UNIVERSITY OF CALIFORNIA SAN DIEGO

Integrating multiplexed metaproteomics to discover novel therapeutic avenues targeting the IBD microbiota

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biomedical Sciences

by

Robert Hardie Mills

Committee in charge:

Professor David J. Gonzalez, Co-Chair
Professor Rob Knight, Co-Chair
Professor Pieter Dorrestein
Professor Partho Ghosh
Professor Larry Smarr

2020

The Dissertation of Robert Hardie Mills is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Co-Chair

University of California San Diego

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am truly blessed to be in this position today. Reflecting on my journey towards becoming a PhD, there is no path I could have taken to get to this point without the support of many people. This work is a tribute and testament to the mentors that have trained me, and the support of my friends and family.

First of all, I need to thank my thesis advisors, Professor's David Gonzalez and Rob Knight. You both took a chance on me when you decided to take me into your labs, believing in my potential to become a scientist. I can't thank you enough for taking that chance and supporting me throughout the past several years. I will not forget the lessons I have learned from the both of you and am extremely grateful for this opportunity. Throughout my graduate studies I feel that I have had a unique opportunity to truly follow my passions and curiosities, and I have the both of you to thank for supporting this freedom.

David, I am particularly grateful for your consistent belief, despite ups and downs, that I can excel in science. Thank you for your daily presence and guidance in how to navigate and be productive in academia alongside lessons in how to conduct and present research. Rob, thank you for your ideas, insight, and fostering of a truly amazing multi-disciplinary and collaborative work environment. I could not have asked for a more exciting place to pursue graduate research.

In the backdrop I have had amazing support from my family and friends. Maddie, over this time you have become my closest companion and provided constant support. I have somehow found the perfect partner and am so thankful for you. Mom and Dad, I am so lucky to have you both as parents, thank you for all that you have done to support me over this time. Brian, thank you for inspiring me, and providing your guidance and support.

Finally, I'd like to write a brief acknowledgement of other friends and mentors who have made this journey possible. Professor Hiu Chu and Professor Pieter Dorrestein, thank you for your guidance and for allowing me to access and work in your laboratories. I am very grateful for the collaborative environment that you both have provided that has helped make my work stand out. I also would like to acknowledge additional members of my thesis committee, Professor Larry Smarr and Professor Partho Ghosh for their input and guidance in my development over the past few years. Professor Aleksandra Sikora, thank you for inspiring and supporting me as an undergraduate researcher. Last, a thank you to all of my friends in the Biomedical Sciences Program. You guys have made these past few years immeasurably better.

It is an important note that the research contained within this document is the result of collaboration between many researchers, without whom this work would not have been possible.

Chapter 2 is a reprint of the material as it appears in Genome Research, 2020, Robert H. Mills, Jacob M. Wozniak, Alison Vrbanac, Anaamika Campeau, Benoit Chassaing, Andrew Gewirtz, Rob Knight, and David J. Gonzalez. The dissertation author played a primary role in all aspects of the work ranging from the study design, data acquisition, analysis and writing of the manuscript.

Chapter 3 is a reprint of the material as it appears in mSystems, 2019, Robert H. Mills, Yoshiki Vazquez-Baeza, Qiyun Zhu, Lingjing Jiang, James Gaffney, Greg Humphrey, Larry Smarr, Rob Knight and David J. Gonzalez. The dissertation author played a primary role in all aspects of the work ranging from the study design, data acquisition, analysis and writing of the manuscript.

Chapter 4 reflects material of a manuscript as it was reviewed by the journal Nature, in Jan 2020, Robert H. Mills, Parambir S. Dulai, Yoshiki Vázquez-Baeza, Qiyun Zhu, Greg Humphrey, Lindsay DeRight Goldasich, MacKenzie Bryant, Robert A. Quinn, Andrew T. Gewirtz, Benoit Chassaing, Hiutung Chu, William J. Sandborn, Pieter C. Dorrestein, Rob Knight, and David J. Gonzalez. The dissertation author played a primary role in aspects of the work ranging from the study design, data acquisition, analysis and writing of the manuscript.

Chapter 5 is a reprint of the material as it appears in Molecular and Cellular Proteomics, 2019, Hao Q. Tran, Robert H. Mills, Nicole V. Peters, Mary K. Holder, Geert J. de Vries, Rob Knight, Benoit Chassaing David J. Gonzalez, and Andrew T. Gewirtz. The dissertation author played a primary role in aspects of the work ranging from metaproteome data acquisition, data analysis and the writing of the manuscript.

Chapter 6 contains preliminary ideas and writing to form the basis of a grant application. The dissertation author played a primary role in the conceptualization and writing of this section. This work also contains editing contributions from Carlos Gonzalez and David J. Gonzalez.

VITA

2015 B.S. in Microbiology, Oregon State University, U. S. A.

2020 Ph.D. in Biomedical Sciences, University of California San Diego, U. S. A.

PUBLICATIONS

In print

1. **Robert H. Mills**, Jacob Wozniak, Alison Vrbanac, Anaamika Campeau, Benoit Chassaing, Andrew Gewirtz, Rob Knight and David J. Gonzalez. Organ level protein networks as a reference for the host effects of the microbiome (2020). *Genome Research.*
2. Alan M. O'Neill, Teruaki Nakatsuji, Asumi Hayachi, Michael R. Williams, **Robert H. Mills**, David J. Gonzalez and Richard L. Gallo. Mining human skin commensal bacteria for novel antimicrobials that selectively kill *Cutibacterium acnes* (2020). *Journal of Investigative Dermatology.*
3. Hao Q. Tran*, **Robert H. Mills***, Nicole V. Peters, Mary K. Holder, Geert J. de Vries, Rob Knight, Benoit Chassaing, David J. Gonzalez and Andrew T. Gewirtz. Associations of the fecal microbial proteome composition and proneness to diet-induced obesity (2019). *Molecular and Cellular Proteomics*. **Featured on Cover** ****Co-first authorship***
4. **Robert H. Mills**, Yoshiki Vazquez-Baeza, Qiyun Zhu, James Gaffney, Greg Humphrey, Larry Smarr, Rob Knight and David J. Gonzalez. Evaluating Metagenomic Prediction of the Metaproteome in a 4.5 Year Study of a Crohn's Patient (2019). *mSystems.*
5. Robert A. Quinn, Sandeep Adem, **Robert H. Mills**, William Comstock, Lindsay DeRight Goldasich, Gregory Humphrey, Alexander A. Aksenov, Ricardo da Silva, Gail Ackerman, David J. Gonzalez, Doug Conrad, Anthony J. O'Donoghue, Rob Knight and Pieter C. Dorrestein. Neutrophilic Proteolysis Alters the Cystic Fibrosis Lung Neutrophilic proteolysis in the cystic fibrosis lung correlates with a pathogenic microbiome (2019). *Microbiome.*
6. John D. Lapek*, **Robert H. Mills***, Jacob M. Wozniak, Anaamika Campeau, Ronnie H. Fang, Xiaoli Wei, Kristen van de Groep, Araceli Perez-Lopez, Nina M. van Sorge, Manuela Raffatellu, Rob Knight, Liangfang Zhang and David J. Gonzalez. Defining Host Responses during Systemic Bacterial Infection through Construction of a Murine Organ Proteome Atlas (2018). *Cell Systems*. ****Co-first authorship***
7. Aleksandra E. Sikora, **Robert H. Mills**, Jacob V. Weber, Adel Hamza, Bryan W. Passow, Andrew Romaine, Zachary A. Williamson, Robert W. Reed, Ryszard A. Zielke and Konstantin V. Korotkov. Peptide Inhibitors Targeting the *Neisseria gonorrhoeae* Pivotal Anaerobic Respiration Factor AniA (2017). *Antimicrobial Agents and Chemotherapy.*

8. Daniel Petras, Louis-Felix Nothias, Robert Quinn, Theodore Alexandrov, Nuno Bandeira, Amina Bouslimani, Gabriel Castro-Falcón, Liangyu Chen, Tam Dang, Dimitrios Floros, Vivian Hook, Neha Garg, Nicole Hoffner, Yike Jiang, Clifford Kapono, Irina Koester, Rob Knight, Christopher Leber, Tie-Jun Ling , Tal Luzzatto-Knaan, Laura-Isobel McCall, Aaron McGrath, Michael Meehan, Jonathan Merritt, **Robert H. Mills**, Jamie Morton, Sonia Podvin, Ivan Protsyuk, Trevor Purdy, Kendall Satterfield, Stephen Searles, Sahil Shah, Sarah Shires, Dana Steffen, Margot White, Jelena Todoric, Robert Tuttle, Aneta Wojnicz, Valerie Sapp, Fernando Vargas, Jin Yang, Chao Zhang and Pieter Dorrestein. Mass Spectrometry-Based Visualization of Molecules Associated with Human Habitats (2016). *Analytical Chemistry*.

In preparation

9. **Robert H. Mills***, Parambir Dulai*, Yoshiki Vazquez-Baeza, Qiyun Zhu, Greg Humphrey, Lindsay DeRight Goldasich, Robert A. Quinn, Andrew Gewirtz, Benoit Chassaing, Hiutung Chu, William Sandborn, Pieter C. Dorrestein, Rob Knight and David J. Gonzalez. Meta-omics Reveals Microbiome Driven Proteolysis as a Contributing Factor to Severity of Ulcerative Colitis Disease Activity. *In revision at Nature*. ***Co-first authorship***
10. Jacob M. Wozniak, **Robert H. Mills,** Josh Olson, JR Caldera, Marvic Carrillo-Terrazas, Gregory D. Poore, Chih-Ming Tsai, Fernando Vargas, Rob Knight, Pieter C. Dorrestein, George Liu, Victor Nizet, George Sakoulas, Warren Rose and David J. Gonzalez. Mortality Risk Profiling of Staphylococcus aureus Bacteremia by Deep Multi-omic Serum Analysis Reveals New Predictive and Pathogenic Molecular Signatures. *In revision at Cell.*
11. Alison Vrbanac, Kathryn A. Patras, Alan Jarmusch, **Robert H. Mills**, Samuel Shing, Robert A. Quinn, Fernando Vargas, David J. Gonzalez, Pieter C. Dorrestein, Rob Knight and Victor Nizet. Organism-wide Changes in the Metabolome and Microbiome Following a Single Dose of Antibiotic. *In Revision.*
12. Anaamika Campeau, **Robert H. Mills**, Marie Blanchette, Kaja Bajc, Mario Malfavon, Roeben Munji, Liwen Deng, Bryan Hancock, Katryn Patras, Joshua Olson, Victor Nizet, Richard Daneman, Kelly Doran, and David J. Gonzalez. Multi-Dimensional Proteome Profiling of Blood-Brain Barrier Perturbation by Group B Streptococcus. *In revision.*
13. **Robert H. Mills**, Marvic Carrillo-Terrazas, Yash Mittal, Brian A. Yee, Christella E. Widjaja, Fernando Vargas, Kevin Nguyen, Kelly Weldon, Julia Gauglitz, Gail Ackerrmann, Gregory Humphrey, Lindsay DeRight-Goldasich, Tara Schwartz, Austin D. Swafford, Corey A. Siegel, Gene W. Yeo, John Chang, Pradipta Ghosh, Pieter C. Dorrestein, Rob Knight, David J. Gonzalez, Parambir S. Dulai. Host-microbiome response to hyperbaric oxygen treatment in ulcerative colitis. *In preparation.*

ABSTRACT OF THE DISSERTATION

Integrating multiplexed metaproteomics to discover novel therapeutic avenues targeting the IBD microbiota

by

Robert Hardie Mills

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2020

Professor David Gonzalez, Co-Chair
Professor Rob Knight, Co-Chair

We are more than humans. Our mammalian cells are intimately associated with roughly an equivalent number of microbial cells, which contain more than 100 times more genes than mammalian cells. There is constant cross-communication between our cells and these microbial cells through the molecules (i.e. metabolites and proteins) that are produced by microbe and man. We are currently amidst a revolution in our

understanding of these interactions through the use of large-scale systems approaches utilizing technological developments in sequencing and mass spectrometry. While sequencing and metabolomic profiling are rapidly becoming standard practice in the field, the use of mass-spectrometry based proteomics is less commonly applied when considering host-microbiome interactions. Given the central role that proteins play in all biological organisms, the integration of this field into the wider context of microbiome research promises abundant insights.

This work describes the use of technological improvements in the field of quantitative multiplexed proteomics for understanding host-microbiome interactions. In the first chapter, I introduce the current state of human microbiome research, what diseases it is associated with and the technologies used to study it. Further, I describe the use of proteomics to study multispecies communities (metaproteomics). In the second chapter, I utilize proteomics to understand the response of different organ systems to a lack of microbes in mice. In the third chapter, I evaluate the differences that result from using genomic versus proteomic technology when studying the microbiome of a patient with a disease strongly connected to the microbiome, inflammatory bowel disease (IBD). In the fourth chapter, I study a cohort of IBD patients using a combination of six –omic datasets, and through the use of metaproteomics, identify a new therapeutic treatment avenue for IBD patients targeting bacterial proteases. I further evaluate this therapeutic approach experimentally in colonic epithelial cells and germ-free mice. The fifth chapter outlines the potential of quantitative multiplexed metaproteomics to better understand other microbiota-associated diseases. Specifically, I observe a potential of the technology to understand and predict obesity outcomes in mice. In the final chapter, I discuss the implications of this work.

Chapter 1

General Introduction

Summary

The human microbiome is becoming increasingly recognized to be intertwined with human health. Connections between the microbiome and diseases such as obesity and inflammatory bowel disease have primarily been described through genomic technologies. However, new methods of characterizing complex host-microbiome connections are emerging: shotgun metagenomics, metatranscriptomics, metaproteomics and metabolomics. This introduction will briefly survey the emerging relationship between the microbiome and human health, and the technologies that we use to understand these complex systems. I further emphasize the current state of proteomics for understanding the microbiome as I highlight this emerging technology throughout the dissertation.

1.1 Technology, Human Health, and the Microbiome

Transformations in our understanding of human health have often been accompanied by technological advancements. Our understanding of the microorganisms living on and around us stems from innovation in microscopy that Antonie van Leeuwenhoek made in 1676. After a series of technological advancements, scientist Robert Koch in the 1880's were able to cultivate isolated microorganisms and show that specific microorganisms, including *Bacillus anthracis*, caused disease. Through these observations, germ theory was created and a war on microorganisms began. This knowledge of the microbial world transformed public health with life saving developments like improved sanitation, antibiotics and vaccines. The impact of these technological developments may partially be contributing to a rise in the average life expectancy of a US citizen, which increased from 39 years in 1880 to a current average of 79 years.

The rise in the average lifespan over the past century is undoubtedly an amazing human accomplishment; however, with it has come new health complications we are just starting to understand. Rates of diseases such as Inflammatory Bowel Disease (IBD), diabetes and obesity have all dramatically increased in recent decades, and primarily within developed countries. With the entrance of new technologies peering into the communities of microorganisms on and inside us, we are now starting to realize that each of these conditions may be linked to a "dysbiosis" in our gut microbiome. As laid out by Martin Blaser in his book "Missing Microbes", it could be that the very innovations helping extend our lifespan are causing a collapse in our microbial communities and leading to these very diseases.

Efficient detection and profiling of microbial communities has been a relatively recent development that has taken large leaps in technological development. The first large-scale efforts to catalog the entire repertoire of microbes associated with different locations of the human body was the NIH funded Human Microbiome Project (HMP), results of which were published in 2012 (Human Microbiome Project 2012a; Human Microbiome Project 2012b)(Human Microbiome Project, 2012a, 2012b). These projects were aided by a jump to "next-generation" sequencing platforms like Illumina, which have allowed data collection on a massive scale. Perhaps just as important for the rise of the microbiome field were advances in bioinformatics. One example is the UniFrac distance metric, a widely adopted tool which provided an ecological metric accounting for the phylogeny of a given microbial community (Lozupone and Knight 2005).

Additionally, bioinformatics software packages like Qiime (Caporaso et al. 2010) and Mothur (Schloss et al. 2009), have made analysis of microbiome data reproducible and more accessible to a wider audience.

Once a baseline "healthy" microbial community was established by the Human Microbiome Project, associations between an array of health conditions and altered microbiota proliferated. The example at the forefront for demonstrating the power of the microbiome to shape human health is demonstrated by *Clostridium difficile* infection (CDI). Symptoms of CDI can range from mild diarrhea to severe life-threatening inflammation of the colon, and the infection is typically treated through antibiotics. However, the bacterium can become extremely resistant to antibiotics, leaving few treatment options. With this challenge, physicians turned to augmenting patients's microbiome through fecal microbiota transplant (FMT). Amazingly, studies report 81-100 % effectiveness for FMT transplant while vancomycin may be only 31 % effective (Czepiel et al. 2019). These results demonstrate the importance of a "healthy" microbiome in preventing overgrowth of particular pathogens, as is the case in CDI.

Another condition for which the microbiome is increasingly becoming recognized as a contributing factor to is IBD. Though the disease is speculated to have been present in ancient times, it has been found to be primarily associated with industrialized countries and has an increasing incidence rate (Mulder et al. 2014). Despite extensive research, the underlying etiology of the disease is still unknown today and appears quite variable (Zhang and Li 2014). Genetic screens have identified a variety of mutations in certain genes that increase the risk for developing of IBD. Many of these genes belong to systems involved in host-microbe interactions, some of the most prominent being the

NOD2 and ATG16L1 gene (Xavier and Podolsky 2007). Microbiome studies have now identified that shifts in the microbial community structure correlate with disease states (Frank et al. 2007), that particular taxa appear enriched or depleted in IBD patients (Walters et al. 2014), and that an individual patient's microbiome fluctuates over time (Lloyd-Price et al. 2019). Further connecting the gut microbiota and IBD are reports of the success of FMT in IBD. Though not as effective as in CDI, the ~30% response rate of FMT in IBD further cements the role of the microbiota in IBD (Browne and Kelly 2017). Still, many of the molecular mechanisms mediating host-microbe interactions in IBD remain unknown.

To address questions regarding the mechanisms and molecules mediating host-microbiome interactions, new technologies are needed. The field is currently in a transition, shifting from simply profiling the community through techniques like 16S rRNA gene amplicon sequencing (16S), to profiling the entire repertoire of genes present in a community through shotgun sequencing (metagenomics) and to next integrating the analysis of other types of molecules. There is now great interest in the analysis of multiple molecular profiles in tandem to better characterize the role of the microbiome (Jansson and Baker 2016). To this end we are not only seeing further developments in sequencing methods such as metatranscriptomics, but also mass-spectrometry based methods to characterize the metabolites and proteins present in a microbiome sample. The field's move in this direction is exemplified in the expansion to the NIH's HMP, the integrative HMP (iHMP), whose results were reported in a series of high-profile manuscripts tracking microbiome dynamics in IBD (Lloyd-Price et al. 2019), type-2 diabetes (Zhou et al. 2019) and preterm birth (Fettweis et al. 2019) in 2019.

1.2 The Use of Meta –omics to Study the Microbiome

Given the importance of systems-scale analyses in the microbiome field today, it is critical to understand what can be done with each meta-omic technique, how each data type is collected, and the limitations or future directions of the technologies. To date, methods are dominated by two main technologies, sequencing and mass-spectrometry, whose workflows are modified depending on the particular molecular analysis researchers wish to perform. Understanding some basic technical aspects of each of these technologies, and the limitations of each are important for understanding how best to apply and combine results from each. Sequencing is used for 16S, metagenomic and metatranscriptomic data collection, allowing the profiling of the community structure through a marker gene analysis (16S), the entire complement of genes present in a multi-species community (metagenome) or the complement of transcripts present in a complex sample (metatranscriptome). Mass spectrometry techniques analyze the molecules of a complex community of organisms further down the central dogma through profiling the proteins present (metaproteomics) or the metabolites present (metabolomics) in a given sample.

The principal concept behind modern sequencing technologies are the use of fluorescent nucleotides that are sequentially incorporated onto a DNA template strand and emitting light that can be detected. Next-generation sequencing technologies were introduced in the 2000's expanding upon this same concept in a highly parallel fashion where instead of analyzing one sequence at a time, millions were analyzed. Modern technologies include Illumina, Ion torrent and Pacific Biosciences (PacBio). Each of these platforms have unique strategies for amplifying DNA related to an early stage of the preparation of sequencing libraries which results in either linear (Illumina and Ion torrent) or circular libraries (PacBio). The technologies are further differentiated by steps further into the process related to template generation where libraries will be either amplified by bridge amplification (Illumina), emulsion PCR (Ion torrent), or directly detected without amplification (PacBio) (Buermans and den Dunnen 2014). Each technique has certain advantages and applications, notably the high output of data that can be generated with the Illumina platform and the length of sequencing reads using the PacBio platform. With rapid progression and decreased cost of these technologies, we can efficiently capture more genetic information on living systems than ever before.

Mass spectrometry (MS) technology has also been rapidly advancing from its initial designs in the early 20th century (Griffiths 2008). MS is based on the ability to infer properties of ionized molecules through comparing the mass to charge ratio of a given molecule (m/z). This technology becomes ever more useful when applying tandem mass spectrometry, where the instrument can perform multiple rounds of MS, breaking a parent molecule into smaller parts, and measuring the m/z of each of these fragment ions (a process often referred to as MS2). This information is the basis of how researchers can identify what molecule is present, as one can compare these MS2 spectra to the spectra of how a known molecule fragments.

After around a century of developments in instrumentation some of the more widely used and important MS technologies today include the linear ion trap and orbitrap mass analyzers. Linear ion traps allow for capture of ions along a linear track through the shifting of electrical frequencies in a set of quadropole rods. Instruments utilizing linear ion trap instruments such as the Thermo LTQ can have high sensitivity alongside rapid acquisition of measurements (Eliuk and Makarov 2015). The instruments utilizing orbitrap technology provide high-resolution mass measurements by subjecting ions to orbital motion and calculating the m/z values through a Fourier transformation (Eliuk and Makarov 2015). Notably, a new series of instruments starting with the 2015 Thermo Orbitrap Fusion, combines a quadropole mass filter, and Orbitrap and linear ion trap mass analyzers to provide a "best of both worlds" scenario where you can concurrently use each mass analyzer to simultaneously isolate ions and detect ions (Eliuk and Makarov 2015). These technologies are at the forefront and provide researchers with unprecedented levels of information on peptides, lipids and small molecules present in a given sample.

From every process between sample selection to final analysis, there are decisions to be made which might affect conclusions from a sequencing or mass spectrometry experiment. Because many of these concerns have been well summarized and documented for sequencing analyses before (Quince et al. 2017; Knight et al. 2018)(Knight et al., 2018; Quince, Walker, Simpson, Loman, & Segata, 2017) we will instead focus on processes that may be important to consider when performing analysis of multiple –omic types. Sequencing studies have unique biases in comparison to mass spectrometry. Both sequencing and mass spectrometry experiments start with an extraction. The choice of extraction protocol is particularly important as it strongly influences 16S results (Brooks et al. 2015). As opposed to mass spectrometry, where the material one analyzes was present in the beginning of the process, sequencing techniques rely on PCR amplification. Various aspects of PCR amplification can result in biases in the data (Gohl et al. 2016) including the choice of primers (of particular importance in

16S studies (Walker et al. 2015)), the creation of chimera sequences (Haas et al. 2011), and host contamination (Marotz et al. 2018). For mass spectrometry experiments, it is important to consider that the organic solvent used for extraction can greatly influence the metabolite identifications (Want et al. 2006; Dettmer et al. 2011)(Dettmer et al., 2011; Want et al., 2006), that run-to-run variability can have a very large effect (Bittremieux et al. 2018), that instruments and acquisition settings greatly vary, that some molecules will not ionize well (Leito et al. 2008), and that the methods used can influence how accurate quantitative comparisons are (Pappireddi et al. 2019).

The fact that each technique has its own set of advantages and disadvantages highlights the potential impact of combining multiple data types. No method is perfect for all scenarios and pulling from the strengths of each approach can lead to stronger hypotheses. The following few paragraphs serve as a brief introduction to the applications of each data type, starting with DNA based methods and ending on the applications of metabolomics in microbiome research.

Collection of 16S data is currently an enticing entry point for microbiome studies for several reasons. The 16S method uses PCR amplification of the evolutionarily well-conserved 16S rRNA gene through selection of primers flanking selected regions of the gene representative of a large range of bacteria and archaea. Aside from typically costing less than the other data types, it also has had a large amount of bioinformatic development and better-established standard practices. Software development projects like Qiime (Hall and Beiko 2018) and Mothur (Schloss et al. 2009) now make it relatively easy to go from raw data to insight. As a testament to this notion, it was even shown that it was possible to complete a 16S and metabolomics analysis in less than 48 hours (Quinn et al. 2016). Further, thanks to efforts such as the Greengenes, which provides reference phylogenies for full-length 16S rRNA genes (McDonald et al. 2012), there are well-established methods for microbial community comparison that account for phylogeny. The importance of bioinformatics methods including a phylogenetic perspective is highlighted in the rise of the UniFrac distance metric. UniFrac is a metric for calculating beta-diversity (the between sample differences in community structures), that incorporated the evolutionary history of the organisms present in the community. The original UniFrac manuscript now has over 4700 citations (Lozupone and Knight 2005). The importance of a phylogenetic perspective also holds true for measures of alpha-diversity, a measure of within sample community structure. Unlike other data types, alpha-diversity metrics using a phylogenetic context such as Faith's PD are now routinely implemented in 16S analysis (Faith 1994). Further, the current amount of 16S data collected and available for reanalysis through platforms like Qiita (Gonzalez et al. 2018) provides another advantage for 16S.

Also residing at the DNA level for microbiome analyses is shotgun metagenomics. Here, instead of only amplifying a selected region of DNA, researchers will amplify all available DNA in short segments (often around 150 nucleotide). These segments can be combined into longer segments called "contigs" or directly mapped to a reference database for analysis (Quince et al. 2017). This option provides distinct advantages from 16S, namely the potential of getting higher-resolution taxonomic annotations and a profile of the functional state of a microbial community. With this platform it is also possible to simultaneously analyze viral and eukaryotic genes. Having a functional perspective can provide useful insights such as identifying metabolic differences (Franzosa et al. 2018), antibiotic gene cassettes (Berglund et al. 2019), or biosynthetic pathways (Aleti et al. 2019). Of note, there have been tools created to predict functional profiles from16S sequences, such as PICRUSt (Langille et al. 2013). Though the additional information gained from transitioning from 16S to metagenomic sequencing is appealing, the amount of a data generated in a metagenomic sequencing run can become cumbersome, require intensive computing power, and fewer standard practices make analysis difficult.

Given that the identity of transcripts and proteins are encoded within the genome, the question arises of whether there is further utility in profiling these molecules. Can we not predict levels of transcripts and proteins from identifying the genes present? Research has shown that there are fairly strong correlations between metagenomic and metatranscriptomic data (Spearman's *r* = 0.76), with notable differences in a few categories such as bacterial ribosomal and chaperone proteins (Franzosa et al. 2014). The relationship between genes and protein abundances, as well as transcript to protein abundances have shown much lower correlations with average Spearman correlations at or below *r* = 0.3 (Lloyd-Price et al. 2019; Mills et al. 2019)(Lloyd-Price et al., 2019; Mills et al., 2019). However, these relationships are still being investigated and there is a need to evaluate these relationships with new statistical methods given that the use of traditional statistical methods have inflated false-discovery rates in highly sparse data sets often produced in microbiome research (Weiss et al. 2016).

Metatranscriptomics has been shown to be of use in several applications. Transcripts have the advantage of only being present for actively transcribing organisms, while a large amount of reads from metagenomics could be derived from dead cells.

However, there are methods being developed to separate out dead cells that can overcome this limitation of metagenomics (Marotz et al. 2018). Additionally, one important step necessary for metatranscriptomics is the removal of ribosomal RNA which can constitute up to 90% of all data when not removed before sequencing (Shakya et al. 2019). Another important consideration for metatranscriptomics is the stability of mRNA, which makes sample collection and storage an added issue (Reck et al. 2015). Metatranscriptomics while not as routinely used, could be used to detect immediate transcriptional reprograming upon a perturbation. One example of its successful application when compared to in-parallel metagenomic analysis comes from Cullender and colleagues. Cullender et al., were able to demonstrate that wild type and TLR5$^{-/-}$ (a gene required to produce a response to flagella) mice had distinct differences in metatranscriptomic pathways related to flagella while metagenomics showed no difference (Cullender et al. 2013). Other examples include identifying *Faecalibacterium praustniztii* as the primary contributor to the transcriptional pool of some IBD patients, despite a complex community being identified through metagenomics (Franzosa et al. 2018). In total, the field has had successful applications of metatranscriptomics ranging from host-microbe interactions to characterizing microbial responses to environmental conditions (Shakya et al. 2019).

One transcriptomic technique that aims to close the correlation gap between transcripts and proteins is Ribo-seq. This technology reveals short segments of RNA that are being translated by ribosomes. Results in the field have shown some examples of transcript and protein abundances reaching a Spearman correlation of $r = 0.8$, though there was a limited overlap of the proteins and transcripts identified (Liu et al. 2017). It must be noted that to date, there have been no known applications of this technology in the microbiome field.

The preferred analytical method for characterizing proteins and metabolites is MS. Because proteins have fundamental roles in carrying out most functional processes in cells, the proteome component can be thought of as the “functional state” of a community, while the metabolome is thought as close to a phenotypic readout given that many molecules can have affects in their surrounding environment. Despite both typically being analyzed through MS analysis, the metabolomic and proteomic workflows are quite distinct. Both workflows typically involve cell lysis, but the workflows start to differ after this point. Metabolites are typically extracted using various different organic solvents depending on the type of metabolite. Part of this process precipitates out the proteins from the sample. Proteomic processing is typically more elaborate and involves multiple purification steps, enzymatic digestion, and fractionation.

The computational methods of identifying a protein and metabolite are also quite distinct. Peptides fragment in predictable ways that can be matched to *in silico* predictions. This property affords proteomics the ability to discern high and low quality spectra by using false-discovery methods. A popular approach to preventing false discoveries is the reverse database search, which computes theoretical spectra from the complement sequences of your expected protein database and only allow a controlled percentage of spectra to match these unexpected spectra (Elias and Gygi 2010). In metabolomics, fragments can be matched to a reference library of previously fragmented molecules (Wang et al. 2016). This makes the metabolite molecular identification process quite dependent upon previously acquired libraries of the fragmentation of known molecules. Historically, identification rates of MS2 in metabolomics datasets have been quite low (>5%), however novel methods for inferring related spectra have shown potential to provide estimated molecular classes to around 70% of MS2 (Djoumbou Feunang et al. 2016; van der Hooft et al. 2016)(Djoumbou Feunang et al., 2016; van der Hooft, Wandy, Barrett, Burgess, & Rogers, 2016)caveat - _ENREF_4_65e fact that, unlike peptides, whose origins can potentially be identified, it is difficult to determine the origins of a metabolite. This poses a major challenge in complex microbial communities as understanding which microbes are producing which metabolites is important for mechanistic insight into host-microbe interactions. A more detailed review of the current challenges in metabolomics can be found in Schrimpe-Rutledge *et al*., 2016 (Schrimpe-Rutledge et al. 2016). The history, technical challenges and applications of proteomics in the microbiome field will be elaborated on further in the next section.

There is currently a great amount of interest in the microbiome field regarding the identification of molecules created by the microbiome that can influence the host. Notably, there have been robust associations between microbial derived trimethylamine-N-oxide (TMAO) and cardiovascular disease (Tang et al. 2013). Other metabolites of interest include microbially modified bile acids, which interact with the FXR receptor and might have roles in colorectal cancer and IBD (Jia et al. 2018; Quinn et al. 2020)(Jia, Xie, & Jia, 2018; Quinn et al., 2020). The microbiota also digest dietary fibers producing short-chain fatty acids, which are both a major nutritional source to colonic epithelial cells and can have a variety of physiological effects through mechanisms related to histone deacetylases and G-protein coupled receptors (Koh et al. 2016).

The merging of all of these data types is also of great interest, but there are currently very few examples of this being performed. Some notable exceptions are the large-scale projects of the iHMP, as previously mentioned. The current methods of inter-correlating different -omic data types have been limited to creating large cross correlation networks (Lloyd-Price et al. 2019). Researchers have also performed analyses such as a Procrustes analysis, which can be used to compare beta-diversity distributions (Gower 1975). New methods are currently being developed that utilize machine-learning techniques for novel insight and inter-omic relationships (Grapov et al. 2018). For example, a tool was recently published which identifies microbe metabolite interactions through neural networks (Morton et al. 2019).

1.3 Proteomics and the Microbiome

The use of proteomics to study the microbiome is still a relatively under-utilized approach in comparison to genetic-based microbiome analyses. The microbiome field has seen an explosion of research in the past decade with the average number of publications per year now over 10,000 (Fig. 1.1A). Given the comparatively limited number of research in this area today, progress is needed to adequately benchmark and create standard practices in the field (Zhang and Figeys 2019). Despite this, the number of publications utilizing metaproteomics to investigate the microbiome is also entering exponential growth, and the number of publications per year has recently breached 100 (Fig. 1.1B). Though there are likely several reasons for the increased interest in the field, the gaining traction might be related to recent progress improving upon the limitations of the technology identified in early studies.

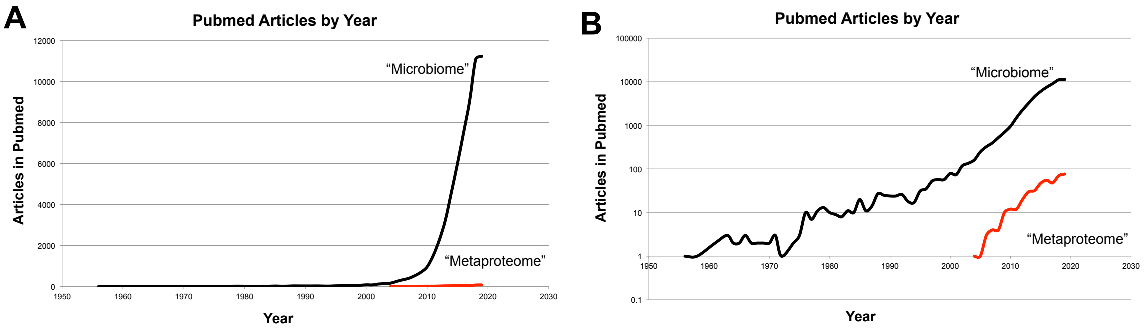


Figure 1.1 Pubmed articles using keywords "Microbiome" or "Metaproteome". A, Microbiome publications has vastly exceeded the number of metaproteome publications. B, Using a logarithmic scale reveals that the number of metaproteome publications is exponentially increasing.

The early history of the metaproteome field highlights challenges related to sample size limitations and depth of analysis. One of the earliest metaproteomic studies performed was by Klaassens et al. in 2007. Here, the researchers analyzed 6 infant fecal samples by 2D gel MALDI-TOF and identified 200 protein spots (Klaassens et al. 2007). Researchers soon moved from a gel-based approach toward a broader reaching, "shotgun" metaproteomic approach, starting with an n=2 study of human fecal samples in 2008 where 1534 proteins were identified (Verberkmoes et al. 2009). Other notable early studies in the field came in 2012 with the first temporal study (Kolmeder et al. 2012), and an early study integrating metagenomic and metaproteomic to study Crohn's disease (Erickson et al. 2012). Another early exploratory study collected 16S, metagenome, metatranscriptome, metabolome and metaproteome data on one patient at 6 time points before and after antibiotic exposure (Perez-Cobas et al. 2013). Each of these early studies was performed with limitations in sample size and identified up to 3,000 proteins (Fig. 1.2).

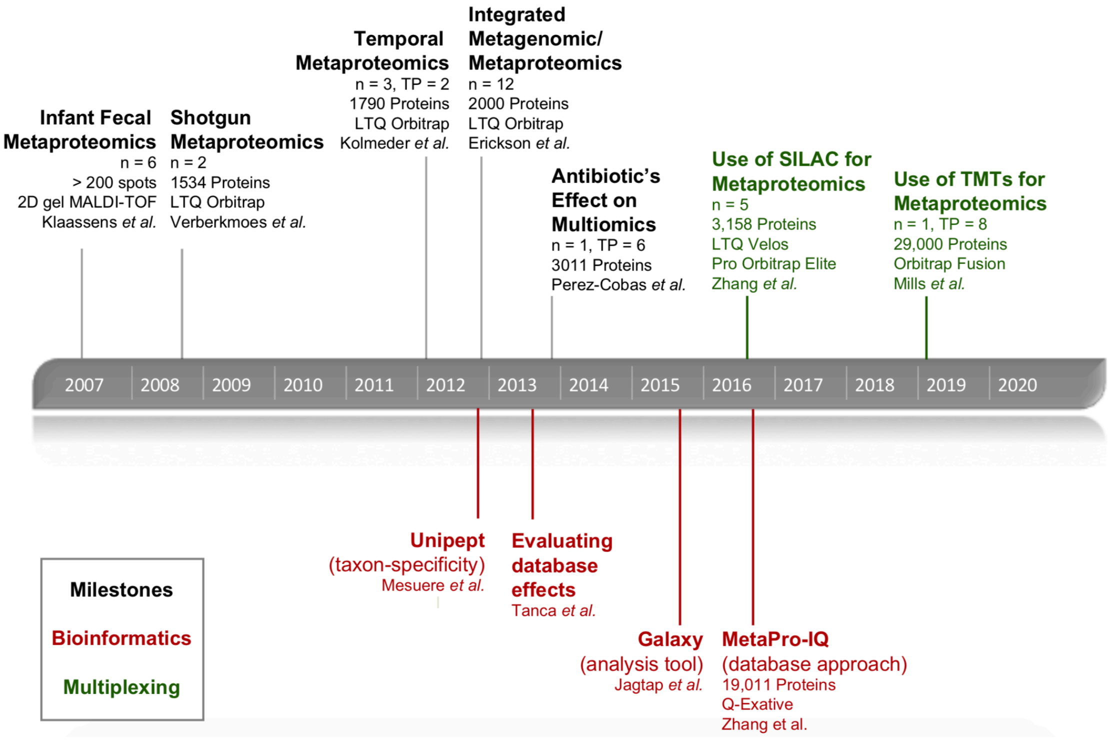


Figure 1.2 Milestone studies in the human metaproteome. A timeline of developments in the field of human metaproteomics is shown. Each milestone is colored by association to having been an early application (black), an important new bioinformatics development (red), or introduction of a new multiplexing technology to the field (green). The sample sizes, instrument used for data aquisition and number of identified proteins are highlighted for most studies.

In the backdrop of these early studies was a need to improve methodology in the field. In the preparation before analyzing samples on a mass spectrometer, methods needed to be established regarding best practices for cell lysis, and how to enrich or deplete human cells from samples. As with genomic based studies, the lysis protocol used can bias the taxonomic composition of the proteins you identify given the relative difficulty it takes to lyse either gram positive or gram negative bacterial cells (Wang et al. 2020). To address whether or not to analyze the human component of the metaproteome, work has been done to evaluate methods that use differential centrifugation (Tanca et al. 2015) or filtration approaches (Xiong et al. 2015). Once a sample is ready to analyze, there are additional considerations given the complexity of the microbiome. Some early studies attempted to address this complexity by using extensive fractionation methods including a 2-dimensional liquid chromatography approach requiring 22 hours of instrument time per sample (Erickson et al. 2012).

There are additional areas of consideration for metaproteomic versus proteomic studies related to the computational processing of spectral data. In particular, new spectral search identification schemes and choices of protein database methodology has been an area of much development. Choosing a protein reference database can be roughly split into two categories. The two main categories are to take either an untargeted approach using a public standardized reference database, or to build a custom database. The public standardized reference database is advantageous because of being easily accessible for researchers without sequencing data available, and the standardized methodology that makes cross-comparisons between studies possible (Zhang and Figeys 2019). There are also several bioinformatics tools online to help facilitate searching data through one of these approaches (Cheng et al. 2017; Beyter et al. 2018; Blank et al. 2018)(Beyter, Lin, Yu, Pieper, & Bafna, 2018; Blank et al., 2018; K. Cheng et al., 2017). For building a custom refuilding a custom reference database there are several options. It is possible to use 16S to identify microbes present in the samples, and then compile the genomes of identified species into a reference database. Another strategy is to perform shotgun metagenomic sequencing followed by compiling the data into a customized database for metaproteomic analysis (Tanca et al. 2016).

The choice of database methodology can have impacts on downstream results. It was recognized early on that the size of some of the databases used in the field was much larger than a standard proteomics search, which results in the decoy database (a routinely used strategy for controlling false discoveries (Elias and Gygi 2007)) filling at a faster rate. To circumvent this, researchers started to employ a multi-step iterative search approach that could increase spectral matches (Jagtap et al. 2013). Further approaches are being explored building upon iterative searches from a parallel search approach of individual samples (Zhang et al. 2016b), or a partitioned database approach (Beyter et al. 2018). Tanca and colleagues have been notably important in the early benchmarking efforts to understand the effects of database methodology (Tanca et al. 2013; Tanca et al. 2016)(Tanca et al., 2013; Tanca et al., 2016).

Further complicating metaproteomic studies is added complexity for properly assigning peptides to an array of highly similar proteins. Methods like unipept have emerged to analyze peptides for their lowest common ancestors, therefore increasing the confidence that a given peptide was not found in other taxa of the same level (Mesuere et al. 2015). However, a common approach is to group peptides into proteins and analyze the data based on a summed protein abundance. Given the greatly increased possible proteins in a metaproteomic study, there is a higher likelihood of having multiple proteins containing overlapping peptides. It is standard to group these proteins together, but it may be of particular importance to report the unique peptides assigned to each protein group for metaproteome studies (Zhang and Figeys 2019). For these approaches, having a metagenomic guided database may be preferred as the metagenome can provide a higher resolution for annotating proteins and provide *a priori* evidence that a given protein could be present in a given sample. Creating a project-specific database can help constrain and define the potential protein identifications, thus making peptide assignment more analogous to a traditional proteomics experiment.

One technological development in the field that may address both the limited sample sizes and sparsity of metaproteome data is multiplexing. Multiplexing is an MS approach where multiple samples can be run simultaneously. These approaches consist of several strategies of labeling peptides, either by chemical tags, or isotopes that are metabolically incorporated which can be later resolved by the MS. The first metaproteomic study to incorporate a multiplexed approach came in 2016 using a Stable Isotope Labeling with Amino acids in Cell culture (SILAC) method (Zhang et al. 2016a). However, there are serious limitations to the SILAC method, notably that these studies typically multiplex 2 samples at once, and that it requires the label to be incorporated in growing organisms, meaning that biases related to culturing samples might bias results. Given these drawbacks an alternative approach might be the use of Tandem Mass Tags (TMTs), which can be incorporated to peptides during MS sample preparation (Thompson et al. 2003). TMTs can now provide the ability to multiplex up to 16 samples in one mass spectrometry run (Li et al. 2020), and have recently been incorporated into metaproteomic studies (Liu et al. 2019; Mills et al. 2019).

Another major advantage to a multiplexing approach for the metaproteome field is the ability to decrease sparsity. While label-free approaches may have a larger dynamic range than label based approaches, the complexity of the metaproteome results in a lack of the statistical power needed to identify significantly changing features given the number missing values present, as has been shown for phosphoproteomics (Hogrebe et al. 2018). Applying further approaches in combination to TMTs like multinotch MS3 for quantifying the isotopic labels can also improve the sensitivity of the analysis, while retaining more quantitative accuracy from TMT approaches using MS2 (McAlister et al.

2014). As a testament to this approach, we analyzed data from two recently published metaproteomic datasets that used either a label-free or an MS3-based TMT approach (Lloyd-Price et al. 2019; Mills et al. 2019). By taking a subset of 24 samples from each dataset and searching the spectra through an identical database search approach, we found that the multiplexing approach improved overall protein identifications (Figure 1.3A), greatly improved the number of proteins identified per sample (Figure 1.3B) and dramatically reduced sparsity from 85% missing values to 16.5% missing values (Figure 1.3C). While other factors including the different study designs, fractionation and acquisition settings may also influence these results, the labeling approach may be a key component to managing missing values in metaproteomics.

Given all of these improvements coming alongside advances in mass spectrometry technology in general, we are seeing dramatic increases in what metaproteomics can do. The field has come a long way in protein depth from some of the earliest studies identifying 200 "protein spots," to the studies in the early 2010's which identified around 2,000 proteins to recent studies which can quantify over 50,000 proteins in a single study (Zhang et al. 2018). With the incorporation of multiplexing technology to allow for larger-scale studies, it is likely that these numbers will only grow in the near future.

