMICS TO EXPLORE GUT MICROBIOME FUNCTION

2.1 Introduction

The large-scale detection of the complete suite of proteins from environmental samples, termed metaproteomics, is an extension of mass spectrometry-based proteomics where proteins are extracted directly from an environmental sample instead of a microbial isolate or multicellular tissue sample (Wilmes and Bond 2004). In proteomics, proteins isolated from a sample are digested into peptides, separated via liquid chromatography, measured by mass spectrometry, and detected by matching resultant mass spectra to theoretical spectra derived from a protein sequence database (Figure 2.1). Ultimately, the goal of a proteomics experiment is to maximize the number of quantifiable proteins detectable with reasonable confidence and minimal ambiguity. In metaproteomics, there is the additional challenge of inferring the function and taxonomy of detected proteins. This chapter will discuss the main elements needed to detect proteins and infer their function in gut microbiomes. Many of the items in this chapter can be generalized to the analysis of other environmental samples, and as such most of the chapter is relatively general in terms of sample origin. This chapter also contains some preliminary analyses that were useful in defining the methods for the later chapters but were not included there.

2.2 Preparation of samples for detection by mass spectrometry

The main purpose of analyzing an environmental sample by metaproteomics is to characterize the composition and/or function of a specific type of microbial community, like a gut or soil microbiome. The problem is that even though a microbial community is likely to have a typical composition in terms of phyla, the composition of a microbial community can change across habitats, space within a habitat, and time. For example, in the gut >70% of the microbial species can be different between individuals (Ley et al. 2006), the composition of the gut microbiome can alter within a single day (David et al. 2013), and the community at the start of the colon can be distinct from the community of at the end of the colon (Eckburg et al. 2005).

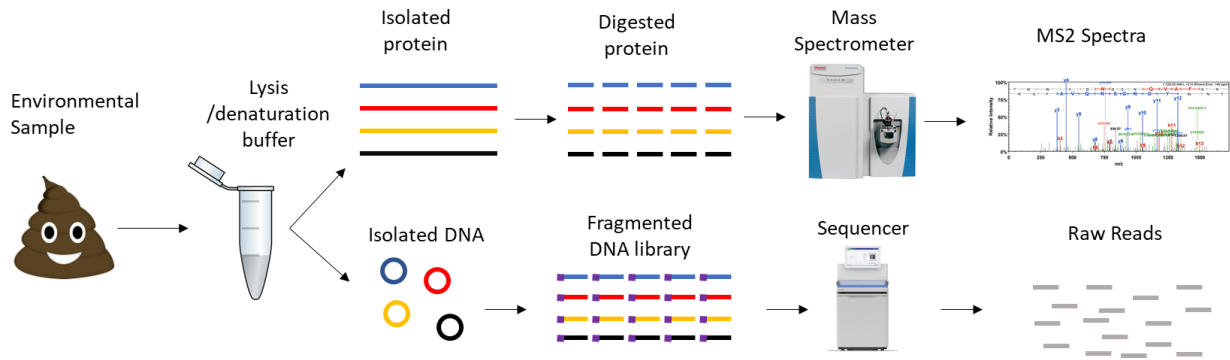


Figure 2.1: Workflow for taking DNA and protein from a fecal sample and processing it in the mass spectrometer or sequencer

This implies that it is impossible to know the composition of a specific gut microbiome *a priori* and that specific care needs to be taken in the selection of samples before the extraction of proteins.

2.2.1 Considerations for samples selection

Eckburg et al. showed that the composition of the microbiome in the colon changes spatially, increasing in diversity from beginning to end; however, they also found that fecal material represented the best overall average of the gut microbiome (Eckburg et al. 2005). This implies that for studying the whole gut microbiome, fecal material is the best sample; however, if the study question is more localized, a more targeted sampling is needed. For example, Zhang et al. used mucosal luminal interface aspirates to compare three separate parts of the colon, both with and without inflammation, providing a better understanding of the localized effects of inflammatory bowel disease (Zhang et al. 2018). In this dissertation, the goal was to look at the overall gut microbiome, so fecal samples were used.

An additional consideration is how the sample is collected. The gut microbiome exists in an anaerobic environment (Flint et al. 2012); however, as soon as fecal material leaves the body, it becomes exposed to an aerobic environment. Since some of the bacteria in the gut are facultative anaerobes while others are strict anaerobes, this creates some concern that the microbial composition of the sample will change the more time passes between a bowel movement and the processing of the sample. Some studies have shown that for population profiling, the community does not change much within the first 24 hours (Tedjo et al. 2015); however, what may be true for population profiling may not be correct for functional analysis

using proteins. Exposure to an aerobic environment automatically alters the metabolic processes available, and the impact of the host on the microbiome is no longer retained. One solution to this is to inactivate the microbiome using a chemical solution such as ethanol (Song et al. 2016). This approach is particularly popular for field studies, where access to freezers is limited (Tung et al. 2015). In the lab and the clinical setting, freezing is more common. For adult humans, the typical procedure is self-collection at home, followed by immediate freezing and shipment to the collection center where the samples are placed at -80°C within 24 hours. This was done in chapter 3. With preterm infants, the infants are in intensive care, so collection by professionals and freezing shortly after is possible. This was done in chapter 4. With mice, samples are typically collected in metabolic cages over 24 hours, this raises some concern that the samples will have been exposed to an aerobic environment for too long. In the case of chapter 5, fecal material was extracted surgically directly from the colon.

2.2.2 Isolation of protein from environmental samples

For the most part, the isolation of proteins from environmental samples is the same as it is from an isolate or a tissue sample. Before this, however, certain types of samples need special attention. For example, preterm infants often have a higher than usual load of human protein in the fecal samples. In this case, to get a better representation of microbial proteins, the samples can be size filtered to enrich the bacterial cells at the expense of the larger Eukaryotic cells (Xiong et al. 2015). In soil, specific procedures need to be followed for the removal of humic acid (Qian and Hettich 2017). After these steps, protein isolation can proceed as usual. These kinds of steps are not needed for adult fecal samples. There are four main steps to prepare samples for metaproteomics analysis: cellular lysis, isolation of protein, digestion, and cleanup. This process is similar to the extraction of DNA except that for DNA, many of these steps have been made routine by commercial kits. For metaproteomics, it is more typical for researchers to use their own in-house protocol.

For lysis, samples are usually treated initially with chemical disruption using the detergent sodium dodecyl sulfate (SDS) in tris-HCl lysis buffer at pH 8.0. SDS has the additional benefit of denaturing the proteins and moving the fecal samples into solution. Next, samples are treated with physical disruption either in the form of bead beating or sonication. Sonication has been reported to achieve more efficient lysis; however, bead beating lyses cells sufficiently well, is

compatible with DNA extraction, and can enable multiplexed processing of more samples at the same time (Zhang et al. 2018). Finally, samples are heated to complete extraction and denaturation by boiling for a specified period. These three steps should lyse cells sufficiently to proceed to protein isolation; however, before that, the solution is typically treated with dithiothreitol (DTT) to break disulfide bonds between cysteine residues. Iodoacetamide (IAA) is then used to modify the cysteine residues so that the bond cannot be reformed.

After these steps, protein is isolated by precipitating it out of solution. This can be done by adding trichloroacetic acid (TCA) or a solution made of chloroform, methanol, and water. Removing the supernatant isolates the protein, which can be reconstituted in a proteolysis buffer, such as an ammonium bicarbonate solution with sodium deoxycholate (SDC) as the denaturant. Once reconstituted, protein concentration can be measured by bicinchoninic acid (BCA) protein assay or a NanoDrop instrument set an UV absorbance of 205 nm or 280 nm. This concentration is then used for determining the trypsin concentration needed for digestion into peptides. Trypsin is the most common protease used for protein digestion in proteomics because it consistently cleaves proteins at the lysine and arginine residues, narrowing the set of possible peptides. Once the peptides have been digested, the samples are cleaned up and desalted. Size filtration removes under-digested protein. SDC is removed using acid and ethyl acetate solution, which precipitates the detergent allowing the rest of the solution to be extracted. Typically, at this step, formic acid is added to protonate the peptides. The final peptide solution is quantified so that a consistent amount of peptides can be added to the liquid chromatography system coupled to the mass spectrometer.

2.3 Detection of proteins by mass spectrometry in environmental samples

Mass spectrometers are made up of three basic components: a source that converts molecules into ions and delivers them into the mass spectrometer, a mass analyzer which separates ions by mass-charge-ratio (m/z), and a detector which records the m/z of the ions along with their relative abundance (intensity). Using these components, a mass spectrometer measures the m/z and intensity of ions and plots these values on a histogram (mass spectrum) with m/z along the x-axis and intensity (relative or absolute) along the y-axis. A simple m/z is not sufficient information with which to distinguish peptide sequences. Thus, peptides are usually further interrogated with an extension of this concept called tandem mass spectrometry

(MS/MS). In MS/MS, ions are isolated in a mass analyzer by their initial m/z (MS1) and then fragmented, producing a second mass spectrum of the fragmented ions (MS2).

Low-energy collision induced dissociation (CID) is an ion fragmentation method that can be conducted in a linear-ion-trap (LTQ) mass analyzer. CID reliably fragments peptides at the carbon to nitrogen bonds yielding fragmented ions traditionally termed b and y ions (Roepstorff and Fohlman 1984, Hunt et al. 1986). These ions yield a mass spectrum that can be used to partially sequence the isolated peptide (Figure 2.2). Other fragmentation methods yield different sets of ions; however, CID is the most used for proteomics (Kim and Pandey 2012). Fragment ions are numbered sequentially from smallest to largest with the b fragment ions numbered from the N to the C-terminus, while the y ions are numbered from C to the N-terminus of a peptide sequence. These ions are numbered from smallest to largest, and amino acids can be predicted by subtracting a b or y ions by the preceding b or y ion, respectively (Figure 2.2) (Hunt et al. 1986).

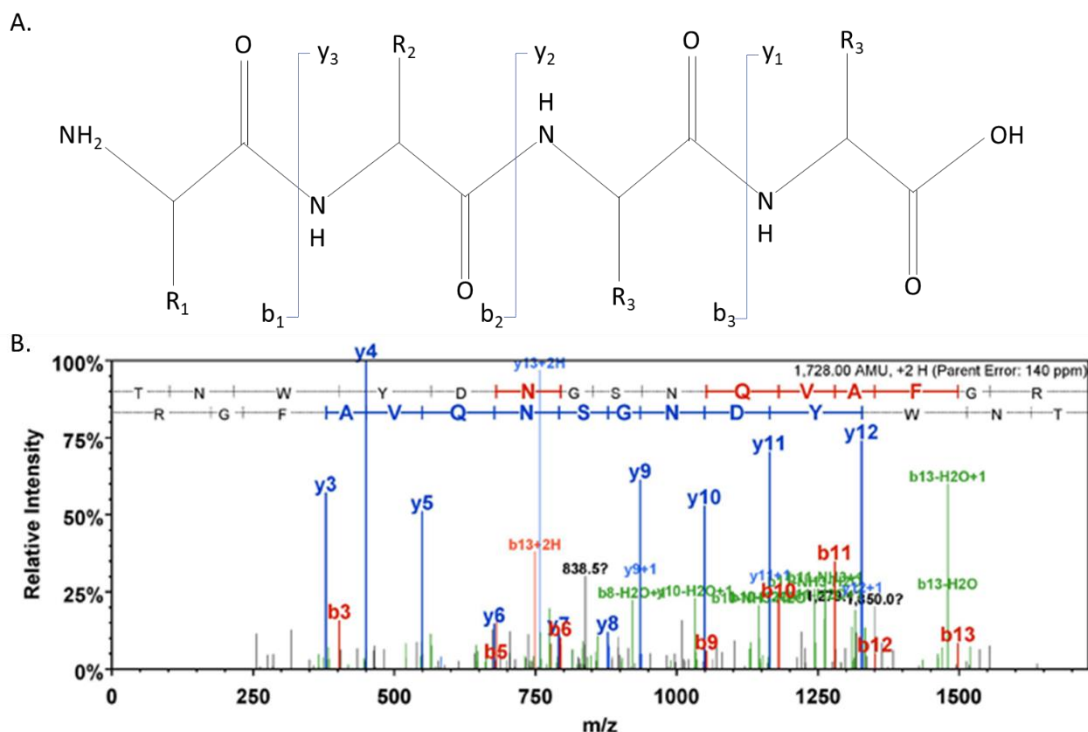


Figure 2.2: Fragmentation by CID yields b and y ions, which can be used to sequence a peptide. (A) Locations at which b or y ions can be fragmented and their numbering scheme from the N to C terminus. (B) An example MS2 scan where amino acids are inferred based on the difference in mass between y or b ions respectively.

2.3.1 Instrument considerations for detection of proteins using mass spectrometry

Due to limitations in instrumentation and informatics, not all ions can be isolated, fragmented, and detected at once. As such, a scheme for selecting ions for fragmentation across time is necessary. For most proteomics measurements, ions are selected for fragmentation in a data dependent manner, called data dependent acquisition (DDA) (Courchesne et al. 1998). First, a full MS1 scan is recorded, and then ions are selected for fragmentation scans in order of abundance from highest to lowest intensity. Typically, between 5-20 ions are sequentially selected for fragmentation between MS1 measurements. Ions selected for fragmentation are added to an exclusion list for a defined period so that the same abundant ions are not continuously selected for fragmentation (Wang and Li 2008). An alternative method called data independent acquisition exists (DIA) (Chapman et al. 2014). DIA fragments all the ions within a given m/z window, providing the opportunity for more extensive fragmentation of ions; however, the bioinformatics has not caught up to dealing with the heterogeneity of DIA derived MS2 spectra in environmental samples (Saito et al. 2019).

The number of high-quality spectra generated by DDA can be increased by improvements in the speed, fragmentation efficiency, mass resolution, and mass accuracy of the instrument (Olsen et al. 2007, Mann and Kelleher 2008). Speed allows for more ions to be measured in a narrower time frame. Fragmentation efficiency allows for better fragmentation of precursor ions leading to a richer MS2 spectrum and allowing for better sequencing of the peptide by b and y ions. Mass resolution determines the narrowness of peaks, calculated as the full-width-at-half-maximum. The resolution allows ions of similar m/z to be better distinguished, allowing for improved calculation of charge, better isolation of ions, and more accurate quantification. Mass accuracy determines how close the measured m/z is to the real m/z , typically calculated in parts-per-million (ppm). Mass accuracy narrows the window of possible peptides that can match to a spectrum making it easier to infer a peptide sequence from the spectrum. Improvements in speed typically come at the cost of resolution and accuracy.

To address these performance issues, hybrid instruments have been developed over the last two decades to utilize the strengths of different mass analyzers. One of the first of these instruments to revolutionize the field of proteomics was the LTQ-Orbitrap (Makarov et al. 2006). At its most basic, this instrument contains a linear ion trap mass analyzer for the rapid isolation and fragmentation of ions and an orbitrap mass analyzer for high accuracy and high-resolution

measurement of ions. This instrument can be run in a high-high mode where both precursor ions and fragment ions are measured in the orbitrap while the ions are fragmented in the ion trap, but this comes at the cost of speed. Alternatively, there is a high-low mode where the precursor is measured in the orbitrap. However, the fragment ions are fragmented and measured more quickly at lower resolution and accuracy in the ion trap. The high-low mode allows for the best tradeoff between resolution, accuracy, and speed for maximizing the number of detected peptides. This instrument was used in chapters 3 and 4 of this dissertation.

An improvement on this model is the Q-Exactive line of instruments (Michalski et al. 2011). This line of instruments replaces the linear ion trap with a quadrupole and uses an HCD cell for fragmentation. In this instrument, the quadrupole isolates ions by m/z , while the HCD cell fragments the ions in succession. Ions are then measured simultaneously in the orbitrap. This allows for high-high measurements at a faster speed than the high-low measurements of the LTQ-orbitrap. The HCD also fragments ions more efficiently than the CID implemented in the LTQ, allowing for better peptide sequencing in the MS2 (Olsen et al. 2007). This process leads to faster measurements in general, but also a higher proportion of high-quality measurements both of which lead to fewer measurements needed in total and shorter overall measurement times for the same number of peptides. This instrument was used in Chapter 5 of this dissertation.

Further development into hybrid instruments has led to tribrid lines of instruments, which add a linear ion trap to the instrument. The additional mass analyzer adds more options for fragmentation, allowing for MS3 and beyond tandem measurements. This type of fragmentation is particularly important for quantification by isobaric labeling (Ting et al. 2011). As all quantification in this dissertation was label-free, this line of instruments was not needed for the dissertation.

For DDA to work, the mass spectrum needs a continuous source of ions. Soft ionization sources, which allow for the ionization of intact peptides without fragmentation, were developed in the 1980s. Matrix assisted laser desorption/ionization (MALDI) ionizes peptides by using laser energy to convert peptides into ions from a matrix (Karas et al. 1985). The main advantage of this approach is that it can ionize large molecules. Its disadvantage is that it is not easy to maintain a constant flow of ions. The original metaproteomics experiment used MALDI to ionize peptides located on electrophoresis gel spots using a quadrupole time-of-flight mass spectrometer and was able to detect several hundred peptides (Wilmes and Bond 2004). An

alternative method of ionization, called electrospray, was created around the same time as MALDI (Whitehouse et al. 1985). In this method, protonated peptides in solution flow through an emitter with a high voltage applied to it. When the ions reach the tip of the emitter, they aerosolize spraying into a heated capillary for desolvation, after which ions are guided into a mass analyzer using electrostatic fields. The advantage of this approach is that it allows for the continuous flow of ions into the mass spectrometer and can be coupled directly to a liquid chromatography system for the online separation of peptides (Washburn et al. 2001).

An unseparated mixture of peptides will simply deliver peptides continuously at the same time to the mass spectrometer, leading to the detection of only the most abundant peptides. To increase peptide measurement depth, peptides need to be delivered to the mass spectrometer separated over time (Peng and Gygi 2001). The most popular way of doing this is by reverse-phase liquid chromatography. In reverse-phase, the emitter/analytical column is packed with a stationary phase made up of C18 hydrophobic beads that capture peptides as they flow through the column. Then a solution gradient (mobile phase) from low hydrophobicity (typically water) to high hydrophobicity (typically acetonitrile) is applied to the column. As the solution increases in hydrophobicity, peptides are released from the stationary phase in order of their hydrophobicity. The more gradual the gradient, the more time the measurement takes, the better the separation between peptides, and the more peptides can be detected. This separation is limited by the concentration of peptides delivered to the column, as too much separation relative to the concentration of peptides can lead to a decrease in signal. One solution for this is to add an orthogonal chromatography step before reverse-phase, called 2D liquid chromatography (Washburn et al. 2001, Hinzke et al. 2019). In this dissertation, the additional chromatography step had a strong cation exchange stationary phase. Peptides attach to the stationary phase by charge and are released by increasing concentrations of salt, such as ammonium acetate. Peptides are then released on to the analytical reverse-phase column stepwise using an increasing concentration of salt. The first metaproteomics study to measure over 1,000 proteins in a human gut microbiome was conducted using this 2D-LC-MS/MS method (Verberkmoes et al. 2008).

2.3.2 Algorithms for confident detection of peptides by tandem mass spectra

Unfortunately, CID does not yield a complete spectrum, thus making it difficult to predict the full sequence of a peptide-based on this information alone. Two main algorithms have been

developed to infer peptide sequences based on their CID fragmentation pattern. In the sequence tag approach, a series of sequential fragment ions can be used to create a partial sequence tag (Mann and Wilm 1994). This tag can be used in conjunction with the m/z of the precursor ion to predict possibilities for the remaining amino acids, and then a sequence database can be used to identify possible matching sequences. The alternative method, called database searching, predicts a series of theoretical peptide spectra based on a database made up of protein sequences likely to be in the sample (Eng et al. 1994). Then real spectra are matched to these theoretical spectra, and the peptide spectral matches (PSMs) are scored based on the closeness of the match. Modified peptides can also be detected by database searching (Yates et al. 1995). To detect modifications, mass shifts specific to modifications of amino acids can be added to the theoretical spectra list. Two of the most common modifications included in a database search are oxidation of methionine and carbamidomethylation of cysteine. Carbamidomethylation of cysteine is due to the addition of IAA during sample preparation. Database searching was initially implemented in a software package called Sequest, but since then, a variety of algorithms have been developed which follow the same principle with different scoring methods (Tabb 2015). For a peptide to be detected by the database search method, it must be present in the protein database; however, the sequence tag approach can provide at least a partial sequence forming the basis for *de novo* peptide sequencing. The database searching approach maintains more traction in the proteomics community due to its ability to reliably detect more peptides with high confidence.

To ensure that only PSMs with highly confident scores are used to infer proteins, a PSM false discovery rate (FDR) can be calculated, $FDR = FP/(TP+FP)$ where TP represents true positives and FP represents false positives (Elias and Gygi 2010). False positives can be predicted by engaging in a target-decoy competition by creating a decoy protein database of similar size and structure to the target protein database (Elias and Gygi 2007). The decoy database, often generated by “reversing” all of the normal sequences, is then included in the database search with the target database. Based on the results, a score threshold that yields the desired FDR can be used for the identification of peptides. The accuracy of FDR prediction can be improved using a semi-supervised machine learning approaches as implemented in Percolator (Käll et al. 2007). The advantage of the semi-supervised approach is that multiple scoring schemes can be implemented together in a test set to produce a false discovery rate

prediction, which becomes more accurate the more PSMs are included in the analysis. False discovery rates are often set at 1% for high confidence and 5% for medium confidence.

High or medium confidence peptides can be used to infer identified proteins. One way to do this is to use the two peptide rule to provide confidence, and the unique peptide rule to distinguish proteins that share peptides (Erickson et al. 2012); however, innovations in protein false discovery rate calculation allow for the combination of multiple factors including uniqueness, number of peptides, and PSM score in the inference of identified proteins (Figure 2.3). In this case, proteins with only a single peptide detected, but with high confidence, can be inferred at a reasonable false discovery rate (Gupta and Pevzner 2009, Serang et al. 2010). In the case of Eukaryotic organisms, the unique peptide rule is sometimes not reasonable, as multiple proteoforms make it difficult to resolve certain proteins with a unique peptide. In this case, proteins that share peptides are grouped into a protein group, called a master protein, and inferred together. In metaproteomes, however, the need to distinguish detected proteins by organism outweighs the benefit of protein groups, and as such, the unique peptide rule is critical.

2.3.3 *Creation of a protein database for database searching*

The most crucial consideration for detecting proteins in an environmental sample is the protein database. This is most clearly exemplified by a study that showed that using different protein databases on the same measurement could have a profound impact on the taxonomic composition of the sample inferred from the metaproteome (Timmins-Schiffman et al. 2017). This study implied that using a generic protein database for database searching has the potential for generating erroneous results. Database searching cannot detect peptides that are not in the database, leading to missed PSMs if the protein database is incomplete. FDR calculation relies

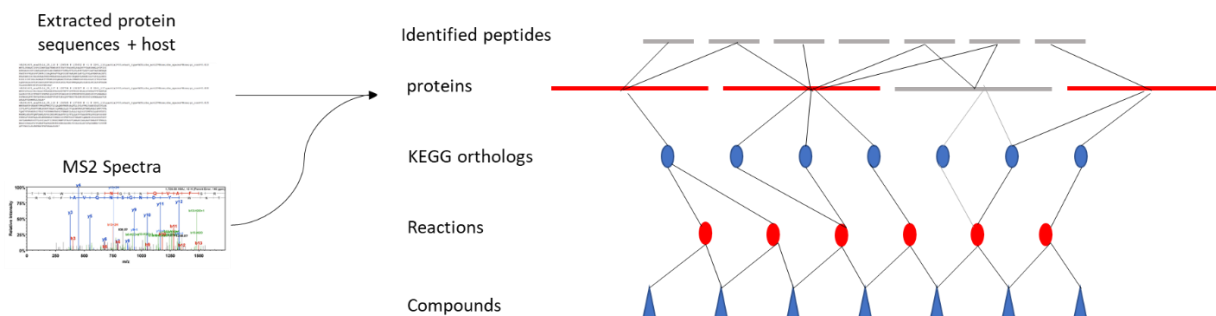


Figure 2.3: Workflow for the inference of proteins from peptides based on database searching. Proteins can then be used to infer metabolic function based on their annotations.

on a significant separation in score between true positives and false positives. The less representative the database is of the sample, the lower the separation in score between true and false positives. In a typical proteomics study, this is not that great of a problem. If the genome of the organism has been sequenced, a mostly complete protein database can be created for the proteomic analysis. Even if the organism is not sequenced, the genomes of closely related organisms have been used to some success (Waridel et al. 2007). In the case of environmental samples, however, it is impossible to know the genetic composition of the sample in advance because prior sampling of the same community type does not imply the same gene content.

A solution to this problem that is gaining substantial traction in the metaproteomics community is the use of system-specific gene catalogs. Due to the HMP and Metahit projects, human microbiomes have been heavily sequenced, and extensive repositories of their gene content have been created (Qin et al. 2010, The Human Microbiome Project 2012). Several groups have suggested using these catalogs for an initial database search, and to use the results from that search to create sample-specific databases (Jagtap et al. 2013, Muth et al. 2015, Zhang et al. 2016). Then raw data of each sample can be re-searched against its specific protein database for FDR calculation. This approach, however, is not broadly applicable. For most types of microbial communities, many of the organisms remain un-sequenced and, as such, are not well represented in public repositories.

To illustrate that many microbiome communities remain underrepresented in sequence repositories, a simple experiment was conducted. Protein databases derived from a human preterm infant metagenome project, a mouse metagenome project, and a switchgrass microbiome project were each used to see how represented they were in major public repositories of sequencing information. The human and mouse metagenomes are from Chapter 4 and Chapter 5, respectively. Payal Chirania from the Hettich lab kindly provided the switchgrass metagenome. A blast search was conducted between the protein sequences in these three databases and the UniParc database, which is representative of all known protein sequences (The UniProt 2018). The percent sequence identity for the top hit of each protein was recorded (Figure 2.4). For the human metagenome, the median sequence matched with 100% identity. This match suggests that using a catalog is a feasible approach for human gut microbiomes; however, for the mouse and switchgrass metagenomes, the median sequence matched with less than 80% identity to the UniParc database. This result suggests that the mouse and switchgrass metagenomes were not

well represented at even the genus level in the primary sequence repositories. Based on this information, the catalog approach appears to not be appropriate for samples taken from environments other than the human body, including the mouse, which has a reasonably well-studied gut microbiome. The mouse result was surprising, given that the mouse gut microbiome has a published gene catalog (Xiao et al. 2015). The certainty of using a gene catalog on the human gut microbiome is also not so clear. There were still protein sequences in the human protein database that matched to UniParc proteins with less than 20% identity, and a recent manuscript found thousands of new species uncovered from non-traditionally sequenced populations (Pasolli et al. 2019). Therefore, even analyzing a human metaproteome without prior sequencing is risky since there is a chance that the community has not already been sequenced. To ensure accurate protein sequence representation in the protein database, parallel metagenomic sequencing with metaproteomics is recommended. There is some resistance to this in the metaproteomics community, as the assembly of metagenomes is expensive and requires additional expertise (Zhang et al. 2016); however, this is a weak argument as over the last decade academic and commercial sequencing centers have become ubiquitous, and all the tools needed

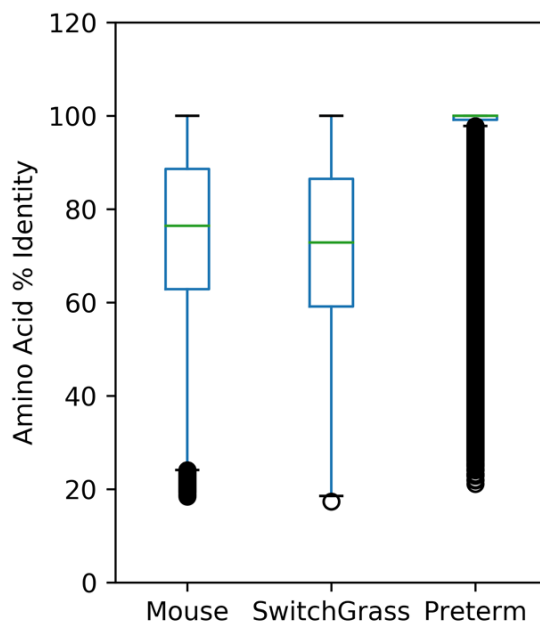


Figure 2.4: Percent amino acid identity between protein databases and their top hits in UniParc. Protein databases where derived from metagenomes that where part of a mouse, switchgrass, and human preterm infant study.

to assemble raw sequence data into discrete metagenome assembled genomes (MAGs) have been made freely available and easy to use (Figure 2.5).

Metagenome assembly typically proceeds as follows. Raw sequencing reads are quality checked and trimmed if necessary. In the case of fecal samples, host DNA can be removed using an aligner; however, in this dissertation, it was found that host sequences are easily identifiable, so this may not be necessary. Cleaned short-reads can then be extended into contigs or scaffolds using metagenome specific assemblers (Peng et al. 2012, Li et al. 2015). At this point, bacterial gene predictors, like MetaGeneMark or Prodigal, can predict genes in contigs with greater than 1,000 sequences (Hyatt et al. 2010, Zhu et al. 2010). Although some progress has been made recently, Eukaryotic gene calling remains difficult (West et al. 2018). These gene calls can be used to make a basic metagenome derived protein database as was done in Chapter 3; however, clustering the contigs into genomes can improve the metagenomes. Using a sequence aligner, reads from multiple samples can be mapped back to contigs to estimate differential coverage (Li and Durbin 2009). Then contigs can be clustered into genomes using differential coverage and tetranucleotide frequency (Kang et al. 2019). Quality of genomes can be assessed for completeness and redundancy (contamination) using single-copy genes (Parks et al. 2015), and redundant genomes can be removed by average nucleotide identity (ANI) retaining the highest quality genome (Olm et al. 2017). Using these steps, it is relatively routine to extract a large number of medium quality genomes, defined as greater than 50% complete and less than 10% contaminated (Bowers et al. 2017). Genomes that do not quite meet this threshold can be further refined by individually clustering their contigs by tetranucleotide frequency and genome coverage and removing outliers (Figure 2.6) (Eren et al. 2015). A pipeline along these lines was developed for Chapter 5, while the protein database for Chapter 4 was derived from an older, more manually derived set of metagenome assembled genomes (Brown et al. 2018).

Not all metagenomes are created equal. A shallow metagenome can lead to an incomplete gene list leading to false-negative protein identifications. To mitigate the creation of a shallow

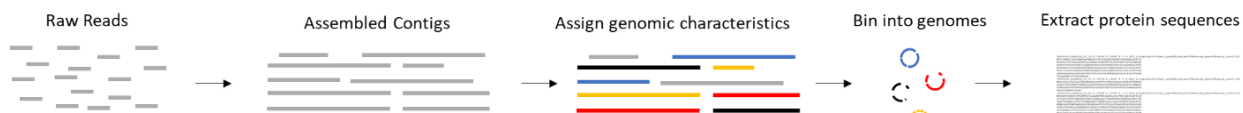


Figure 2.5: Typical workflow for the construction of protein databases from genome resolved metagenomics

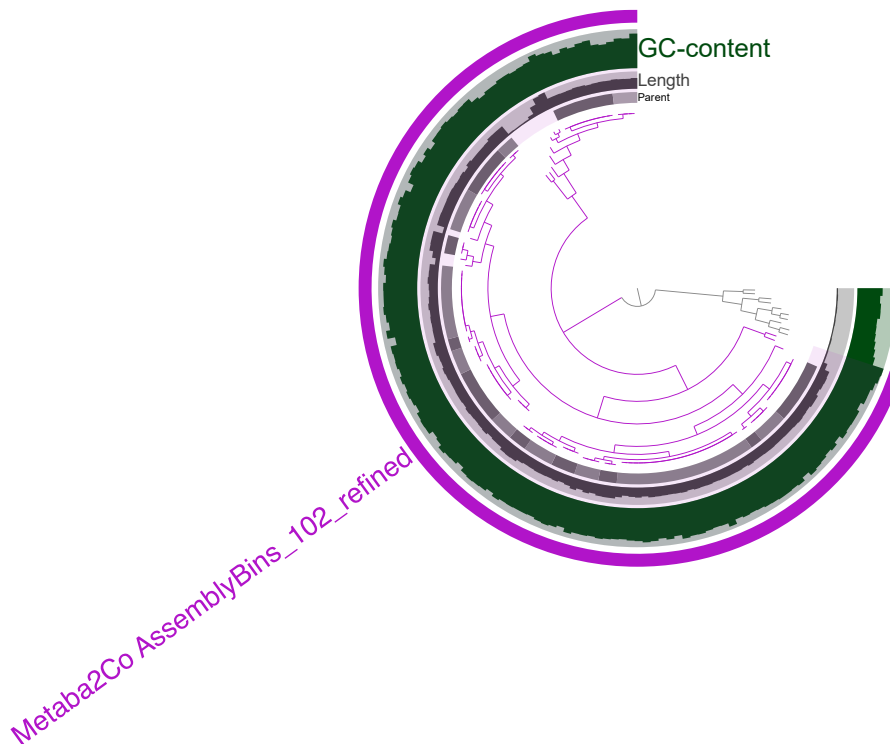


Figure 2.6: Example of refining a metagenome assembled genome in Anvi'o. Contigs are clustered by tetranucleotide frequency and outlier contigs are removed. This genome was adjusted from a completion score 97.5% and a contamination score of 18.38% to a completion score of 96.79% and a contamination of 0%.

metagenome, samples can be preemptively sequenced using 16S rRNA sequencing to predict the size of the community to be profiled. This information can be used to predict how much sequencing is needed to achieve a specific coverage for all the genomes in the community, as was done in Chapter 5. Another solution is to combine metagenomes assembled from multiple samples from the same experiment. To test the effect of combining metagenomes, preliminary searches were conducted on the data from Chapter 4. Each sample from a single individual was searched against its sample-specific metagenome and a few different formulations combining all the metagenomes from that individual. It was found that combining metagenomes led to a significant increase in the number of peptides detected (Figure 2.7). This implies that it is not so much database size, but the appropriateness of the database that affects the number of proteins detected confirming the need for parallel sequencing for metaproteomic analysis. Combining metagenomes has the added benefit of comparing metaproteomes from different samples on the same footing since all the metaproteomes had equal opportunity to detect the same peptides.

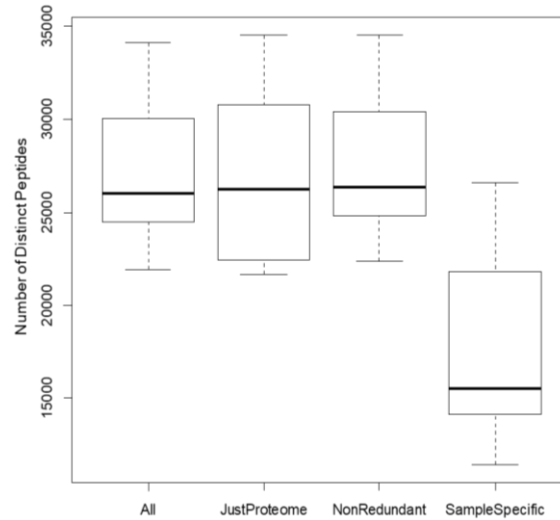


Figure 2.7: Number of distinct peptides detected using different protein database formulations. All the metaproteomes from a single preterm infant were searched using different formulations of the infants metagenome: just the sample specific metagenome, all the metagenomes from the infant, all the metagenomes that had an associated proteome, and all the metagenomes clustered at 100% identity. Sample specific had a significantly lower number of peptides detected ($P < 0.01$).

A problem with combining metagenomes is that gene callers provide sequential identifiers for genes as they are predicted. The addition of sequential identifiers means that the same protein sequence across two metagenomes will have a different identifier. Clustering across all metagenomes involved in a single metaproteomics study can resolve this redundancy by creating a non-redundant list of protein sequence identifiers (Fu et al. 2012, Xiong et al. 2015). Removing redundancy is also necessary so that non-redundant proteins can be inferred by peptides unique to the sequence, not the identifier. There is some debate about the ideal amino acid percent identity for this clustering approach. In Chapter 3, the clustering cut-off was set at 90% identity because this was the point at which the number of additional unique peptides per sample plateaued. In Chapter 4, the cut-off was set at 100% identity to ensure subspecies level discrimination between proteins. In contrast, for Chapter 5, this cut-off was adjusted to 98% identity after it was found that 97% of protein sequences could still be resolved down to the species level.

2.4 Assigning function and taxonomy of origin to proteins detected in environmental samples

Detected proteins do not independently provide any biological information. All information about the taxonomy and function of detected proteins is inferred from the protein sequences found in the protein database. After the detection of proteins in an environmental sample, the next step is to infer the function and taxonomy of origin of said proteins and to use that information to infer functional differences between conditions.

2.4.1 *Assignment of taxonomy based on database*

Assigning taxonomy to proteins detected in a metaproteome is not simple. A suggestion that is gaining substantial prominence in the metaproteomics community is to match the detected peptides to a sequence repository and report the lowest taxonomic rank that the peptide is unique to, as implemented in UniPept and METATRYP (Mesuere et al. 2012, Saito et al. 2015). UniPept uses the UniProt database as the background for this analysis, while METATRYP uses a custom domain-specific database. The problem with this approach is that it assumes that a peptide that is unique to a database is also unique to the entire domain of life. Preliminary analysis in Chapter 5 of this dissertation showed at least one instance of a peptide being traced to a specific strain, where that strain was not in the metagenomes, and the actual protein that was detected by that peptide was from a species whose genera was not present in UniProt. The organism predicted by UniPept and the actual organism of origin ended up being from the same family; however, this simple observation shows how UniPept like approaches are capable of erroneous results. An alternative solution is to use the full sequence to identify the taxa of origin. Several metagenomics tools can do this either by exact match or by sequence similarity searches (Huson et al. 2007, Kim et al. 2016, Menzel et al. 2016). Even the Kyoto Encyclopedia of Genes and Genomes (KEGG) internal annotator, Ghost KOALA, offers genus and phylum level assignments based on the sequences that match the protein sequence of interest (Kanehisa et al. 2016). These approaches, however, fall into the same trap as UniPept by assuming that a reference database is representative of the environmental sample in question. Also, these approaches assume that simply because two protein sequences have highly similar sequences that the overall genomes these sequences belong to will have the same evolutionary history; however, it is well known that the phylogenetic history of a specific gene often does not align with

phylogenetic history of the species as a whole (Maddison 1997, Dewhurst et al. 2005, Szöllősi et al. 2014, Zhu et al. 2019).

All the genes in a metagenome assembled genome rarely match the same species when compared to a sequence repository. This can be shown simply by blast searching all the proteins found on any random scaffold in a metagenome. As illustrated in the table below, the protein sequences will not necessarily match the same species (Table 2.1). A solution to this problem is to use metagenome assembled genomes to assign taxonomy to protein sequences in microbial communities. Using this method, one can assign taxonomy based on the average taxonomical assignment of all the genes in a genome as proposed by the Banfield group, or one can assign taxonomy by placing the whole genome in a phylogenetic tree made up of all the genomes found in NCBI, as proposed by the Genome Taxonomy Database (Brown et al. 2018, Parks et al. 2018, Chaumeil et al. 2019). These approaches also have the added advantage of being able to assign a protein to a species group even if its genome of origin is not found in any major sequence

Table 2.1: A series of protein identifiers taken from a randomly selected scaffold from a human preterm infant metagenome. Each protein was searched using Ghost KOALA and blast.

Gene	Ghost KOALA		NCBI Blast
	Genus	Whole Genome Taxonomy	
b019-d013_scaffold_31_16	Lelliottia	Enterobacter cloacae 7	Enterobacter hormaechei
b019-d013_scaffold_31_17	Enterobacter	Enterobacter cloacae 7	Enterobacter cloacae
b019-d013_scaffold_31_18	Enterobacter	Enterobacter cloacae 7	Enterobacter hormaechei
b019-d013_scaffold_31_19	Enterobacter	Enterobacter cloacae 7	Enterobacter cloacae
b019-d013_scaffold_31_20	Enterobacter	Enterobacter cloacae 7	Enterobacter cloacae
b019-d013_scaffold_31_21	Enterobacter	Enterobacter cloacae 7	Enterobacter hormaechei
b019-d013_scaffold_31_22	Enterobacter	Enterobacter cloacae 7	Enterobacter cloacae
b019-d013_scaffold_31_23	Enterobacter	Enterobacter cloacae 7	Enterobacter cloacae
b019-d013_scaffold_31_24	Enterobacter	Enterobacter cloacae 7	Enterobacter cloacae
b019-d013_scaffold_31_25	Lelliottia	Enterobacter cloacae 7	Enterobacter hormaechei
b019-d013_scaffold_31_26	Leclercia	Enterobacter cloacae 7	Enterobacter hormaechei
b019-d013_scaffold_31_27	Leclercia	Enterobacter cloacae 7	Enterobacter hormaechei
b019-d013_scaffold_31_28	Enterobacter	Enterobacter cloacae 7	Enterobacter hormaechei
b019-d013_scaffold_31_29	Leclercia	Enterobacter cloacae 7	Enterobacter hormaechei
b019-d013_scaffold_31_30	Kluyvera	Enterobacter cloacae 7	Enterobacter cloacae
b019-d013_scaffold_31_31	Leclercia	Enterobacter cloacae 7	Enterobacter hormaechei
b019-d013_scaffold_31_32	Enterobacter	Enterobacter cloacae 7	Enterobacter cloacae
b019-d013_scaffold_31_33	Enterobacter	Enterobacter cloacae 7	Enterobacter cloacae
b019-d013_scaffold_31_34	Leclercia	Enterobacter cloacae 7	Enterobacter hormaechei

repositories. It is the opinion of the author of this dissertation that this should be the norm not only for metaproteomics but for microbiome analysis in general.

2.4.2 Assignment of function to detected proteins

The assignment of function to a protein sequence requires two basic elements: a language by which to define a biological function, also known as an ontology, and a method for transferring information between proteins that have an experimentally confirmed function and proteins that do not. There are many different ontologies, but the two most well-known are the genome ontology (GO) (Gene Ontology 2015), and the enzyme commission (E.C) (Bairoch 2000). The advantage of E.C. is that it offers a set of unique identifiers for all known enzymatic reactions with experimental evidence, while its main disadvantage is that it only provides functional information for proteins that are enzymes. GO provides a common language for all biological functions performed by genes; however, it is organized as a hierarchy that combines broad and specific biological functions, and any function a gene is associated with is added to that gene, whether it performs that function explicitly or if it is just associated. This ambiguity makes it difficult to infer the specific function of a given gene. The primary purpose of GO is to identify enriched functions based on a set of genes, typically called gene set enrichment analysis (Mi et al. 2019).

The most common method of functional transfer is homology, defined as proteins that have a common ancestor. Within homology, orthologs are homologs found in different organisms, while paralogs are homologs found within the same organism, possibly due to a gene duplication event (Li et al. 2003). It is typically assumed that orthologs are better choices for functional transfer even though paralogs often have more similar sequences. There are several publicly available databases of orthologous protein groups that provide functional annotations based on experimental evidence from member proteins, e.g., KEGG, eggNOG, COG (Tatusov et al. 2000, Huerta-Cepas et al. 2016, Kanehisa et al. 2016). Orthologous proteins are grouped using sequence similarity. The most basic approach is an “all versus all” blast followed by Markov clustering (Enright et al. 2002). This approach has two advantages. Blast includes the probability of amino acids switching between each other in its scoring method, offering a more evolutionarily relevant score, and Markov clustering performs a random walk through the network of similar proteins cutting any spurious links. This clustering is meant to account for

promiscuous domains. This approach is the fundamental basis for eggNOG and COG, while KEGG orthology uses a more loosely defined threshold based on the functionality of proteins.

KEGG uses E.C. numbers as the basis for a fairly extensive library of enzymatic reactions assigned to its orthologs, while eggNOG taps into the KEGG orthology for enzymatic information and GO for more general functional information. Since this Ph.D. process started in 2014, functional transfer with one of these databases has become relatively routine, as most of them now offer tools for annotating proteins with their orthologous groupings, after which associated functions such as E.C. or GO can be added to the queried protein sequences (Figure 2.3). Today, it is possible to annotate over a million sequences in under 24 hours.

2.4.3 *Analyzing function in environmental samples:*

After functional ontologies have been assigned to proteins, the next step is to infer functional differences between conditions. In a typical proteomics experiment, this is done by identifying proteins that are significantly different in abundance between conditions and then using the proteins that are up or down in that condition as gene sets for enrichment analysis (Poudel et al. 2019). In the case of a single organism, differences in protein abundance reflect changes in gene expression, and it is a fair assumption that changes in protein abundance reflect changes in function. In microbiomes, however, protein function is highly redundant between organisms, as will be shown in Chapter 3, and differences in protein abundance can be more reflective of changes in community composition, as will be shown in Chapter 5. In many cases, the functions enriched for proteins that are up or down between conditions are the same. This similarity invalidates the utility of gene set enrichment analysis. Since the main strength of the genome ontology is gene set enrichment, it is likely not the best approach for looking at changes in function in microbial communities. An alternative approach is to use protein abundance to quantitate specific functions such as those associated with KEGG orthologs and do a quantitative differential function analysis.

There are two main label-free approaches to quantitate proteins. One is to count the number of spectra assigned to a protein (McIlwain et al. 2012), and the other is to use the MS1 feature intensity of detected peptides to determine protein abundance (Argentini et al. 2016). Typically, MS1 feature intensity is considered the more accurate approach; however, recent evidence suggests that spectral counting may be at least as accurate as MS1 features in

metaproteomes (Kleiner et al. 2017), and spectral counts are substantially easier to calculate. Results from these quantifications can be summed into specific functional ontologies to create a functional abundance, which can then be compared statistically. One issue with this approach is that it is a fairly confusing metric since it is driven by the expression of the proteins within their originating organism, the abundance of the originating organism, and how many proteins are being detected with that function overall. An alternative quantification method could be to simply count the number of proteins associated with a function. This metric simply tabulates the number of proteins expressed above the limit of detection for a specific function and may be more indicative of a bulk transition in gut microbiome function.