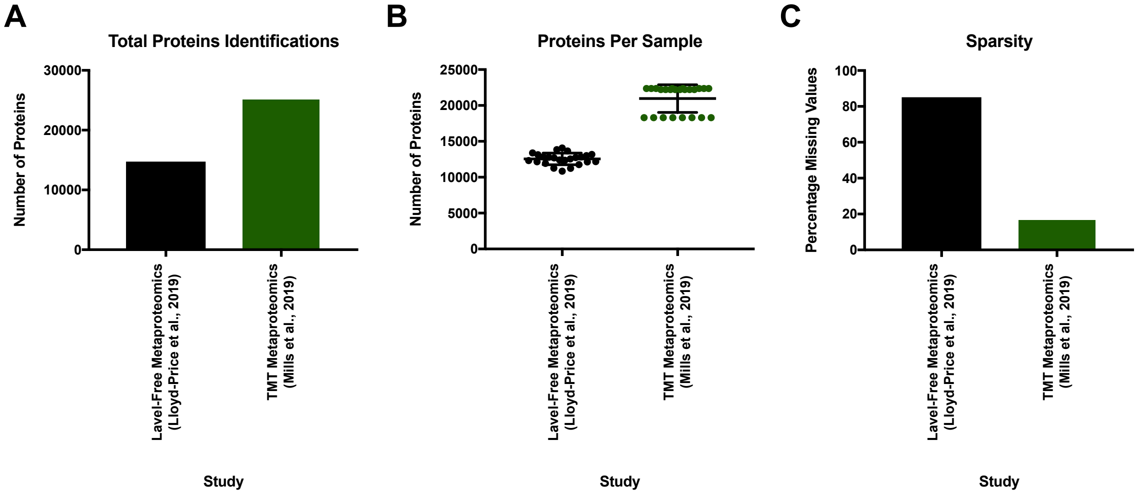


Figure 1.3 TMT-Metaproteomics decreases the sparsity of metaproteomics data. A subset of 24 IBD patient fecal samples from a label-free study and 24 fecal samples from an IBD patient case study which used TMT-based multiplexed metaproteomics were analyzed by identical search parameters. A, The number of proteins identified in two metaproteome studies using either label-free methods or TMT-based multiplexed proteomics. B, The number of proteins quantified per sample between these two studies. Mean and standard deviations are shown. C, The percentage of missing values within each study is shown.

Future directions for the field will be to incorporate information on post-translational modifications. This remains one of the primary advantages to the proteomics field given the known importance of modifications like phosphorylation on signaling activity states, glycosylation on the physiochemical properties of cells, and other events like acetylation and ubiquitination (Macek et al. 2019). The metaproteomics field has yet to incorporate this level of information. However, there are new reports of work toward profiling acetylation states (Zhang et al. 2019), and modified bioinformatic workflows for PTM identification (Cheng et al. 2020).

From observing the early history of the field, it is easy to see why metaproteomics was not a preferred method for the early microbiome studies. The lack of depth, amount of expense, and need for benchmarked methodologies all made it fall behind genomic approaches. However, recent studies in the field give hope that the technology might be at a coming-of-age stage and might soon be adopted by the greater microbiome community in the near future. Already we are seeing strong applications of the technology for identifying bacterial digestive enzymes (Patnode et al. 2019), and investigating the host-microbiome interplay in IBD (Zhang et al. 2018). We suspect many more important discoveries to be made by investigating the microbiome through proteomics, especially when leveraged along side other meta –omic technologies.

1.4 References

Aleti G, Baker JL, Tang X, Alvarez R, Dinis M, Tran NC, Melnik AV, Zhong C, Ernst M, Dorrestein PC et al. 2019. Identification of the Bacterial Biosynthetic Gene Clusters of the Oral Microbiome Illuminates the Unexplored Social Language of Bacteria during Health and Disease. *mBio* **10**.

Berglund F, Osterlund T, Boulund F, Marathe NP, Larsson DGJ, Kristiansson E. 2019. Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome* **7**: 52.

Beyter D, Lin MS, Yu Y, Pieper R, Bafna V. 2018. ProteoStorm: An Ultrafast Metaproteomics Database Search Framework. *Cell Syst* **7**: 463-467 e466.

Bittremieux W, Tabb DL, Impens F, Staes A, Timmerman E, Martens L, Laukens K. 2018. Quality control in mass spectrometry-based proteomics. *Mass Spectrom Rev* **37**: 697-711.

Blank C, Easterly C, Gruening B, Johnson J, Kolmeder CA, Kumar P, May D, Mehta S, Mesuere B, Brown Z et al. 2018. Disseminating Metaproteomic Informatics Capabilities and Knowledge Using the Galaxy-P Framework. *Proteomes* **6**.

Brooks JP, Edwards DJ, Harwich MD, Jr., Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P et al. 2015. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol* **15**: 66.

Browne AS, Kelly CR. 2017. Fecal Transplant in Inflammatory Bowel Disease. *Gastroenterol Clin North Am* **46**: 825-837.

Buermans HP, den Dunnen JT. 2014. Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta* **1842**: 1932-1941.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335-336.

Cheng K, Ning Z, Zhang X, Li L, Liao B, Mayne J, Figeys D. 2020. MetaLab 2.0 enables accurate post-translational modifications profiling in metaproteomics. *bioRxiv* doi:10.1101/753996: 753996.

Cheng K, Ning Z, Zhang X, Li L, Liao B, Mayne J, Stintzi A, Figeys D. 2017. MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome* **5**: 157.

Cullender TC, Chassaing B, Janzon A, Kumar K, Muller CE, Werner JJ, Angenent LT, Bell ME, Hay AG, Peterson DA et al. 2013. Innate and adaptive immunity interact to quench microbiome flagellar motility in the gut. *Cell Host Microbe* **14**: 571-581.

Czepiel J, Drozdz M, Pituch H, Kuijper EJ, Perucki W, Mielimonka A, Goldman S, Wultanska D, Garlicki A, Biesiada G. 2019. Clostridium difficile infection: review. *Eur J Clin Microbiol Infect Dis* **38**: 1211-1221.

Dettmer K, Nurnberger N, Kaspar H, Gruber MA, Almstetter MF, Oefner PJ. 2011. Metabolite extraction from adherently growing mammalian cells for metabolomics studies: optimization of harvesting and extraction protocols. *Anal Bioanal Chem* **399**: 1127-1139.

Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, Fahy E, Steinbeck C, Subramanian S, Bolton E et al. 2016. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminform* **8**: 61.

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**: 207-214.

Elias JE, Gygi SP. 2010. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol* **604**: 55-71.

Eliuk S, Makarov A. 2015. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu Rev Anal Chem (Palo Alto Calif)* **8**: 61-80.

Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, Shah M, Halfvarson J, Tysk C, Henrissat B et al. 2012. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* **7**: e49138.

Faith DP. 1994. Phylogenetic pattern and the quantification of organismal biodiversity. *Philos Trans R Soc Lond B Biol Sci* **345**: 45-58.

Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, Huang B, Arodz TJ, Edupuganti L, Glascock AL et al. 2019. The vaginal microbiome and preterm birth. *Nat Med* **25**: 1012-1021.

Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. 2007. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A* **104**: 13780-13785.

Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N et al. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* **15**: 962-968.

Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D et al. 2014. Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci U S A* **111**: E2329-2338.

Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, Gould TJ, Clayton JB, Johnson TJ, Hunter R et al. 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat Biotechnol* **34**: 942-949.

Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vazquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB et al. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. *Nat Methods* **15**: 796-798.

Gower JC. 1975. Generalized Procrustes Analysis. *Psychometrika* **40**: 33-51.

Grapov D, Fahrmann J, Wanichthanarak K, Khoomrung S. 2018. Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine. *OMICS* **22**: 630-636.

Griffiths J. 2008. A brief history of mass spectrometry. *Anal Chem* **80**: 5678-5683.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E et al. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21**: 494-504.

Hall M, Beiko RG. 2018. 16S rRNA Gene Analysis with QIIME2. *Methods Mol Biol* **1849**: 113-129.

Hogrebe A, von Stechow L, Bekker-Jensen DB, Weinert BT, Kelstrup CD, Olsen JV. 2018. Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nat Commun* **9**: 1045.

Human Microbiome Project C. 2012a. A framework for human microbiome research. *Nature* **486**: 215-221.

Human Microbiome Project C. 2012b. Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207-214.

Jagtap P, Goslinga J, Kooren JA, McGowan T, Wroblewski MS, Seymour SL, Griffin TJ. 2013. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **13**: 1352-1357.

Jansson JK, Baker ES. 2016. A multi-omic future for microbiome studies. *Nat Microbiol* **1**: 16049.

Jia W, Xie G, Jia W. 2018. Bile acid-microbiota crosstalk in gastrointestinal inflammation and carcinogenesis. *Nat Rev Gastroenterol Hepatol* **15**: 111-128.

Klaassens ES, de Vos WM, Vaughan EE. 2007. Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. *Appl Environ Microbiol* **73**: 1388-1392.

Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall LI, McDonald D et al. 2018. Best practices for analysing microbiomes. *Nat Rev Microbiol* **16**: 410-422.

Koh A, De Vadder F, Kovatcheva-Datchary P, Backhed F. 2016. From Dietary Fiber to Host Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites. *Cell* **165**: 1332-1345.

Kolmeder CA, de Been M, Nikkila J, Ritamo I, Matto J, Valmu L, Salojarvi J, Palva A, Salonen A, de Vos WM. 2012. Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. *PLoS One* **7**: e29913.

Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R et al. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**: 814-821.

Leito I, Herodes K, Huopolainen M, Virro K, Kunnapas A, Kruve A, Tanner R. 2008. Towards the electrospray ionization mass spectrometry ionization efficiency scale of organic compounds. *Rapid Commun Mass Spectrom* **22**: 379-384.

Li J, Van Vranken JG, Pontano Vaites L, Schweppe DK, Huttlin EL, Etienne C, Nandhikonda P, Viner R, Robitaille AM, Thompson AH et al. 2020. TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat Methods* **17**: 399-404.

Liu CW, Chi L, Tu P, Xue J, Ru H, Lu K. 2019. Isobaric Labeling Quantitative Metaproteomics for the Study of Gut Microbiome Response to Arsenic. *J Proteome Res* **18**: 970-981.

Liu TY, Huang HH, Wheeler D, Xu Y, Wells JA, Song YS, Wiita AP. 2017. Time-Resolved Proteomics Extends Ribosome Profiling-Based Measurements of Protein Synthesis Dynamics. *Cell Syst* **4**: 636-644 e639.

Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ et al. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**: 655-662.

Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228-8235.

Macek B, Forchhammer K, Hardouin J, Weber-Ban E, Grangeasse C, Mijakovic I. 2019. Protein post-translational modifications in bacteria. *Nat Rev Microbiol* **17**: 651-664.

Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. 2018. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome* **6**: 42.

McAlister GC, Nusinow DP, Jedrychowski MP, Wuhr M, Huttlin EL, Erickson BK, Rad R, Haas W, Gygi SP. 2014. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal Chem* **86**: 7150-7158.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610-618.

Mesuere B, Debyser G, Aerts M, Devreese B, Vandamme P, Dawyndt P. 2015. The Unipept metaproteomics analysis pipeline. *Proteomics* **15**: 1437-1442.

Mills RH, Vázquez-Baeza Y, Zhu Q, Jiang L, Gaffney J, Humphrey G, Smarr L, Knight R, Gonzalez DJ. 2019. Evaluating Metagenomic Prediction of the Metaproteome in a 4.5-Year Study of a Patient with Crohn's Disease. *mSystems* **4**: e00337-00318.

Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, Swenson TL, Van Goethem MW, Northen TR, Vazquez-Baeza Y et al. 2019. Learning representations of microbe-metabolite interactions. *Nat Methods* **16**: 1306-1314.

Mulder DJ, Noble AJ, Justinich CJ, Duffin JM. 2014. A tale of two diseases: the history of inflammatory bowel disease. *J Crohns Colitis* **8**: 341-348.

Pappireddi N, Martin L, Wuhr M. 2019. A Review on Quantitative Multiplexed Proteomics. *Chembiochem* **20**: 1210-1224.

Patnode ML, Beller ZW, Han ND, Cheng J, Peters SL, Terrapon N, Henrissat B, Le Gall S, Saulnier L, Hayashi DK et al. 2019. Interspecies Competition Impacts Targeted Manipulation of Human Gut Bacteria by Fiber-Derived Glycans. *Cell* **179**: 59-73 e13.

Perez-Cobas AE, Gosalbes MJ, Friedrichs A, Knecht H, Artacho A, Eismann K, Otto W, Rojo D, Bargiela R, von Bergen M et al. 2013. Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut* **62**: 1591-1601.

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* **35**: 833-844.

Quinn RA, Melnik AV, Vrbanac A, Fu T, Patras KA, Christy MP, Bodai Z, Belda-Ferre P, Tripathi A, Chung LK et al. 2020. Global chemical effects of the microbiome include new bile-acid conjugations. *Nature* **579**: 123-129.

Quinn RA, Navas-Molina JA, Hyde ER, Song SJ, Vazquez-Baeza Y, Humphrey G, Gaffney J, Minich JJ, Melnik AV, Herschend J et al. 2016. From Sample to Multi-Omics Conclusions in under 48 Hours. *mSystems* **1**.

Reck M, Tomasch J, Deng Z, Jarek M, Husemann P, Wagner-Dobler I, Consortium C. 2015. Stool metatranscriptomics: A technical guideline for mRNA stabilisation and isolation. *BMC Genomics* **16**: 494.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537-7541.

Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. 2016. Untargeted Metabolomics Strategies-Challenges and Emerging Directions. *J Am Soc Mass Spectrom* **27**: 1897-1905.

Shakya M, Lo CC, Chain PSG. 2019. Advances and Challenges in Metatranscriptomic Analysis. *Front Genet* **10**: 904.

Tanca A, Palomba A, Deligios M, Cubeddu T, Fraumene C, Biosa G, Pagnozzi D, Addis MF, Uzzau S. 2013. Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PLoS One* **8**: e82981.

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

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

Tang WH, Wang Z, Levison BS, Koeth RA, Britt EB, Fu X, Wu Y, Hazen SL. 2013. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N Engl J Med* **368**: 1575-1584.

Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C. 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**: 1895-1904.

van der Hooft JJ, Wandy J, Barrett MP, Burgess KE, Rogers S. 2016. Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci U S A* **113**: 13738-13743.

Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, Lefsrud MG, Apajalahti J, Tysk C, Hettich RL et al. 2009. Shotgun metaproteomics of the human distal gut microbiota. *ISME J* **3**: 179-189.

Walker AW, Martin JC, Scott P, Parkhill J, Flint HJ, Scott KP. 2015. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome* **3**: 26.

Walters WA, Xu Z, Knight R. 2014. Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Lett* **588**: 4223-4233.

Wang J, Zhang X, Li L, Ning Z, Mayne J, Schmitt-Ulms C, Walker K, Cheng K, Figeys D. 2020. Differential Lysis Approach Enables Selective Extraction of Taxon-Specific Proteins for Gut Metaproteomics. *Anal Chem* **92**: 5379-5386.

Wang M Carver JJ Phelan VV Sanchez LM Garg N Peng Y Nguyen DD Watrous J Kapono CA Luzzatto-Knaan T et al. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**: 828-837.

Want EJ, O'Maille G, Smith CA, Brandon TR, Uritboonthai W, Qin C, Trauger SA, Siuzdak G. 2006. Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry. *Anal Chem* **78**: 743-752.

Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, Xia LC, Xu ZZ, Ursell L, Alm EJ et al. 2016. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J* **10**: 1669-1681.

Xavier RJ, Podolsky DK. 2007. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448**: 427-434.

Xiong W, Giannone RJ, Morowitz MJ, Banfield JF, Hettich RL. 2015. Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut. *J Proteome Res* **14**: 133-141.

Zhang X, Deeke SA, Ning ZB, Starr AE, Butcher J, Li J, Mayne J, Cheng K, Liao B, Li LY et al. 2018. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nature Communications* **9**.

Zhang X, Figeys D. 2019. Perspective and Guidelines for Metaproteomics in Microbiome Studies. *J Proteome Res* doi:10.1021/acs.jproteome.9b00054.

Zhang X, Ning Z, Mayne J, Deeke SA, Li J, Starr AE, Chen R, Singleton R, Butcher J, Mack DR et al. 2016a. In Vitro Metabolic Labeling of Intestinal Microbiota for Quantitative Metaproteomics. *Anal Chem* **88**: 6120-6125.

Zhang X, Ning Z, Mayne J, Deeke SA, Walker K, Farnsworth CL, Stokes MP, Mack D, Stintzi A, Figeys D. 2019. Deep characterization of the protein lysine acetylation in human gut microbiome and its alterations in patients with Crohn's disease. *bioRxiv* doi:10.1101/772483: 772483.

Zhang X, Ning ZB, Mayne J, Moore JI, Li J, Butcher J, Deeke SA, Chen R, Chiang CK, Wen M et al. 2016b. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **4**.

Zhang YZ, Li YY. 2014. Inflammatory bowel disease: pathogenesis. *World J Gastroenterol* **20**: 91-99.

Zhou W, Sailani MR, Contrepois K, Zhou Y, Ahadi S, Leopold SR, Zhang MJ, Rao V, Avina M, Mishra T et al. 2019. Longitudinal multi-omics of host-microbe dynamics in prediabetes. *Nature* **569**: 663-671.

Chapter 2

Organ level protein networks as a reference for the host effects of the microbiome

2.1 Abstract

Connections between the microbiome and health are rapidly emerging in a wide range of diseases. However, a detailed mechanistic understanding of how different microbial communities are influencing their hosts is often lacking. One method researchers have used to understand these effects are germ-free mouse models. Differences found within the organ systems within these model organisms may highlight generalizable mechanisms that microbiome dysbioses have throughout the host. Here, we applied multiplexed, quantitative proteomics on the brains, spleens, hearts, small intestines and colons of conventionally raised and germ-free mice, identifying associations to colonization state in over 7,000 proteins. Highly ranked associations were constructed into protein-protein interaction networks and visualized onto an interactive 3D mouse model for user-guided exploration. These results act as a resource for microbiome researchers hoping to identify host effects of microbiome colonization on a given organ of interest. Our results include validation of previously reported effects in xenobiotic metabolism, the innate immune system and glutamate-associated proteins while simultaneously providing organism-wide context. We highlight organism-wide differences in mitochondrial proteins including consistent increases in NNT, a mitochondrial protein with essential roles in influencing levels of NADH and NADPH, in all analyzed organs of conventional mice. Our networks also reveal new associations for further exploration including protease responses in the spleen, high-density lipoproteins in the heart, and glutamatergic signaling in the brain. In total, our study provides a resource for microbiome researchers through detailed tables and visualization of the protein-level effects of microbial colonization on several organ systems.

2.2 Introduction

The gut microbiome is emerging as a critical component of human health. It has been shown that the microbial communities colonizing our bodies play important roles in the immune development of infants(Milani et al. 2017) and the regulation of the innate immune system(Thaiss et al. 2016). Further, a dysbiosis of the gut microbiome has been correlated with many diseases including Inflammatory Bowel Disease (IBD)(Sartor and Wu 2017), diabetes(Tilg and Moschen 2014), obesity(Bouter et al. 2017), cardiovascular disease(Ahmadmehrabi and Tang 2017) and mental health disorders(Nguyen et al. 2018). Microbial production or modification of metabolites such as bile acids, choline derivatives, vitamins and lipids provide some insight into the underlying host-microbe interactions in these diseases(Nicholson et al. 2012). However, many mechanisms mediating these disease states remain unknown.

Germ-free (GF) mouse models, wherein a mouse is raised without any exposure to microbes, have been an invaluable tool for assessing causal effects in microbiome research(Bhattarai and Kashyap 2016). GF models also provide an opportunity to understand the fundamental effects of microbial colonization at an organismal scale. Systems level analyses of the tissues of GF mice have been performed, but these studies have generally highlighted a select few organ tissues. Protein-level studies have shown varying responses to colonization along different regions of the gastrointestinal (GI) tract (Lichtman et al. 2016), changes in drug metabolizing proteins in livers and kidneys(Kuno et al. 2016), and differences in circulating fatty acids from an analysis of serum and livers(Kindt et al. 2018). They have also shown that microbial colonization alters post-translational modifications including histone acetylation and methylation in liver, colon, and adipose tissue(Krautkramer et al. 2016), as well as lysine acetylation in the gut and liver(Simon et al. 2012). An important transcriptomic study revealed a strong connection between colonization and increased *Nnt*, a mitochondrial protein that has functions in redox homeostasis and biosynthetic pathways through the generation of NADH and NADP+(Mardinoglu et al. 2015). The authors found *Nnt* transcripts increased in several conventional mouse tissues including sections of the small intestine, colon and liver, which correlated with significant alterations in host amino acid levels and glutathione metabolism(Mardinoglu et al. 2015). Other related studies found transcript differences in the brain(Diaz Heijtz et al. 2011), and further highlighted the GI-dependent transcript effects of microbial colonization in *Myd88* deficient mice(Larsson et al. 2012).

Here, we sought to further detail the protein effects of microbiota colonization occurring both inside and outside of the GI tract as a reference for microbiome researchers interested in a given host protein, organ system or protein-network. Associations of highly ranked proteins were constructed into protein-protein interaction networks for the brain, spleen, heart, small intestine and colon, as well as a global network. While the brains and gastrointestinal tract of GF mice have been characterized in several studies, other organs such as the heart and spleen may be of interest given the emerging roles of the microbiota in atherosclerosis(Karlsson et al. 2012) and immune development(Chung et al. 2012). We hypothesized that applying methods for improved accuracy in quantitative proteomics (Ting et al. 2011) would further define the influence of the microbiota in each of these organs. With this, we hope to reveal microbiota-induced changes that could underlie disease states. Our results validate a body of literature in the field, provide visualization tools for contextualizing the organism-wide effects of microbial colonization, and identify several new host-microbiota associations for further investigation.

2.3 Results

Construction of protein-protein interaction networks

To determine the protein-level consequences of microbial colonization, three biological replicates of five different tissues (brain, small intestine, colon, spleen, heart) were analyzed from conventionally-raised or GF mice. Multiplexed proteomic analysis of tissue homogenates resulted in the quantification of 7752 proteins overall, of which 4663 were quantified across all samples. These 4663 proteins were used for downstream analysis. A separate pilot study of the brain tissue resulted in the quantification of 6203 proteins.

On a per-organ basis, we contrasted the protein abundances of tissue collected from conventional mice against tissue from GF mice. We ranked the association of each protein to colonization state by accounting for both significance level and fold-change. Interaction networks were built in order to identify groups of proteins whose expression was modulated by microbial colonization. We first constructed organ-specific protein interaction networks containing all the proteins with a highly ranked ($|\pi| > 1$) association to the colonization state of a given organ (Fig. 2.1). Functional enrichments within these organ networks were then assessed by gene-set enrichment analysis (Fig. 2.2). To

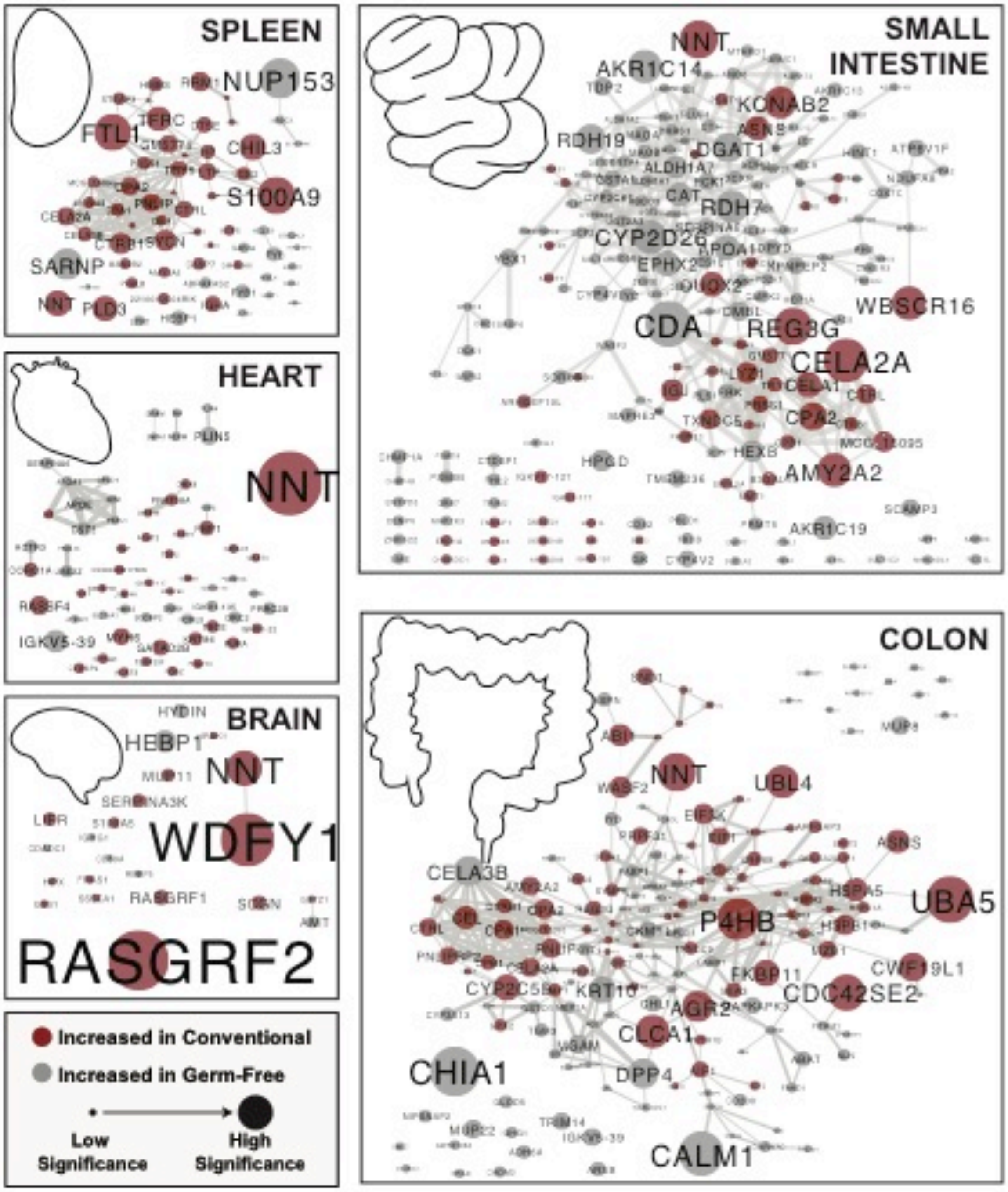


Figure 2.1 Organ-specific protein networks modulated by microbial colonization. Proteins significantly increased in either germ-free or conventional animals were analyzed for interactions through String-db. Edges in each node represent the combined score accounting for all interaction sources. The edges are sized by the combined score with the minimum threshold being 0.4 (of a maximum confidence 1). Nodes represent gene names of significant proteins with a minimum statistical cutoff of |π| >1. Red indicates a significantly higher presence in conventional mice and grey indicates the opposite. The nodes are sized by the level of significance as assessed by π-score.

identify overlapping mechanisms occurring throughout all organs, we compiled all highly ranked proteins within each organ ($|\pi| > 1$) into a single protein network (Fig. 2.3).

Colonization state of the mouse appeared to have larger impacts on organs in direct contact with the gut microbiota, namely the small intestine and colon. The GI organs analyzed had an average of 210 proteins associated with colonization state while the three organs outside of the GI tract averaged 52. The brain yielded the lowest number of associated proteins with only 22. GI tract organs also displayed a higher percentage of interconnectivity, with an average of 71% of associated proteins within GI organs having a moderate-confidence connection to another associated protein within the organ, while organs outside the GI tract averaged 39% (Fig. 2.1). We hypothesize that the interconnectivity and number of associations to colonization is related to the direct contact of GI organs with microbiota. However, it is possible that including a higher portion of intestinal tissue may have influenced these results.

Validation of protein-networks through previously reported associations

Our networks provided support for previously reported broad-scale effects on GI organs, as well as abundance shifts from specific transcripts or proteins. As shown in a previous proteomic study(Lichtman et al. 2016), the small intestine and colon displayed distinct changes as a result of microbial colonization. One example of this was the larger portion of proteins associated with GF status within the small intestine (66%) than in the colon (42%, Fig. 2.1).

Other studies identified similar results at a pathway and individual protein level. After a literature search for protein or transcript differences within GF models, we identified seven publications with related findings. In brief,

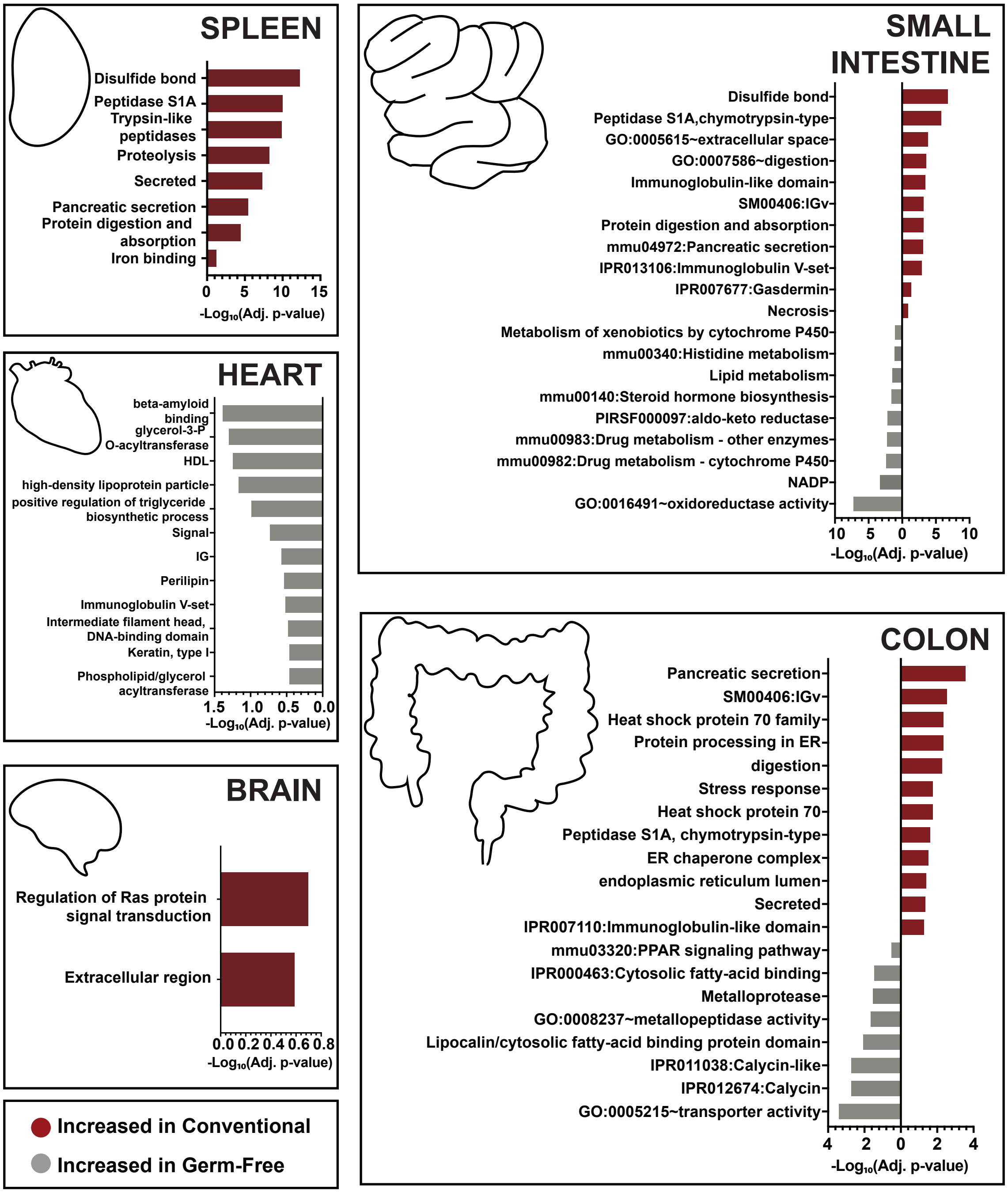


Figure 2.2 Functional enrichments associated with microbial colonization within each organ system. Proteins significantly increased in either germ-free or conventional animals within each organ were analyzed for functional enrichments using DAVID. All proteins identified within the experiment were used as a background. Displayed are barplots showing the -Log$_{10}$(adj. p-values) associated with selected functional groupings. Benjamini-hochberg correction was applied to account for multiple hypothesis testing. The bars are plotted in red if they are associated with the proteins enriched among conventional mice and grey if they are associated with germ-free mice.

changes in xenobiotic metabolism were reported in the GI, liver and kidney, both at the transcriptional level(Fu et al. 2017) and the protein level(Kuno et al. 2016). Our networks also highlighted changes in glutamate related proteins which were previously reported at the transcriptional level(El Aidy et al. 2013). We also report the regulation of innate immune proteins including the antimicrobial peptide REG3G(Larsson et al. 2012), and the regulation of NNT(Mardinoglu et al. 2015), which will be discussed further below.

Organ-specific network results

Functional enrichment analysis of GI tract organs resulted in stronger and more diverse associations to colonization status than organs outside the GI tract. Several enrichments emerged with potential links to redox shifts in the small intestine. These included disulfide bonds, oxidoreductase activity, NADP, and xenobiotic metabolism through cytochrome P450s (Fig. 2.2, Fig. 2.3). Functional differences in the colon highlighted pancreatic secretion, immunoglobulins, proteins of the heat shock protein 70 family, digestion, and stress as the functions increased in conventional mice (Fig. 2.2, Fig. 2.3). GF colons had enrichments for transporter activity, Calycin, fatty-acid binding, metalloproteases and peroxisome proliferator-activated receptor (PPAR) signaling (Fig. 2.2). Together, these results may indicate stress response as an important factor mediating host-microbiota interactions in the colon and redox states influencing interactions of the small intestine.

Organs outside of the GI tract have been less studied in regards to their regulatory responses to the microbiome. Within the spleen, proteins increased in conventional mice were found to be part of a highly connected functional protein network consisting primarily of pancreatic digestive enzymes (CELA2A, CELA3B, CELA1, CPA1, CPA2, CTRB1, CTRL, CTSE, TRY5, etc.), iron-binding proteins (FST1, TFRC, STEAP3, FECH, LTF and HP) and innate immune mediators (LCN2, S100A9, NGP, ITGAM and CHIL3) (Fig. 2.1, Fig. 2.3). Many of the digestive enzymes have roles in the GI tract and were similarly associated to colonization state within the small intestine and colon (Fig. 2.3). Iron-binding and immune proteins were similarly regulated in other organs (Fig. 2.3), but the differential regulation of LTF (Lactotransferrin) and FSLT1 (Follistatin-related protein 1) were primarily restricted to the spleen. Given the biological roles of the spleen, the influence of these proteins in mediating immune processes may be of significant interest.

The networks associated with the brain and the heart displayed fewer interconnected proteins. Only 34% and 27% of proteins had a connection within the heart network and brain network respectively. However, a group of primarily GF-associated proteins that included APOA1 and APOE was found among the heart proteins (Fig. 2.1). These results may suggest changes in lipid profiles in the heart; increased high-density lipoproteins (HDL) and chylomicrons in GF compared to conventional mice. The brain showed limited functional enrichments. However, both RASGRF1 and RASGRF2 were increased among conventional mice. These proteins may be of interest given implications in glutamatergic excitatory synaptic signaling, and in facilitating long-term potentiation leading to enhanced memory, learning, and synaptic plasticity(Drake et al. 2011; Schwechter et al. 2013).

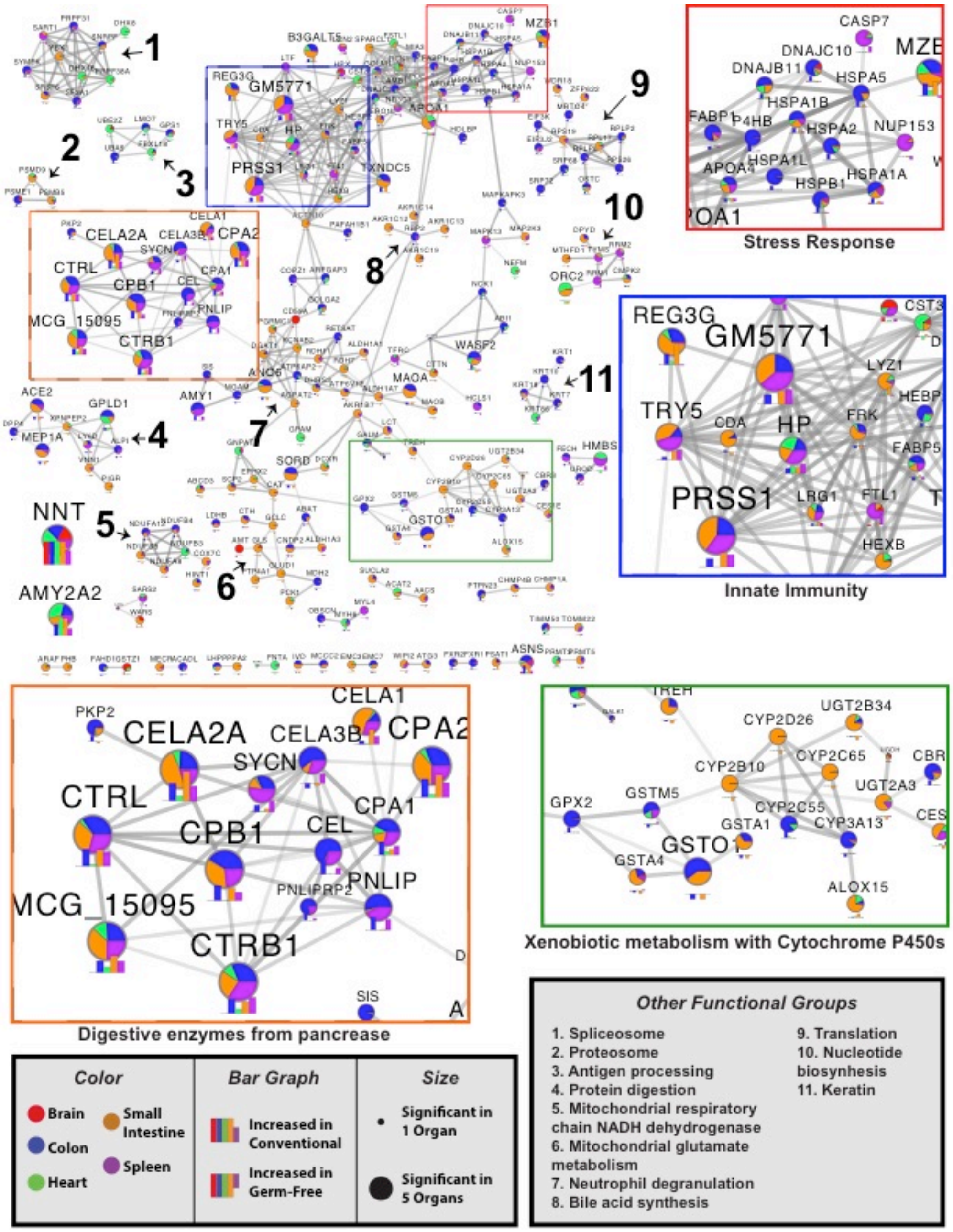


Figure 2.3 Combined organ protein networks modulated by microbial colonization. Proteins significantly increased in either GF or conventional animals were analyzed for interactions through String-db. Edges in each node represent the combined score accounting for all interaction sources. The edges are sized by the combined score with the minimum threshold being 0.8 (of a maximum confidence 1). Nodes represent gene names of proteins with a highly ranked association (a minimum statistical cutoff of |π| >1) within at least one organ. Nodes are sized by the number of organs the protein had a strong association with. The level of association of each node to a particular organ is colored according to the fraction of the total |π|-score each organ provides. Within each node is a bar plot of the π-scores for each organ within the node. Putative functional groupings within the network are highlighted. Select sections are highlighted in colored boxes and shown in 2× zoom.

Organism-wide network results

One prominent finding from our generalized network was a significant increase in NNT in all conventional tissues. This protein is a key regulator of generalized biosynthetic processes, and is related to glutamate synthesis(Mardinoglu et al. 2015). Statistically, NNT was among the strongest relationships found within all organs analyzed (π = 3.8, 4, 2.3, 7, and 3.7 for small intestine, colon, spleen, heart and brain respectively). We also identified sub-networks related to mitochondrial glutamate metabolism and mitochondrial respiratory chain NADH dehydrogenase, with most of the related proteins downregulated in the small intestine of conventional mice (Fig. 2.3). In addition, proteins relating to the mitochondrial reduction of glutathione, GLUD1 and GLS had confirmed associations to previous transcriptomic analysis. While this sub-network was largely derived from small intestine proteins, there was evidence that this system may also be affected in the brain through AMT, which is involved in the mitochondrial metabolism of glycine.

The protein networks identified generalized differences among all organs related to the innate and adaptive immune systems. Proteins related to innate immunity tended to be increased in conventional mice, while proteins associated with neutrophil degranulation were associated with germ-free mice (Fig. 2.3). REG3G, a protein associated with Toll-like receptor (TLR) signaling subsequent to pathogen-associated molecular pattern (PAMP) activation in Paneth cells, was increased among conventional mice. Additionally, proteins related to antigen processing were moderately enriched in GF mice.

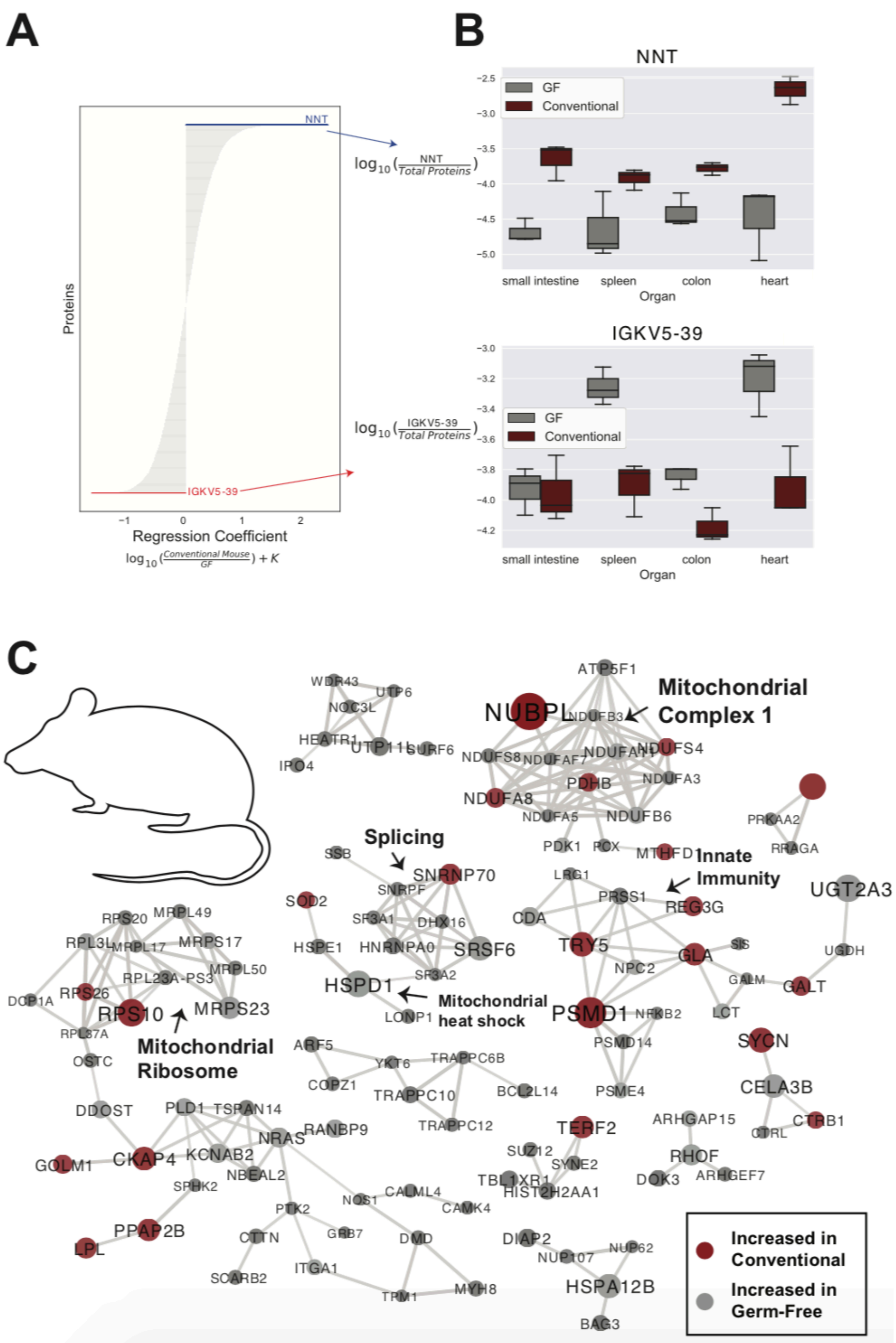


Figure 2.4 Organism level protein networks modulated by the microbiome. A multinomial regression controlling for organ and microbial colonization state of the mice was used to assess proteins associated with colonization status. (**A**) Proteins ranked by regression coefficient; proteins with coefficients of the greatest magnitude are most associated with colonization status. Proteins with positive coefficients are more abundant in conventional mice, while proteins with negative coefficients are more abundant in GF mice. (**B**) Log abundance of NNT or IGKV5-39 over the entire proteome in each organ. (**C**) Protein-protein interaction networks from the top ranked proteins from the multinomial regression associated with both conventional and GF status when controlling for organ and mouse. The top 150 proteins associated with both GF and conventional status were analyzed (300 proteins total), and proteins with high confidence interactions (0.8) are shown. Nodes are sized by the absolute value of the regression coefficient and colored by association to GF (grey) or conventional (red) status. Putative functional groupings are indicated.

We next evaluated protein relationships to GF status at an organismal-level by applying a compositionally aware multinomial regression technique. This technique accounts for organ type to assess organism-wide protein associations through ranks. Our top ranked protein associated with conventional status was NNT, while the protein with the strongest association to GF status was IGKV5-39, a protein involved in immune response and immunoglobin production (Fig. 2.4A). As observed from the traditional statistical approach, NNT was significantly higher in conventional mice within all organs included in the regression (Fig. 2.4B). IGKV5-39 was more strongly associated with the spleen, colon and heart than the small intestine (Fig. 2.4B).