2.5 Conclusions

The main objective of a metaproteomics study is to use the suite of measured/identified proteins to infer the functionality of an environmental sample and trace that function back to its taxonomies of origin based on the annotation of protein sequences. This chapter has presented many of the considerations needed to reach that point and suggested some methodologies which were used in the following chapters. This dissertation focuses on describing metabolic function in gut microbiomes. As such, the taxonomy of origin was not particularly emphasized in any of the Chapters; however, phyla of origin could be traced for individual functions across all chapters, and in Chapters 4 and 5, functions could be traced back to species or even subspecies of origin.

3 METAPROTEOMICS REVEALS PERSISTENT AND PHYLUM-REDUNDANT METABOLIC FUNCTIONAL STABILITY IN ADULT HUMAN GUT MICROBIOMES OF CROHN'S REMISSION PATIENTS DESPITE TEMPORAL VARIATIONS IN MICROBIAL TAXA, GENOMES, AND PROTEOMES

Text and figures were adapted from:

Blakeley-Ruiz, J.A., Erickson, A.R., Cantarel, B.L., Xiong, W., Adams, R., Jansson, J.K., Fraser, C.M., Hettich, R.L. Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes. *Microbiome* 7, 18 (2019). <https://doi.org/10.1186/s40168-019-0631-8>

J. Alfredo Blakeley-Ruiz contributions include data analysis, writing of original manuscript, revision, and response to reviewers.

3.1 Introduction

The composition of the gut microbiota can vary quite substantially across time and individuals. Many factors have been shown to impact this variation, including diet, geography, age, genetic relatedness, and health status (Erickson et al. 2012, Yatsunenکو et al. 2012, David et al. 2013). Given the number of possible host factors influencing the gut microbiota, it becomes difficult to tease apart what precisely separates the gut microbiota in healthy versus unhealthy individuals. One potential approach to this complicated question is to focus on discrete “metabolic modules” in a gut microbiome, rather than taxa membership or genomic inventory. Many different bacteria share the ability to perform similar metabolic functions. Hence, microbiomes with very different taxonomic and protein compositions could share a substantial functional similarity, suggesting that functional redundancy in microbial membership may provide an environmental health advantage (Moya and Ferrer 2016). Since proteins are critical participants in the functional activity of life, the direct detection of microbiome-relevant proteins via metaproteomics is ideally suited to help tease apart this problem, and to this end, several studies have demonstrated the ability of LC-MS/MS to measure gut metaproteomes from fecal material, providing a framework for potentially observing gut microbiome function via direct protein detection (Verberkmoes et al. 2008, Erickson et al. 2012, Xiong et al. 2015, Zhang et al. 2018).

A few studies have investigated the gut metaproteomes of preterm infants and healthy adults across time and these studies have revealed variability in both the protein identities as well as overall metabolic activities (Kolmeder et al. 2012, Kolmeder et al. 2016, Xiong et al. 2017). Recently, the temporal dynamics of the gut microbiome composition was determined by 16S rRNA gene sequencing in an inflammatory bowel disease cohort, that included subjects with Crohn's disease (CD) having inflammation either in the ileum (ICD) or the colon (CCD), as well as subjects with ulcerative colitis and healthy individuals (Halfvarson et al. 2017). All of the subjects with IBD exhibited more volatility in microbiome composition over time compared to healthy individuals, but the degree of dysbiosis from healthy was significantly higher for the ICD patients that had undergone surgery. The question remaining is how this volatility influences functions carried out by the gut microbiome.

To examine this issue, an integrated metagenomic/metaproteomic approach was designed to investigate the longitudinal stability and variability in the gut metaproteomes of Crohn's diseased patients post resection-surgery. In contrast to previous studies, the focus of this work seeks to go beyond taxa, gene, and protein profiling to investigate microbial metabolic activity at a higher functional level to observe how persistent and redundant function is maintained across varying microbial gut populations. Fecal samples were collected from several adult individuals in remission (post resection surgery) over the course of one year (25 samples in total, and five samples per individual). Metagenomic and metaproteomic data revealed persistent and phylum-redundant metabolic functions despite a significant level of variability in taxa, genes, and proteins.

3.2 Methods

3.2.1 Patient Cohort

This study focused on five human subjects (labeled as P58, P68, P33, P92, and P104), each with a history of Crohn's Disease and resection surgery, which were selected (as detailed below) from a larger Swedish cohort that has been described previously (Halfvarson et al. 2017). Five subjects were selected who had multiple fecal samples collected over one year. Representatives of both sexes were included in the study (3 females and 2 males) and the subjects were all adults (youngest, born 1967; oldest born 1944). All subjects had undergone resection surgery prior to 2008 (Table 3.1).

Community DNA preparation

Stool (fecal) samples were self-collected and shipped within one day to Dr. Jonas Halfvarsson at the Orebro University Hospital in Orebro, Sweden, where they were immediately frozen at -70°C upon arrival. The samples were stored frozen until use, and then small portions were excised and thawed immediately prior to DNA extraction to avoid freeze-thaw damage. DNA was extracted from 250 mg of each fecal sample in duplicate. For processing, samples were thawed at 4°C and, in aliquots of 0.15 g per tube, resuspended in 1 ml of 1× phosphate-buffered saline. Cell lysis was initiated with two enzymatic incubations, first, using 5 µl of lysozyme (10 mg ml⁻¹; Amresco, Solon, OH), 13 µl of mutanolysin (11.7 U µl⁻¹; Sigma-Aldrich), and 3 µl of lysostaphin (4.5 U µl⁻¹; Sigma-Aldrich) for an incubation of 30 min at 37°C and, second, using 10 µl proteinase K (20 mg ml⁻¹; Research Products International, Mt. Prospect, IL), 50 µl 10% SDS, and 2 µl RNase (10 mg ml⁻¹) for an incubation of 45 min at 56°C. After the enzyme treatments, cells were disrupted by bead beating in tubes with lysing matrix B (0.1-mm silica spheres; MP Biomedicals, Solon, OH), at 6 m s⁻¹ for 40 s at room temperature in a FastPrep-24 (MP Biomedicals). The resulting crude lysate was processed using the ZR fecal DNA miniprep kit (Zymo, Irvine, CA) according to the manufacturer's recommendations. The samples were eluted with 100 µl of ultrapure water into separate tubes. DNA concentrations in the samples were measured using the Quant-iT PicoGreen double-stranded DNA (dsDNA) assay kit (Molecular Probes, Invitrogen, Carlsbad, CA) (Zhang et al. 2018).

3.2.2 Shotgun metagenomic sequencing and assembly

DNA isolation from stool samples yielded 3 – 5 µg of purified metagenomic DNA from each of the fifteen samples. All metagenomic samples were sequenced using the Illumina platform. Illumina libraries were prepared with the DNA Prep Kit (Illumina, San Diego, CA) following a variation of the manufacturer's protocol. Following library construction, each sample was subjected to cluster amplification (cBOT) and paired-end sequencing using an Illumina HiSeq2000 according to manufacturer specifications. Raw sequence data were processed using a combination of Illumina RTA/CASAVA software for base-calling and quality scoring and in-house QC pipelines to filter and truncate low-quality reads. Sequences were assembled using

Table 3.1: Description of Crohn's disease patient cohort.
AB stands for antibiotics and MG stands for metagenome.

Patient #	Sample Name	Day	AB	MG	Birth Year	Gender	Smoker	Extent of Disease	Stricturing
33	A022	0		Yes	1944	F	Yes	Ileocolon	No
33	A64	82							
33	A120	175		Yes					
33	A177	259							
33	B75	361		Yes					
33	B190	441	Y		1949	F	No	Ileocolon	Yes
58	A49	0		Yes					
58	B93	73							
58	A150	171		Yes					
58	B11	263							
58	B135	352		Yes	1952	M	Ex	Ileocolon	Yes
68	A52	0	Y	Yes					
68	A105	91	Y						
68	A158	175	Y	Yes					
68	B40	274	Y						
68	B152	357		Yes	1961	F	Yes	Terminal Ileum	Yes
92	A42	0		Yes					
92	A91	88							
92	A144	177		Yes					
92	B6	282							
92	B125	367		Yes	1967	M	No	Terminal Ileum	Yes
104	A30	0		Yes					
104	A77	84							
104	A131	173		Yes					
104	A194	259	Y						
104	B99	362	Y	Yes					
104	B187	427							

NEWBLER by subject after read reduction using khmer (Kumar and Blaxter 2010, Crusoe et al. 2015). Genes were predicted using MetaGeneMark (Zhu et al. 2010). Default parameters were used for all assembly related software. Amino acid sequences of all the predicted genes were then compiled into five individual specific protein databases (each one consisting of a single concatenated, de-replicated metagenome of all three time-points per individual). The metagenome sequence data can be retrieved using the following URL for the NCBI SRA data deposit, under project ID 46321:

http://www.ncbi.nlm.nih.gov/sites/entrez?db=bioproject&cmd=Retrieve&dopt=Overview&list_uids=46321

3.2.3 Metaproteomics sample collection

Each fecal sample for metaproteome measurements (~130 mg for each sample) was solubilized in 1 mL SDS lysis buffer (4% w/v SDS, 100 mM Tris·HCl (pH 8.0), 10 mM dithiothreitol (DTT)), sonically disrupted (40% amplitude, 10s pulse with 10s rest, 2 min total pulse time), incubated at 95°C for 5 min, and centrifuged at 21,000 x g. An aliquot of each crude protein extract was quantified using a bicinchoninic acid (BCA)-based protein assay kit (Pierce), and yielded about 4 mg/mL protein for each sample. The crude protein extract was precipitated by trichloroacetic acid (TCA), pelleted by centrifugation, and washed with ice-cold acetone to remove lipids and excess SDS, as described previously (Erickson et al. 2012). The protein precipitates were resolubilized via sonication in 500 uL of 8M urea in 100 mM Tris·HCl (pH 8.0) and reduced by incubation with DTT at a final concentration of 10 mM for 1 hr at room temperature. Samples were normalized for total protein at this step by using 1 mg of crude protein for each sample, which was then diluted further with 100mM Tris·HCl and 10mM CaCl₂ (pH 8.0) to a final urea concentration below 4M. Proteolytic digestions were initiated with sequencing grade trypsin (1/100, w/w; Promega) and incubated overnight at room temperature. A second aliquot of trypsin was added (1/100) after the reactions were diluted with 100mM Tris·HCl (pH 8.0) to a final urea concentration below 2M. Following digestion, the peptides were acidified (protonated) in 200mM NaCl and 0.1% formic acid, filtered with a 10kDa molecular weight cutoff spin column (Sartorius) to remove under digested proteins. Final peptide solutions were then quantified using bicinchoninic acid (BCA)-based protein assay kit (Pierce) to enable uniform sample injection onto the LC column.

3.2.4 LC-MS/MS

Peptide mixtures were analyzed in technical replicate measurements via two-dimensional liquid chromatography tandem mass spectrometry (LC/LC-MS/MS) on an LTQ-Orbitrap-Elite mass spectrometer (ThermoFisher Scientific). Peptides (100 μ g per sample) were loaded and separated on-line using a bi-phasic 2D (strong-cation exchange (SCX) and C18 reverse phase (RP))-LC column. Each peptide sample was first washed off-line to remove residual urea and NaCl, and was then placed in-line and analyzed via 22-h 2D-LC-MS/MS. All samples were analyzed by 11 salt pulses (5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 25%, 35%, 50%, and 100% of 500 mM ammonium acetate) each followed by a 110 min gradient to 50% solvent B (70% acetonitrile, 30% HPLC grade water, 0.1% formic acid) with the following profile: 0 to 10% solvent B in 10 min, 10 to 35% solvent B in 75 min, and 35 to 50% solvent B in 25 min. Mass spectral data were acquired using Xcalibur in data-dependent acquisition mode for each chromatographic separation. One precursor MS scan was acquired in the Orbitrap at 30K resolution followed by ten data-dependent MS/MS scans (m/z 400-1700) at 35% normalized collision energy with dynamic exclusion enabled at 1.

3.2.5 Informatics and quantification

Custom-built FASTA target-decoy databases were generated for each individual by combining that individual's specific protein database mentioned above with the human genome and common contaminants. The MS raw data for each sample was searched against the individual specific protein database using MyriMatch/IDPicker with a PSM FDR filter of less than 2% (Tabb et al. 2007, Ma et al. 2009). MyriMatch automatically concatenates a reversed database to the forward version prior to searching to enable proper false discovery rate (FDR) calculation. Protein sequences were clustered into protein groups at $\geq 90\%$ sequence identity using USEARCH (v5.0) (Edgar 2010). A protein group was classified as identified if it had at least one unique peptide and two distinct peptides, as resolved with in-house scripts. Protein groups were assigned spectra using spectral balancing, as previously described (Abraham et al. 2012, Young Jacques et al. 2015, Xiong et al. 2017).

3.2.6 Data analysis

The protein databases were functionally annotated with Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologous groups, and phylum and genus level taxonomical

assignments using Ghost KOALA (Kanehisa et al. 2016). Phylum level taxonomy and KEGG orthologous groups were assigned to each protein group using the Ghost KOALA annotations of the seed sequence for each protein group. The KEGG assignments to each identified protein group were further used to infer and quantify human gut microbiome specific metabolic modules using GOMixer (Darzi et al. 2015, Vieira-Silva et al. 2016). A metabolic module was inferred if $\geq 33.3\%$ of the enzymatic steps in the module were covered within a single phylum. To examine statistical variance of these samples, Spearman correlation coefficient was calculated by loading a tab-delimited file containing identified protein groups and their abundance per technical replicate into Perseus (v1.5.8.5.). The subsequent correlation matrix was rendered in a hierarchically clustered heatmap (**Figure 2**) using the Euclidian distance metric in gplots (Tyanova et al. 2016, Warnes et al. 2016, R Core Team 2017). Functional influence of each phylum was calculated by taking the total number of KEGG orthologous groups identified with protein evidence in each phylum by sample and dividing that by the total number of KEGG orthologous groups found in the sample's protein database.

3.3 Results

3.3.1 *Sequence-guided sample selection and metagenome assembly*

16S rRNA sequences were previously generated and published for 135 fecal samples from a Swedish cohort of IBD patients and healthy subjects (Halfvarson et al. 2017). From this same sample collection, five individuals with ileal Crohn's disease (ICD) were selected for deeper analysis of microbial function by metagenomics and metaproteomics because these ICD individuals showed the greatest microbiome volatility and were the most disparate compared to healthy individuals (Halfvarson et al. 2017). Shotgun metagenomic sequence data were generated from three fecal samples per each individual, corresponding to an initial time point and at 2 subsequent 6-month intervals. Metaproteome measurements were conducted on the same samples, plus an additional two intervening time-points for each individual; ensuring that the protein data was adequately represented by metagenomes on either side of the sampling dates.

3.3.2 Individual gut metaproteomes reveal substantial proteome variability

A total of 14,850 non-redundant protein groups were identified in this study from the 25 samples. Of these protein groups, 494 were human, and 14,356 were microbial. A total of 732-2,900 microbial and 119-222 human protein groups were found in each sample (Figure 3.1A). Although relatively few human protein groups were identified in total number, they were quite abundant in most samples and comprised a substantial proportion of the assigned spectra. (Figure 3.1 B). The measurement correlation for all technical replicates had an R^2 greater than 0.9 and a slope within 0.1 of 1, indicating high technical reproducibility (Figure 3.2).

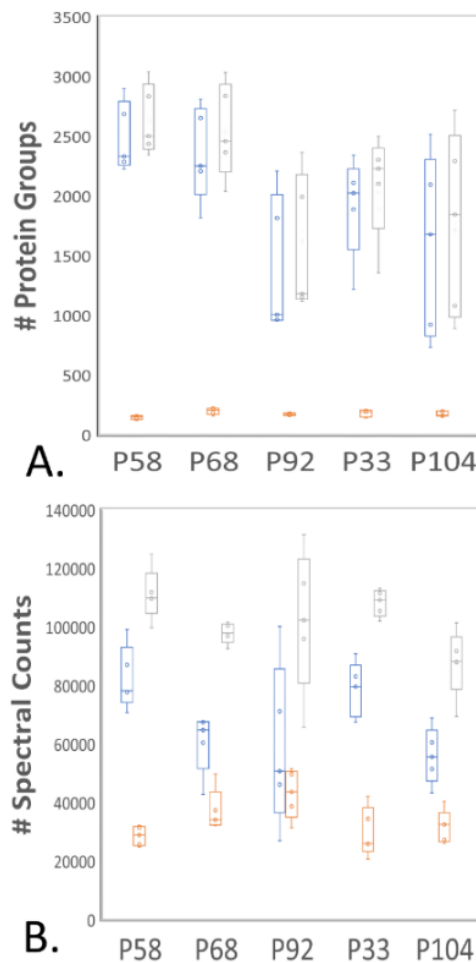


Figure 3.1: Number of proteins detected in Chron's subjects. Box plots of (A) the number of microbial (blue), human (orange), and total (grey) protein groups per sample and (B) the number of spectral counts assigned to each of those groups per sample.

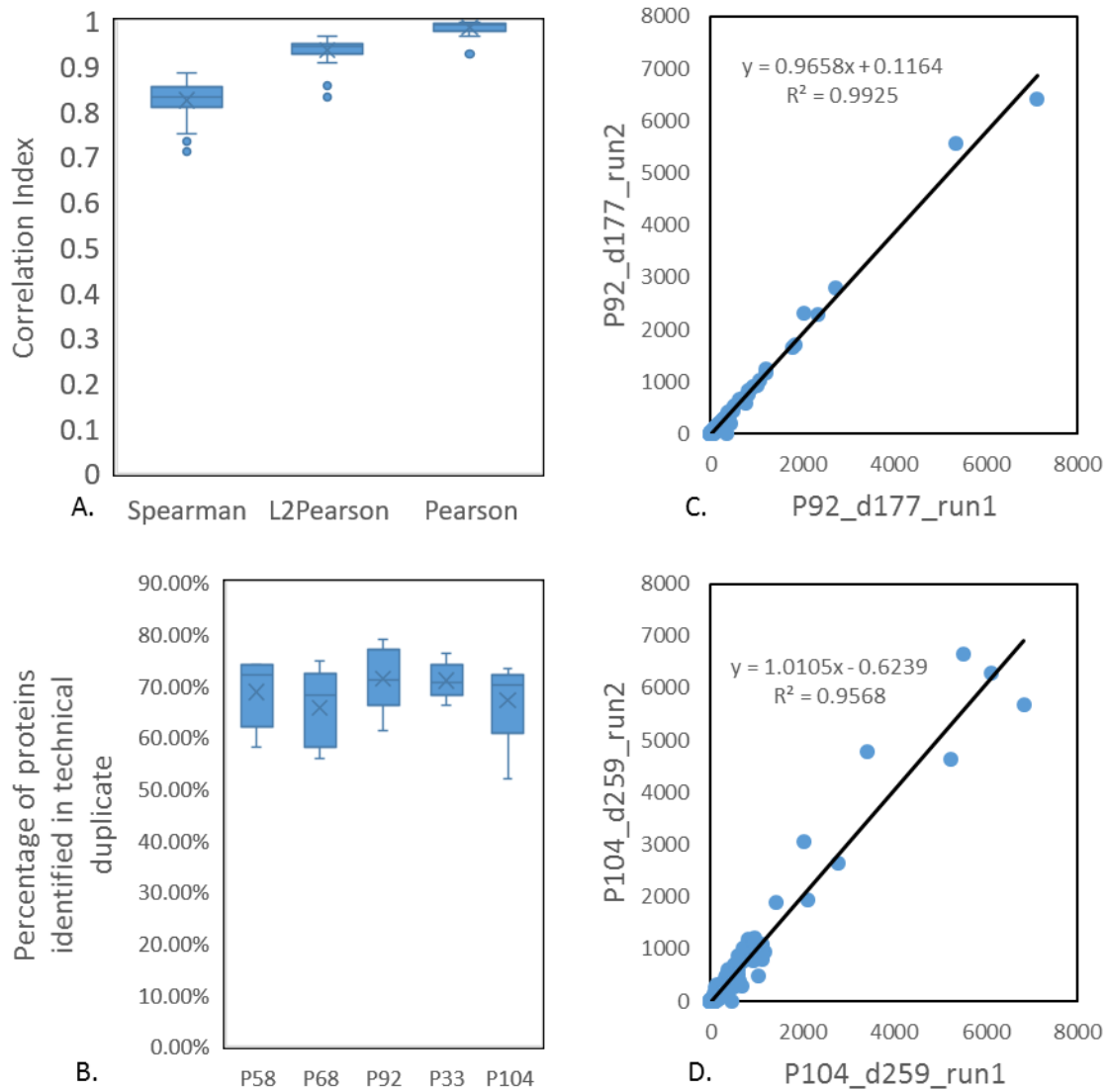


Figure 3.2: Technical reproducibility of Chron's samples from several perspectives. (A) Distribution of Spearman Correlation Index of raw spectra, Pearson Correlation Index of Log2 transformed spectra, and Pearson Correlation Index of raw spectra for each sample. (B) Percentage of the total protein groups identified in each sample that were found in both technical duplicates. (C) Linear fit of raw spectra identified in both technical replicates of sample P92_d177. Sample P92_d177 had the highest technical reproducibility per Spearman correlation index. (D) Linear fit of raw spectra identified in both technical replicates of P104_d259. P104_d259 had the lowest spearman correlation.

As evident by the number of protein groups identified in each sample, qualitative variation between samples was high. Relatively few *identical* microbial protein groups were observed in all samples. Additionally, only 168 microbial protein groups were observed at least once in every individual. Even within a single individual, qualitative variation was high. In individual P104, only 7% of the microbial protein groups identified in that individual were observed across all time points (Figure 3.3). Individuals P33, P104, and P92 had the most varied metaproteomes across time, whereas P58 and P68 had more stable metaproteomes. However, even though individual P58 had the most qualitatively stable metaproteome, less than 36% of its microbial protein groups were found across all time points (Figure 3.3).

With this much variation, it was deemed that Spearman correlation coefficients were best suited to quantify the similarity or dissimilarity of the metaproteomes between measurements (Figure 3.4). Hierarchical clustering of the correlation between measurements revealed a personalized metaproteome. Spearman correlation coefficients not only clustered strongly by technical replicated measurements from the same sample, but also between measurements of the different time points of the same individual. All measurements clustered more closely to samples from the same individual than samples from any other individual (Figure 3.4). These results are similar to a previous study that examined the metaproteomes of healthy adult individuals over time (Kolmeder et al. 2016). The idea of personalized gut metaproteomes appears to be a distinct feature of adult metaproteomes, as a recent study of preterm infants showed that the gut metaproteomes of preterm infants did not necessarily cluster by individual over time (Xiong et al. 2017), indicating that individual specific metaproteomes develop between early infancy and adulthood.

Overall, these results suggest that there is an underlying continuity in metaproteomes from the same individual. However, due to the protein variability found within the individuals even with the most similar metaproteomes (P58 and P68), any real observations of functional similarity between samples are likely more meaningful by observing common identifiers of function, such as KEGG orthologous groups, instead of the individual taxa, genes, or proteins.

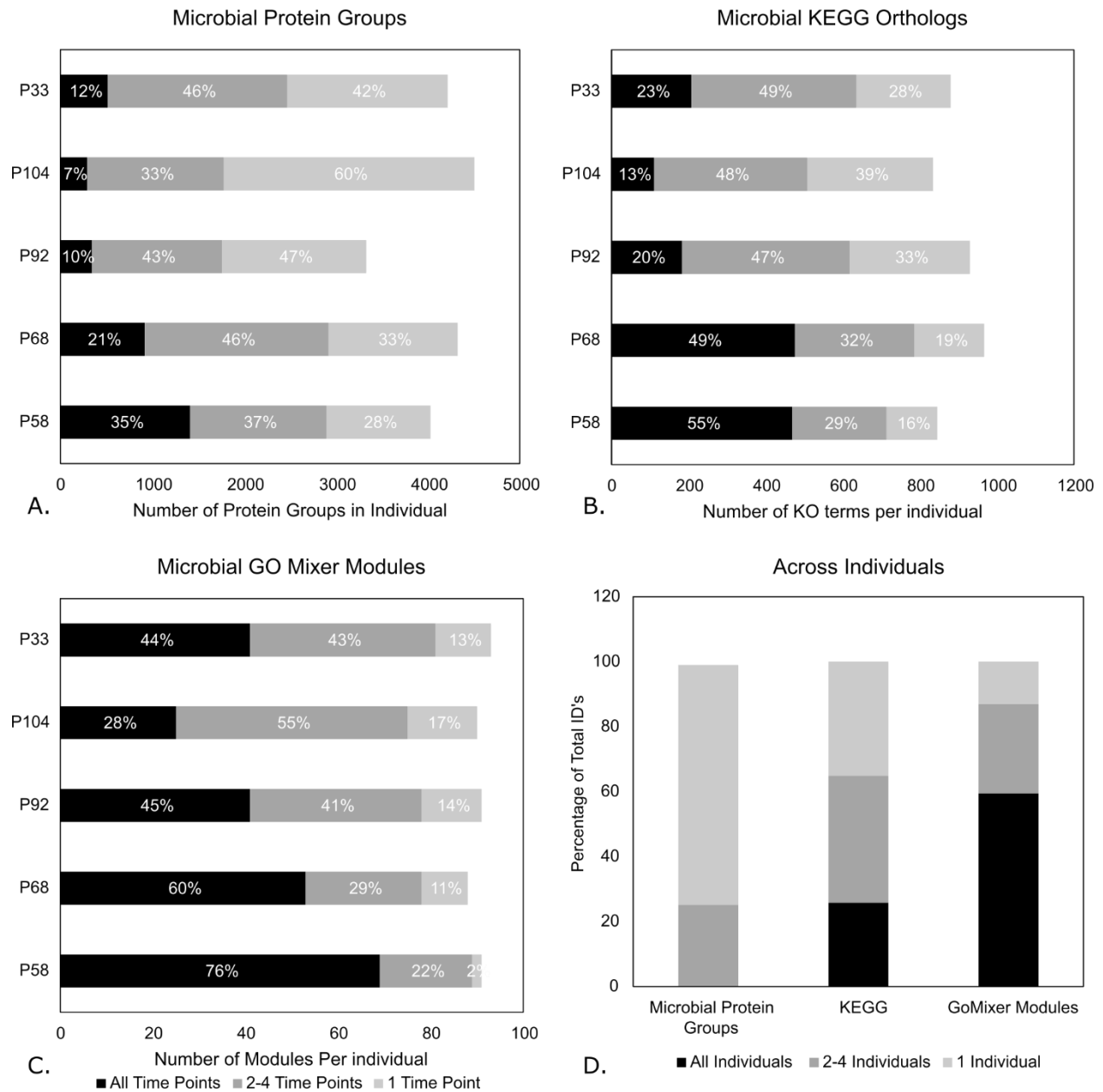


Figure 3.3: Variation in the protein groups identified across time and individuals, and how that variation decreases with functional groups. (A) Shows the number of microbial protein groups that were observed across all time points, more than one time point, and only one time point for each individual. (B and C) Show the same thing but for KEGG orthologous groups and GoMixer modules respectively. (D) Highlights the percentage of microbial protein groups, KEGG orthologs, and GoMixer modules that are found across all individuals, more than one individual, and only one individual.

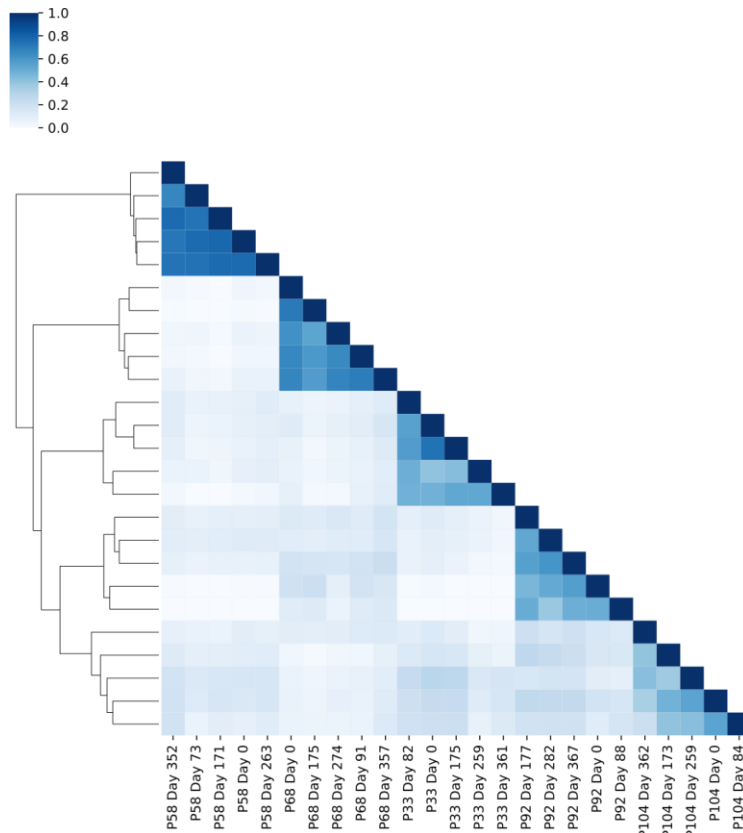


Figure 3.4: Heatmap illustrating the hierarchical clustering of the spearman correlation coefficient between mass spectrometry runs.

3.3.3 Description of the Seven Microbial Proteins Identified in All Samples

While variation was observed within individuals at the protein group level, it was even higher between individuals. Out of 14,850 protein groups, only 60 were identified in all samples. Of those protein groups, 53 were human, and the remaining 7 were microbial. All of these microbial protein groups mapped to the genus *Bacteroides*, and, in most cases, these protein groups were relatively low abundance in the sample, often having less than 100 spectral counts (Figure 3.5).

All 7 of these protein groups were queried for their known function and conserved domains using UniProt and InterPro Scan respectively (Finn et al. 2017, Pundir et al. 2017, The UniProt Consortium 2017) (Figure 3.6).

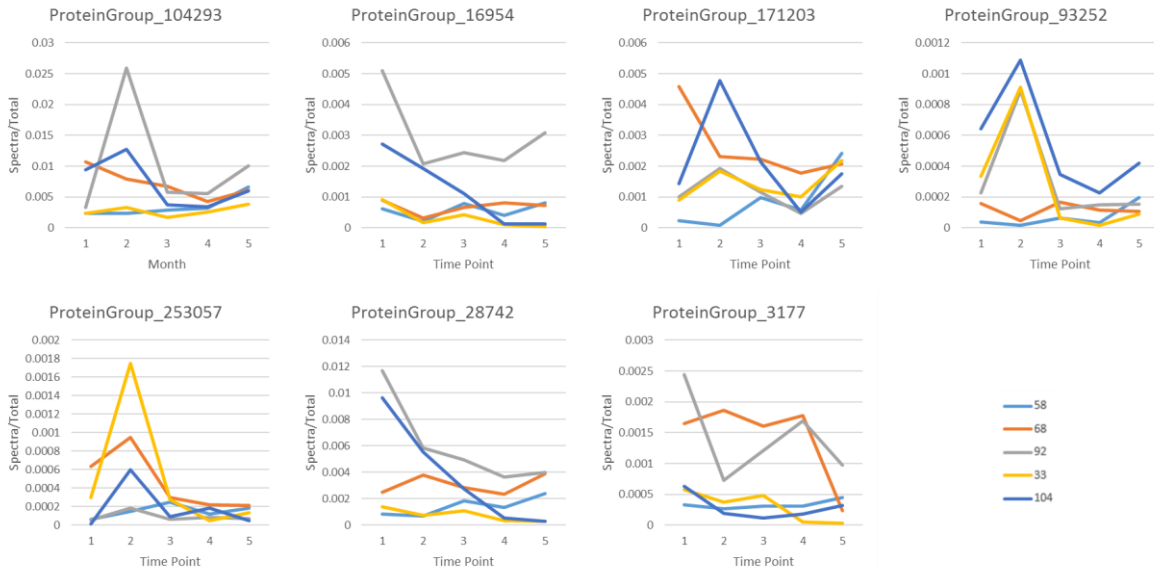


Figure 3.5: Abundance across time of the seven protein groups that were consistently observed across all samples.

Three of these protein groups (104293, 93252, 16954) were found to be related to metabolism. ProteinGroup_104293 had the most spectral counts and contained two domains related to Glyceraldehyde 3-phosphate dehydrogenase, which catalyzes one of the reducing steps in glycolysis. ProteinGroup_16954 contains a FAD binding domain and is annotated as a succinate dehydrogenase/fumarate reductase subunit, which reduces fumarate into succinate or oxidizes succinate into fumarate, and according to KEGG is involved in a variety of microbial metabolic processes, including the TCA cycle, short-chain fatty acid metabolism, carbon fixation, and carbon metabolism (Kanehisa et al. 2017). ProteinGroup_93252 is annotated as phosphoserine aminotransferase, which according to KEGG is a component of the transition from Glycerate 3-phosphate to serine and vice versa, and as a result, plays a role in the degradation and biosynthesis of amino acids. Together these three protein groups cover a wide range of microbial metabolism, including central metabolism, fermentation, and amino acid metabolism. Of the remaining four protein groups, ProteinGroup_253057 is annotated as a L5 ribosomal protein, while protein groups 28742, 171203, and 3177 are proteins of limited known function with low-quality annotations based on their domains. ProteinGroup_28742 could potentially be an antigen. All the sequences that make up this protein group contain a tetratricopeptide repeat domain. This domain

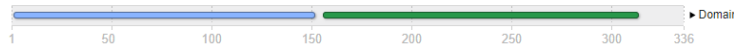
Protein family membership

- IPR020831 Glyceraldehyde/Erythrose phosphate dehydrogenase family
- IPR006424 Glyceraldehyde-3-phosphate dehydrogenase, type I

Homologous superfamilies



Domains and repeats



Detailed signature matches

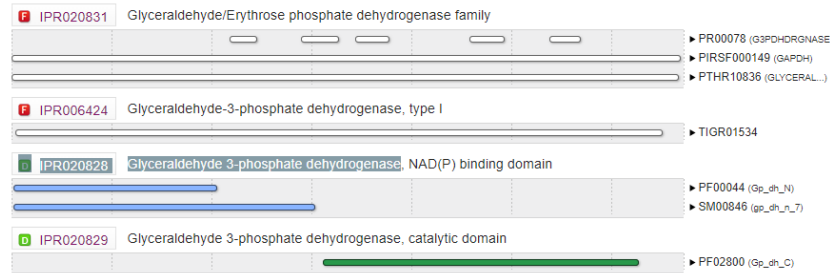


Figure 3.6: Example InterPro Scan result for 7 protein groups found across all Crohn's disease study samples. Specifically, InterPro Scan Result for ProteinGroup_104293 representative sequence. Sequence submitted 2/27/2018. Revealed domains IPR020828 and IPR020829 which are the NAD(P) binding and catalytic domains respectively for Glyceraldehyde 3-phosphate dehydrogenase.

is a common domain often involved in protein-protein interactions and is an important component of several virulence factors among other types of proteins (Cervený et al. 2013). Along those lines, the submitted name for one of the perfect hits in UniPropt for the sequences clustered under ProteinGroup_28742 was “Immunoreactive 53 kDa antigen PG123”. Further research is necessary to characterize the function of these sequences fully. The sequences clustered under ProteinGroup_3177 contained TonB dependent receptor domains and are likely to be a part of the SusC/RagA family. Interestingly, TonB and SusC were shown to be significantly more abundant in ileal Crohn's disease in a previous publication (Erickson et al. 2012). The UniPropt hits to the sequences clustered under ProteinGroup_171203 were typically either “Uncharacterized protein” or “NifU-related domain containing protein”. Linking from the UniPropt annotation to the InterPro annotation revealed that the sequences clustered under this protein group were homologous to the domain labeled IPR002871. This domain is reported to be the N-Terminal of NifU, which contains a 2Fe-2S cluster, and is implicated in nitrogen fixation systems, although the exact biochemical function remains unknown (Ouzounis et al. 1994, Finn et al. 2017). A homolog to this domain was highlighted previously as a core human gut

metaproteome protein (Kolmeder et al. 2012). Given its robust observation in gut metaproteome studies, further insight into the function of this family of proteins could reveal novel insight into gut community function. The realization that almost half of the proteins identified across all samples in our database remain poorly characterized highlights the challenge of using proteomics to understand microbiome function when many if not most of the protein sequences remain poorly understood

3.3.4 Assignment of KEGG Orthologous groups reveals remarkable functional redundancy among protein groups

KEGG can be used to infer the functional and taxonomic association of protein groups based on the biochemical function of the orthologous groups (KO terms) (Aoki-Kinoshita and Kanehisa 2009), and was employed here to classify and interrogate the metaproteome information. For greater than 99% of protein groups, all the amino acid sequences associated with that protein group mapped to the same KEGG orthologous group. Similarly, for greater than 99% of all protein groups, all the amino acid sequences in that group mapped to the same bacterial phylum. This trend was not true for other lower taxonomic levels. For this reason, the seed protein's Ghost KOALA KO and phylum annotation was used for all subsequent analyses. This was deemed to be the most accurate possible function and taxonomy that could be assigned to each protein group.

An average of 165,451 ORF's per protein database were assigned a KO term. This translated into an average of 5,504 predicted KO terms per individual, indicating high functional redundancy across protein sequences. Similarly, this high functional redundancy was also observed across identified protein groups. Of the identified protein groups, 10,172 were assigned a KO term, but this only amounted to a total of 1,573 KO terms. These results are consistent with the concept that many of the different protein groups identified likely have the same function. This is supported by the much more expansive HMP project, which reveal similar available metabolic functions despite varied taxonomy (The Human Microbiome Project 2012).

Within an individual, 16% of the KO terms predicted by each individual's protein database had protein evidence in at least one of their five samples. For context, less than 2% of the protein groups predicted for each individual had protein evidence in at least one sample. The majority of the observed function was centered around four phyla: Firmicutes, Bacteroidetes,

Actinobacteria, or Proteobacteria. These phyla represented the greatest percentage of the total number of KO terms predicted by each sample's respective protein database. Similarly, Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria, along with Verrucomicrobia were the most represented phyla in previous human gut metaproteomic studies (Kolmeder et al. 2016, Tanca et al. 2017). Although they were observed in this study, Verrucomicrobia did not represent a significant proportion of the identified protein groups. This could be related to compromised health state of the individuals in the current study as compared to the other studies of healthy individuals. While not as dominant as the other four phyla, protein groups associated with Fusobacteria and Euryarchaeota in P58 and P33, respectively, represented an uncharacteristically high percentage of the total number of KO terms predicted by their associated protein databases.

Variability between samples decreased when the qualitative analysis was constrained to the KO term level. Between individuals, 25% of the KO terms identified across all samples were seen at least once in every individual. There was also lower intra-individual variability. A greater percentage of the total number of detected KO terms in each individual were also found across time when compared to the protein groups (Figure 3.3). These followed essentially the same trend as the protein group level observations, with P104, P33, and P92 being more variable than P58 and P68 (Figure 3.3). This suggests two conclusions: the trends in terms of the most variable individuals across time hold true through to the KO term level, and there is overall less qualitative variation between samples when focusing on functional groups instead of protein groups (Figure 3.3).

3.3.5 Inferred metabolic modules demonstrate persistence of metabolic function in a dynamic microbiome

To expand the search for functional similarity between samples, the metabolic dynamics of this metaproteomic dataset were explored using a map of human microbiome derived metabolic modules created by the Raes lab (Vieira-Silva et al. 2016), which represent a set of reactions responsible for the conversion of one compound to another and are inferred via enzyme evidence from the metaproteomic datasets. This metabolic map is available in an online tool called "GOMixer" at <http://www.raeslab.org/gomixer/>, and is illustrated for this study in Figure 3.7. 116 out of a total of 133 metabolic modules in "GOMixer" were inferred for specific phyla using the Ghost KOALA annotations for each protein group. A variety of phyla contributed

proteomic evidence for these modules; however, most of the inferred modules were populated with protein groups that originated in Firmicutes, Bacteroidetes, Proteobacteria, or Actinobacteria. Two or more phyla provided evidence for most of the modules observed in Figure 3.7, where the green lines indicate modules that have protein evidence in two or more phyla, and the red lines indicate modules that have protein evidence in all four of the main phyla described above. Together these modules indicate protein evidence for high redundancy in metabolic function between gut microbiome phyla.

An inspection of the inferred modules reveals several interesting points about the gut microbiome's metabolic function in these five individuals. First, despite substantial variation in the protein groups observed in each sample, these modules demonstrate persistence of *metabolic function* across time and individuals. 69 modules were observed in all individuals, of which 18

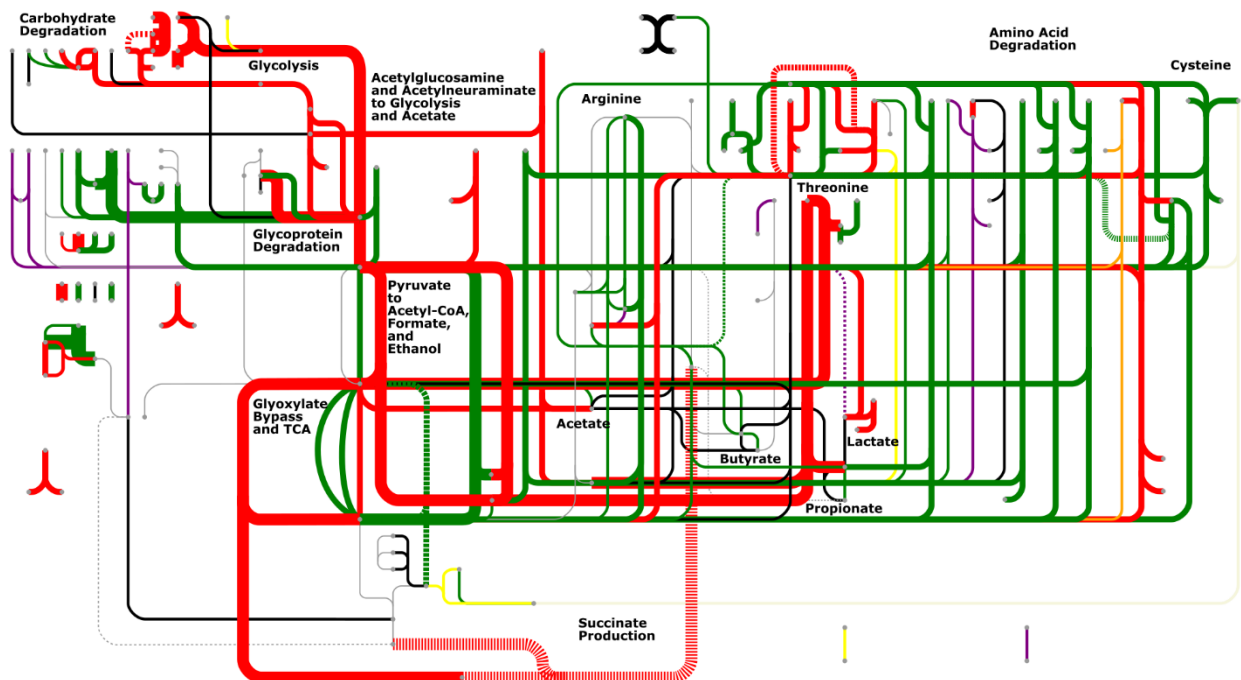


Figure 3.7: Metabolic map of all the GOMixer modules rendered in the GOMixer web application. Highlighted lines represent all the modules that were inferred with peptide evidence in at least one sample. Line thickness represents how many samples the module was observed in with the thickest lines representing modules that were represented across all samples, the medium thickest lines being represented by modules inferred in all individuals, and the least thick lines representing modules not seen in all individuals. Red colored lines represent modules that were observed in Firmicutes, Actinobacteria, Bacteroidetes, and Proteobacteria. Green lines represent modules that were observed in at least two of the phyla. Black lines represent modules only represented in Firmicutes. Yellow lines represent modules only observed in Bacteroidetes. Purple lines represent modules only represented in Proteobacteria. Two modules that were unique to other phyla were MF0027 inferred in Euryarcheota and MF0010 inferred in Fusobacteria. These were colored beige and orange, respectively.

were observed in all samples. Tracking through these 18 modules reveals a clear path from carbohydrate, lipid, and amino acid degradation to central metabolism and finally the production of fermentation products (Figure 3.7). Second, this persistence of function is redundant across multiple phyla, but was not always observable in the same sample. All 69 modules except for one were observed in two or more phyla, and all of the 18 modules observed across all samples were observed in two or more phyla, but not always in the same sample. In some cases, the protein evidence for a module comes from one phylum but then changes to another phylum at a different time (Figure 3.8). Third, the gut microbiome metabolism is not driven by a set of discrete linear pathways, but a web of interconnected reactions facilitated by a network of enzymes that connect multiple molecules across multiple pathways. For example, the identified protein groups suggest that lactose (MF0048) degrades into glucose before following the classical glycolysis pathway. In contrast, glycerol (MF0107, MF0108, MF0109) and fucose (MF0124) degrade into glyceraldehyde 3-phosphate, which is the precursor to the pay-off phase of glycolysis. Similarly, protein enzyme pathway detection suggests that threonine degradation (MF0029 and MF0030) skips glycolysis by degrading directly into Acetyl-CoA and other fermentation products, including propionate and formate. These modules represent a major portion of the metabolic function of the gut microbiome. The population of these modules from the measured proteomes provides evidence for microbial-driven degradation of complex molecules across multiple samples, including all 20 amino acids, and a large repertoire of carbohydrates, such as starch, fucose, xylose, arabinose, lactose, and glycerol. Degradation modules connect to energy producing modules such as glycolysis, glyoxylate bypass, and fermentation, for which there was consistent protein evidence. These energy-producing modules culminate into a variety of gut relevant metabolites, including pyruvate, acetyl-CoA, succinate, formate, propanoyl-CoA, CO₂, acetate, butyrate, and propionate.