Next we assessed the 150 top and bottom ranked proteins associated with GF status from the multinomial regression and created a protein-protein interaction network (Fig. 4C). Of interest was a cluster of proteins related to redox states in mitochondria. Several proteins including ECSIT, NDUFS4, NUBPL, NDUFA11, NDUFB6, NDUFA8, ATPAF1 are all related to mitochondrial complex 1 of the electron transport chain. Other evidence of the organism-wide impact of microbial colonization on mitochondria included shifts in mitochondrial ribosomal proteins (MRPS23, MRPL50, MRPS17, MRPL49 and MRPL17) and proteins related to mitochondrial heat shock response (HSPD1, HSPE1, SOD2, LONP1). The data presented here suggests that microbial colonization may have organismal-level impacts on mitochondrial function.

Interactive 3D Visualization of Associations to Colonization Status

To encourage user interaction with our data, we created a web-based display of our results projected onto a 3D model of a mouse. This interactive display allows for user-guided exploration of the protein and pathway-level associations to colonization status. To access the model, users should access the ‘ili webserver (https://ili.embl.de/), then drag-and-drop the texturization file and a supplemental table for either the proteins or pathways identified in this study. With this tool users can search for proteins and pathways of interest and project the association scores onto all the organs analyzed in this study.

For the protein-level visualization, each protein is listed by the protein name with any Gene Ontology (GO) molecular function terms associated with the protein. With this, users are able to search for a protein of interest or identify proteins of interest by molecular functions. The ‘ili-compatible table of association scores (Supplemental Table S4) uses the π-statistic for conventional/GF status. These scores range from 6.98 (significantly increased in conventional mice) to -5.11 (significantly increased in GF mice). An example use case is displayed in Figure 5A, which depicts the strong association to conventional status for NNT within all organs.

The pathway-level visualizations help to summarize and explore the gene set enrichment analyses. On a per-organ basis, an association to conventional or GF status for each functional category was calculated by comparing the statistical strength of gene set enrichments. These association scores have been summarized in an ‘ili-compatible file in supplemental table. Pathway association scores ranged from 12.90 (highly associated to conventional status) to -7.25 (highly associated with GF status). We demonstrate the use of this tool to visualize the associations to “Oxidoreductase” in Figure 2.5B.

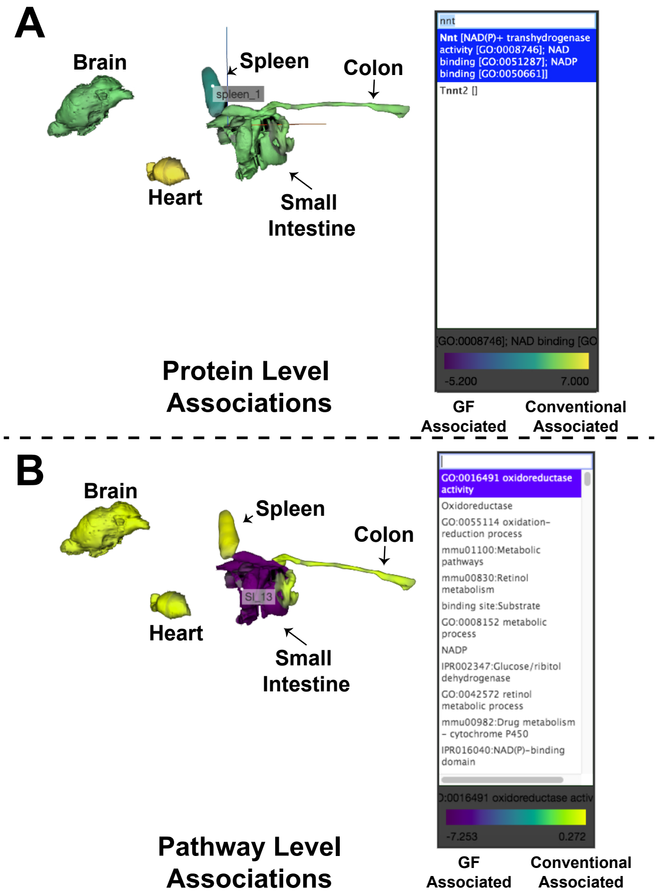

A
Brain
Spleen
spleen_1
Colon
Heart
Small Intestine
nnt
Nnt [NAD(P)+ transhydrogenase activity [GO:0008746]; NAD binding [GO:0051287]; NADP binding [GO:0050661]]
Tnnt2 []
[GO:0008746]; NAD binding [GO
-5.200
7.000
Protein Level Associations
GF Associated
Conventional Associated
B
Brain
Spleen
Colon
SI_13
Heart
Small Intestine
GO:0016491 oxidoreductase activity
Oxidoreductase
GO:0055114 oxidation–reduction process
mmu01100:Metabolic pathways
mmu00830:Retinol metabolism
binding site:Substrate
GO:0008152 metabolic process
NADP
IPR002347:Glucose/ribitol dehydrogenase
GO:0042572 retinol metabolic process
mmu00982:Drug metabolism – cytochrome P450
IPR016040:NAD(P)–binding domain
):0016491 oxidoreductase activ
-7.253
0.272
Pathway Level Associations
GF Associated
Conventional Associated

Figure 2.5 Interactive 3D Visualization of Associations to Colonization Status. A 3D mouse model was generated for use on the web-based ili platform (https://ili.embl.de/). (**A**) An example use case for the protein-level association visualizations are shown through plotting the π-score enrichment for conventional colonization status. (**B**) An example use case for the pathway level association visualizations are shown through highlighting the enrichment scores for "Oxidoreductase". Pathway association scores were generated through –Log$_{10}$(Benjamini-corrected p-values) of the conventional organs minus the GF organs.

2.4 Discussion

Our multi-organ analysis was utilized to generate protein interaction maps, and 3D visualization tools that researchers can use to understand organ-specific and organism-wide changes that may underlie host-microbiome interactions. From our networks we can identify common themes found from previous analyses of GF tissue. For example, we found changes in innate and adaptive immune responses to microbial colonization(Larsson et al. 2012), changes in the glutamine and glutamate pathway(El Aidy et al. 2013; Mardinoglu et al. 2015), and changes in xenobiotic degradation pathways (Kuno et al. 2016; Fu et al. 2017; Kindt et al. 2018). This study contributes to the field by moving toward an organism-wide understanding of host-microbiome interactions. Here, we incorporate new statistical and visualization tools for multi-organ analysis and include understudied organs from outside of the GI tract. As roles for the microbiome are expanding into immunity (Thaiss et al. 2016), cardiovascular disease(Ahmadmehrabi and Tang 2017), and mental health disorders(Nguyen et al. 2018), defining the influence of the microbiome on these tissues may be of importance. Indeed, our networks provided several putative organism-level roles for the microbiome.

Our protein networks highlight the potential of organism-wide redox state changes being linked to the microbiome. This is perhaps best highlighted in the distal increases of Nnt associated with microbial colonization. NNT is a mitochondrial protein with well-described roles in glutathione redox reactions through the conversion of NADPH to NADP+(Ronchi et al. 2013). Glutathione is interconverted with glutathione disulfide, collectively representing the most abundant redox pair in the body(Circu and Aw 2011). Abundances of this redox pair are often used as an indication of the general redox state(Circu and Aw 2011), which is involved in a large variety of biological processes(Circu and Aw 2011; Birben et al. 2012; Ray et al. 2012). Evidence of intestinal redox differences in GF animals has been known since the 1970s(Koopman et al. 1975; Celesk et al. 1976)(Celesk, Asano, & Wagner, 1976; Koopman, Janssen, & van Druten, 1975) and evidence of microbiota-dependent regulation of *Nnt* in the GI tract has been previously described(Mardinoglu et al. 2015). Here we find evidence of increased NNT throughout the entire conventional mouse. Additionally, both the traditional and multivariate statistical methods used for the identification of organism-wide effects found networks of proteins related to mitochondria, including mitochondrial complex I proteins. Mitochondrial complex I activity might also be linked to redox states as complex I activity was shown to be dependent on glutathione transport into the mitochondria(Kamga et al. 2010). In summary, many of our protein networks may have been influenced by redox states, including shifts in mitochondrial proteins, the innate immune processes such as neutrophil degranulation, and degradation of xenobiotics through cytochrome P450s (Kramer and Darley-Usmar 2015).

Though our strongest relationships to microbial colonization were found within the GI tract, which was unsurprisingly the focus of most previous GF organ analyses(Larsson et al. 2012; El Aidy et al. 2013; Mardinoglu et al. 2015; Lichtman et al. 2016), the analyses of the brain, spleen and heart did yield several interesting results. Within the heart, our analyses indicated a relationship to HDL and the germ-free state. Our results within the heart may be of importance given the increasing literature regarding microbiota regulation of lipids and lipoproteins including HDL(Nakaya and Ikewaki 2018), as well as the microbial links to metabolites leading to atherosclerosis(Wang et al. 2011b). Our findings may indicate a mechanism in which microbiota influence the makeup of the heart through changes in lipoproteins.

The data of the spleen proteomes revealed networks of proteases with similar increases within the small intestine and colon in the presence of the microbiota. These networks also suggested potential links between the gut microbiota and pancreatic secretion. Pancreatic secretion of enzymes is thought to be largely regulated through circulating hormones(Singh and Webster 1978). The gut microbiome has several potential links to pancreatitis and pancreatic cancer, which may be mediated through sensing of microbial compounds such as lipopolysaccharide through TLR4(Leal-Lopes et al. 2015). Here, we have observed a link between microbial colonization and increased pancreatic secretion of digestive proteins, which may be of interest for further investigation.

The organism-wide effects of microbial colonization illustrated in our networks may have implications in several disease states. Dysregulation of Complex I has been implicated in microbiome related diseases including Ulcerative Colitis(Haberman et al. 2019) and there is accumulating evidence of mitochondrial dysfunction in Crohn’s disease(Mottawea et al. 2016). This association between the microbiome and mitochondria are thought to be mediated through three key microbiome metabolites: short-chain fatty acids, the urolithins and lactate(Franco-Obregon and Gilbert 2017). While speculative, it is possible that the IBD microbiota may influence these interactions. Glutamatergic signaling has been suggested as a potential target for treating mood disorders(Zarate et al. 2010). Our identification of proteins in the brain related to glutamatergic signaling and glutamate (i.e. RASGRF proteins and NNT) may be of relevance to the discussions surrounding utilization of the microbiome to treat mood disorders(Mangiola et al. 2016).

There are likely many unknown mechanisms mediating host-microbiota interactions. Our detailed maps and visualization tools for understanding the organ-level impacts of microbial colonization give insight into the unique and common protein-level changes occurring throughout mice. These networks suggest changes throughout the mice related to mitochondrial dysfunction and redox states. Though fewer changes were found in organs apart from the GI tract, our protein networks of the spleen, brain and heart may provide insight for researchers establishing connections between the microbiota and diseases related to these organ systems. We hope these networks and visualization tools may be useful in the microbiome research community to help dissect the specific effects that microbial communities have in a given organ-system, protein or protein network of interest. In total, we view our study as a step toward better understanding the role of the microbiota in health and disease.

2.5 Materials and Methods

Gnotobiotic Mice

Three male germ-free C57/BL6 mice were kept under germ-free conditions in a Park Bioservices isolator in GSU's germ-free facility, and three male conventional C57/BL6 were kept in regular housing at GSU's animal facility. At 5 months of age, mice were euthanized and organs were collected followed by immediate snap-freezing. All mice were bred and housed at Georgia State University, Atlanta, Georgia, USA. under institutionally-approved protocols (IACUC # A14033).

Protein digestion and TMT labeling

All organ proteome methods were preformed as previously described(Lapek et al. 2018). Snap frozen organs (stored at -80 ºC beforehand) were suspended in PBS and homogenized using a Mini BeadBeater (Biospec). Our final analysis contained 3 biological replicates of each colonization condition per organ, and no technical replicates were collected given strong correlation observed between biological replicates in our past work (Lapek et al. 2018). Organ homogenates were lysed in 1 mL of buffer composed of 75 mM NaCl (Sigma-Aldrich), 3% sodium dodecyl sulfate (SDS, Thermo Fisher Scientific), 1 mM NaF (Sigma-Aldrich), 1 mM beta-glycerophosphate (Sigma-Aldrich), 1 mM sodium orthovanadate (Sigma-Aldrich), 10 mM sodium pyrophosphate (Sigma-Aldrich), 1 mM phenylmethylsulfonyl fluoride (PMSF, Sigma-Aldrich), and a Complete Mini EDTA free protease inhibitors (1 tablet per 10 mL, Roche) in 50 mM HEPES (Sigma-Aldrich), pH 8.5(Wessel and Flugge 1984). To ensure full lysis, homogenates were passed through a 21-gauge syringe 20 times. Insoluble debris was then pelleted by centrifugation for 5 minutes at 14,000 rpm. Supernatants were transferred to new tubes and an equal volume of 8 M urea in 50 mM HEPES, pH 8.5 was added to each sample. Samples were then vortexed and further lysed through two 10-second intervals of probe sonication at 25% amplitude.

Proteins were reduced with dithiothreitol (DTT, Sigma-Aldrich) and alkylated with iodoacetamide (IAA, Sigma-Aldrich)(Wessel and Flugge 1984). Proteins were next precipitated via methanol-chloroform precipitation(Wessel and Flugge 1984). Precipitated proteins were re-solubilized in 300 μL of 1 M urea (Thermo Fisher Scientific) in 50 mM HEPES, pH 8.5. Proteins were digested in a two-step digestion process. First, 3 μg of LysC (Wako) was added to each sample, and samples were digested overnight at room temperature. Next, 3 μg of trypsin was added, and samples were digested for six hours at 37 ºC. Digests were acidified with trifluoroacetic acid (TFA, Pierce) to quench the digestion reaction. Peptides were desalted with C18 Sep-Paks (Waters) as previously described(McAlister et al. 2014). Concentration of desalted peptides was determined using a Pierce Quantitative Colorimetric Peptide Assay (Thermo

Fisher Scientific), and peptides were aliquoted into 50 µg portions. Aliquots were dried under vacuum and stored at -80 ºC until they were labeled with TMT reagents.

Peptides were labeled with 10-plex TMT reagents (Thermo Fisher Scientific) (Thompson et al. 2003; McAlister et al. 2014) as previously described(Wang et al. 2011a). TMT reagents were reconstituted in dry acetonitrile (Sigma-Aldrich) at 20 µg/µL. Dried peptides were re-suspended in 30% dry acetonitrile in 200 mM HEPES, pH 8.5, and 7 µL of the appropriate TMT reagent was added to peptides. Reagents 126 and 131 (Thermo Fisher Scientific) were used to label peptide aliquots composed of an equal concentration of every sample within all mass spec runs. These composite samples acted as a reference within all mass spec runs for data normalization purposes described later. Remaining reagents were used to label samples in random order with no bias regarding animal of origin, organ or colonization status. This randomization was performed to prevent known batch-effects in mass-spectrometry experiments(Brenes et al. 2019). The brain samples were analyzed as a pilot experiment before the other organs. For brains, all procedures were performed as above, though the TMT experiment was separate from other organs. Within the brain TMT experiment, one channel, 129C, consisted of a 50 µg average of all peptides from brain samples. Labeling was carried out for 1 hour at room temperature, and was quenched by adding 8 µL of 5% hydroxylamine (Sigma-Aldrich). TMT-labeled peptides from each of the organ samples were acidified by adding 50 µL of 1% TFA and subsequently combined into a composite sample per TMT 10-plex experiment. During the pilot experiment with the brains, the pooled samples were desalted and fractionated using a High pH Reversed-Phase Peptide Fractionation Kit (Pierce) per manufacturer instructions. For the larger studied, samples were pooled per

10-plex experiment, desalted with C18 Sep-Paks, and further fractionated as described below.

Collection of LC-MS2/MS3 spectra for protein identification and quantification

Data acquisition methods were performed as previously described(Lapek et al. 2018). Sample fractionation for excluding the brains, was performed by basic pH reverse-phase liquid chromatography with concatenated fractions as previously described(Wang et al. 2011a). Briefly, samples were re-suspended in 5% formic acid/5% acetonitrile and separated over a 4.6 mm × 250 mm C18 column (Thermo Fisher Scientific) on an Ultimate 3000 HPLC fitted with a fraction collector, degasser and variable wavelength detector. The separation was performed over a 22% to 35%, 60-minute linear gradient of acetonitrile in 10 mM ammonium bicarbonate (Thermo Fisher Scientific) at 0.5 mL/min. The resulting 96 fractions were combined as previously described(Wang et al. 2011a). All fractions were dried under vacuum and re-suspended in 5% formic acid/5% acetonitrile and analyzed by liquid chromatography (LC)-MS^2/MS^3 for identification and quantitation.

All LC-MS^2/MS^3 experiments were carried out on an Orbitrap Fusion (Thermo Fisher Scientific) with an in-line Easy-nLC 1000 (Thermo Fisher Scientific) and chilled autosampler. Home-pulled, home-packed columns (100 μm ID × 30 cm, 360 μm OD) were used for analysis. Analytical columns were triple-packed with 5 μm C4 resin, 3 μm C18 resin, and 1.8 μm C18 resin (Sepax) to lengths of 0.5 cm, 0.5 cm, and 30 cm respectively. Peptides were loaded at 500 bar and eluted with a linear gradient of 11% to 30% acetonitrile in 0.125% formic acid over 165 minutes at a flow rate of 300 nL/minute, with the column heated to 60 ºC. Nano-electrospray ionization was performed by applying 2000 V through a stainless-steel T-junction at the inlet of the microcapillary column.

The mass spectrometer was run in data-dependent mode, where a survey scan was performed over 500-1200 m/z at a resolution of 60,000 in the Orbitrap. Automatic gain control (AGC) was set to 2×10^5 for MS^1 with a maximum ion injection time of 100 ms. The S-lens RF was set to 60 and centroided data was collected. Top-N mode was used to select the most abundant ions in the MS^1 scan for MS^2 and MS^3 with N set to 10.

The decision tree option was used for MS^2 analysis, using charge state and m/z range as qualifiers. Ions carrying 2 charges were analyzed from the m/z range of 600-1200, and ions carrying 3 and 4 charges were selected from the m/z range of 500-1200. An ion intensity threshold of 5×10^4 was used. MS^2 spectra were obtained using quadrupole isolation at a 0.5 Th window and fragmented using Collision Induced Dissociation with a normalized collision energy of 30%. Fragment ions were detected and centroided data collected in the linear ion trap using rapid scan rate with an AGC target of 1×10^4 and maximum ion injection time of 35 ms.

MS^3 analysis was performed using synchronous precursor selection (SPS) enabled to maximize TMT quantitation sensitivity(McAlister et al. 2014). A maximum of 10 MS^2 precursors was specified for the SPS setting, which were simultaneously isolated and fragmented for MS^3 analysis. Higher-Energy Collisional Dissociation fragmentation was used for MS^3 analysis with a normalized collision energy of 55%. Resultant fragment ions (MS^3) were detected in the Orbitrap at a resolution of 60,000 with a low mass cut-off of 110 m/z. AGC for MS^3 spectra was set to 1×10^5 with a maximum ion injection time of

100 ms. Centroided data were collected, and MS^2 ions between the range of 40 m/z below and 15 m/z above the precursor m/z were excluded by SPS.

Data processing and normalization

Data were processed using Proteome Discover 2.1 (Thermo Fisher Scientific). MS^2 data were searched against UniProt mouse databases (downloaded 7/2/2018 and 5/11/2017 for the brain analysis and other organs respectively) using the Sequest algorithm(Huttlin et al. 2010). A decoy search was also conducted with sequences in reverse order(Huttlin et al. 2010) was specified and 0.6 Da tolerance for MS^2 fragments. Static modification of TMT 10-plex tags on lysine and peptide n-termini (+229.162932 Da) and carbamidomethylation of cysteines (+57.02146 Da) were specified. Variable oxidation of methionine (+15.99492 Da) was also included in search parameters. Data were filtered to 1% peptide and protein level false discovery rates with the target-decoy strategy through Percolator (Xiao et al., 2014).

TMT reporter ion intensities were extracted from MS^3 spectra for quantitative analysis, and signal-to-noise values were used for quantitation. Spectra were filtered and summed as previously described(Xiao et al. 2014). Data were normalized in a multi-step process, whereby they were first normalized to the pooled standards (TMT-126 and -131) for each protein, and then to the median signal across the pooled standards from all experiments(Xiao et al. 2014). An average of these normalizations was used for the next step. As the brain samples did not contain a composite bridge channel for normalization, raw signal/noise ratios were normalized by the average signal of the given protein divided by the median of all protein averages. To account for slight differences in amounts of protein labeled, these values were then normalized to the median of the entire dataset and reported as final normalized summed signal-to-noise ratios per protein per sample.

Statistical analysis

Bioinformatic analysis was performed in Python (version 3.5.1) and records are available online in Jupyter Notebooks (https://github.com/rhmills/Germ-free-organ-proteomics). To prevent statistical artifacts generated due to the various methods of dealing with missing values, only proteins with quantification in all samples were used. A Student's *t*-test with unequal variance was performed through the package, SciPy (https://www.scipy.org). For ranking purposes, we evaluated associations to colonization state through π–score, which accounts for both fold change and p-value(Xiao et al. 2014). A statistical cutoff of $|\pi| > 1$ was chosen based on previous work(Tran et al. 2019). This statistical measure corresponds to a significance level of $\alpha \sim 0.05$, and allowed for moderate stringency while including an adequate number of proteins for protein network construction and functional enrichment analysis.

Protein-protein interaction networks were created through STRING-db(Shannon et al. 2003). Associations between proteins were determined through default settings, accounting for textmining, experiments, databases, co-expression, neighborhood, gene fusion and co-occurrence. Connections were restricted to interactions between proteins within the query list only. Networks were subsequently visualized through Cytoscape (version 3.5.1)(Shannon et al. 2003). Edges within protein networks were based on the combined evidence scores, with thicker edges indicating higher confidence. Per-organ networks were performed with a medium minimum confidence (0.4) to visualize connectivity through maximizing potential connections, while combined organ networks utilized a high minimum confidence (0.8) to identify putative functional groupings.

Functional enrichment analysis was performed through the DAVID server(Huang da et al. 2009) to identify significant groups of proteins per organ, split between Germ-free and colonized states. Parameters were set as previously described(Tran et al. 2019). Benjamini-hochberg corrected p-values were reported for the most significant groupings. Barplots were visualized through GraphPad Prism (version 7.0b).

Songbird(Morton et al. 2019) was used to implement the multinomial regression analysis with organ, mouse, and colonization status used in the regression formula. Model parameters were as follows: 10,000 epochs, batch size of 5, differential prior of 1.0, learning rate of 0.001, gradient clipping size of 10, and proteins with >5 counts were included. As the brain tissue was processed as an independent pilot study, we could not include this data as part of the multinomial regression analysis given the independent normalization used in each experiment.

3D Mouse Model

The 3D Mouse model was generated as described previously(Quinn et al. 2019). In brief, MRI images were acquired at the UCSD Center for Functional MRI from a euthanized 8-week-old female C57BL/6 mouse with a Bruker 7T/20 MRI scanner using a quadrature birdcage transceiver. MRI Parameters were as follows: 3D FLASH protocol with TE/TR=6 ms/15 ms and matrix size 128×64×156, field of view prescribed to match the body size. InVesalius (https://link.springer.com/chapter/10.1007/978-3-319-27857-5_5) was used to trace individual organs in each MRI slice to generate the 3D model. The model was then processed with Blender (https://www.blender.org/) for smoothing.

Interactive display of associations can be done through the following steps: 1) Accessing the ‘ili-web browser (https://ili.embl.de) (Protsyuk et al. 2018). 2) Uploading the 3D mouse model visualization file 3) Uploading either the protein-based associations table or the pathway-based associations table.

Pathway-level association scores were determined by comparing the – Log_{10}(Benjamini-Hochberg corrected p-values) for each functional enrichment in Supplemental Table S2. The statistical strength of the functional enrichment in conventional tissue was subtracted from the statistical strength of the functional enrichment in the GF tissue. The following formula summarizes the calculation:

Association Score = (-Log_{10}(Adj. p-value for conventional status) – -Log_{10}(Adj. p-value for GF Status))

Data Access

The proteomic data generated in this study have been submitted to the Mass Spectrometry Interactive Virtual Environment (MassIVE) Repository (https://massive.ucsd.edu) under the study ID MSV000083874. Code is available through GitHub (https://github.com/rhmills/Germ-free-organ-proteomics), and as Supplemental Code.

Chapter 2 is a reprint of the material as it appears in Genome Research, 2020, Robert H. Mills, Jacob M. Wozniak, Alison Vrbanac, Anaamika Campeau, Benoit Chassaing, Andrew Gewirtz, Rob Knight, and David J. Gonzalez. The dissertation author played a primary role in all aspects of the work ranging from the study design, data acquisition, analysis and writing of the manuscript.

2.6 References

Ahmadmehrabi S, Tang WHW. 2017. Gut microbiome and its role in cardiovascular diseases. *Curr Opin Cardiol* **32**: 761-766.

Bhattarai Y, Kashyap PC. 2016. Germ-Free Mice Model for Studying Host-Microbial Interactions. *Methods Mol Biol* **1438**: 123-135.

Birben E, Sahiner UM, Sackesen C, Erzurum S, Kalayci O. 2012. Oxidative stress and antioxidant defense. *World Allergy Organ J* **5**: 9-19.

Bouter KE, van Raalte DH, Groen AK, Nieuwdorp M. 2017. Role of the Gut Microbiome in the Pathogenesis of Obesity and Obesity-Related Metabolic Dysfunction. *Gastroenterology* **152**: 1671-1678.

Brenes A, Hukelmann J, Bensaddek D, Lamond AI. 2019. Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. *Mol Cell Proteomics* **18**: 1967-1980.

Celesk RA, Asano T, Wagner M. 1976. The size pH, and redox potential of the cecum in mice associated with various microbial floras. *Proc Soc Exp Biol Med* **151**: 260-263.

Chung H, Pamp SJ, Hill JA, Surana NK, Edelman SM, Troy EB, Reading NC, Villablanca EJ, Wang S, Mora JR et al. 2012. Gut immune maturation depends on colonization with a host-specific microbiota. *Cell* **149**: 1578-1593.

Circu ML, Aw TY. 2011. Redox biology of the intestine. *Free Radic Res* **45**: 1245-1266.

Diaz Heijtz R, Wang S, Anuar F, Qian Y, Bjorkholm B, Samuelsson A, Hibberd ML, Forssberg H, Pettersson S. 2011. Normal gut microbiota modulates brain development and behavior. *Proc Natl Acad Sci U S A* **108**: 3047-3052.

Drake NM, DeVito LM, Cleland TA, Soloway PD. 2011. Imprinted Rasgrf1 expression in neonatal mice affects olfactory learning and memory. *Genes Brain Behav* **10**: 392-403.

El Aidy S, Derrien M, Merrifield CA, Levenez F, Dore J, Boekschoten MV, Dekker J, Holmes E, Zoetendal EG, van Baarlen P et al. 2013. Gut bacteria-host metabolic interplay during conventionalisation of the mouse germfree colon. *ISME J* **7**: 743-755.

Franco-Obregon A, Gilbert JA. 2017. The Microbiome-Mitochondrion Connection: Common Ancestries, Common Mechanisms, Common Goals. *mSystems* **2**.

Fu ZD, Selwyn FP, Cui JY, Klaassen CD. 2017. RNA-Seq Profiling of Intestinal Expression of Xenobiotic Processing Genes in Germ-Free Mice. *Drug Metab Dispos* **45**: 1225-1238.

Haberman Y, Karns R, Dexheimer PJ, Schirmer M, Somekh J, Jurickova I, Braun T, Novak E, Bauman L, Collins MH et al. 2019. Ulcerative colitis mucosal transcriptomes reveal mitochondriopathy and personalized mechanisms underlying disease severity and treatment response. *Nat Commun* **10**: 38.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44-57.

Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villen J, Haas W, Sowa ME, Gygi SP. 2010. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**: 1174-1189.

Kamga CK, Zhang SX, Wang Y. 2010. Dicarboxylate carrier-mediated glutathione transport is essential for reactive oxygen species homeostasis and normal respiration in rat brain mitochondria. *Am J Physiol Cell Physiol* **299**: C497-505.

Karlsson FH, Fak F, Nookaew I, Tremaroli V, Fagerberg B, Petranovic D, Backhed F, Nielsen J. 2012. Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nat Commun* **3**: 1245.

Kindt A, Liebisch G, Clavel T, Haller D, Hormannsperger G, Yoon H, Kolmeder D, Sigruener A, Krautbauer S, Seeliger C et al. 2018. The gut microbiota promotes hepatic fatty acid desaturation and elongation in mice. *Nat Commun* **9**: 3760.

Koopman JP, Janssen FG, van Druten JA. 1975. Oxidation-reduction potentials in the cecal contents of rats and mice. *Proc Soc Exp Biol Med* **149**: 995-999.

Kramer PA, Darley-Usmar VM. 2015. The emerging theme of redox bioenergetics in health and disease. *Biomed J* **38**: 294-300.

Krautkramer KA, Kreznar JH, Romano KA, Vivas EI, Barrett-Wilt GA, Rabaglia ME, Keller MP, Attie AD, Rey FE, Denu JM. 2016. Diet-Microbiota Interactions Mediate Global Epigenetic Programming in Multiple Host Tissues. *Mol Cell* **64**: 982-992.

Kuno T, Hirayama-Kurogi M, Ito S, Ohtsuki S. 2016. Effect of Intestinal Flora on Protein Expression of Drug-Metabolizing Enzymes and Transporters in the Liver and Kidney of Germ-Free and Antibiotics-Treated Mice. *Mol Pharm* **13**: 2691-2701.

Lapek JD, Jr., Mills RH, Wozniak JM, Campeau A, Fang RH, Wei X, van de Groep K, Perez-Lopez A, van Sorge NM, Raffatellu M et al. 2018. Defining Host

Responses during Systemic Bacterial Infection through Construction of a Murine Organ Proteome Atlas. *Cell Syst* doi:10.1016/j.cels.2018.04.010.

Larsson E, Tremaroli V, Lee YS, Koren O, Nookaew I, Fricker A, Nielsen J, Ley RE, Backhed F. 2012. Analysis of gut microbial regulation of host gene expression along the length of the gut and regulation of gut microbial ecology through MyD88. *Gut* **61**: 1124-1131.

Leal-Lopes C, Velloso FJ, Campopiano JC, Sogayar MC, Correa RG. 2015. Roles of Commensal Microbiota in Pancreas Homeostasis and Pancreatic Pathologies. *J Diabetes Res* **2015**: 284680.

Lichtman JS, Alsentzer E, Jaffe M, Sprockett D, Masutani E, Ikwa E, Fragiadakis GK, Clifford D, Huang BE, Sonnenburg JL et al. 2016. The effect of microbial colonization on the host proteome varies by gastrointestinal location. *ISME J* **10**: 1170-1181.

Mangiola F, Ianiro G, Franceschi F, Fagiuoli S, Gasbarrini G, Gasbarrini A. 2016. Gut microbiota in autism and mood disorders. *World J Gastroenterol* **22**: 361-368.

Mardinoglu A, Shoaie S, Bergentall M, Ghaffari P, Zhang C, Larsson E, Backhed F, Nielsen J. 2015. The gut microbiota modulates host amino acid and glutathione metabolism in mice. *Mol Syst Biol* **11**: 834.

McAlister GC, Nusinow DP, Jedrychowski MP, Wuhr M, Huttlin EL, Erickson BK, Rad R, Haas W, Gygi SP. 2014. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal Chem* **86**: 7150-7158.

Milani C, Duranti S, Bottacini F, Casey E, Turroni F, Mahony J, Belzer C, Delgado Palacio S, Arboleya Montes S, Mancabelli L et al. 2017. The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. *Microbiol Mol Biol Rev* **81**.

Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, Zengler K, Knight R. 2019. Establishing microbial composition measurement standards with reference frames. *Nat Commun* **10**: 2719.

Mottawea W, Chiang CK, Muhlbauer M, Starr AE, Butcher J, Abujamel T, Deeke SA, Brandel A, Zhou H, Shokralla S et al. 2016. Altered intestinal microbiota-host mitochondria crosstalk in new onset Crohn's disease. *Nat Commun* **7**: 13419.

Nakaya K, Ikewaki K. 2018. Microbiota and HDL metabolism. *Curr Opin Lipidol* **29**: 18-23.

Nguyen TT, Kosciolek T, Eyler LT, Knight R, Jeste DV. 2018. Overview and systematic review of studies of microbiome in schizophrenia and bipolar disorder. *J Psychiatr Res* **99**: 50-61.

Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, Pettersson S. 2012. Host-gut microbiota metabolic interactions. *Science* **336**: 1262-1267.

Protsyuk I, Melnik AV, Nothias LF, Rappez L, Phapale P, Aksenov AA, Bouslimani A, Ryazanov S, Dorrestein PC, Alexandrov T. 2018. 3D molecular cartography using LC-MS facilitated by Optimus and 'ili software. *Nat Protoc* **13**: 134-154.

Quinn RA, Vrbanac A, Melnik AV, Patras KA, Christy M, Nelson AT, Aksenov A, Tripathi A, Humphrey G, da Silva R et al. 2019. Chemical Impacts of the Microbiome Across Scales Reveal Novel Conjugated Bile Acids. *bioRxiv* doi:10.1101/654756: 654756.

Ray PD, Huang BW, Tsuji Y. 2012. Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. *Cell Signal* **24**: 981-990.

Ronchi JA, Figueira TR, Ravagnani FG, Oliveira HC, Vercesi AE, Castilho RF. 2013. A spontaneous mutation in the nicotinamide nucleotide transhydrogenase gene of C57BL/6J mice results in mitochondrial redox abnormalities. *Free Radic Biol Med* **63**: 446-456.

Sartor RB, Wu GD. 2017. Roles for Intestinal Bacteria, Viruses, and Fungi in Pathogenesis of Inflammatory Bowel Diseases and Therapeutic Approaches. *Gastroenterology* **152**: 327-339 e324.

Schwechter B, Rosenmund C, Tolias KF. 2013. RasGRF2 Rac-GEF activity couples NMDA receptor calcium flux to enhanced synaptic transmission. *Proc Natl Acad Sci U S A* **110**: 14462-14467.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498-2504.

Simon GM, Cheng J, Gordon JI. 2012. Quantitative assessment of the impact of the gut microbiota on lysine epsilon-acetylation of host proteins using gnotobiotic mice. *Proc Natl Acad Sci U S A* **109**: 11133-11138.

Singh M, Webster PD. 1978. Neurohormonal control of pancreatic secretion. A review. *Gastroenterology* **74**: 294-309.

Thaiss CA, Zmora N, Levy M, Elinav E. 2016. The microbiome and innate immunity. *Nature* **535**: 65-74.

Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C. 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**: 1895-1904.

Tilg H, Moschen AR. 2014. Microbiota and diabetes: an evolving relationship. *Gut* **63**: 1513-1521.

Ting L, Rad R, Gygi SP, Haas W. 2011. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* **8**: 937-940.

Tran HQ, Mills RH, Peters NV, Holder MK, de Vries GJ, Knight R, Chassaing B, Gonzalez DJ, Gewirtz AT. 2019. Associations of the fecal microbial proteome composition and proneness to diet-induced obesity. *Mol Cell Proteomics* doi:10.1074/mcp.RA119.001623.

Wang Y, Yang F, Gritsenko MA, Wang Y, Clauss T, Liu T, Shen Y, Monroe ME, Lopez-Ferrer D, Reno T et al. 2011a. Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11**: 2019-2026.

Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, Dugar B, Feldstein AE, Britt EB, Fu X, Chung YM et al. 2011b. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**: 57-63.

Wessel D, Flugge UI. 1984. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal Biochem* **138**: 141-143.

Xiao Y, Hsiao TH, Suresh U, Chen HI, Wu X, Wolf SE, Chen Y. 2014. A novel significance score for gene selection and ranking. *Bioinformatics* **30**: 801-807.

Zarate C, Jr., Machado-Vieira R, Henter I, Ibrahim L, Diazgranados N, Salvadore G. 2010. Glutamatergic modulators: the future of treating mood disorders? *Harv Rev Psychiatry* **18**: 293-303.

Chapter 3

Evaluating Metagenomic Prediction of the Metaproteome in a 4.5 Year Study of a Crohn’s Patient

3.1 Abstract

Although genetic approaches are the standard in microbiome analysis, proteome-level information is largely absent. This discrepancy warrants a better understanding of the relationship between gene copy number and protein abundance, as this is crucial information for inferring protein level changes from metagenomic data. As it remains unknown how metaproteomic systems evolve during dynamic disease states, we leveraged a 4.5-year fecal time series of a single patient with colonic Crohn's disease. Utilizing multiplexed quantitative proteomics and shotgun metagenomic sequencing of eight time points in technical triplicate, we quantified over 29,000 protein groups and 110,000 genes and compare them to five protein biomarkers of disease activity. Broad-scale observations were consistent between data types, including overall clustering by Principal Coordinates Analysis and fluctuations in Gene Ontology terms related to Crohn's disease. Through linear regression, we determined genes and proteins fluctuating in conjunction with inflammatory metrics. We discovered conserved taxonomic differences relevant to Crohn's disease including a negative association of *Faecalibacterium* and a positive association of *Escherichia* to calprotectin. Despite concordant genera associations, the specific genes correlated with these metrics were drastically different between metagenomic and metaproteomic datasets. This resulted in the generation of unique functional interpretations dependent on the data type, with metaproteome evidence for previously investigated mechanisms of dysbiosis. One such example was a connection between urease enzymes, amino acid metabolism and the local inflammation state within the patient. This proof-of-concept approach prompts further

investigation of the metaproteome and its relations to the metagenome in biologically complex systems such as the microbiome.

3.2 Introduction

Due to the growing evidence for a connection between microbial communities and human health, exploration of the microbiome has rapidly expanded in the past decade. To date, the primary avenue for studying the microbiome has been through genomic technologies (Turnbaugh et al. 2007; Thompson et al. 2017; McDonald et al. 2018)(Turnbaugh et al. 2007; Thompson et al. 2017; McDonald et al. 2018)(McDonald et al., 2018; L. R. Thompson et al., 2017; Turnbaugh et al., 2007). These techniques help gain an understanding of who and how abundant the microbial constituents are and can define their associated metabolic potential. However, gene copy numbers are not representative of protein levels due to the complex systems governing when and how much of a given protein should be present (Pandey and Mann 2000). Further, RNA expression has been well documented to have limited correlation to protein abundance within many eukaryotes and bacteria (Liu et al. 2016). These relationships have not been thoroughly investigated in the context of the complex communities inhabiting the human gut microbiome, thus limiting the utility of DNA (or even RNA)-based analyses for understanding microbiome function.

Metaproteomics is an emerging technique that directly characterizes proteins from multi-species matrices. There has been over a decade of development of the field (Klaassens et al. 2007; Verberkmoes et al. 2009; Kolmeder and de Vos 2014; Zhang et al. 2017), though most studies have been limited in scope due in part to complex technical hurdles: lack of proteome coverage (Zhang et al. 2017), sample sizes typically below 20 samples (Kolmeder and de Vos 2014), limited reference database selection (Tanca et al., 2013; Tanca et al., 2016; Zhang et al., 2016), and peptide assignment to proteins of similar identity (Tanca et al. 2013). The introduction of new methods and mass spectrometers have dramatically increased the number of quantifiable peptides and proteins, allowing for a greater than 20 fold increased coverage of the metaproteome in the past few years (Zhang et al., 2017; Zhang et al., 2018). Here, we leverage Tandem Mass Tag (TMT) technology, allowing higher throughput by combining up to 11 samples within one mass spectrometry (MS) experiment, without the necessity of culturing (Thompson et al. 2003). In addition, TMT workflows utilize Synchronous Precursor Selection (SPS), LCMS^2/MS^3-based quantitation workflow to increase accuracy and reduce the sparsity associated with label-free proteomics (Ting et al. 2011). This combination has enabled unprecedented, deep characterization of proteomes at large scales (Lapek et al., 2017; Lapek et al., 2018; Weekes et al., 2014). In comparison to current metagenomic technology, the metaproteome field is still limited in depth of coverage and throughput. Nevertheless, performing direct protein-level analysis through advances in MS may allow for new insights into complex biological systems.

Here we utilized these technical advances to better understand the relationship between fluctuations in microbiome protein expression and fluctuations in microbiome gene content. Crohn’s disease (CD), a subtype of Inflammatory Bowel Disease (IBD), represents a chronic, autoimmune condition associated with large fluctuations in the microbiome (Gevers et al., 2014; Halfvarson et al., 2017; Manichanh, Borruel, Casellas, & Guarner, 2012; Walters, Xu, & Knight, 2014). A study in 2012 was the first to integrate the metagenome and metaproteome in the context of IBD (Erickson et al. 2012).

The results indicated that in six Crohn's patients, Ileal Crohn's Disease (ICD) had a unique metaproteome from colonic Crohn's disease (CCD) (Erickson et al. 2012). Subsequently, a meta-analysis of human single nucleotide polymorphisms from 30,000 IBD patients corroborated the split between ICD and CCD (Cleynen et al. 2016). While further metaproteome studies have been conducted on the human gut microbiome of IBD (Presley et al. 2012; Juste et al. 2014; Zhang et al. 2018), few integrate and compare results from metagenome and metaproteome data.

A distinguishable aspect of our study is a shift from contrasting IBD cohorts with healthy subjects to exploring a time series perspective from a single patient. Previous studies investigated metaproteome stability in the context of healthy subjects (Kolmeder et al. 2012; Kolmeder et al. 2016), however these studies were limited to time periods at or below one year. Herein, we tracked the disease activity of our patient through the abundances of several sub-components of the immune system, which form the basis of several clinical tests used to monitor IBD disease activity (Chang, Malter, & Hudesman, 2015; Iskandar & Ciorba, 2012; Mosli et al., 2015; Vermeire, Van Assche, & Rutgeerts, 2004). These proteins include C-Reactive Protein (CRP), lysozyme, Secretory Immunogloblin A (sIgA), calprotectin, and lactoferrin (Table 3.1). Our experimental design contains one patient and eight time points, with a focus on the comparisons between metagenomic and metaproteomic data. By tracking IBD episodic dynamics through the metagenome and metaproteome, we identified a set of bacterial taxa, and functional groups that are time-correlated with immunological biomarkers in our patient. Further we evaluate metagenomic prediction of the metaproteome, and identify unique aspects of function accessible through metaproteomics.

Table 3.1 Roles of immunological proteins of interest. IL-6, interleukin-6; TNF-α, tumor necrosis factor alpha.

Protein	Role
CRP	An acute phase response protein produced by the liver upon stimulation by IL-6, TNF-α and IL-1-β and a common clinical marker of general inflammation (Vermeire et al. 2004). It is found both in human blood serum and stool.
Lysozyme	A glycoside hydrolase used in the innate immune system for hydrolysis of Gram-positive bacterial cell walls (van der Sluys Veer et al. 1998). Measurements of lysozyme in the stool of patients with IBD have shown some correlation to disease activity in colonic IBD (van der Sluys Veer et al. 1998).
Secretory IgA	The most abundant antibody in the human colon and helps tightly control the relationship between commensal microbes and the host by delaying or abolishing the ability of microbes to adhere to the epithelium (Corthesy 2013).
Calprotectin	An antimicrobial protein that sequesters manganese to prevent the growth of pathogenic microbes that require these metals (Brophy and Nolan 2015). Consisting of two subunits, S100A8 and S100A9, calprotectin is an important molecule to the innate immune system constituting 40% of the cytoplasmic proteins in neutrophils (Brophy and Nolan 2015). Fecal calprotectin levels have been described as a stronger indicator of endoscopic activity than CRP, and has potential for identifying endoscopic remission (Chang et al. 2015; Mosli et al. 2015; Joy et al. 2017; Dai et al. 2018).
Lactoferrin	An antimicrobial glycoprotein, and a major component of the secondary granules of neutrophils (Dai et al. 2018). lactoferrin's antimicrobial properties are a result of iron sequestration, and has potential for both discriminatory and activity tests in the clinic (Mosli et al. 2015; Dai et al. 2018).

3.3 Results

Patient Information

The N=1 patient was a non-smoker male. He was diagnosed in 2011, at age 63, with CCD by Dr. William J. Sandborn at the University of California Health System. The inflamed region of the colon was determined, via colonoscopy and abdominal MRIs, to be confined to 6-8" of the sigmoid colon. Specifically, a 2012 colonoscopy revealed that this region had extensive diverticulosis and inflammatory focal ulceration, inflammatory pseudopolyps, and patchy friability not associated with the diverticular orifices. During the time interval covered in this work (12/28/2011 to 5/22/2016), the patient had one period of antibiotic therapy, ciprofloxacin 500 mg administered twice daily and metronidazole 250 mg administered three times daily for one month starting 1/31/2012. During that period, the patient was also taking 40 mg Prednisone daily. In another 4-month period from August through November, 2013, the patient had a simultaneous course of Lialda (anti-inflammatory) and Uceris (budesonide) administered at 9 mg daily. During the reported period, the patient had episodic symptoms of rectal bleeding, abdominal cramps, bloating, and malaise. Lastly, there was no surgery performed on the patient during the time period covered by this work.