The metaproteomic experimental evidence for these specific metabolic modules reveals a detailed functional view of the gut microbiome ecosystem, in which microbes acquire energy by degrading complex molecules from both the host and the host's diet, and then this degradation process ultimately leads to metabolites, such as butyrate, acetate and propionate, that the host can use for energy. It appears that the majority of this metabolic function is not specific to an individual bacterial strain, species, genus, or phylum but is instead shared between phyla. Furthermore, the phylum driving an individual module, as evidenced by proteomic data, can

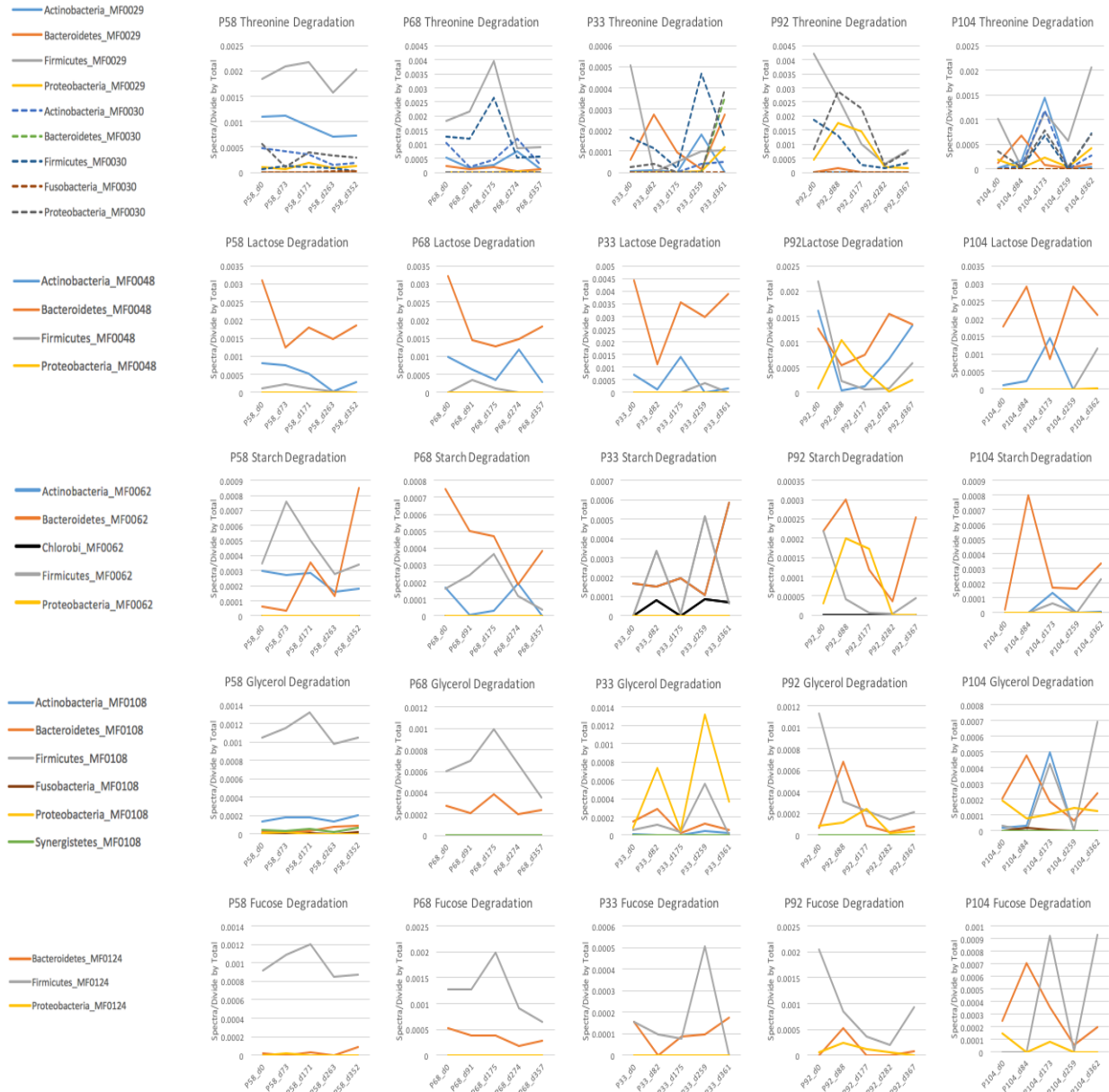


Figure 3.8: Trends by phylum in modules pertaining to degradation of food products that were consistently observed across all samples. MF0029 and MF0030 represent different pathways that degrade Threonine the only amino acid degradation pathway that was seen across all samples. MF0048 represents Lactose degradation. MF0062 represents starch degradation a critical polysaccharide food source that comes from plants. MF0108 represents glycerol degradation, the only lipid degradation pathway that is observed across all samples. MF0124 represent Fucose degradation which is a mucosal glycan that the host produces and that the bacteria in the gut microbiome consume. were observed in two or more phyla, but not always in the same sample.

change dynamically across time and individuals (Figure 3.8). This manuscript provides a detailed analysis of a few of these key modules, but the full set of modules that were inferred with proteomic evidence across time and individuals is detailed in Supplementary Spreadsheet 3 of Blakeley-Ruiz et al. 2019.

3.3.6 Observation of acetylglucosamine and acetylneuraminate indicate the microbiome can feed itself at the expense of the host

Detailed analysis of Figure 3.7, and the information in Supplementary Spreadsheet 3 of Blakeley-Ruiz et al. 2019, prompted interest in two specific modules, MF003 and MF005. These modules are observed in all individuals across multiple phyla and highlight metabolites that enter central metabolism at multiple positions in the pathway. Specifically, these modules represent the degradation of acetylglucosamine and acetylneuraminate into D-Fructose 6-phosphate, pyruvate, and acetate. Acetylglucosamine and acetylneuraminate are sugar amine hybrids found in the host mucosal layer, human milk oligosaccharides, and in animal food products that are key host nutrients, and presumably are used by the gut microbiome as a source of energy (Wang and Brand-Miller 2003, Marcobal et al. 2011, Tailford et al. 2015). Interestingly, providing N-acetylglucosamine as a supplement has been suggested as a potential method for alleviating inflammatory bowel disease (Salvatore et al. 2000). MF0003 and MF0005 enter central metabolism at three separate levels, and merge with each other at N-Acetyl-D-glucosamine 6-phosphate, highlighting a simple demonstration of the interconnected nature of gut microbiome metabolism (Figure 3.9).

Evidence for the presence of MF003 and MF005 was observed across all individuals and in *Firmicutes*, *Proteobacteria*, *Bacteroidetes*, and *Actinobacteria*, thereby demonstrating the persistence of function despite the dynamic nature of the gut microbiome. Evidence for N-acetylneuraminate lyase (K01639) was observed in *Firmicutes* (P68, P92, P58, and P104), *Proteobacteria* (92), and *Bacteroidetes* (P68, P33, P58, and P104), but not in *Actinobacteria*. Evidence for the production of N-Acetyl-D-glucosamine 6-phosphate was limited in both modules; however, evidence for N-acetylmannosamine-6-phosphate 2-epimerase and N-acetylmannosamine kinase (K00885 and K01788) was observed in *Proteobacteria* on day 88 in P92 and in *Actinobacteria* on day 173 in P104. Similarly, evidence for the phosphotransferase system (K02802, K02803, and K02804) was only observed in *Firmicutes* in individuals P68 and

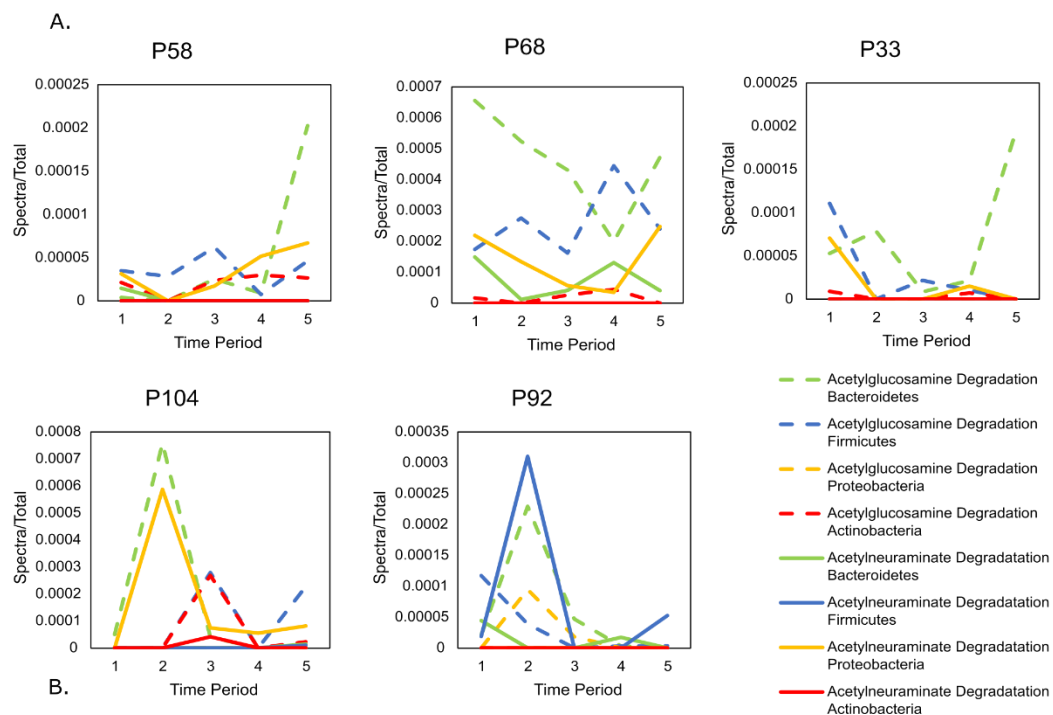
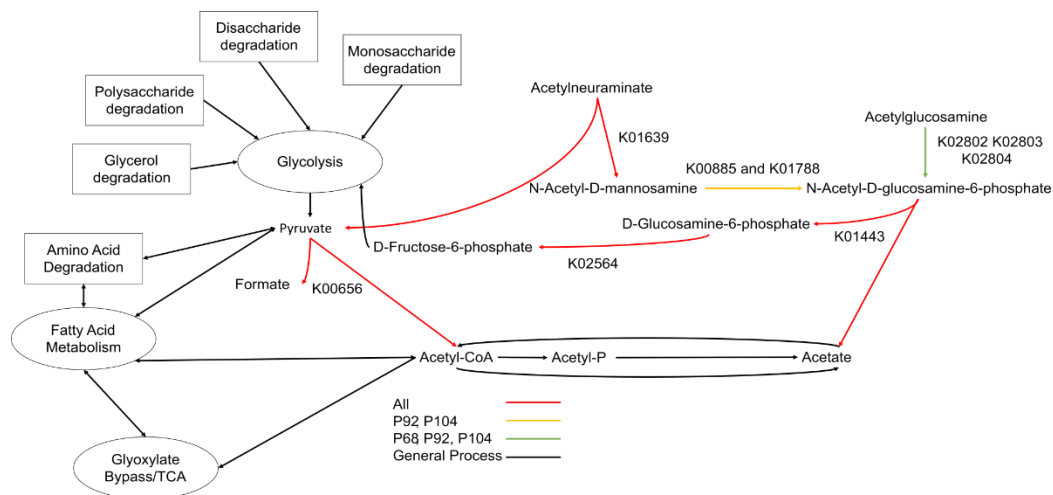


Figure 3.9: Detailed pathway describing modules MF0003 and MF0005, which together degrade acetylglucosamine and acetylneuraminate into pyruvate, fructose-6-phosphate, and acetate. All the KEGG ortholog terms in this figure had peptide evidence in our data set. K00656, while not a part of these modules, had proteomic evidence in every sample of this database, and was the method for converting pyruvate to Acetyl-CoA that there was the most protein evidence for.

P104, and in both *Firmicutes* and *Proteobacteria* in individual P92. Despite the relatively sparse evidence for enzymes specifically involved in the production of N-Acetyl-D-glucosamine 6-phosphate, mentioned above, there was substantial evidence for enzymes associated with the utilization of N-Acetyl-D-glucosamine 6-phosphate. N-acetylglucosamine-6-phosphate deacetylase (K01443) was observed across all individuals in *Firmicutes* (P68, P92, P33, P58, P104), *Proteobacteria* (P92), *Actinobacteria* (P33, P58, P104), and *Bacteroidetes* (P68, P92, P33, P104). Glucosamine-6-phosphate deaminase (K02564) was also observed across all individuals in *Firmicutes* (P68, P92, P33, P58, P104), *Actinobacteria* (P68, P58, P104), *Proteobacteria* (P92), and *Bacteroidetes* (P68, P92, P33, 58, P104). Together these enzymes show consistent presence of both modules across all individuals.

Of interest, N-acetylglucosamine-6-phosphate deacetylase (K01443) was highlighted as a core enzyme in a previous healthy cohort study (Kolmeder et al. 2016). Combined with the fact that this enzyme is found across multiple phyla, this suggests that it may be a key source of energy for gut microbiomes. In contrast, K01639 and K01639 were also observed in all individuals in this study, but were not in the core set from this previous study, possibly suggesting a potentially different utilization of D-Glucosamine 6-phosphate.

These enzymes serve as a representative case for both the redundant and dynamic nature of microbiome metabolic functions. For example, most of these enzymes were observed in one of the four phyla at least once, but the modules were not always inferred in the same phyla across individuals. These modules also highlight the web-like nature of gut microbiome metabolism. The degradation of acetylglucosamine and acetylneuraminate connect at N-Acetyl-D-glucosamine 6-phosphate, and together these modules release products into central metabolism at three separate points: D-Fructose 6-phosphate, pyruvate, and acetate. D-Fructose 6-phosphate is a precursor to the pay-off phase of glycolysis. Pyruvate is the end product of the pay-off phase of glycolysis and is the starting point for a variety of fermentation pathways as well as the TCA cycle and glyoxylate bypass. Acetate is one of the three main short chain fatty acids produced by the gut microbiome, and is a source of energy for both the host and the microbiome. It can be converted to Acetyl-CoA at the cost of ATP or via transferase with the help of butyryl-CoA or propionyl-CoA (Vanderwinkel et al. 1968). Conversion by transferase produces butyrate and propionate which can also be converted by the host into energy. Together these modules

represent a sample case of how many different phyla contain similar metabolic modules that can utilize host derived molecules, such as acetylglucosamine and acetylneuraminate.

3.3.7 Observation of short-chain fatty acid associated modules indicate that the host is fed by a changing set of pathways and bacteria

Just as acetylglucosamine and acetylneuraminate feed the microbiome, short-chain fatty acids, particularly acetate, propionate, and butyrate, are thought to be one of the primary methods by which the gut microbiome feeds its host. “GOMixer” highlights 9 modules as acetate (MF0113), propionate (MF0121, MF0122, MF0123, MF0125, MF0126), or butyrate (MF0114, MF0116, MF0117) producing. All these modules were observed with peptide evidence in at least one sample. Some of these modules do not represent the direct production of short-chain fatty acids, but instead, represent the production of precursors to CoA-transferase and kinase reactions. Only MF0113, MF0116, MF0117, MF0125, and MF0126 connect the direct production of short-chain fatty acids via CoA-transferase, kinase, or a similar reaction. The average abundance of the KO terms in each module divided by the total depicts the persistence of function amidst a dynamic microbial ecosystem. Particularly propionate and acetate producing modules persist across time and individuals while being driven by different phyla or modules (Figure 3.10). Evidence for enzymes involved in the direct production of butyrate was observed only in individuals P58, P68, and P92. This supports previous proteomic studies, which indicated that the abundance of butyrate-producing enzymes is significantly lower in patients with Crohn’s disease (Erickson et al. 2012).

Utilizing the KEGG orthologous groups that had representation in the modules and their associated reactions in KEGG, a reaction network was manually produced (Figure 3.10). This network highlights evidence for dynamic and interconnected microbial metabolic function. Acetate production via acetate kinase (MF0113) had peptide evidence across all individuals. One set of enzymes can produce acetate via a two-step process that involves first converting acetyl-CoA to acetyl-phosphate (K13788, K00625, K15024) and then converting acetyl-phosphate to acetate (K00925), producing ATP in the process. These reactions are reversible (Wolfe 2005), and there was evidence for the enzymes involved in both steps across all individuals and almost all the samples. According to the KEGG database and the literature, given propanoyl-CoA as the starting material, this module will also produce propionate, providing

peptide evidence for the potential production of propionate across all individuals as well (Heßlinger et al. 1998). This highlights the interconnectedness of gut metabolism, as the same enzymes can plug into different fermentation routes of the broader metabolic network given different starting material.

There was also limited evidence for a propionate specific kinase (K00932) in individual P92, and the production of propionate via CoA transferase (K01026) in P33 and P58. Production of butyrate or propionate by CoA-transferase requires the transfer of the coenzyme from butyrate or propionate to another acid typically acetate, but it could also be lactate. This is an interconnected process. For example, propionate CoA-transferase, K01026, according to KEGG reactions, transfers CoA between lactate, propionate, or acetate to produce lactoyl-CoA, acetyl-CoA, or propanoyl-CoA. K01026 further highlights the interconnectedness of gut microbiome metabolism as Propanoyl-CoA is converted into propionate by converting acetate or lactate into acetyl-CoA or lactoyl-CoA respectively, and vice versa.

Butanoyl-CoA can be converted into butyrate by two main methods, presenting a good example of the dynamics of microbial metabolic function. Acetate CoA-transferase, (K01034 and K01035) converts butanoyl-CoA to butyrate by converting acetate into Acetyl-CoA and vice versa. Butyrate kinase is a two-step process that involves the conversion of Butanoyl-CoA to Butanoyl-Phosphate (K00634), followed by the production of butyrate via the kinase (K0929). In each individual, except P92, which had no peptide evidence of butyrate production, only one method or the other was observed. P58 and P68 had evidence for the production of butyrate via transferase, and P104 and P33 had evidence for the production of butyrate via kinase. Albeit, P33 only had peptide evidence for K00634, but not the actual kinase. This presents an example of persistence of function despite a dynamic gut microbiome. There was peptide evidence for butyrate production across most individuals; however, half of the individuals had peptide evidence for a different module than the other half. Although not typically included in discussions about short-chain fatty acid production in the gut, it should be noted that peptide evidence for the production of the smallest short-chain fatty acid, formate, was observed via K00656 across all samples, thus providing persistent peptide evidence for the conversion of pyruvate to acetyl-CoA across a variety of phyla.

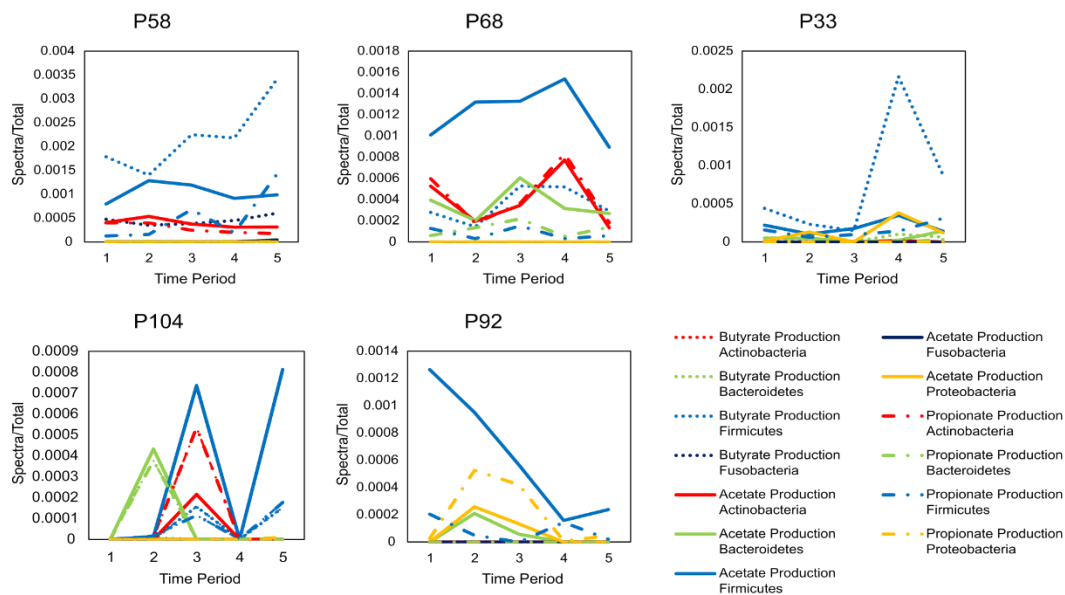
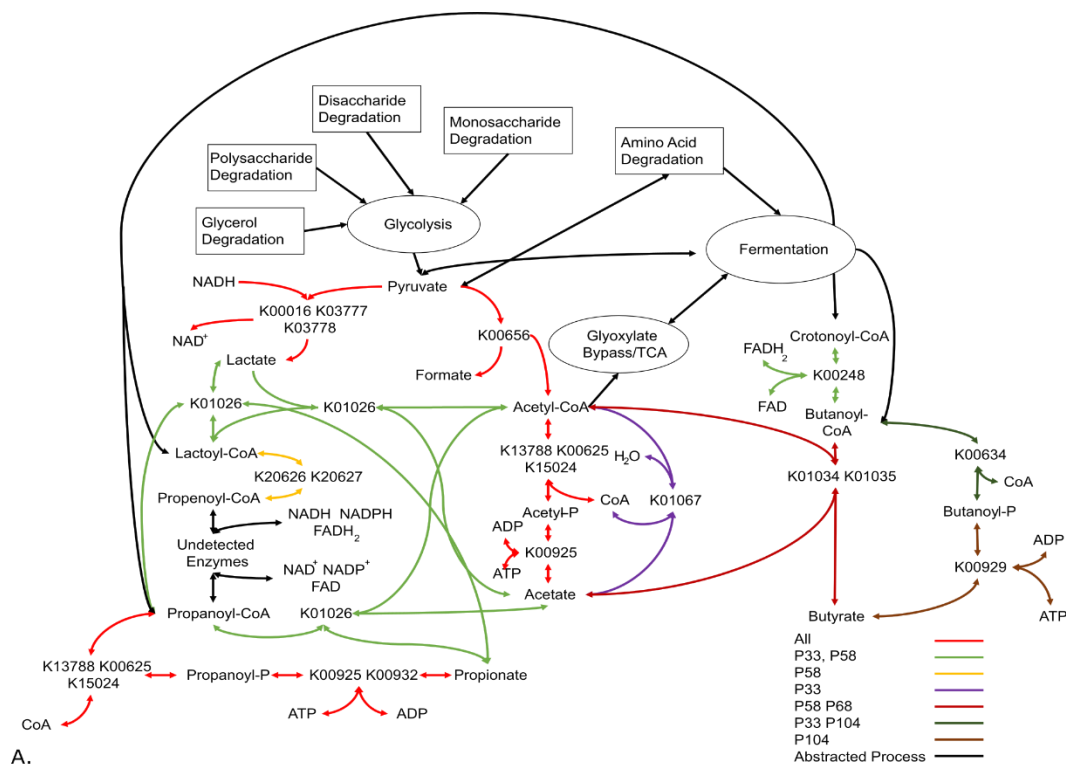


Figure 3.10: Representation of direct routes to short chain fatty acids and acetyl-CoA with proteomic evidence. All KEGG Ortholog terms mentioned in this figure had proteomic evidence in at least one sample, although many had evidence across all individuals and most samples. For example, K00656 had proteomic evidence in all samples and in Firmicutes, Actinobacteria, Proteobacteria, and Bacteroidetes, but not always at the same time. These enzymes form a sizable representation of the pathways to these metabolites with proteomic evidence; however, not all enzymes that can produce these molecules are included.

K00656 has been consistently observed in previous metaproteomic studies, along with acetate producing enzymes K00625 and K00925, (Kolmeder et al. 2016, Tanca et al. 2017) indicating that acetate production via these enzymes may be ubiquitous regardless of human host health status. Kolmeder et al. 2016, highlights Propionate CoA-Transferase and Butyryl-CoA dehydrogenase as being core enzymes, although these are not observed across all individuals in this dataset. This correlates to previous research which has shown that short-chain fatty acid producing proteins especially butyrate, are less abundant in Crohn's diseased patients (Erickson et al. 2012).

3.4 Discussion

Measurement of the metaproteomes of Crohn's diseased patients (all post resection-surgery) over the course of a year revealed substantial variability in the protein groups observed across time and individuals. There was significant qualitative variability even within the same individuals, which supports the dynamic nature of the microbiome's composition, as was previously observed in the same individuals (Halfvarson et al. 2017). The gut metaproteomes of these individuals were distinctly personalized, with some individuals exhibiting less variability across time than others.

This study has demonstrated that despite this variability, many metabolic functions are consistently observed across diverging sequence space, suggesting that sequential variability may not be a good indicator of metabolic functional variability. Individual KO terms were found to be more consistent across time and individuals than protein groups. The majority of KO terms observed originated from Firmicutes, Actinobacteria, Proteobacteria, and Bacteroidetes, and when placed in a metabolic context, consistent and somewhat redundant metabolic functions relating to the degradation and fermentation of food products were observed.

Although this study did not focus on taxonomical resolution below the phylum level (for reasons noted above), it is evident that almost all the functions observed across all individuals were observed in multiple phyla. This provides compelling evidence that these functions are not specific to any one phylum, genus, or species, and may suggest that functional redundancy across taxa is a hallmark of robust gut microbiome stability. These modules reveal the interconnectedness of gut microbiome metabolism, suggesting that overall gut microbiome operation should be viewed in a network context focused on metabolic function. By combining

metagenome resolved genomes (MAGs) with metaproteomes it could be possible in the near future to acquire genome constrained proteomes for complex communities providing a framework by which proteome guided metabolic reconstruction could be done. Although using a much broader taxonomical level due to technical constraints, this chapter provides a first look at how metabolic function is shared across taxa, and persists despite different identified protein groups between samples, and a taste of what MAG driven community proteomics could look like in the near future.

Although this study was not designed to directly contrast healthy individuals versus individuals with Crohn's disease, these specific diseased individuals have been contrasted with healthy individuals in a prior study (Erickson et al. 2012), and thus were selected here to encompass the widest range of taxonomical variability. The data revealed that microbiomes of these post-surgery individuals had significant variability in taxa, genes, and proteins; however, key metabolic modules associated with central metabolism were seen in all samples, even though the phyla of origin was often different. End-point metabolic modules, such as short-chain fatty acid production, were seen intermittently across samples, although only butyrate-producing modules were absent in one individual.

Specifically, there was evidence that there is persistent and phylum-redundant metabolic functional stability in these individuals. This approach provides a distinct observational viewpoint of following metabolic reactions and compounds/enzymes as a marker of overall microbiome activity, even if populations and specific genes/proteins vary.

4 METABOLIC REACTIONS INFERRED BY METAPROTEOMICS SHOW CONNECTED METABOLIC NETWORKS AND FUNCTIONAL DIFFERENCES BETWEEN INFANTS THAT DEVELOP NECROTIZING ENTEROCOLITIS AND THOSE THAT DO NOT

Authors of an upcoming manuscript

J. Alfredo Blakeley-Ruiz, Stephen Grady, Carissa Bleker, Michael Langston, Robert L. Hettich

J. Alfredo Blakeley-Ruiz contributions included metaproteomics measurement, data analysis, writing of the original manuscript, and revision

Special acknowledgment to Weili Xiong and Chris Brown for access to their dissertation data.

4.1 Introduction

Disruption of the gut's microbial composition is associated with a variety of metabolic and immunological diseases including diabetes (Stewart et al. 2018), obesity (Pfalzer et al. 2015), inflammatory bowel diseases (Erickson et al. 2012), irritable bowel syndrome (Rajilić-Stojanović et al. 2011), and necrotizing enterocolitis (NEC) (Stewart et al. 2016). Though specific taxa have been associated with these diseases, e.g., *Klebsiella* in the case of NEC (Olm et al. 2020), the inherent variability in gut microbiome composition across individuals makes it challenging to observe specific microbial taxa correlated with a disease consistently. While similar phyla are present across individuals, a recent investigation of over six hundred twins in the UK showed the average microbial species to be found in only two individuals (Visconti et al. 2019). Despite this inherent variability in composition, microbial function appears to be more conserved between individuals (Blakeley-Ruiz et al. 2019). An alternative and less explored approach is to examine changes in gut microbiome function with regards to these diseases regardless of taxonomic composition.

At birth, the gut microbiome is sparsely populated but rapidly increases in community complexity over the first two years of life before plateauing (Yatsunenko et al. 2012). Early in life, particularly in preterm infants, the microbial community remains more varied. Within an individual preterm infant, the gut microbiome can be as different from day to day as it is between individuals (Xiong et al. 2017). Despite this variance, preterm infants, because of their less complex gut microbiomes, have been a useful proving ground for exploring leading-edge

techniques to study the microbiome. The extraction of individual genomes from fecal sample metagenomes was developed and improved using preterm infants (Vieira-Silva et al. 2016). Metaproteomics techniques have been combined with genome-level resolution metagenomic techniques to explore protein expression in these genomes (Brooks et al. 2015, Young Jacque et al. 2015, Xiong et al. 2017). Mass spectrometry-based metaproteomics can detect and quantify thousands of microbial proteins from an environmental sample (Verberkmoes et al. 2008). Metaproteomics compliments metagenomics by measuring the functional activity of the cell, while metagenomics only measures the functional potential of the cell (Kleiner 2019). Metaproteomics also detects host proteins in conjunction with microbial proteins providing an avenue by which host function can be observed alongside microbial function (Xiong et al. 2017).

One of the most extensive integrated metagenomic and metaproteomic studies to date is a time-course study conducted on a cohort of preterm infants that did or did not develop NEC (Brown et al. 2018). Although *Klebsiella* was observed to be more abundant in NEC afflicted infants, no specific species detected in NEC was found to be in all infants that developed NEC. Several proteins, however, were found to be significantly different in abundance between infants that developed NEC than those that did not. A more recent analysis of just the metagenomic data from this study used gradient boosted machine learning to determine the most significant predictors of NEC. The top predictors found were bacterial replication rate, KEGG metabolic modules, and *Klebsiella*; however, the study did not go into much detail about the predictive metabolic modules (Olm et al. 2019).

This analysis seeks to explore further the microbial metabolic function of these preterm infants by exploring the specific metabolic reactions, coded by the KEGG ortholog annotations, of proteins previously measured in (Brown et al. 2018). Reaction to compound data from KEGG can, in turn, be used to build metabolic networks (Huss and Holme 2007). In this manuscript, a reaction is defined as the conversion of a specific set of one or more compounds into another specific set of one or more compounds as described by the KEGG reaction, compound, and glycan databases (Aoki-Kinoshita and Kanehisa 2009, Kanehisa et al. 2016). For example, R01688 represents butyrate kinase, which catalyzes the conversion of butanoylphosphate to butanoic acid, converting ADP to ATP in the process. In this analysis, it was hypothesized that combining metagenomic data with metaproteomic data could produce highly connected metabolic networks (Figure 2.3) and that the reactions that make up these networks would be

more conserved across individuals than the individual proteins or orthologous groupings that predict them. These reactions could be assigned to the host or the microbiome, and reactions supported by the literature could be shown to differentiate between NEC and non-NEC infants. Finally, the metabolic networks could be used to extract pathways of reactions over or under-represented in NEC.

4.2 Methods

4.2.1 Data acquisition

An integrated metaproteomics/metagenomics dataset describing 17 preterm infants over the first approximately 80 days of life was used for this study. Six of the infants eventually developed NEC. For each infant, between 3 to 12 samples were compiled for a total of 91 samples. All samples were collected between Day of Life (DOL) 6 to 86 and covered a gestational age range of 25 and 39 weeks. All metaproteomes were collected in technical duplicate using a 24 hour 11 salt-pulse 2D-LC-MS/MS method on an orbitrap-elite mass spectrometer (Xiong et al. 2017, Brown et al. 2018). Raw proteome data from all infants except one have been published previously (Xiong et al. 2017, Brown et al. 2018). Through email communication with the authors of (Brown et al. 2018), predicted protein sequences from 204 metagenomes with relevant metadata were compiled as well, between 5 and 22 per infant.

4.2.2 Protein database construction and annotation

Predicted protein sequences derived from all the metagenomes available for each were compiled into infant specific protein databases and clustered at 100% sequence identity using cd-hit to remove redundancy (Fu et al. 2012). All predicted proteins from all metagenomes were also compiled into a combined database and clustered at 100% sequence identity to create a sequence-specific identifier to link identical sequences between individuals. The scaffolds from the metagenomes had previously been binned into metagenome assembled genomes (MAGs) (Brown et al. 2018). These MAGs had been assigned a subspecies level taxonomy and had been grouped into subspecies groups using a 98% average nucleotide identity (ANI). The original scaffold of each predicted protein was retained in each protein identifier and used to match each protein to its subspecies group if one was available. Proteins with a subspecies group were assigned the genus of that subspecies group. Proteins with no corresponding MAG were assigned

as unknown since it is imprecise to determine the genus of origin of a single protein sequence without the context of other sequences from the same genome. KEGG orthology terms were assigned to protein sequences using the eggNOG database's internal annotator emapper (version 1.0.3-3-g3e22728), default parameters except for -m diamond, which uses the diamond instead of the hmmer algorithm to assign orthology (Kanehisa et al. 2016, Huerta-Cepas et al. 2017).

4.2.3 Protein inference

All mass spectrometry raw files were converted to mzML format using msconvert (proteowizard release 3.0.11252) (Chambers et al. 2012). Each infant specific microbiome protein database was combined with a human proteome, and a list of common contaminants to use for spectral matching. Each metaproteome sample was collected in technical duplicate, and each of these measurements was searched using the crux framework (v. 3.2-0bf523f) (McIlwain et al. 2014). The databases were indexed using tide-index (--missed-cleavages 2, --mods-spec C+57.02146,2M+15.9949 --clip-nterm-methionine T --decoy-format protein-reverse), the mass spectra were searched against the index using tide-search (--compute-sp T, --exact-p-value T, --score-function both, --precursor-window 10, --precursor-window-type ppm), results were quality scored using percolator (--spectral-counting-fdr 0.02, --test-fdr 0.02, --train-fdr 0.02, --protein T), and quantified with NSAF values using spectral-counts (--parsimony simple) (Käll et al. 2007, Serang et al. 2010, Diamant and Noble 2011). A protein was inferred if it had a least 1 unique spectrum at a peptide spectral match false discovery rate (FDR) q-value of less than 0.02 and a fido protein FDR q-value of less than 0.05. Technical duplicate measurements were combined by averaging the NSAF of proteins detected in both technical replicates. If the sample was only detected in one replicate, then the NSAF of that replicate was used.

4.2.4 Construction of graph database

Using the R package KEGGREST, reaction, compound, and glycan objects were downloaded from KEGG (Tenenbaum 2020). These were used to build a bipartite metabolic graph in which one partite set consisted of reaction, and the other consisted of compounds (compound and glycan objects). Vertices between partite sets were linked if the compound was a substrate or product of the reaction. For example, continuing with butyrate kinase, butanoic acid, C00246, and butanoylphosphate, C02527, were each linked to KEGG reaction R01688, and by extension, to each other. To this foundational graph, infants, samples, and identified proteins

were added (Figure 4.1). An edge was created between a protein and a reaction if they shared an EC or KO number, and an edge was added between a protein and a sample if a peptide from the sample had been mapped to the protein. Note that samples link to multiple proteins, and specific proteins may be linked to multiple reactions. For convenience, this graph was built and accessed as a labeled property graph database using Neo4J (<http://neo4j.org>).

4.2.5 Data Analysis

KEGG orthologs, reactions, and compounds associated with detected proteins were extracted from the graph database using Cypher and Py2neo (<https://py2neo.org/>). Using the Python pandas package (<https://pandas.pydata.org/>), proteins, KEGG orthologs, reactions, and compounds were assessed by how many samples/individuals in which they appeared and whether they were microbial or host. In addition, the Jaccard index of similarity was calculated between samples for microbial and host proteins, KEGG orthologs, reactions, and compounds ($J = (A \cup B) / (|A| + |B| - A \cup B)$) (Jaccard 1908). For each sample, a metabolic network was reconstructed by creating an edge list of reactions to compounds where a protein associated with a given reaction was present in the samples. To remove high degree connections that do not reflect the flow of carbon, a list of currency metabolites, such as ATP, was used to remove those edges. These reactions-to-compound graphs were analyzed for connected components to identify fully connected paths, and bipartite projections were used to create compound-to-compound and reaction-to-reaction networks using the python-igraph package (<https://igraph.org/python>). The number of reactions and proteins over time relative to NEC was compared using linear mixed-effects models from the R package (lme4), and ANOVA from the R package (lmerTest). Reactions were quantified in each sample by summed NSAF of associated

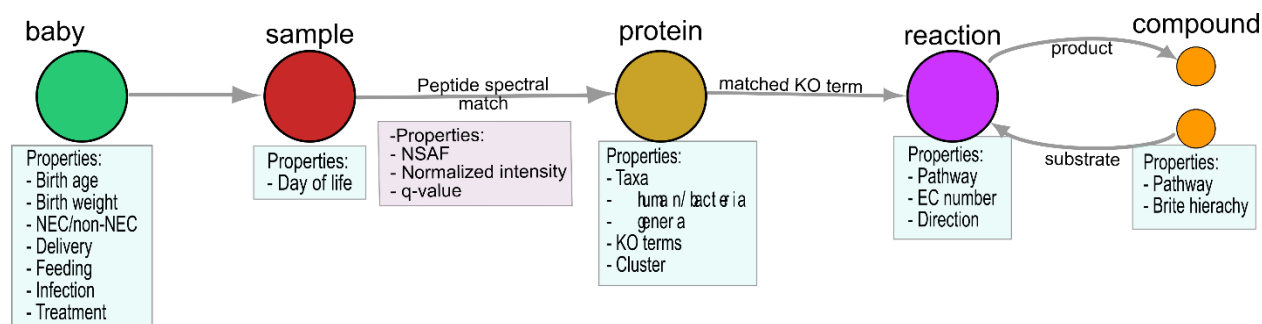


Figure 4.1: Structure of baby-sample-protein-reaction-compound graph database

proteins and by counting the number of proteins detected for each reaction in each sample. Each of these metrics was normalized by $\log_2$ transformation followed by subtracting the mean for each sample. Reactions were then compared between NEC and non-NEC infants using a linear mixed-effects model, `lmer (Reaction~NEC*Time + (1|Individual))`, and corrected by Benjamini-Hochberg procedure ($Q < 0.25$ for all p-values). Figures were rendered using a combination of python and R packages `seaborn`, `pandas`, `ggplot2`, and `jtools`. In-house python scripts conducted any data processing not associated with a published bioinformatics tool.

4.3 Results

4.3.1 *Reactions are substantially more conserved between samples than protein sequences*

To build a comprehensive database for protein identification, predicted protein sequences were compiled from 204 metagenomes independently sampled, sequenced, and assembled from 17 preterm infants, 6 of whom eventually developed NEC (Brown et al. 2018). The combined protein database of these metagenomes contained a total of 4,350,713 distinct protein accessions, which clustered into 804,297 non-redundant protein sequences at 100% sequence identity (Li and Godzik 2006). Variability between samples at the metagenome level was high. The median number of individuals that a sequence was observed in was one, and the median number of samples a sequence was observed in was two (Figure 4.2a).

Based on this information, instead of searching against all sequences from all metagenomes, the metagenomes specific to each individual were compiled to form individual specific microbial protein databases clustered at 100% amino acid identity to remove redundancy. These individual specific databases were combined with human and contaminant databases, also clustered 100% identity, and then searched against 182 technical duplicated metaproteomic 2D-LC-MS/MS measurements from 91 samples. In total, 439,434 non-redundant proteins were identified, of which 415,448 were microbial (52% of the total non-redundant database); however, only a median of 17,794 microbial and 6,996 host proteins was identified per sample, and, the median microbial protein was only observed in three samples and one individual (Figure 4.2c).

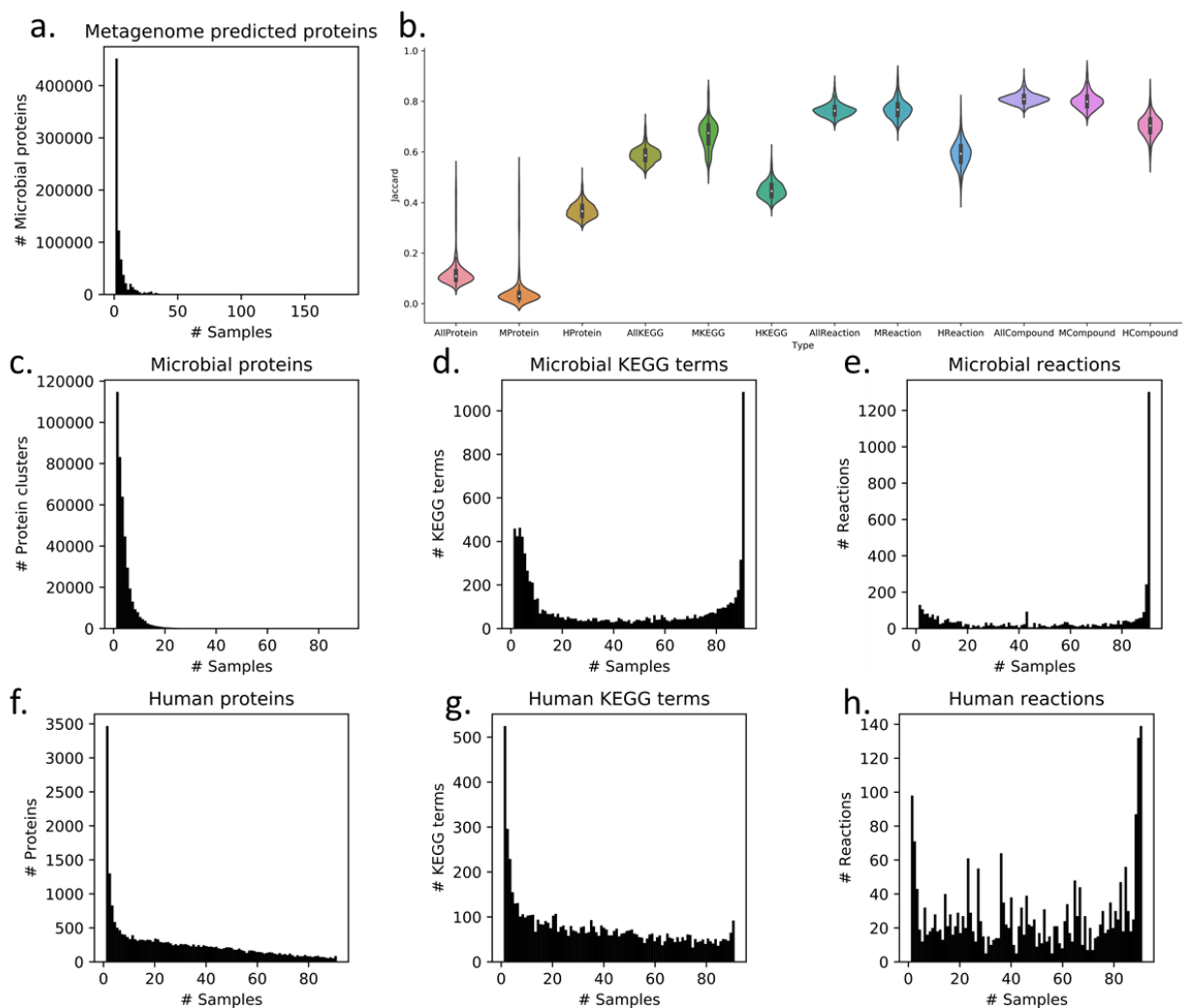


Figure 4.2: Observed genes are radically different across samples, but reactions are more conserved. (a) Histogram shows the number of samples in which each non-redundant protein was sequenced in the metagenome. (b) Violin plot of the Jaccard index of similarity between samples at the protein, KEGG, reaction, and compound level. (c-h) Histograms show the number of samples each human or microbial protein, KEGG term, and reaction was observed in.

To confirm that KEGG orthologs were more conserved than proteins across samples, and to show that reactions were further conserved, microbial proteins were annotated with KEGG ortholog terms using the emapper pipeline from the eggNOG database (Huerta-Cepas et al. 2016, Huerta-Cepas et al. 2017). These KEGG terms (KO) were used to connect protein identifiers to reactions, and reactions to associated compounds they convert. Together all this data was used to build a baby-sample-protein-KO-reaction-compound graph database (Figure 4.1).

Identified proteins were annotated with a total of 14,013 KEGG orthologous groups, of which 1,882 were shared between human and microbial proteins, and 8,845 were exclusive to microbial proteins. These KO terms pointed to a total of 4,933 reactions, of which 1,813 were shared between proteins from the microbiome and the host, and 2,435 were unique to proteins from the microbiome. The median microbial KO term were observed in 37 out of 91 samples, and 12% were observed in all samples. The median microbial reaction were observed in 76 out of 91 samples, 30% were observed across all samples. The compounds catalyzed by these reactions were even more conserved, with the median compound found in 86 samples and all individuals, and 40% of compounds having an associated enzyme in all samples, suggesting multiple types of reactions converting the same compounds. Thus, in a three-step transition from identified sequence to a specific function, one can observe that specific metabolic functions are substantially more conserved across samples than their protein sequences of origin (Figure 4.2c-e). A similar but less extreme trend is observed with the human proteins (Figure 4.2f-h). This makes sense since reactions from the host are less likely to be attached to multiple proteins than reactions from the multiple organisms in the microbiome. Together, these trends support the hypothesis that discrete metabolic function rather than protein sequence information may be a meaningful approach for the analysis of sparse microbiome data.

To look at the pairwise similarity in detected proteins, KO terms, reactions, and compounds between samples, the Jaccard similarity index was calculated between samples for both the host and microbial versions of each of the metrics (Figure 4.2b). A similar trend was observed with samples sharing a median 3% of microbial proteins while sharing >70% of reactions with each other. Host proteins were less extreme, sharing a median 36% of host proteins between samples and a median 59% of host reactions between samples. These results show that the trends observed above hold on a pairwise basis, and reactions are indeed more

likely to be shared between samples, thus suggesting a more stable space for quantitative comparison between samples.

4.3.2 KEGG annotation of detected proteins enables the generation of connected metabolic networks

A key goal of this project was to determine if metabolic pathways could be reconstructed *de novo* using reaction to compound data derived exclusively from the proteomes in any given sample. A bipartite reaction-to-compound graph was created for each sample based on the reactions associated with the annotations of each detected protein. These networks each had a dominant single connected component. Many of these connections were driven by so-called currency metabolites, such as ATP, ADP, NAD, NADH, etc. A list of 37 of these types of metabolites was compiled and removed from the graphs. While this did indeed break some of the connections, the resultant graphs maintained the same structure with a single dominant connected component (Figure 4.3a). Using a bipartite projection, each of these bipartite graphs was collapsed into a reaction-to-reaction graph and a compound-to-compound graph (Figure 4.3b and 4.3c). By doing a connected component analysis of each graph, it was determined that >80% of the compounds in each graph were present within the single largest connected component, and >86% of the reactions were present in the single connected component (Figure 4.3d). This result indicates that beyond being able to reconstruct pathways, fully connected metabolic networks can be reconstructed exclusively using metaproteomic data. This offers both the potential of pathway reconstruction of up-regulated functions and the possibility of metaproteome guided metabolic modeling. Further analysis of these networks found that the maximum distance between any compound in the compound to the compound connected component was 20, suggesting a maximum pathway length of 20 between any two compounds.

4.3.3 Network analysis reveals numerous host-microbe connections

To explore the overall dynamics of the metabolic interaction between the microbiome and the host, reactions were divided into reactions that were shared between the host and the microbiome, reactions specific to the host, and reactions specific to the microbiome. 13% of reactions were found to be unique to the host in all samples, while 36% were shared between the host and the microbiome, and 49%, were unique to the microbiome (Figure 4.4). Many reactions

that were shared, however, were not shared in all samples. Within any sample, the percentage of reactions exclusive to the microbiome ranged from 45% to 75%, and the percentage of reactions exclusive to the host ranged from 9% to 25%. The percentage of reactions associated with the microbiome significantly increased over time, while the percentage of reactions shared or unique to the host decreased over time. Interestingly, further analysis of these results showed an opposite trend for infants that eventually developed NEC, with the percentage of microbial reactions decreasing over time, while the percentage of host proteins increased over time (Figure 4.5). This result was only observed when Baby 61, the only NEC infant with data past day of life 80, was included in the analysis.

To determine if host, shared, or microbial reactions were more likely to be connected to a similar reaction type or a different reaction type, the number of host-host, host-microbe, host-shared, or microbe-shared edges in each sample's reaction-reaction network was calculated. On average 26% of edges containing a host reaction were found to be host-host, while 52% were host-microbe. In contrast, on average, 51% of edges with a microbial reaction were microbe-microbe, while 18% were host-microbe. Exploring these trends among infants that did not develop NEC over time revealed a significant increase in the percentage of microbe-microbe edges relative to the number of total microbe edges (Figure 4.5). Similarly, the percentage of host-microbe edges relative to total host edges also increased over time. Conversely, the percentage of host-microbe edges relative to total microbial edges and the percentage of host-host edges relative to total host edges decreased overtime. Interestingly, infants that developed NEC showed an opposite trend, but only when baby 61 was included in the analysis. This result provides further evidence that ordinarily, the host cedes functionality overtime to the microbiome, and that NEC may have a different trajectory.

To quantify the redundancy of reactions across taxa and samples, the number of genera including the host associated with each reaction was calculated both across samples and within each sample. Nineteen percent of all reactions were exclusive to one genus when considering both host and microbial reactions. This number dropped to six percent when host-specific reactions were excluded, and reactions shared between host and microbiota were included. Shared reactions are not exclusive to one genus since they include the genus *Homo* and at least one other microbial genus. When these shared reactions were excluded as well, it was found that eleven percent of reactions were unique to at least one genus across all samples. Together this

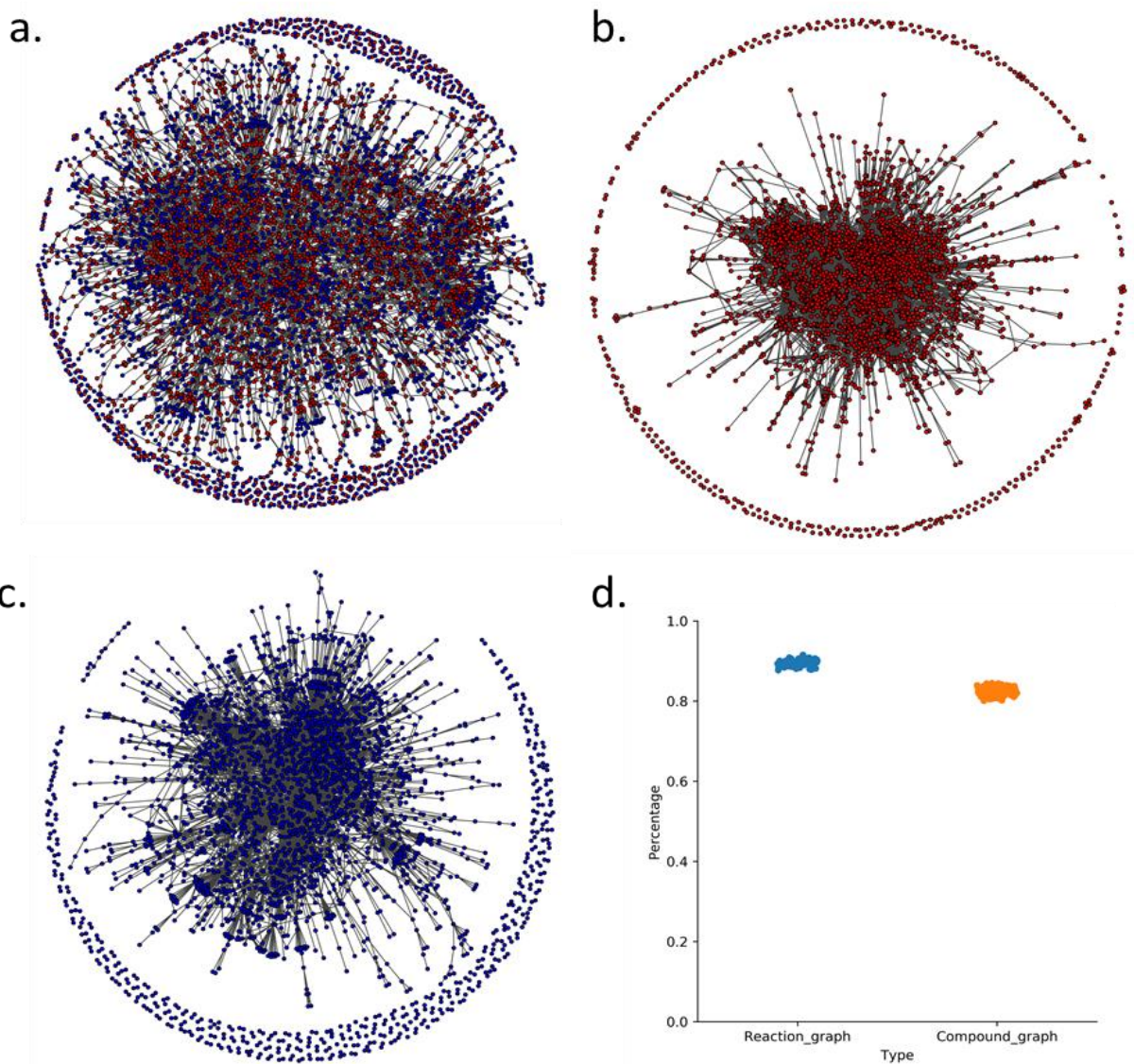


Figure 4.3: Reaction to compound metabolic networks reveal primarily a single connected component. (a) Representative compound to reaction graph for a single sample, S_3_25.1. (b) Reaction to reaction graph of the same sample after a bipartite projection. (c) Compound to compound graph of same sample after a bipartite projection. (d) Categorical dot plot depicting the number of nodes in the largest component divided by the total number of nodes for both the compound and reaction graphs. Red nodes represent reactions and blue nodes represent compounds.

implies that across all the metabolic networks, only 6% of reactions were unique to specific reactions across time, suggesting that most reactions can be catalyzed by more than one genus.

Temporally, a greater percentage of reactions were unique to a single genus within any one sample. A mean 35% of reactions in any one sample were unique to a single genus when considering the host, and a mean 22% of reactions were unique to a single genus when excluding reactions derived only from host proteins. Analyzing these two data sets, it seemed that overall, the number of reactions unique to one genus decreased across time (Figure 4.5). When including the host, this appeared to be driven by infants that did not develop NEC, providing further evidence for the idea that the host cedes function to the microbiome as time progresses. Interestingly, when excluding host-specific reactions, an opposite trend was observed, where the downward trend was driven by infants that developed NEC. Potentially this could be driven by several microbial genera responding to the gut conditions that lead to NEC.

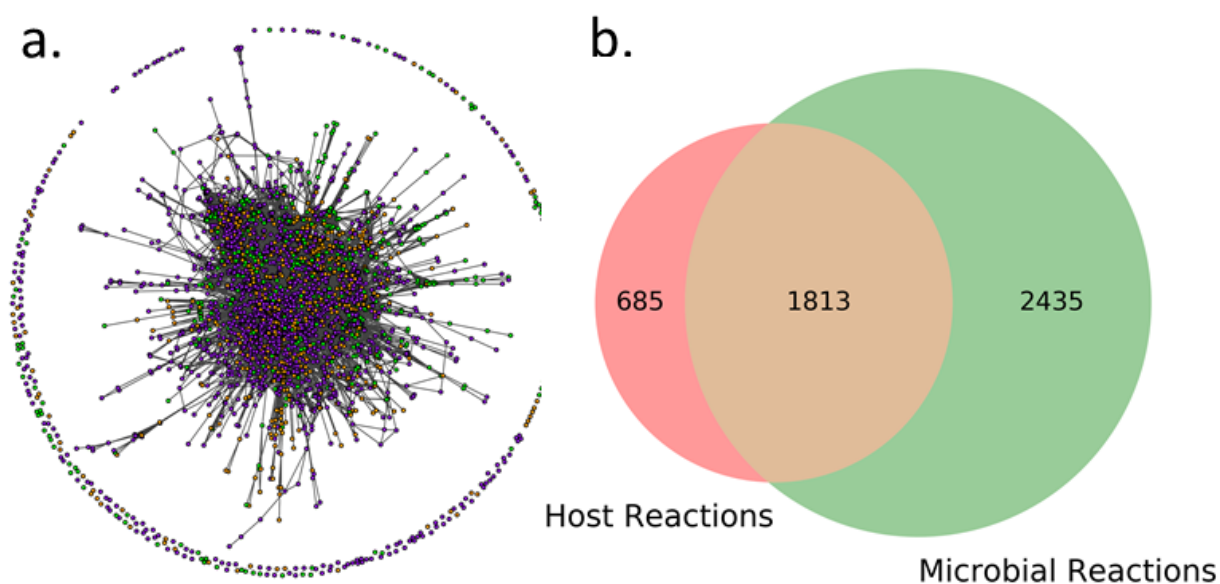


Figure 4.4: A substantial number of reactions are shared between host and microbiome. (a) Reaction to reaction graph for S_3.25.1 showing microbial (purple), shared (orange), and host (green) reactions. (b) Venn diagram showing number of reactions that are microbial, shared, or host across all samples.

4.3.4 *Linear mixed-effects models reveal reactions associated with the development of Necrotizing Enterocolitis*

To explore the functional differences between infants that developed NEC versus infants that did not, KEGG reactions with protein evidence were quantified via two methods. First, the NSAF abundance of all proteins associated with KEGG reactions was summed for each sample, referred to as $NSAF_k$. Second, the number of proteins detected for each KEGG reaction was counted for each sample, referred to as PN. Both quantifications were similarly normalized, and missing values imputed. Though these two metrics may appear similar, they measure different things. $NSAF_k$, as a protein abundance metric, measures at least three things: number of proteins expressed that are annotated with the reactions, the abundance of the organisms producing the proteins, and the relative abundance of the protein within the organisms. PN measures only one thing, the number of distinct proteins sequences above the limit of detection, annotated with the reaction. Thus, $NSAF_k$ is an abundance metric driven by multiple factors, while PN is more reflective of a bulk microbiome functional transition and this difference was visible in the results.

To account for random effects caused by multiple measures from the same individual and the effect of time, each reaction was tested using a linear mixed-effects model with NEC and time as interacting fixed effects, and the infant of origin as a random effect. All p-values were compared using a Benjamini-Hochberg procedure, and $p < 0.05$ was found to be equivalent to a $Q < 0.25$. This was set as the significance cutoff along with an absolute $\log_2$ fold change of 0.5 or higher.

This approach allowed the identification of reactions that significantly increase or decrease over time. Using the PN, 2,583 reactions were found to be significantly increasing or decreasing across time, and 301 were found to be differentially changing across time between NEC and non-NEC. Among reactions that were significantly changing over time, all but six decreased in the relative number of proteins across time. This suggests a general trend toward decreasing the relative number of proteins per reaction over time. An explanation for this could be the increase in the number of reactions across time, while the number of proteins remains constant, suggesting fewer proteins available for each reaction. Using $NSAF_k$, there was a more even distribution, 885 reactions were seen to be increasing or decreasing across time, of which 39% were decreasing over time. Reproducing the results observed in Brooks et al. 2015 for one

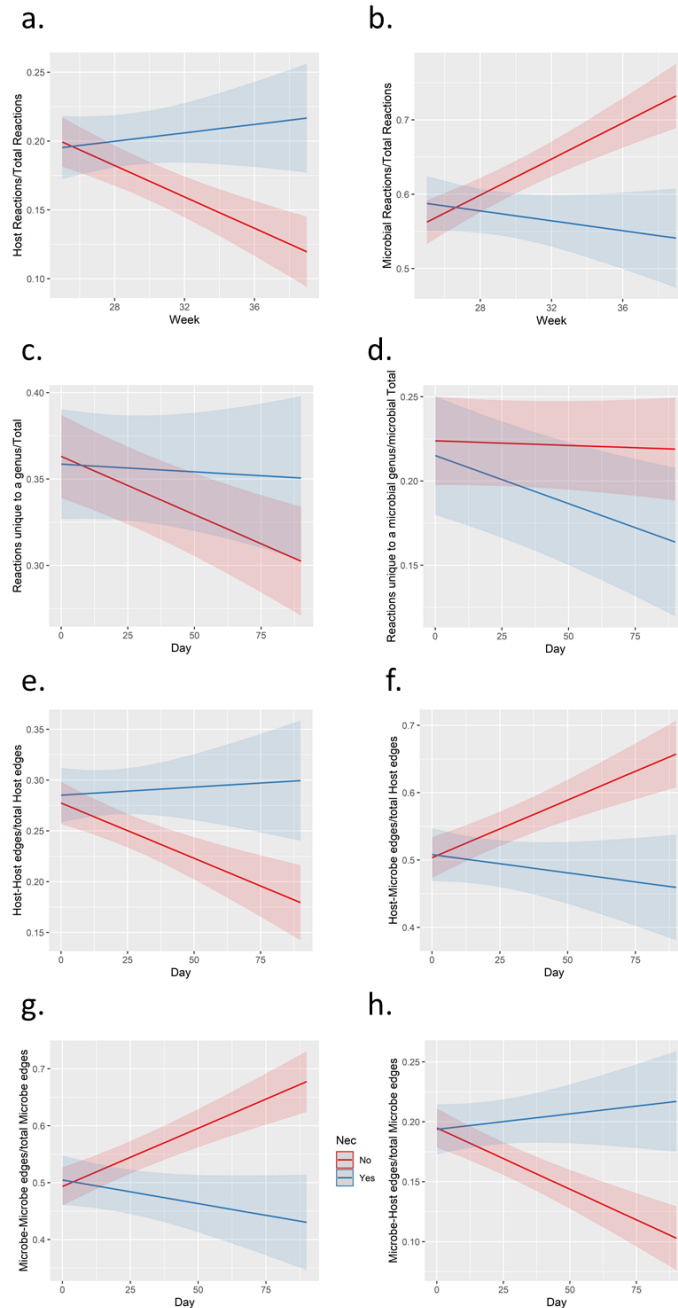


Figure4.5: Ratio of host to microbial reactions changes across time. Regression plots showing a variety of metrics across time shaded region is 95% confidence interval as predicted by jtools. (a) Ratio of host reactions relative to total reactions across time. (b) Ratio of microbial reactions relative to total reactions across time. (c) Number of reactions unique to a genera/relative to total reactions across time. (d) Number of reactions unique to a genera, when only considering microbial or shared reactions. Using all the reaction to reaction networks it was calculated how many edges were microbe to microbe or host to host. (e) Number of host to host edges relative to total number of edges with a host reaction. (f) Number of host to microbe edges relative to the total number of edges with a host reaction. (g) Number of microbe to microbe edges relative to total edges with a microbial reactions. (h) Number of microbe to host edges relative to total number of edges with microbial reactions.

infant, reactions associated with beta-oxidation decreased across time, while reactions associated with glycolysis increased across time, suggesting a transition from aerobic to anaerobic metabolism (Brooks et al. 2015). The primary purpose of this analysis, however, was to determine reactions that were consistently more or less abundant across time in infants that develop NEC relative to those infants that did not develop NEC. The remainder of the analysis will focus on those.

When considering reactions significantly distinct between NEC and non-NEC infants, different but complementary trends were observed between the two-abundance metrics (Figure 4.6). Using NSAF_k, 347 reactions were observed to be different in abundance between NEC and control, of which 95% were more abundant across time in non-NEC infants than in infants that developed NEC, and 68% were host or shared between the microbiome and the host. Using PN, 107 reactions were observed to be significantly different between non-NEC and infants that eventually developed NEC. Ninety-seven percent of these had more relative protein in NEC than in non-NEC, and seventy-one percent were exclusive to the microbiome. This difference may be driven by the inherent differences in what is being quantified between the two metrics.

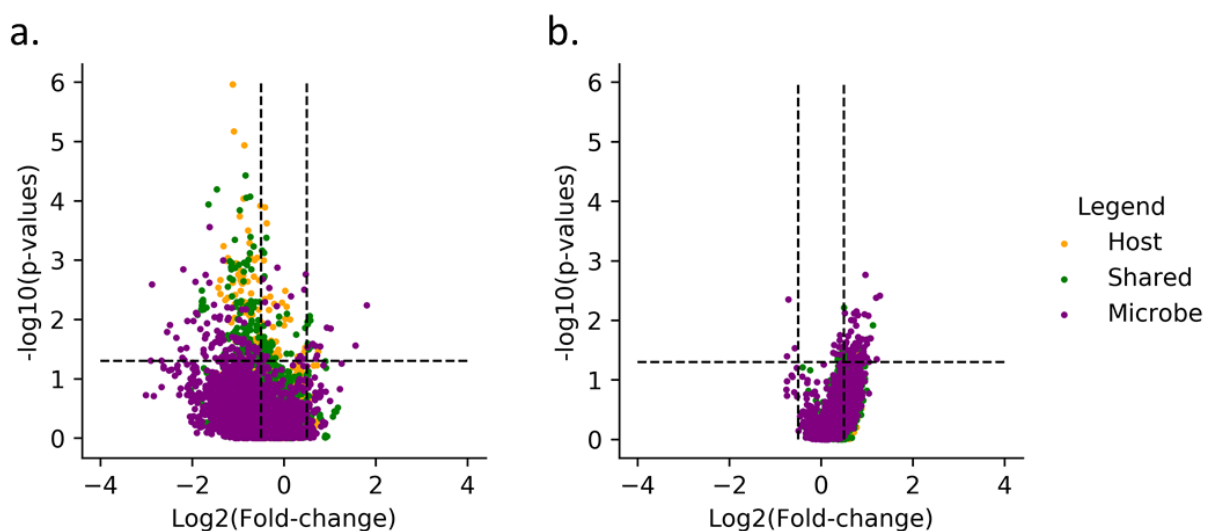


Figure 4.6: Reactions significantly different between NEC and non-NEC. (a) $-\log_{10}(\text{p-values})$ versus $\log_2(\text{fold-change})$ for reaction abundance calculated via summed NSAF. (b) $-\log_{10}(\text{p-values})$ versus $\log_2(\text{fold-change})$ for reaction abundance calculated via relative number of proteins.

4.3.5 Pathway reconstruction reveals differences between NEC and non-NEC

To explore the relationship between reactions significantly different between NEC or non-NEC using either PN or NSAF, a reaction-to-compound network was created for each of the two metrics (Figure 4.7 and Figure 4.8). These two networks revealed several significant host, shared, and microbial reactions connected by similar compounds that either indicated a pathway, multiple reactions acting on the same compound independent of a pathway, or singleton reactions with no other connected reactions. The networks tended to follow the trends observed in Figure 4.6, with the NSAF_k network reflecting more host reactions and reactions that were up among healthy infants, while the PN network reflected more microbial reactions that also had more relative proteins in NEC rather than non-NEC infants. Upon manual inspection, several reaction groupings of interest were uncovered.

In the network indicative of reactions significantly different between NEC and non-NEC by NSAF_k, host reactions associated with the biosynthesis of mucin and bile acids were more abundant in healthy infants than those that developed NEC. Several groupings of reactions were

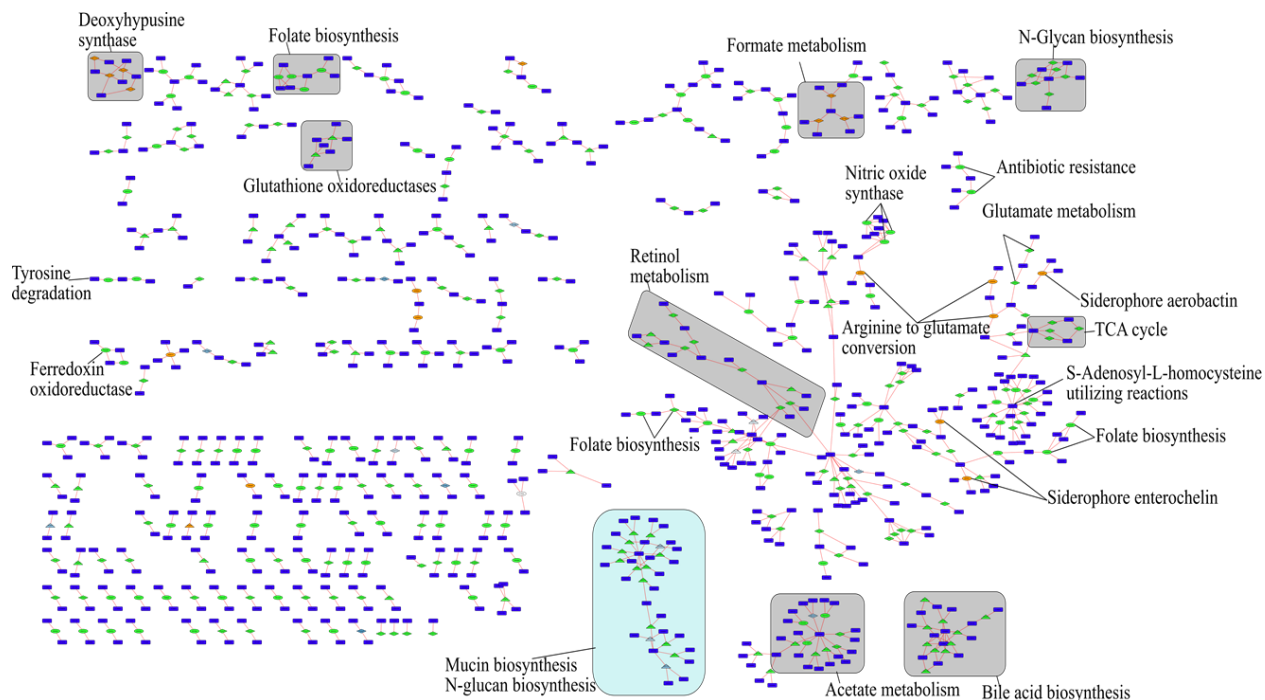


Figure 4.7: Reaction to compound network made up of all the reactions that are significantly different between NEC and non-NEC using summed NSAF abundance. Green nodes are up in infants that do not develop NEC and orange nodes are up in infants that develop NEC. Square blue nodes are compounds, circle nodes are microbial reactions, diamond nodes are shared reactions, and triangle nodes are host reactions.

shared or a mix of host, shared, and microbial reactions. Among these, deoxyhypusine synthase, acetate metabolism, and retinol metabolism appeared to be up in infants that did not develop NEC, while formate appeared to be up in infants that developed NEC. Further inspection of deoxyhypusine synthase, however, revealed that this trend was driven by host proteins as the shared aspect was driven by proteins unknown origin in baby 70, one of the healthy infants. Among exclusively microbial reactions, nitric oxide synthase, folate biosynthesis, and antibiotic resistance appeared to be up in healthy infants. In contrast, arginine degradation and the biosynthesis of siderophores enterochelin and aerobactin were found to be more abundant in infants that developed NEC.

In the PN, microbial reactions involved with the biosynthesis of siderophores were also observed; however, in this case, all the reactions between chorismate and enterochelin were found to have more relative proteins in NEC than in non-NEC babies indicating a strong siderophore signal, and the value of using multiple methods for functional quantification. Apart from the siderophore biosynthesis, microbial reactions associated with the catabolism of amino acids and lipoyl synthase were found to have more relative protein in NEC than non-NEC. Quinone oxidoreductase was found to have more relative microbial proteins in infants that did not develop NEC. Reactions associated with the enzyme lactosylceramide alpha-2,3-sialyltransferase, which converts lactosyl ceramides to glucosylceramides or extends

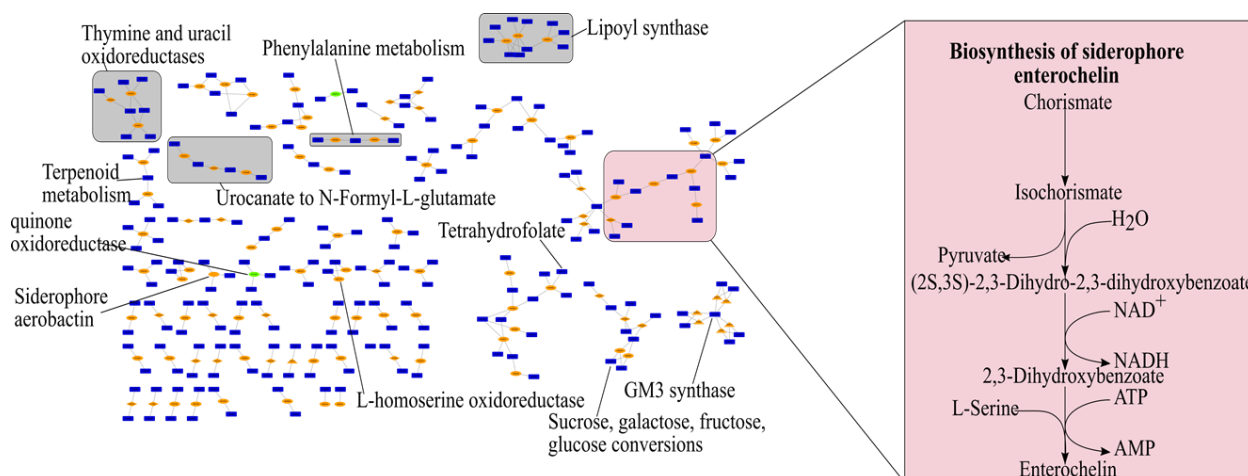


Figure 4.8: Compound to reaction network of reactions that are significantly different in abundance using relative numbers of proteins. Green nodes are up in infants that do not develop NEC and orange nodes are up in infants that develop NEC. Square blue nodes are compounds, circle nodes are microbial reactions, diamond nodes are shared reactions, and triangle nodes are host reactions. In addition, pathway from chorismate to enterochelin, a siderophore, is depicted.

glucosylceramides, had more relative host proteins in infants that developed NEC than those that did not.

4.4 Discussion

Previous studies have shown that, though highly variable in composition, gut microbiomes are much more conserved in function using either broad functional categories, metabolic modules, or orthologous groupings (The Human Microbiome Project 2012, Blakeley-Ruiz et al. 2019). This study takes these findings one step further, demonstrating that the most specific entities of metabolic functions, the reactions themselves, are more likely to be shared between samples in preterm infants than orthologous groupings. This indicates that specific function rather than protein sequences may be a more stable space for analysis of microbiome functions than orthology or individual proteins in highly variable microbiomes.

Linked together by the compounds that they catalyze, these reactions were used to reconstruct metabolic networks. Though metaproteomes have previously been used to reconstruct metabolic networks and even model metabolism (Tobalina et al. 2015), this study revealed that these reaction networks could be used to consistently build sample-specific metabolic networks dominated by a single connected component even after the removal of currency metabolites like ATP. Although a flux balance analysis was beyond the scope of this study (Orth et al. 2010), this study showed that reasonably deep metaproteomes are a powerful tool for guiding comprehensive metabolic modeling of microbial systems.

A previous study successfully reconstructed 773 genome-scale metabolic models of bacteria derived from human gut microbiomes (Magnúsdóttir et al. 2017). After comparing these models to the human genome, it was found that approximately 8% of metabolic reactions were shared between the host and the reconstructed microbial models. Due to the relatively high percentage of human proteins in preterm infants relative to adults, the metaproteomes in this study yielded fairly deep human and microbiome proteomes allowing for the concurrent analysis of reactions of microbial, host, or shared origin. In this study, approximately 36% of the total reactions detected were shared between the host and the microbiome, while only 13% were unique to the host showing that even in less complex microbial communities in the gut, the microbiome contributes the greater share of metabolic functional capacity. Within any given sample, these percentages shifted towards more microbial reactions at the expense of host

reactions over time. This was further reflected in an increase in microbial edges over time relative to host edges. Though NEC infants showed an opposite trend, this was primarily driven by the single infant whose sampling extended beyond day 80 after birth.

Division of detected reactions into host, shared, and microbial groupings allowed for the concurrent analysis of host and microbial functions associated with the development of NEC or the absence of NEC. This analysis was conducted by two metrics: normalized summed NSAF abundance, $NSAF_k$, and the relative number of proteins, PN, and these metrics offered complimentary information. Reactions uncovered by these two metrics provided a variety of insights into gut functionality in necrotizing enterocolitis. Among microbial reactions that were less abundant in NEC were reactions involved in nitric oxide synthase, folate biosynthesis, and quinone oxidoreductase. Nitric oxide is involved in healthy construction of the mucosal layer (Brown et al. 1993), suppression of inflammation, protection against gastric damage (Calatayud et al. 1999), and previously shown to be correlated with healthy tissues in Necrotizing enterocolitis (Nowicki et al. 2006, Lanas 2008). Folates are critical co-factors for the biosynthesis of a variety of molecules in the host (Rossi et al. 2011). A major source of folates are bacteria in the gut microbiome (Engevik et al. 2019), and its deficiency has been associated with microbiome dysbiosis (Pan et al. 2017). A probiotic with a quinone oxidoreductase has been shown to recover the microbiome from dysbiosis (Pan et al. 2019).

In the case of microbial reactions that were up in NEC, several reactions associated with the degradation of amino acids were observed, and infants with NEC have been shown to have lower amino acid serum concentrations (Becker et al. 2000). The most interesting microbiome signal that was up in NEC was the biosynthesis of siderophores. Siderophores are iron-chelating molecules that assist microbes in iron harvesting (Wandersman and Delepelaire 2004). Brown et al. 2018 previously detected iron harvesting proteins as up in NEC as well (Brown et al. 2018); however, in the current study, the entire pathway from chorismate to enterochelin was visible as up in NEC, suggesting multiple organisms taking up this function in the lead up to necrotizing enterocolitis. Infants with NEC have been shown to have a predisposition for being unable to uptake iron as efficiently as infants that do not develop NEC (Göpel et al. 2018). Iron homeostasis dysregulation is closely associated with inflammation, and this has been shown to lead to a higher fecal iron load when colitis is induced in mice (Semrin et al. 2006, Wessling-Resnick 2010, Verma and Cherayil 2017, Mahalhal et al. 2018). Increased iron harvesting is

closely associated with pathogen virulence and has been suggested as a potential target to combat infections (Page 2019). Treatment with lactoferrin, an iron-binding protein, has also been shown to ameliorate NEC (Raghuveer et al. 2002, Telang 2018). Interestingly chorismate is a major source of folate in the gut, suggesting reduced folate biosynthesis may be a result of chorismate being repurposed towards siderophore production (Engevik et al. 2019). Together these results indicate disruption iron homeostasis may be involved in triggering microbial virulence leading to NEC.

On the host side, the mucosal layer forms an important protective layer preventing the microbial invasion of the host; however, in this study reactions associated with biosynthesis of mucin were down in NEC suggesting a disruption in production of mucin in infants that develop NEC, which could offer a host explanation for the disruption of the mucosal layer in NEC (Hodzic et al. 2017). Less clear cut, high bile acid load is associated with NEC (Halpern and Dvorak 2008, Hulzebos et al. 2017), but the results in this study indicate that bile acid production is down in NEC. One explanation could be that since infants that develop NEC are predisposed to iron uptake deficiency, possibly NEC infants are also not as effective as taking up bile-acids leading to bile-acid accumulation and a decrease in bile acid production. A further study showed higher variance in bile-acids available, suggesting a disruption of bile acid homeostasis rather than a general increase (Knapp et al. 2020).

Together these results present a picture of the crosstalk between the host and microbiome in the formation of NEC. In the lead up to the development of NEC, the host is predisposed to not uptake iron and bile acids. This leads to disruption of bile-acid and iron homeostasis, which disrupts folate biosynthesis and mucosal layer maintenance. This opens the door for invasion by pathogenic iron harvesting bacteria. These results, however, have some limitations. The identified reactions are putative annotations created by orthology based functional transfer. Each protein could point to many reactions leading to some overlap in quantification that may be indicative of more than one reaction. The experimental setup with only six NEC infants and non-consistent time-point sampling makes the results predictive, but not confirmatory. Nevertheless, this study presents a strong case for the use of reactions when comparing metaproteomes between conditions when exploring host and microbiome functions concurrently. Since adult metaproteomes contain substantially less host protein relative to microbial protein, this analysis

presents an argument for concurrent proteome measurement of colon tissue and fecal material to create a comparable host/microbiome functional repertoire.

5 MORPHINE ADMINISTRATION DIFFERENTIALLY ALTERS GUT MICROBIOME COMPOSITION AND FUNCTION IN LEAN AND OBESE MICE

Authors of an upcoming manuscript

J. Alfredo Blakely-Ruiz, Carlee S. McClintock, Him K. Shrestha, Suresh Poudel, Zamin K. Yang, James J. Choo, Richard Giannone, Mircea Podar, Helen A. Baghdoyan, Ralph Lydic Robert L. Hettich

J. Alfredo Blakeley-Ruiz contributions included coauthor of grant, listed co-investigator on grant, experimental design, experimental measurements, data analysis, writing of original manuscript, and revision

5.1 Introduction

Gut microbiome disruption correlates with altered metabolism (Donohoe et al. 2011), chronic pain (Rajilić–Stojanović et al. 2011), and increased inflammation (Lloyd-Price et al. 2019). Obesity also correlates with increased inflammatory cytokines (Poret et al. 2018, Elisia et al. 2020) and chronic pain (Hitt et al. 2007), while disrupting the gut microbiome (Ridaura et al. 2013). Obesity triggered disruption of the microbiome manifests as a decrease in the overall richness/diversity of the gut microbiome and a decrease in the abundance of Bacteroidetes bacteria relative to Firmicutes bacteria (Ley et al. 2005, Turnbaugh et al. 2009, Le Chatelier et al. 2013); however, some of these findings have proven difficult to reproduce in meta-analyses using body-mass-index (BMI) as the delimiter between lean and obese (Finucane et al. 2014, Walters et al. 2014, Sze and Schloss 2016). Obesity, however, is not one type of disease, and general metrics for obesity like BMI do not capture the individual nuances of different types of obesity, such as genetic versus diet-induced obesity. Different kinds of obesity affect the microbiome distinctly both in the magnitude of the effect and in the specific taxonomical changes in composition. This could explain some of these inconsistencies as genetic obesity appears to increase the weight of mice more rapidly than high-fat diet, while high-fat-diet induced obesity appears to have a greater magnitude impact on the microbiome (Carmody et al. 2015, Pfalzer et al. 2015).

Opioids are a standard treatment for chronic pain, yet opioids also have a significant effect on the gut microbiome and specific inflammatory cytokine markers. Specifically,

morphine has been shown to increase the abundance of *Enterococcus* bacteria and the cytokine IL-6, while suppressing the abundance of *Lachnospiracea* bacteria (Banerjee et al. 2016, Acharya et al. 2017, Wang et al. 2018). The above information indicates a relationship between obesity and opioid use through their correlations with microbiome disruption, chronic pain, and inflammation; however, to the authors' knowledge, no study has looked at the effects of morphine on the microbiome in the context of obesity and no study has looked at the effects of morphine on microbiome function.

Given the coexisting epidemics of obesity and opioid abuse, it is of paramount importance to seek a better understanding of the mechanisms by which obesity and opioids interact to promote inflammation and alter gut microbiomes composition and function (Figure 5.1). Using a combined metagenomics and metaproteomics experimental design, the present study sought to determine the impact of chronic morphine administration on the gut microbiome composition and function in mice with genetically endowed versus diet-induced obesity.

5.2 Methods

5.2.1 Study Design and Sample Collection

Adult male db/db (Stock# 000697; n=6), wild-type B6 (Stock# 000664; n=6), and B6 mice with diet induced obesity (DIO) (Stock No: 380050; n=6) were purchased from The Jackson Laboratory (Bar Harbor, ME). The mouse feed was purchased from Research Diets, Inc., (New Brunswick, NJ, USA). The B6 and db/db mice were fed normal Teklad 8640 rodent chow containing 5% fat (Envigo, Madison, WI), the DIO mice were fed a 60% fat diet (Research Diets catalog number D12492). Environmental housing conditions for all mice included a 12:12 light:dark cycle, maintenance of thermo-neutral room temperature, and free access to food and water.

After a post-shipment acclimation period, mice were weighed, anesthetized with isoflurane, and implanted with Alzet osmotic pumps delivering 10 mg/kg/day of morphine sulfate, at a rate of 0.5 μ L/h, or saline (vehicle control) and then maintained for two weeks. All mice were weighed daily and individually housed for 24 hours in metabolic cages to collect feces and urine. The urine samples were used to confirm morphine metabolism. Finally, all mice were anesthetized with isoflurane and terminal samples (blood plasma, colon tissue, and colon fecal) material were surgically taken from all 18 mice at the end of the study. Terminal blood was

collected from right ventricle in EDTA-treated collection tubes then centrifuged at 3000 rpm. Plasma was transferred to clean tubes and frozen at -80°C. Colon tissue was resected, fecal material evacuated, then both sample types were frozen immediately upon transfer to clean tubes and stored at -80°C.

5.2.2 Protein Extraction

Feces extracted by metabolic cage or directly from the colon were treated identically. Between 8-50 mg of fecal material were placed in an eppendorf tube. Approximately 200 μ L of 0.115 mm zirconium oxide beads were added to each tube, followed by 350-400 μ L of lysis buffer (2% SDS in 100mM Tris-HCL buffer) in each tube. Samples were lysed by bead beating in a Next Advance bead beater at speed 10 for 5 minutes. Samples were centrifuged at 21,000 g for 10 minutes. Next, 200 μ L of supernatant was pipetted for protein extraction, and the remainder (with beads) were used for DNA extraction. Protein extraction supernatant was heated at 85°C for 10 minutes, incubated with 25mM DTT for 20 minutes, and then incubated with 75 mM IAA for 20 minutes. Protein was then extracted using a chloroform methanol extraction protocol: addition of 800 μ L of methanol, 200 μ L of chloroform, and 600 μ L of water followed by 10 minutes of centrifugation at 21,000 g. Supernatant was disposed of and remaining pellet was washed using methanol. Dried pellet was reconstituted in sodium deoxycholate (SDC) solution made up of 4% SDC in 100 mM ammonium bicarbonate. Protein concentration was measured using bicinchoninic acid (BCA) assay. Each sample was adjusted to be 250 μ g of total protein in 500 μ L solution. Peptides were digested twice using a 1:75 (wt/wt) trypsin ratio, once for three hours and once overnight. After digested, solution was filtered to remove undigested peptides. Digestion reaction was quenched using 1% formic acid to lower pH, then solution was cleared with an ethyl acetate wash. Peptide concentration was determined via BCA assay. Peptides were then lyophilized/dried via SpeedVac and reconstituted in Solution A (95% LC-MS-grade H₂O, 5% LC-MS-grade acetonitrile, 0.1% formic acid) at a concentration of 1 μ g/ μ L.

5.2.3 DNA Extraction

DNA extraction was performed using a Quick-DNA Fecal/Soil Microbe kit (Zymo Research, Irvine CA), according to manufacturer's protocol. The extracted DNA was further purified using the DNA Clean & Concentrator kit (Zymo Research) and quantified by fluorometry using a Qubit assay (ThermoFisher Scientific).

5.2.4 16S rRNA Amplicon Sequencing and Analysis

For 16S rRNA amplicon sequencing, the V4 hypervariable region of the 16S rRNA gene was amplified using universal primers derived from the original 515F and 806R primers (5' GTGCCAGCMGCCGCGGTAA and 5' GGACTACHVGGGTWTCTAA, respectively) (Bates et al. 2011) fused to Illumina sequencing adapters, following the procedure developed by (Lundberg et al. 2013). We also supplemented the reaction with a 5% of a modified 515F primer that increases the coverage of TM7/Saccharibacteria (5' GTGCCAGCMGCCGCGGTCA) (Liang et al. 2018). 12-nucleotide barcode sequences were incorporated into the second stage amplification reaction to enable sample multiplexing. The final amplicons were pooled, purified using Agencourt Ampure XP beads and quantified by Qubit. A pooled amplicon sample was sequenced (2 x 250 nt) on an Illumina MiSeq instrument (Illumina Inc, San Diego, CA) using a v2 500 cycle kit.

Raw sequence reads were trimmed of the PCR primers/adaptors using cutadapt (<https://doi.org/10.14806/ej.17.1.200>) and joined using the Qiime script join_paired_ends.py. Joined reads were processed using the qiime2 (v. 2019.1.0) framework (Bolyen et al. 2019). Reads were demultiplexed using the demux plug-in and denoised using the dada2 plug-in creating non-redundant amplicon sequence variants with a table of read counts per sample (Callahan et al. 2016). A phylogenetic tree was created using the raxml-rapid-bootstrap plug-in with default settings (Stamatakis 2014). Based on the data summary, diversity metrics Shannon (alpha-diversity), Simpson (alpha-diversity), Faith (phylogenetic alpha-diversity), weighted and unweighted UniFrac (phylogenetic beta-diversity), Bray-Curtis (beta-diversity considering abundance and presence versus absence), and Jaccard (beta-diversity considering just presence and absence) were calculated with the diversity plugin using a rarefaction of 50,000 reads (Faith 1992, Lozupone et al. 2007). Taxonomy was assigned to amplicon sequence variants using the qiime 2 Silva database classifier (Pruesse et al. 2007, Bokulich et al. 2018). Amplicon sequence variants were filtered if they had less than 50 reads across all samples, and these robust amplicon sequence variants were used in subsequent statistical analysis using the gneiss plug-in, and MicrobiomeAnalyst, <https://www.microbiomeanalyst.ca/> (Dhariwal et al. 2017, Morton et al. 2017). In MicrobiomeAnalyst, a random forest classifier was used to determine bacterial families most associated with each experimental group, and gneiss generated a heatmap of the overall 16S data.

5.2.5 *Whole Genome Metagenomic Sequencing and Analysis*

Metagenomic sequencing was performed at the Sequencing Center of HudsonAlpha Institute (Huntsville, AL). Genomic DNA from 18 colon fecal samples was sheared and used to generate libraries that were sequenced (2 x 150 nt) on an Illumina NovaSeq instrument. The raw reads quality checked and trimmed using Trim Galore https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Reads were then assembled with an aim to extract metagenome assembled genomes using a modified version of the Olm et al. procedure (Olm et al. 2017). In brief, all 18 samples were co-assembled using Megahit with the parameters k-min set to 31 and k-step set to 10 (Li et al. 2015). Mouse and human DNA was removed using bbsplit against a masked human and mouse databases with the parameter minid set to 0.95 similar to as described here, <http://seqanswers.com/forums/showthread.php?t=42552>. Cleaned sequence reads were single sample assembled using idba_ud with default parameters (Peng et al. 2012). Reads from all samples were mapped to all assemblies using bwa mem piped to samtools creating sorted bam files (Li and Durbin 2009, Li et al. 2009). Subsequent bam files were used to bin contigs from all assemblies by coverage and tetranucleotide frequency using Metabat2 (Kang et al. 2019). Genomes were quality checked using checkM (Parks et al. 2015). Taxonomy was assigned to all genomes that passed the threshold >50% complete <25% contaminated using gtdb-tk, which uses pplacer and ANI to classify genomes based on a database of over 150,000 genomes from NCBI and other sources (Parks et al. 2018, Chaumeil et al. 2019). Over 54% of genomes were >50% complete, and <10% contaminated out of the box, minimum standard for a medium quality genome so no further curation was done (Bowers et al. 2017). Select genomes that were >50% complete >10% contaminated were refined using Anvi'o by tetranucleotide frequency clustering (Eren et al. 2015). Finally, all medium quality genomes, >50% complete and <10% contaminated were dereplicated using dRep at 99.5% ANI and 95% ANI to create non-redundant genome and species clusters respectively (Olm et al. 2017). For each group, the best quality non-redundant genome was selected as a representative genome. Raw data and representative genomes are available in BioProject **PRJNA603829**. Protein sequences from the assemblies were extracted from the checkM result and annotated using emapper from the eggNOG database (Huerta-Cepas et al. 2017, Huerta-Cepas et al. 2018). Abundance of each representative genome was calculated using the inner-quantile mean coverage (Q2Q3 mean coverage) of each genome as calculated by Anvi'o.