Selection of Immunological Proteins of Interest

The immunological proteins, fecal C-Reactive Protein (CRP), lysozyme, sIgA, calprotectin, and lactoferrin were selected for their unique properties and clinical applications in IBD. We observed similar expression patterns over time for calprotectin, lactoferrin, and sIgA (Fig. 3.1a). Lactoferrin and sIgA abundances were the most strongly correlated to calprotectin (Pearson $r = 0.96, 0.50$ respectively), which led to overlapping results in downstream analysis. Because calprotectin is more widely used for the assessment of IBD (Chang et al. 2015), we focus primarily on relationships to calprotectin over those found with lactoferrin and sIgA.

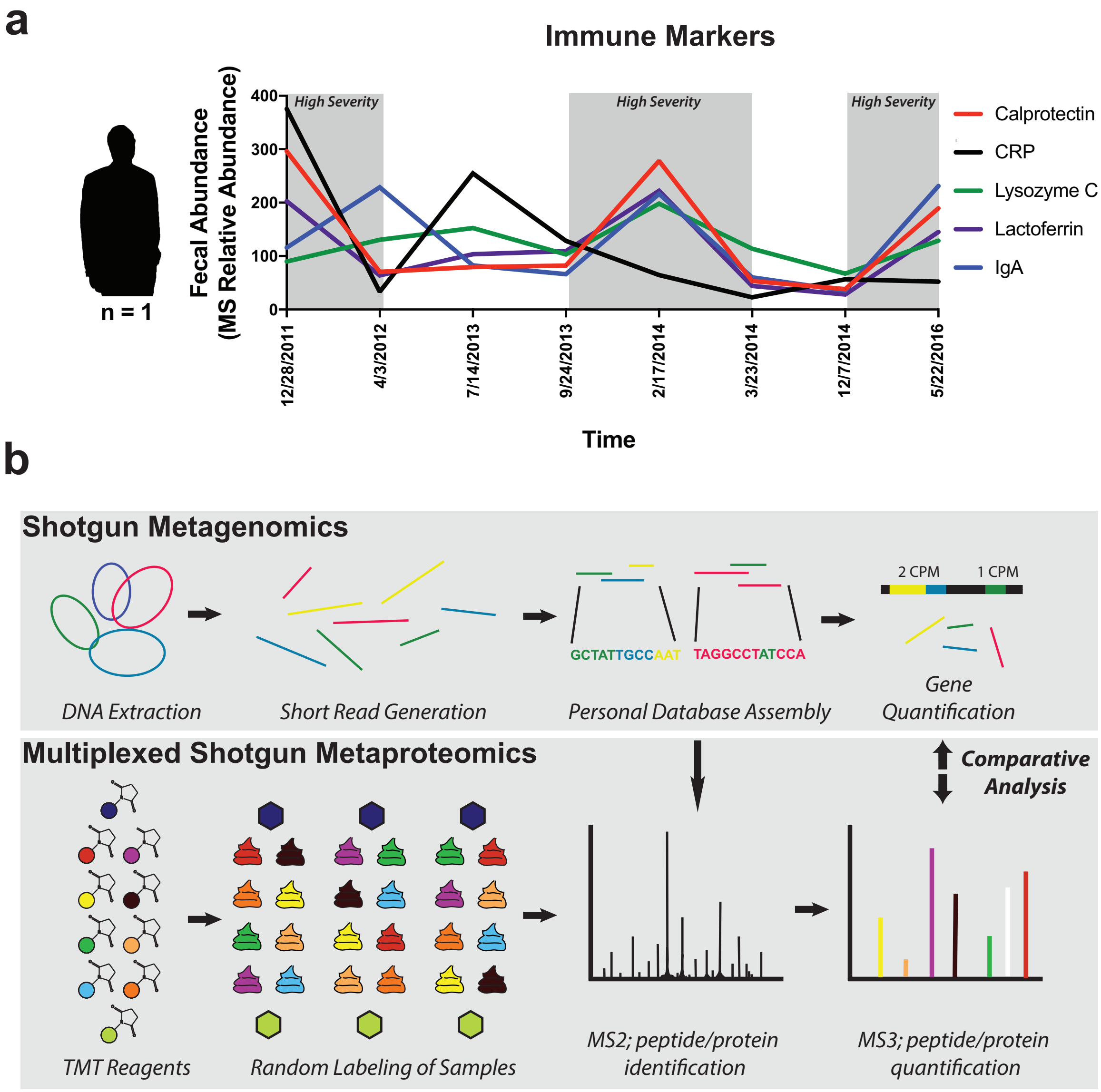


Figure 3.1 Study design. a, Immune markers associated with samples. Mass spectrometry based relative abundances of fecal calprotectin, CRP, Lysozyme, lactoferrin, and secretory IgA are plotted on the left y-axis for each of the eight time points in this study. **b**, Workflow schematic describing omic methods. Shotgun sequencing and metaproteomic methods were performed in parallel for the analysis of eight selected samples. Both methods were performed in technical triplicate for evaluation of technical variability. Tandem Mass Tag (TMT) labeling of tryptic peptides was performed for three mass spectrometry experiments. Green and purple hexagons represent composite samples used as controls, while other colors represent the random labeling of samples using the remaining TMT reagents. Shotgun sequencing reads were combined and assembled into a shared reference database (Personal Database Assembly) for assigning gene counts (in counts per million, CPM) and protein abundances. Not shown is MS1, which was used for precursor selection.

Technical comparisons between –omic types and protein database methodology

As discussed above, eight fecal samples from our patient were collected over a time period from 2011 to 2016 representing a wide range of disease activity. Samples were processed in technical triplicate through shotgun metagenomic sequencing and a proteomic workflow using TMT mediated liquid chromatography triple-stage MS (LC-MS^3) (Fig. 3.1b).

To address the lack of standardized database methodology (Tanca et al., 2013; Tanca et al., 2016), two different protein reference database approaches were used for analysis of LC-MS^3 data. Our first approach utilized the shotgun metagenomic reads generated within the study to create a personalized database (pDB) containing 1.3 million protein coding regions (Erickson et al. 2012). Through alignment of our protein coding regions to taxonomic and functional databases, the pDB provided genera level annotations for 80% of genes, and functional annotations to KEGG Orthologous (KO) groups in 15% of genes. The pDB approach was crucial for comparison between metagenomic and metaproteomic data types as it allowed for a shared reference for gene and protein abundances. For comparison, we separately performed a two-step search method (Zhang et al. 2016) of the MS data using a public database of gut microbial genes (Integrated Gene Catalog, IGC) (Li et al. 2014). Our methods resulted in 123,806 predicted open reading frames (ORFs) from the pDB with DNA quantification and 29,370 with protein quantification. A search through both databases yielded a similar number of peptides and proteins with 113,373 total unique peptides and 72.5% of peptides shared between pDB and IGC database methodology. The degree of overlap in peptides was consistent with previous findings (Zhang et al. 2016).

Notably, a lack of shared sequences between samples is a known trait of microbiome studies (Tsilimigras and Fodor 2016). We observed that the TMT-based metaproteomic methods provided quantification measurements within all samples for a larger percentage of proteins (52% of proteins identified from the pDB, 65% of proteins identified from the IGC) than the metagenomic techniques provided for gene quantifications (4%). This increased overlap is likely a result of TMT multiplexing methods, which are known to reduce sparsity when compared to label-free MS (Rosa Viner 2013). Our methods also allowed for in parallel quantification of nearly 1000 human proteins. Human protein quantification is an important advantage of metaproteomics, especially in light of recent results showing the ability of human proteins to distinguish IBD patients from controls (Zhang et al. 2018). It is important to note that the database used for protein assignment can result in different functional annotations. For example we observed that the IGC approach identified 83% more unique KEGG Orthologous (KO) groups than the pDB approach. This discrepancy in peptide matching is an ongoing area of investigation in computational biology (Tanca et al. 2013; Tanca et al. 2016; Zhang et al. 2016; Xiao et al. 2018).

The technical and biological variability within each dataset was assessed through Principal Coordinates Analysis (PCoA) using the Bray-Curtis distance metric (Caporaso et al. 2010). To overcome structural artifacts from the missing values within TMT experiments, only the proteins common to all samples were used in this analysis. After

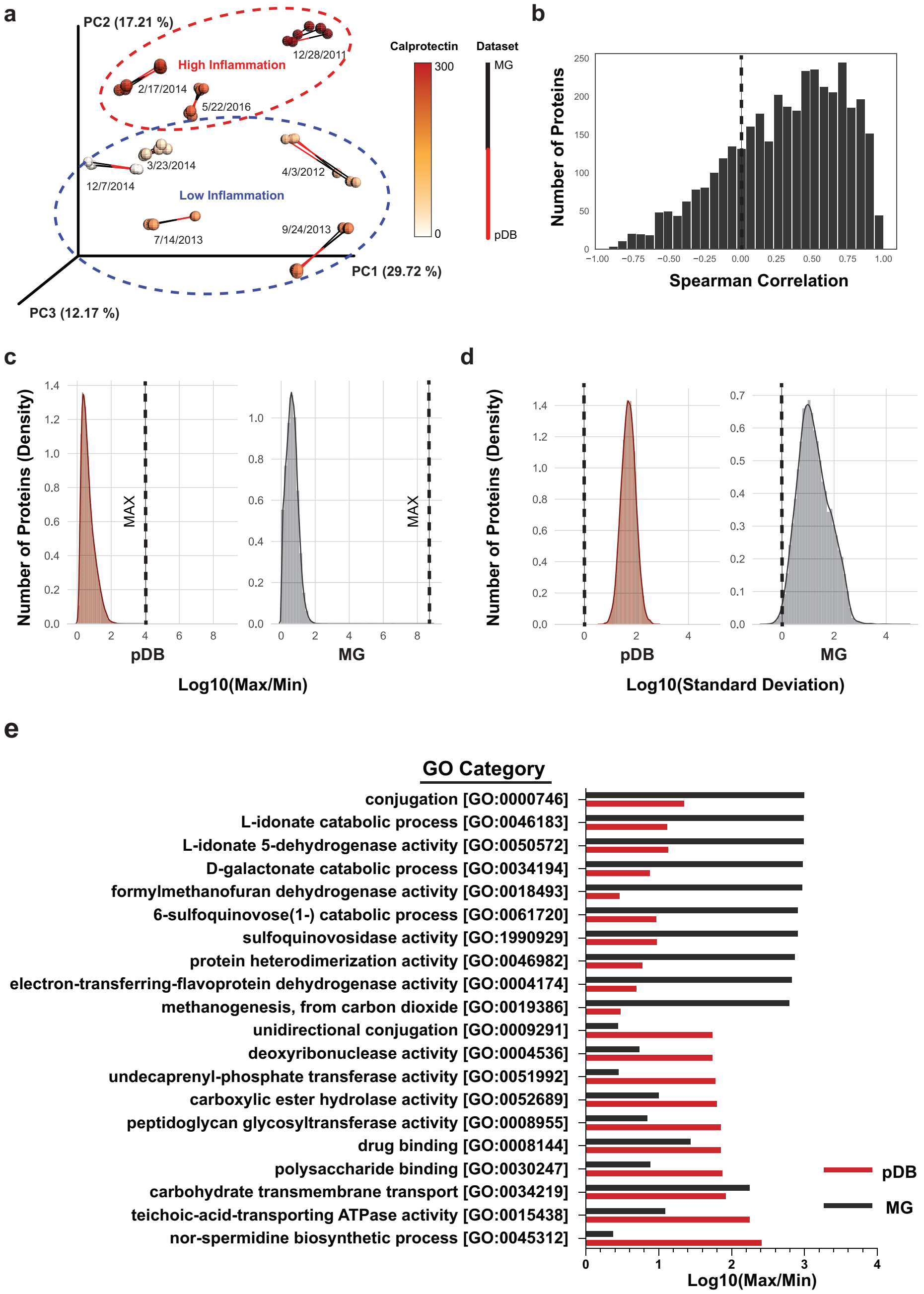


Figure 3.2 Broad-scale data type comparisons. a, Procrustes analysis comparing clustering of the metaproteome and metagenome. Bray-Curtis distance metric was used on both the metagenome and metaproteome (only proteins common to all samples, pDB database) to assess technical and biological variability within and between datasets. Samples are colored by calprotectin relative abundances. **b,** Distribution of Spearman correlations comparing metagenomic and metaproteomic fluctuations. The x-axis displays the Spearman correlation (ρ) and the y-axis displays the number of gene-protein pair within a range of Spearman correlation values. **c,** Dynamic range comparison. Histograms fitted with a Gaussian kernel density estimate are displayed at the gene and protein level. The Log10 of the maximum value for each protein or gene divided by the minimum value is plotted on the x-axis. The number of proteins corresponding to each max/min range are plotted on the y-axis. **d,** Variability comparison. As described in (c) but according to the standard deviation of each gene or protein. **e,** GO categories with largest fluctuations. Proteins and genes were summed according to their GO categories and the maximum to minimum were compared. The highest metagenomic fluctuations are recorded on the top and the highest metaproteomic fluctuations are displayed on the bottom.

this adjustment, a comparison between our datasets was performed using Procrustes analysis and a Mantel test (Fig. 3.2a). The Procrustes analysis transforms two distance matrices from corresponding samples to compare distributions. These tests showed minimal technical variability and a strong association between the two data types (Mantel test $p < 0.001$). We also observed clustering based on high or low inflammation state (Fig. 3.2a). Group differences between high and low inflammation state were not statistically significant, likely a result of the small number of samples analyzed. Though not significant, the metaproteome showed a stronger association to inflammation state than the metagenome (Pseudo-F = 1.54 for metaproteome, Pseudo-F = 1.19 for metagenome).

To investigate the relationship between gene and protein level fluctuations, the data was subset to the 3598 ORFs with both copy number and protein abundance data. Spearman correlations were assessed between the protein and gene abundances in each of the samples. Overall, the Spearman correlations were normally distributed around $\rho = 0.317$ (Fig. 3.2b). This limited correlation highlights the added value a metaproteomic approach can have in cases such as CD, where disease severity is associated with fluctuations in the microbiome (Gevers et al. 2014). We next investigated data type comparisons from a functional perspective by summing abundances by Gene Ontology (GO) and KO annotations and performing Spearman correlations between the genes and protein abundances. This analysis resulted in an approximately normal distribution near $\rho = 0.140$ for both annotation types. These weak correlations might have been expected given that our approach was based on comparing DNA to protein, as even RNA abundances are often weakly correlated to protein abundance (Maier et al. 2009).

We further investigated data type differences by comparing the distribution of dynamic ranges and standard deviations. Ratios of maxima to minima showed that both data types demonstrate a normal distribution centered around 4.4 for proteins and 11 for gene copy numbers (Fig. 3.2c). The maximum to minimum ratios reached up to 9,400 for proteins and 129 million for gene copy number (Fig. 3.2c), indicating a much greater dynamic range in the latter. These dynamic ranges may indicate the extent to which microbial genes and proteins can change over time within an individual. However, this result may be influenced by the differences in depth of coverage, in which the metagenome is approaching more complete coverage than the metaproteome, and the less abundant genes only detected by the metagenomic methods may have a greater dynamic range. The standard deviations of the genes and proteins were normally distributed but displayed differences in averages and variances (Fig. 3.2d). The metagenome had larger variance in the distribution of standard deviations, potentially indicating more variability within that platform (variance = 0.36, 0.074 for MG and pDB). Still, this result may also be influenced by the differences in the depth of coverage. Distribution of the maxima to minima for GO and KO sums shared similar distributions between data types. The largest fluctuations in GO terms were greater than 100-fold for proteins and 1000-fold for genes (Fig. 3.2e). Large changes were observed in categories of interest such as drug binding for proteins and methanogenesis (Scanlan et al. 2008) for genes. This was likely the result of the presence then absence of two archaeal methanogens, *Methanobrevibacter smithii* and *Methanosphaera stadtmanae* (Gaci et al. 2014)*,* whose genes were on average, 15 times more abundant in the first collected time point (12/28/2011) than any other sample.

These results give some indication of the fundamental dynamics of genes and proteins, but are surely influenced by the techniques used in the study design.

Copy number prediction of protein abundances by functional categories

Because proteins have consistent roles (Toyama and Hetzer 2013), we expected that certain functional categories would have a stronger correlation between gene content and protein expression. We tested this hypothesis using several different functional databases for a comprehensive analysis. After removing human proteins and subdividing individual genes by Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (eggNOG) functional category, the distribution of gene to protein Spearman correlations was largely consistent with the overall mean ρ ~ 0.3 (Fig. 3.3a). Categories with the largest number of features shared, such as “Energy production and conversion”, “Carbohydrate transport and metabolism” and “Translation, ribosomal structure and biogenesis” all had distributions centered near Spearman ρ ~ 0.3. Other categories with fewer features had more variability in their average correlation values. Less abundant categories included “Cell cycle control”, which had a lower average correlation and “Inorganic ion transport and metabolism” which had a higher average correlation (Fig. 3.3a). This indicates that there were not broad-scale functional group differences distinguishable from the overall low, but positive correlation observed between all genes and proteins.

In addition to individual gene correlations, we also evaluated inter-omic relationships between the abundances of entire gene categories. We assessed these relationships through summing protein and gene abundances by GO annotation and performing Spearman correlations. There was large variability (σ = 0.445) in the

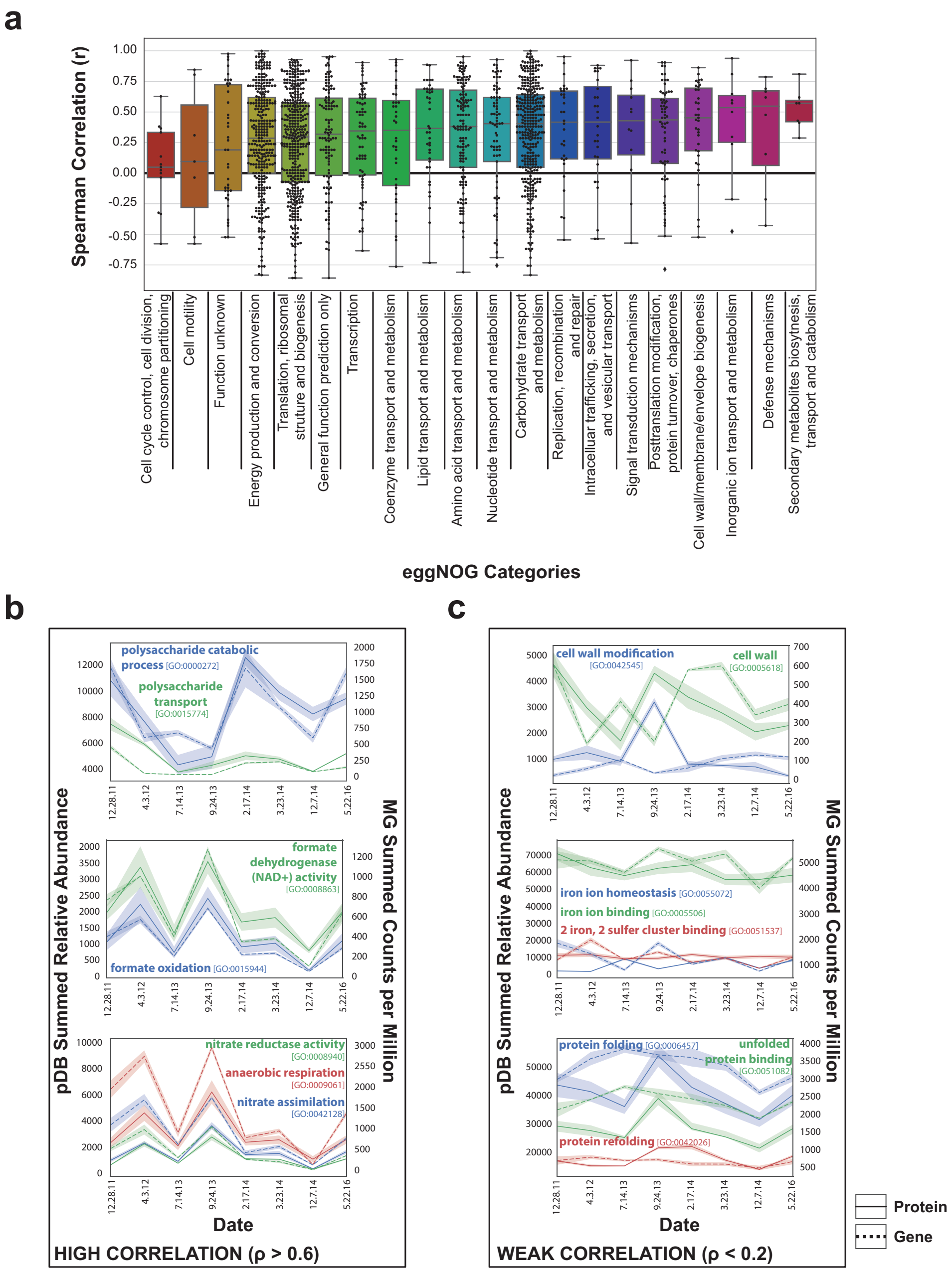


Figure 3.3 Functional categories with strong or weak genomic prediction of proteome fluctuation. a, Box plots demonstrating the distribution of Spearman correlations per gene which have an associated eggNOG functional category. Spearman correlation (ρ) between the summed metagenomic CPM per time point with the average relative abundance of associated metaproteomic protein is displayed. Summary statistics for this data can be found in Supplementary Table 1. **b,** Summed GO categories with strong genomic and proteomic correlation. **c,** Summed GO categories with weak genomic and proteomic correlation.

correlations of different functional groupings with an average Spearman ρ = 0.135. Despite the low overall correlation, themes of GO categories with similar correlations were present. Several GO terms related to polysaccharide, formate, and anaerobic respiration all had strong positive correlations above $\rho = 0.6$ (Fig. 3.3b). Other categories had consistently low, or even negative, correlations below $\rho = 0.2$. Cell wall and membrane proteins, metal binding proteins and chaperones were among the categories with poor correlations (Fig. 3.3c). These results suggest that there are some categories of genes that better represent protein expression levels, which may be the result of constitutive versus inducible expression. However, the techniques used also influence particular categories, such as membrane proteins, whose hydrophobic nature presents a challenge to MS workflows (Chandramouli and Qian 2009). All of the described categories had greater than 200 proteins and genes contributing to these relationships, which indicates that this was not related to differences based on high or low abundance proteins.

Taxonomic correlations with inflammatory markers are largely shared at the protein and gene level

We next sought to determine whether fluctuations related to inflammatory markers were conserved between genes and proteins. Taxonomic assignments for the pDB database were assigned based on the protein sequences to ensure consistent assignments for both datasets. Genus level compositions were significantly different in the metagenome but not in the metaproteome (Friedman test p = 8.9e-5 and 0.69 respectively) (Fig. 3.4a, Fig. 3.4b). Dominant genera included *Escherichia*, *Bacteroides*, *Faecalibacterium* and *Alistipes* (Fig. 3.4a). The metaproteome composition was intentionally not adjusted for lowest common ancestor of peptides (Mesuere et al. 2015) for easier interpretation of the abundances used for metagenome comparisons. Metaproteome taxonomic composition plots adjusted for lowest common ancestor also

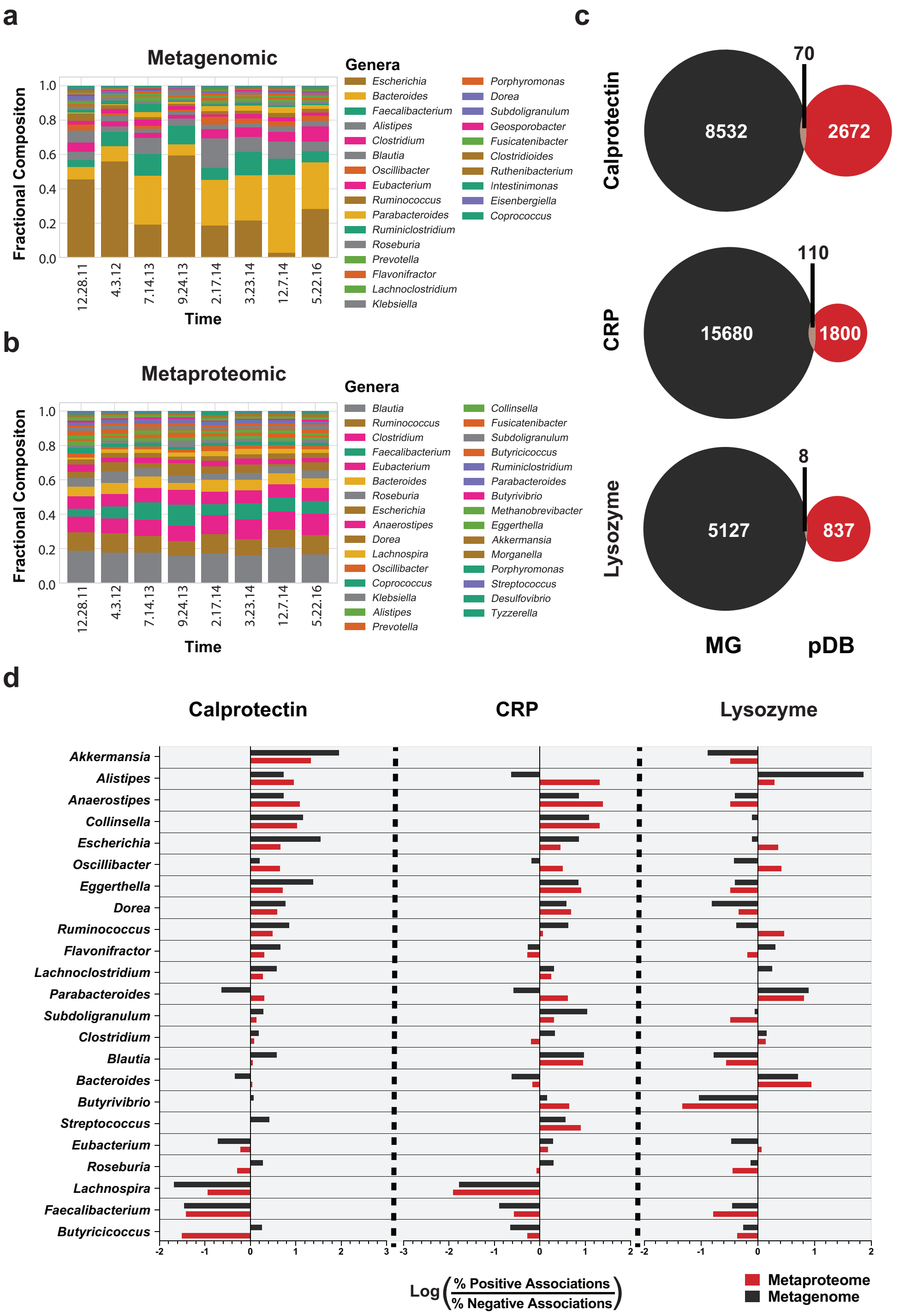


Figure 3.4 Genus level associations to clinical markers. a, Genera level barplot displaying the fractional composition of the most abundant genera (> 0.03) in the metagenome and (**b**) the metaproteome in each of the samples analyzed. **c,** Comparison of genes and proteins significantly associated to each clinical marker. Venn diagrams are displayed showing the number of genes and proteins with a large effect size (|*r*| > 0.7) to clinical markers based on linear regression. **d,** Genera associated with clinical markers. The associated proteins with genus level taxonomy from (c) were compared by the log ratio of the composition of positive and negative proteins. The log ratio is plotted on the x-axis for each clinical marker and bars represent the association of each genus. Metaproteome values are plotted in red and metagenome values are plotted in black.

displayed stable compositions, though certain genera, such as Blautia, had a notably different composition after the adjustment.

To evaluate the relationship between species related to inflammation in CD and our biomarkers of interest, we evaluated each immune protein against a previously defined microbial dysbiosis index (Gevers et al. 2014). This index was developed using hundreds of samples from both Crohn's patients and healthy controls to predict CD severity through log ratios of the species increased and decreased within CD (Gevers et al. 2014). Nineteen of the species defined in the index were found in our dataset. These included *Escherichia coli* and *Fusobacterium nucleatum,* which are increased in CD, and *Faecalibacterium prausnitzii, Eubacterium ractale,* and *Bacteroides vulgatus*, which are decreased in CD. After summing gene and protein abundances and determining the relationship between log ratios and each biomarker, fecal calprotectin had the strongest association with the microbial dysbiosis index in both the metagenome and metaproteome. This result was not statistically significant, which was likely a result of either the small sample size, or extrapolating methods developed from hundreds of patients onto a single subject.

Linear regressions against inflammatory markers were performed on each gene and protein. To evaluate our results, we compared the positively and negatively associated genes with large effect sizes (Cohen 1988) (correlation coefficient, $|r| > 0.7$). Interestingly, the individual genes and proteins associated with each of the inflammatory markers were largely unique with only 0.5% (188/34,836) of associations shared between data types (Fig. 3.4c). When accounting for only the genes and proteins quantified in both datasets, 10% (188/1,814) of the strong associations were shared between datasets.

Despite the lack of overlap in the individual identities of the genes and proteins correlated with each clinical marker, we observed consistent trends in the taxonomic annotations among the correlated genes and proteins. With over 800 genes and proteins strongly correlated to each marker ($|r| > 0.7$), we contrasted the taxonomic composition of the positive and negative correlations. Several genera had >30-fold difference between compositions (Fig. 3.4d). Genera-level differences were largely conserved between data types in both direction and magnitude of association (Fig. 3.4d). *Akkermansia* and *Anaerostipes* had the strongest pro-inflammatory relationship, while *Faecalibacterium* and *Butyricicoccus* had the largest anti-inflammatory relationship as assessed through the number of proteins positively or negatively correlated to calprotectin (Fig. 3.4d). Several genus level trends were conserved between CRP and calprotectin such as *Alistipes*, *Anaerostipes*, *Faecalibacterium* and *Lachnospira*, while lysozyme had largely different associated genera. Contextually, the number of proteins and genes used to generate these associations is important for the interpretation of these results as some associations were based on very few observations.

Lysozyme is a component of the innate immune response that targets Gram-positive cell walls. Interestingly, proteins and genes correlated with lysozyme levels had large phylum-level changes. Bacteroidetes is a Gram-negative phylum, while Firmicutes is largely Gram-positive (Winter et al. 2013). The Gram-positive Firmicutes were 1.4 fold enriched among negative associations to lysozyme in both gene and proteins, while the Gram-negative Bacteroidetes were 4.3-fold and 8.9-fold enriched among positively correlated proteins and genes, respectively. Even though there were more than 800 genes and proteins from Firmicutes and Bacteroidetes that were correlated to lysozyme, very few from other phyla, such as the Gram-negative Proteobacteria, and Gram-positive Actinobacteria were observed. To validate these observations at the genus level, Gram-stain information was cross referenced (Markowitz et al. 2012). Although there were genera with both Gram-negative and Gram-positive species, the genus level associations to lysozyme largely reflected the phylum level observations.

Comparing functional interpretations of the genes and proteins associated with immunological biomarkers

Using the same identifications from linear regressions that provided the genus level results, we next compared broad-scale functional groupings. The broad scale functional associations were weaker in comparison to the genus associations. This observation may represent a broad versus fine scale categorization. Illustrating this point, the largest difference among the genera associations was 90 fold, while the largest difference between functional groupings using assignments to the eggNOG database was 12 fold (Fig. 3.4d, Fig. 3.5a). Analyzing a broader taxonomic category, we observed that the maximum difference among phyla was 8.9 fold, considerably closer to the 12 fold maximum for eggNOG categories. An additional consideration for this result is the annotation rate for functional assignments. Only 15% of observed ORFs had an identifiable function and this lower annotation rate may bias the results.

Despite the weaker associations of functional categories, several functional relationships to the disease markers were of interest. In total, 19 eggNOG (12 from the metaproteome, 7 from the metagenome) had differences of 3 fold or greater (Fig. 3.5a). Comparing the categories with associations to different immune markers provided insight

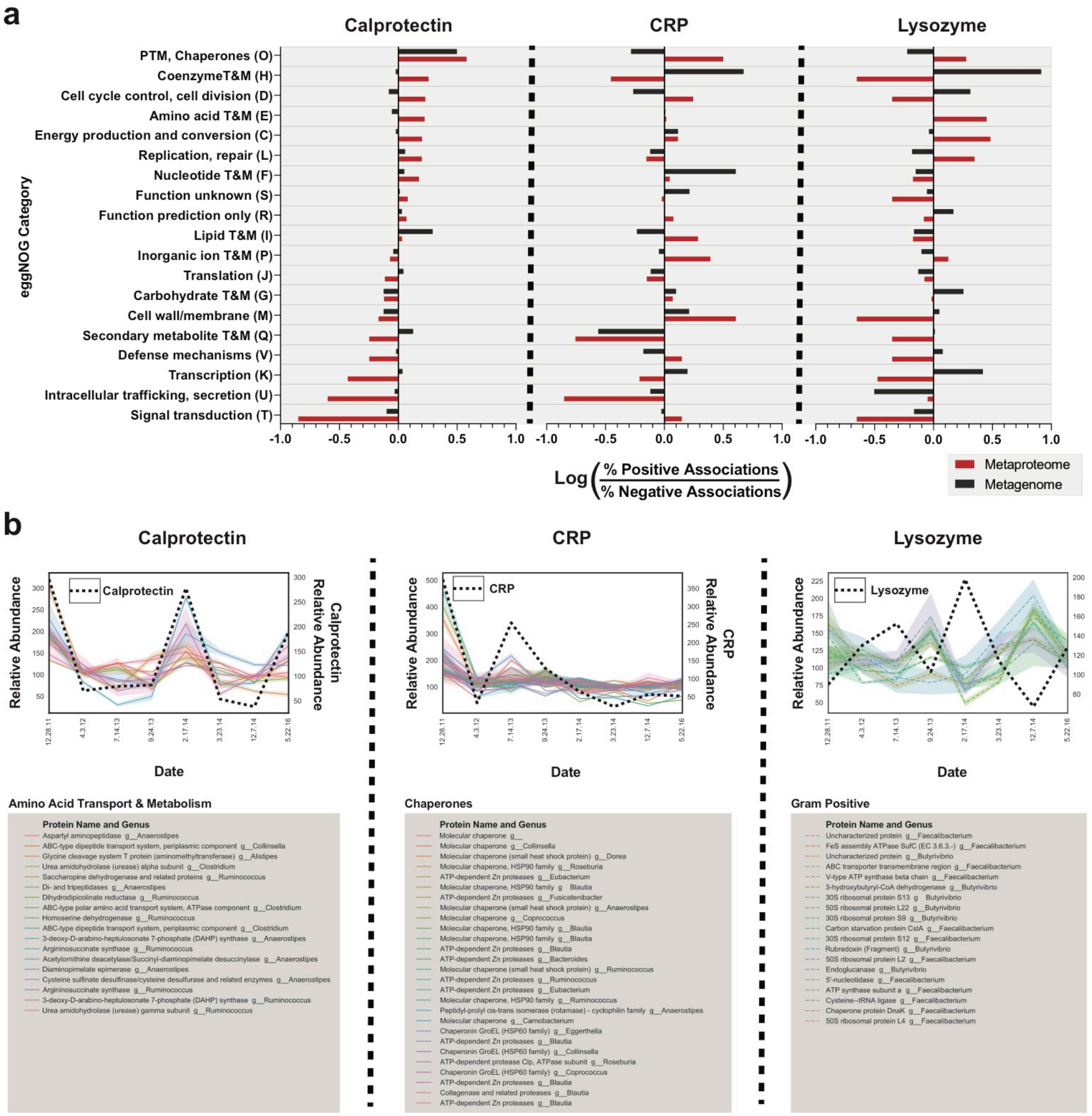


Figure 3.5 Functional associations to clinical markers. a, Functions associated with clinical markers. Linear regressions to clinical markers were performed and the number of proteins or genes derived from each functional group with a large effect size (|*r*| > 0.7) were compared**.** The log ratio of the composition of positive and negative proteins is plotted on the x-axis for each clinical marker. Metaproteome values are plotted in red and metagenome values are plotted in black. **b,** Time series plots of selected proteins of interest. Protein abundances of one finding from each clinical marker are shown. A legend describing the protein names and associated genus is shown below each graph.

into how different data types might influence functional interpretation. For example, metagenomic data had several strong functional associations that were not confirmed by protein abundances. One such category, “Nucleotide transport and metabolism”, had 147 genes positively correlated with CRP, and 0 negatively correlated genes, indicating a positive association to CRP. The metaproteome data for this category had almost no association to CRP (Fig. 3.5a), with 6 proteins negatively correlated and 38 proteins positively correlated. We suspect that nucleotide metabolism undergoes protein expression independent of inflammatory conditions. The underlying reasons for this observation need to be further investigated.

Biologically relevant relationships were observed in the metaproteome that were not detectable in the metagenome. Free amino acids and urease enzymes have previously been associated with gut dysbiosis and Crohn's disease (Ni et al. 2017). Interestingly, the metaproteome data identified a functional association of amino acid metabolism proteins to calprotectin, while this observation was absent in the metagenomic data (Fig. 3.5a). This observation included several urease proteins, and transporters for free amino acids, many of which were derived from the genera that had positive associations with inflammation (Fig. 3.5b). These ureases and transporters thus represent interesting targets for further investigation, and provide further evidence of a previously established connection (Ni et al. 2017).

Another observation that was exclusively related to the metaproteome data was the relationship of chaperone proteins to several of the inflammatory metrics. There were 15 chaperone proteins with similar trends in expression to CRP (Fig. 3.5b). This corresponded to post-translational modification and chaperone proteins having a 3.2 fold higher representation in positively associated proteins, and a 1.9 fold lower representation in genes (Fig. 3.5a). This unique observation from our patient's fecal metaproteome is a potential indication of microbial stress occurring in response to the acute phase response, and may indicate a need for the microbiome to refold proteins.

Because lysozyme targets Gram-positive cell walls, we expected correlated genes and proteins to be influenced by taxonomy and to have functions related to cell walls or membranes. However, cell wall proteins were under-represented in the metaproteomic dataset relative to their occurrence in the metagenomic dataset. Of the cell wall proteins associated with lysozyme, two were related to cell wall biosynthesis (COG1088, COG0463), a Glycosyl transferase and a dTDP-glucose 4-6-dehydratase. In this case, the binding of lysozyme to peptidoglycan may disrupt the binding of these cell wall/membrane/envelope biogenesis proteins leading to the observed negative correlation. Even though we were not able to detect many membrane or cell wall proteins related to lysozyme, 15 negatively correlated proteins from the butyrate producing (Sitkin and Pokrotnieks 2018), Gram-positive genera *Faecalibacterium* and *Butyrivibrio* were identified (Fig. 3.5c). These proteins included 5 ribosomal proteins, which may indicate decreased translation occurring in the presence of lysozyme.

In addition to analyzing calprotectin, CRP and lysozyme levels, we also evaluated sIgA and lactoferrin. Secretory IgA is secreted in large quantities in the intestine for maintaining favorable microbial compositions (Corthesy 2013) and lactoferrin sequesters iron as an antimicrobial response (Dai et al. 2018). We observed similar expression patterns of lactoferrin, sIgA and calprotectin. The similar expression resulted in minimal differences in both genus and functional relationships between calprotectin, lactoferrin and sIgA. Proteins positively associated with lactoferrin ($|r| > 0.7$) had a larger portion of GO terms related to iron (15.5% of 470 positive associations, 10% of 233 negative associations). Many of these proteins were pyruvate oxidoreductases, which are used in anaerobic bacteria for forming acetyl CoA from pyruvate (Chabriere et al. 1999). These are crucial enzymes for certain anaerobic bacteria, and have been suggested as potential drug targets (Chabriere et al. 1999). This result suggests that a connection exist between iron sequestering host proteins and microbial proteins in our patient that are dependent on iron as a cofactor.

3.4 Discussion

Our investigation of the fundamental relationship between changes in the metagenome and the metaproteome reveal important considerations for interpreting these data types. Currently, studies using shotgun metagenomics to dissect the functions of the microbiome are becoming more prevalent (Quince et al. 2017), and this study shows that differences at the gene level may not reflect differences at the protein level. Though discordance between RNA and protein expression is widely acknowledged for individual species (Pandey and Mann 2000), the relationships between DNA and protein content in the complex ecology of the microbiome is less understood. As these systems have rarely been studied in parallel, it is possible that communities of microbes influence fundamental relationships between genes and proteins that have been previously established in monoculture settings. Although the metaproteomics field is improving in depth of coverage (Zhang et al. 2017) and scope (Zhang et al. 2018), the technical hurdles MS presents often make DNA based studies a more practical, higher throughput solution. That being the case, functional insight from metagenomic studies requires a consideration of the relationship between protein abundances and metagenomic copy numbers.

Our results, although limited to a single patient, suggest that there is a degree of general agreement between changes in the metagenome and changes in the metaproteome. However, the relationship is overall weak for individual genes/proteins (our average Spearman ρ = 0.3). In single species context, bacterial systems have generally shown correlations between mRNA and proteins to range from ρ = 0.5 – 0.6 (Maier et al. 2009). Our experimental estimates place DNA to protein correlations in complex microbial systems to be notably lower. These associations do not appear to have obvious biases between large-scale functional groupings, but have certain trends in finer-resolution functional groupings such as individual GO terms. An important notion in the field of IBD, formate- and nitrate-related categories had large fluctuations and consistent trends between the two data types. Formate oxidation has been implicated as a metabolic signature of inflammation-associated dysbiosis (Hughes et al. 2017), indicating that metagenomic studies may predict protein abundances within this system. We do not expect the consistency between formate oxidation genes and proteins is a result of constitutive expression as, at least in *E. coli*, related genes such as formate hydrogenases are regulated by the presence of formate (Bohm et al. 1990). Nitrate based anaerobic respiration is implicated in promoting the growth of facultative anaerobes such as the Enterobacteriaceae, which can lead to microbial dysbiosis and intestinal inflammation (Winter et al. 2013). Tables of the identified eggNOG and GO terms are provided, which describe how well metagenomic copy number predicted protein abundances within each identified category.

Identifying the genes and proteins with similar expression trends to certain inflammatory and immune markers revealed that there were large differences in genus level associations that were biologically relevant and generally consistent between data types. *Faecalibacterium* is a genus depleted in IBD (Gevers et al., 2014; Takahashi et al., 2016), which appears to have anti-inflammatory effects, possibly mediated by butyrate production (Sokol et al. 2008). Both data types had a strong negative correlation in numerous *Faecalibacterium* proteins to our biomarker for local inflammation, calprotectin. While it was previously shown that there were consistent trends between these data types showing increased *Faecalibacterium* in healthy patients (Erickson et al. 2012), our results show these relationships can occur within a patient through time corresponding to the current level of inflammation. Other trends were also found for well-documented genera with inflammatory roles in IBD (Gevers et al. 2014), including *E. coli*, which is of particular interest because of its adherent-invasive properties in CD (Barnich & Darfeuille-Michaud, 2007; Palmela et al., 2018). Interestingly, these shared trends were found from almost entirely different genes. This may potentially indicate that the underlying bacterial abundance influences both of these data types, while the individual proteins expressed at certain times are not directly associated with the amount of corresponding genetic material present. If this is the case, it is possible that functional associations made through some broad-scale categories, such as eggNOG, may have different results depending on the data type. This concept is supported by our results that indicate smaller and less consistent associations to broad-scale groupings than associations at the genera level.

Our analysis of clinical biomarkers was useful for understanding the biology associated with each immune component. As calprotectin had the strongest association to the microbial dysbiosis index (Gevers et al. 2014), it suggests that calprotectin may be a better indication of microbial imbalances. Interestingly, CRP has been reported to be a less useful diagnostic than fecal calprotectin for intestinal inflammation (Chang et al.

2015). CRP levels may be a better indication of systemic inflammation, and here we have observed that many bacterial chaperone proteins may be increased in correspondence. With the abundances of lysozyme, we observed taxonomic trends consistent with its biological function of acting upon cell walls. In general, predominately Gram-positive genera and phyla had a larger portion of anti-correlated genes and proteins, while Gram-negative bacteria had an opposite association.

Our observed discrepancies between gene and protein levels may have large implications for data interpretation, but it is important to replicate these results in a larger cohort of IBD patients. As certain GO categories present strong correlation between data types, it suggests that it may be possible to develop a metagenomic-metaproteomic reference guide for creating stronger functional hypotheses. This guide may be used to outline which groups of genes have strong or weak association to protein abundances.

The relationship between genes and proteins may be influenced by several factors. Correlation between DNA and protein abundances might be reflecting DNA from dormant or dead cells (Jansson and Baker 2016), which may lead to a higher correlation because the cells are not actively producing or secreting proteins. Other factors may include constitutive versus inducible genes or the stability of the proteins. For example, chaperone proteins were found in high abundance which may be a result of their high stability, and stable concentrations within the cell (Henderson et al. 2006). Ultimately, the associations between -omic datasets are influenced by the nature of the data collection techniques and normalization, and further benchmarking is necessary. Although, there are significant challenges in integrating multi-omic data types (Palsson and Zengler 2010), further understanding these relationships is of paramount importance as the microbiome field progresses.

Our study presents several technical findings of interest. Leverage of the modern TMT based LC-MS^3 quantification platform provided a highly accurate quantification method for comparison with gene counts. Our workflow for mediating comparisons between metagenomic and metaproteomic data expands upon our knowledge of data type differences and acts as a bioinformatic and technological update to previous studies (Erickson et al. 2012). Additionally, the use of technical triplicates validates the reproducibility of these methods and helped increase our confidence in the quantification values at both the metagenomic and metaproteomic level. However, outside validation from other technological pipelines may be necessary to further understand these biological systems. Our results are also derived from a small number of samples from one patient, and the time points are spread out over large time spans. This design provided unique opportunities, but limit our interpretation of the data to a single individual.