5.2.6 Metaproteomic Measurement

Colon extracted fecal samples were measured on a Q-Exactive Plus (Thermo Fisher Scientific, Bremen, Germany) using an auto-sampled online two-dimensional chromatography system coupled to an electrospray emitter. Briefly, a three-phase back column was created using 150 μ m ID polyimide-coated fused silica containing 5cm of Aqua C18, 10cm of Luna strong cation exchange (SCX), and 5cm of Aqua C18 (Phenomenex, Torrance, CA). This frit was then connected to a 30cm C18 analytical column with an electrospray emitter at the end. A 12 μ g of peptide solution, per sample from the previously detailed extraction method was loaded onto the back-column via an UltiMate 3000 HPLC auto sampler over a 30 minute load step, then 4 subsequent salt pulse steps of 35, 50, 100, 500 nano molar ammonium acetate were used to push peptides onto the analytical column. Peptides were then separated over a 150-minute reverse phase gradient, while being sprayed onto the mass spectrometer. A final 500 nano molar ammonium acetate salt pulse was used as a wash step prior to the load of the next sample. Peptides were detected using data dependent acquisition, with 5 peptides isolated for MS2 fragmentation per MS1. Peptides were fragmented using HCD fragmentation, and previously detected mass-to-charge ratios, were placed on a 30 second exclusion list.

5.2.7 Protein identification

All prodigal gene calls extracted by running checkM on the medium quality representative genomes were extracted to create a protein database for proteomic analysis. The UniProt mouse proteome was added to this protein database, and all sequences were clustered at 98% identity using cd-hit to remove redundancy (Fu et al. 2012). After, a list of common protein contaminants was added to the protein database. Protein identifications were conducted using the crux (v. 3.2.8aa66a2) framework (McIlwain et al. 2014). Initially an index of the protein database was created using tide-index with the settings: --missed-cleavages 2 --mods-spec C+57.02146,2M+15.9949 --clip-nterm-methionine T --decoy-format protein-reverse. Then an initial search was conducted using tide-search with a precursor window of 10 ppm. After this search using an in-house python script, a new protein database was created for each sample based on all proteins that were potentially matched to a spectrum. With these protein databases, each sample was searched against a sample specific database using tide-search with the following parameters: --compute-sp T --exact-p-value T --score-function both --precursor-

window 10 --precursor-window-type ppm --num-threads 10 --mz-bin-width 1.0005079 (Diament and Noble 2011, Howbert and Noble 2014). Final search results were all submitted together to percolator for false discovery rate (FDR) calculation (Käll et al. 2007). For a protein to be identified, it needed to have at least one unique peptide identified with a peptide spectral match FDR of less than 1%. Protein FDR was also calculated, but not used in initial filtering. Retention time was extracted for all peptide spectral matches using an in-house script, and MS1 apex feature intensity was calculated for each peptide spectral match with a q-value < 0.01 using moFF (Argentini et al. 2016). The moFF match between runs feature was specifically not used due to issues with false hits. Protein abundance for each sample was calculated by summing intensities for all unique MS1 features. Then all quantified proteins with a unique peptide were loaded onto a matrix. This matrix was log₂ transformed, loess normalized, and mean centered using Inferno (Polpitiya et al. 2008). Proteins were also similarly quantified by spectral counting when appropriate.

5.2.8 Richness and principle component comparison between conditions and omics methods

Frequency of robust ASV's from qiime 2, inner-quantile of mean coverage of representative genomes from Anvi'o, summed spectral count from genome specific proteins, and individual protein intensity were all independently submitted to log transformation, loess normalization, and mean centering in Inferno (Polpitiya et al. 2008). Subsequently missing values were imputed as zero, and principle components were calculated using base R. Finally, PCA plots were rendered using ggplot2's scatter plot feature. To determine richness of each measurement. Number of quantified filtered ASV's in each sample, Genomes with an inner-quantile mean coverage >2, and number of proteins were counted per sample. Significant difference between conditions was calculated using Base R aov and TukeyHSD functions, and boxplots were rendered using ggplot2.

5.2.9 Evaluating newness of representative genome's gene calls

The newness of protein sequences taken from representative genomes was evaluated against a previously compiled mouse microbiome gene catalog (7/17/2015), nr (10/16/2019), environmental nr (9/28/2019), and UniParc (9/19/2019). All these databases were downloaded on October 16, 2019 (Xiao et al. 2015, The UniProt 2018). A diamond blastp was then conducted between the protein database used for metaproteomic analysis and these four databases

(Buchfink et al. 2015). Proteins that were not matched were recorded, and the top percent identity for each matched protein was also recorded.

5.2.10 Differential protein analysis

Normalized proteins intensities were filtered down to proteins that were quantified at least 3 times in a condition. These proteins were then submitted to serial T-tests between B6 control, and an experimental condition (B6 T, db/db S, db/db T, DIO S, DIO T). Using python, t-statistic, confidence interval, p-value, and Benjamini-Hochberg procedure was calculated for each comparison (Benjamini and Hochberg 1995). Proteins that passed an Q of 0.1 were considered to be significantly different between two conditions. Prior to t-test calculation, missing values were imputed with a random number down shifted 2.8 standard deviations from the mean and within half a standard deviation. The goal was to avoid comparisons between zero and a large number, while ensuring that any situation where a protein was quantified in all samples in one condition and none in the other the protein would be significant. All proteins in this set had a protein FDR below 5%.

Hypergeometric tests were used to determine if any species were significantly enriched on one side of the comparison or the other. Among these species that had a $q < 0.01$ species that had greater than 15 significant proteins in one condition were isolated for future analysis. These species were also compared by number of proteins in a condition versus the inner quantile mean coverage of a representative genome of that species in the same conditions.

5.2.11 Functional analysis

COG broad functional categories and KEGG Orthologs were extracted for each protein from the emapper results (Tatusov et al. 2000, Kanehisa et al. 2016). The KEGG ftp was used to connect KEGG orthologs to pathways, module, reactions, and Brite hierarchies. For each grouping the number of quantified proteins per sample was counted. For Broad COG categories number of proteins was divided by number of proteins to create a percentage, and statistical analysis was conducted by Two-way ANOVA and TukeyHSD in base R. For Pathways, Modules, and Brite hierarchies all groupings were counted in one matrix, and the count data was transformed and normalized by centered log ratio after filtering any category that did not have more than 4 proteins in at least one sample. The same was done for KO terms, but once for all KO terms with at least 4 proteins in a condition, and again for all KO terms with at least 7

proteins in a condition. The second filtering was done to help elevate more abundant KO terms. All of these results were statistically analyzed by Two-Way ANOVA in Perseus (Tyanova et al. 2016), and confirmed by Two-Way ANOVA and TukeyHSD in R. P-values within each analysis were stacked and FDR was controlled by Benjamini-Hochberg procedure ($Q < 0.25$).

5.2.12 Cytokine Assays and Analysis

Plasma samples collected as previously described and colon tissue (cryo grinded then protein extracted as previously described) homogenates were analyzed for inflammatory factors using Milliplex MAP Mouse Cytokine/Chemokine Magnetic Bead Panel (Millipore Sigma, Burlington, MA) premixed 32-plex kit according to the protocol. All samples were 0.2 μ m filtered, and plasma volume was doubled (25 μ L) to achieve detectable levels for most cytokines. Tissue supernatant volumes were based on protein concentration by NanoDrop readings of individual samples to achieve 15 μ g total protein per well. Assay samples were aliquoted in triplicate; after protocol completion, plates were read on a Luminex analyzer.

Absolute quantitation in pg/mL for each cytokine (limit of detection 3.2 pg/mL) was achieved using Luminex software provided polynomial standard curves. Values below 3.2 pg/mL were imputed with a random number between 0 and 3.2 pg/mL. Exploratory data analysis using PLS-DA in the MixOmics R package was used to hypothesize which cytokines were significantly more abundant in each line, and in morphine treatment (Rohart et al. 2017). Two-way ANOVA followed by Tukey Honest Significant Difference (HSD) was used to confirm hypotheses.

5.2.13 Urinary Metabolites

Mouse urine collected during 24-hour metabolic cage duration was analyzed by LC-MS/MS on a Shimadzu LCMS-8040 in multiple reaction monitoring (MRM) mode for the following analyte transitions (m/z): morphine (285.70 > 201.00, 285.70 > 152.35), morphine-3- β -D-glucuronide (462.20 > 286.15, 462.20 > 183.15, 462.20 > 58.20), hydromorphone (285.75 > 184.95, 285.75 > 157.00), hydromorphone-3- β -D-glucuronide (462.20 > 286.10, 462.20 > 184.95), and normorphine (272.20 > 152.00, 272.20 > 165.00, 272.20 > 121.00). Spiked internal standards morphine-D6 (292.00 > 201.05, 292.00 > 152.95, 292.00 > 165.00) and morphine-3- β -D-glucuronide-D6 (465.25 > 289.10, 465.25 > 201.10, 465.25 > 183.00) added to each sample at the same concentration for standardization and normalization. All standards obtained from

Cerilliant (Round Rock, TX) were compiled into a stock solution used for serial dilution into aqueous urine matrix to achieve a nine-level calibration range: 10 to 5,000 ng/mL for morphine, hydromorphone, and normorphine; 50 to 25,000 ng/mL for morphine-3- β -D-glucuronide and hydromorphone-3- β -D-glucuronide. Urinary drug testing analysis confirmed morphine dosing of experimental treatment subjects in addition to hepatic drug metabolism among all three mouse lines.

5.3 Results

To compare microbiome changes due to morphine in the context of different kinds of obesity, three lines of mice from the same genetic background were selected. The control group comprised C57BL/6J mice (B6; n=6) consuming a standard mouse diet. The two groups of obese mice included C57BL/6J with diet induced obesity (DIO; n=6) caused by consuming a 60% fat diet, and B6.BKS(D)-LepR^{db}/J (db/db; n=6) consuming a standard mouse diet (Figure 5.1). The db/db mice have a mutated gene that codes for the leptin receptor, which confers susceptibility to obesity and type-2 diabetes (Coleman 1978). Three mice from each of the three lines were treated either with a saline vehicle control, or morphine administered continuously for two weeks via chronically implanted Alzet pumps. At the end of the treatment, feces were collected from the metabolic cages (24 hours) and surgically excised from the colon. The three mouse lines had significantly different weights, with db/db mice heavier than DIO and DIO heavier than B6 mice (Figure 5.1). To evaluate morphine metabolism, urine was collected in the metabolic cages and measured with a targeted mass spectrometry assay. Morphine metabolites were detected only in mice treated with morphine, although the db/db mouse had significantly lower levels for these metabolites. (Figure 5.1).

To show that inflammatory effects of morphine and obesity were observable in this study, plasma (systemic) and protein from colon tissue (localized) were analyzed with a panel of mouse chemokines and cytokines (Figure 5.2). Most analytes were not detectable; however, markers of inflammation were observed to be elevated in db/db, DIO, or morphine treated mice. Each group had different elevated analytes, and morphine's inflammatory effect could only be observed in the lean mice (Figure 5.2). There were some conflicting results for example the mouse homolog of IL-8, KC, was more abundant in db/db plasma; however, the anti-inflammatory cytokine IL-10 was more abundant in db/db colon tissue. This could be caused by the colon tissue attempting

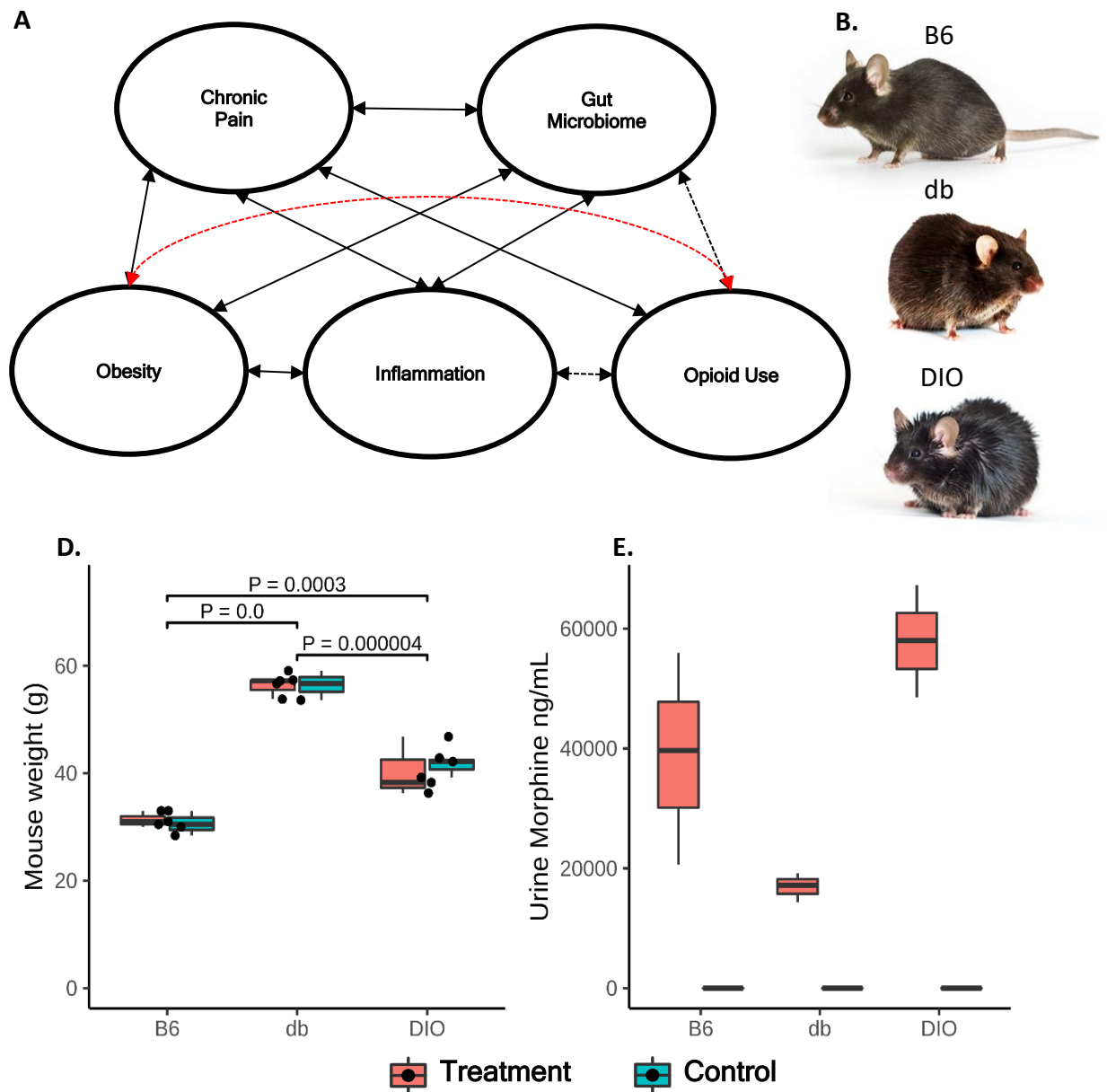


Figure 5.1: Experimental design for mouse study. (A) Flow chart of starting hypotheses of bidirectional interaction between pain, obesity, inflammation, opioid use and the gut microbiome. (B) Side by side comparison of db mice vs B6 Mice. (C) Table describing experimental setup. (D) Boxplot showing weight in grams of mice. P-Values calculated by TukeyHSD in R. (E) Boxplot depicting morphine quantities in mouse urine Analyzed by triple quadrupole mass spectrometer. Images are from the Jackson Laboratory website.

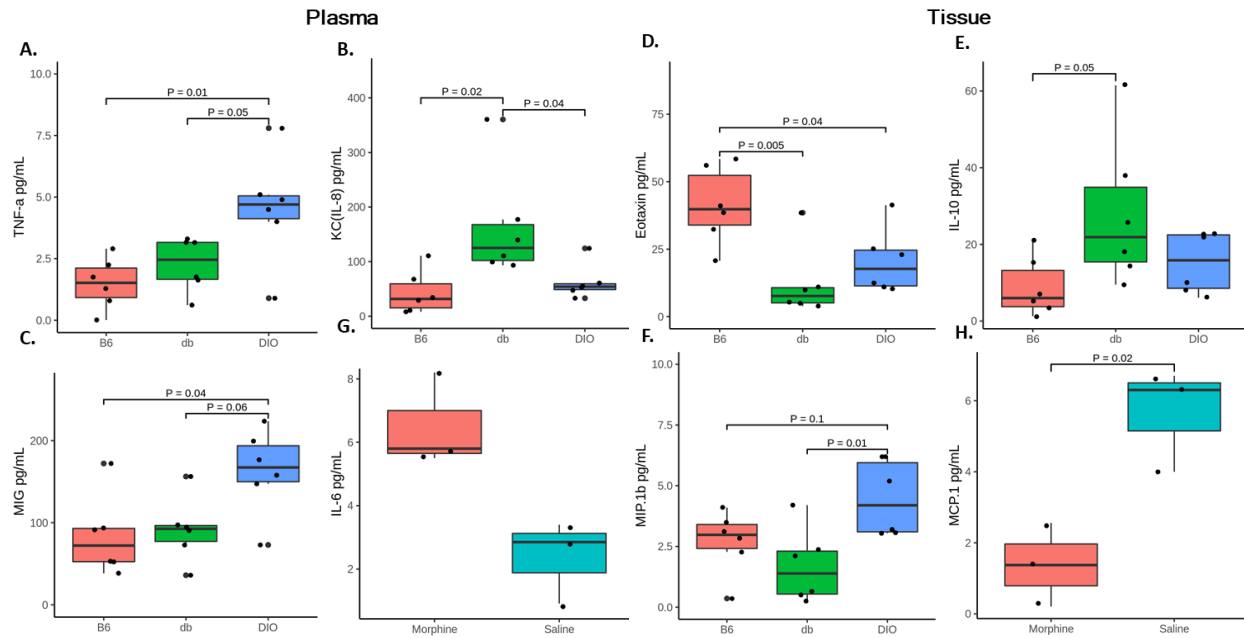


Figure 5.2: Cytokine analysis reveals different markers of inflammation for different types of obesity and morphine. (A-C) Boxplots comparing abundance of TNF-a, KC, and MIG in plasma across different mouse lines. (D-F) Boxplots comparing abundance of Eotaxin, IL-10, and MIP.1b in colon tissue across different mouse models. P-Values calculated by two-factor ANOVA (Model versus Treatment) followed by Tukey HSD. (G) Boxplot comparing abundance of IL-6 in plasma between B6 model morphine samples to B6 model saline samples. (H) Boxplot comparing abundance of MCP.1 in colon tissue samples between B6 model morphine samples versus B6 model saline samples. P-values calculated by welch t-test.

to compensate for the systemic inflammatory state in the db/db mice. While pro-inflammatory IL-6 was seen to be elevated in morphine treated B6 mice; pro-inflammatory MCP.1 was seen to be elevated in colon tissue of saline treated B6 mice relative to morphine treated B6 mice. Eotaxin a chemokine associated with the development of the mucosal layer was elevated in B6 mouse tissue relative to db/db and DIO tissue potentially eluding to the microbiome disruption caused by the two types of obesity (Loktionov 2019). These results show clear but distinctly driven systemic inflammation due to DIO, db/db, and morphine, with some conflicting results at the localized level.

5.3.1 Comprehensive microbiome characterization reveals a mouse gut microbiome membership that is underrepresented in public repositories

To provide an overview of the mouse gut microbiome's composition and guide subsequent whole genome metagenomic sequencing, 16S rRNA amplicons derived from metabolic cage and colon fecal samples were sequenced to saturation revealing 157- 467

amplicon sequence variants (ASV's) per sample and 1,546 in total. While previous reports indicated that collection over 24 hours should not have a major impact on the composition of the microbiome (Lauber et al. 2010, Song et al. 2016), the anerobic nature of the gut microbiome provided concern that this would have an undue effect on some of the more detailed results. Results from the two samples types aligned with the literature in principle. Sample types were overall indistinguishable by beta-diversity analysis; however, 350 out of 1,546 amplicon sequence variants (ASV's) detected in total were not detected in the colon fecal samples, differences between morphine and saline were more observable in the colon fecal samples, and alpha diversity results conflicted between the two sample types (Figure 5.3). Thus, subsequent analysis was done using the colon fecal samples as the best representative of the native microbiome.

Matching ASV's to the Silva the database revealed that 95% of the ASV's could not be classified at the species level. Many sequences matched with less than 95% nucleotide identity when compared to the NCIB 16S rRNA database (Figure 5.4) (Pruesse et al. 2007, Bolyen et al. 2019). This implies that many of the bacteria present in these samples remain uncultured and do

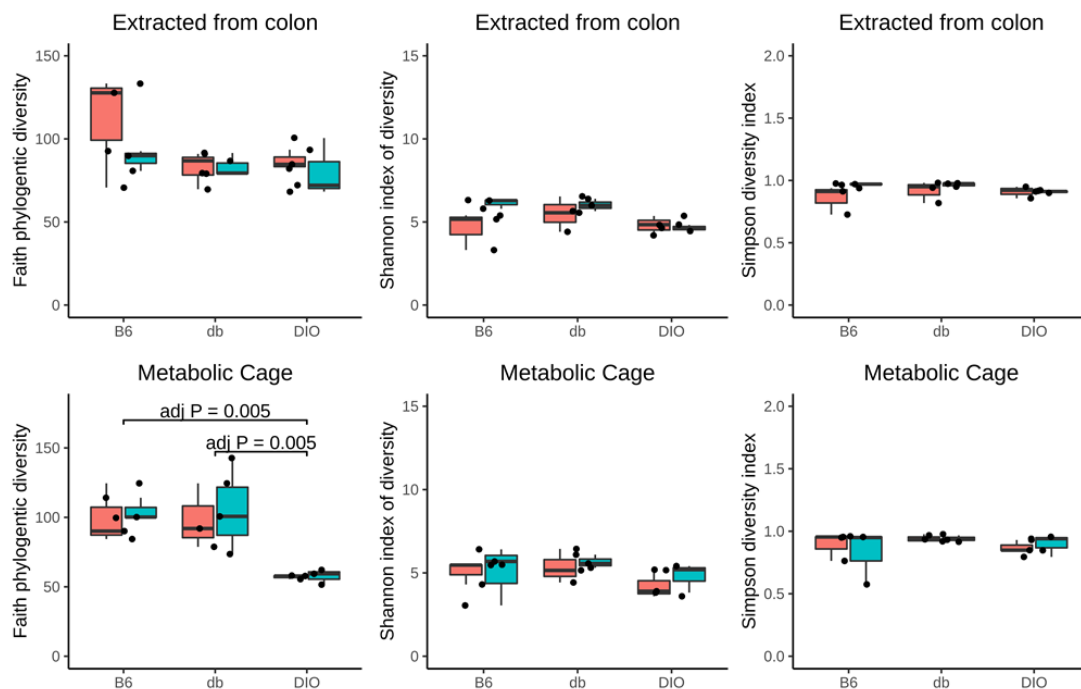


Figure 5.3: Alpha diversity metrics for samples collected directly from colon versus in a metabolic cage.

not have a representative in the NCBI database reinforcing the need for obtaining protein sequences using metagenomics.

Metagenomic sequencing of each colon extracted microbiome sample resulted in an minimum 250 million reads per sample (2 X 150 bp). Sequence binning and curation extracted 590 medium-quality metagenome-assembled genomes (MAGs), as determined by previously defined standards (Bowers et al. 2017). Between 215-467 of the genomes were quantifiable per sample suggesting a thorough extraction of MAGs from the metagenome based on the 16S results (Figure 5.4). The MAGs represent 460 distinct species with a 95% ANI threshold (Olm et al. 2020). Only 7% of the genomes are represented in the Genome Taxonomy Database, which is compiled from RefSeq and GenBank (Parks et al. 2018, Chaumeil et al. 2019) (Figure 5.4). As the number of named species assembled was even lower, this indicates that the mouse microbiome is still underrepresented in sequence databases. Supporting this, blast searches of all MAG derived protein sequences against the UniParc, NR databases, and mouse catalog yielded mean top-blast alignments <90 percent amino acid identity (Figure 5.4) (Xiao et al. 2015, Coordinators 2018, The UniProt 2018).

Predicted encoded proteins from the MAGs were used as the reference database for metaproteomic identifications. In total, 1,212,277 non-redundant protein sequences clustered at 98% identity were derived from the metagenome assembled genomes and combined with host proteins to generate a protein sequence database for searching against spectra. 97% of these non-redundant proteins were resolvable down to the species level making species-resolved metaproteomics on previously uncharacterized species possible. Using a 2D-LC-MS/MS method, 115,628 proteins were detected with at least one unique peptide, with 11,112-16,238 non-redundant proteins detected in each sample (peptide spectral match $q < 0.01$) (Figure 5.4). Less than 10% of the proteins detected in each sample belonged to the host, the remainder were bacterial. Approximately 70% of these proteins were assigned a functional annotation using the eggNOG database, enabling subsequent comparative metabolic inferences across the mice and treatments.

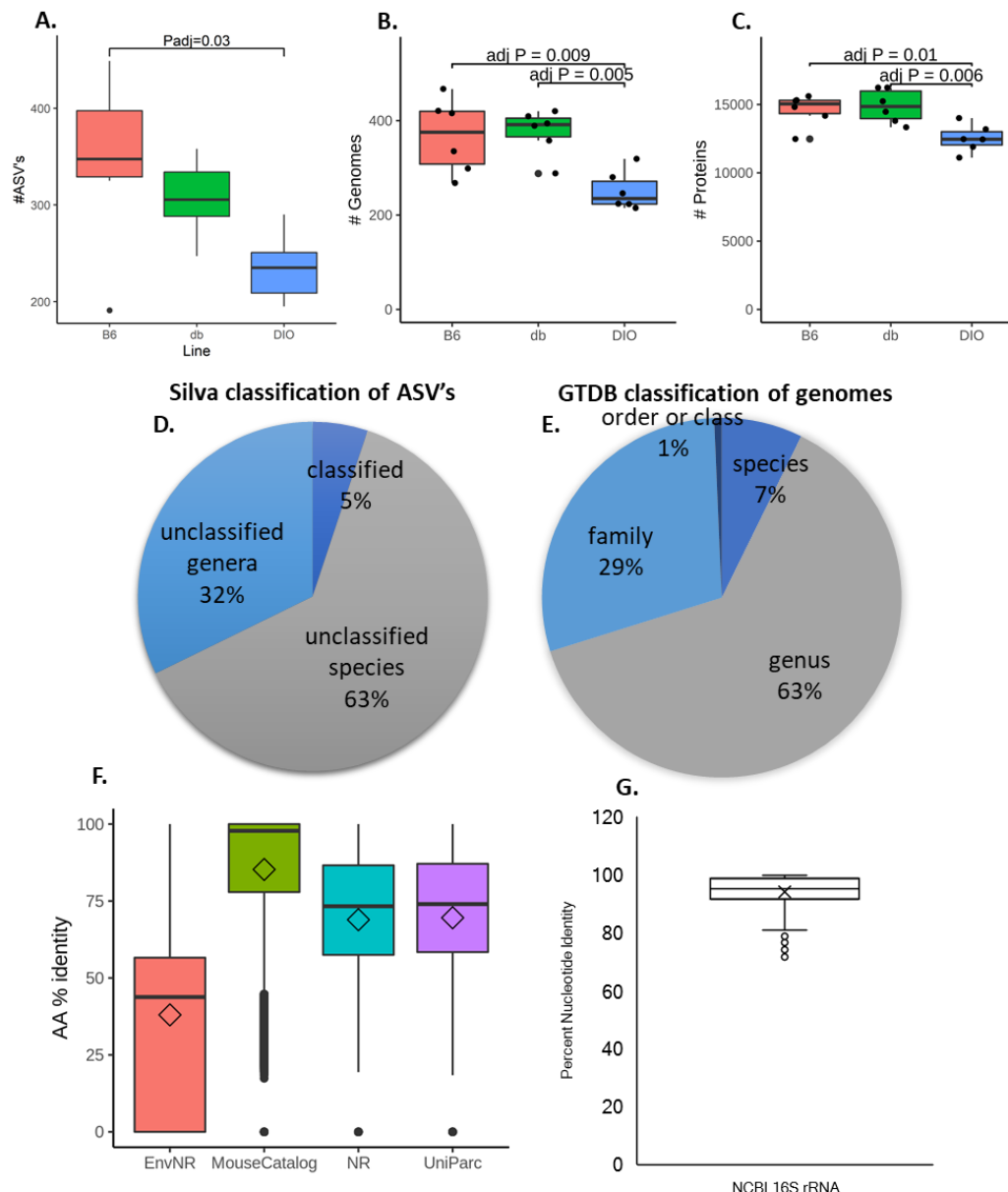


Figure 5.4: Comprehensive profiling of the mouse gut microbiome reveals a microbiome underrepresented in public repositories. (A-C) Boxplots depicting the number of ASV's, quantifiable genomes, and proteins detected in each sample respectively. (D) pie chart depicting the percentage of ASV's classified using the Silva database down to the species level, unclassified at the species level, and then unclassified at the genera level. (E) pie chart depicting the percentage of genomes classified down to the species, genus, family, class, or order level. All genomes were classified at the phylum level. (F) boxplot depicting amino acid percent identity match of each protein to environmental NR, mouse catalog, NR, and UniParc databases. (G) boxplot of nucleotide percent identity match of each ASV to the NCBI 16S rRNA database.

5.3.2 *Morphine alters the gut microbiome composition but is eclipsed by obesity effects*

To see if morphine or different kinds of obesity had any effects on the diversity of gut microbial communities, 16S data was evaluated using varying alpha diversity metrics. Except for the Faith phylogenetic diversity of amplicon sequence variants taken from the metabolic cage samples, no metric of alpha diversity indicated any differences in within sample diversity between conditions (Figure 3.3). The total number of amplicon sequence variants, genomes, and protein detected in each sample was significantly lower in DIO suggesting lower community richness due to high-fat-diet in agreement with a recent meta-analysis (Figure 3.4) (Bisanz et al. 2019). Given that db/db mice have a significantly greater weight than DIO mice (Figure 3.1), this implies that the source of obesity drives decreases in microbial population richness not obesity itself. Morphine did not appear to have any effect on alpha-diversity or richness of the gut microbiome.

To evaluate changes in microbial composition between conditions, mean-centered log transformations of ASV abundance and genome abundance calculated by metagenomic coverage or metaproteomic spectral counts were evaluated by principle component analysis (Figure 5.5). Abundances derived from all three omics showed the same data profile. Microbial composition is affected most by DIO separating along the first component. db/db has the second greatest effect by separating along the second component. Morphine was shown to have a lesser effect separating along the third component, but only in the lean mice. These results were corroborated by hierarchical clustering of 16S and metagenomic data, and analysis of beta-diversity metrics from the 16S data (Figure 5.6, Figure 5.7).

5.3.3 *Differential metaproteomics reveals that taxa-level protein abundances were specific to sample conditions.*

To uncover details behind changes in composition correlated with morphine treatment or different kinds of obesity, the proteomics data was evaluated for proteins that were significantly different between conditions. Protein presence versus absence was shown to be the driving effect in the differential comparison between samples, and these differences appeared to be driven by the abundance of originating species not the functionality of the organisms. Differential abundance proteomic analysis revealed 7,324 proteins quantified in all samples from at least

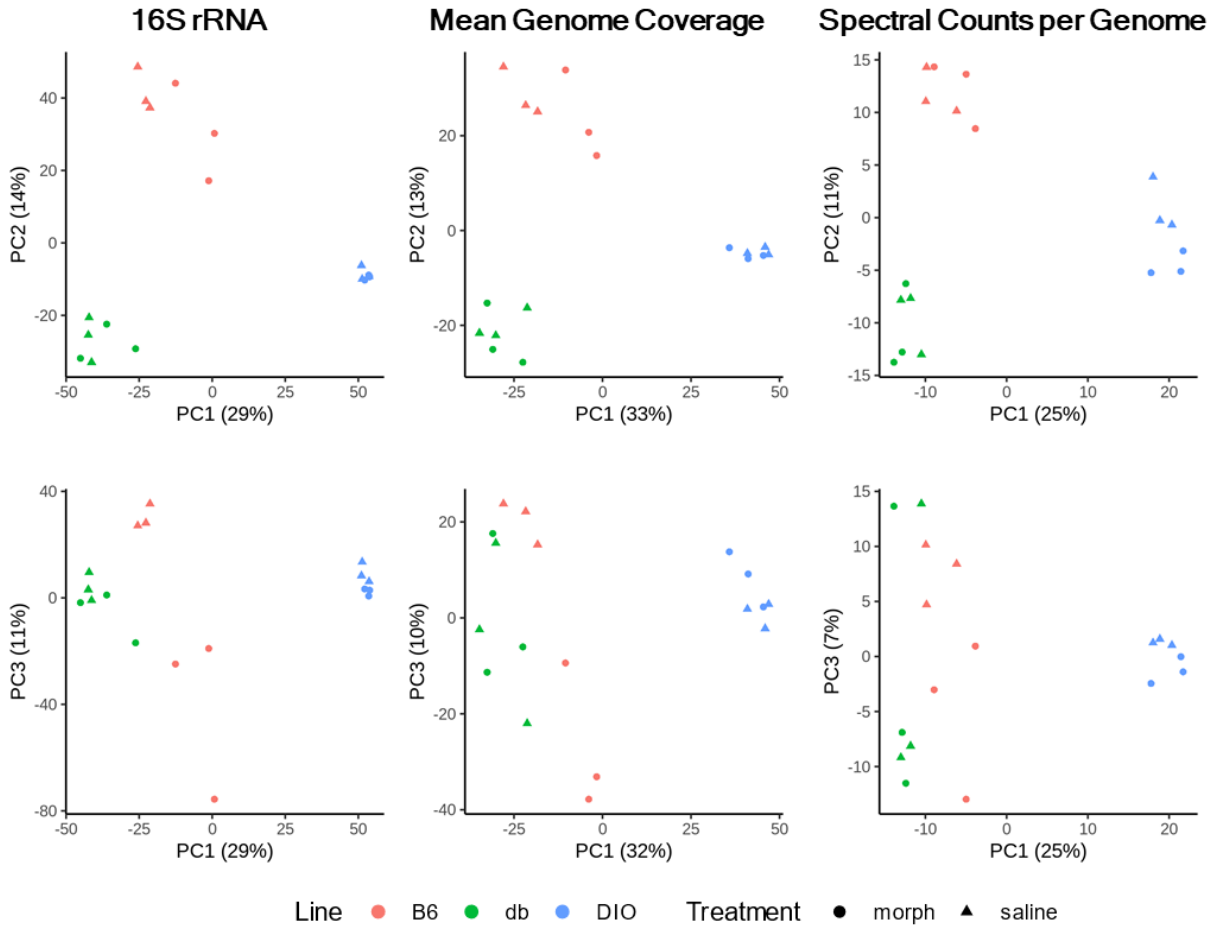


Figure 5.5: Diet induced obesity alters microbiome composition more than genetic obesity or morphine treatment. Principle component analysis of log transformed mean centered counts derived from 16S rRNA read coverage, genome coverage, and metaproteome derived mass spectra assigned to genomes.

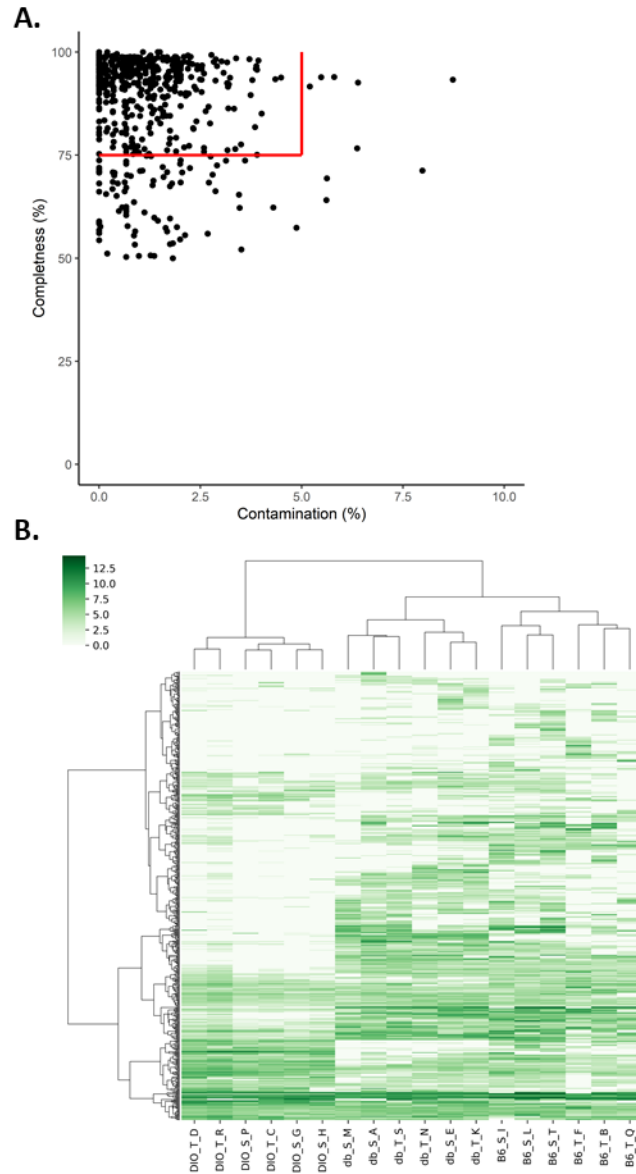


Figure 5.6: Medium quality genomes cluster by experimental group using hierarchical clustering. (A) Scatter plot of % contamination versus % completion of each genomes as calculated by checkM. Redlines depict a 75% completion and 5% contamination score. 83% of genomes are above and below that threshold respectively. (B) Heatmap of normalized inner-quantile of mean coverage per genomes as calculated by Anvi'o. Dendrograms calculated by ward clustering.

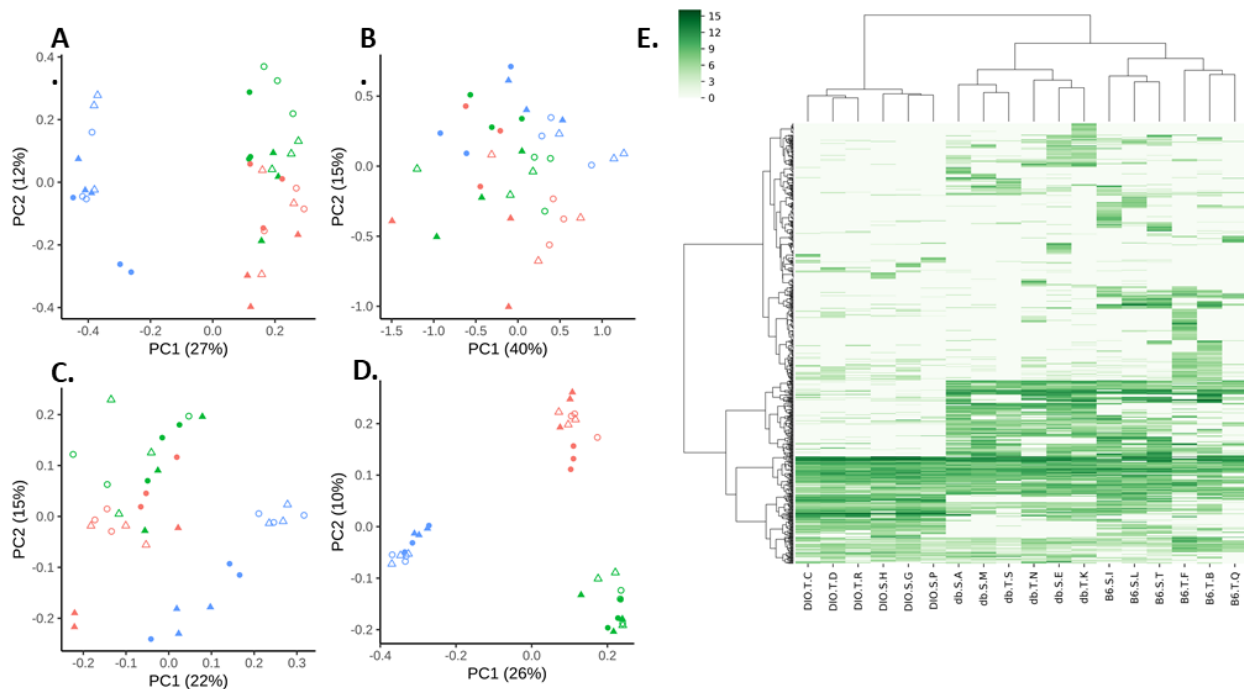


Figure 5.7: Beta-diversity analysis of mouse 16S data. (A-D) PCoA plots of bray-curtis, unweighted UniFrac, weighted UniFrac, and Jaccard respectively. (E) Ward clustered heatmap of all the ASV's collected from colon feces samples.

one condition. From these, 3,208 were significantly different in abundance between B6 saline and another condition (Benjamini-Hochberg procedure on student t-test, $Q < 0.1$). In almost all cases, a significant protein was present in all three samples of one condition, and absent from all three samples in one or more other conditions.

Specific species tended to be enriched among proteins abundant in B6 saline relative to another condition or vice-versa, and these tended to reflect the relative number of proteins detected for that species in that condition. The pattern of these numbers of proteins aligned with the genome coverage of that species (Figure 5.8, Figure 5.9). Thus, in a fashion similar to spectral counting, the number of proteins seem to be a good indicator of abundance for species biomass (Kleiner et al. 2017). This differential protein analysis suggested that specific species from the Lachnospiraceae family (labeled *Eubacterium* sp. or *Lachnospiraceae bacterium* for new genera) were suppressed by morphine (only in B6), db/db, and DIO, mice with the most Lachnospiraceae species suppressed by B6 morphine and DIO (Figure 5.8). This aligned with random forest classification of the ASV's, which suggested that high abundance of the

Lachnospiraceae family was most indicative of B6 saline (Figure 5.10). Previous studies in which morphine was administered reported similar results (Acharya et al. 2017); however, this does not mean that all Lachnospiraceae bacteria followed this trend. For example, *Lachnospiraceae bacterium 266_1* and *Eubacterium sp. 307_1* were more enriched among proteins more abundant in DIO than in B6 saline. Caution should be employed when making assumptions about trends relating to broader taxonomic classifications that are driven by specific species.

5.3.4 Both diet and morphine alter microbiome metabolic functions in specific ways

Since protein enumeration appeared to align to species abundance, it was inferred that this would apply to functional classifications as well. To explore changes in overall microbial function between conditions, the number of proteins was tabulated for COG broad functional categories, KEGG modules/pathways, and KEGG orthologs (Tatusov et al. 2000, Kanehisa et al. 2017). Broad functional classifications were mostly unchanged between individuals, which aligns with previous literature (The Human Microbiome Project 2012); however, the carbohydrate metabolism and transport had a significantly lower proportion of proteins in DIO mice and mice treated with morphine (Figure 5.11). This indicates an overall metabolic effect on the microbiome due to morphine and high-fat-diet, but not due to genetic obesity. PCA analysis of normalized number of proteins for KEGG orthologs confirmed the above result for DIO mice by showing that the DIO mice regardless of treatment were distinct from db/db and B6 mice via the first principle component, while B6 and db/db mice did not separate along any components (Figure 5.12). The metabolic shift due to morphine observed at the COG level was not reflected at the KEGG ortholog level as morphine treated mice, even in B6, were indistinguishable from saline treated mice by PCA.

To acquire a finer level of detail, KEGG pathways and modules tabulated by number of proteins per sample were filtered so that each pathway had to have at least 4 proteins detected in one of its samples. These results were normalized by centered-log-ratio and evaluated by two-way ANOVA comparing the factors mouse line, morphine, and their interactions. Significance was defined as a Benjamini-Hochburg procedure $Q < 0.25$.

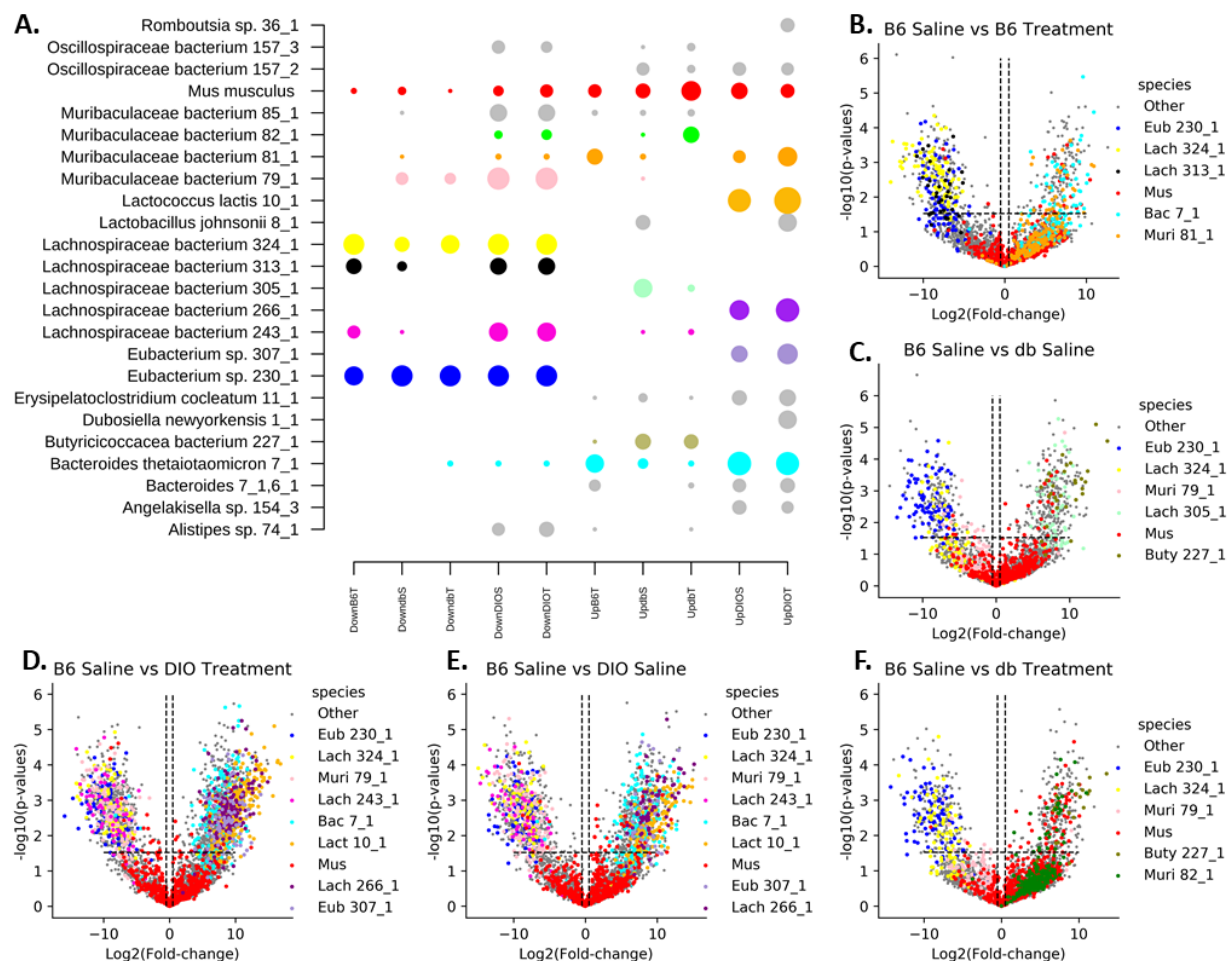


Figure 5.8: Differential protein abundance analysis reveals specific *Lachnospiraceae* species that favor B6 saline over B6 morphine and different kinds of obesity. (A) Bubble plot of species with at least 15 or more significant proteins in a comparisons between B6 saline and other experimental conditions. (B-E) Volcano plots of each binary comparison between B6 saline and another experimental condition. Dotted black lines depict a $\log_2$ fold change of 0.5 or greater, and Benjamini Hochberg procedure $Q < 0.1$ threshold. P-values were calculated by student T-test only comparing proteins that had at least 3 in one condition. Colors are species specific.

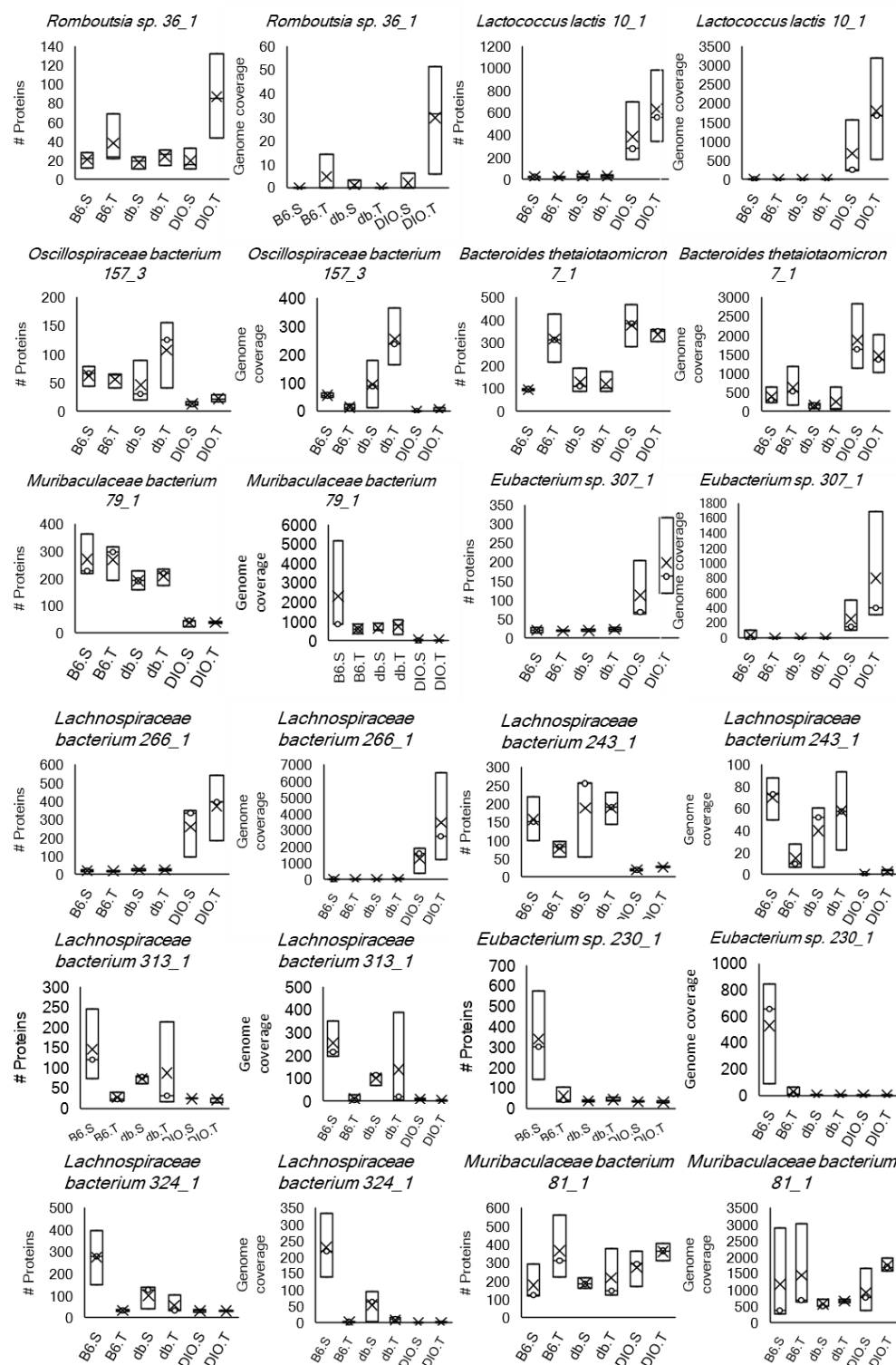


Figure 5.9: Species enriched for up-regulated proteins align with number proteins and genome sequence coverage of representative genomes. Boxplots depicting the number of proteins and inner-quantile mean coverage for representative species from Figure 4.

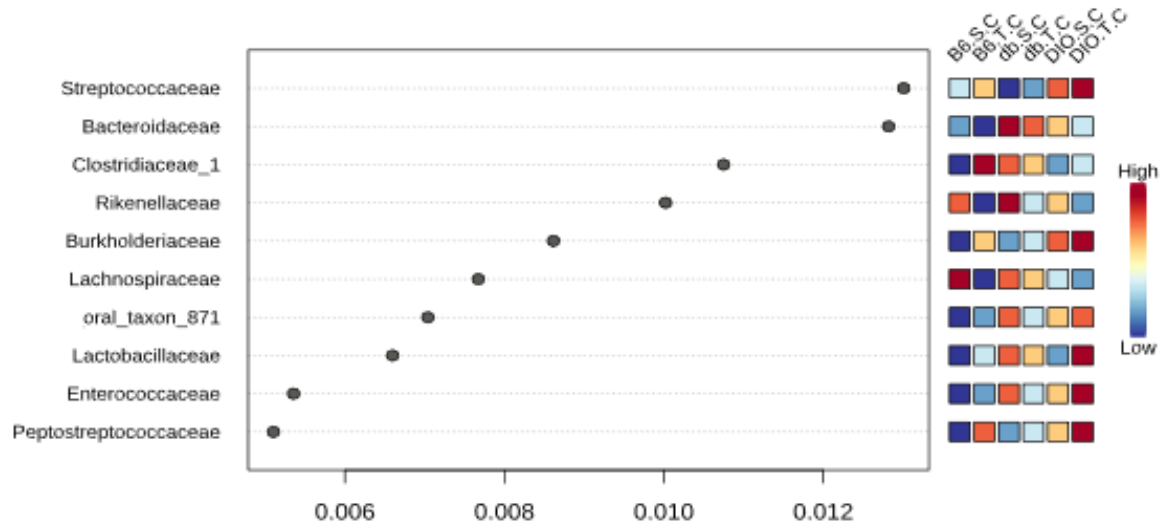


Figure 5.10: Random forest classification of 16S data show high abundance of Lachnospiraceae to be the best predictor of B6 saline treated mice.

Three KEGG pathways/modules were significantly different between morphine treated mice and saline treated mice. The prolactin signaling pathway (ko04917) and the nucleotide sugar biosynthesis module (M00554) had more protein in saline treated mice than morphine treated mice, while an ascorbate biosynthesis module (M00129) had significantly more protein in morphine treated mice relative to saline treated mice (Figure 5.11). Morphine promotes the production of prolactin (Muraki and Tokunaga 1978); however, the prolactin signaling pathway had more proteins in saline treated mice relative to morphine treated mice. This result was driven by microbial orthologs of the protein GaLT (K00965), which is suppressed by prolactin (Figure 5.11) (Halperin et al. 2008). The effect of morphine treatment on GaLT and subsequently the prolactin signaling pathway appeared driven by the saline treated db/db and B6 mice, but not the DIO mice. Although not identified by this analysis, KEGG maintains a morphine addiction pathway, ko05032. This pathway was specified to see if there was a morphine effect, and no microbial orthologs were found to be associated to this pathway; however, overall there were more host proteins ($P < 0.05$) associated with this pathway in morphine treated B6 mice relative to saline treated B6 mice (Figure 5.11).

Forty- KEGG pathways/modules were found significantly different between mouse lines, and 31 of these trended similar for db and B6 suggesting that high fat has a greater effect on microbiome function than genetic obesity. Among these pathways' pectin degradation, pentose

and glucuronide interconversions, ascorbate and alderate metabolism, and methanogenesis pathways all had relatively more proteins in B6 and db/db than in DIO (Figure 5.11). Conversely, glutathione metabolism, aminobenzoate degradation, fat digestion and absorption, and thyroid hormone synthesis pathways were all seen to have more relative proteins in DIO than in db/db and B6.

A similar analysis was conducted for KEGG orthologous groups; however, most of the significant orthologs were driven by a few proteins detected in one condition and not detected in another, which is not reflective of a bulk microbiome transition. Despite this, a couple interesting orthologs stood out. K00383 glutathione reductase was significantly greater in DIO relative to B6 and db. Expressed by *Lactococcus lactis* 10_1, among others, glutathione reductase helps *Lactococcus lactis* survive oxidative stress (Li et al. 2003), while high fat diet induces oxidative stress (Du et al. 2012). This could explain why *Lactococcus lactis* 10_1 is able to thrive in DIO (Figure 5.8), while overall community richness decreases (Figure 5.3). The large conductance mechano-sensitive channel, K03282, which protects cells against extreme turgor was the most significant difference between saline treated mice and morphine treated mice, with more proteins in morphine treated mice relative to saline treated mice (Figure 5.11) (Levina et al. 1999).

To better observe the larger scale and more robust transitions in microbiome function, KEGG orthologs (KO) that did not have 7 or more proteins in at least one sample were filtered out. After re-normalization, the majority of significant KEGG orthologs showed differences between mouse lines (Benjamini-Hochburg procedure $Q < 0.25$) and had similar trending between db/db and B6 mice. Several of the KO terms that were significantly up in db/db and B6 relative to DIO represented enzymes that specifically release sugars from known plant fibers including pectin, starch, arabinogalactan, and xylan (Figure 5.12). This suggests a preference for plant matter as a source of nutrients under B6 and db/db conditions as opposed to DIO. Though no clear pathways, there were a variety of conversion enzymes for the types of sugars that are released from the above fibers including several enzymes associated with glycolysis, methanogenesis, and fermentation further suggesting a well-nourished microbiome in B6 and db/db.

In contrast, KO terms with significantly more proteins in DIO mice relative to B6 and/or db/db mice suggested that high fat diet triggered starvation stress and transitions the microbiome away from foraging the diet towards foraging the host for nutrients. Trimeric autotransporter adhesin and flagellar assembly factor, virulence factors involved in host cell invasion and

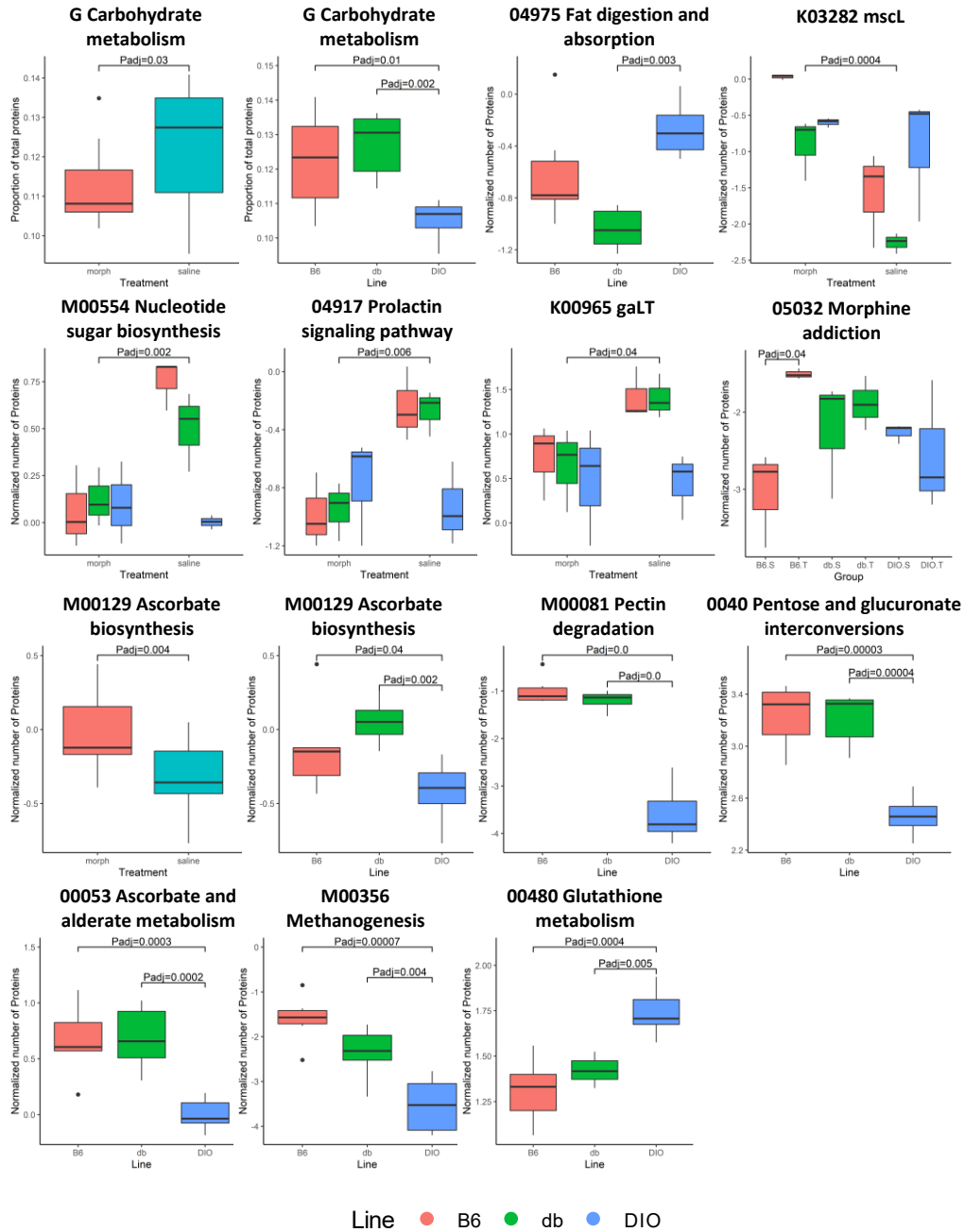


Figure 5.11: Counting proteins into functional categories reveals functional shifts between morphine versus saline and B6/db versus DIO. Boxplots depicting representative broad functional categories, KEGG pathways and modules, and specific KEGG terms that are significantly different between experimental conditions after the first filtration. At the top the broad COG category G represents the relative proportion of proteins that belong to that category. For the KEGG orthologs K00965 and K02382 these are the centered-log-ratio of tabulated proteins after filtering out all KO terms that did not have 4 proteins in at least one samples. The remaining boxplots represent KEGG pathways or modules, which also had their proteins tabulated and normalized by centered-log-ratio after filtering any pathways or modules that did not have at least 4 proteins in at least one sample. All p-values are from TukeyHSD after two-way ANOVA

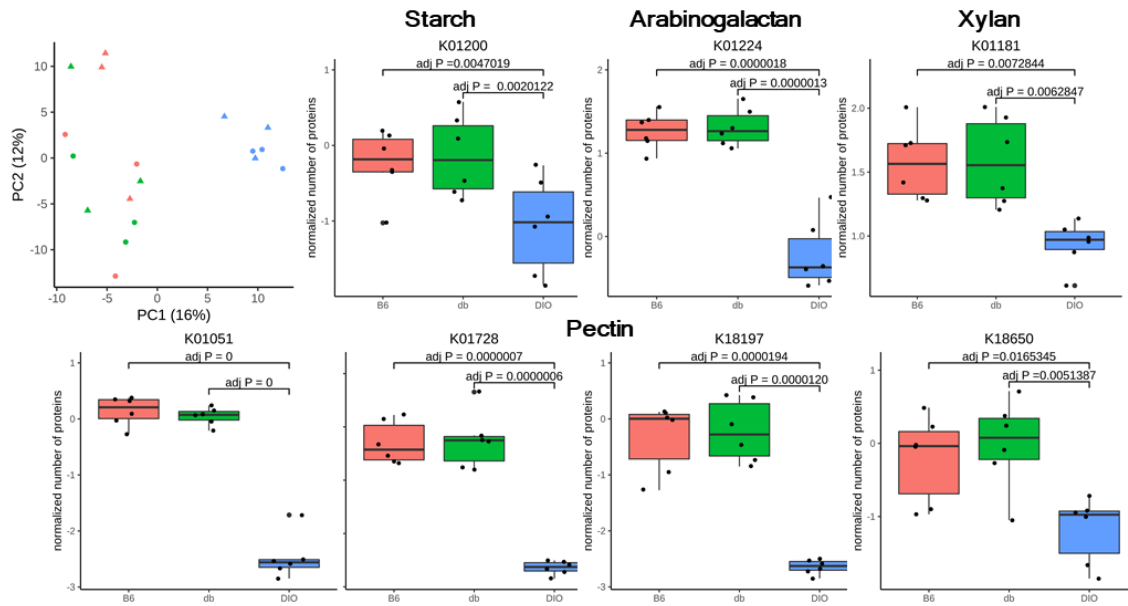


Figure 5.12: Counting proteins into KEGG orthologs reveals diet fiber degrading enzymes are more present in db and B6 than DIO. Number of quantified proteins per sampler were counted for each KEGG ortholog group. Kegg orthologs that had less 7 proteins max in a sample were filtered out. The remaining KEGG orthologs were normalized by center log ratio. All P-values are calculated by TukeyHSD after two-way ANOVA. All enzymes pass a Benjamini Hochberg procedure $Q < 0.25$. Starting at the top principle component analysis reveals that DIO has a distinct functional profile from db and B6, while B6 and db are overall indistinguishable. Boxplots depict specific KEGG Orthologs that degrade starch, xylan, arabinogalactan, or pectin and have significantly more normalized proteins in db and B6 than DIO.

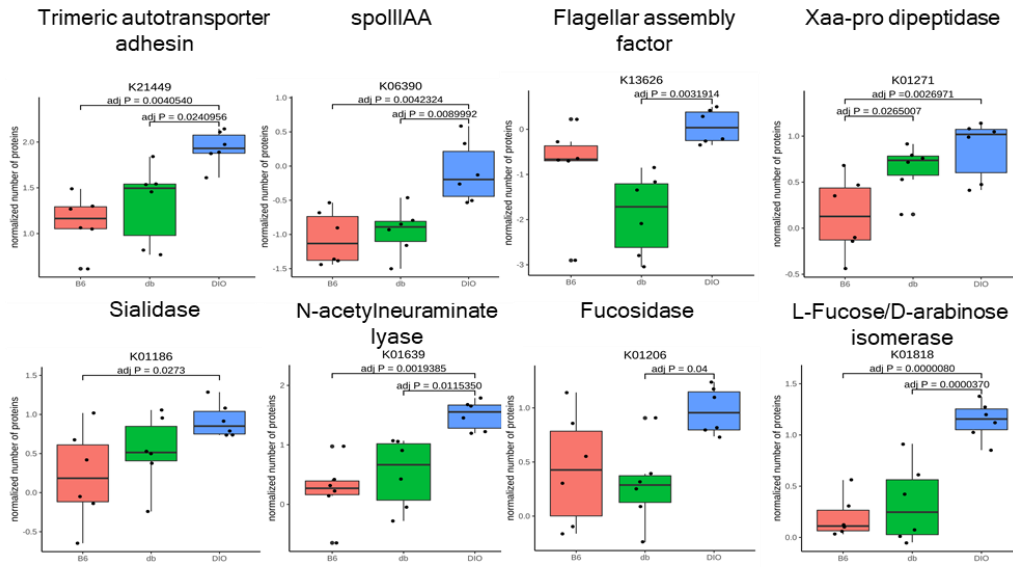
adherence to components of the extracellular matrix such as glycoproteins (Müller et al. 2011, Haiko and Westerlund-Wikström 2013), as well as SpoIIIAA, a protein critical for initiation of sporulation (Kellner et al. 1996), each had significantly more normalized proteins in DIO than db/db mice and/or B6 mice. Many orthologs with more proteins in DIO came from the KEGG amino sugar and nucleotide sugar pathway, which contains enzymes that catalyze fucose, acetylneuraminate, and acetylglucosamine (Figure 5.13). Mucins are glycoproteins that make up a major portion of the intestinal mucosal layer, and mucins' serine and threonine residues connect to glycans made up of fucose, acetylneuraminate, and acetylglucosamine, along with acetylgalactosamine and galactose (Holmen Larsson et al. 2013). Acetylgalactosamine and galactose catalyzing enzymes, however, were not significantly different between mouse lines, which may suggest that those enzymes are not well annotated or that the high fat diet microbiome prefers fucose, acetylneuraminate, or acetylglucosamine as sources of nutrition.

It is known that several glycoside hydrolases, hexosaminidase, fucosidase, galactosidase, and sialidase, specifically release the different components of mucin glycans (Marcobal et al. 2013). Galactosidases, though present, did not appear to have different abundances between conditions; however, a sialidase ortholog, K01186, and a fucosidase ortholog, K01206, each had more normalized proteins in DIO versus B6 or db/db, respectively (by $P < 0.05$ but not by Benjamini-Hochberg procedure), supporting the fucose and acetylneuraminate catalyzing enzymes that were up in DIO. The hexosaminidase KEGG ortholog K12373, however, did have significantly more normalized protein in DIO relative to B6 and db/db, and is specifically annotated to release acetylglucosamine from glycans. Additionally, the chitinase ortholog K01183, which is also annotated to release acetylglucosamine from glycans, specifically chitin, had significantly more protein in DIO and db/db relative to B6. Although there is no known chitin in mammalian guts, it has been shown that chitin may be an evolutionary precursor to mucin (Nakashima et al. 2018), and bacterial chitinase when combined with proteases has been shown to have mucolytic properties (Sanders et al. 2007). The xaa-pro dipeptidase ortholog K01271, initially isolated from guinea-pig intestinal mucous (O'Cuinn and Fottrell 1975), was found to have significantly more normalized protein in DIO and db/db mice than B6 mice providing the protease for the chitinase.

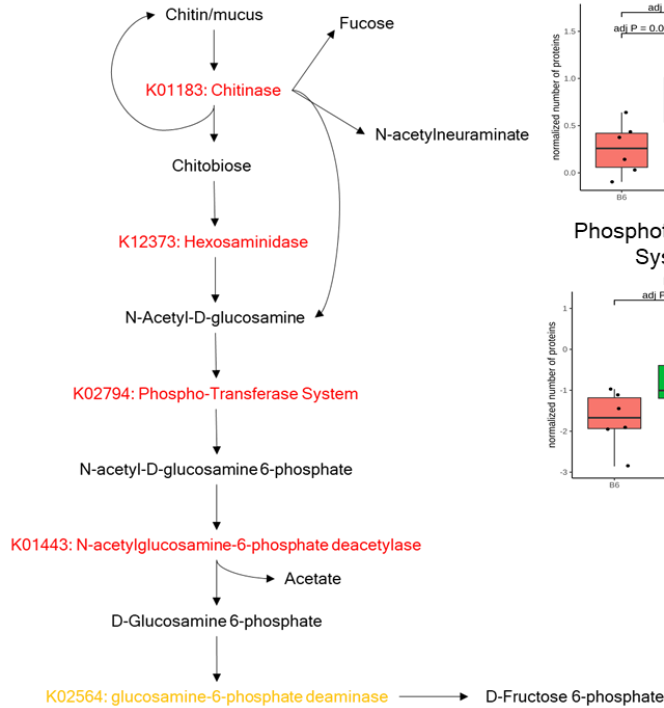
Using KEGG data surrounding the KEGG **amino sugar and nucleotide sugar pathway**, a new pathway from mucin to fructose-6-phosphate via acetylglucosamine was proposed (Figure 5.13). This pathway proposes the release of acetylglucosamine from mucin via chitinase, hexosaminidase, and xaa-pro dipeptidase orthologs, followed by the import of acetylglucosamine into the cell by phosphotransferase system orthologs, the deacetylation of now acetylglucosamine-6-phosphate by N-acetylglucosamine-6-phosphate deacetylase orthologs, and the conversion of glucosamine-6-phosphate to fructose-6-phosphate by glucosamine 6-phosphate deaminase orthologs. The evidence for all of these orthologs came from many different species across the different samples, and all these orthologs had higher relative protein in DIO versus db/db and/or B6.

Some elements of this pathway require some qualification. The traditional phosphotransferase system (PTS) for acetylglucosamine was not significantly different in abundance; however, the mannose PTS system EIIAB component, shown to import acetylglucosamine into microbial cells, was significantly more abundant in DIO relative B6 and

A.



B.



C.

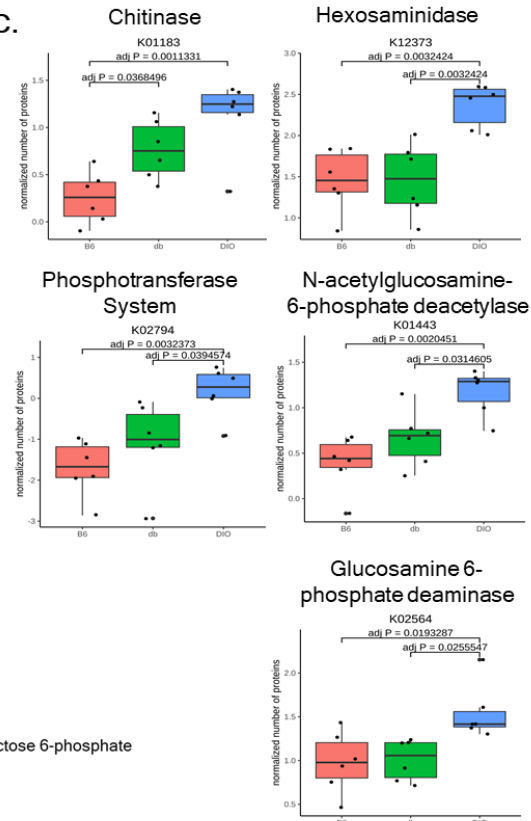


Figure 5.13: Counting number of proteins in KEGG orthologs reveals a complete pathway from mucus to D-Fructose 6-phosphate that is more abundant in DIO than B6 or db. (A) Boxplots supporting the notion that high fat drives the microbiome towards starvation stress and degradation of mucin. (B) Proposed pathway from mucin to fructose-6-phosphate. (C) Specific KEGG orthologs supporting the pathway. Here KEGG orthologs were filtered out if they did not have 7 proteins in at least one sample. Then they were normalized by centered-log-ratio. Finally p-values were calculated by TukeyHSD following two-way ANOVA.

db/db (Moye et al. 2014). Glucosamine 6-phosphate deaminase, the final step of the pathway, was not significant by Benjamini-Hochberg procedure, but was by P-value ($P=0.01$).

Increased foraging of the host mucosal layer due to poor nutritional diets has been shown for low fiber and westernized diet in several studies (Desai et al. 2016, Schroeder et al. 2018). Though it has been suggested that the capacity for mucosal foraging may be widespread, these studies for the most part are species driven, and specifically *Bacteroidetes thetaiotamicron* and *Akkermansia muciniphila* have been shown to forage the mucosal layer for nutrients (Sonnenburg et al. 2005, Berry et al. 2013). Both *Bacteroidetes thetaiotamicron* and *Akkermansia muciniphila* were detected in this study, and several of the orthologs described above were detected for these species, but they were not the primary driving factor for the observation that high fat diet drives the microbiome towards host foraging. An important component of the pathway described above was that it included the incorporation of free acetylglucosamine into the cell prior to being incorporated into the central metabolism via glycolysis. This suggests that acetylglucosamine could be cleaved from mucus by one species and used for energy by another. These results suggest that transition towards host foraging is not driven by any one species but instead by a bulk microbiome concerted effort.

5.4 Discussion

This study sought to compare the effects of morphine on the composition and function of the gut microbiome in the context of two different kinds of obesity. Although a few studies have observed morphine to have a significant impact on the composition of the gut microbiome (Acharya et al. 2017, Wang et al. 2018), the authors are not aware of any studies that have done so in the context of obesity or have examined how morphine changes the functionality of gut microbiota. In this study, high-fat-diet disrupted the gut microbiome's composition by a greater magnitude than morphine or genetic obesity, as evidenced by hierarchical clustering and principle component analysis of taxonomic features independently quantified and analyzed from 16S, metagenomic, and metaproteomic data. Supported by multiple lines of evidence, the already disturbed microbiome in db/db and DIO mice masks the morphine effect, implying that more mild perturbations of the gut microbiome may not be observable when the microbiome has already been disturbed.

As previously observed, specific Lachnospiracea species appeared to be suppressed by morphine and the cytokine IL-6 was more abundant in morphine treated B6 mice relative to saline treated B6 mice (Acharya et al. 2017). Unlike previous observations, *Enterococcus* was not observed to be significantly more abundant in morphine treated mice regardless of line (Wang et al. 2018). Studies have shown that the overall composition of the microbiome varies substantially by individual (Ley et al. 2005, The Human Microbiome Project 2012, Blakeley-Ruiz et al. 2019), which can help to explain why specific taxa may not shift in abundance due to the same perturbation across studies where the starting composition the microbiome might be quite different. The message here is that morphine does indeed seem to alter the gut microbiome composition in a consistent manner within this study, though shifted by a lower magnitude relative to the dramatic shift induced by high-fat diet.

Functionally, morphine appeared to suppress microbial carbohydrate metabolism; however, this effect was only observed at the broad category level not between KEGG orthologs as no samples separated due to morphine treatment in principle component analysis of tabulated KEGG orthologs. Nevertheless, the presence of morphine did appear to suppress or promote proteins of a few host and microbial pathways/orthologs. In particular, the KEGG Prolactin pathway had significantly more proteins in saline treated mice than in morphine treated mice. This was primarily driven by microbial orthologs of the host protein GaLT. Host GaLT is suppressed by prolactin in its short form (Halperin et al. 2008), and prolactin is known to be increased by morphine treatment (Muraki and Tokunaga 1978). This suggests host hormones may also sometimes suppress the microbial orthologs of their host targets. The KEGG ortholog most significantly affected by morphine was the large conductance mechano-sensitive channel, K03282, which protects cells against extreme turgor (Levina et al. 1999). One of the side effects of morphine is increased urinary retention, which in extreme cases can lead to low serum osmolality (hyponatremia), a decrease in serum sodium concentration leading to an excess water relative to solute (Petersen et al. 1982, Petros et al. 1993, Langfeldt and Cooley 2003). This along with the prolactin results suggests that known impacts of morphine on the host may affect microbial functionality as well; however, it seems that this did not manifest in any overall shifts similar to those observed in the compositional analyses. This implies that morphine may have an even more mild effect on microbial function than it does on microbial composition.

A couple of studies have shown high-fat-diet induced obesity to have a greater magnitude effect on microbiome composition than genetically induced obesity (Carmody et al. 2015, Pfalzer et al. 2015). The data from this study supports this assertion by beta-diversity analyses, principle component analyses, and hierarchical clustering in all three of the major omics measured. Furthermore, different species were either suppressed or promoted by high-fat diet or diet induced obesity relative to lean mice. Further supporting the notion that genetic obesity changes the composition of the gut microbiome distinctly from high fat diet.

A few studies have used metaproteomics to explore the effects of obesity on the microbiome (Ferrer et al. 2013, Kolmeder et al. 2015, Zhang et al. 2016, Guirro et al. 2018). To the authors knowledge, however, no study has looked at effects of different kinds of obesity on microbial function from a proteomics perspective. In this study, principle component analysis of KEGG orthologous groups indicated that genetic obese mice and lean mice were indistinguishable, while high-fat-diet clearly separated along the first component implying a change in microbial function due to high-fat-diet not obesity itself. Supporting this interpretation, in most cases, KEGG orthologs marked as significantly different between lines were not significantly different between db and B6, but were different in DIO relative B6 to db. This high fat diet induced change in microbial metabolic function manifested as a decrease in dietary fiber degrading enzymes, and an increase in mucin metabolizing enzymes. High-fat-diet decreases availability of nutrients favored by the microbiome in exchange for fat. This dearth of carbohydrates could lead to starvation induced stress and push gut microbes toward foraging the mucosal layer for sugar. Previous studies have suggested a transition to nutrient foraging from the host mucosal layer due to poor diet with different methods, but those studies have been species focused (Sonnenburg et al. 2005, Berry et al. 2013, Desai et al. 2016, Schroeder et al. 2018). A novel aspect of the presented result is evidence that this transition was not driven by a specific subgroup of species, but by a variety of species from different phylogenetic groupings. For example, 43, 47, 71, and 97 distinct species had protein evidence for hexosaminidase, chitinase, n-acetylglucosamine-6-phosphate deacetylase, and glucosamine 6-phosphate deaminase, respectively, indicating community, not species, driven transitions in function. This functional shift would not be visible when focusing on individual proteins or species instead of community function.

The previous metaproteomic obesity studies did not identify any bulk microbiome shifts in metabolic function. Some studies did notice increases in flagellar complexes and other adhesion related proteins and surmised that this was related to an increase in adhesion to the host extracellular matrix (Ferrer et al. 2013); however, they did not capture that the reason was to degrade mucin. Multiple factors could have contributed to this, including a shallow measured metaproteome (Ferrer et al. 2013), too broad functional classification (Kolmeder et al. 2015, Zhang et al. 2016), ambiguity caused by using BMI as a delimiter (Ferrer et al. 2013, Kolmeder et al. 2015), or splitting the functional analysis between taxa (Guirro et al. 2018). Instead these studies focused substantial effort towards the previously mentioned Firmicutes/Bacteroidetes ratio, ultimately producing conflicting results. In the current study, there were no phyla that were statistically different in relative abundance between conditions. Since the beta-diversity analysis implied specific organisms were more important than phylogenetic groupings, no further effort was given to this topic. *Bacteroides thetaiotaomicron*, a Bacteroidetes, and *Eubacterium* sp. 307_1, a Firmicute, were found to favor high fat diet over lean. Conversely, *Muribaculaceae bacterium* 79_1, a Bacteroidetes, and *Eubacterium* sp. 230_1, a Firmicute, favored lean over high fat diet. While early studies seemed to find that more Bacteroidetes and less Firmicutes were indicative of a healthy gut (Ley et al. 2005, Turnbaugh et al. 2009), this connection does not seem to hold up in recent studies (Finucane et al. 2014, Walters et al. 2014). Perturbations affect bacterial species, not their phylogenetic groupings.

Despite the existence of mouse microbiome gene catalog(Xiao et al. 2015), and numerous studies exploring the gut microbiome of mice (Ley et al. 2005, Turnbaugh et al. 2008, Zhang et al. 2016), this study found that the majority of genes detected by metagenomic/metaproteomic analysis came from novel species not represented in the NCBI or UniProt repositories. This agrees with recent publications suggesting thousands of new bacterial species continue to be uncovered by metagenomics, even in systems considered to be well defined (Brown et al. 2015, Parks et al. 2017, Pasolli et al. 2019). This highlights the importance of continuing to sequence metagenomic samples even in types of environments that are well studied as it seems that there is still more sequence space to uncover. Thus, highlighting the importance of sequencing DNA from new environmental samples prior to metaproteomic analysis. Although it is possible to detect peptides from reference databases using relatively any sample, those identifications can be entirely erroneous if the proteins present in the sample are not represented in the database

(Timmins-Schiffman et al. 2017). In the hope that the mouse microbiome will be better covered in public repositories in the future, this study has contributed all genomes and their respective proteins to GenBank.

The objective of this study was to uncover differences in gut microbiome composition and function due to morphine treatment in the context of genetic or high-fat-diet induced obesity. Towards that end, this study found several changes in microbial composition that were consistent with previous studies. Furthermore, this study showed that an observable change in microbial composition does not necessarily translate to an observable change in microbiome function. While morphine, genetic obesity, and high-fat-diet were all shown to have a significant impact on total microbial composition, a discernable functional shift was only observable for high-fat-diet when considering KEGG orthologs. Ultimately, this study shows that caution should be employed when attributing a microbiome change to specific taxonomic features. This study highlights the importance of looking at microbiome function concurrently with taxonomic composition. Gut microbiomes can be quite varied between individuals regardless of physiology, microbiome functionality tends to be more stable (The Human Microbiome Project 2012)

6 CONCLUSIONS AND PERSPECTIVES ON METAPROTEOMICS RESEARCH

The goal of this dissertation was to advance gut microbiome research by examining how to analyze changes in metabolic function using metaproteomics data. The first step in this process was to identify a framework by which differences between experimental conditions could be analyzed. In chapter 3, multiple fecal samples taken from five adult individuals throughout a year were measured using metaproteomics. The goal was to investigate the stability of gut metaproteomes across time and individuals. These metaproteomes were highly variable between individuals. Only seven microbial proteins were identified across all samples, and even within an individual as few as 7% of the proteins were identified across all samples collected from that individual. This variation in proteins detected presented an intractable analysis problem. How could the metaproteomes between samples be compared quantitatively with so few common features?

KEGG orthologs were found to be more likely to be shared between samples across time than protein groups, suggesting a better space for comparison between samples. Using the KEGG orthologs to infer gut-specific metabolic modules, paths from carbohydrate and amino acid degradation to short-chain fatty acid production were shown to be present across all samples. These paths were multi-phyletic and different phyla across time, and individuals would drive these metabolic modules while the function persisted. This result suggested that the metabolic modules were not driven by any specific phyla but by the microbial community as a whole. Detailed inspection of the component KEGG orthologs of these modules led to the observation that a network of enzymatic reactions linked together by the reactants and products underpinned these metabolic modules. This observation led to the hypothesis that, guided by metaproteomics, metabolic networks could be used to analyze a gut microbiome's metabolic function independent of preconceived pathways and modules.

In Chapter 4, this hypothesis was tested using a gut metaproteome collected from preterm infants that was recently published as part of a collaboration between the Hettich, Banfield, and Morowitz groups (Xiong et al. 2017, Brown et al. 2018). This metaproteome had several advantages over the previous data set. It remains one of the deepest metaproteomes collected to date, allowing for a more comprehensive representation of microbial metabolism. This metaproteome was underpinned by a thorough effort in genome resolved metagenomics,

allowing proteins to be traced back to their genus, species, or even subspecies of origin. The preterm infant gut microbiome is of lower complexity than the adult gut microbiome, allowing for a more comprehensive view of gut microbiome function. Finally, this study had two experimental groups, allowing for metabolic function to be quantitatively compared between conditions.

In Chapter 4, the proteins detected were quite distinct between samples; however, the reactions that these proteins coded for were much more conserved between samples. Interestingly, reactions were also more conserved than the KEGG orthologs that derived the reactions implying that the specific function was more conserved across samples than the orthologous groupings. These reactions could be linked together by the compounds they catalyze to form mostly connected metabolic networks even after the removal of currency metabolites. Using information from the metaproteome, it was possible to distinguish between reactions that were from the host, the microbiome, or shared between the microbiome and the host, providing a framework for analyzing host and microbiome functionality concurrently. Mixed-effects modeling uncovered metabolic groupings of reactions from both the host and the microbiome that differentiated infants that developed necrotizing enterocolitis from those that did not. This analysis uncovered a pathway from chorismate to the siderophore enterochelin, confirming that key pathways could be uncovered using the compound to reaction data present in KEGG. Predefined KEGG pathways and modules were still indispensable in providing them with biological context.

In Chapter 5, all the technical concepts discussed in the previous chapters were applied to a new system, the mouse gut microbiome. In this chapter, the goal was to examine the effects of morphine on two different types of obesity using an integrated metagenomics/metaproteomics approach. Genome resolved metagenomics revealed the species uncovered from these mice to be underrepresented in NCBI and other public repositories. This result validated the suspicion that concurrent metagenomic and metaproteomic measurements were necessary for the metaproteomics analysis. Differential protein abundance analysis found that significant differences in protein abundance were more indicative of changes in taxonomical abundance than function. Thus, perturbations of microbial function were instead analyzed by quantifying KEGG reactions and KEGG orthologs. Reactions significantly different between conditions could all be found among the significantly different KEGG orthologs, while KEGG orthologs

also covered functions like transportation and sporulation. In this analysis, predefined KEGG pathways and modules were indispensable for inferring the underlying biological meaning of the over or under represented KEGG orthologs. Regardless, a new pathway degrading mucin into fructose-6-phosphate via acetylglucosamine was proposed using compound to reaction data. This pathway could not have been uncovered without further inspection beyond the predefined KEGG pathway information.

Chapter 5 showed that at the taxonomical level, the diet induced obesity gut microbiome was as distinct from the genetic obesity gut microbiome as it was from the lean gut microbiome. Morphine had a detectable effect on the lean mouse gut microbiome that was not observed in the obese mice. The above effects were, for the most part, masked at the functional level, and only the differences caused by diet-induced obesity were particularly striking as a high in fat diet triggered a shift in metabolic sources of sugar.

Ultimately, these chapters provide a framework by which the microbiome, in general, can be compared functionally. Using annotations from KEGG or other databases that contain identifiers for specific biological functions, shifts in specific functions could be compared between experimental conditions using metaproteomic data. These shifts in function could be traced back to their species or even subspecies of origin using genome resolved metagenomic data. More often than not, tracing functions back to species of origin showed different species performing the same function across time and individuals confirming the need to look at shifts in microbiome function concurrent with taxonomical differences.

6.1 Getting a deeper metaproteome

When this dissertation commenced, state-of-the-art for metaproteomics was to detect a few thousand proteins per sample, as exemplified by chapter 3. This was a significant feat driven by the availability of concurrent metagenomes with metaproteomes, high mass accuracy instruments, and chromatography. As this dissertation work progressed, these numbers were improved by an order of magnitude to greater than ten thousand proteins per sample. The original analysis of the preterm infant metaproteomics data reported no more than 6,000 proteins per sample. After re-analysis, the same data reported tens of thousands of proteins per sample. This was driven by a couple of adjustments in how the raw data was processed. Combining multiple metagenomes from the same study instead of just searching against the sample-specific

metagenome yielded an increase in the number of distinct peptides detected per sample, as shown in Figure 2.7. A manuscript published by the Noble lab indicated that calculating the exact p-value as a scoring metric during database searching and combining that with other scoring metrics yielded a higher number of peptides passing the FDR threshold (Howbert and Noble 2014). Finally, the criteria of one unique peptide and at least two different peptides decreased the number of proteins detected by greater than three quarters, without a comparable improvement in FDR, essentially making this criterion too conservative. Together these adjustments led to tens of thousands of proteins detected per sample.

Even so, metaproteomics remains far from the comprehensive detection of proteins in gut microbiomes. Ten thousand proteins detected is only about one to three percent of the proteins predicted by a typical metagenome. While it is fair to assume that not all predicted genes will be expressed as proteins, most predictions assume that there are at least 100,000 to 200,000 distinct proteins in an adult fecal sample at any given time, meaning that metaproteomics still needs to improve by an order of magnitude to achieve comprehensive measurement (Zhang and Figeys 2019). In 16S rRNA sequencing, a rarefaction curve is typically used to show that the microbial community has been adequately sampled. Usually, it takes two orders of magnitude more sequences measured than there are amplicon sequence variants to saturate the measurement. In contrast, a typical metaproteomics measurement currently has no more than 200,000 scans measured in total per sample, and never more than 1 million. To the author's knowledge, no study has attempted to measure 1 million let alone 10 million spectra per sample. An interesting approach would be to use metagenomics to first predict how many proteins need to be detected to comprehensively measure a metaproteome, and then proceed to measure two orders of magnitude more spectra than that number to see if that number can be matched. It should be noted that metaproteomics is, in many ways, an inherently different measurement than sequencing. Sequences are detected via random sampling, while peptides are separated over time and detected in a data dependent manner, thus providing a more systematic albeit still somewhat stochastic sampling. Thus, assumptions about the amount of sampling needed to achieve the same dynamic range in protein measurements versus sequencing measurements may not align one to one. It is still reasonable, however, to assume that if one wants to measure a comprehensive proteome, one should have several times more scans than the number of proteins one expects to detect.

6.2 Modeling gut microbiome metabolism

Microbiome research is reaching the point where the next phase is to be able to comprehensively observe shifts in microbial metabolism over time and across perturbations. One proposal to do this is to create genome-scale metabolic models of the constituent organisms in a microbial community (Magnúsdóttir et al. 2017). The problem with this approach is that it cannot handle the dynamic aspects of the microbiome. No matter how many genome-scale metabolic models are produced, a different set of species will be present in the next individual. A good genome-scale models account for all the genes in a genome. However, it is already expected that all the genes will not be expressed at any given time and that there might be substantial function shared across a gut microbiome, as observed in Chapter 4. Others have suggested using metaproteomics to create context-specific metabolic network reconstructions based on metaproteomics information (Tobalina et al. 2015). This was performed in Chapter 4 yielding mostly connected sample-specific metabolic networks across 91 samples. Thus, the field is within striking distance of metaproteomics guided metabolic network reconstruction followed by metabolic modeling.