From a biological perspective, our results provide evidence that certain proteins and genera are correlated or anti-correlated with immunoprotein markers of inflammation. While the taxonomic insights we observed were conserved between data types, our functional interpretations differed. This personalized perspective also demonstrates the extent of variability occurring within an individual, an important consideration to control for in studies with larger cohorts. In total, our study investigates the relationships between metagenomic and metaproteomic methods and highlights important considerations for interpretation of meta –omic data.

3.5 Methods

Longitudinal sample collection

Naturally passed fecal samples were collected and immediately stored without buffer at -80 °C. Eight samples were selected. A personal symptom log was generated at the time each fecal sample was passed. Additionally, the weight and BMI of the patient was determined on the day associated with each sample.

Generation of metagenomic reads

Samples were extracted according to the Earth Microbiome Project (Thompson et al. 2017) protocol using the QIAGEN MagAttract PowerSoil DNA Kit as previously described (Marotz et al. 2017). Briefly, swabbed fecal material was plated into 96-well PowerBead® DNA Plates containing garnet beads. DNA extraction was performed once on each of the eight samples according to the manufacturer's instructions, with an additional incubation at 65°C for 10 minutes following the addition of lysis solution and immediately prior to shaking (QIAGEN® TissueLyser® II; QIAGEN® catalogue: 85300). Magnetic DNA purification was performed using the KingFisher™ Flex™ Purification System. Then, whole-genome shotgun libraries were made using the Nextera DNA Library Prep Kit (Illumina, San Diego, CA, USA), at a 1:10 miniaturized reaction volume. Unique barcodes were used per triplicate totaling 24 metagenomic samples. Median insert size by sample ranged from 183 bp to 366 bp. Libraries were sequenced using Illumina MiSeq paired-end (2 × 250 bp) sequencing, filling a total of one lane.

Processing of metagenomic reads for a shared reference library (pDB)

Because typical metagenomics and metaproteomics workflows require a reference database, it was necessary to create from scratch using a minimal approach a single reference database that could be used for both metagenomics and metaproteomics from the individualized data. All reads from the technical triplicates of each sample were concatenated. Next the MEGAHIT alignment program (Hyatt et al. 2010) was utilized for assembling short reads into larger contigs. Assembled contigs were searched for possible coding regions through the program Prodigal (Hyatt et al. 2010). Next, the program Diamond (Buchfink et al. 2015) was used for gene alignment to the uniref50 database (Suzek et al. 2015). Finally, the most likely uniref50 entry, determined through bitScore, was used for the functional annotations. KEGG orthology annotations were cross-referenced using GhostKOALA (Kanehisa et al. 2016). Taxonomic assignments were determined by Diamond alignment (Buchfink et al. 2015) to an in-house library of microbial genomes. Taxonomy was assigned from the amino acid translated sequence of each predicted ORF in the pDB. This database was used as a reference database for both mass spectrometry data and sequencing data. Scripts used for data processing are available online (https://github.com/knightlab-analyses/Crohns-MG-MP-Comparisons).

Generating copy numbers of metagenomic genes

The program Salmon (Patro et al. 2017) was applied to determine the reads present for each gene from the pDB. First, an index was created with Salmon inputting the pDB fasta file. Next, reads were aligned to this index in quasi-mapping mode for each of the 24 metagenomic samples. The results were represented in counts per million sequences, with missing values padded as zeroes.

Protein abundances from the shared reference library (pDB)

The generation of mass spectra data is described below. Spectral data was searched against the pDB with a concatenated human reference library (uniprot.org, accessed 11-28-16) using Proteome Discoverer 2.1 (Thermo Fisher Scientific). Further data processing is described below.

Protein digestion and TMT labeling

Fecal samples were measured out to ~0.5 g and suspended in 10 mL of ice-cold, sterilized TBS. Samples were suspended through vortexing and homogenized through a blender apparatus. A 20 μM vacuum, steriflip (Milipore) filter was used to remove particulate from the samples. Cells were pelleted through centrifugation at 4000 rpm for 10 min. Next, cells were lysed in 2 mL of buffer containing 75 mM NaCl (Sigma), 3% sodium dodecyl sulfate (SDS, Fisher), 1 mM NaF (Sigma), 1 mM beta-glycerophosphate (Sigma), 1 mM sodium orhtovanadate (Sigma), 10 mM sodium pyrophosphate (Sigma), 1 mM phenylmethylsulfonyl fluoride (PMSF, Sigma), and 1X Complete Mini EDTA free protease inhibitors (Roche) in 50 mM HEPES (Sigma), pH 8.5(Wessel and Flugge 1984). An equal volume of 8M Urea in 50 mM HEPES, pH 8.5 was added to each sample. Cell lysis was achieved through two 10 second intervals of probe sonication at 25% amplitude. Proteins were then reduced with dithiothreitol (DTT, Sigma), alkylated through iodoacetamide (Sigma), and quenched as previously described (Wessel and Flugge 1984). Proteins were then precipitated via chloroform-methanol precipitation and protein pellets were dried (Wessel and Flugge 1984). Protein pellets were re-suspended in 1M urea in 50 mM HEPES, pH 8.5 and digested overnight at room temperature with LysC (Wako) (Lapek et al. 2018). A second, 6-hour digestion using trypsin at 37 ºC was performed and the reaction was stopped through addition of 10% trifluoroacetic acid (TFA, Pierce). Samples were then desalted through C18 Sep-Paks (Waters) and eluted with a 40% and 80% Acetonitrile solution containing 0.5 % Acetic Acid (Lapek et al.

2018). Concentration of desalted peptides was determined with a BCA assay (Thermo Scientific). 50 μg aliquots of each sample were dried in a speed-vac, additional bridge channels consisting of 25 μg from each sample were created and 50 μg aliquots of this solution were used in duplicate per TMT-10 plex as previously described (Lapek et al. 2018). These bridge channels were used to control for labeling efficiency, inter-run variation, mixing errors and the heterogeneity present in each sample(Thompson et al. 2003). Each sample or bridge channel was resuspended in 30% dry acetonitrile in 200 mM HEPES, pH 8.5 for TMT labeling with 7 μL of the appropriate TMT reagent (Thompson et al. 2003). Reagents 126 and 131 (Thermo Scientific) were used to bridge between mass spec runs. Remaining reagents were used to label samples in random order. Labeling was carried out for 1 hour at room temperature, and quenched by adding 8 μL of 5% hydroxylamine (Sigma). Labeled samples were acidified by adding 50 μL of 1% TFA. After TMT labeling each 10-plex experiment was combined and desalted through C18 Sep-Paks and dried in a speed-vac.

Basic pH reverse-phase liquid chromatography sample fractionation

Sample fractionation was performed by basic pH reverse-phase liquid chromatography with concatenated fractions as previously described (Wang et al. 2011). Briefly, samples were re-suspended in 5% formic acid/5% acetonitrile and separated over a 4.6 mm x 250 mm C18 column (Thermo Scientific) on an Ultimate 3000 HPLC fitted with a fraction collector, degasser, and variable wavelength detector. The separation was performed over a 22% to 35%, 60-minute linear gradient of acetonitrile in 10 mM ammonium bicarbonate (Fisher) at 0.5 mL/min. The resulting 96 fractions were combined as previously described(Wang et al. 2011). Fractions were dried under vacuum and re-suspended in 5% formic acid/5% acetonitrile and analyzed by liquid chromatography (LC)-MS^2/MS^3 for identification and quantitation.

LC-MS^2/MS^3 for protein identification and quantitation

All LC-MS^2/MS^3 experiments were carried out on an Orbitrap Fusion (Thermo Fisher Scientific) with an in-line Easy-nLC 1000 (Thermo Fisher Scientific) and chilled autosampler. Separation and acquisition settings were as previously defined (Huttlin et al. 2010).

Proteomic data processing

Data was processed using Proteome Discoverer 2.1 (Thermo Fisher Scientific). MS^2 data was searched against the pDB and Uniprot human database (uniprot.org, accessed 11-28-16). The Sequest searching algorithm(Huttlin et al. 2010) was used to align spectra to database peptides. A precursor mass tolerance of 50 ppm(Huttlin et al., 2010) was specified and 0.6 Da tolerance for MS^2 fragments. Included in the search parameters was static modification of TMT 10-plex tags on lysine and peptide n-termini (+229.162932 Da), carbamidomethylation of cysteines (+57.02146 Da), and variable oxidation of methionine (+15.99492 Da). Raw data was searched at a peptide and protein false discovery rate of 1% using a reverse database search strategy (Gupta & Pevzner, 2009).

TMT reporter ion intensities were extracted from MS^3 spectra for quantitative analysis and signal-to-noise values were used for quantitation. Additional stringent filtering was used removing any moderate confidence peptide spectral matches (PSMs), or ambiguous PSM assignments. Additionally, any peptides with a spectral interference above 25% were removed, as well as any peptides with an average signal to noise ratio less than 10. Proteins matching only one high confidence PSM were not removed in accordance with false discovery rate benchmarking (Gupta and Pevzner 2009). As metaproteome data contains a high degree of similarity in identity between proteins, several decisions were made to reduce false assignments. Standardized methods in Proteome Discoverer (Version 2.1) preferentially assign peptides to proteins that previously had peptides reported. If this does not resolve the assignment, the peptide is assigned to the longest protein. Additionally, a duplicate peptide filter was applied according to the Proteome Discoverer report. Normalization occurred as previously described (Zhang et al. 2016). Briefly, relative abundances are normalized first to the pooled standards for each protein and then to the median signal across the pooled standard. An average of these normalizations was used for the next step. To account for slight differences in amounts of protein labeled, these values were then normalized to the median of the entire dataset and reported as final normalized summed signal-to-noise ratios per protein per sample.

Use of integrated gene catalog for reference library comparison

The integrated reference catalog was downloaded from http://meta.genomics.cn/meta/home (accessed 12-22-2016). A two-step database search method was utilized (Zhang et al. 2016). Briefly, the full database was used as a first pass screen. Second, both forward and reverse database identifications were used to create a study specific database. This database was used to search mass spectrometry data and identifications were filtered at a 1% peptide and protein FDR.

Data analysis

Data analysis was performed in python version 3.5 (www.python.org), and records of the code are available in corresponding Jupyter Notebooks for this project (https://github.com/knightlab-analyses/Crohns-MG-MP-Comparisons). All displayed metaproteomic data was generated using the pDB metaproteomic data unless otherwise specified. Qiime was used for Principle Coordinates Analysis (Caporaso et al. 2010). Spearman correlations were performed through the python package, pandas (http://pandas.pydata.org/). Linear regressions were performed on metagenome sums and metaproteome averages against the metaproteome abundances of each of the biomarker abundances. Protein and gene associations were ranked by the associated coefficient of correlation, and taxonomic and functional annotations of the top associated genes and proteins ($|r| < 0.7$) were compared. Linear regressions were performed using the python package scipy (https://www.scipy.org). Friedman tests were also performed through scipy, comparing genus compositions within the metagenome and metaproteome between samples.

Ethics Statement

The patient had stool samples collected under the consent of two protocols: HRPP #141853 American Gut Project and HRPP #150275 Evaluating the Human Microbiome. Both protocols were approved by University of California San Diego's Human Research Protection Program (HRPP). Written informed consent on dissemination of the result and scientific publication are also included in the approved protocols, and as obtained from the patient.

Data availability

Proteomic data and supplementary files are available online at massive.ucsd.edu (study ID MSV000082113). Metagenomic data is available through EBI https://www.ebi.ac.uk/ena under the study identifiers PRJEB28712 (ERP110957).

Chapter 3 is a reprint of the material as it appears in mSystems, 2019, Robert H. Mills, Yoshiki Vazquez-Baeza, Qiyun Zhu, Lingjing Jiang, James Gaffney, Greg Humphrey, Larry Smarr, Rob Knight and David J. Gonzalez. The dissertation author played a primary role in all aspects of the work ranging from the study design, data acquisition, analysis and writing of the manuscript.

3.6 References

Barnich N, Darfeuille-Michaud A. 2007. Adherent-invasive Escherichia coli and Crohn's disease. *Curr Opin Gastroenterol* **23**: 16-20.

Bohm R, Sauter M, Bock A. 1990. Nucleotide sequence and expression of an operon in Escherichia coli coding for formate hydrogenlyase components. *Mol Microbiol* **4**: 231-243.

Brophy MB, Nolan EM. 2015. Manganese and Microbial Pathogenesis: Sequestration by the Mammalian Immune System and Utilization by Microorganisms. *Acs Chem Biol* **10**: 641-651.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59-60.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335-336.

Chabriere E, Charon MH, Volbeda A, Pieulle L, Hatchikian EC, Fontecilla-Camps JC. 1999. Crystal structures of the key anaerobic enzyme pyruvate:ferredoxin oxidoreductase, free and in complex with pyruvate. *Nat Struct Biol* **6**: 182-190.

Chandramouli K, Qian PY. 2009. Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum Genomics Proteomics* **2009**.

Chang S, Malter L, Hudesman D. 2015. Disease monitoring in inflammatory bowel disease. *World J Gastroenterol* **21**: 11246-11259.

Cleynen I, Boucher G, Jostins L, Schumm LP, Zeissig S, Ahmad T, Andersen V, Andrews JM, Annese V, Brand S et al. 2016. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* **387**: 156-167.

Cohen J. 1988. *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates, Hillsdale, N.J.

Corthesy B. 2013. Multi-faceted functions of secretory IgA at mucosal surfaces. *Front Immunol* **4**.

Dai C, Jiang M, Sun MJ. 2018. Fecal markers in the management of inflammatory bowel disease. *Postgrad Med* doi:10.1080/00325481.2018.1503919: 1-10.

Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, Shah M, Halfvarson J, Tysk C, Henrissat B et al. 2012. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* **7**: e49138.

Gaci N, Borrel G, Tottey W, O'Toole PW, Brugere JF. 2014. Archaea and the human gut: new beginning of an old story. *World J Gastroenterol* **20**: 16062-16078.

Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M et al. 2014. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**: 382-392.

Gupta N, Pevzner PA. 2009. False discovery rates of protein identifications: a strike against the two-peptide rule. *J Proteome Res* **8**: 4173-4181.

Halfvarson J, Brislawn CJ, Lamendella R, Vazquez-Baeza Y, Walters WA, Bramer LM, D'Amato M, Bonfiglio F, McDonald D, Gonzalez A et al. 2017. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol* **2**: 17004.

Henderson B, Allan E, Coates AR. 2006. Stress wars: the direct role of host and bacterial molecular chaperones in bacterial infection. *Infect Immun* **74**: 3693-3706.

Hughes ER, Winter MG, Duerkop BA, Spiga L, Furtado de Carvalho T, Zhu W, Gillis CC, Buttner L, Smoot MP, Behrendt CL et al. 2017. Microbial Respiration and Formate Oxidation as Metabolic Signatures of Inflammation-Associated Dysbiosis. *Cell Host Microbe* **21**: 208-219.

Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villen J, Haas W, Sowa ME, Gygi SP. 2010. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**: 1174-1189.

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

Iskandar HN, Ciorba MA. 2012. Biomarkers in inflammatory bowel disease: current practices and recent advances. *Transl Res* **159**: 313-325.

Jansson JK, Baker ES. 2016. A multi-omic future for microbiome studies. *Nat Microbiol* **1**: 16049.

Joy V, Vora AA, Mascialino B. 2017. F-Calprotectin Use in Inflammatory Bowel Disease (Ibd) Is Characterized by Improved Diagnostic Accuracy, Less Patient Harm and Decreased Costs, Compared with Conventional Serological Markers and Colonoscopy - the Us Perspective. *Value Health* **20**: A251-A251.

Juste C, Kreil DP, Beauvallet C, Guillot A, Vaca S, Carapito C, Mondot S, Sykacek P, Sokol H, Blon F et al. 2014. Bacterial protein signals are associated with Crohn's disease. *Gut* **63**: 1566-1577.

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

Klaassens ES, de Vos WM, Vaughan EE. 2007. Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. *Appl Environ Microbiol* **73**: 1388-1392.

Kolmeder CA, de Been M, Nikkila J, Ritamo I, Matto J, Valmu L, Salojarvi J, Palva A, Salonen A, de Vos WM. 2012. Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. *PLoS One* **7**: e29913.

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

Kolmeder CA, Salojarvi J, Ritari J, de Been M, Raes J, Falony G, Vieira-Silva S, Kekkonen RA, Corthals GL, Palva A et al. 2016. Faecal Metaproteomic Analysis Reveals a Personalized and Stable Functional Microbiome and Limited Effects of a Probiotic Intervention in Adults. *PLoS One* **11**: e0153294.

Lapek JD, Jr., Greninger P, Morris R, Amzallag A, Pruteanu-Malinici I, Benes CH, Haas W. 2017. Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities. *Nat Biotechnol* **35**: 983-989.

Lapek JD, Jr., Mills RH, Wozniak JM, Campeau A, Fang RH, Wei X, van de Groep K, Perez-Lopez A, van Sorge NM, Raffatellu M et al. 2018. Defining Host Responses during Systemic Bacterial Infection through Construction of a Murine Organ Proteome Atlas. *Cell Syst* doi:10.1016/j.cels.2018.04.010.

Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T et al. 2014. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* **32**: 834-841.

Liu Y, Beyer A, Aebersold R. 2016. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**: 535-550.

Maier T, Guell M, Serrano L. 2009. Correlation of mRNA and protein in complex biological samples. *Febs Letters* **583**: 3966-3973.

Manichanh C, Borruel N, Casellas F, Guarner F. 2012. The gut microbiota in IBD. *Nat Rev Gastroenterol Hepatol* **9**: 599-608.

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P et al. 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* **40**: D115-122.

Marotz C, Amir A, Humphrey G, Gaffney J, Gogul G, Knight R. 2017. DNA extraction for streamlined metagenomics of diverse environmental samples. *Biotechniques* **62**: 290-293.

McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y et al. 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3**.

Mesuere B, Debyser G, Aerts M, Devreese B, Vandamme P, Dawyndt P. 2015. The Unipept metaproteomics analysis pipeline. *Proteomics* **15**: 1437-1442.

Mosli MH, Zou G, Garg SK, Feagan SG, MacDonald JK, Chande N, Sandborn WJ, Feagan BG. 2015. C-Reactive Protein, Fecal Calprotectin, and Stool Lactoferrin for Detection of Endoscopic Activity in Symptomatic Inflammatory Bowel Disease Patients: A Systematic Review and Meta-Analysis. *Am J Gastroenterol* **110**: 802-819; quiz 820.

Ni J, Shen TD, Chen EZ, Bittinger K, Bailey A, Roggiani M, Sirota-Madi A, Friedman ES, Chau L, Lin A et al. 2017. A role for bacterial urease in gut dysbiosis and Crohn's disease. *Sci Transl Med* **9**.

Palmela C, Chevarin C, Xu Z, Torres J, Sevrin G, Hirten R, Barnich N, Ng SC, Colombel JF. 2018. Adherent-invasive Escherichia coli in inflammatory bowel disease. *Gut* **67**: 574-587.

Palsson B, Zengler K. 2010. The challenges of integrating multi-omic data sets. *Nature Chemical Biology* **6**: 787-789.

Pandey A, Mann M. 2000. Proteomics to study genes and genomes. *Nature* **405**: 837-846.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**: 417-419.

Presley LL, Ye JX, Li XX, LeBlanc J, Zhang ZP, Ruegger PM, Allard J, McGovern D, Ippoliti A, Roth B et al. 2012. Host-microbe relationships in inflammatory bowel disease detected by bacterial and metaproteomic analysis of the mucosal-luminal interface. *Inflammatory Bowel Diseases* **18**: 409-417.

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* **35**: 833-844.

Rosa Viner RB, Michael Blank, John Rogers. 2013. Increasing the Multiplexing of Protein
Quantitation from 6- to 10-Plex with Reporter Ion Isotopologues.

Scanlan PD, Shanahan F, Marchesi JR. 2008. Human methanogen diversity and incidence in healthy and diseased colonic groups using mcrA gene analysis. *BMC Microbiol* **8**: 79.

Sitkin S, Pokrotnieks J. 2018. Clinical Potential of Anti-inflammatory Effects of Faecalibacterium prausnitzii and Butyrate in Inflammatory Bowel Disease. *Inflamm Bowel Dis* doi:10.1093/ibd/izy258.

Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermudez-Humaran LG, Gratadoux JJ, Blugeon S, Bridonneau C, Furet JP, Corthier G et al. 2008. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci U S A* **105**: 16731-16736.

Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt C. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**: 926-932.

Takahashi K, Nishida A, Fujimoto T, Fujii M, Shioya M, Imaeda H, Inatomi O, Bamba S, Sugimoto M, Andoh A. 2016. Reduced Abundance of Butyrate-Producing Bacteria Species in the Fecal Microbial Community in Crohn's Disease. *Digestion* **93**: 59-65.

Tanca A, Palomba A, Deligios M, Cubeddu T, Fraumene C, Biosa G, Pagnozzi D, Addis MF, Uzzau S. 2013. Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PLoS One* **8**: e82981.

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

Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C. 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**: 1895-1904.

Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G et al. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**: 457-463.

Ting L, Rad R, Gygi SP, Haas W. 2011. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* **8**: 937-940.

Toyama BH, Hetzer MW. 2013. Protein homeostasis: live long, won't prosper. *Nat Rev Mol Cell Biol* **14**: 55-61.

Tsilimigras MC, Fodor AA. 2016. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol* **26**: 330-335.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The human microbiome project. *Nature* **449**: 804-810.

van der Sluys Veer A, Brouwer J, Biemond I, Bohbouth GE, Verspaget HW, Lamers CB. 1998. Fecal lysozyme in assessment of disease activity in inflammatory bowel disease. *Dig Dis Sci* **43**: 590-595.

Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, Lefsrud MG, Apajalahti J, Tysk C, Hettich RL et al. 2009. Shotgun metaproteomics of the human distal gut microbiota. *ISME J* **3**: 179-189.

Vermeire S, Van Assche G, Rutgeerts P. 2004. C-reactive protein as a marker for inflammatory bowel disease. *Inflamm Bowel Dis* **10**: 661-665.

Walters WA, Xu Z, Knight R. 2014. Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Lett* **588**: 4223-4233.

Wang Y, Yang F, Gritsenko MA, Wang Y, Clauss T, Liu T, Shen Y, Monroe ME, Lopez-Ferrer D, Reno T et al. 2011. Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11**: 2019-2026.

Weekes MP, Tomasec P, Huttlin EL, Fielding CA, Nusinow D, Stanton RJ, Wang EC, Aicheler R, Murrell I, Wilkinson GW et al. 2014. Quantitative temporal viromics: an approach to investigate host-pathogen interaction. *Cell* **157**: 1460-1472.

Wessel D, Flugge UI. 1984. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal Biochem* **138**: 141-143.

Winter SE, Lopez CA, Baumler AJ. 2013. The dynamics of gut-associated microbial communities during inflammation. *EMBO Rep* **14**: 319-327.

Xiao J, Tanca A, Jia B, Yang R, Wang B, Zhang Y, Li J. 2018. Metagenomic Taxonomy-Guided Database-Searching Strategy for Improving Metaproteomic Analysis. *J Proteome Res* doi:10.1021/acs.jproteome.7b00894.

Zhang X, Chen W, Ning Z, Mayne J, Mack D, Stintzi A, Tian R, Figeys D. 2017. Deep Metaproteomics Approach for the Study of Human Microbiomes. *Anal Chem* **89**: 9407-9415.

Zhang X, Deeke SA, Ning Z, Starr AE, Butcher J, Li J, Mayne J, Cheng K, Liao B, Li L et al. 2018. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nat Commun* **9**: 2873.

Zhang X, Ning ZB, Mayne J, Moore JI, Li J, Butcher J, Deeke SA, Chen R, Chiang CK, Wen M et al. 2016. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **4**.

Chapter 4

Meta–omics Reveals Microbiome Driven Proteolysis as a Contributing Factor to Severity of Ulcerative Colitis Disease Activity

4.1 Abstract

Ulcerative colitis has a significant global burden(Fumery et al. 2018), and is characterized by an aberrant immune response directed towards the gut microbiota(Sartor and Wu 2017). Current treatment options exclusively target host inflammatory pathways and are often ineffective in managing disease(Dulai et al. 2014). To better understand host-microbiome interactions governing ulcerative colitis (UC), we collected and analyzed six fecal or serum based –omic datasets from 40 UC patients displaying a wide range of clinical, endoscopic, and histologic disease activity. All meta-omics displayed large-scale shifts related to disease activity, with the metabolome and metaproteome best predicting disease severity. After broad-scale analyses, metaproteomics identified a striking association between *Bacteroides* proteases and disease severity. Increased peptide fragments in the metapeptidome of active UC patients further implicated bacterial proteolysis. *Bacteroides vulgatus,* enriched in the metagenomic analysis, disrupted intestinal epithelial permeability *in vitro*, and protease inhibition was sufficient to restore epithelial barrier. Furthermore, transplantation of fecal material from UC patients into germ-free mice resulted in increased colitis, and oral administration of protease inhibitors attenuated disease severity. Our findings highlight the potential therapeutic approach for UC by targeting microbial proteases to ameliorate intestinal barrier dysfunction and restore mucosal integrity.

4.2 Main

Ulcerative colitis (UC), an inflammatory bowel disease (IBD), is characterized by chronic inflammation of the colon, with severity of mucosal inflammation being associated with a higher risk of work disability, hospitalization, colorectal cancer, and colectomy(Fumery et al. 2018). Non-specific immunosuppressive agents targeting the host, such as steroids, thiopurines, and/or biologics, are used to offset the natural history of disease in patients with moderate-severe inflammation. These therapies are, however, associated with significant risks and often ineffective in adequately managing disease(Dulai et al. 2014). Genomic technologies have identified associations between microbial dysbiosis, or temporal shifts in composition, and UC severity(Sartor and Wu 2017; Schirmer et al. 2018; Shen et al. 2018). While recent efforts extended profiling of microbiota in UC beyond genomics(Lloyd-Price et al. 2019), it remains poorly understood if these shifts are causal or associative in nature, and which mechanisms govern pathogenic roles of the microbiome in UC. Metaproteomics is a developing mass spectrometry (MS) method for the comprehensive analysis of the proteins expressed by a community of organisms(Verberkmoes et al. 2009). We predict that the integration of a contemporary metaproteomics platform with other technologies could allow for a more in-depth understanding of host-microbiome interactions governing UC severity and the identification of novel microbial therapeutic targets(Erickson et al. 2012; Jansson and Baker 2016; Franzosa et al. 2018; Zhang et al. 2018).

In the study herein, we recruited 40 UC patients from a single academic IBD center (UC San Diego) who underwent extensive phenotyping with clinical disease activity indices and blinded assessments of endoscopic and histologic severity(Lewis et al. 2008; Dulai et al. 2015; Narula et al. 2018). Individual patient matched serum and fecal samples were subset for genomic, metabolomic, serum proteomic, and metaproteomic analyses, and previously established methods for shared database assembly and quantification were used for integration(Mills et al. 2019). Notably, application of our multiplexing metaproteomic methods provided increased depth and a greater than 10-fold increase in proteins quantified per sample in comparison to the proteome data available from the Human Microbiome Project's IBD multi-omics database (Li et al. 2014; Zhang et al. 2016; Lloyd-Price et al. 2019).

Meta-omic associations with UC severity

All meta-omics displayed large-scale shifts related to disease activity, with the metabolome and metaproteome best predicting UC disease. No distinct taxonomic shifts were observed with increasing disease severity. Using linear regression we identified 3,636 proteins and 62,982 genes that were moderately associated to clinical disease severity (Cohen 1988). The metagenome demonstrated the largest genera level pro-inflammatory relationships with *Escherichia* and *Veillonella,* and the metaproteome demonstrated human and *Bacteroides* proteins dominated the positive associations. There were 528 *Bacteroides* proteins that positively correlated to disease severity, which constituted nearly 60% of the positively associated, non-human proteins. The metagenome largely reflected the direction and magnitude of the genera associations identified in the metaproteome, however, *Bacteroides* genes showed a much weaker relationship to high disease severity relative to the metaproteome (Fig. 4.1a). Interestingly, positive correlations to *Bacteroides* contrasted *Faecalibacterium*, which plays protective roles in IBD(Sokol et al. 2008). The relationship between *Bacteroides* and *Faecalibacterium* was supported in amplicon sequencing data (Fig. 4.1b), but unrelated to the changes observed in community structure (Fig. 4.1c). The most abundant *Bacteroides* species included *B. vulgatus*, *B. dorei*, *B. uniformis*, *B. ovatus*, *B.*

thetaiotaomicron and *B. fragilis* (Fig. 4.1d), which are among the most prevalent *Bacteroides* species isolated from healthy human subjects(Kulagina et al. 2012). Functionally, *Bacteroides* enzymes correlated to disease activity while *Faecalibacterium* membrane transporters correlated to remission (Fig. 4.1e-f). The enzymes identified largely had protease activity, so we further searched for proteins with Gene Ontology (GO) annotations containing "protease" or "peptidase" terms. We found that 78% (132/169) of *Bacteroides* proteases had a positive correlation ($r > 0$) to the partial Mayo score. There were 45 distinct proteases derived from 34 species of *Bacteroides*. These proteases were grouped, revealing 10 serine, 9 metallo, and 4 cysteine peptidases with a range of activities including 5 di-peptidases, an endopeptidase, sialidase and signal peptidase (Fig. 4.1g). As serine and metalloproteases largely function in the extracellular space(Vergnolle 2016), we hypothesized these proteases may play roles in extracellular proteolysis.

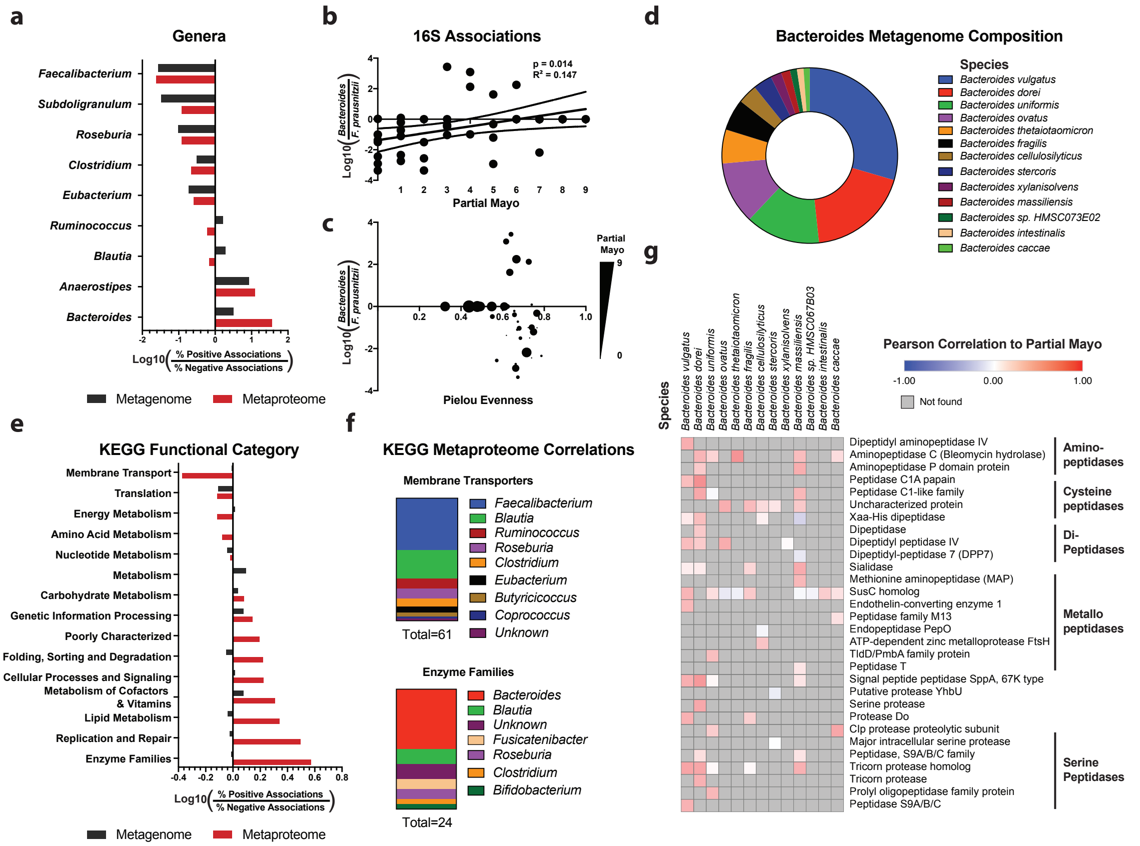


Figure 4.1 Bacteroides proteases are correlated to disease severity. a, Comparison of metagenomic and metaproteomic genera associations to severity**.** Linear correlation to partial Mayo clinical severity was performed on genes and proteins in the microbial metagenome and metaproteome and the genera of moderately associated (|r| > 0.3) genes and proteins were compared. Each bar represents the Log2 ratio of positively associated to negatively associated genes or proteins per genus. **b,** Ratios of microbes of interest correlate to severity. Sctterplots of the Log10 ratio of 16S reads per sample are plotted by partial Mayo severity. Reads associated with *Faecalibacterium prausnitzii* as well as a sequence associated with the Bacteroides genus are shown with a best fit line, a 95% confidence interval and the associated R^2 and p-values. **c,** Evenness does not correlate to ratios of microbes of interest. A scatter plot is shown with the 16S based Pielou Evenness on the x-axis and the Log10 ratio of Bacteroides to Faecalibacterium prausnitzii 16S reads. Samples are sized according to their associated partial Mayo score. **d,** Bacteroides species profile. The total CPM (counts per million) of genes derived from each Bacteroides species was determined and the most abundant species are displayed. **e,** Comparison of metagenomic and metaproteomic KEGG functional associations to severity. An identical analysis to (a), binned by KEGG functional group assignments. **f,** Genera composition of KEGG categories of interest. Stacked barplots displaying the composition of genera within the significantly associated proteins corresponding to KEGG categories of interest. Negatively correlated proteins with KEGG annotations for membrane transport are shown with a bias for *Faecalibacterium*. Positively correlated proteins with annotations for enzyme families are shown to have bias for Bacteroides. **g,** Correlation of Bacteroides proteases to disease severity. Correlation of Bacteroides proteases to disease severity. A heatmap is displayed showing the correlation of the metaproteome abundance of proteases identified from the most abundant Bacteroides species to the partial Mayo clinical scoring. Pearson correlation (r) is represented on a red (high correlation) to blue (low correlation) scale. Proteases not found in the metaproteome are colored gray. Proteases are clustered by gene ontology categorization.

Evidence of proteolysis at high disease severity

Using *de novo* identification of short peptide fragments, we observed an increased presence of peptides among high severity samples (Fig. 4.2a, (Zhang et al. 2012)). Network analyses of human proteins in serum and fecal samples showed a functional enrichment for peptidase activity and inhibition, including neutrophil proteases (Fig. 4.2b). The known cleavage patterns of Neutrophil elastase and Proteinase-3(O'Donoghue et al. 2013) were not strong signals in the metapeptidome data (Fig. 4.2c), indicating neutrophil proteases may not be the primary drivers of proteolysis. As we observed a strong correlation between *Bacteroides* proteases and severity, it is possible that the host peptide fragments (Fig. 4.2d) were a result of cleavage from bacterial proteases. Serum to fecal comparisons of Serpin A1 ratios demonstrated a highly significant association with disease activity even among patients with apparent endoscopic healing (Fig. 4.2e, (Strygler et al. 1990)), highlighting the importance of proteolysis in UC pathology across all spectrums of disease state. We identified several *Bacteroides* proteases, including dipeptidyl peptidase IV, which has orthologs in *P. gingivalis* and cleaves X-Pro or X-Ala dipeptides from N-terminal polypeptides(Kumagai et al. 2000). Supporting the activity of these proteases was four dipeptides which significantly correlated to disease severity, including two X-Pro species. We also observed an abundance of *Bacteroides* TonB related proteins that correlated to disease severity (Fig. 4.1g) and lack of histologic remission despite apparent endoscopic healing (Fig. 4.2f). TonB recruits proteins to the outer membrane, and related proteins from *Bacteroides* have been described as immunogenic(Wei et al. 2001) and suggested as a biomarker for Crohn's disease(Juste et al. 2014). SusC is a TonB-dependent porin that *Bacteroides* species use for the binding

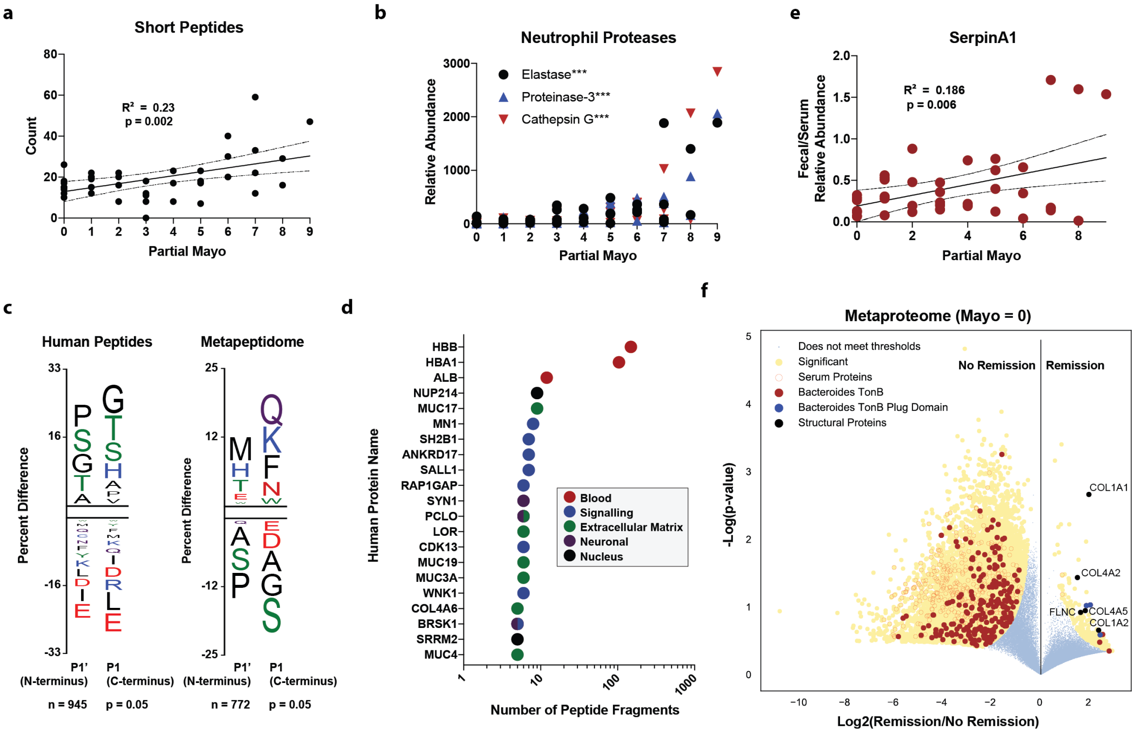


Figure 4.2 Bacterial and host proteolysis correlates to disease severity. a, Peptide fragments are more abundant at high severity. The number of peptides identified is represented on the y-axis and the partial Mayo clinical score is represented on the x-axis. Significance and Pearson correlation of the linear relationship is shown with a 95% CI drawn surrounding the best-fit line. **b,** Neutrophil proteases are significantly correlated to disease severity. The proteome relative abundance is illustrated on the y-axis with the partial Mayo scoring on the x-axis. *** Indicates a $p < 0.001$ for the linear relationship between each protease. **c,** Peptide termini indicate unique proteolysis of human and microbial proteins. The frequency of each amino acid within the N and C terminus of human and de-novo peptides was compared to either the human proteome or the total amino acid content of de novo peptides. The Y-axis represents the percent difference of each residue and the letter indicates the amino acid associated with the difference. The N and C terminus are shown separately and each residue is colored by chemical property (Green = polar, Black = Hydrophobic, Red = Acidic, Blue = Basic, Purple = Neutral). **d,** The number of peptide fragments from human proteins indicates potential targets of proteolysis. The gene symbol for the human proteins with the most short peptides present are shown on the y-axis and the quantity of peptides is shown on a log10 transformed x-axis. The proteins are colored by the observed dominant categories. Proteins fitting into multiple categories have both colors represented. **e,** SerpinA1 fecal to serum ratios correlate to disease severity. The ratio of proteome relative abundances of SerpinA1 in the fecal and serum are plotted on the y-axis with the partial Mayo severity plotted on the x-axis. Significance and Pearson correlation of the linear relationship is shown with a 95% CI drawn surrounding the best-fit line. **f,** The metaproteome of patients with mucosal healing have large fluctuations associated with histological remission. A volcano plot depicting the Log2(fold change) and log10(p-value) for each protein in the metaproteome. Significance was determined by a |π|(Xiao et al. 2014) > 1, and protein groups of interest are highlighted in the legend.

and degradation of complex carbohydrates(Reeves et al. 1996; Martens et al. 2009). TonB-dependent proteins have also been shown to mediate binding of *Bacteroides* to extracellular matrix proteins, including collagen(Pauer et al. 2009).

Protease inhibition prevents *B. vulgatus* disruption of human intestinal epithelial barrier

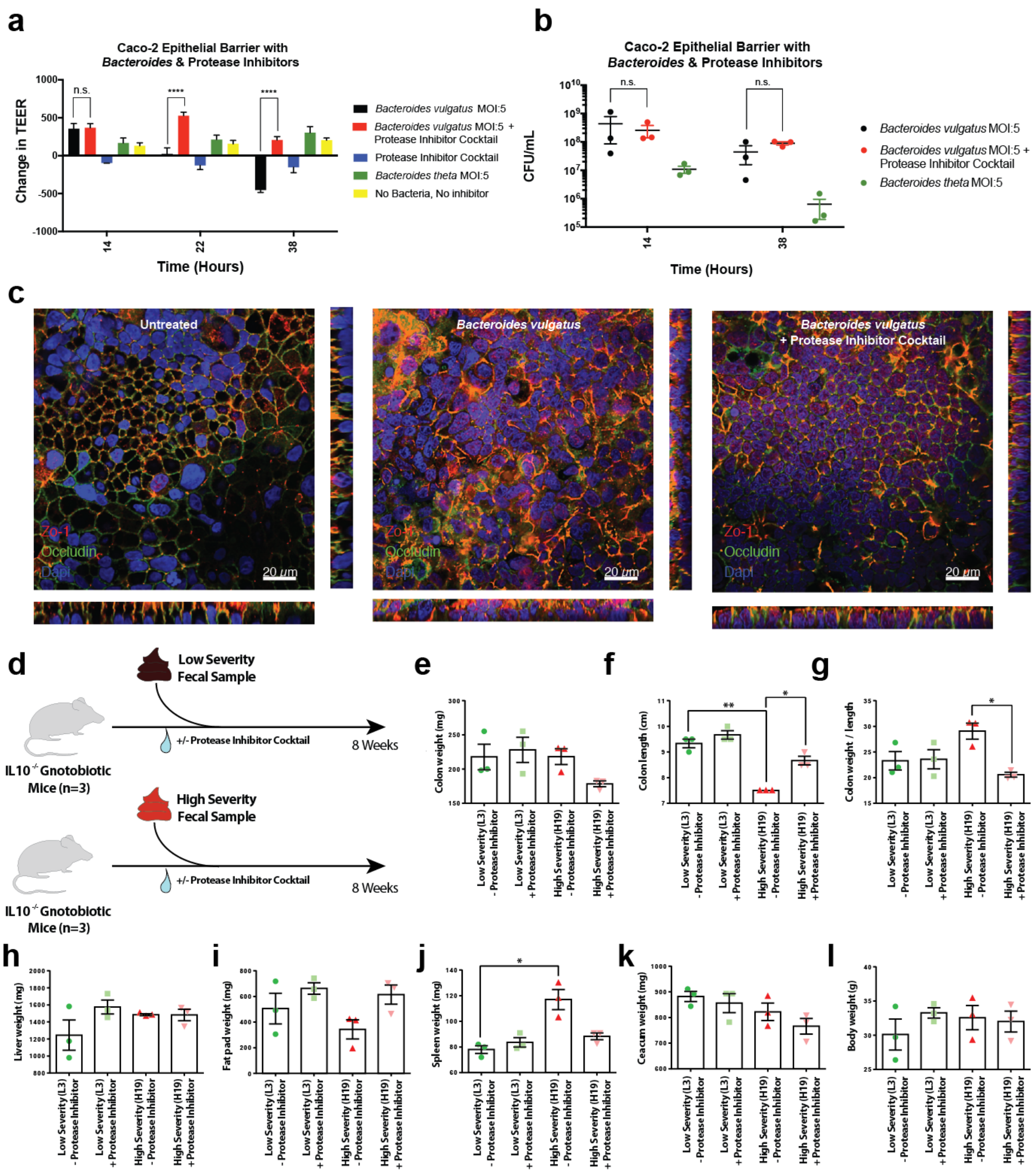
Given the enrichment of *Bacteroides* proteases from our meta-omics analyses, we assessed the six most abundant *Bacteroides* species for effects on intestinal barrier using *in vitro* Caco-2 epithelial monolayers. Our results showed a significant decrease in transepithelial electrical resistance (TEER) after 38 hours of incubation with the two most abundant *Bacteroides* species, *Bacteroides vulgatus* and *Bacteroides dorei*, while other species increased TEER. We next assessed the contribution of protease activity in disruption of epithelial permeability through the addition of a protease inhibitor cocktail to the *B. vulgatus* incubated cells, and found a dramatically increased TEER at both 22 and 38 hours post infection (Adjusted p-value < 0.0001, Fig. 4.3a). The phenotype was not due to effects on bacterial growth or viability, as colony forming units (CFUs) were not significantly different between the *B. vulgatus* wells treated with or without protease inhibitor cocktail (Adjusted p-value = 0.98, Fig. 4.3b).