Chapter 4 provides a proof of principle for metabolic network reconstruction, but not for metaproteomics guided metabolic modeling. To achieve this, a more constrained community is ideal where the sources of carbon are well-controlled, the entire community is well-defined with complete genomes, and the community does not change in composition over time. Two options for this could be gnotobiotic mice with defined communities or gutless worms with well-known symbionts (Ridaura et al. 2013, Kleiner et al. 2018).

To achieve metaproteomics guided metabolic modeling, a few things are needed. A method for estimating the comprehensiveness of a metaproteomic measurement. A method for controlling the carbon sources of the microbial community so that a flux balancing analyzing analysis can be done. An adjustment to the flux balancing analysis that does not necessarily assume that all the bacteria are continuously increasing in biomass. A method for dynamic compartmentalization of the metabolic network by species. Achieving these needs, however, requires several advances in the field of microbial ecology.

To date, it is not known what percentage of a microbe's genome is expected to be expressed in a community setting. Using defined microbial communities, this could be explored by measuring combined isolates from a community and then the community itself using the same

mass spectrometry approach. This could inform whether the percentage of the proteome from each species detected is driven by the measurement or by a change in the percentage of the proteome that is being used in isolation versus a community setting. Mammals consume a complex diet with multiple carbon sources that could be utilized by different bacteria. Stable isotope fingerprinting and stable isotope labeling are methods by which the carbon source for constituent members of the microbiome can be predicted (Kleiner et al. 2018). Applying this method to a microbial community with a defined diet could aid in the construction of a carbon source medium for a flux balance analysis. The Banfield group has developed a method for predicting the replication rate of bacteria based on the sequencing coverage information of a genome (Brown et al. 2016). This method could be used to determine which bacteria in a community are increasing in biomass versus which bacteria are merely surviving. This could provide a guide for both the compartmentalization of bacteria and for the parameters of a flux balance analysis.

6.3 Final thoughts

It is an exciting time to be studying microbial ecology. Despite tremendous funding initiatives over the last two decades, there is still a lot to learn. In a simple example, if one were to sequence the soil samples from all the trees at the University of Tennessee, many new microbial species would likely be uncovered. The microbiome research community is just beginning to uncover the vast amount of microbial diversity that is out there (Parks et al. 2017, Pasolli et al. 2019). The tools that enable these kinds of analyses are now mature enough that anyone familiar with the Unix command line and access to the internet can implement them. Beyond sequencing, much remains to be understood about how microbial communities function. For example, what percentage of a microbe's genome is expressed in a community setting versus when grown in isolation and to what degree do microbes share metabolites in community settings. Metaproteomics is a powerful tool to answer these pertinent questions. Already it has been shown that as much as 45% of the proteins in a microbe's genome can be detected in a metaproteomics study (Xiong et al. 2017); however, it remains to be determined whether the remaining species in the community are merely not expressing as many genes as that species or if there is an issue with the limit of detection of the mass spectrometry measurement. Resolving

this issue will lead to comprehensive metaproteomic profiling, uncovering the degree to which microbes share metabolic functions.

7 REFERENCES

- Abraham, P., R. Adams, R. J. Giannone, U. Kalluri, P. Ranjan, B. Erickson, M. Shah, G. A. Tuskan and R. L. Hettich (2012). "Defining the Boundaries and Characterizing the Landscape of Functional Genome Expression in Vascular Tissues of Populus using Shotgun Proteomics." Journal of Proteome Research **11**(1): 449-460. DOI: 10.1021/pr200851y.
- Acharya, C., N. S. Betrapally, P. M. Gillevet, R. K. Sterling, H. Akbarali, M. B. White, D. Ganapathy, A. Fagan, M. Sikaroodi and J. S. Bajaj (2017). "Chronic opioid use is associated with altered gut microbiota and predicts readmissions in patients with cirrhosis." Aliment Pharmacol Ther **45**(2): 319-331. DOI: 10.1111/apt.13858.
- Aoki-Kinoshita, K. F. and M. Kanehisa (2009). Using KEGG in the Transition from Genomics to Chemical Genomics. Bioinformatics for Systems Biology. S. Krawetz. Totowa, NJ, Humana Press: 437-452.
- Argentini, A., L. J. E. Goeminne, K. Verheggen, N. Hulstaert, A. Staes, L. Clement and L. Martens (2016). "moFF: a robust and automated approach to extract peptide ion intensities." Nature Methods **13**(12): 964-966. DOI: 10.1038/nmeth.4075.
- Arrieta, M.-C., L. T. Stiemsma, N. Amenyogbe, E. M. Brown and B. Finlay (2014). "The intestinal microbiome in early life: health and disease." Frontiers in immunology **5**: 427-427. DOI: 10.3389/fimmu.2014.00427.
- Bairoch, A. (2000). "The ENZYME database in 2000." Nucleic Acids Res **28**(1): 304-305. DOI: 10.1093/nar/28.1.304.
- Ballegaard, M., A. Bjergstrøm, S. Brøndum, E. Hylander, L. Jensen and K. Ladefoged (1997). "Self-Reported Food Intolerance in Chronic Inflammatory Bowel Disease." Scandinavian Journal of Gastroenterology **32**(6): 569-571. DOI: 10.3109/00365529709025101.

- Banerjee, S., G. Sindberg, F. Wang, J. Meng, U. Sharma, L. Zhang, P. Dauer, C. Chen, J. Dalluge, T. Johnson and S. Roy (2016). "Opioid-induced gut microbial disruption and bile dysregulation leads to gut barrier compromise and sustained systemic inflammation." Mucosal Immunology **9**(6): 1418-1428. DOI: 10.1038/mi.2016.9.
- Bates, S. T., D. Berg-Lyons, J. G. Caporaso, W. A. Walters, R. Knight and N. Fierer (2011). "Examining the global distribution of dominant archaeal populations in soil." The ISME Journal **5**(5): 908-917. DOI: 10.1038/ismej.2010.171.
- Becker, R. M., G. Wu, J. A. Galanko, W. Chen, A. R. Maynor, C. L. Bose and J. M. Rhoads (2000). "Reduced serum amino acid concentrations in infants with necrotizing enterocolitis." The Journal of Pediatrics **137**(6): 785-793. DOI: 10.1067/mpd.2000.109145.
- Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society: Series B (Methodological) **57**(1): 289-300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.
- Berry, D., B. Stecher, A. Schintlmeister, J. Reichert, S. Brugiroux, B. Wild, W. Wanek, A. Richter, I. Rauch, T. Decker, A. Loy and M. Wagner (2013). "Host-compound foraging by intestinal microbiota revealed by single-cell stable isotope probing." PNAS **110**(12): 4720-4725. DOI: 10.1073/pnas.1219247110.
- Bisanz, J. E., V. Upadhyay, J. A. Turnbaugh, K. Ly and P. J. Turnbaugh (2019). "Meta-Analysis Reveals Reproducible Gut Microbiome Alterations in Response to a High-Fat Diet." Cell Host & Microbe **26**(2): 265-272.e264. DOI: <https://doi.org/10.1016/j.chom.2019.06.013>.
- Bisgaard, H., N. Li, K. Bonnelykke, B. L. K. Chawes, T. Skov, G. Paludan-Müller, J. Stokholm, B. Smith and K. A. Krogfelt (2011). "Reduced diversity of the intestinal microbiota during infancy is associated with increased risk of allergic disease at school age." Journal

of Allergy and Clinical Immunology **128**(3): 646-652.e645. DOI:
<https://doi.org/10.1016/j.jaci.2011.04.060>.

Blakeley-Ruiz, J. A., A. R. Erickson, B. L. Cantarel, W. Xiong, R. Adams, J. K. Jansson, C. M. Fraser and R. L. Hettich (2019). "Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes." Microbiome **7**(1): 18.

Bokulich, N. A., B. D. Kaehler, J. R. Rideout, M. Dillon, E. Bolyen, R. Knight, G. A. Huttley and J. Gregory Caporaso (2018). "Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin." Microbiome **6**(1): 90. DOI: 10.1186/s40168-018-0470-z.

Bolyen, E., J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar, Y. Bai, J. E. Bisanz, K. Bittinger, A. Brejnrod, C. J. Brislawn, C. T. Brown, B. J. Callahan, A. M. Caraballo-Rodríguez, J. Chase, E. K. Cope, R. Da Silva, C. Diener, P. C. Dorrestein, G. M. Douglas, D. M. Durall, C. Duvallet, C. F. Edwardson, M. Ernst, M. Estaki, J. Fouquier, J. M. Gauglitz, S. M. Gibbons, D. L. Gibson, A. Gonzalez, K. Gorlick, J. Guo, B. Hillmann, S. Holmes, H. Holste, C. Huttenhower, G. A. Huttley, S. Janssen, A. K. Jarmusch, L. Jiang, B. D. Kaehler, K. B. Kang, C. R. Keefe, P. Keim, S. T. Kelley, D. Knights, I. Koester, T. Kosciorek, J. Kreps, M. G. I. Langille, J. Lee, R. Ley, Y.-X. Liu, E. Loftfield, C. Lozupone, M. Maher, C. Marotz, B. D. Martin, D. McDonald, L. J. McIver, A. V. Melnik, J. L. Metcalf, S. C. Morgan, J. T. Morton, A. T. Naimey, J. A. Navas-Molina, L. F. Nothias, S. B. Orchanian, T. Pearson, S. L. Peoples, D. Petras, M. L. Preuss, E. Priesse, L. B. Rasmussen, A. Rivers, M. S. Robeson, P. Rosenthal, N. Segata, M. Shaffer, A. Shiffer, R. Sinha, S. J. Song, J. R. Spear, A. D. Swafford, L. R. Thompson, P. J. Torres, P. Trinh, A. Tripathi, P. J. Turnbaugh, S. Ul-Hasan, J. J. J. van der Hooft, F. Vargas, Y. Vázquez-Baeza, E. Vogtmann, M. von Hippel, W. Walters, Y. Wan, M. Wang, J. Warren, K. C. Weber, C. H. D. Williamson, A. D. Willis, Z. Z. Xu, J. R.

Zaneveld, Y. Zhang, Q. Zhu, R. Knight and J. G. Caporaso (2019). "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2." Nature Biotechnology 37(8): 852-857. DOI: 10.1038/s41587-019-0209-9.

Bowers, R. M., N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Elie-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, C. The Genome Standards, N. C. Kyrpides, L. Schriml, G. M. Garrity, P. Hugenholtz, G. Sutton, P. Yilmaz, F. Meyer, F. O. Glöckner, J. A. Gilbert, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. Murat Eren, L. Schriml, J. F. Banfield, P. Hugenholtz and T. Woyke (2017). "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea." Nature Biotechnology 35: 725. DOI: 10.1038/nbt.3893

<https://www.nature.com/articles/nbt.3893#supplementary-information>.

Brooks, B., R. S. Mueller, J. C. Young, M. J. Morowitz, R. L. Hettich and J. F. Banfield (2015). "Strain-resolved microbial community proteomics reveals simultaneous aerobic and anaerobic function during gastrointestinal tract colonization of a preterm infant." Frontiers in Microbiology 6: 654-654. DOI: 10.3389/fmicb.2015.00654.

Brown, C. T., L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams and J. F. Banfield (2015). "Unusual biology across a group comprising more than 15% of domain Bacteria." Nature 523(7559): 208-211. DOI: 10.1038/nature14486.

- Brown, C. T., M. R. Olm, B. C. Thomas and J. F. Banfield (2016). "Measurement of bacterial replication rates in microbial communities." Nat Biotechnol **34**(12): 1256-1263. DOI: 10.1038/nbt.3704.
- Brown, C. T., W. Xiong, M. R. Olm, B. C. Thomas, R. Baker, B. Firek, M. J. Morowitz, R. L. Hettich and J. F. Banfield (2018). "Hospitalized Premature Infants Are Colonized by Related Bacterial Strains with Distinct Proteomic Profiles." mBio **9**(2).
- Brown, J. F., A. C. Keates, P. J. Hanson and B. J. Whittle (1993). "Nitric oxide generators and cGMP stimulate mucus secretion by rat gastric mucosal cells." American Journal of Physiology-Gastrointestinal and Liver Physiology **265**(3): G418-G422. DOI: 10.1152/ajpgi.1993.265.3.G418.
- Buchfink, B., C. Xie and D. H. Huson (2015). "Fast and sensitive protein alignment using DIAMOND." Nat Methods **12**(1): 59-60. DOI: 10.1038/nmeth.3176.
- Burger-van Paassen, N., A. Vincent, Patrycja J. Puiman, M. van der Sluis, J. Bouma, G. Boehm, Johannes B. van Goudoever, I. van Seuningen and Ingrid B. Renes (2009). "The regulation of intestinal mucin MUC2 expression by short-chain fatty acids: implications for epithelial protection." Biochemical Journal **420**(2): 211-219. DOI: 10.1042/BJ20082222.
- Cahenzli, J., Y. Köller, M. Wyss, Markus B. Geuking and Kathy D. McCoy (2013). "Intestinal Microbial Diversity during Early-Life Colonization Shapes Long-Term IgE Levels." Cell Host & Microbe **14**(5): 559-570. DOI: <https://doi.org/10.1016/j.chom.2013.10.004>.
- Calatayud, S., M. J. Sanz, A. Canet, R. Bello, F. D. de Rojas and J. V. Esplugues (1999). "Mechanisms of gastroprotection by transdermal nitroglycerin in the rat." Br J Pharmacol **127**(5): 1111-1118. DOI: 10.1038/sj.bjp.0702649.

- Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson and S. P. Holmes (2016). "DADA2: High-resolution sample inference from Illumina amplicon data." Nature Methods **13**(7): 581-583. DOI: 10.1038/nmeth.3869.
- Carmody, R. N., G. K. Gerber, J. M. Luevano, Jr., D. M. Gatti, L. Somes, K. L. Svenson and P. J. Turnbaugh (2015). "Diet dominates host genotype in shaping the murine gut microbiota." Cell host & microbe **17**(1): 72-84. DOI: 10.1016/j.chom.2014.11.010.
- Caspi, R., R. Billington, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, P. E. Midford, Q. Ong, W. K. Ong, S. Paley, P. Subhraveti and P. D. Karp (2018). "The MetaCyc database of metabolic pathways and enzymes." Nucleic Acids Research **46**(D1): D633-D639. DOI: 10.1093/nar/gkx935.
- Cervený, L., A. Strásková, V. Danková, A. Hartlová, M. Cecková, F. Staud and J. Stulík (2013). "Tetratricopeptide Repeat Motifs in the World of Bacterial Pathogens: Role in Virulence Mechanisms." Infection and Immunity **81**(3): 629-635. DOI: 10.1128/IAI.01035-12.
- Chambers, M. C., B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M. Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb and P. Mallick (2012). "A cross-platform toolkit for mass spectrometry and proteomics." Nat Biotechnol **30**(10): 918-920. DOI: 10.1038/nbt.2377.
- Chapman, J. D., D. R. Goodlett and C. D. Masselon (2014). "Multiplexed and data-independent tandem mass spectrometry for global proteome profiling." Mass Spectrometry Reviews **33**(6): 452-470. DOI: 10.1002/mas.21400.

- Chaumeil, P.-A., A. J. Mussig, P. Hugenholtz and D. H. Parks (2019). "GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database." Bioinformatics. DOI: 10.1093/bioinformatics/btz848.
- Clemente, Jose C., Luke K. Ursell, Laura W. Parfrey and R. Knight (2012). "The Impact of the Gut Microbiota on Human Health: An Integrative View." Cell **148**(6): 1258-1270. DOI: <https://doi.org/10.1016/j.cell.2012.01.035>.
- Coleman, D. L. (1978). "Obese and diabetes: Two mutant genes causing diabetes-obesity syndromes in mice." Diabetologia **14**(3): 141-148. DOI: 10.1007/BF00429772.
- Coordinators, N. R. (2018). "Database resources of the National Center for Biotechnology Information." Nucleic acids research **46**(D1): D8-D13. DOI: 10.1093/nar/gkx1095.
- Courchesne, P. L., M. D. Jones, J. H. Robinson, C. S. Spahr, S. McCracken, D. L. Bentley, R. Luethy and S. D. Patterson (1998). "Optimization of capillary chromatography ion trap-mass spectrometry for identification of gel-separated proteins." Electrophoresis **19**(6): 956-967. DOI: 10.1002/elps.1150190611.
- Crusoe, M. R., H. F. Alameldin, S. Awad, E. Boucher, A. Caldwell, R. Cartwright, A. Charbonneau, B. Constantinides, G. Edverson, S. Fay, J. Fenton, T. Fenzl, J. Fish, L. Garcia-Gutierrez, P. Garland, J. Gluck, I. González, S. Guermond, J. Guo, A. Gupta, J. R. Herr, A. Howe, A. Hyer, A. Härpfer, L. Irber, R. Kidd, D. Lin, J. Lippi, T. Mansour, P. McA'Nulty, E. McDonald, J. Mizzi, K. D. Murray, J. R. Nahum, K. Nanlohy, A. J. Nederbragt, H. Ortiz-Zuazaga, J. Ory, J. Pell, C. Pepe-Ranne, Z. N. Russ, E. Schwarz, C. Scott, J. Seaman, S. Sievert, J. Simpson, C. T. Skennerton, J. Spencer, R. Srinivasan, D. Standage, J. A. Stapleton, S. R. Steinman, J. Stein, B. Taylor, W. Trimble, H. L. Wiencko, M. Wright, B. Wyss, Q. Zhang, e. zyme and C. T. Brown (2015). "The khmer software package: enabling efficient nucleotide sequence analysis [version 1; referees: 2 approved, 1 approved with reservations]." F1000Research **4**(900). DOI: doi: 10.12688/f1000research.6924.1.

Darzi, Y., G. Falony, S. Vieira-Silva and J. Raes (2015). "Towards biome-specific analysis of meta-omics data." The Isme Journal **10**: 1025. DOI: 10.1038/ismej.2015.188

<https://www.nature.com/articles/ismej2015188#supplementary-information>.

Das, B., T. S. Ghosh, S. Kedia, R. Rampal, S. Saxena, S. Bag, R. Mitra, M. Dayal, O. Mehta, A. Surendranath, S. P. L. Travis, P. Tripathi, G. B. Nair and V. Ahuja (2018). "Analysis of the Gut Microbiome of Rural and Urban Healthy Indians Living in Sea Level and High Altitude Areas." Scientific Reports **8**(1): 10104. DOI: 10.1038/s41598-018-28550-3.

David, L. A., C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, S. B. Biddinger, R. J. Dutton and P. J. Turnbaugh (2013). "Diet rapidly and reproducibly alters the human gut microbiome." Nature **505**: 559. DOI: 10.1038/nature12820

<https://www.nature.com/articles/nature12820#supplementary-information>.

den Besten, G., K. van Eunen, A. K. Groen, K. Venema, D.-J. Reijngoud and B. M. Bakker (2013). "The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism." Journal of Lipid Research **54**(9): 2325-2340. DOI: 10.1194/jlr.R036012.

Desai, M. S., A. M. Seekatz, N. M. Koropatkin, N. Kamada, C. A. Hickey, M. Wolter, N. A. Pudlo, S. Kitamoto, N. Terrapon, A. Muller, V. B. Young, B. Henrissat, P. Wilmes, T. S. Stappenbeck, G. Núñez and E. C. Martens (2016). "A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility." Cell **167**(5): 1339-1353.e1321. DOI: 10.1016/j.cell.2016.10.043.

Dewhirst, F. E., Z. Shen, M. S. Scimeca, L. N. Stokes, T. Boumenna, T. Chen, B. J. Paster and J. G. Fox (2005). "Discordant 16S and 23S rRNA gene phylogenies for the genus *Helicobacter*: implications for phylogenetic inference and systematics." Journal of bacteriology **187**(17): 6106-6118. DOI: 10.1128/JB.187.17.6106-6118.2005.

Dhariwal, A., J. Chong, S. Habib, I. L. King, L. B. Agellon and J. Xia (2017).

"MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data." Nucleic Acids Res **45**(W1): W180-w188. DOI: 10.1093/nar/gkx295.

Diamant, B. J. and W. S. Noble (2011). "Faster SEQUEST Searching for Peptide Identification from Tandem Mass Spectra." Journal of Proteome Research **10**(9): 3871-3879. DOI: 10.1021/pr101196n.

Dill-McFarland, K. A., Z.-Z. Tang, J. H. Kemis, R. L. Kerby, G. Chen, A. Palloni, T. Sorenson, F. E. Rey and P. Herd (2019). "Close social relationships correlate with human gut microbiota composition." Scientific Reports **9**(1): 703. DOI: 10.1038/s41598-018-37298-9.

Donohoe, Dallas R., N. Garge, X. Zhang, W. Sun, Thomas M. O'Connell, Maureen K. Bunger and Scott J. Bultman (2011). "The Microbiome and Butyrate Regulate Energy Metabolism and Autophagy in the Mammalian Colon." Cell Metabolism **13**(5): 517-526. DOI: <https://doi.org/10.1016/j.cmet.2011.02.018>.

Dorofeyev, A. E., I. V. Vasilenko, O. A. Rassokhina and R. B. Kondratiuk (2013). "Mucosal Barrier in Ulcerative Colitis and Crohn's Disease." Gastroenterology Research and Practice **2013**: 431231. DOI: 10.1155/2013/431231.

Du, Z., Y. Yang, Y. Hu, Y. Sun, S. Zhang, W. Peng, Y. Zhong, X. Huang and W. Kong (2012). "A long-term high-fat diet increases oxidative stress, mitochondrial damage and apoptosis in the inner ear of D-galactose-induced aging rats." Hear Res **287**(1-2): 15-24. DOI: 10.1016/j.heares.2012.04.012.

- Dunn, A. B., S. Jordan, B. J. Baker and N. S. Carlson (2017). "The Maternal Infant Microbiome: Considerations for Labor and Birth." MCN. The American journal of maternal child nursing **42**(6): 318-325. DOI: 10.1097/NMC.0000000000000373.
- Earley, H., G. Lennon, Á. Balfe, J. C. Coffey, D. C. Winter and P. R. O'Connell (2019). "The abundance of Akkermansia muciniphila and its relationship with sulphated colonic mucins in health and ulcerative colitis." Scientific Reports **9**(1): 15683. DOI: 10.1038/s41598-019-51878-3.
- Eckburg, P. B., E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson and D. A. Relman (2005). "Diversity of the Human Intestinal Microbial Flora." Science **308**(5728): 1635-1638. DOI: 10.1126/science.1110591.
- Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**(19): 2460-2461. DOI: 10.1093/bioinformatics/btq461.
- Elias, J. E. and S. P. Gygi (2007). "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry." Nature Methods **4**(3): 207-214. DOI: 10.1038/nmeth1019.
- Elias, J. E. and S. P. Gygi (2010). "Target-decoy search strategy for mass spectrometry-based proteomics." Methods in Molecular Biology **604**: 55-71. DOI: 10.1007/978-1-60761-444-9_5.
- Elisia, I., V. Lam, B. Cho, M. Hay, M. Y. Li, J. Kapeluto, T. Elliott, D. Harris, L. Bu, W. Jia, H. Leung, W. Mohn and G. Krystal (2020). "Exploratory examination of inflammation state, immune response and blood cell composition in a human obese cohort to identify potential markers predicting cancer risk." PLOS ONE **15**(2): e0228633. DOI: 10.1371/journal.pone.0228633.

- Eng, J. K., A. L. McCormack and J. R. Yates (1994). "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database." J Am Soc Mass Spectrom **5**(11): 976-989. DOI: 10.1016/1044-0305(94)80016-2.
- Engevik, M. A., C. N. Morra, D. Röth, K. Engevik, J. K. Spinler, S. Devaraj, S. E. Crawford, M. K. Estes, M. Kalkum and J. Versalovic (2019). "Microbial Metabolic Capacity for Intestinal Folate Production and Modulation of Host Folate Receptors." Frontiers in Microbiology **10**(2305). DOI: 10.3389/fmicb.2019.02305.
- Enright, A. J., S. Van Dongen and C. A. Ouzounis (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Research **30**(7): 1575-1584. DOI: 10.1093/nar/30.7.1575.
- Eren, A. M., Ö. C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin and T. O. Delmont (2015). "Anvi'o: an advanced analysis and visualization platform for 'omics data." PeerJ **3**: e1319. DOI: 10.7717/peerj.1319.
- Erickson, A. R., B. L. Cantarel, R. Lamendella, Y. Darzi, E. F. Mongodin, C. Pan, M. Shah, J. Halfvarson, C. Tysk, B. Henrissat, J. Raes, N. C. Verberkmoes, C. M. Fraser, R. L. Hettich and J. K. Jansson (2012). "Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease." PLOS ONE **7**(11): e49138. DOI: 10.1371/journal.pone.0049138.
- Faith, D. P. (1992). "Conservation evaluation and phylogenetic diversity." Biological Conservation **61**(1): 1-10. DOI: [https://doi.org/10.1016/0006-3207\(92\)91201-3](https://doi.org/10.1016/0006-3207(92)91201-3).
- Fallani, M., S. Amarri, A. Uusijarvi, R. Adam, S. Khanna, M. Aguilera, A. Gil, J. M. Vieites, E. Norin, D. Young, J. A. Scott, J. Doré, C. A. Edwards and t. I. team (2011). "Determinants of the human infant intestinal microbiota after the introduction of first complementary foods in infant samples from five European centres." Microbiology **157**(5): 1385-1392. DOI: <https://doi.org/10.1099/mic.0.042143-0>.

- Ferrer, M., A. Ruiz, F. Lanza, S.-B. Haange, A. Oberbach, H. Till, R. Bargiela, C. Campoy, M. T. Segura, M. Richter, M. von Bergen, J. Seifert and A. Suarez (2013). "Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure." Environmental Microbiology **15**(1): 211-226. DOI: 10.1111/j.1462-2920.2012.02845.x.
- Finn, R. D., T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork, A. J. Bridge, H.-Y. Chang, Z. Dosztányi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G. L. Holliday, H. Huang, X. Huang, I. Letunic, R. Lopez, S. Lu, A. Marchler-Bauer, H. Mi, J. Mistry, D. A. Natale, M. Necci, G. Nuka, C. A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S. C. Potter, N. D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P. D. Thomas, Silvio C. E. Tosatto, C. H. Wu, I. Xenarios, L.-S. Yeh, S.-Y. Young and A. L. Mitchell (2017). "InterPro in 2017—beyond protein family and domain annotations." Nucleic Acids Research **45**(D1): D190-D199. DOI: 10.1093/nar/gkw1107.
- Finucane, M. M., T. J. Sharpton, T. J. Laurent and K. S. Pollard (2014). "A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter." PloS one **9**(1): e84689-e84689. DOI: 10.1371/journal.pone.0084689.
- Flint, H. J., K. P. Scott, P. Louis and S. H. Duncan (2012). "The role of the gut microbiota in nutrition and health." Nature Reviews Gastroenterology & Hepatology **9**(10): 577-589. DOI: 10.1038/nrgastro.2012.156.
- Flores, G. E., J. G. Caporaso, J. B. Henley, J. R. Rideout, D. Domogala, J. Chase, J. W. Leff, Y. Vázquez-Baeza, A. Gonzalez, R. Knight, R. R. Dunn and N. Fierer (2014). "Temporal variability is a personalized feature of the human microbiome." Genome Biol **15**(12): 531. DOI: 10.1186/s13059-014-0531-y.

- Fu, L., B. Niu, Z. Zhu, S. Wu and W. Li (2012). "CD-HIT: accelerated for clustering the next-generation sequencing data." Bioinformatics **28**(23): 3150-3152. DOI: 10.1093/bioinformatics/bts565.
- Gene Ontology, C. (2015). "Gene Ontology Consortium: going forward." Nucleic acids research **43**(Database issue): D1049-D1056. DOI: 10.1093/nar/gku1179.
- Gilbert, J. A., M. J. Blaser, J. G. Caporaso, J. K. Jansson, S. V. Lynch and R. Knight (2018). "Current understanding of the human microbiome." Nature medicine **24**(4): 392-400. DOI: 10.1038/nm.4517.
- Goodrich, J. K., S. C. Di Rienzi, A. C. Poole, O. Koren, W. A. Walters, J. G. Caporaso, R. Knight and R. E. Ley (2014). "Conducting a microbiome study." Cell **158**(2): 250-262. DOI: 10.1016/j.cell.2014.06.037.
- Goodrich, J. K., J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, J. T. Bell, T. D. Spector, A. G. Clark and R. E. Ley (2014). "Human genetics shape the gut microbiome." Cell **159**(4): 789-799. DOI: 10.1016/j.cell.2014.09.053.
- Göpel, W., J. Drese, T. K. Rausch, N. Twisselmann, B. Bohnhorst, A. Müller, A. Franz, A. Ziegler, C. Härtel, E. Herting and N. for the German Neonatal (2018). "Necrotizing enterocolitis and high intestinal iron uptake due to genetic variants." Pediatric Research **83**(1): 57-62. DOI: 10.1038/pr.2017.195.
- Guiro, M., A. Costa, A. Gual-Grau, J. Mayneris-Perxachs, H. Torrell, P. Herrero, N. Canela and L. Arola (2018). "Multi-omics approach to elucidate the gut microbiota activity: Metaproteomics and metagenomics connection." Electrophoresis **39**(13): 1692-1701. DOI: 10.1002/elps.201700476.

- Gulhane, M., L. Murray, R. Lourie, H. Tong, Y. H. Sheng, R. Wang, A. Kang, V. Schreiber, K. Y. Wong, G. Magor, S. Denman, J. Begun, T. H. Florin, A. Perkins, P. Ó. Cuív, M. A. McGuckin and S. Z. Hasnain (2016). "High Fat Diets Induce Colonic Epithelial Cell Stress and Inflammation that is Reversed by IL-22." Scientific Reports **6**(1): 28990. DOI: 10.1038/srep28990.
- Gupta, N. and P. A. Pevzner (2009). "False discovery rates of protein identifications: a strike against the two-peptide rule." Journal of proteome research **8**(9): 4173-4181. DOI: 10.1021/pr9004794.
- Haiko, J. and B. Westerlund-Wikström (2013). "The role of the bacterial flagellum in adhesion and virulence." Biology **2**(4): 1242-1267. DOI: 10.3390/biology2041242.
- Halfvarson, J., C. J. Brislawn, R. Lamendella, Y. Vázquez-Baeza, W. A. Walters, L. M. Bramer, M. D'Amato, F. Bonfiglio, D. McDonald, A. Gonzalez, E. E. McClure, M. F. Dunkleberger, R. Knight and J. K. Jansson (2017). "Dynamics of the human gut microbiome in inflammatory bowel disease." Nature Microbiology **2**: 17004. DOI: 10.1038/nmicrobiol.2017.4
<https://www.nature.com/articles/nmicrobiol20174#supplementary-information>.
- Halperin, J., Y. S. Devi, S. Elizur, C. Stocco, A. Shehu, D. Rebourcet, T. G. Unterman, N. D. Leslie, J. Le, N. Binart and G. Gibori (2008). "Prolactin signaling through the short form of its receptor represses forkhead transcription factor FOXO3 and its target gene galt causing a severe ovarian defect." Molecular endocrinology **22**(2): 513-522. DOI: 10.1210/me.2007-0399.
- Halpern, M. D. and B. Dvorak (2008). "Does abnormal bile acid metabolism contribute to NEC?" Seminars in perinatology **32**(2): 114-121. DOI: 10.1053/j.semperi.2008.01.005.
- Heßlinger, C., S. A. Fairhurst and G. Sawers (1998). "Novel keto acid formate-lyase and propionate kinase enzymes are components of an anaerobic pathway in Escherichia coli

- that degrades L-threonine to propionate." Molecular Microbiology **27**(2): 477-492. DOI: 10.1046/j.1365-2958.1998.00696.x.
- Hinzke, T., A. Kouris, R.-A. Hughes, M. Strous and M. Kleiner (2019). "More Is Not Always Better: Evaluation of 1D and 2D-LC-MS/MS Methods for Metaproteomics." Frontiers in Microbiology **10**: 238.
- Hitt, H. C., R. C. McMillen, T. Thornton-Neaves, K. Koch and A. G. Cosby (2007). "Comorbidity of obesity and pain in a general population: results from the Southern Pain Prevalence Study." J Pain **8**(5): 430-436. DOI: 10.1016/j.jpain.2006.12.003.
- Hodzic, Z., A. M. Bolock and M. Good (2017). "The Role of Mucosal Immunity in the Pathogenesis of Necrotizing Enterocolitis." Frontiers in Pediatrics **5**(40). DOI: 10.3389/fped.2017.00040.
- Holmen Larsson, J. M., K. A. Thomsson, A. M. Rodriguez-Pineiro, H. Karlsson and G. C. Hansson (2013). "Studies of mucus in mouse stomach, small intestine, and colon. III. Gastrointestinal Muc5ac and Muc2 mucin O-glycan patterns reveal a regiospecific distribution." Am J Physiol Gastrointest Liver Physiol **305**(5): G357-363. DOI: 10.1152/ajpgi.00048.2013.
- Howbert, J. J. and W. S. Noble (2014). "Computing exact p-values for a cross-correlation shotgun proteomics score function." Mol Cell Proteomics **13**(9): 2467-2479. DOI: 10.1074/mcp.O113.036327.
- Huerta-Cepas, J., K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering and P. Bork (2017). "Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper." Molecular Biology and Evolution **34**(8): 2115-2122. DOI: 10.1093/molbev/msx148.

- Huerta-Cepas, J., D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, M. Kuhn, L. J. Jensen, C. von Mering and P. Bork (2016). "eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences." Nucleic Acids Research **44**(D1): D286-D293. DOI: 10.1093/nar/gkv1248.
- Huerta-Cepas, J., D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, Lars J. Jensen, C. von Mering and P. Bork (2018). "eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses." Nucleic Acids Research **47**(D1): D309-D314. DOI: 10.1093/nar/gky1085.
- Hulzebos, C. V., A. G. J. F. van Zoonen, J. B. F. Hulscher, T. E. Schat, E. M. W. Kooi, M. Koehorst, R. Boverhof, P. F. M. Krabbe, A. K. Groen and H. J. Verkade (2017). "Fecal Bile Salts and the Development of Necrotizing Enterocolitis in Preterm Infants." PLOS ONE **12**(1): e0168633-e0168633. DOI: 10.1371/journal.pone.0168633.
- Hunt, D. F., J. R. Yates, J. Shabanowitz, S. Winston and C. R. Hauer (1986). "Protein sequencing by tandem mass spectrometry." PNAS **83**(17): 6233. DOI: 10.1073/pnas.83.17.6233.
- Huson, D. H., A. F. Auch, J. Qi and S. C. Schuster (2007). "MEGAN analysis of metagenomic data." Genome Research **17**(3): 377-386. DOI: 10.1101/gr.5969107.
- Huss, M. and P. Holme (2007) "Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks." IET Systems Biology **1**, 280-285.
- Hyatt, D., G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer and L. J. Hauser (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification." BMC Bioinformatics **11**(1): 119. DOI: 10.1186/1471-2105-11-119.

- Iraporda, C., A. Errea, D. E. Romanin, D. Cayet, E. Pereyra, O. Pignataro, J. C. Sirard, G. L. Garrote, A. G. Abraham and M. Rumbo (2015). "Lactate and short chain fatty acids produced by microbial fermentation downregulate proinflammatory responses in intestinal epithelial cells and myeloid cells." Immunobiology **220**(10): 1161-1169. DOI: <https://doi.org/10.1016/j.imbio.2015.06.004>.
- Jaccard, P. (1908). "Nouvelles recherches sur la distribution florale." Bull. Soc. V and. Sci. Nat **44**: 223-270.
- Jagtap, P., J. Goslinga, J. A. Kooren, T. McGowan, M. S. Wroblewski, S. L. Seymour and T. J. Griffin (2013). "A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies." Proteomics **13**(8): 1352-1357. DOI: 10.1002/pmic.201200352.
- Jakobsdottir, G., J. Xu, G. Molin, S. Ahn   and M. Nyman (2013). "High-Fat Diet Reduces the Formation of Butyrate, but Increases Succinate, Inflammation, Liver Fat and Cholesterol in Rats, while Dietary Fibre Counteracts These Effects." PLOS ONE **8**(11): e80476. DOI: 10.1371/journal.pone.0080476.
- Jiang, P., B. Smith, N. Qvist, C. Nielsen, J. M.-F. Wan, W.-H. Sit, T. K. Jensen, H. Wang and P. T. Sangild (2013). "Intestinal proteome changes during infant necrotizing enterocolitis." Pediatric Research **73**(3): 268-276. DOI: 10.1038/pr.2012.182.
- Johansson, M. E. V. (2012). "Fast Renewal of the Distal Colonic Mucus Layers by the Surface Goblet Cells as Measured by In Vivo Labeling of Mucin Glycoproteins." PLOS ONE **7**(7): e41009. DOI: 10.1371/journal.pone.0041009.
- Johansson, M. E. V., H. E. Jakobsson, J. Holm  n-Larsson, A. Sch  tte, A. Ermund, A. M. Rodr  guez-Pi  eiro, L. Arike, C. Wising, F. Svensson, F. B  ckhed and G. C. Hansson (2015). "Normalization of Host Intestinal Mucus Layers Requires Long-Term Microbial Colonization." Cell Host & Microbe **18**(5): 582-592. DOI: 10.1016/j.chom.2015.10.007.

- Johansson, M. E. V., H. Sjövall and G. C. Hansson (2013). "The gastrointestinal mucus system in health and disease." Nature Reviews Gastroenterology & Hepatology **10**(6): 352-361. DOI: 10.1038/nrgastro.2013.35.
- Jorth, P., K. H. Turner, P. Gumus, N. Nizam, N. Buduneli and M. Whiteley (2014). "Metatranscriptomics of the Human Oral Microbiome during Health and Disease." mBio **5**(2).
- Käll, L., J. D. Canterbury, J. Weston, W. S. Noble and M. J. MacCoss (2007). "Semi-supervised learning for peptide identification from shotgun proteomics datasets." Nature Methods **4**(11): 923-925. DOI: 10.1038/nmeth1113.
- Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato and K. Morishima (2017). "KEGG: new perspectives on genomes, pathways, diseases and drugs." Nucleic Acids Research **45**(D1): D353-D361. DOI: 10.1093/nar/gkw1092.
- Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe (2016). "KEGG as a reference resource for gene and protein annotation." Nucleic Acids Research **44**(D1): D457-D462. DOI: 10.1093/nar/gkv1070.
- Kanehisa, M., Y. Sato and K. Morishima (2016). "BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences." Journal of Molecular Biology **428**(4): 726-731. DOI: <https://doi.org/10.1016/j.jmb.2015.11.006>.
- Kang, D. D., F. Li, E. Kirton, A. Thomas, R. Egan, H. An and Z. Wang (2019). "MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies." PeerJ **7**: e7359. DOI: 10.7717/peerj.7359.

- Karas, M., D. Bachmann and F. Hillenkamp (1985). "Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules." Analytical Chemistry **57**(14): 2935-2939. DOI: 10.1021/ac00291a042.
- Karu, N., L. Deng, M. Slae, A. C. Guo, T. Sajed, H. Huynh, E. Wine and D. S. Wishart (2018). "A review on human fecal metabolomics: Methods, applications and the human fecal metabolome database." Analytica Chimica Acta **1030**: 1-24. DOI: <https://doi.org/10.1016/j.aca.2018.05.031>.
- Kellner, E. M., A. Decatur and C. P. Moran, Jr. (1996). "Two-stage regulation of an anti-sigma factor determines developmental fate during bacterial endospore formation." Mol Microbiol **21**(5): 913-924. DOI: 10.1046/j.1365-2958.1996.461408.x.
- Kim, D., L. Song, F. P. Breitwieser and S. L. Salzberg (2016). "Centrifuge: rapid and sensitive classification of metagenomic sequences." Genome research **26**(12): 1721-1729. DOI: 10.1101/gr.210641.116.
- Kim, M.-S. and A. Pandey (2012). "Electron transfer dissociation mass spectrometry in proteomics." Proteomics **12**(4-5): 530-542. DOI: 10.1002/pmic.201100517.
- Kleiner, M. (2019). "Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities." mSystems **4**(3): e00115-00119. DOI: 10.1128/mSystems.00115-19.
- Kleiner, M., X. Dong, T. Hinzke, J. Wippler, E. Thorson, B. Mayer and M. Strous (2018). "Metaproteomics method to determine carbon sources and assimilation pathways of species in microbial communities." PNAS **115**(24): E5576. DOI: 10.1073/pnas.1722325115.

- Kleiner, M., E. Thorson, C. E. Sharp, X. Dong, D. Liu, C. Li and M. Strous (2017). "Assessing species biomass contributions in microbial communities via metaproteomics." Nature Communications **8**(1): 1558. DOI: 10.1038/s41467-017-01544-x.
- Knapp, S., A. Kehring, J. Stepp, C. M. Calton, S. M. Gephart, S. Bandlamuri, K. E. Boyle, G. I. Dietz, H. Johnson, R. E. Romo, M. Spencer, A. D. Bedrick and M. D. Halpern (2020). "Elevated Coefficient of Variation in Total Fecal Bile Acids Precedes Diagnosis of Necrotizing Enterocolitis." Scientific Reports **10**(1): 249. DOI: 10.1038/s41598-019-57178-0.
- Knight, R., A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciulek, L.-I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso and P. C. Dorrestein (2018). "Best practices for analysing microbiomes." Nature Reviews Microbiology **16**(7): 410-422. DOI: 10.1038/s41579-018-0029-9.
- Kolmeder, C. A., M. de Been, J. Nikkilä, I. Ritamo, J. Mättö, L. Valmu, J. Salojärvi, A. Palva, A. Salonen and W. M. de Vos (2012). "Comparative Metaproteomics and Diversity Analysis of Human Intestinal Microbiota Testifies for Its Temporal Stability and Expression of Core Functions." PLOS ONE **7**(1): e29913. DOI: 10.1371/journal.pone.0029913.
- Kolmeder, C. A., J. Ritari, F. J. Verdam, T. Muth, S. Keskitalo, M. Varjosalo, S. Fuentes, J. W. Greve, W. A. Buurman, U. Reichl, E. Rapp, L. Martens, A. Palva, A. Salonen, S. S. Rensen and W. M. de Vos (2015). "Colonic metaproteomic signatures of active bacteria and the host in obesity." Proteomics **15**(20): 3544-3552. DOI: 10.1002/pmic.201500049.
- Kolmeder, C. A., J. Salojärvi, J. Ritari, M. de Been, J. Raes, G. Falony, S. Vieira-Silva, R. A. Kekkonen, G. L. Corthals, A. Palva, A. Salonen and W. M. de Vos (2016). "Faecal Metaproteomic Analysis Reveals a Personalized and Stable Functional Microbiome and

- Limited Effects of a Probiotic Intervention in Adults." PLOS ONE **11**(4): e0153294. DOI: 10.1371/journal.pone.0153294.
- Koropatkin, N. M., E. A. Cameron and E. C. Martens (2012). "How glycan metabolism shapes the human gut microbiota." Nature Reviews Microbiology **10**(5): 323-335. DOI: 10.1038/nrmicro2746.
- Kumar, S. and M. L. Blaxter (2010). "Comparing de novo assemblers for 454 transcriptome data." BMC Genomics **11**(1): 571. DOI: 10.1186/1471-2164-11-571.
- Lanas, A. (2008). "Role of nitric oxide in the gastrointestinal tract." Arthritis Research & Therapy **10 Suppl 2**(Suppl 2): S4-S4. DOI: 10.1186/ar2465.
- Langfeldt, L. A. and M. E. Cooley (2003). "Syndrome of inappropriate antidiuretic hormone secretion in malignancy: review and implications for nursing management." Clin J Oncol Nurs **7**(4): 425-430. DOI: 10.1188/03.Cjon.425-430.
- Lauber, C. L., N. Zhou, J. I. Gordon, R. Knight and N. Fierer (2010). "Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples." FEMS Microbiology Letters **307**(1): 80-86. DOI: 10.1111/j.1574-6968.2010.01965.x.
- Le Chatelier, E., T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J. M. Batto, S. Kennedy, P. Leonard, J. Li, K. Burgdorf, N. Garup, T. Jørgensen, I. Brandslund, H. B. Nielsen, A. S. Juncker, M. Bertalan, F. Levenez, N. Pons, S. Rasmussen, S. Sunagawa, J. Tap, S. Tims, E. G. Zoetendal, S. Brunak, K. Clément, J. Doré, M. Kleerebezem, K. Kristiansen, P. Renault, T. Sicheritz-Ponten, W. M. de Vos, J. D. Zucker, J. Raes, T. Hansen, P. Bork, J. Wang, S. D. Ehrlich and O. Pedersen (2013). "Richness of human gut microbiome correlates with metabolic markers." Nature **500**(7464): 541-546. DOI: 10.1038/nature12506.

- Levina, N., S. Tötemeyer, N. R. Stokes, P. Louis, M. A. Jones and I. R. Booth (1999). "Protection of *Escherichia coli* cells against extreme turgor by activation of MscS and MscL mechanosensitive channels: identification of genes required for MscS activity." The EMBO journal **18**(7): 1730-1737. DOI: 10.1093/emboj/18.7.1730.
- Ley, R. E., F. Bäckhed, P. Turnbaugh, C. A. Lozupone, R. D. Knight and J. I. Gordon (2005). "Obesity alters gut microbial ecology." PNAS **102**(31): 11070-11075. DOI: 10.1073/pnas.0504978102.
- Ley, R. E., P. J. Turnbaugh, S. Klein and J. I. Gordon (2006). "Human gut microbes associated with obesity." Nature **444**(7122): 1022-1023. DOI: 10.1038/4441022a.
- Li, D., C.-M. Liu, R. Luo, K. Sadakane and T.-W. Lam (2015). "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph." Bioinformatics **31**(10): 1674-1676. DOI: 10.1093/bioinformatics/btv033.
- Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics **25**(14): 1754-1760. DOI: 10.1093/bioinformatics/btp324.
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079. DOI: 10.1093/bioinformatics/btp352.
- Li, L., C. J. Stoeckert, Jr. and D. S. Roos (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Research **13**(9): 2178-2189. DOI: 10.1101/gr.1224503.
- Li, W. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics **22**(13): 1658-1659. DOI: 10.1093/bioinformatics/btl158.