Confocal microscopy of the intestinal monolayers revealed dramatic impact on the *B. vulgatus* treated epithelial cells, with apparent degradation of tight-junction proteins, Zo-1 and Occludin (Fig. 4.3c). The protease inhibitor cocktail treatment may have specificity for *B. vulgatus's* degradation of Occludin, as visible restoration of Zo-1 was not observed with protease inhibitor treatment. Imaging studies also demonstrated potential impacts on cell morphology and actin networks of the Caco-2 cells treated with *B. vulgatus*. Further studies are needed, but our results suggest a pathological effect of *B. vulgatus* on intestinal cells dependent upon protease activity.

Protease inhibition prevents colonic inflammation from patient derived fecal transplants in germ-free mice

Next we sought to confirm the efficacy of protease inhibition in a germ-free mouse model. We cross-referenced the metagenome with the metaproteome and metapeptidome to identify fecal samples with the lowest and highest *B. vulgatus* and *B. dorei* related protease activity. Notably, we observed limited association between protease abundance and metagenome frequency of *B. vulgatus* and *B. dorei.* A single high severity (H19) and low severity fecal sample (L3) were selected based on several metrics; *B. vulgatus* protease abundance, number of peptide fragments (59 versus 14), male donor status, partial Mayo score (0 versus 7) and *Bacteroides* composition. An 8-week fecal transplant colonization study in 4-week-old male IL10$^{-/-}$ gnotobiotic mice was performed (Fig. 4.3d). Through the duration of the study, mice were fed either water containing a protease inhibitor cocktail or water alone (n=3 per group), after which time mice were sacrificed and macroscopic measurements were taken (Fig. 4.3e-l). The high severity sample induced an average 19% reduction in colon length (p = 0.008, Fig. 4.3f), a 24% increase in colon weight/length (p = 0.072, Fig. 4.3g), and a 50% increase in spleen weight (p = 0.026, Fig. 4.3j), without significant impact on fat pad weight (p = 0.324, Fig. 4.3i), liver weight (p = 0.309, Fig. 4.3h), caecum weight (p = 0.217, Fig. 4.3k) or total body weight (p = 0.442, Fig. 4.3l). Interestingly, the measurements most impacted by the high severity fecal sample significantly shifted toward the low severity groups in the protease inhibitor group (colon length p = 0.020, Fig. 4.3f; colon

Figure 4.3 Protease inhibition ameliorates *Bacteroides vulgatus* disruption of epithelial cell resistance and *in vivo* colitis induced by UC patient fecal transplantation. a, Protease inhibitor cocktail significantly reduces the Caco-2 resistance reduction when co-culturing *Bacteroides vulgatus*. Caco-2 cells were grown in monolayers on a transwell for 2.5 weeks before inoculating *Bacteroides vulgatus* or *Bacteroides thetaiotamicron* at a multiplicity of infection (MOI) of ~5. Transepithelial electrical resistance (TEER) was measured at the given hours post inoculation. Shown are representative barplots and standard error of the mean (SEM) from 3 biological replicates containing 3 technical replicates within each experiment. **b,** Protease inhibitor cocktail does not significantly influence the number of colony forming units during Caco-2 co-culturing with *Bacteroides vulgatus*. Colony forming units from above the transwell insert were estimated through serial dilution and plating onto BHI-S plates under anaerobic conditions. Plotted are the mean CFUs from each experimental condition from three biological replicates containing 3 technical replicates per experiment. **c,** Representative images from confocal microscopy of the transwell experiments. Following 38 hours of co-culturing, the Caco-2 transwell inserts were fixed and stained for immunofluorescence of tight junction proteins, Zo-1 and Occludin. A representative image from untreated Caco-2 cells, Caco-2 cells co-cultured with *Bacteroides vulgatus*, and Caco-2 cells co-cultured with *Bacteroides vulgatus* and a protease inhibitor cocktail are shown. **d**, Experimental design of humanized IL10$^{-/-}$ mouse study. Fecal samples from one high severity patient and one low severity patient were transplanted into 6x gnotobiotic mice per patient sample. During 8-weeks of colonization, a protease inhibitor cocktail was continuously administered through the drinking water of 3 mice per patient sample. Mice were sacrificed after 8-weeks and macroscopic organ measurements were taken. **e-l** Barplots showing the mean and standard error of the mean are shown for colon weight (**e**), colon length (**f**), colon weight/length (**g**), liver weight (**h**), fat pad weight (**i**), spleen weight (**j**), ceacum weight (**k**), and body weight (**l**). * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001, **** p-value < 0.0001.


a
Caco-2 Epithelial Barrier with
Bacteroides & Protease Inhibitors
Change in TEER
1000
500
0
-500
-1000
n.s.
****
****
14
22
38
Time (Hours)
Bacteroides vulgatus MOI:5
Bacteroides vulgatus MOI:5 + Protease Inhibitor Cocktail
Protease Inhibitor Cocktail
Bacteroides theta MOI:5
No Bacteria, No inhibitor
b
Caco-2 Epithelial Barrier with
Bacteroides & Protease Inhibitors
CFU/mL
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
n.s.
n.s.
14
38
Time (Hours)
Bacteroides vulgatus MOI:5
Bacteroides vulgatus MOI:5 + Protease Inhibitor Cocktail
Bacteroides theta MOI:5
c
Untreated
Bacteroides vulgatus
Bacteroides vulgatus + Protease Inhibitor Cocktail
Zo-1
Occludin
Dapi
20 μm
d
Low Severity Fecal Sample
+/- Protease Inhibitor Cocktail
8 Weeks
IL10$^{-/-}$ Gnotobiotic Mice (n=3)
High Severity Fecal Sample
+/- Protease Inhibitor Cocktail
8 Weeks
IL10$^{-/-}$ Gnotobiotic Mice (n=3)
e
Colon weight (mg)
f
Colon length (cm)
**
*
g
Colon weight / length
*
h
Liver weight (mg)
i
Fat pad weight (mg)
j
Spleen weight (mg)
*
k
Ceacum weight (mg)
l
Body weight (g)
Low Severity (L3) - Protease Inhibitor
Low Severity (L3) + Protease Inhibitor
High Severity (H19) - Protease Inhibitor
High Severity (H19) + Protease Inhibitor

weight/length p = 0.023, Fig. 4.3g; spleen weight p = 0.055, Fig. 4.3j). These studies reveal that the microbiome derived from severe UC patients who express high *Bacteroides* protease activity may induce pathological changes through protease activity, and that protease inhibition may have potential as a therapeutic intervention in severe UC.

4.3 Discussion

Here, we effectively collect and translate one of the most comprehensive meta-omic profiles of UC patients to date into a hypothesis of biological and therapeutic value. Through integrating fecal metaproteomics, metabolomics, 16S gene amplicon sequencing, shotgun metagenomic sequencing, metapeptidomics, and serum proteomics, in addition to *in vitro* and *in vivo* validation, we demonstrate that certain members of the microbiome, such as *Bacteroides vulgatus*, may contribute to exacerbating disease activity through protease activity. Further, given the promise of our *in vitro* and *in vivo* experiments, this study sets the stage for further investigation of protease inhibition as a novel therapeutic approach in UC.

To generate our hypothesis, we utilized several innovative -omic advances that may be of broad interest. The core of our findings stemmed from our previously developed integrated approach for comparing metagenomic and metaproteomic data(Mills et al. 2019). This allowed the identification of discrepancies between the more traditionally collected metagenomic and multiplexed metaproteomic data sets. Given that previous high profile IBD data sets that included metaproteomic data(Lloyd-Price et al. 2019) used methods that generated an order of magnitude more missing values (i.e.

sparsity), we had high confidence and interest in further investigating findings absent in these studies(Lloyd-Price et al. 2019). One striking observation uniquely highlighted in our study was that ~60% of microbial proteins correlating to disease activity were derived from *Bacteroides*. While metapeptidomic data is rarely collected in microbiome studies, this data provided an important complementary tool for identifying that proteolysis, potentially derived from *Bacteroides* proteases, was correlated to severity. By integrating metagenomic data, we provided a genomic context to our findings and identified *Bacteroides* species of interest for our *in vitro* studies. Other -omic profiles (serum proteomics, metabolomics, and 16S) further corroborated and contextualized the core hypothesis of *Bacteroides* derived proteolysis as a contributing factor to UC severity.

Our novel findings on *Bacteroides* were derived in the backdrop of several previously described observations. An early metaproteomic study identified *Bacteroides* proteins as markers of CD(Juste et al. 2014), although genomic approaches only occasionally implicate *Bacteroides*(Schirmer et al. 2018; Vich Vila et al. 2018), and *Bacteroides* functional role in IBD was not well established(Wexler 2007). *Bacteroides* typically reside in the outer mucosal layer of the colon(Donaldson et al. 2016), and are described as decreased in IBD(Zhou and Zhi 2016). There is some evidence that commensal *Bacteroides* species can induce colitis in mouse models(Bloom et al. 2011), although they are typically beneficial unless outside of the gut(Wexler 2007). As the *in vitro* studies found a disruptive phenotype for only *Bacteroides vulgatus* and *Bacteroides dorei* (which are highly related species(Bakir et al. 2006)), and that this phenotype was only ameliorated through certain protease inhibitors, it may be that only a few

Bacteroides species and proteases are related to inducing severity. This is highlighted by cross-referencing the metagenome with metaproteomics and metapeptidomics where even among the high severity UC patients, only a sub-set were noted to have high *Bacteroides vulgatus* or *dorei* protease activity despite having comparable frequency.

Interestingly, *Bacteroides* protein expression was strongly correlated to clinical severity while *Bacteroides* DNA showed only modest correlation. One possible explanation for this difference would be the production of *Bacteroides* outermembrane vesicles. Supporting this were many correlated outermembrane proteins including variants of SusC, a membrane protein with potential roles in binding mucus glycans(Wexler 2007). *Bacteroides* are thought to produce membrane vesicles abundant in proteases(Elhenawy et al. 2014), further strengthening this hypothesis. Recently, extracellular vesicles were linked to IBD, with *Bacteroides* proteins being the majority of bacterial extracellular proteins(Zhang et al. 2018). Our data suggests that further studies into the links between *Bacteroides* membrane vesicles and IBD may be of interest.

Extracellular matrix remodeling(Shimshoni et al. 2015) and protease activity(Vergnolle 2016) are known molecular events in IBD, but current treatments are focused on targeting host inflammatory pathways(Ordas et al. 2012). Bacterial protease inhibition may present a novel therapeutic strategy that prevents downstream tissue destruction and influx of immune cells(Shimshoni et al. 2015; Vergnolle 2016). Protease studies within IBD have predominately focused on host derived proteases, although some have recognized the potential contribution of bacterial proteases(Steck et al. 2012). The most efficacious approach in our studies involved broad-spectrum protease inhibition, targeting serine and cysteine proteases. The positive outcomes may be a combination of preventing both host and bacterial based proteolysis. However, the phenotypes observed specific to *Bacteroides vulgatus* and our fecal transplant studies suggest that the bacterial contribution to proteolysis may be of high importance.

The multidimensional meta-omic integration shown here serves as a landmark study in comparative –omics and the development of hypotheses from large-scale data integration. Starting with broad-scale analysis and further refining our studies according to an observation of interest led to important findings within each dataset. The efficacy of protease inhibition *in vitro* and *in vivo* validates the utility of our approach and opens new areas of investigation into UC pathology and treatment.

4.4 Methods

Patient population and clinical diagnostics

UC patients were selected from a convenience sampling biobank at the University of California at San Diego (UCSD: PI Dulai). In this biobank patients consent to longitudinal data collection on patient demographics (age, gender, ethnicity), disease characteristics (prior surgeries, disease-related complications, phenotype classification according to Montreal sub-classifications), current and prior treatments (corticosteroids, immunomodulators, biologics), and clinical disease activity (patient reported outcomes using the partial Mayo score and endoscopic scores). Alongside this data collection patients agree to stool, serum, and mucosal biopsy collection. When endoscopy is performed as part of routine practice, stool is collected within 24 hours prior to endoscopy and serum is collect the day of endoscopy. At each endoscopy a physician with advanced training in IBD performs a detailed endoscopic disease activity assessment using the Mayo endoscopic sub-score and the Ulcerative Colitis Endoscopic Index of Severity (UCEIS), without knowledge of the clinical disease activity score or biomarker data. Routine standard of care biopsies are scored using the Geboes score by a pathologist with training and expertise in IBD, who is blinded to clinical, biomarker, and endoscopic data and scores. Further information regarding clinical, endoscopic and histologic activity scoring have been previously discussed(Dulai et al. 2015). All serum and stool samples are aliquoted within 24 hours of collection to avoid future freeze-thaw cycles, and samples are stored at -80 ºC until future analyses.

DNA extraction

Frozen samples were thawed and transferred into 96-well plates containing garnet beads and extracted using Qiagen MagAttract DNA kit adapted for magnetic bead purification as previously described(Marotz et al. 2017). DNA was eluted in 100 μl Qiagen elution buffer.

16S gene amplicon sequencing

16S rRNA gene amplicon sequencing was performed according to the Earth Microbiome Project. Briefly, the V4 region of the 16S rRNA gene (515f/806r) was amplified from 1 ul DNA per sample in triplicate(Caporaso et al. 2012; Thompson et al. 2017). Amplicons were quantified with Quant-iT™ PicoGreen™ dsDNA Assay Kit, and 240 ng, or maximum 15 ul, of each sample was pooled into a final library and cleaned using the QIAquick PCR Purification Kit. Paired-end sequencing was performed on the Illumina MiSeq using MiSeq Reagent Kit v3 (300-cycle).

Shotgun metagenomic sequencing

Extracted DNA was quantified with PicoGreen™ dsDNA Assay Kit, and 1 ng of input, or maximum 3.5 μl, gDNA was used in a 1:10 miniaturized Kapa HyperPlus protocol. Per sample libraries were quantified and pooled at equal nanomolar concentration. The pooled library was cleaned with the QIAquick PCR Purification Kit and size selected for fragments between 300 and 700 bp on the Sage Science PippinHT. The pooled library was sequenced as a paired-end 150-cycle run on an Illumina HiSeq4000 v2 at the UCSD IGM Genomics Center.

Processing of metagenomic reads for a shared reference library

Because typical metagenomics and metaproteomics workflows require a reference database, it was necessary to create from scratch using a minimal approach a single reference database that could be used for both metagenomics and metaproteomics from the individualized data. All reads from each sample were concatenated. Next the MEGAHIT alignment program(Li et al. 2015) was utilized for assembling short reads into larger contigs. Assembled contigs were searched for possible coding regions through the program Prodigal(Hyatt et al. 2010). Next, the program Diamond(Buchfink et al. 2015) was used for gene alignment to the uniref50 database. Finally, the most likely uniref50 entry, determined through bitScore, was used for the functional annotations. KEGG orthology annotations were cross-referenced using GhostKOALA(Kanehisa et al. 2016). Taxonomic assignments were determined by Diamond alignment(Buchfink et al. 2015) to an in-house library of microbial genomes. This database was used as a reference database for both metaproteomic data and shotgun sequencing data. Scripts used for data processing are available online (https://github.com/knightlab-analyses/uc-severity-multiomics).

Unweighted UniFrac analysis of shotgun metagenomic data

Taxonomic profiling of shotgun sequences was performed using Centrifuge 1.0.3 with default parameter settings against the aforementioned in-house microbial genome database. The numbers of reads mapped to individual reference genomes per sample were summarized into a BIOM table. Genomes mapped by less than 0.01% reads per sample were dropped. The beta diversity of samples was assessed using the unweighted UniFrac metric as implemented in QIIME(Caporaso et al. 2010), with reference to the phylogenetic tree of the microbial genomes (also available at: https://github.com/biocore/wol). The resulting distance matrix was visualized with PCoA, and the hypothesis was tested using PERMANOVA and Adonis as implemented in QIIME(Caporaso et al. 2010).

Generating copy numbers of metagenomic genes

The program Salmon (Patro et al. 2017) was applied to determine the reads present for each gene from the shared reference library described above. First, an index was created with Salmon inputting the shared reference library's fasta file. Next, reads were aligned to this index in quasi-mapping mode for each of the 40 metagenomic samples. The results were represented in counts per million sequences, with missing values padded as zeroes.

Serum collection, depletion and analysis

Seppro human depletion kits were used according to manufacturer protocols for depletion of highly abundant proteins. After thawing samples on ice, 14 uL of serum was applied to columns following the depletion protocol, and the wash and elution fractions were combined to increase the total protein content. After depletion, protein was processed as described below, with the exception of a TCA precipitation(Koontz 2014) being used in place of chloroform methanol extraction. After data collection and processing, large variability was observed dependent on serum coloring, and 7 samples with study identifiers L7, L15, L13, L8, L18, L6 and H17 (which were colored red likely because of the presence of blood in the serum) were removed for PCoA visualization.

Protein preparation

Fecal samples were measured out to ~0.5 g and suspended in 5 mL of ice-cold, sterile TBS. Samples were vortexed until completely suspended. Two 20 μM vacuum, steriflip (Milipore) filters were used per sample to remove particulate. Cells were pelleted through centrifugation at 4000 rpm for 10 min at 4 ºC. Next, cells were lysed in 2 mL of buffer containing 75 mM NaCl (Sigma), 3% sodium dodecyl sulfate (SDS, Fisher), 1 mM NaF (Sigma), 1 mM beta-glycerophosphate (Sigma), 1 mM sodium orhtovanadate (Sigma), 10 mM sodium pyrophosphate (Sigma), 1 mM phenylmethylsulfonyl fluoride (PMSF, Sigma), and 1X Complete Mini EDTA-free protease inhibitors (Roche) in 50 mM HEPES (Sigma), pH 8.5(Villen and Gygi 2008). An equal volume of 8M Urea in 50 mM HEPES, pH 8.5 was added to each sample. Cell lysis was achieved through two 15-second intervals of probe sonication at 25% amplitude. Proteins were then reduced with dithiothreitol (DTT, Sigma), alkylated through iodoacetamide (Sigma), and quenched as previously described(Haas et al. 2006). Proteins were next precipitated via chloroform-methanol precipitation and protein pellets were dried(Wessel and Flugge 1984). Protein pellets were re-suspended in 1M urea in 50 mM HEPES, pH 8.5 and digested overnight at room temperature with LysC (Wako)(Van Rechem et al. 2015). A second, 6-hour digestion using trypsin at 37 ºC was performed and the reaction was stopped through addition of 10% trifluoroacetic acid (TFA, Pierce). Samples were then desalted through C18 Sep-Paks (Waters) and eluted with a 40% and 80% Acetonitrile solution containing 0.5% Acetic Acid(Tolonen 2014). Concentration of desalted peptides was determined with a BCA assay (Thermo Scientific). 50 μg aliquots of each sample were dried in a speed-vac. Additionally bridge channels consisting of 25 μg from each sample were created and a 50 μg aliquots of this solution were used in duplicate per Tandem Mass Tag (TMT) 10 plex MS experiment as previously described(Lapek et al. 2018). These bridge channels were used to control for labeling efficiency, inter-run variation, mixing errors and the heterogeneity present in each sample(Tolonen et al. 2011). Each sample or bridge channel was resuspended in 30% dry acetonitrile in 200 mM HEPES, pH 8.5 for TMT labeling with 7 μL of the appropriate TMT reagent(Thompson et al. 2003). Reagents 126 and 131 (Thermo Scientific) were used to bridge between mass spec runs. Remaining reagents were used to label samples in random order. Labeling was carried out for 1 hour at room temperature, and quenched by adding 8 μL of 5% hydroxylamine (Sigma). Labeled samples were acidified by adding 50 μL of 1% TFA. After TMT labeling each 10-plex experiment was combined and desalted through C18 Sep-Paks and dried in a speed-vac.

Generation and processing of proteomic data through LC- LC-MS^2/MS^3

Basic pH reverse-phase liquid chromatography (LC) followed by data acquisition through LC-MS^2/MS^3 was performed as previously described(Lapek et al. 2018). Briefly, 60-minute linear gradients of acetonitrile were performed on C18 columns using an Ultimate 3000 HPLC (Thermo Scientific). Subsequently, 96 fractions were combined as previously described(Wang et al. 2011), and further separation of fractions was performed with an in-line Easy-nLC 1000 (Thermo Fisher Scientific) and a chilled autosampler. LC-MS^2/MS^3 data acquisition and separation setting were as previously defined(Lapek et al. 2017).

Data was processed using Proteome Discoverer 2.1 (Thermo Fisher Scientific). MS^2 data was searched against the shared metagenomic database and Uniprot Human database (uniprot.org, accessed 5/11/2017). The Sequest searching algorithm(Eng et al. 1994) was used to align spectra to database peptides. A precursor mass tolerance of 50 parts per million (ppm)(Beausoleil et al. 2006; Huttlin et al. 2010) was specified and 0.6 Da tolerance for MS^2 fragments. Included in the search parameters was static modification of TMT 10-plex tags on lysine and peptide n-termini (+229.162932 Da), carbamidomethylation of cysteines (+57.02146 Da), and variable oxidation of methionine (+15.99492 Da). Raw data was searched at a peptide and protein false discovery rate of 1% using a reverse database search strategy(Peng et al. 2003; Elias et al. 2005; Elias and Gygi 2007).

TMT reporter ion intensities were extracted from MS^3 spectra for quantitative analysis and signal-to-noise values were used for quantitation. Additional stringent filtering was used removing any moderate confidence peptide spectral matches (PSMs), or ambiguous PSM assignments. Additionally, any peptides with a spectral interference above 25% were removed, as well as any peptides with an average signal to noise ratio less than 10. As metaproteome data contains a high degree of homology between proteins, several decisions were made to reduce false assignments for the metaproteome dataset. The standardized methods in Proteome Discoverer (Version 2.1) preferentially assign peptides to proteins that previously had peptides reported. If this does not resolve the assignment, the peptide is assigned to the longest protein. After the first search, all proteins reported in forward or reverse datasets were filtered into a smaller database for a second search as previously described(Zhang et al. 2016). This method effectively decreased the search space from a database of 748 mb to 21.8 mb. Any PSMs assigned to proteins from the reverse databases were removed. Additionally, a duplicate peptide filter was performed according to the Proteome Discoverer report. Relative abundances are normalized first to the pooled standards for each protein and then to the median signal across the pooled standard. An average of these normalizations was used for the next step. To account for slight differences in amounts of protein labeled, these values were then normalized to the median of the entire dataset and reported as final normalized summed signal-to-noise ratios per protein per sample. When indicated, a lowest common ancestor approach was used for taxonomic bar plots accounting for only peptides unique to a particular taxa(Mesuere et al. 2015).

Metabolite extraction and LC-MS2

Metabolites were extracted by adding a 1:5 weight to volume solution of 70% methanol infused with a 5 μM internal standard sulfamethoxine. The samples were briefly vortexed to mix and stored at 4°C overnight. Extracts were then centrifuged at 4000 rpm for 5 minutes to pellet particulate matter and the supernatant was removed for MS analysis. The extracts were diluted 1:4 in a 96 well plate in pure methanol prior to injection.

LC-MS/MS was performed on a Bruker Daltonics® Maxis qTOF mass spectrometer (Bruker, Billerica, MA USA) with a ThermoScientific UltraMate 3000 Dionex UPLC (Fisher Scientific, Waltham, MA USA). Metabolites were separated using a Kinetex 2.6 µm C18 (30 x 2.10 mm) UPLC column with a guard column. Mobile phases were A 98:2 and B 2:98 ratio of water and acetonitrile containing 0.1% formic acid and a linear gradient from 0 to 100% for a total run time of 840 s at a flow rate of 0.5 mL min^{-1} were used. The mass spectrometer was calibrated daily using Tuning Mix ES-TOF (Agilent Technologies) at a 3 mL min^{-1} flow rate. For accurate mass measurements, lock mass internal calibration used a wick saturated with hexakis (1H,1H,3H-tetrafluoropropoxy) phosphazene ions (Synquest Laboratories, *m/z* 922.0098) located within the source. Full scan MS spectra (*m/z* 50 – 2000) were acquired in the qTOF and the top ten most intense ions in a particular scan were fragmented using collision induced dissociation at 35 eV for +1 ions and 25 eV for +2 ions in the collision cell. Data dependent automatic exclusion protocol was used so that an ion was fragmented when it was first detected, then twice more, but not again unless its intensity was 2.5x the first fragmentation. This exclusion method was cyclical, being restarted after every 30 seconds.

Metabolite annotation through GNPS

Data was converted to the .mzXML format using the Bruker Data Analysis software and uploaded to GNPS through the MassIVE server under ID MSV000082457. Molecular networking was performed as follows: precursor and fragment ion mass tolerance 0.03 Da, minimum cosine score of 0.65, minimum matched fragment ions of 4, and minimum cluster size of 2. GNPS library searching was performed with the same minimum matched peaks and cosine score. All library hits were inspected for quality with the mirror plot feature in GNPS. Area under the curve feature abundances were calculated to produce a metabolome buckettable with the mzMine software. Parameters were as follows: MS1 noise level of 5000 counts, MS2 noise level of 150 counts, m/z tolerance of 0.03 Da for chromatogram building with a minimum time span of 0.1 min, isobaric peaks were deconvoluted with a minimum height of 5000 counts and using the base-line cutoff algorithm, peaks were deisotoped with the same mass tolerance, a 0.1 min retention time tolerance and a maximum isotopic peak pattern of 4, peaks were aligned with the same mass tolerance and retention time tolerance and filtered for at least 3 peaks in a sample and gap filling was performed to produce the final buckettable for statistical analysis.

Generation of metapeptidome data

LC-MS/MS .mzXML formatted files were loaded into PEAKS Studio 8.5(Zhang et al. 2012) for *de novo* identification and searching against the Uniprot human protein database. *De novo* error tolerance parameters were used according to PEAKS default qTOF settings, 0.1 Da parent mass error tolerance, 0.1 Da fragment mass error tolerance. The search settings included no added restriction enzymes, variable dehydration, Acetylation (N-Term), Oxidation (M), and Ubiquitination. The max variable post-translational modifications per peptide was set to 3. *De novo* sequences were filtered to keep only those with an average local confidence above 85%

For human peptides, Label free quantification was run through PEAKS Studio 8.5(Zhang et al. 2012). A 1% FDR cutoff was used integrating peaks with a 20 ppm mass error tolerance and a 6 min retention time window. Peptides were searched against the human protein database (uniprot.org, accessed 05/11/2017) for identification. Quantification was normalized to the total ion chromatograph.

Meta -omic data analysis

Data analysis was performed in python (version 3.5), and records of the code are available in corresponding Jupyter Notebooks for this project (https://github.com/knightlab-analyses/uc-severity-multiomics). Beta-diversity plots were performed using QIIME 2(Bolyen et al. 2019) (Version 2018.4) using the "qiime diversity core-metrics" command. All ADONIS and PERMANOVA statistical analyses of Beta-diversity was performed using the QIIME 1(Caporaso et al. 2010) "compare categories.py" command.

16S fastq were split, demultiplexed, trimmed to 150 base pairs, demultiplexed and processed through deblur using QIITA(Gonzalez et al. 2018) (Study ID 11549). A denovo phylogenetic tree was formed for 16S data using the reference hits through QIIME 2(Bolyen et al. 2019) (version 2018.4) commands "qiime alignment mafft", "qiime alignment mask", "qiime phylogeny fasttree" and "qiime phylogeny midpoint-root". 16S alpha-diversity was generated using QIIME 2(Bolyen et al. 2019) (Version 2018.4) through the command "qiime diversity core-metrics-phylogenetic". Kruskal-Wallis significance tests for alpha diversity were performed in QIIME 2(Bolyen et al. 2019) (Version 2018.4) using the "qiime diversity alpha-group-significance" command. Linear regressions between alpha diversity scores and quantitative categories were performed using the linregress command from the python package scipy (https://www.scipy.org).

Linear regression of metagenome and metaproteome to partial Mayo were also performed using the linregress command as above. Before performing regression, missing values from the metagenome were padded with zeros and the minimum value per protein was used for missing values in the metaproteome.

The program iceLogo's web application(Colaert et al. 2009) was used for consensus sequence analysis. The first and last amino acids from peptides with an average local confidence over 85% were analyzed against a background using the percentage scoring system. For metapeptidome consensus sequences, all residues from peptides with over 85% average local confidence were used as background. For human consensus sequences, the precompiled Homo sapiens Swiss-Prot database was used.

Random forest regressions were performed using QIIME 2(Bolyen et al. 2019) (Version 2018.11) using the sample-classifier regress-sample command. The test size was set to 0.1. Statistics and importance scores for each feature within the 100 independent analyses were compiled.

Caco-2 transwell studies

Caco-2 cell transwell studies were conducted essentially as previously described(Wang et al. 2005). Briefly, Caco-2 cells (passage number ranging from 14-30) were plated into collagen coated 6.5 mm inserts with 0.4 μm pores (Fisher Scientific). Cells were then cultured for 2.5 weeks prior to bacterial inoculation, changing media every 2 days. A day before inoculation, media was changed to media without antibiotics and when indicated, protease inhibitors were dissolved at a given concentration. TEER was measured prior to inoculation of bacteria, and measurements at each following timepoint referenced the original TEER measurement prior to inoculation. Transwell plates were allowed to equilibrate to room temperature for 20 minutes before each TEER timepoint. CFU estimates were performed through serial dilution of 10 μLs of media from inside of the transwell insert. Mammalian cell culture media consisted of DMEM with L-Glutamine (Corning) with 10% heat-inactivated fetal bovine serum, 100 μM sodium pyruvate (Corning), 0.75% sodium bicarbonate, 1X Insulin-Transferrin-Selenium (Gibco), 238.3 μM HEPES, and 1x Penicillin Streptomycin (Thermo). An antibiotic free version of the media consisted was used during bacterial inoculation containing the same contents with the exception of 2% heat-inactivated fetal bovine serum. A day prior to inoculation media was switched to the antibiotic free version, with or without protease inhibitors at the given concentrations. Protease inhibitors tested included Roche cOmplete EDTA-free protease inhibitor cocktail (Sigma).

Bacteroides strains derived from ATCC were used for *vulgatus*, *fragilis*, *uniformis*, and *ovatus*. *Bacteroides dorei* was derived from the Human Microbiome Project strain #717, *Bacteroides dorei* CL02T00C15. For inoculation, *Bacteroides* cultures were grown overnight in Brain-heart-infusion (BHI) broth supplemented with Vitamin K and Hemin. Cultures were spun down at 8000g, and resuspended in DMEM. Inoculations were performed through normalization by OD600 at an estimated multiplicity of infection of 5.

Confocal microscopy

At the end point of transwell studies (38 hours post bacterial innoculation), cells were fixed and prepared for immunofluorescence as follows. Caco-2 cells were fixed on the transwell membrane at 37 °C for 10 minutes in 1 mL 4% Paraformaldehyde (Thermo) in PHEM (60 mM Piperzine-1,4-bis[2-ethanesulfonic Acid] Monosodium Salt, pH 6.9 [TCI Chemicals], 25 mM HEPES (Tremelling et al.), 10 mM EGTA [Oakwood Chemical], 2 mM MgCl2 x 6H2O (Tremelling et al.)). Cells were next permeabilized for 5 minutes in PHEM with 0.5% Triton X-100 (Fisher) at room temperature. Next, 3x 5-minute washes were performed in PHEM containing 0.1% Triton X-100 at room temperature. Cells were next blocked for 30 minutes in 1 mL AbDil (150 mM NaCl (Tremelling et al.), 20 mM Tris-HCl, pH 7.4 [JT Baker], 0.1% Triton X-100 (Tremelling et al.), 2% Bovine serum albumin [Gemini Bioproducts]) at room temperature. After blocking, primary antibodies for Occludin (Thermo, catalog number 33-1500, 0.5 μg/mL) and ZO-1 (Thermo, catalog number 61-7300, 1.5 μg/mL) were added into AbDil and left in a humidified chamber overnight at 4 °C. Cells were next washed 4x in PHEM containing 0.1% Triton X-100 for 5 minutes at room temperature. After washing, secondary antibodies, Rhodamine Red Donkey Anti-Rabbit (Jackson ImmunoResearch, Code Number 711-295-152), and Alexa Fluor 488 Donkey Anti-Mouse (Jackson ImmunoResearch) were diluted to 3 μg/mL in AbDil containing a 1:1000 dilution of Phalloidin-iFluor 647 (abcam, ab176759) and 1 μg/mL DAPI (Thermo). Secondary antibodies were incubated for 1 hour at room temperature in a humidified chamber. Following secondary antibody incubation, cells were again washed 3x in PHEM containing 0.1% Triton X-100 for 5 minutes at room temperature. Finally, cells were rinsed in PHEM, removed from transwell insert and fixed onto microscope slides for imaging.

Cells were imaged using a Nikon A1R HD confocal with a four-line (405nm, 488nm, 561nm, and 640nm) LUN-V laser engine and DU4 detector using bandpass and longpass filters for each channel (450/50, 525/50, 595/50 and 700/75), mounted on a Nikon Ti2 using an Apo 60x 1.49 NA objective, or a C2 Plus confocal with a similar four-line LUN-4 laser engine and a DUV-B detector operating in virtual bandpass mode. Images stacks were acquired with the galvo scanning mode on both confocals, and Z-steps of 0.2 μm. To avoid cross-talk between channels, Z-stacks were acquired of the

DAPI and Rhodamine Red channels first, and the AlexaFluor 488 and Phalloidin-iFluor 647 channels were acquired subsequently. The laser powers used were 1.5% for the 405 nm laser 2% for the 488 nm laser, 1.5% for the 561 nm laser and 1.5% for the 640 nm laser.

Gnotobiotic mouse fecal transplant studies

Germ-free C57BL/6 IL10-/- male mice (C57BL/6NTac-*Il10*^{em8Tac}; Taconic model GF-16006) were maintained in isolated ventilated cages Isocages (Techniplast, West Chester, PA, USA) (Hecht et al., 2014). At 5-6 weeks of age, mice were orally administered with 200 µL of fecal suspension from a patient with a high disease severity and a high proteolysis activity (H19) or from a patient with a low disease severity and a low proteolysis activity (L3). Transplanted mice were housed in isolated ventilated cages, Isocages and fed autoclaved Purina Rodent Chow # 5021. All mice were housed at Georgia State University (Atlanta, Georgia, USA) under institutionally approved protocols (IACUC # A18006). Mice were then weighted, euthanized, and colon length, colon weight, spleen weight, liver weight, adipose weight and ceacum weight were measured.

Data availability

Metabolomic data, Proteomic data and supplementary files are available online at https://massive.ucsd.edu (study ID MSV000082094). Genomic data is being uploaded through EBI https://www.ebi.ac.uk/ena.

Chapter 4 reflects material of a manuscript as it was reviewed by the journal Nature, in Jan 2020, Robert H. Mills, Parambir S. Dulai, Yoshiki Vázquez-Baeza, Qiyun

Zhu, Greg Humphrey, Lindsay DeRight Goldasich, MacKenzie Bryant, Robert A. Quinn, Andrew T. Gewirtz, Benoit Chassaing, Hiutung Chu, William J. Sandborn, Pieter C. Dorrestein, Rob Knight, and David J. Gonzalez. The dissertation author played a primary role in aspects of the work ranging from the study design, data acquisition, analysis and writing of the manuscript.

4.5 References

Bakir MA, Sakamoto M, Kitahara M, Matsumoto M, Benno Y. 2006. Bacteroides dorei sp. nov., isolated from human faeces. *Int J Syst Evol Microbiol* **56**: 1639-1643.

Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. 2006. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* **24**: 1285-1292.

Bloom SM, Bijanki VN, Nava GM, Sun L, Malvin NP, Donermeyer DL, Dunne WM, Jr., Allen PM, Stappenbeck TS. 2011. Commensal Bacteroides species induce colitis in host-genotype-specific fashion in a mouse model of inflammatory bowel disease. *Cell Host Microbe* **9**: 390-403.

Bolyen E Rideout JR Dillon MR Bokulich NA Abnet CC Al-Ghalith GA Alexander H Alm EJ Arumugam M Asnicar F et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**: 852-857.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59-60.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335-336.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**: 1621-1624.

Cohen J. 1988. *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates, Hillsdale, N.J.

Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K. 2009. Improved visualization of protein consensus sequences by iceLogo. *Nat Methods* **6**: 786-787.

Donaldson GP, Lee SM, Mazmanian SK. 2016. Gut biogeography of the bacterial microbiota. *Nat Rev Microbiol* **14**: 20-32.

Dulai PS, Levesque BG, Feagan BG, D'Haens G, Sandborn WJ. 2015. Assessment of mucosal healing in inflammatory bowel disease: review. *Gastrointest Endosc* **82**: 246-255.

Dulai PS, Siegel CA, Colombel JF, Sandborn WJ, Peyrin-Biroulet L. 2014. Systematic review: Monotherapy with antitumour necrosis factor alpha agents versus combination therapy with an immunosuppressive for IBD. *Gut* **63**: 1843-1853.

Elhenawy W, Debelyy MO, Feldman MF. 2014. Preferential packing of acidic glycosidases and proteases into Bacteroides outer membrane vesicles. *MBio* **5**: e00909-00914.

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**: 207-214.

Elias JE, Haas W, Faherty BK, Gygi SP. 2005. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods* **2**: 667-675.

Eng JK, McCormack AL, Yates JR. 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**: 976-989.

Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, Shah M, Halfvarson J, Tysk C, Henrissat B et al. 2012. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* **7**: e49138.

Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ et al. 2018. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* doi:10.1038/s41564-018-0306-4.

Fumery M, Singh S, Dulai PS, Gower-Rousseau C, Peyrin-Biroulet L, Sandborn WJ. 2018. Natural History of Adult Ulcerative Colitis in Population-based Cohorts: A Systematic Review. *Clin Gastroenterol Hepatol* **16**: 343-356 e343.

Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vazquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB et al. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. *Nat Methods* **15**: 796-798.

Haas W, Faherty BK, Gerber SA, Elias JE, Beausoleil SA, Bakalarski CE, Li X, Villen J, Gygi SP. 2006. Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol Cell Proteomics* **5**: 1326-1337.

Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villen J, Haas W, Sowa ME, Gygi SP. 2010. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**: 1174-1189.

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

Jansson JK, Baker ES. 2016. A multi-omic future for microbiome studies. *Nat Microbiol* **1**: 16049.

Juste C, Kreil DP, Beauvallet C, Guillot A, Vaca S, Carapito C, Mondot S, Sykacek P, Sokol H, Blon F et al. 2014. Bacterial protein signals are associated with Crohn's disease. *Gut* **63**: 1566-1577.

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

Koontz L. 2014. TCA precipitation. *Methods Enzymol* **541**: 3-10.

Kulagina EV, Efimov BA, Maximov PY, Kafarskaia LI, Chaplin AV, Shkoporov AN. 2012. Species Composition of Bacteroidales Order Bacteria in the Feces of Healthy People of Various Ages. *Biosci Biotech Bioch* **76**: 169-171.

Kumagai Y, Konishi K, Gomi T, Yagishita H, Yajima A, Yoshikawa M. 2000. Enzymatic properties of dipeptidyl aminopeptidase IV produced by the periodontal pathogen Porphyromonas gingivalis and its participation in virulence. *Infect Immun* **68**: 716-724.

Lapek JD, Jr., Lewinski MK, Wozniak JM, Guatelli J, Gonzalez DJ. 2017. Quantitative Temporal Viromics of an Inducible HIV-1 Model Yields Insight to Global Host Targets and Phospho-Dynamics Associated with Vpr. *Mol Cell Proteomics* doi:10.1074/mcp.M116.066019.

Lapek JD, Jr., Mills RH, Wozniak JM, Campeau A, Fang RH, Wei X, van de Groep K, Perez-Lopez A, van Sorge NM, Raffatellu M et al. 2018. Defining Host

Responses during Systemic Bacterial Infection through Construction of a Murine Organ Proteome Atlas. *Cell Syst* doi:10.1016/j.cels.2018.04.010.

Lewis JD, Chuai S, Nessel L, Lichtenstein GR, Aberra FN, Ellenberg JH. 2008. Use of the noninvasive components of the Mayo score to assess clinical response in ulcerative colitis. *Inflamm Bowel Dis* **14**: 1660-1666.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**: 1674-1676.

Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T et al. 2014. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* **32**: 834-841.

Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ et al. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**: 655-662.

Marotz C, Amir A, Humphrey G, Gaffney J, Gogul G, Knight R. 2017. DNA extraction for streamlined metagenomics of diverse environmental samples. *Biotechniques* **62**: 290-293.

Martens EC, Koropatkin NM, Smith TJ, Gordon JI. 2009. Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. *J Biol Chem* **284**: 24673-24677.

Mesuere B, Debyser G, Aerts M, Devreese B, Vandamme P, Dawyndt P. 2015. The Unipept metaproteomics analysis pipeline. *Proteomics* **15**: 1437-1442.

Mills RH, Vázquez-Baeza Y, Zhu Q, Jiang L, Gaffney J, Humphrey G, Smarr L, Knight R, Gonzalez DJ. 2019. Evaluating Metagenomic Prediction of the Metaproteome in a 4.5-Year Study of a Patient with Crohn's Disease. *mSystems* **4**: e00337-00318.

Narula N, Alshahrani AA, Yuan Y, Reinisch W, Colombel JF. 2018. Patient-Reported Outcomes and Endoscopic Appearance of Ulcerative Colitis: A Systematic Review and Meta-Analysis. *Clin Gastroenterol Hepatol* doi:10.1016/j.cgh.2018.06.015.

O'Donoghue AJ, Jin Y, Knudsen GM, Perera NC, Jenne DE, Murphy JE, Craik CS, Hermiston TW. 2013. Global substrate profiling of proteases in human neutrophil extracellular traps reveals consensus motif predominantly contributed by elastase. *PLoS One* **8**: e75141.

Ordas I, Eckmann L, Talamini M, Baumgart DC, Sandborn WJ. 2012. Ulcerative colitis. *Lancet* **380**: 1606-1619.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**: 417-419.

Pauer H, Ferreira Ede O, dos Santos-Filho J, Portela MB, Zingali RB, Soares RM, Domingues RM. 2009. A TonB-dependent outer membrane protein as a Bacteroides fragilis fibronectin-binding molecule. *FEMS Immunol Med Microbiol* **55**: 388-395.

Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. 2003. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* **2**: 43-50.

Reeves AR, D'Elia JN, Frias J, Salyers AA. 1996. A Bacteroides thetaiotaomicron outer membrane protein that is essential for utilization of maltooligosaccharides and starch. *J Bacteriol* **178**: 823-830.

Sartor RB, Wu GD. 2017. Roles for Intestinal Bacteria, Viruses, and Fungi in Pathogenesis of Inflammatory Bowel Diseases and Therapeutic Approaches. *Gastroenterology* **152**: 327-339 e324.

Schirmer M, Denson L, Vlamakis H, Franzosa EA, Thomas S, Gotman NM, Rufo P, Baker SS, Sauer C, Markowitz J et al. 2018. Compositional and Temporal Changes in the Gut Microbiome of Pediatric Ulcerative Colitis Patients Are Linked to Disease Course. *Cell Host Microbe* **24**: 600-610 e604.

Shen ZH, Zhu CX, Quan YS, Yang ZY, Wu S, Luo WW, Tan B, Wang XY. 2018. Relationship between intestinal microbiota and ulcerative colitis: Mechanisms and clinical application of probiotics and fecal microbiota transplantation. *World J Gastroentero* **24**: 5-14.

Shimshoni E, Yablecovitch D, Baram L, Dotan I, Sagi I. 2015. ECM remodelling in IBD: innocent bystander or partner in crime? The emerging role of extracellular molecular events in sustaining intestinal inflammation. *Gut* **64**: 367-372.

Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermudez-Humaran LG, Gratadoux JJ, Blugeon S, Bridonneau C, Furet JP, Corthier G et al. 2008. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci U S A* **105**: 16731-16736.

Steck N, Mueller K, Schemann M, Haller D. 2012. Bacterial proteases in IBD and IBS. *Gut* **61**: 1610-1618.

Strygler B, Nicar MJ, Santangelo WC, Porter JL, Fordtran JS. 1990. Alpha 1-antitrypsin excretion in stool in normal subjects and in patients with gastrointestinal disorders. *Gastroenterology* **99**: 1380-1387.

Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C. 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**: 1895-1904.

Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G et al. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**: 457-463.

Tolonen AC, and Haas, W. 2014. Quantitative Proteomics Using Reductive Dimethylation for Stable Isotope Labeling. *Jove-J Vis Exp*.

Tolonen AC, Haas W, Chilaka AC, Aach J, Gygi SP, Church GM. 2011. Proteome-wide systems analysis of a cellulosic biofuel-producing microbe. *Mol Syst Biol* **7**: 461.

Tremelling M, Cummings F, Fisher SA, Mansfield J, Gwilliam R, Keniry A, Nimmo ER, Drummond H, Onnie CM, Prescott NJ et al. 2007. IL23R variation determines susceptibility but not disease phenotype in inflammatory bowel disease. *Gastroenterology* **132**: 1657-1664.

Van Rechem C, Black JC, Boukhali M, Aryee MJ, Graslund S, Haas W, Benes CH, Whetstine JR. 2015. Lysine Demethylase KDM4A Associates with Translation Machinery and Regulates Protein Synthesis. *Cancer Discov* **5**: 255-263.

Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, Lefsrud MG, Apajalahti J, Tysk C, Hettich RL et al. 2009. Shotgun metaproteomics of the human distal gut microbiota. *ISME J* **3**: 179-189.

Vergnolle N. 2016. Protease inhibition as new therapeutic strategy for GI diseases. *Gut* **65**: 1215-1224.

Vich Vila A, Imhann F, Collij V, Jankipersadsing SA, Gurry T, Mujagic Z, Kurilshikov A, Bonder MJ, Jiang X, Tigchelaar EF et al. 2018. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci Transl Med* **10**.

Villen J, Gygi SP. 2008. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat Protoc* **3**: 1630-1638.

Wang F, Graham WV, Wang Y, Witkowski ED, Schwarz BT, Turner JR. 2005. Interferon-gamma and tumor necrosis factor-alpha synergize to induce intestinal epithelial barrier dysfunction by up-regulating myosin light chain kinase expression. *Am J Pathol* **166**: 409-419.

Wang Y, Yang F, Gritsenko MA, Wang Y, Clauss T, Liu T, Shen Y, Monroe ME, Lopez-Ferrer D, Reno T et al. 2011. Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11**: 2019-2026.

Wei B, Dalwadi H, Gordon LK, Landers C, Bruckner D, Targan SR, Braun J. 2001. Molecular cloning of a Bacteroides caccae TonB-linked outer membrane protein identified by an inflammatory bowel disease marker antibody. *Infect Immun* **69**: 6044-6054.

Wessel D, Flugge UI. 1984. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal Biochem* **138**: 141-143.

Wexler HM. 2007. Bacteroides: the good, the bad, and the nitty-gritty. *Clin Microbiol Rev* **20**: 593-621.

Xiao Y, Hsiao TH, Suresh U, Chen HI, Wu X, Wolf SE, Chen Y. 2014. A novel significance score for gene selection and ranking. *Bioinformatics* **30**: 801-807.

Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B. 2012. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* **11**: M111 010587.

Zhang X, Deeke SA, Ning ZB, Starr AE, Butcher J, Li J, Mayne J, Cheng K, Liao B, Li LY et al. 2018. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nature Communications* **9**.

Zhang X, Ning ZB, Mayne J, Moore JI, Li J, Butcher J, Deeke SA, Chen R, Chiang CK, Wen M et al. 2016. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **4**.

Zhou Y, Zhi F. 2016. Lower Level of Bacteroides in the Gut Microbiota Is Associated with Inflammatory Bowel Disease: A Meta-Analysis. *Biomed Res Int* **2016**: 5828959.

Chapter 5

Associations of the fecal microbial proteome composition and proneness to diet-induced obesity

5.1 Abstract

Consumption of refined high-fat, low-fiber diets promotes development of obesity and its associated consequences. While genetics play an important role in dictating susceptibility to such obesogenic diets, mice with nearly uniform genetics exhibit marked heterogeneity in their extent of obesity in response to such diets. This suggests non-genetic determinants play a role in diet-induced obesity. Hence, we sought to identify parameters that predict, and/or correlate with, development of obesity in response to an obesogenic diet. We assayed behavior, metabolic parameters, inflammatory markers/cytokines, microbiota composition, and the fecal metaproteome, in a cohort of mice (n=50) prior to, and the 8 weeks following, administration of an obesogenic high-fat low-fiber diet. Neither behavioral testing nor quantitation of inflammatory markers broadly predicted severity of diet-induced obesity. Although, the small subset of mice that exhibited basal elevations in serum IL-6 (n=5) were among the more obese mice in the cohort. While fecal microbiota composition changed markedly in response to the obesogenic diet, it lacked the ability to predict which mice were relatively prone or resistant to obesity. In contrast, fecal metaproteome analysis revealed functional and taxonomic differences among the proteins associated with proneness to obesity. Targeted interrogation of microbiota composition data successfully validated the taxonomic differences seen in the metaproteome. While future work will be needed to determine the breadth of applicability of these associations to other cohorts of animals and humans, this study nonetheless highlights the potential power of gut microbial proteins to predict and perhaps impact development of obesity.

5.2 Introduction

Obesity is an emerging 21^{st} century epidemic. Obesity, and the disease states it drives, including type 2 diabetes, cardiovascular disease, and liver disease threaten to overwhelm healthcare systems(Apovian 2016). Thus, obesity is a contemporary medical concern that poses a grave public health crisis in dire need of a solution. The increased incidence in obesity is thought to have been driven by broad societal changes that have resulted in reduced physical activity and increased availability of palatable low-cost energy-rich foods(Stelmach-Mardas et al. 2016). Yet the extent to which individuals develop obesity in such an environment is highly heterogeneous. Variations in individual genetics contribute to, but are insufficient to fully explain, such heterogeneity. For example, studies characterizing weight-discordant monozygotic twins has shown that individuals with shared environmental, physical activity, and genetic factors display heterogeneity in adiposity(Naukkarinen et al. 2014). Similarly, rat-based studies show marked heterogeneity in weight gain and adiposity in response to obesogenic diets even when using highly inbred animals in a well-controlled environment(Archer et al. 2003; de La Serre et al. 2010). Better understanding non-genetic factors that influence proneness to obesity might aid the identification of individuals at-risk for development of obesity and can yield modifiable factors to ameliorate this disease state.

A number of factors that are at least partially independent of genetics are proposed to influence proneness to diet-induced obesity (DIO). One potential central nexus of such factors is inflammation, impacting metabolic signaling pathways including insulin and leptin(Hotamisligil 2017), which have well-established roles in feeding behavior. Inflammation is also suggested to promote behavioral patterns such as anxiety-like and anti-social behaviors that can impact food consumption(Jeon and Kim 2018). While numerous elements impact inflammation, one increasingly appreciated factor is the gut microbiota(Ley et al. 2006; Turnbaugh et al. 2006; Turnbaugh et al. 2008; Turnbaugh et al. 2009; Turnbaugh 2017), which is the collective term for the large diverse community of microorganisms that inhabit the gastrointestinal tract. Indeed, in humans, gut microbiota composition is associated with obesity. One way microbiota composition influences metabolic signaling is via lipopolysaccharide (LPS), which activates pro-inflammatory signaling via Toll-like receptor 4 (TLR4) resulting in production of molecules including tumor necrosis factor alpha (TNF-α), and interleukin-6 (IL-6). These molecules interfere with leptin and insulin signaling, wherein LPS derived from gamma-proteobacteria is a particularly potent activator of TLR4(Aygun et al. 2005). Another host-microbiota interaction implicated in inflammation and obesity is the sensing of flagella through TLR5, which keeps motile bacteria in-check by a range of mechanisms including production of antimicrobial peptides and promoting production of anti-flagella immunoglobulins that help regulate the microbiota in the healthy gut(Cullender et al. 2013). In addition to its impacts on inflammation, microbiota composition is also reported to influence energy harvest from ingested food(Turnbaugh et al. 2006; El Kaoutari et al. 2013). Hence, in light of its ability to impact inflammation, metabolism, and behavior, gut microbiota composition might provide a means of identifying host proneness to obesity when presented with an obesogenic diet.

Here, we sought to identify microbiota-based markers that might predict proneness to diet-induced obesity, specifically exposing mice to a western-style, low-fiber high fat diet (HFD). Both targeted and untargeted approaches were utilized including 16S rRNA gene amplicon sequencing for microbial community profiling and a Tandem Mass Tag (TMT) based multiplexed mass spectrometry (MS) approach for analysis of the fecal metaproteome. Additionally, we measured behavior, inflammatory markers, and metabolic parameters. Notably, we show that the fecal metaproteome appears to be a promising candidate for distinguishing mice with differential responses to obesogenic diets. Collectively, this study provides insight into potential mechanisms regarding the host-microbiota interactions mediating response to HFD exposure, and highlights putative biomarkers for predicting DIO.

5.3 Results

Stratification and characterization of mice prone, or resistant, to HFD-induced metabolic syndrome

The primary goal of this study was to elucidate factors that predict and possibly govern, susceptibility to developing obesity in response to administration of an obesogenic diet. Hence, we designed a prospective study wherein 50, 3-week old female C57BL/6 mice, housed 5 mice per cage, were subjected to metabolic monitoring, including behavior analysis and sample collection over a 3-week period. During this time, the mice were fed standard grain-based chow (GBC), which is comprised of relatively unrefined ingredients. The cohort of mice was then switched to a diet compositionally low in fiber (5%) and high in fat (35% by mass, 60% by calories), herein referred to as an obesogenic diet or high-fat diet (HFD) for an 8-week period. Prior to, during and after administration of the high-fat diet, sample collection and monitoring was performed as outlined in Fig. 5.1A. In accord with our previous rodent-based studies(de La Serre et al.

2010), the extent of obesity following the obesogenic diet was quite heterogeneous with many mice weighing between 20-25 grams, which is the approximate weight of age-matched GBC-fed mice of this strain/gender. In contrast, some mice appeared to dramatically gain weight over the course of the experiment with final weights over 30 grams. Therefore, based on their final body weight, mice were stratified into tertiles as being prone, intermediate, or relatively resistant to being obese following exposure to an obesogenic diet (Fig. 5.1B). First, we examined if mice prone or resistant to DIO clustered within cages but did not observe a distribution pattern to support this possibility (Fig. 5.1C). Nor were these groupings significantly related to the initial weight of the mice (Fig. 5.1D). The total weight gain over the period of exposure to the obesogenic diet for resistant mice was about 40%. This observation is approximately the expected age-related weight gain of GBC-fed mice during this period, while prone mice increased in weight by about 70% during this period (Fig. 5.1E). Fat mass, as determined by magnetic resonance imaging (MRI), prior to, during, or at the end of exposure to HFD, was highly correlated with body weight within the cohort (Fig. 5.1F). Accordingly, post-euthanasia weight of the periovarian fat pad, which has classically been used to assess adiposity in mice correlated closely (r^2 = 0.8229) with final body weights confirming our stratifications reflected degree of adiposity (Fig. 5.1G). Final body weights were also correlated with fasting glucose concentration (r^2 = 0.2218), suggesting mice that were prone to diet-induced obesity were also prone to its downstream consequences (Fig. 5.1H). Lastly, in light of the appreciation that low-grade intestinal inflammation can promote adiposity and its consequences, we measured weight/length ratio of the colon(Vijay-Kumar et al. 2010; Carvalho et al. 2012). This measurement was also

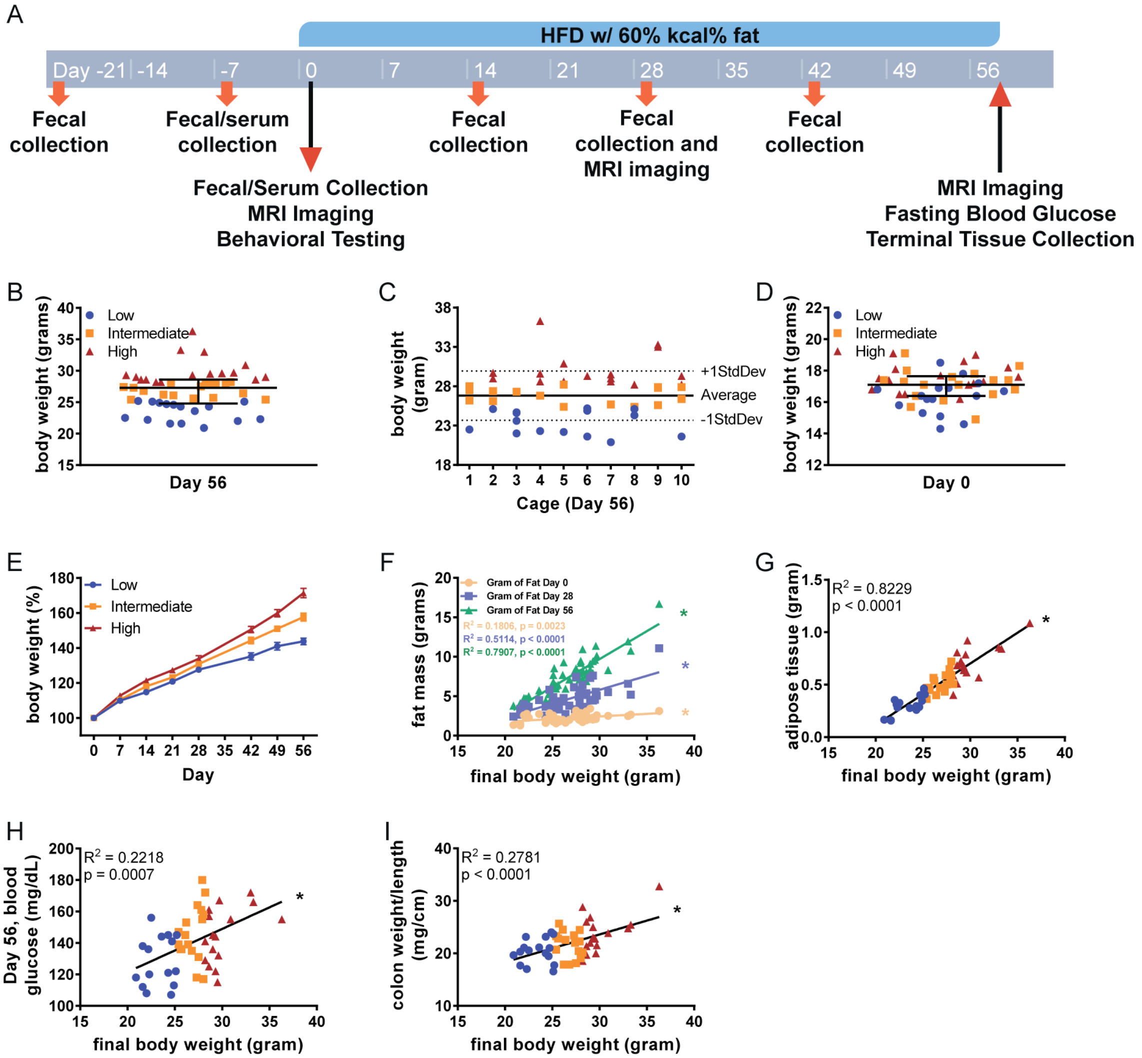


Figure 5.1 Stratification and characterization of mice prone, or resistant, to HFD-induced metabolic syndrome. (**A**) 3-5-week old, female C57BL/6 mice were purchased from The Jackson Laboratory and housed for three weeks before high-fat diet administration in order to favor microbiota stabilization. Subsequently, animals were treated with high-fat diet (60% kcal from fat) for 8 weeks. Serum collection occurred on days -7, 0, and 56. Body weight measurements occurred prior to every flagellin administration. Fecal collection occurred on days -21 and -7, then every other week starting on day 0. (**B**) Mice were identified as low, intermediate, or high responders based on if their final body weight fell within the first, second, or third tertile, respectively. (**C**) Final body weights of mice by cage. (**D**) Initial weights of mice. (**E**) Body weights were measured weekly and expressed as relative values, day 0 (pre high-fat diet treatment) being defined as 100%. Final body weights were correlated to (**F**) fat mass by MRI, (**G**) epididymal adipose weight, (**H**) day 56 fasting blood glucose, and (**I**) colon weight/length ratio. Data are the means +/- S.E.M. (*N*=50). Significance was determined using linear regression analysis (**p*≤0.05).

correlated with final body weight ($r^2 = 0.2781$; Fig. 5.1I), supporting the notion that the obese mice were in a state of low-grade gut inflammation.

Associations of inflammatory markers/mediators and proneness to obesity.

Low-grade inflammation is reported to associate with, and promote obesity(Cani et al. 2007; Vijay-Kumar et al. 2010). Accordingly, we investigated levels of pro-inflammatory mediators to determine if they might mark mice that would be prone to becoming obese following exposure to an obesogenic diet. Hence, we measured levels of fecal lipocalin-2 (Lcn-2), which is a broadly dynamic marker of gut inflammation(Chassaing et al. 2012). Levels of fecal Lcn-2 did not correlate with final body weights when measured 14 days prior (r^2 = 0.0156) to exposure or 4 weeks after the initiation of the diet (r^2 = 0.0074; Fig. 5.2A, B). Additionally, the levels of serum pro-inflammatory cytokines CXCL1 and IL-6 when measured 7 days prior to administration of the obesogenic diet were also not correlated to final body weight (r^2 = 0.0177, 0.022 respectively, Fig. 5.2C, D). However, 4 of the 5 mice that displayed detectable serum IL-6 at this time point were in the top tertile of obesity following the diet suggesting the subset of mice displaying this parameter might be more prone to DIO. To further investigate this subset of mice, we tested for differences within various parameters associated with diet-induced obesity between the subsets of mice with or without detectable IL-6. At day -7, several of these parameters were consistent with the possibility that detection of IL-6 can discriminate high or low responders, but ultimately, none reached statistical significance. Nevertheless, such elevations in IL-6 were not maintained when assayed after 8-weeks of diet (Fig. 5.2E). Other findings supporting the notion that obesity is associated with low-grade inflammation included a modest correlation after 8-weeks of diet between body weights and CXCL1, which is a chemokine expressed by many cell types and often used as a general serum marker of

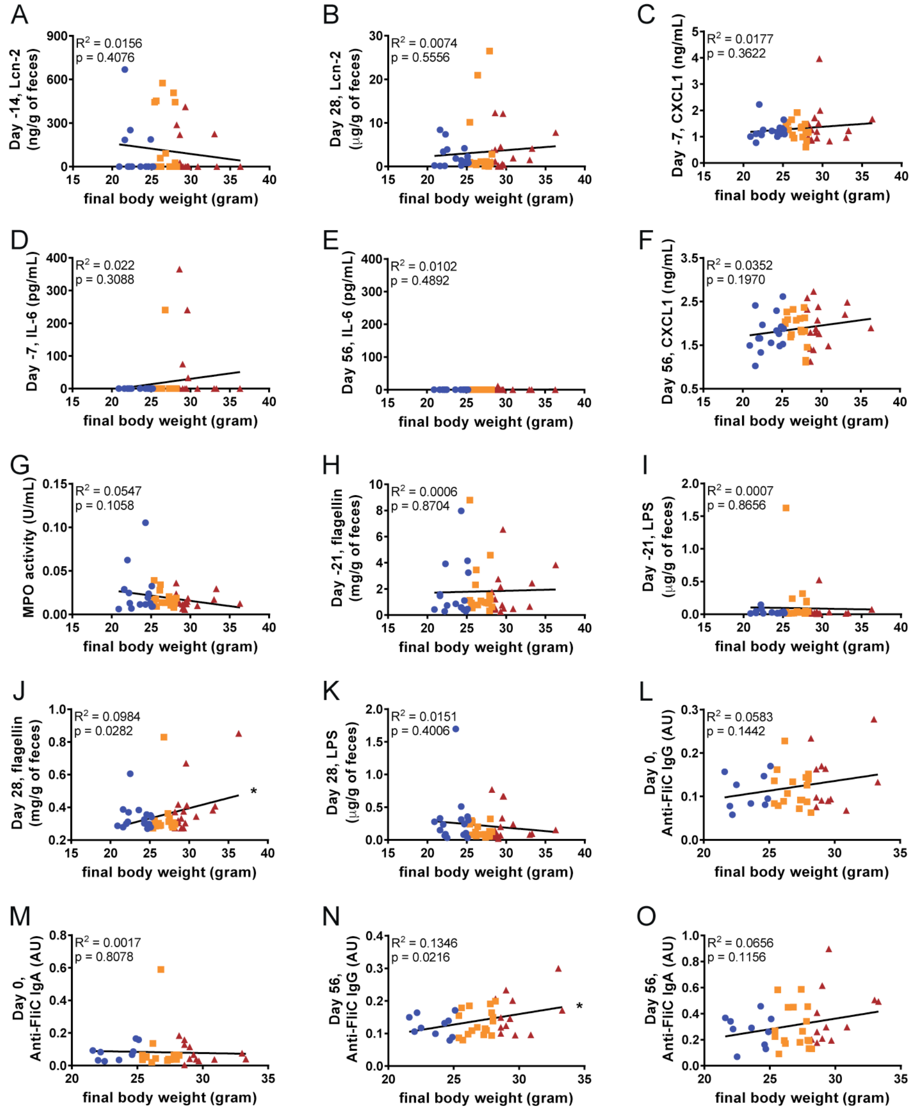


Figure 5.2 Associations of inflammatory markers/mediators and proneness to obesity. Final body weights correlated to fecal lipocalin-2 at (**A**) day -14 and (**B**) 28, analyzed using ELISA kits. Additionally, final body weights were correlated to serum cytokines CXCL1 and IL-6 at (**C-D**) day -7 and (**E-F**) day 56, analyzed using ELISA kits. Final body weights were also correlated to (**G**) colonic myeloperoxidase levels, as well as, fecal flagellin and lipopolysaccharide at (**H-I**) day -21 and (**J-K**) day 28 using HEK 293 cells expressing mTLR5 or mTLR4 measuring bioactive flagellin and lipopolysaccharide, respectively. Serum anti-flagellin IgG and IgA were also quantified using ELISA techniques at days 0 (**L-M**) and 56 (**N-O**). (*N*=50). Significance was determined using linear regression analysis (**p*≤0.005).

low-grade inflammation (r^2 = 0.0352, Fig. 5.2F). In contrast, there was no correlation between final body weights and levels of intestinal myeloperoxidase (MPO), which is a widely used marker of classic inflammation in the intestine(Masoodi et al. 2012) (Fig. 5.2G).

Gut bacterial components, flagellin and lipopolysaccharide (LPS), are well known for their inflammatory properties(Hayashi et al. 2001; Jones et al. 2001). Fecal levels of each were measured 3 weeks prior to administration of the obesogenic diet with neither correlating with final body weight (Fig. 5.2H, I). However, flagellin, but not LPS, were modestly correlated with final body weight when measured 4 weeks following initiation of the obesogenic diet (r^2 = 0.0984, 0.0151 respectively, Fig. 5.2J, K). Moreover, there was a correlation in levels of anti-flagellin antibodies at the time of diet administration (for IgG but not IgA) and 8 weeks following exposure to obesogenic diet (Fig. 5.2L-O). Levels of anti-flagellin antibodies likely reflect exposure of the immune system to this molecule, which can be influenced by both levels of flagellin in the gut, bacterial-epithelial distance, and intestinal permeability(Sanders et al. 2006; Ziegler et al. 2008). Together, these studies did not reveal a reliable predictive marker of proneness to diet-induced obesity but suggest exposure to bacterial products, such as flagellin, might have some predictive power.

Quantitative measures of behavior did not predict obesity proneness

The gut-brain axis is increasingly appreciated to play a role in the pathogenesis of many neurological and metabolic diseases(Dinan and Cryan 2017). Hence, we investigated the extent to which certain behavioral parameters are able to predict proneness to weight gain. Compulsive behavior and activity level were measured in a home cage behavior test, and time spent digging, time spent grooming, and total distance travelled were quantified. Additionally, anxiety-like behavior was assessed using the open field test, represented by time spent in the center zone and distance travelled in the center of the open field arena. Ultimately, none of these measures had a significant ability to predict extent of obesity in response to the obesogenic diet.

Impact of DIO on fecal metaproteome.

We next turned to a contemporary metaproteomics approach to study the fecal protein composition of our cohort of mice. While administration of an obesogenic diet is well known to rapidly alter gut microbiota species composition(Hildebrandt et al. 2009), whether it might also impact the fecal metaproteome, let alone whether the fecal metaproteome might predict responsiveness to such a diet, has not been described. While metaproteomic analysis presents the challenges of discriminating host and bacterial proteins from potentially millions of proteins, the field is an area of rapid growth currently developing standard methodology(Zhang and Figeys 2019).

To this end, we applied our recently developed TMT-based metaproteomic methods(Mills et al. 2019), in combination with a two-step database search method(Zhang et al. 2016) on feces from mice that developed the highest and lowest degree of obesity (n= 4 mice per condition). Our analysis included specimens collected before (day 0) and after 56 days of exposure to the obesogenic diet. The final data included quantitation of 13,975 proteins of which 1,108 were derived from mice.

For a broad scale perspective of the data, an unsupervised Principle Coordinates analysis of the metaproteome data using the Bray-Curtis distance metric exhibited clear separation of samples before and after diet administration, reflecting a dramatic impact of the obesogenic diet on the overall fecal metaproteome (Fig. 5.3A). This analysis also exhibited clustering at the 56 day time point discriminating high and low response to diet (Permanova pseudo-F = 1.99, p = 0.058). Using K-means clustering, we identified 6 protein clusters, some of which were associated with increased representation of particular taxa and functions (Fig. 5.3B). These groupings include Group 4, which appeared to show an increased presence of Clostridiales and lipid transport and metabolism proteins in high responder mice after exposure to HFD (Fig. 5.3B). These groupings provide putative taxonomic associations to the functional differences observed before and after administration of HFD.

Comparing all samples before and after HFD exposure made evident that there were widespread changes in the fecal metaproteome. By using a statistical ranking method accounting for both fold change and t-test p-values, π-score (Xiao et al. 2014), we observed that 58% (3670/6311) of proteins displayed a high level of association to diet exposure ($|\pi| > 1$, Fig. 5.3C). The proteins associated with the dietary intervention contained large differences in their taxonomic and functional annotations. Taxonomic differences included a larger portion of proteins from Clostridiales and Bacteroidales before HFD exposure while a large portion (~40%) of proteins enriched after 8-weeks exposure were derived from Lactobacillales (Fig. 5.3D).

Functional categorization of the proteins associated to the dietary intervention was performed through the Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (eggNOG) database. These studies revealed a very strong association between proteins related to motility and HFD exposure as expression levels of 141 proteins were reduced following exposure to HFD with only 3 proteins increased after HFD, resulting in a 32-fold difference (Fig. 5.3E). In accord, we note that, on average, levels of fecal flagellin decreased by about 5-fold when measured 21 days preceding or 4 weeks following administration of the obesogenic diet (Fig. 5.2H, J). Further, when subsetting all 680 flagellin proteins from the metaproteome dataset, we observed statistically

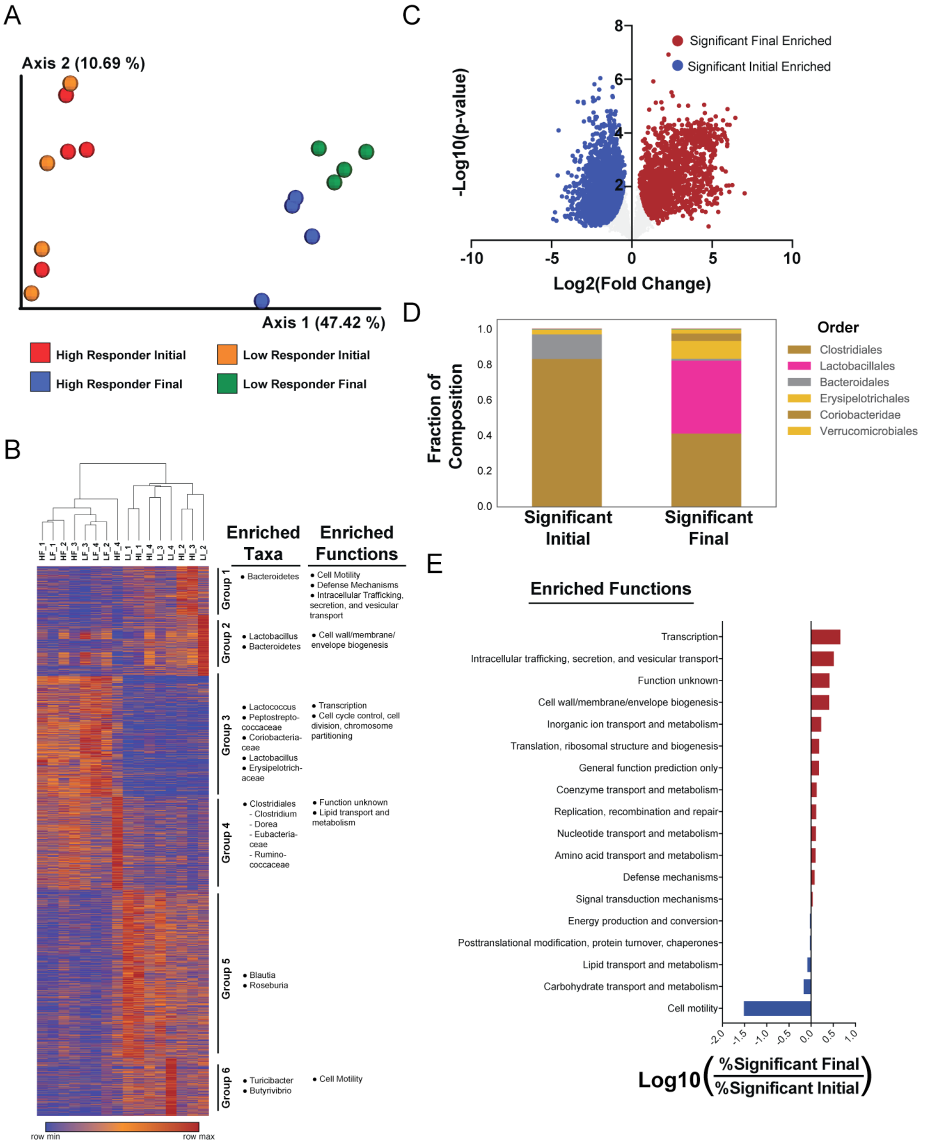


Figure 5.3 Impact of DIO on fecal metaproteome. (**A**) Principal coordinates analysis (PCoA) of metaproteome data using the Bray-Curtis distance metric (**B**) Protein relative abundance heatmap. Samples are clustered by 1-Pearson correlation and proteins are grouped using KMeans clustering. Relative abundances per protein are colored on a spectrum with red as row maxima and blue as row minima. Functional and taxonomic bias within each KMeans cluster is displayed on the right. (**C**) Volcano plot of metaproteome response to HFD. Fold change and t-test significance of each protein are plotted. Overall significance was set at |π-score| > 1. (**D**) Taxonomic composition of significant proteins. (**E**) Functional bias in significant proteins. Compositions of eggNOG annotations between proteins enriched in the final and initial time points were compared and the log ratio of high abundance categories (>10 proteins) is shown. Sample names in (**A**-**B**) are annotated H for High Responder, L for Low Responder, then I for Initial time point, F for Final time point, with 1,2, 3, or 4 for replicate number.

significant decreases in abundance for both high and low responding mice ($p < 0.0001$). Functional assessment of the proteins enriched after HFD exposure resulted in weaker associations, the strongest of which was a 1.5-fold increased representation of

Transcription proteins (Fig. 5.3E).

Functional and taxonomic characterization of fecal metaproteome in low- and high-responder mice fed the obesogenic diet.

We next focused the analysis on discovering patterns in the fecal metaproteome that might have preceded or accompanied degree of responsiveness to the obesogenic diet. Toward this end, we examined the broad-scale functional composition of each sample's metaproteome through the eggNOG database. This revealed only modest variance amongst the samples (Fig. 5.4A). In contrast, viewing the composition of taxonomic orders in this manner revealed differences, both preceding and following diet exposure, that associated with a high- and low-response to the diet (Fig. 5.4B). There were 424 highly ranked proteins ($|\pi| > 1$) differentiating high and low responders at the initial time point (Fig. 5.4C). These proteins had large differences in their taxonomic origins with all proteins distinguishing the low responders belonging to Clostridiales while high responders had over 50% of proteins derived from Bacteroidales and Lactobacillales (Fig. 5.4D). Functionally, the proteins distinguishing high responders had a 14-fold enrichment in "Posttranslational modification, protein turnover, and chaperones" proteins, and a 5.6-fold enrichment in "Cell motility" proteins (Fig. 5.4E). Many of the posttranslational modification, protein turnover, and chaperone proteins with the largest differences between high and low responders were chaperone proteins, a potential indication of a microbial stress response. The increased representation of cell motility proteins was a result of a subset of flagella, mostly derived from the order Clostridiales, that were significantly increased in high responders ($p < 0.0001$).

We also noted unique sets of carbohydrate metabolism and transport proteins

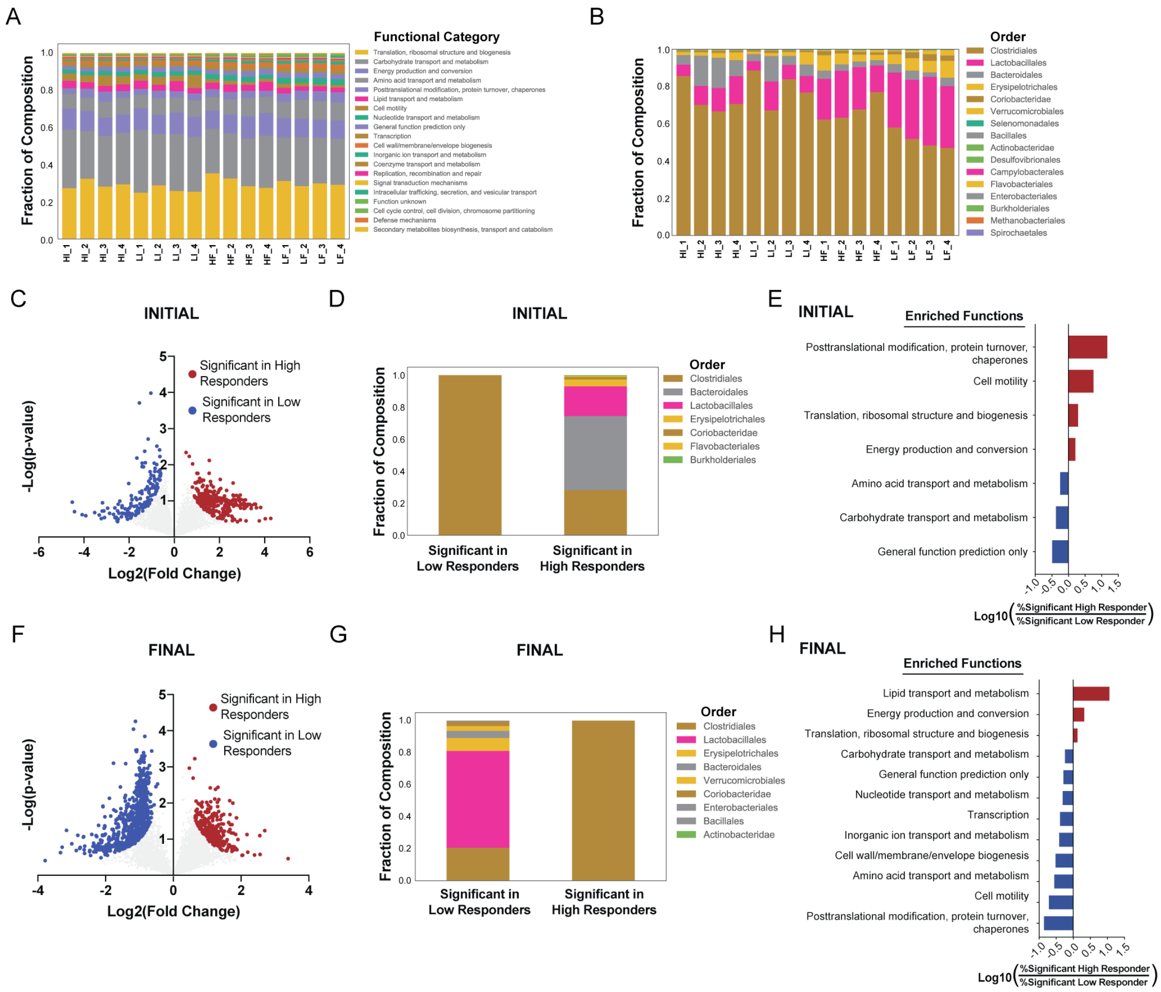


Figure 5.4 Functional and taxonomic characterization of fecal metaproteome in low- and high-responder mice fed the obesogenic diet. (**A**) Functional Composition. (**B**) Taxonomic Composition. (**C-E**) Comparison of significant proteins from high and low responders at the initial timepoint. (**C**) Volcano plot displaying the fold change and t-test significance of each protein in the metaproteome. Significance was set at |π-score| > 1. (**D**) Taxonomic composition of significant proteins. (**E**) Barplots demonstrating the functional bias in significant proteins. Compositions of eggNOG annotations were compared between high and low responders and the log ratios of the high abundance categories (>10 proteins) are shown. (**F-H**) Comparison of significant proteins from high and low responders at the final timepoint. Same analysis as (**C-E**) for the final time point. Sample names in (**A-B**) are annotated H for High Responder, L for Low Responder, then I for Initial time point, F for Final time point, with 1,2, 3, or 4 for replicate number.

differentially expressed at the initial time point. High responders had increased expression of Bacteroidale metabolism proteins including isomerases, kinases and aldolases, while Clostridiales uniquely had an increased expression of sugar transporters within low responders. This could be an indication of unique energy harvesting capacities in the microbiome present before the onset of HFD treatment(Turnbaugh et al. 2006).

In the samples collected following administration of the obesogenic diet, there were 970 proteins distinguishing high and low responders (Fig. 5.4F). In contrast to proteins discriminating responses at the onset of HFD exposure, the proteins corresponding to high response were derived entirely from Clostridiales, while nearly 60% of proteins in low responders were derived from Lactobacillales (Fig. 5.4G). Changes following the obesogenic diet associated with a high response to the diet included an 11-fold increased representation of lipid transport and metabolism proteins (Fig. 5.4H). It is possible that the increase in lipid metabolism proteins from Clostridiales mediates more efficient harvesting of energy from lipids in high responding mice.

Analysis of fecal mouse proteins.

In addition to analyzing microbial proteins from the metaproteome, we next subset the data to determine associations within the fecal mouse proteome. In total, 699 host proteins were quantified within all samples and therefore included in the statistical analyses. A large portion (77%) of mouse proteins were strongly influenced by HFD exposure, and 92% of those proteins were increased after HFD (Fig. 5.5A). Using DAVID functional enrichment(Huang da et al. 2009), we identified significant (Bonferroni adjusted p-values < 0.05) enrichment of digestion and myosin proteins before HFD treatment and mitochondrial proteins after HFD. Notably, myosin proteins occupied 6 of the top 10 proteins associated with the initial samples (Fig. 5.5B). Phosphorylated myosin light chains have previously been linked to intestinal permeability after HFD exposure(de La Serre et al. 2010). Thus, our observed decrease in heavy chain myosin proteins may be related to changes in intestinal permeability. In regards to the increase of mitochondrial proteins, it was shown that HFD results in mitochondrial dysfunction(Miotto et al. 2018), and our data likely reflects this phenomena.

We next looked for mouse fecal proteins that might discriminate high and low responders prior to HFD administration. Here, 109 mouse proteins strongly differed between the groups, all of which were enriched in the DIO prone mice (Fig. 5.5C). Of these proteins, 40 (37%) were related to immunoglobulin. This strong enrichment for immunoglobulin genes was confirmed through DAVID, which showed a significant, 3-fold enrichment (Bonferroni p-value = 7.0E-11) for Immunoglobulin V-set proteins. This enrichment of immunoglobulin variable domains was also illustrated in the top 20 proteins associated with a heightened response to HFD (Fig. 5.5D). These findings

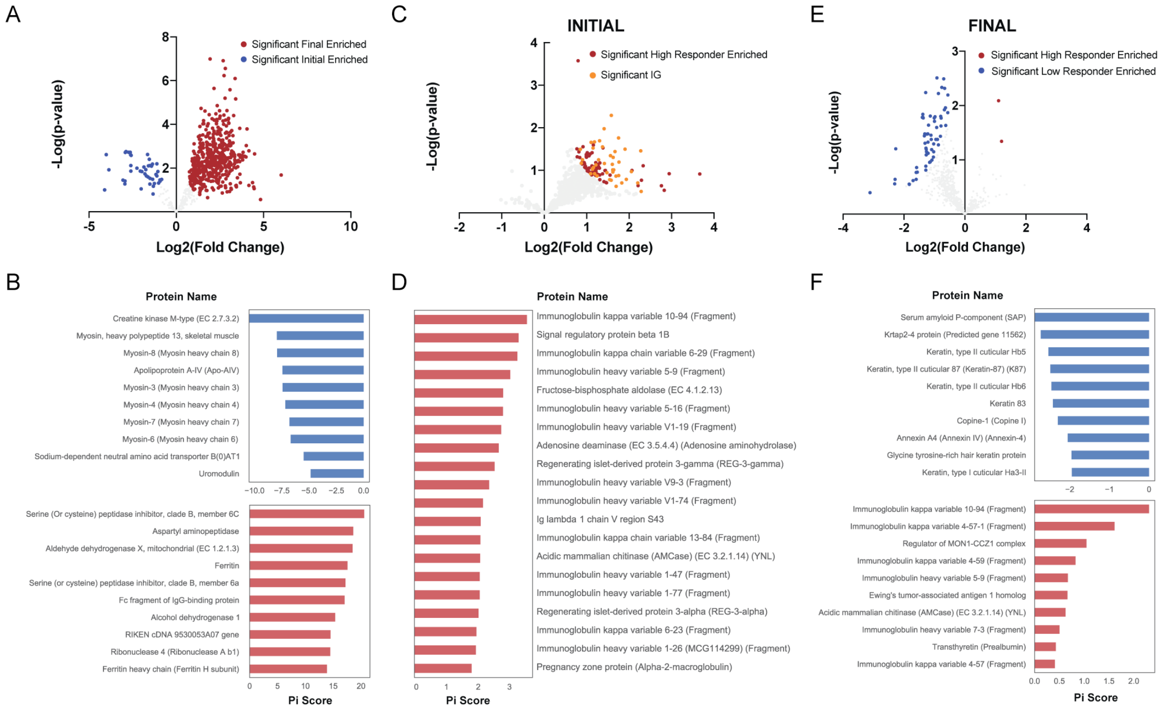


Figure 5.5 Analysis of mouse fecal proteome. After sub-setting the mouse derived proteins from the metaproteome data, differentially expressed proteins were determined using a statistical cut-off of |π-score| > 1. Volcano plots are shown demonstrating the log2 fold change (x-axis) and log10 p-value (y-axis) for (**A**) differences between final and initial samples, (**C**) differences between high and low responders and the initial time point, and (**E**) differences between high and low responders at the final time point. The π-score of the most significant proteins from each analysis are shown below each volcano plot in bar plots (**B**,**D**,**F**).

further illustrate the link between low-grade inflammation and DIO, as this is a potential indication of increased immune activity in high responder mice, before administration of HFD.

Applying the same analysis to samples collected after 8-weeks exposure to HFD also revealed interesting insight into proneness to HFD exposure. After the dietary intervention, 63 mouse proteins were highly ranked in their ability to discriminate between high and low responders. All but two of these proteins were enriched within the low responders (Fig. 5.5E). Functional analysis showed a significant 6-fold enrichment (Bonferroni p-value = 4.0E-8) for keratin within the proteins enriched within low responders. This increase of keratin could be an indication of greater colonic stress(Helenius et al. 2016) in low responders at the final time point. High responders at this time had several immunoglobulin proteins within the top discriminatory proteins, though most were only modestly associated (Fig. 5.5F). Of note, many of the immunoglobulin proteins were among the strongest discriminatory proteins in high responders at both the initial and final day.

Analysis of microbiota composition vs. proneness and severity to DIO.

Lastly, we examined the potential of fecal microbiota composition, as analyzed by 16S rRNA gene sequencing to identify and/or reflect proneness to DIO. Visualization of fecal microbiota composition of all 50 mice at all time points by unweighted UniFrac revealed the expected dramatic difference in microbiota composition before and following administration of the obesogenic diet (p = 0.001; Fig. 5.6A). This analysis also showed clear, but much more modest differences between the 5 and 8-week post-dietary change time points (Fig. 5.6A). In contrast, using this approach to examine differences in beta diversity did not identify differences in microbiota composition in high or low-responders either prior to (p = 0.977; Fig. 5.6B), or following administration of the obesogenic diet (p = 0.323; Fig. 5.6C). Rather, in accord with other diets, 8-week administration of the diet, which provided mice an additional 8 weeks to share their microbiota with their cage-mates, we observed strong cage clustering of microbiota compositions (p = 0.001; Fig. 5.6D). Nonetheless, levels of alpha-diversity, prior to administration of the obesogenic diet were moderately but significantly correlated (r^2 = 0.0873, p = 0.0394) with final body weights (Fig. 5.6E) suggesting that microbial community structure had some ability to predict proneness to DIO. An analogous but not significant trend was observed 8-weeks post-dietary change (Fig. 5.6F).

That overall assessment of microbiota composition lacked ability to identify high- and low- responder mice does not preclude the possibility that select OTUs might provide such power. Hence, we selected specific OTUs whose abundance was enriched or depleted at time 0 in the mice that developed the greatest degree of obesity in response to HFD. This yielded an array of bacterial groups, those of which had the ten lowest p values represented here. However, determining whether these differences are reproducible and/or biologically significant will require further experimentation.