- Li, Y., J. Hugenholtz, T. Abee and D. Molenaar (2003). "Glutathione protects *Lactococcus lactis* against oxidative stress." Applied and Environmental Microbiology **69**(10): 5739-5745. DOI: 10.1128/aem.69.10.5739-5745.2003.
- Liang, X., J. M. Whitham, E. K. Holwerda, X. Shao, L. Tian, Y.-W. Wu, V. Lombard, B. Henrissat, D. M. Klingeman, Z. K. Yang, M. Podar, T. L. Richard, J. G. Elkins, S. D. Brown and L. R. Lynd (2018). "Development and characterization of stable anaerobic thermophilic methanogenic microbiomes fermenting switchgrass at decreasing residence times." Biotechnology for Biofuels **11**(1): 243. DOI: 10.1186/s13068-018-1238-1.
- Lin, H. V., A. Frassetto, E. J. Kowalik Jr, A. R. Nawrocki, M. M. Lu, J. R. Kosinski, J. A. Hubert, D. Szeto, X. Yao, G. Forrest and D. J. Marsh (2012). "Butyrate and Propionate Protect against Diet-Induced Obesity and Regulate Gut Hormones via Free Fatty Acid Receptor 3-Independent Mechanisms." PLOS ONE **7**(4): e35240. DOI: 10.1371/journal.pone.0035240.
- Lloyd-Price, J., C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, D. Casero, H. Courtney, A. Gonzalez, T. G. Graeber, A. B. Hall, K. Lake, C. J. Landers, H. Mallick, D. R. Plichta, M. Prasad, G. Rahnavard, J. Sauk, D. Shungin, Y. Vázquez-Baeza, R. A. White, J. Bishai, K. Bullock, A. Deik, C. Dennis, J. L. Kaplan, H. Khalili, L. J. McIver, C. J. Moran, L. Nguyen, K. A. Pierce, R. Schwager, A. Sirota-Madi, B. W. Stevens, W. Tan, J. J. ten Hoeve, G. Weingart, R. G. Wilson, V. Yajnik, J. Braun, L. A. Denson, J. K. Jansson, R. Knight, S. Kugathasan, D. P. B. McGovern, J. F. Petrosino, T. S. Stappenbeck, H. S. Winter, C. B. Clish, E. A. Franzosa, H. Vlamakis, R. J. Xavier, C. Huttenhower and I. Investigators (2019). "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases." Nature **569**(7758): 655-662. DOI: 10.1038/s41586-019-1237-9.

- Locey, K. J. and J. T. Lennon (2016). "Scaling laws predict global microbial diversity." PNAS **113**(21): 5970. DOI: 10.1073/pnas.1521291113.
- Loktionov, A. (2019). "Eosinophils in the gastrointestinal tract and their role in the pathogenesis of major colorectal disorders." World Journal of Gastroenterology **25**(27): 3503-3526. DOI: 10.3748/wjg.v25.i27.3503.
- Lozupone, C. A., M. Hamady, S. T. Kelley and R. Knight (2007). "Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities." Applied and Environmental Microbiology **73**(5): 1576. DOI: 10.1128/AEM.01996-06.
- Lundberg, D. S., S. Yourstone, P. Mieczkowski, C. D. Jones and J. L. Dangl (2013). "Practical innovations for high-throughput amplicon sequencing." Nature Methods **10**(10): 999-1002. DOI: 10.1038/nmeth.2634.
- Ma, Z.-Q., S. Dasari, M. C. Chambers, M. D. Litton, S. M. Sobecki, L. J. Zimmerman, P. J. Halvey, B. Schilling, P. M. Drake, B. W. Gibson and D. L. Tabb (2009). "IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering." Journal of Proteome Research **8**(8): 3872-3881. DOI: 10.1021/pr900360j.
- Maddison, W. P. (1997). "Gene Trees in Species Trees." Systematic Biology **46**(3): 523-536. DOI: 10.1093/sysbio/46.3.523.
- Magnúsdóttir, S., A. Heinken, L. Kutt, D. A. Ravcheev, E. Bauer, A. Noronha, K. Greenhalgh, C. Jäger, J. Baginska, P. Wilmes, R. M. T. Fleming and I. Thiele (2017). "Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota." Nature Biotechnology **35**(1): 81-89. DOI: 10.1038/nbt.3703.
- Mahalhal, A., J. M. Williams, S. Johnson, N. Ellaby, C. A. Duckworth, M. D. Burkitt, X. Liu, G. L. Hold, B. J. Campbell, D. M. Pritchard and C. S. Probert (2018). "Oral iron exacerbates

- colitis and influences the intestinal microbiome." PLOS One **13**(10): e0202460. DOI: 10.1371/journal.pone.0202460.
- Makarov, A., E. Denisov, A. Kholomeev, W. Balschun, O. Lange, K. Strupat and S. Horning (2006). "Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer." Analytical Chemistry **78**(7): 2113-2120. DOI: 10.1021/ac0518811.
- Mann, M. and N. L. Kelleher (2008). "Precision proteomics: The case for high resolution and high mass accuracy." PNAS **105**(47): 18132. DOI: 10.1073/pnas.0800788105.
- Mann, M. and M. Wilm (1994). "Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags." Analytical Chemistry **66**(24): 4390-4399. DOI: 10.1021/ac00096a002.
- Marcobal, A., M. Barboza, Erica D. Sonnenburg, N. Pudlo, Eric C. Martens, P. Desai, Carlito B. Lebrilla, Bart C. Weimer, David A. Mills, J. B. German and Justin L. Sonnenburg (2011). "Bacteroides in the Infant Gut Consume Milk Oligosaccharides via Mucus-Utilization Pathways." Cell Host & Microbe **10**(5): 507-514. DOI: <https://doi.org/10.1016/j.chom.2011.10.007>.
- Marcobal, A., A. M. Southwick, K. A. Earle and J. L. Sonnenburg (2013). "A refined palate: Bacterial consumption of host glycans in the gut." Glycobiology **23**(9): 1038-1046. DOI: 10.1093/glycob/cwt040.
- McIlwain, S., M. Mathews, M. S. Bereman, E. W. Rubel, M. J. MacCoss and W. S. Noble (2012). "Estimating relative abundances of proteins from shotgun proteomics data." BMC Bioinformatics **13**(1): 308. DOI: 10.1186/1471-2105-13-308.
- McIlwain, S., K. Tamura, A. Kertesz-Farkas, C. E. Grant, B. Diamant, B. Frewen, J. J. Howbert, M. R. Hoopmann, L. Käll, J. K. Eng, M. J. MacCoss and W. S. Noble (2014). "Crux:

- Rapid Open Source Protein Tandem Mass Spectrometry Analysis." Journal of Proteome Research **13**(10): 4488-4491. DOI: 10.1021/pr500741y.
- McNeil, N. I. (1984). "The contribution of the large intestine to energy supplies in man." The American Journal of Clinical Nutrition **39**: 338-342.
- Menzel, P., K. L. Ng and A. Krogh (2016). "Fast and sensitive taxonomic classification for metagenomics with Kaiju." Nature Communications **7**(1): 11257. DOI: 10.1038/ncomms11257.
- Mesuer, B., B. Devreese, G. Debyser, M. Aerts, P. Vandamme and P. Dawyndt (2012). "Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples." J Proteome Res **11**(12): 5773-5780. DOI: 10.1021/pr300576s.
- Mi, H., A. Muruganujan, X. Huang, D. Ebert, C. Mills, X. Guo and P. D. Thomas (2019). "Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0)." Nature Protocols **14**(3): 703-721. DOI: 10.1038/s41596-019-0128-8.
- Michalski, A., E. Damoc, J.-P. Hauschild, O. Lange, A. Wiegand, A. Makarov, N. Nagaraj, J. Cox, M. Mann and S. Horning (2011). "Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer." Molecular & Cellular Proteomics **10**(9): M111.011015-M011111.011015. DOI: 10.1074/mcp.M111.011015.
- Morton, J. T., J. Sanders, R. A. Quinn, D. McDonald, A. Gonzalez, Y. Vázquez-Baeza, J. A. Navas-Molina, S. J. Song, J. L. Metcalf, E. R. Hyde, M. Lladser, P. C. Dorrestein and R. Knight (2017). "Balance Trees Reveal Microbial Niche Differentiation." mSystems **2**(1): e00162-00116. DOI: 10.1128/mSystems.00162-16.

- Moya, A. and M. Ferrer (2016). "Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance." Trends in Microbiology **24**(5): 402-413. DOI: <https://doi.org/10.1016/j.tim.2016.02.002>.
- Moye, Z. D., R. A. Burne and L. Zeng (2014). "Uptake and metabolism of N-acetylglucosamine and glucosamine by *Streptococcus mutans*." Applied and Environmental Microbiology **80**(16): 5053-5067. DOI: 10.1128/AEM.00820-14.
- Müller, N. F., P. O. Kaiser, D. Linke, H. Schwarz, T. Riess, A. Schäfer, J. A. Eble and V. A. J. Kempf (2011). "Trimeric autotransporter adhesin-dependent adherence of *Bartonella henselae*, *Bartonella quintana*, and *Yersinia enterocolitica* to matrix components and endothelial cells under static and dynamic flow conditions." Infection and Immunity **79**(7): 2544-2553. DOI: 10.1128/IAI.01309-10.
- Muraki, T. and Y. Tokunaga (1978). "Effects of morphine on the serum prolactin levels of morphine-tolerant and nontolerant male rats and of the in vitro release of pituitary prolactin." Jpn J Pharmacol **28**(6): 803-810. DOI: 10.1254/jjp.28.803.
- Muth, T., A. Behne, R. Heyer, F. Kohrs, D. Benndorf, M. Hoffmann, M. Lehtevä, U. Reichl, L. Martens and E. Rapp (2015). "The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation." Journal of Proteome Research **14**(3): 1557-1565. DOI: 10.1021/pr501246w.
- Nagpal, R., R. Mainali, S. Ahmadi, S. Wang, R. Singh, K. Kavanagh, D. W. Kitzman, A. Kushugulova, F. Marotta and H. Yadav (2018). "Gut microbiome and aging: Physiological and mechanistic insights." Nutrition and Healthy Aging **4**(4): 267-285. DOI: 10.3233/NHA-170030.
- Nakashima, K., S. Kimura, Y. Ogawa, S. Watanabe, S. Soma, T. Kaneko, L. Yamada, H. Sawada, C.-H. Tung, T.-M. Lu, J.-K. Yu, A. Villar-Briones, S. Kikuchi and N. Satoh (2018). "Chitin-based barrier immunity and its loss predated mucus-colonization by

- indigenous gut microbiota." Nature Communications **9**(1): 3402. DOI: 10.1038/s41467-018-05884-0.
- Nowicki, P. T., K. M. Reber, P. J. Giannone, C. A. Nankervis, S. Hammond, G. E. Besner and D. A. Caniano (2006). "Intestinal O₂ Consumption in Necrotizing Enterocolitis: Role of Nitric Oxide." Pediatric Research **59**(4): 500-505. DOI: 10.1203/01.pdr.0000203094.27615.5f.
- O'Cuinn, G. and P. F. Fottrell (1975). "Purification and characterization of an aminoacyl proline hydrolase from guinea-pig intestinal mucosa." Biochimica et Biophysica Acta (BBA) - Enzymology **391**(2): 388-395. DOI: [https://doi.org/10.1016/0005-2744\(75\)90262-4](https://doi.org/10.1016/0005-2744(75)90262-4).
- Olm, M. R., N. Bhattacharya, A. Crits-Christoph, B. A. Firek, R. Baker, Y. S. Song, M. J. Morowitz and J. F. Banfield (2019). "Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria." Science Advances **5**(12): eaax5727. DOI: 10.1126/sciadv.aax5727.
- Olm, M. R., C. T. Brown, B. Brooks and J. F. Banfield (2017). "dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication." The ISME Journal **11**: 2864. DOI: 10.1038/ismej.2017.126
<https://www.nature.com/articles/ismej2017126#supplementary-information>.
- Olm, M. R., A. Crits-Christoph, S. Diamond, A. Lavy, P. B. Matheus Carnevali and J. F. Banfield (2020). "Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries." mSystems **5**(1): e00731-00719. DOI: 10.1128/mSystems.00731-19.
- Olsen, J. V., B. Macek, O. Lange, A. Makarov, S. Horning and M. Mann (2007). "Higher-energy C-trap dissociation for peptide modification analysis." Nature Methods **4**(9): 709-712. DOI: 10.1038/nmeth1060.

- Orth, J. D., I. Thiele and B. Ø. Palsson (2010). "What is flux balance analysis?" Nature Biotechnology **28**(3): 245-248. DOI: 10.1038/nbt.1614.
- Ouzounis, C., P. Bork and C. Sander (1994). "The modular structure of NifU proteins." Trends in Biochemical Sciences **19**(5): 199-200. DOI: [https://doi.org/10.1016/0968-0004\(94\)90021-3](https://doi.org/10.1016/0968-0004(94)90021-3).
- Page, M. G. P. (2019). "The Role of Iron and Siderophores in Infection, and the Development of Siderophore Antibiotics." Clinical Infectious Diseases **69**(Supplement_7): S529-S537. DOI: 10.1093/cid/ciz825.
- Pan, H., R. Guo, Y. Ju, Q. Wang, J. Zhu, Y. Xie, Y. Zheng, T. Li, Z. Liu, L. Lu, F. Li, B. Tong, L. Xiao, X. Xu, E. L.-H. Leung, R. Li, H. Yang, J. Wang, H. Zhou, H. Jia and L. Liu (2019). "A single bacterium restores the microbiome dysbiosis to protect bones from destruction in a rat model of rheumatoid arthritis." Microbiome **7**(1): 107. DOI: 10.1186/s40168-019-0719-1.
- Pan, Y., Y. Liu, H. Guo, M. S. Jabir, X. Liu, W. Cui and D. Li (2017). "Associations between Folate and Vitamin B12 Levels and Inflammatory Bowel Disease: A Meta-Analysis." Nutrients **9**(4). DOI: 10.3390/nu9040382.
- Parada Venegas, D., M. K. De la Fuente, G. Landskron, M. J. González, R. Quera, G. Dijkstra, H. J. M. Harmsen, K. N. Faber and M. A. Hermoso (2019). "Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases." Frontiers in Immunology **10**(277). DOI: 10.3389/fimmu.2019.00277.
- Parks, D. H., M. Chuvpochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil and P. Hugenholtz (2018). "A standardized bacterial taxonomy based on genome phylogeny

- substantially revises the tree of life." Nature Biotechnology **36**: 996. DOI: 10.1038/nbt.4229
<https://www.nature.com/articles/nbt.4229#supplementary-information>.
- Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz and G. W. Tyson (2015). "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes." Genome Research **25**(7): 1043-1055.
- Parks, D. H., C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz and G. W. Tyson (2017). "Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life." Nature Microbiology **2**(11): 1533-1542. DOI: 10.1038/s41564-017-0012-7.
- Pasolli, E., F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, M. C. Collado, B. L. Rice, C. DuLong, X. C. Morgan, C. D. Golden, C. Quince, C. Huttenhower and N. Segata (2019). "Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle." Cell **176**(3): 649-662.e620. DOI: <https://doi.org/10.1016/j.cell.2019.01.001>.
- Peng, J. and S. P. Gygi (2001). "Proteomics: the move to mixtures." Journal of Mass Spectrometry **36**(10): 1083-1091. DOI: 10.1002/jms.229.
- Peng, Y., H. C. M. Leung, S. M. Yiu and F. Y. L. Chin (2012). "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth." Bioinformatics **28**(11): 1420-1428. DOI: 10.1093/bioinformatics/bts174.
- Petersen, T. K., S. E. Husted, L. Rybro, B. A. Schurizek and M. Wernberg (1982). "URINARY RETENTION DURING I.M. AND EXTRADURAL MORPHINE ANALGESIA." British Journal of Anaesthesia **54**(11): 1175-1178. DOI: <https://doi.org/10.1093/bja/54.11.1175>.

Petros, J. G., J. K. Mallen, K. Howe, E. B. Rimm and R. J. Robillard (1993). "Patient-controlled analgesia and postoperative urinary retention after open appendectomy." Surg Gynecol Obstet **177**(2): 172-175.

Pfalzer, A. C., P.-D. C. Nesbeth, L. D. Parnell, L. K. Iyer, Z. Liu, A. V. Kane, C. Y. O. Chen, A. K. Tai, T. A. Bowman, M. S. Obin, J. B. Mason, A. S. Greenberg, S.-W. Choi, J. Selhub, L. Paul and J. W. Crott (2015). "Diet- and Genetically-Induced Obesity Differentially Affect the Fecal Microbiome and Metabolome in Apc1638N Mice." PLOS ONE **10**(8): e0135758-e0135758. DOI: 10.1371/journal.pone.0135758.

Polpitiya, A. D., W. J. Qian, N. Jaitly, V. A. Petyuk, J. N. Adkins, D. G. Camp, 2nd, G. A. Anderson and R. D. Smith (2008). "DAnTE: a statistical tool for quantitative analysis of -omics data." Bioinformatics **24**(13): 1556-1558. DOI: 10.1093/bioinformatics/btn217.

Poret, J. M., F. Souza-Smith, S. J. Marcell, D. A. Gaudet, T. H. Tzeng, H. D. Braymer, L. M. Harrison-Bernard and S. D. Primeaux (2018). "High fat diet consumption differentially affects adipose tissue inflammation and adipocyte size in obesity-prone and obesity-resistant rats." International Journal of Obesity **42**(3): 535-541. DOI: 10.1038/ijo.2017.280.

Poudel, S., R. J. Giannone, A. T. Farmer, S. R. Campagna, A. N. Bible, J. L. Morrell-Falvey, J. G. Elkins and R. L. Hettich (2019). "Integrated Proteomics and Lipidomics Reveal That the Swarming Motility of *Paenibacillus polymyxa* Is Characterized by Phospholipid Modification, Surfactant Deployment, and Flagellar Specialization Relative to Swimming Motility." Frontiers in Microbiology **10**(2594). DOI: 10.3389/fmicb.2019.02594.

Pride, D. T., J. Salzman, M. Haynes, F. Rohwer, C. Davis-Long, R. A. White, 3rd, P. Loomer, G. C. Armitage and D. A. Relman (2012). "Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome." The ISME journal **6**(5): 915-926. DOI: 10.1038/ismej.2011.169.

- Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies and F. O. Glöckner (2007). "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB." Nucleic Acids Research **35**(21): 7188-7196. DOI: 10.1093/nar/gkm864.
- Pundir, S., M. J. Martin and C. O'Donovan (2017). UniProt Protein Knowledgebase. Protein Bioinformatics: From Protein Modifications and Networks to Proteomics. C. H. Wu, C. N. Arighi and K. E. Ross. New York, NY, Springer New York: 41-55.
- Qian, C. and R. L. Hettich (2017). "Optimized Extraction Method To Remove Humic Acid Interferences from Soil Samples Prior to Microbial Proteome Measurements." Journal of Proteome Research **16**(7): 2537-2546. DOI: 10.1021/acs.jproteome.7b00103.
- Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, M. Antolin, F. Artiguenave, H. Blottiere, N. Borruel, T. Bruls, F. Casellas, C. Chervaux, A. Cultrone, C. Delorme, G. Denariáz, R. Dervyn, M. Forte, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, A. Jamet, C. Juste, G. Kaci, M. Kleerebezem, J. Knol, M. Kristensen, S. Layec, K. Le Roux, M. Leclerc, E. Maguin, R. Melo Minardi, R. Oozeer, M. Rescigno, N. Sanchez, S. Tims, T. Torrejon, E. Varela, W. de Vos, Y. Winogradsky, E. Zoetendal, P. Bork, S. D. Ehrlich, J. Wang and H. I. T. C. Meta (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65. DOI: 10.1038/nature08821.
- R Core Team (2017). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing.

- Raghuveer, T. S., E. M. McGuire, S. M. Martin, B. A. Wagner, C. J. Rebouché, G. R. Buettner and J. A. Widness (2002). "Lactoferrin in the Preterm Infants' Diet Attenuates Iron-Induced Oxidation Products." Pediatric Research **52**(6): 964-972. DOI: 10.1203/00006450-200212000-00024.
- Rajilić–Stojanović, M., E. Biagi, H. G. H. J. Heilig, K. Kajander, R. A. Kekkonen, S. Tims and W. M. de Vos (2011). "Global and Deep Molecular Analysis of Microbiota Signatures in Fecal Samples From Patients With Irritable Bowel Syndrome." Gastroenterology **141**(5): 1792-1801. DOI: <https://doi.org/10.1053/j.gastro.2011.07.043>.
- Ravisankar, S., R. Tatum, P. M. Garg, M. Herco, P. S. Shekhawat and Y. H. Chen (2018). "Necrotizing enterocolitis leads to disruption of tight junctions and increase in gut permeability in a mouse model." BMC Pediatrics **18**(1): 372. DOI: 10.1186/s12887-018-1346-x.
- Rea, K., S. M. O'Mahony, T. G. Dinan and J. F. Cryan (2017). The Role of the Gastrointestinal Microbiota in Visceral Pain. Gastrointestinal Pharmacology. B. Greenwood-Van Meerveld. Cham, Springer International Publishing: 269-287.
- Ridaura, V. K., J. J. Faith, F. E. Rey, J. Cheng, A. E. Duncan, A. L. Kau, N. W. Griffin, V. Lombard, B. Henrissat, J. R. Bain, M. J. Muehlbauer, O. Ilkayeva, C. F. Semenkovich, K. Funai, D. K. Hayashi, B. J. Lyle, M. C. Martini, L. K. Ursell, J. C. Clemente, W. Van Treuren, W. A. Walters, R. Knight, C. B. Newgard, A. C. Heath and J. I. Gordon (2013). "Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice." Science **341**(6150).
- Roepstorff, P. and J. Fohlman (1984). "Proposal for a common nomenclature for sequence ions in mass spectra of peptides." Biomedical Mass Spectrometry **11**(11): 601. DOI: 10.1002/bms.1200111109.

- Rohart, F., B. Gautier, A. Singh and K.-A. Lê Cao (2017). "mixOmics: An R package for 'omics feature selection and multiple data integration." PLOS Computational Biology **13**(11): e1005752. DOI: 10.1371/journal.pcbi.1005752.
- Rosner, J. (2014). "Ten Times More Microbial Cells than Body Cells in Humans?" Microbe Magazine **9**: 47-47. DOI: 10.1128/microbe.9.47.2.
- Rossi, M., A. Amaretti and S. Raimondi (2011). "Folate production by probiotic bacteria." Nutrients **3**(1): 118-134. DOI: 10.3390/nu3010118.
- Saito, M. A., E. M. Bertrand, M. E. Duffy, D. A. Gaylord, N. A. Held, W. J. Hervey, R. L. Hettich, P. D. Jagtap, M. G. Janech, D. B. Kinkade, D. H. Leary, M. R. McIlvin, E. K. Moore, R. M. Morris, B. A. Neely, B. L. Nunn, J. K. Saunders, A. I. Shepherd, N. I. Symmonds and D. A. Walsh (2019). "Progress and Challenges in Ocean Metaproteomics and Proposed Best Practices for Data Sharing." Journal of Proteome Research **18**(4): 1461-1476. DOI: 10.1021/acs.jproteome.8b00761.
- Saito, M. A., A. Dorsk, A. F. Post, M. R. McIlvin, M. S. Rappé, G. R. DiTullio and D. M. Moran (2015). "Needles in the blue sea: Sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome." PROTEOMICS **15**(20): 3521-3531. DOI: 10.1002/pmic.201400630.
- Salvatore, S., R. Heuschkel, S. Tomlin, S. E. Davies, S. Edwards, J. A. Walker-Smith, I. French and S. H. Murch (2000). "A pilot study of N-acetyl glucosamine, a nutritional substrate for glycosaminoglycan synthesis, in paediatric chronic inflammatory bowel disease." Alimentary Pharmacology & Therapeutics **14**(12): 1567-1579. DOI: 10.1046/j.1365-2036.2000.00883.x.
- Sanders, N. N., V. G. Eijssink, P. S. van den Pangaart, R. J. Joost van Neerven, P. J. Simons, S. C. De Smedt and J. Demeester (2007). "Mucolytic activity of bacterial and human

- chitinases." Biochim Biophys Acta **1770**(5): 839-846. DOI: 10.1016/j.bbagen.2007.01.011.
- Schoch, C. L., K. A. Seifert, S. Huhndorf, V. Robert, J. L. Spouge, C. A. Levesque, W. Chen, C. Fungal Barcoding and L. Fungal Barcoding Consortium Author (2012). "Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi." PNAS **109**(16): 6241-6246. DOI: 10.1073/pnas.1117018109.
- Schroeder, B. O., G. M. H. Birchenough, M. Ståhlman, L. Arike, M. E. V. Johansson, G. C. Hansson and F. Bäckhed (2018). "Bifidobacteria or Fiber Protects against Diet-Induced Microbiota-Mediated Colonic Mucus Deterioration." Cell Host & Microbe **23**(1): 27-40.e27. DOI: <https://doi.org/10.1016/j.chom.2017.11.004>.
- Semrin, G., D. S. Fishman, A. Bousvaros, A. Zholudev, A. C. Saunders, C. E. Correia, E. Nemeth, R. J. Grand and D. A. Weinstein (2006). "Impaired intestinal iron absorption in Crohn's disease correlates with disease activity and markers of inflammation." Inflammatory Bowel Diseases **12**(12): 1101-1106. DOI: 10.1097/01.mib.0000235097.86360.04.
- Sender, R., S. Fuchs and R. Milo (2016). "Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans." Cell **164**(3): 337-340. DOI: 10.1016/j.cell.2016.01.013.
- Serang, O., M. J. MacCoss and W. S. Noble (2010). "Efficient Marginalization to Compute Protein Posterior Probabilities from Shotgun Mass Spectrometry Data." Journal of Proteome Research **9**(10): 5346-5357. DOI: 10.1021/pr100594k.
- Smith, P. M., M. R. Howitt, N. Panikov, M. Michaud, C. A. Gallini, M. Bohlooly-Y, J. N. Glickman and W. S. Garrett (2013). "The Microbial Metabolites, Short-Chain Fatty Acids, Regulate Colonic T_{reg} Cell Homeostasis." Science **341**(6145): 569.

Song, S. J., A. Amir, J. L. Metcalf, K. R. Amato, Z. Z. Xu, G. Humphrey and R. Knight (2016). "Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies." mSystems **1**(3): e00021-00016. DOI: 10.1128/mSystems.00021-16.

Song, S. J., C. Lauber, E. K. Costello, C. A. Lozupone, G. Humphrey, D. Berg-Lyons, J. G. Caporaso, D. Knights, J. C. Clemente, S. Nakielnny, J. I. Gordon, N. Fierer and R. Knight (2013). "Cohabiting family members share microbiota with one another and with their dogs." eLife **2**: e00458-e00458. DOI: 10.7554/eLife.00458.

Sonnenburg, J. L., J. Xu, D. D. Leip, C.-H. Chen, B. P. Westover, J. Weatherford, J. D. Buhler and J. I. Gordon (2005). "Glycan Foraging in Vivo by an Intestine-Adapted Bacterial Symbiont." Science **307**(5717): 1955. DOI: 10.1126/science.1109051.

Stamatakis, A. (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." Bioinformatics **30**(9): 1312-1313. DOI: 10.1093/bioinformatics/btu033.

Stewart, C. J., N. J. Ajami, J. L. O'Brien, D. S. Hutchinson, D. P. Smith, M. C. Wong, M. C. Ross, R. E. Lloyd, H. Doddapaneni, G. A. Metcalf, D. Muzny, R. A. Gibbs, T. Vatanen, C. Huttenhower, R. J. Xavier, M. Rewers, W. Hagopian, J. Toppari, A.-G. Ziegler, J.-X. She, B. Akolkar, A. Lernmark, H. Hyoty, K. Vehik, J. P. Krischer and J. F. Petrosino (2018). "Temporal development of the gut microbiome in early childhood from the TEDDY study." Nature **562**(7728): 583-588. DOI: 10.1038/s41586-018-0617-x.

Stewart, C. J., N. D. Embleton, E. C. L. Marrs, D. P. Smith, A. Nelson, B. Abdulkadir, T. Skeath, J. F. Petrosino, J. D. Perry, J. E. Berrington and S. P. Cummings (2016). "Temporal bacterial and metabolic development of the preterm gut reveals specific signatures in health and disease." Microbiome **4**(1): 67. DOI: 10.1186/s40168-016-0216-8.

Suchodolski, J. S., C. G. Ruaux, J. M. Steiner, K. Fetz and D. A. Williams (2005). "Assessment of the qualitative variation in bacterial microflora among compartments of the intestinal

- tract of dogs by use of a molecular fingerprinting technique." Am J Vet Res **66**(9): 1556-1562. DOI: 10.2460/ajvr.2005.66.1556.
- Sun, M., W. Wu, Z. Liu and Y. Cong (2017). "Microbiota metabolite short chain fatty acids, GPCR, and inflammatory bowel diseases." Journal of Gastroenterology **52**(1): 1-8. DOI: 10.1007/s00535-016-1242-9.
- Sze, M. A. and P. D. Schloss (2016). "Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome." mBio **7**(4).
- Szöllősi, G. J., E. Tannier, V. Daubin and B. Boussau (2014). "The Inference of Gene Trees with Species Trees." Systematic Biology **64**(1): e42-e62. DOI: 10.1093/sysbio/syu048.
- Tabb, D. L. (2015). "The SEQUEST family tree." Journal of the American Society for Mass Spectrometry **26**(11): 1814-1819. DOI: 10.1007/s13361-015-1201-3.
- Tabb, D. L., C. G. Fernando and M. C. Chambers (2007). "MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis." Journal of Proteome Research **6**(2): 654-661. DOI: 10.1021/pr0604054.
- Tailford, L. E., E. H. Crost, D. Kavanaugh and N. Juge (2015). "Mucin glycan foraging in the human gut microbiome." Frontiers in Genetics **6**(81). DOI: 10.3389/fgene.2015.00081.
- Tanca, A., M. Abbondio, A. Palomba, C. Fraumene, V. Manghina, F. Cucca, E. Fiorillo and S. Uzzau (2017). "Potential and active functions in the gut microbiota of a healthy human cohort." Microbiome **5**(1): 79. DOI: 10.1186/s40168-017-0293-3.
- Tatusov, R. L., M. Y. Galperin, D. A. Natale and E. V. Koonin (2000). "The COG database: a tool for genome-scale analysis of protein functions and evolution." Nucleic Acids Research **28**(1): 33-36. DOI: 10.1093/nar/28.1.33.

- Tedjo, D. I., D. M. A. E. Jonkers, P. H. Savelkoul, A. A. Masclee, N. van Best, M. J. Pierik and J. Penders (2015). "The Effect of Sampling and Storage on the Fecal Microbiota Composition in Healthy and Diseased Subjects." PLOS ONE **10**(5): e0126685. DOI: 10.1371/journal.pone.0126685.
- Telang, S. (2018). "Lactoferrin: A Critical Player in Neonatal Host Defense." Nutrients **10**(9): 1228. DOI: 10.3390/nu10091228.
- Tenenbaum, D. (2020). KEGGREST: Client-side REST access to KEGG, Bioconductor.
- The Human Microbiome Project, C. (2012). "Structure, Function and Diversity of the Healthy Human Microbiome." Nature **486**(7402): 207-214. DOI: 10.1038/nature11234.
- The UniProt, C. (2018). "UniProt: a worldwide hub of protein knowledge." Nucleic Acids Research **47**(D1): D506-D515. DOI: 10.1093/nar/gky1049.
- The UniProt Consortium (2017). "UniProt: the universal protein knowledgebase." Nucleic Acids Research **45**(D1): D158-D169. DOI: 10.1093/nar/gkw1099.
- Timmins-Schiffman, E., D. H. May, M. Mikan, M. Riffle, C. Frazar, H. R. Harvey, W. S. Noble and B. L. Nunn (2017). "Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns." The ISME Journal **11**(2): 309-314. DOI: 10.1038/ismej.2016.132.
- Ting, L., R. Rad, S. P. Gygi and W. Haas (2011). "MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics." Nature Methods **8**(11): 937-940. DOI: 10.1038/nmeth.1714.
- Tobalina, L., R. Bargiela, J. Pey, F.-A. Herbst, I. Lores, D. Rojo, C. Barbas, A. I. Peláez, J. Sánchez, M. von Bergen, J. Seifert, M. Ferrer and F. J. Planes (2015). "Context-specific metabolic network reconstruction of a naphthalene-degrading bacterial community

- guided by metaproteomic data." Bioinformatics **31**(11): 1771-1779. DOI: 10.1093/bioinformatics/btv036.
- Topping, D. L. and P. M. Clifton (2001). "Short-Chain Fatty Acids and Human Colonic Function: Roles of Resistant Starch and Nonstarch Polysaccharides." Physiological Reviews **81**(3): 1031-1064. DOI: 10.1152/physrev.2001.81.3.1031.
- Tung, J., L. B. Barreiro, M. B. Burns, J.-C. Grenier, J. Lynch, L. E. Grieneisen, J. Altmann, S. C. Alberts, R. Blekhman and E. A. Archie (2015). "Social networks predict gut microbiome composition in wild baboons." eLife **4**: e05224. DOI: 10.7554/eLife.05224.
- Turnbaugh, P. J., F. Bäckhed, L. Fulton and J. I. Gordon (2008). "Diet-Induced Obesity Is Linked to Marked but Reversible Alterations in the Mouse Distal Gut Microbiome." Cell Host & Microbe **3**(4): 213-223. DOI: <https://doi.org/10.1016/j.chom.2008.02.015>.
- Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight and J. I. Gordon (2009). "A core gut microbiome in obese and lean twins." Nature **457**(7228): 480-484. DOI: 10.1038/nature07540.
- Tyanova, S., T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann and J. Cox (2016). "The Perseus computational platform for comprehensive analysis of (prote)omics data." Nature Methods **13**: 731. DOI: 10.1038/nmeth.3901
<https://www.nature.com/articles/nmeth.3901#supplementary-information>.
- Van Herreweghen, F., P. Van den Abbeele, T. De Mulder, R. De Weirtdt, A. Geirnaert, E. Hernandez-Sanabria, R. Vilchez-Vargas, R. Jauregui, D. H. Pieper, C. Belzer, W. M. De Vos and T. Van de Wiele (2017). "In vitro colonisation of the distal colon by *Akkermansia muciniphila* is largely mucin and pH dependent." Beneficial Microbes **8**(1): 81-96. DOI: 10.3920/BM2016.0013.

Vanderwinkel, E., P. Furmanski, H. C. Reeves and S. J. Ajl (1968). "Growth of *Escherichia coli* on fatty acids: Requirement for coenzyme A transferase activity." Biochemical and Biophysical Research Communications **33**(6): 902-908. DOI: [https://doi.org/10.1016/0006-291X\(68\)90397-5](https://doi.org/10.1016/0006-291X(68)90397-5).

Verberkmoes, N. C., A. L. Russell, M. Shah, A. Godzik, M. Rosenquist, J. Halfvarson, M. G. Lefsrud, J. Apajalahti, C. Tysk, R. L. Hettich and J. K. Jansson (2008). "Shotgun metaproteomics of the human distal gut microbiota." The ISME Journal **3**: 179. DOI: 10.1038/ismej.2008.108
<https://www.nature.com/articles/ismej2008108#supplementary-information>.

Verma, S. and B. J. Cherayil (2017). "Iron and inflammation - the gut reaction." Metallomics : integrated biometal science **9**(2): 101-111. DOI: 10.1039/c6mt00282j.

Vieira-Silva, S., G. Falony, Y. Darzi, G. Lima-Mendez, R. Garcia Yunta, S. Okuda, D. Vandeputte, M. Valles-Colomer, F. Hildebrand, S. Chaffron and J. Raes (2016). "Species–function relationships shape ecological properties of the human gut microbiome." Nature Microbiology **1**: 16088. DOI: 10.1038/nmicrobiol.2016.88
<https://www.nature.com/articles/nmicrobiol201688#supplementary-information>.

Vinolo, M. A. R., H. G. Rodrigues, R. T. Nachbar and R. Curi (2011). "Regulation of Inflammation by Short Chain Fatty Acids." Nutrients **3**(10): 858-876. DOI: 10.3390/nu3100858.

Visconti, A., C. I. Le Roy, F. Rosa, N. Rossi, T. C. Martin, R. P. Mohn, W. Li, E. de Rinaldis, J. T. Bell, J. C. Venter, K. E. Nelson, T. D. Spector and M. Falchi (2019). "Interplay between the human gut microbiome and host metabolism." Nature Communications **10**(1): 4505. DOI: 10.1038/s41467-019-12476-z.

- Walters, W. A., Z. Xu and R. Knight (2014). "Meta-analyses of human gut microbes associated with obesity and IBD." FEBS Letters **588**(22): 4223-4233. DOI: 10.1016/j.febslet.2014.09.039.
- Wampach, L., A. Heintz-Buschart, A. Hogan, E. E. L. Muller, S. Narayanasamy, C. C. Laczny, L. W. Hugerth, L. Bindl, J. Bottu, A. F. Andersson, C. de Beaufort and P. Wilmes (2017). "Colonization and Succession within the Human Gut Microbiome by Archaea, Bacteria, and Microeukaryotes during the First Year of Life." Frontiers in Microbiology **8**: 738-738. DOI: 10.3389/fmicb.2017.00738.
- Wandersman, C. and P. Delepelaire (2004). "Bacterial Iron Sources: From Siderophores to Hemophores." Annual Review of Microbiology **58**(1): 611-647. DOI: 10.1146/annurev.micro.58.030603.123811.
- Wang, B. and J. Brand-Miller (2003). "The role and potential of sialic acid in human nutrition." European Journal Of Clinical Nutrition **57**: 1351. DOI: 10.1038/sj.ejcn.1601704.
- Wang, F., J. Meng, L. Zhang, T. Johnson, C. Chen and S. Roy (2018). "Morphine induces changes in the gut microbiome and metabolome in a morphine dependence model." Scientific Reports **8**(1): 3596. DOI: 10.1038/s41598-018-21915-8.
- Wang, N. and L. Li (2008). "Exploring the Precursor Ion Exclusion Feature of Liquid Chromatography–Electrospray Ionization Quadrupole Time-of-Flight Mass Spectrometry for Improving Protein Identification in Shotgun Proteome Analysis." Analytical Chemistry **80**(12): 4696-4710. DOI: 10.1021/ac800260w.
- Waridel, P., A. Frank, H. Thomas, V. Surendranath, S. Sunyaev, P. Pevzner and A. Shevchenko (2007). "Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automated de novo sequencing." PROTEOMICS **7**(14): 2318-2329. DOI: 10.1002/pmic.200700003.

- Warnes, G. R., B. Bolker, L. Bonebakker, R. Gentleman, W. H. A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz and B. Venables (2016). gplots: Various R Programming Tools for Plotting Data. <https://CRAN.R-project.org/package=gplots>.
- Washburn, M. P., D. Wolters and J. R. Yates Iii (2001). "Large-scale analysis of the yeast proteome by multidimensional protein identification technology." Nature Biotechnology **19**: 242. DOI: 10.1038/85686
https://www.nature.com/articles/nbt0301_242#supplementary-information.
- Wessling-Resnick, M. (2010). "Iron homeostasis and the inflammatory response." Annual Review of Nutrition **30**: 105-122. DOI: 10.1146/annurev.nutr.012809.104804.
- West, P. T., A. J. Probst, I. V. Grigoriev, B. C. Thomas and J. F. Banfield (2018). "Genome-reconstruction for eukaryotes from complex natural microbial communities." Genome Research **28**(4): 569-580. DOI: 10.1101/gr.228429.117.
- Whitehouse, C. M., R. N. Dreyer, M. Yamashita and J. B. Fenn (1985). "Electrospray interface for liquid chromatographs and mass spectrometers." Analytical Chemistry **57**(3): 675-679. DOI: 10.1021/ac00280a023.
- Wilmes, P. and P. L. Bond (2004). "The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms." Environmental Microbiology **6**(9): 911-920. DOI: 10.1111/j.1462-2920.2004.00687.x.
- Wolfe, A. J. (2005). "The Acetate Switch." Microbiology and Molecular Biology Reviews **69**(1): 12-50. DOI: 10.1128/MMBR.69.1.12-50.2005.
- Xiao, L., Q. Feng, S. Liang, S. B. Sonne, Z. Xia, X. Qiu, X. Li, H. Long, J. Zhang, D. Zhang, C. Liu, Z. Fang, J. Chou, J. Glanville, Q. Hao, D. Kotowska, C. Colding, T. R. Licht, D. Wu, J. Yu, J. J. Y. Sung, Q. Liang, J. Li, H. Jia, Z. Lan, V. Tremaroli, P. Dworzynski, H.

- B. Nielsen, F. Bäckhed, J. Doré, E. Le Chatelier, S. D. Ehrlich, J. C. Lin, M. Arumugam, J. Wang, L. Madsen and K. Kristiansen (2015). "A catalog of the mouse gut metagenome." Nature Biotechnology **33**: 1103. DOI: 10.1038/nbt.3353
<https://www.nature.com/articles/nbt.3353#supplementary-information>.
- Xiong, W., P. E. Abraham, Z. Li, C. Pan and R. L. Hettich (2015). "Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota." PROTEOMICS **15**(20): 3424-3438. DOI: 10.1002/pmic.201400571.
- Xiong, W., C. T. Brown, M. J. Morowitz, J. F. Banfield and R. L. Hettich (2017). "Genome-resolved metaproteomic characterization of preterm infant gut microbiota development reveals species-specific metabolic shifts and variabilities during early life." Microbiome **5**(1): 72. DOI: 10.1186/s40168-017-0290-6.
- Xiong, W., R. J. Giannone, M. J. Morowitz, J. F. Banfield and R. L. Hettich (2015). "Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut." Journal of Proteome Research **14**(1): 133-141. DOI: 10.1021/pr500936p.
- Yassour, M., T. Vatanen, H. Siljander, A.-M. Hämäläinen, T. Härkönen, S. J. Ryhänen, E. A. Franzosa, H. Vlamakis, C. Huttenhower, D. Gevers, E. S. Lander, M. Knip and R. J. Xavier (2016). "Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability." Science Translational Medicine **8**(343): 343ra381. DOI: 10.1126/scitranslmed.aad0917.
- Yates, J. R., 3rd, J. K. Eng, A. L. McCormack and D. Schieltz (1995). "Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database." Analytical Chemistry **67**(8): 1426-1436. DOI: 10.1021/ac00104a020.
- Yatsunencko, T., F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder,

- J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight and J. I. Gordon (2012). "Human gut microbiome viewed across age and geography." Nature **486**(7402): 222-227. DOI: 10.1038/nature11053.
- Young Jacque, C., C. Pan, M. Adams Rachel, B. Brooks, F. Banfield Jillian, J. Morowitz Michael and L. Hettich Robert (2015). "Metaproteomics reveals functional shifts in microbial and human proteins during a preterm infant gut colonization case." PROTEOMICS **15**(20): 3463-3473. DOI: 10.1002/pmic.201400563.
- Zhang, X., S. A. Deeke, Z. Ning, A. E. Starr, J. Butcher, J. Li, J. Mayne, K. Cheng, B. Liao, L. Li, R. Singleton, D. Mack, A. Stintzi and D. Figeys (2018). "Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease." Nature Communications **9**(1): 2873. DOI: 10.1038/s41467-018-05357-4.
- Zhang, X. and D. Figeys (2019). "Perspective and Guidelines for Metaproteomics in Microbiome Studies." Journal of Proteome Research **18**(6): 2370-2380. DOI: 10.1021/acs.jproteome.9b00054.
- Zhang, X., L. Li, J. Mayne, Z. Ning, A. Stintzi and D. Figeys (2018). "Assessing the impact of protein extraction methods for human gut metaproteomics." Journal of Proteomics **180**: 120-127. DOI: 10.1016/j.jprot.2017.07.001.
- Zhang, X., Z. Ning, J. Mayne, J. I. Moore, J. Li, J. Butcher, S. A. Deeke, R. Chen, C. K. Chiang, M. Wen, D. Mack, A. Stintzi and D. Figeys (2016). "MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota." Microbiome **4**(1): 31. DOI: 10.1186/s40168-016-0176-z.
- Zhang, Y., A. Brady, C. Jones, Y. Song, T. C. Darton, C. Jones, C. J. Blohmke, A. J. Pollard, L. S. Magder, A. Fasano, M. B. Sztein and C. M. Fraser (2018). "Compositional and Functional Differences in the Human Gut Microbiome Correlate with Clinical Outcome

following Infection with Wild-Type Salmonella enterica Serovar Typhi." mBio **9**(3): e00686-00618. DOI: 10.1128/mBio.00686-18.

Zhu, Q., U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciulek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab and R. Knight (2019). "Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea." Nature Communications **10**(1): 5477-5477. DOI: 10.1038/s41467-019-13443-4.

Zhu, W., A. Lomsadze and M. Borodovsky (2010). "Ab initio gene identification in metagenomic sequences." Nucleic Acids Research **38**(12): e132-e132. DOI: 10.1093/nar/gkq275.

VITA

Jose Alfredo Blakeley-Ruiz was born on November 11, 1988 to Lucinda Blakeley and Dr. Jose A. Blakeley Perez in Bloomington, IN. Alfredo grew up in Plano, Texas then Redmond, WA. In 2011, Alfredo graduated with a Bachelor of Science degree in Biology from the University of Notre Dame. After graduation, he spent three years working at the Infectious Disease Research Institute, in Seattle, on early drug discovery for tuberculosis. He applied and was admitted to the Genome Science and Technology program at University of Tennessee Knoxville campus in 2014. This dissertation will culminate in acquiring a PhD in Biological Sciences from the University of Tennessee, Knoxville in 2020 under the direction of Dr. Robert L. Hettich. Jose Alfredo Blakeley-Ruiz was nominated in Summer 2020 to serve on the College of Science Advisory Council for the University of Notre Dame as a Young Alumni, and will do his initial Postdoctoral work at North Carolina State University, Raleigh under the direction of Dr. Manuel Kleiner.