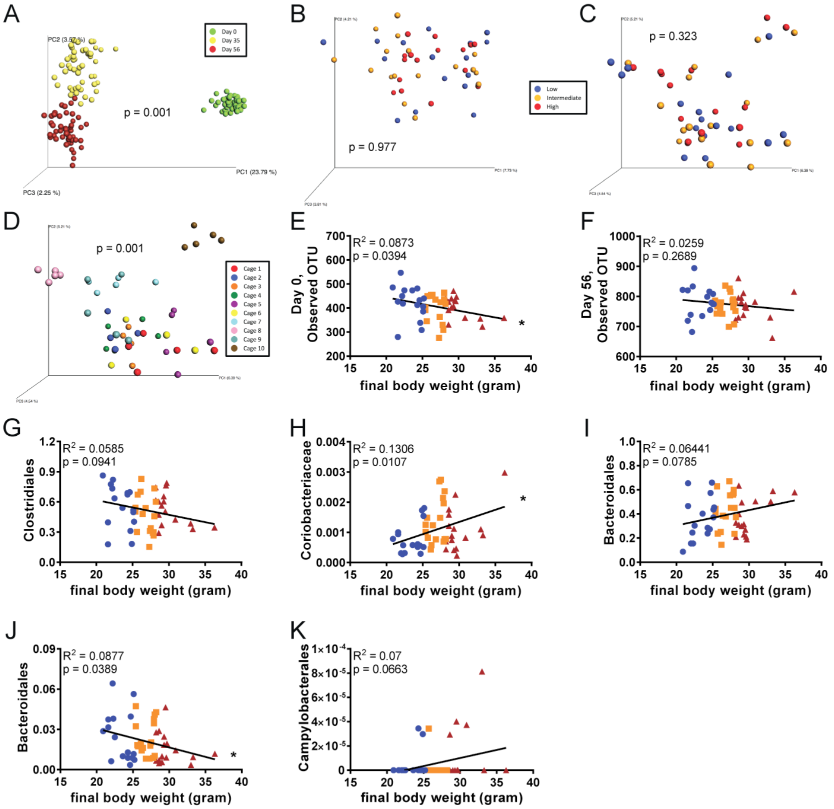


Figure 5.6 Analysis of microbiota composition vs. proneness and severity to DIO. Fecal microbiota composition was analyzed using Illumina sequencing of the V4 region of 16S rRNA genes. Principal coordinates analysis (PCoA) was performed using the unweighted UniFrac distance metric (**A**) over the course of HFD administration, (**B**) day 0, (**C-D**) day 56. Final body weights were correlated to alpha diversity at (**E**) day 0 and (**F**) day 56. Final body weights were correlated to bacterial groups found at the start of HFD administration: (**G**) Clostridiales, (**H**) Coriobacteriaceae, and (**I**) Bacteroidales. Final body weights were also correlated to bacterial groups found at the end of HFD administration: (**J**) Bacteroidales and (**K**) Campylobacterales. (*N*=50). In **A**-**D**, categories were compared and statistical significance of clustering were determined *via* Permanova. In **E**-**K**, significance was determined using linear regression analysis (*$p \leq 0.005$).

Additionally, we examined the ability of bacterial candidates generated by the metaproteomic analysis to predict (day 0), or reflect (day 56), proneness to the obesogenic diet. Of the 12 taxa analyzed in the day 0, 3 groups showed correlations predicted by the proteomic analysis with p values lower than 0.1 (Fig. 5.6G-I) while 9 did not. Regarding the taxa proteomic analysis identified as correlating with extent of obesity in the day 56 samples, 2 taxa correlated as determined with p-values lower than 0.1 (Fig. 5.6J-K) while the 10 others analyzed did not meet this criteria. Thus, overall, while development of approaches to predict proneness to obesity via analysis of the fecal proteome and/or microbiome remains a work in progress, these findings support its potential to contribute to such prognostications.

5.4 Discussion

The goal of this study was to improve understanding of non-genetic determinants of DIO, focusing on parameters that might be impacted by gut microbiota, which is known to play a role in dictating severity of DIO. As obesity is promoted by low-grade, systemic inflammation, which can be driven by exposure to microbiota products(de La Serre et al. 2010), we hypothesized that inflammatory and microbial factors might impact behavior and/or metabolism and thereby predict the extent of DIO displayed by individual hosts. However, the behavioral measures of general activity and anxiety in 6 – 8 week old mice were not predictive of susceptibility to HFD-induced obesity while the inflammatory markers IL-6, MPO, CXCL1, and LCN-2 showed only very limited ability in discriminating proneness to DIO when measured before, during, or after administration of HFD. From a microbial perspective, research has shown roles for LPS and flagellin in inflammation and obesity (Musso et al. 2010; Cullender et al. 2013; Ley and Gewirtz 2016). However, while measuring flagellin levels showed promise for predicting weight after administration of HFD, it was less successful prior to administration. Our results revealed new evidence of host-microbial interactions underlying differential weight gain.

To find microbial factors that may correlate to DIO, we next turned to an untargeted metaproteomic approach. Our results confirm prior research showing large shifts in the overall structure of the metaproteome after administering HFD (Daniel et al.

2014). These broad shifts seem to be driven by proteins derived from Clostridiales and Bacteroidales, which decreased upon exposure to HFD, while the composition of Lactobacillales proteins expands. Interestingly, this difference was not observable in our analysis of the microbiota by 16S sequencing. One possible explanation is the known discrepency between genomic and proteomic technologies(Mills et al. 2019) which is supported by the notion that differences in protein abundance are not directly associated with species composition due to complex regulatory processes. However, other DNA sequence-based studies have also shown significant alterations in Clostridiales and Bacteroidales upon exposure to HFD (de La Serre et al. 2010; Martinez-Guryn et al. 2018), further suggesting a role for these taxonomies in HFD response.

Functionally, the most striking shift with HFD was the decreased abundance of flagella after administration of HFD. Flagellin proteins can be targeted in several ways by the host, including the release of anti-flagellin IgA and anti-flagellin IgG. The levels of Anti-flagellin IgA are anti-correlated with total flagellin load, and are a key mechanism for down regulating motility-related genes (Cullender et al. 2013). While the overall levels of anti-flagellin IgG and IgA in general did not significantly correlate with the obesogenic outcomes, we did see a distinct immune signature within the fecal proteome of the high responder mice. This data may suggest that the mice that gain the most weight have a baseline immune reaction occurring before treatment. Possible antigens of this immune reaction were also identified from the metaproteome data including a subset of flagellin proteins that effectively discriminate high and low responders. However, as most of the identified immunoglobulin subunit regions could be a result of either IgA or IgV (Gemenetzi et al. 2016), it is not clear whether this potential reaction is mediated through TLR5, NOD-like receptor 4, or other mechanisms (Ley and Gewirtz 2016).

Taxonomic differences were also consistent with the idea that flagellin directed immunoglobulin may be shaping the gut of high responders before the onset of HFD. The majority of flagellin proteins identified were derived from Clostridiales, and Bacteroidales do not contain flagella (Lozupone et al. 2012). Here we observed a dominance of Clostridiales proteins enriched in low responders at the onset while Bacteroidales proteins contained a large portion of the high responder metaproteome. If the observed immunoglobulin proteins from the metaproteome were targeting flagella, it may be expected that the portion of Clostridiales proteins would be shifted in favor of Bacteroidales.

In total, our results suggest the ability of host and microbial proteomics to discern subjects particularly prone to developing DIO. Our results indicated significant metaproteome differences between high and low responding mice despite the limited number of samples analyzed. In addition, the taxonomic origins and functional roles of these discriminatory proteins suggested new evidence that host-microbiota interactions may be underlying proneness to DIO. While larger studies are needed to confirm our results, the fecal metaproteome appears to be a promising tool for identifying hosts at risk of weight gain upon exposure to an obesogenic diet.

5.5 Methods

Mice and high-fat diet administration

Female, 3-5 week old C57BL/6 mice were purchased from Jackson Laboratory (Bar Harbor, ME) and maintained at Georgia State University, Atlanta, Georgia, USA under institutionally approved protocol under approved protocols (IACUC # A14033), housed 5 mice per cage, were subjected to metabolic monitoring, including behavior analysis and sample collection over a 3-week period. During this time, the mice were fed standard grain-based chow (GBC), which is comprised of relatively unrefined ingredients. The cohort of mice was then switched to a diet composed of 60% kcal from fat (Research Diet, D12492) for 8 weeks. Mice were then euthanized, and colon length, colon weight, spleen weight and adipose weight were measured. Serum, feces, and organs were collected for downstream analysis.

Fecal metaproteome data acquisition

Fecal samples were measured out to ~0.2 g and suspended in 10 mL of ice-cold, sterilized TBS. A 20 μM vacuum, steriflip (Milipore) filter was used to remove particulate from the samples. Cells were pelleted through centrifugation at 4000 rpm for 10 min. Next, cells were lysed in 2 mL of buffer containing 75 mM NaCl (Sigma), 3% sodium dodecyl sulfate (SDS, Fisher), 1 mM NaF (Sigma), 1 mM beta-glycerophosphate (Sigma), 1 mM sodium orhtovanadate (Sigma), 10 mM sodium pyrophosphate (Sigma), 1 mM phenylmethylsulfonyl fluoride (PMSF, Sigma), and 1X Complete Mini EDTA free protease inhibitors (Roche) in 50 mM HEPES (Sigma), pH 8.5(Villen and Gygi 2008). An equal volume of 8M Urea in 50 mM HEPES, pH 8.5 was added to each sample. Cell lysis was achieved through two 10-second intervals of probe sonication at 25% amplitude. Proteins were then reduced with dithiothreitol (DTT, Sigma), alkylated through iodoacetamide (Sigma), and quenched as previously described(Haas et al. 2006). Proteins were then precipitated via chloroform-methanol precipitation and protein pellets were dried(Wessel and Flugge 1984). Protein pellets were re-suspended in 1M urea in 50 mM HEPES, pH 8.5 and digested overnight at room temperature with LysC (Wako)(Van Rechem et al. 2015). A second, 6-hour digestion using trypsin at 37 ºC was performed and the reaction was stopped through addition of 10% trifluoroacetic acid (TFA, Pierce). Samples were then desalted through C18 Sep-Paks (Waters) and eluted with a 40% and 80% Acetonitrile solution containing 0.5 % Acetic Acid(Tolonen 2014). Concentration of desalted peptides was determined with a BCA assay (Thermo Scientific). 50 μg aliquots of each sample were dried in a speed-vac, additional bridge channels consisting of 25 μg from each sample were created and 50 μg aliquots of this solution were used in duplicate per TMT 10-plex (Thermo Scientific) as previously described(Lapek et al. 2018). These bridge channels were used to control for labeling efficiency, inter-run variation, mixing errors and the heterogeneity present in each sample(Tolonen et al. 2011). Each sample or bridge channel was resuspended in 30% dry acetonitrile in 200 mM HEPES, pH 8.5 for TMT labeling with 7 μL of the appropriate TMT reagent(Thompson et al. 2003). Reagents 126 and 131 (Thermo Scientific) were used to bridge between mass spec runs. Remaining reagents were used to label samples in random order. Labeling was carried out for 1 hour at room temperature, and quenched by adding 8 μL of 5% hydroxylamine (Sigma). Labeled samples were acidified by adding 50 μL of 1% TFA. After TMT labeling, each 10-plex experiment was combined and desalted through C18 Sep-Paks and dried in a speed-vac. Each 10-plex experiment was fractionated using a High pH Reversed-Phase Peptide Fractionation Kit (Pierce) per manufacturer instructions. All LC-MS^2/MS^3 experiments were carried out on an Orbitrap Fusion (Thermo Fisher Scientific) with an in-line Easy-nLC 1000 (Thermo Fisher Scientific) and chilled autosampler. Separation and acquisition settings were as previously defined(Lapek et al.

2017).

Metaproteome data processing

Data was processed using Proteome Discoverer 2.1 (Thermo Fisher Scientific). MS^2 data was searched against a catalog of mouse gut genes(Xiao et al. 2015) (accessed 02/12/2017) containing 2,569,907 entries along side the Uniprot mouse proteome (www.uniprot.org, access date 11/14/2016) which contained 53,374 entries. The Sequest searching algorithm(Eng et al. 1994) was used to align spectra to database peptides. A precursor mass tolerance of 50 ppm(Beausoleil et al. 2006; Huttlin et al. 2010) was specified and 0.6 Da tolerance for MS^2 fragments. Included in the search parameters was static modification of TMT 10-plex tags on lysine and peptide n-termini (+229.162932 Da), carbamidomethylation of cysteines (+57.02146 Da), and variable oxidation of methionine (+15.99492 Da). The search parameters included trypsin as the enzyme used to generate peptides with a maximum of 2 missed cleavages permitted. A two-step database search method was utilized(Zhang et al. 2016) wherein proteins identified in either the forward or reverse database were included in a second search containing 14,368 entries from the original mouse gut gene catalog, and annotations derived from this database were used for downstream analysis of microbial proteins(Xiao et al. 2015). A peptide and protein false discovery rate of 1% was enforced using a reverse database search strategy(Peng et al. 2003; Elias et al. 2005; Elias and Gygi 2007).

TMT reporter ion intensities were extracted from MS^3 spectra for quantitative analysis and signal-to-noise ratios were used for quantitation. Additional stringent filtering was used removing any moderate confidence peptide spectral matches (PSMs), or ambiguous PSM assignments. Additionally, any peptides with a spectral interference above 25% were removed, as well as any peptides with an average signal to noise ratio less than 10. Normalization occurred as previously described(Lapek et al. 2017). Briefly, relative abundances are normalized first to the pooled standards for each protein and then to the median signal across the pooled standard. An average of these normalizations was used for the next step. To account for slight differences in amounts of protein labeled, these values were then normalized to the median of the entire dataset and reported as final normalized summed signal-to-noise ratios per protein per sample.

Behavioral analysis

Three weeks after arrival in the animal facilities, behavior in the open field and in the home cage was assessed in a counter-balanced fashion over the course of two days. Behavioral testing occurred within the last 4 h of the light and quiescent phase and was conducted under illumination of overhead white lighting (between 300-400 lux). Open field arenas were cleaned with 70% ethanol between trials, and home cage bedding was changed after each trial. Behavioral tests were videotape using a Sony camcorder for later analysis by The Observer version XT11 (Noldus Information Technology Inc., Wageningen, The Netherlands). An experiment blinded as to the treatment conditions scored behavioral tests in the Observer.

Open Field Test

Locomotor behavior was assessed in a 43.2 X 43.2 X 30.5cm (WxLxH) Plexiglas arena (Med Associates, Inc., St. Albans, VT) containing 2 arrays of infrared transmitters strips (16 beams each) located on the bottom of the arena (in the X and Y plane). The center zone of the arena was defined as square containing the center 8 beams (e.g., beams 4-12) in the X and Y plane. Each mouse was placed into the arena with its nose facing the wall and allowed to freely investigate for 10 min. The total distance traveled, the total time spent in the center of the arena, and circling behaviors, which are defined as movements below a preset ambulatory threshold, were calculated by Activity Monitor (Med Associates, Inc.) on a computer connected to the open field arenas.

Home Cage Behavior

Mice were placed into a clean housing cage containing 2 cm deep Alpha-dri bedding (Shepherd Specialty Paper, Fibercore, Cleveland, OH, USA) and video recorded for 10 min. An experimenter blinded as to condition scored the occurrence and duration of i) the time spent walking, as defined by locomotion along the bottom of the enclosure, around the arena, ii) grooming, as defined by stroking or scratching the face of body iii) digging, as defined as using the fore- or hind paws to displace the bedding, and iv) the rears, as defined by standing on the hind legs with either the forepaws unsupported or when the forepaws were supported by the walls of the enclosure, were quantified using the Observer.

Fasting blood glucose measurement and body composition measurement.

For fasting blood glucose tolerance test, mice were fasted for 5 hours, and baseline blood glucose were measured by using a Nova Max plus Glucose meter and expressed in mg/dL. Measurement of percent fat mass and lean mass was performed via MRI (Bruker MiniSpec) at day 0, prior to diet treatment, and day 28 and 56, after diet treatment.

Fecal sample preparation for immunoglobulin quantification.

Fecal sample preparation of enzyme-linked immunosorbent assay (ELISA) has been previously described(Cong et al. 1998). 100 mg of fecal pellets were homogenized in 3 mL of collection media consisting of 0.05 mg soybean trypsin inhibitor per ml of a 3:1 mixture of 1X PBS and 0.1 M EDTA, pH 7.4. Following centrifugation at 1800 rpm for 10 minutes, the supernatant was centrifuged again at 14,000 rpm for 15 minutes at 4°C, and final supernatant was collected and stored with 20% glycerol and 2 mM phenylmethylsulfonyl fluoride (Sigma, P-7626) at -20°C until analysis.

Fecal and serum anti-flagellin IgA/IgG

Quantification of anti-flagellin- specific IgA and IgG has been previously described(Sitaraman et al. 2005; Ziegler et al. 2008; Fedirko et al. 2017). Briefly, 96-well microtiter plates (Costar, Corning, New York) were coated with 100 ng/well of laboratory-made flagellin in 9.6 pH bicarbonate buffer overnight at 4° C. Serum samples from mice were then applied either pure or at a 1:100 dilution for 1 hour at 37° C. After incubation and washing, the wells were incubated with either horseradish peroxidase-linked anti-mouse IgG (GE Healthcare Life Sciences, Pittsburgh, Pennsylvania) or horseradish peroxidase-linked anti-IgA (Southern Biotech, Birmingham, Alabama). Quantification of immunoglobulin was then developed by the addition of 3,3',5,5'-Tetramethylbenzidine and the optical density was calculated by the difference between readings at 450nm and 540nm.

Fecal Lcn-2 quantification

As previously described(Chassaing et al. 2012), frozen fecal samples were reconstituted in PBS containing 0.1% Tween 20 at 100 mg/ml and vortexed for 20 min. The homogenate was then centrifuged at 12,000 rpm for 10 min at 4°C. Clear supernatants were collected and stored at −20°C until analysis. Lcn-2 levels were measured in the supernatants using Duoset murine Lcn-2 ELISA kit (R&D Systems, Minneapolis, MN).

Myeloperoxidase quantification

Tissue samples were homogenized in 100 mg/mL of 0.5% hexadecyltrimethylammonium bromide (Sigma, St. Louis, MO) in 50 mM PBS, pH 6.0, as previously described(Chassaing et al. 2012). Following 3 cycles of freeze-thaw at -80°C and 37°C, samples were sonicated and centrifuged at 14,000 rpm for 15 min at 4°C. Supernatants were stored at −20°C until analysis. Myeloperoxidase (MPO) was assayed in the supernatant by adding 1 mg/mL of dianisidine dihydrochloride (Sigma, St. Louis, MO) and 5×10^{-4}% H_2O_2 and the change in optical density measured at 450 nm.

Serum CXCL1 and IL-6 quantification

Serum chemokine (C-X-C motif) ligand 1 (CXCL1) and Interleukin-6 (IL-6) concentrations were determined using Duoset cytokine ELISA kits (R&D Systems, Minneapolis, MN) according to manufacturer's instructions(Chassaing et al. 2012).

Fecal flagellin and lipopolysaccharide load quantification

We quantified flagellin and lipopolysaccharide (LPS) as previously described using human embryonic kidney (HEK)-Blue-mTLR5 and HEK-BluemTLR4 cells, respectively (Invivogen, San Diego, California, USA)(Chassaing et al. 2015; Chassaing et al. 2017). We resuspended fecal material in PBS to a final concentration of 100 mg/mL and homogenized for 10 s using a Mini-Beadbeater-24 without the addition of beads to avoid bacteria disruption. We then centrifuged the samples at 8000 *g* for 2 min, serially diluted the resulting supernatant, and applied to mammalian cells. Purified *E. coli* flagellin and LPS (Sigma, St Louis, Missouri, USA) were used for standard curve determination using HEK-Blue-mTLR5 and HEK-Blue-mTLR4 cells, respectively. After 24 h of stimulation, we applied cell culture supernatant to QUANTI-Blue medium (Invivogen, San Diego, California, USA) and measured alkaline phosphatase activity at 620 nm after 30 min.

Microbiota analysis by 16S rRNA gene sequencing using Illumina MiSeq technology

16S rRNA gene amplification and sequencing were done using the Illumina MiSeq technology following the protocol of Earth Microbiome Project with their modifications to the MOBIO PowerSoil DNA Isolation Kit procedure for extracting DNA (www.earthmicrobiome.org/emp-standard-protocols). Bulk DNA were extracted from frozen extruded feces using a PowerSoil-htp kit from MoBio Laboratories (Carlsbad, California, USA) with mechanical disruption (bead-beating). The 16S rRNA genes, region V4, were PCR amplified from each sample using a composite forward primer and a reverse primer containing a unique 12-base barcode, designed using the Golay error-correcting scheme, which was used to tag PCR products from respective samples(Caporaso et al. 2012). We used the forward primer 515F 5’-*AATGATACGGCGACCACCGAGATCTACAC*TATGGTAATT*GT*GTGCCAGCMGCCG CGGT AA-3’: the italicized sequence is the 5’ Illumina adapter B, the bold sequence is the primer pad, the italicized and bold sequence is the primer linker and the underlined sequence is the conserved bacterial primer 515F. The reverse primer 806R used was 5’-*CAAGCAGAAGACGGCATACGAGAT* XXXXXXXXXXXX AGTCAGTCAG *CCGGAC TACHVGGGTWTCTAAT*-3’: the italicized sequence is the 3’ reverse complement sequence of Illumina adapter, the 12 X sequence is the golay barcode, the bold sequence is the primer pad, the italicized and bold sequence is the primer linker and the underlined sequence is the conserved bacterial primer 806R. PCR reactions consisted of Hot Master

PCR mix (Five Prime), 0.2 μM of each primer, 10-100 ng template, and reaction conditions were 3 min at 95°C, followed by 30 cycles of 45 s at 95°C, 60s at 50°C and 90 s at 72°C on a Biorad thermocycler. Four independent PCRs were performed for each sample, combined, purified with Ampure magnetic purification beads (Agencourt), and products were visualized by gel electrophoresis. Products were then quantified (BIOTEK Fluorescence Spectrophotometer) using Quant-iT PicoGreen dsDNA assay. A master DNA pool was generated from the purified products in equimolar ratios. The pooled products were quantified using Quant-iT PicoGreen dsDNA assay and then sequenced using an Illumina MiSeq sequencer (paired-end reads, 2 × 250 bp) at Cornell University, Ithaca.

16S rRNA gene sequence analysis

Forward and reverse Illumina reads were joined using the fastq-join method(Aronesty 2011; Aronesty 2013), sequences were demultiplexed, quality filtered using Quantitative Insights Into Microbial Ecology (QIIME, version 1.8.0) software package(Caporaso et al. 2010). QIIME default parameters were used for quality filtering (reads truncated at first low-quality base and excluded if: (1) there were more than three consecutive low quality base calls (2), less than 75% of read length was consecutive high quality base calls (3), at least one uncalled base was present (4), more than 1.5 errors were present in the bar code (5), any Phred qualities were below 20, or (6) the length was less than 75 bases). Sequences were clustered to operational taxonomic units (OTUs) using UCLUST algorithm(Edgar 2010) with a 97% threshold of pairwise identity (without the creation of new clusters with sequences that do not match the reference sequences), and taxonomically classified using the Greengenes reference database

13_8(McDonald et al. 2012). A single representative sequence for each OTU was aligned and a phylogenetic tree was built using FastTree(Price et al. 2009). The phylogenetic tree was used for computing the unweighted UniFrac distances between samples(Lozupone and Knight 2005; Lozupone et al. 2006), rarefaction were performed and used to compare abundances of OTUs across samples. Principal coordinates analysis (PCoA) plots were used to assess the variation between experimental group (beta diversity). Alpha diversity curves were determined for all samples using the determination of the number of observed species. LEfSE (LDA Effect Size) was used to investigate bacterial members that drive differences between groups(Segata et al. 2011). Unprocessed sequencing data are deposited in the European Nucleotide Archive under accession number PRJEB33328.

Experimental Design and Statistical Rationale:

Study design

The overall study included 50 mice which were followed longitudinally for monitoring of weight gain and other measures. The sample size was determined for statistical power based on previous publications(Archer et al. 2003; de La Serre et al. 2010) and experience. The metaproteome analysis included four mice with the highest weight gain and the lowest weight gain. These samples sizes were determined to be sufficient based on previous reports of strong differences in related animal models with similar sample sizes (Daniel et al. 2014).

Metaproteome analysis

Metaproteome analysis was performed using python (version 3.5) and records are available online (https://github.com/rhmills/High-Fat_Diet_Metaproteomics_analysis). Extra files associated with the analysis within the notebooks are deposited as supplementary files in the MassIVE (https://massive.ucsd.edu) repository for this study (Study ID: MSV000083891). All analysis was performed on the proteins identified in both TMT 10-plex experiments. Qiime2, version 2019.1 (https://qiime2.org/), was used for principle coordinates analysis through the command "qiime diversity core-metrics" as well as for determining significance of beta-diversity clustering through the command "qiime diversity beta-group-significance". K-means clustering was performed through Morpheus (https://software.broadinstitute.org/morpheus). Enriched and depleted proteins were determined by π–score, which accounts for both fold change and p-value(Xiao et al. 2014). A statistical cutoff for highly ranked associations was set to $|\pi| > 1$ ($\alpha \sim 0.05$), which provided an adequate number of proteins for functional and taxonomic assessment(Mills et al. 2019) while maintaining a moderate stringincy. Volcano plots were visualized using GraphPad Prism (version 7.0b). Mouse protein gene functional enrichment analysis was performed using DAVID(Huang da et al. 2009), with all mouse proteins identified as a background list. The python package, Seaborn (version 0.9.0) was used for boxplots, swarmplots, and catplots. Statistical analysis between groups within the boxplots was performed using ANOVA with Dunnett corrected p-values through GraphPad Prism (version 7.0b).

Statistical analysis

Significance was determined using unpaired two-tailed *t*-test or linear regression analysis (GraphPad Prism software, version 6.01). Differences were noted as significant $^*p \leq 0.05$ for *t*-test or linear regression analysis. For clustering analyzing on principal coordinate plots, categories were compared and statistical significance of clustering was determined *via* Permanova(Caporaso et al. 2010).

Availability of data and materials.

All data generated or analyzed during this study are included in this published article. Metaproteomic data is available through massive (massive.ucsd.edu) under study ID MSV000083891. The data is also available through Proteome Xchange (http://proteomecentral.proteomexchange.org) under the study ID PXD014128.

Chapter 5 is a reprint of the material as it appears in Molecular and Cellular Proteomics, 2019, Hao Q. Tran, Robert H. Mills, Nicole V. Peters, Mary K. Holder, Geert J. de Vries, Rob Knight, Benoit Chassaing David J. Gonzalez, and Andrew T. Gewirtz. The dissertation author played a primary role in aspects of the work ranging from metaproteome data acquisition, data analysis and the writing of the manuscript.

5.6 References

Apovian CM. 2016. Obesity: definition, comorbidities, causes, and burden. *Am J Manag Care* **22**: s176-185.

Archer ZA, Rayner DV, Rozman J, Klingenspor M, Mercer JG. 2003. Normal distribution of body weight gain in male Sprague-Dawley rats fed a high-energy diet. *Obes Res* **11**: 1376-1383.

Aronesty E. 2011. Command-line tools for processing biological sequencing data. http://codegooglecom/p/ea-utils.

Aronesty E. 2013. Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal* **7**: 1-8.

Aygun AD, Gungor S, Ustundag B, Gurgoze MK, Sen Y. 2005. Proinflammatory cytokines and leptin are increased in serum of prepubertal obese children. *Mediators Inflamm* **2005**: 180-183.

Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. 2006. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* **24**: 1285-1292.

Cani PD, Amar J, Iglesias MA, Poggi M, Knauf C, Bastelica D, Neyrinck AM, Fava F, Tuohy KM, Chabo C et al. 2007. Metabolic endotoxemia initiates obesity and insulin resistance. *Diabetes* **56**: 1761-1772.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335-336.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**: 1621-1624.

Carvalho FA, Koren O, Goodrich JK, Johansson ME, Nalbantoglu I, Aitken JD, Su Y, Chassaing B, Walters WA, Gonzalez A et al. 2012. Transient inability to manage proteobacteria promotes chronic gut inflammation in TLR5-deficient mice. *Cell Host Microbe* **12**: 139-152.

Chassaing B, Koren O, Goodrich JK, Poole AC, Srinivasan S, Ley RE, Gewirtz AT. 2015. Dietary emulsifiers impact the mouse gut microbiota promoting colitis and metabolic syndrome. *Nature* **519**: 92-96.

Chassaing B, Srinivasan G, Delgado MA, Young AN, Gewirtz AT, Vijay-Kumar M. 2012. Fecal lipocalin 2, a sensitive and broadly dynamic non-invasive biomarker for intestinal inflammation. *PLoS One* **7**: e44328.

Chassaing B, Van de Wiele T, De Bodt J, Marzorati M, Gewirtz AT. 2017. Dietary emulsifiers directly alter human microbiota composition and gene expression ex vivo potentiating intestinal inflammation. *Gut* **66**: 1414-1427.

Cong Y, Brandwein SL, McCabe RP, Lazenby A, Birkenmeier EH, Sundberg JP, Elson CO. 1998. CD4+ T cells reactive to enteric bacterial antigens in spontaneously colitic C3H/HeJBir mice: increased T helper cell type 1 response and ability to transfer disease. *J Exp Med* **187**: 855-864.

Cullender TC, Chassaing B, Janzon A, Kumar K, Muller CE, Werner JJ, Angenent LT, Bell ME, Hay AG, Peterson DA et al. 2013. Innate and adaptive immunity interact to quench microbiome flagellar motility in the gut. *Cell Host Microbe* **14**: 571-581.

Daniel H, Gholami AM, Berry D, Desmarchelier C, Hahne H, Loh G, Mondot S, Lepage P, Rothballer M, Walker A et al. 2014. High-fat diet alters gut microbiota physiology in mice. *ISME J* **8**: 295-308.

de La Serre CB, Ellis CL, Lee J, Hartman AL, Rutledge JC, Raybould HE. 2010. Propensity to high-fat diet-induced obesity in rats is associated with changes in the gut microbiota and gut inflammation. *Am J Physiol Gastrointest Liver Physiol* **299**: G440-448.

Dinan TG, Cryan JF. 2017. Gut-brain axis in 2016: Brain-gut-microbiota axis - mood, metabolism and behaviour. *Nat Rev Gastroenterol Hepatol* **14**: 69-70.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460-2461.

El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B. 2013. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat Rev Microbiol* **11**: 497-504.

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**: 207-214.

Elias JE, Haas W, Faherty BK, Gygi SP. 2005. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods* **2**: 667-675.

Eng JK, McCormack AL, Yates JR. 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**: 976-989.

Fedirko V, Tran HQ, Gewirtz AT, Stepien M, Trichopoulou A, Aleksandrova K, Olsen A, Tjonneland A, Overvad K, Carbonnel F et al. 2017. Exposure to bacterial products lipopolysaccharide and flagellin and hepatocellular carcinoma: a nested case-control study. *BMC Med* **15**: 72.

Gemenetzi K, Agathangelidis A, Papalexandri A, Medina A, Genuardi E, Moysiadis T, Hatjiharissi E, Papaioannou M, Terpos E, Jimenez C et al. 2016. Distinct Immunogenetic Signatures in IgA Versus IgG Multiple Myeloma. *Blood* **128**.

Haas W, Faherty BK, Gerber SA, Elias JE, Beausoleil SA, Bakalarski CE, Li X, Villen J, Gygi SP. 2006. Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol Cell Proteomics* **5**: 1326-1337.

Hayashi F, Smith KD, Ozinsky A, Hawn TR, Yi EC, Goodlett DR, Eng JK, Akira S, Underhill DM, Aderem A. 2001. The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**: 1099-1103.

Helenius TO, Antman CA, Asghar MN, Nystrom JH, Toivola DM. 2016. Keratins Are Altered in Intestinal Disease-Related Stress Responses. *Cells* **5**.

Hildebrandt MA, Hoffmann C, Sherrill-Mix SA, Keilbaugh SA, Hamady M, Chen YY, Knight R, Ahima RS, Bushman F, Wu GD. 2009. High-fat diet determines the composition of the murine gut microbiome independently of obesity. *Gastroenterology* **137**: 1716-1724 e1711-1712.

Hotamisligil GS. 2017. Inflammation, metaflammation and immunometabolic disorders. *Nature* **542**: 177-185.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44-57.

Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villen J, Haas W, Sowa ME, Gygi SP. 2010. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**: 1174-1189.

Jeon SW, Kim YK. 2018. The role of neuroinflammation and neurovascular dysfunction in major depressive disorder. *J Inflamm Res* **11**: 179-192.

Jones BW, Heldwein KA, Means TK, Saukkonen JJ, Fenton MJ. 2001. Differential roles of Toll-like receptors in the elicitation of proinflammatory responses by macrophages. *Ann Rheum Dis* **60 Suppl 3**: iii6-12.

Lapek JD, Jr., Lewinski MK, Wozniak JM, Guatelli J, Gonzalez DJ. 2017. Quantitative Temporal Viromics of an Inducible HIV-1 Model Yields Insight to Global Host Targets and Phospho-Dynamics Associated with Vpr. *Mol Cell Proteomics* doi:10.1074/mcp.M116.066019.

Lapek JD, Jr., Mills RH, Wozniak JM, Campeau A, Fang RH, Wei X, van de Groep K, Perez-Lopez A, van Sorge NM, Raffatellu M et al. 2018. Defining Host Responses during Systemic Bacterial Infection through Construction of a Murine Organ Proteome Atlas. *Cell Syst* doi:10.1016/j.cels.2018.04.010.

Ley RE, Gewirtz AT. 2016. Corralling Colonic Flagellated Microbiota. *N Engl J Med* **375**: 85-87.

Ley RE, Turnbaugh PJ, Klein S, Gordon JI. 2006. Microbial ecology: human gut microbes associated with obesity. *Nature* **444**: 1022-1023.

Lozupone C, Faust K, Raes J, Faith JJ, Frank DN, Zaneveld J, Gordon JI, Knight R. 2012. Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. *Genome Res* **22**: 1974-1984.

Lozupone C, Hamady M, Knight R. 2006. UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. *BMC bioinformatics* **7**: 371.

Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228-8235.

Martinez-Guryn K, Hubert N, Frazier K, Urlass S, Musch MW, Ojeda P, Pierre JF, Miyoshi J, Sontag TJ, Cham CM et al. 2018. Small Intestine Microbiota Regulate Host Digestive and Absorptive Adaptive Responses to Dietary Lipids. *Cell Host Microbe* **23**: 458-469 e455.

Masoodi I, Kochhar R, Dutta U, Vaishnavi C, Prasad KK, Vaiphei K, Hussain S, Singh K. 2012. Evaluation of fecal myeloperoxidase as a biomarker of disease activity and severity in ulcerative colitis. *Dig Dis Sci* **57**: 1336-1340.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610-618.

Mills RH, Vázquez-Baeza Y, Zhu Q, Jiang L, Gaffney J, Humphrey G, Smarr L, Knight R, Gonzalez DJ. 2019. Evaluating Metagenomic Prediction of the Metaproteome in a 4.5-Year Study of a Patient with Crohn's Disease. *mSystems* **4**: e00337-00318.

Miotto PM, LeBlanc PJ, Holloway GP. 2018. High-Fat Diet Causes Mitochondrial Dysfunction as a Result of Impaired ADP Sensitivity. *Diabetes* **67**: 2199-2205.

Musso G, Gambino R, Cassader M. 2010. Obesity, diabetes, and gut microbiota: the hygiene hypothesis expanded? *Diabetes Care* **33**: 2277-2284.

Naukkarinen J, Heinonen S, Hakkarainen A, Lundbom J, Vuolteenaho K, Saarinen L, Hautaniemi S, Rodriguez A, Fruhbeck G, Pajunen P et al. 2014. Characterising metabolically healthy obesity in weight-discordant monozygotic twins. *Diabetologia* **57**: 167-176.

Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. 2003. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* **2**: 43-50.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641-1650.

Sanders CJ, Yu Y, Moore DA, 3rd, Williams IR, Gewirtz AT. 2006. Humoral immune response to flagellin requires T cells and activation of innate immunity. *J Immunol* **177**: 2810-2818.

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. *Genome Biol* **12**: R60.

Sitaraman SV, Klapproth JM, Moore DA, 3rd, Landers C, Targan S, Williams IR, Gewirtz AT. 2005. Elevated flagellin-specific immunoglobulins in Crohn's disease. *Am J Physiol Gastrointest Liver Physiol* **288**: G403-406.

Stelmach-Mardas M, Rodacki T, Dobrowolska-Iwanek J, Brzozowska A, Walkowiak J, Wojtanowska-Krosniak A, Zagrodzki P, Bechthold A, Mardas M, Boeing H. 2016. Link between Food Energy Density and Body Weight Changes in Obese Adults. *Nutrients* **8**: 229.

Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C. 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**: 1895-1904.

Tolonen AC, and Haas, W. 2014. Quantitative Proteomics Using Reductive Dimethylation for Stable Isotope Labeling. *Jove-J Vis Exp*.

Tolonen AC, Haas W, Chilaka AC, Aach J, Gygi SP, Church GM. 2011. Proteome-wide systems analysis of a cellulosic biofuel-producing microbe. *Mol Syst Biol* **7**: 461.

Turnbaugh PJ. 2017. Microbes and Diet-Induced Obesity: Fast, Cheap, and Out of Control. *Cell Host Microbe* **21**: 278-281.

Turnbaugh PJ, Backhed F, Fulton L, Gordon JI. 2008. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe* **3**: 213-223.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP et al. 2009. A core gut microbiome in obese and lean twins. *Nature* **457**: 480-484.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**: 1027-1031.

Van Rechem C, Black JC, Boukhali M, Aryee MJ, Graslund S, Haas W, Benes CH, Whetstine JR. 2015. Lysine Demethylase KDM4A Associates with Translation Machinery and Regulates Protein Synthesis. *Cancer Discov* **5**: 255-263.

Vijay-Kumar M, Aitken JD, Carvalho FA, Cullender TC, Mwangi S, Srinivasan S, Sitaraman SV, Knight R, Ley RE, Gewirtz AT. 2010. Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science* **328**: 228-231.

Villen J, Gygi SP. 2008. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat Protoc* **3**: 1630-1638.

Wessel D, Flugge UI. 1984. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal Biochem* **138**: 141-143.

Xiao L, Feng Q, Liang S, Sonne SB, Xia Z, Qiu X, Li X, Long H, Zhang J, Zhang D et al. 2015. A catalog of the mouse gut metagenome. *Nat Biotechnol* **33**: 1103-1108.

Xiao Y, Hsiao TH, Suresh U, Chen HI, Wu X, Wolf SE, Chen Y. 2014. A novel significance score for gene selection and ranking. *Bioinformatics* **30**: 801-807.

Zhang X, Figeys D. 2019. Perspective and Guidelines for Metaproteomics in Microbiome Studies. *J Proteome Res* doi:10.1021/acs.jproteome.9b00054.

Zhang X, Ning ZB, Mayne J, Moore JI, Li J, Butcher J, Deeke SA, Chen R, Chiang CK, Wen M et al. 2016. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **4**.

Ziegler TR, Luo M, Estivariz CF, Moore DA, 3rd, Sitaraman SV, Hao L, Bazargan N, Klapproth JM, Tian J, Galloway JR et al. 2008. Detectable serum flagellin and lipopolysaccharide and upregulated anti-flagellin and lipopolysaccharide immunoglobulins in human short bowel syndrome. *Am J Physiol Regul Integr Comp Physiol* **294**: R402-410.

Chapter 6

Concluding remarks and future directions

6.1 Summary

The work in this dissertation outlines the potential impact that multiplexed metaproteomics might have for our understanding of host-microbiome interactions. As shown in Chapter 2, the microbiome has an organism-wide impact on animal physiology. The human microbiome is now associated with a wide-range of diseases and the technology developed herein may reveal new, actionable insights regarding the nature of these associations. The work in Chapter 3 developed a methodological pipeline and evaluated multiplexed metaproteomics against conventionally collected metagenomic data. In Chapter 4, we see the true potential of the technology to provide novel and translationally relevant insights, which might lead to novel therapeutic treatments in IBD. Further we were able to integrate these data and evaluate them in the broader context of meta- omic data, comparing results between shotgun metagenomics, amplicon sequencing, metabolomics, and metaproteomics. Further emphasizing the potential of this technology are the results of Chapter 5, which highlight that multiplexed metaproteomics proved to be a breakthrough technology when trying to find molecular markers that might predict obesity outcomes.

Beyond the promise of the emerging technology were several emerging hypotheses that serve as future research directions. In Chapter 4, we identified a striking association between Ulcerative Colitis disease activity and *Bacteroides* proteins. After further investigation, we identified *Bacteroides vulgatus* proteases as a potential therapeutic target. To date, there remains very little known about *Bacteroides* proteases. The genera are cornerstone stone members of the gut, needed for the commensal degradation of complex-carbohydrates(Foley et al. 2016). It is also known that they have membrane vesicles tightly packed with proteases (Elhenawy et al. 2014). However, it remains unknown how the proteases are regulated. Does a shift in nutrient availability prompt increased production and secretion of proteases? Does the inflammatory state of the gut of someone with Ulcerative Colitis induce the expression of *Bacteroides* proteases? And ultimately, are *Bacteroides vulgatus* proteases a viable therapeutic target for treating Ulcerative Colitis? The remainder of this chapter will outline an experimental design to further investigate these pertinent questions.

6.2 Determine the regulation and activity of *B. vulgatus* proteases relevant to UC

Characterize *B. vulgatus* protease activity relevant to UC

Our preliminary research shed light on a potential role for *B. vulgatus* proteases as contributors to UC pathology. PubMed searches of '*Bacteriodes vulgatus proteases and Ulcerative colitis*' yields 0 articles. Thus, our proposed studies are pioneering given that the role of the microbiome in UC disease etiology is highly understudied and is likely to play a crucial role. This subaim will be a first step in validating the identity of these proteases, their proteolytic characteristics, and the conditions that elicit their expression. We will answer questions including how protease abundance is altered by bacterial culturing conditions, what are the cleavage sites of *B. vulgatus* proteases, and how comparable is the proteolysis of UC fecal sample supernatant and cultured *B. vulgatus* supernatant? A detailed biochemical characterization of *B. vulgatus* proteases will surely aid successive studies that detail proteolytic mechanisms associated with UC disease severity.

Experiment 1.1.1 – Defining proteases unique to *B. vulgatus* and their abundances in different growth conditions: The role of nutrient broth conditions on *B. vulgatus* protease production has not been established. To answer this, we will collect the supernatant of *B. vulgatus* grown under standard anaerobic growth conditions (BHI-S liquid media), aerobic growth in mammalian cell culture media (DMEM +CO2), as well as minimal media broth, all in triplicate. After all supernatant is collected, we will leverage our highly-developed quantitative multiplexed proteomics platform to quantitatively compare all detected *Bacteroides* proteases. The proteomics experiments will be cross-validated with sample matched RNAseq.

Experiment 1.1.2 – Quantifying and characterizing *B. vulgatus* protease activity in culture supernatants: In order to quantify proteolytic potential of *B. vulgatus*, we will first grow *B. vulgatus* under the same conditions previously described in subaim 1.1.1. We will then measure extracullar protease activity in these conditions using standard proteolytic activity assays (EnzCheck Protease activity assay, Molecular Probes)(Popov et al. 2005). To more fully characterize proteases present, we will also compare supernatant proteolytic activity when one (or a combination) of several major protease inhibitor classes to that of uninhibited supernatant. These inhibitors will include a serine protease inhibitor 4(2-Aminoethyl)benzenesulfonyl Fluoride (MP Biomedicals), an aspartic acid proteinase inhibitor Pepstatin A (MP Biomedicals), a metalloprotease inhibitor GM6001 (EMD Millipore), and a cysteine protease inhibitor E-64 (Sigma). Preliminary results have been gathered suggesting primarily serine protease activity of *B. vulgatus* supernatant in overnight cultures using BHI-S broth and anaerobic growth (Fig. 6.1).

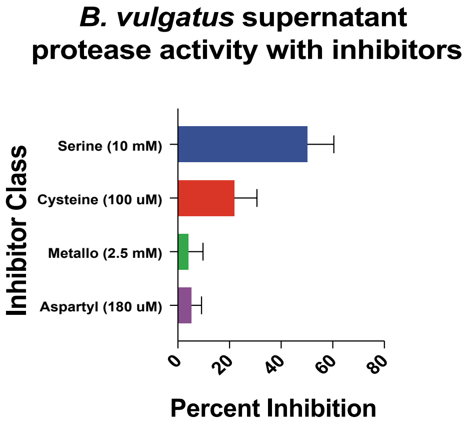


Figure 6.1 Characterizing protease activity in *B. vulgatus* supernatant. Supernatant from overnight cultures of *B. vulgatus* were concentrated and tested for protease activity in the presence of different protease inhibitors.

Experiment 1.1.3 – Determine the preferred cleavage sites of *B. vulgatus* supernatant proteolysis: While the previous aim will determine the type and conditions eliciting *B. vulgatus* protease activity, identifying the patterns of sequences surrounding *B. vulgatus* proteolysis sites will provide a fingerprint to determine *B. vulgatus* proteolysis in patients. In order to determine this, we will catalog the cleavage patterns present in *B. vulgatus* supernatant by utilizing technology our group recently developed to characterize protease activity(Lapek et al. 2019). We will first add *B. vulgatus* supernatant to beads containing a library of bound but proteolytically cleavable peptide substrates. After incubation, we will determine the identity of the peptides present. This will allow us to describe in detail *B. vulgatus'* di-peptidase, exo-, and endopeptidase activity. We will compare these results to those of the same peptides incubated with *B. dorei* or *B. thetaiotamicron.*

Experiment 1.1.4 – Determine patient extracellular proteolysis potential: To test whether *B. vulgatus* supernatant-resident proteases are packaged in outer-membrane vesicles (OMVs), we will first isolate outer membrane vesicles (OMVs) from UC patient fecal samples with high or low levels of *B. vulgatus* proteases as well as control subject stool samples. These OMVs will be characterized via our quantitative TMT-labeled mass spectrometry pipeline. In parallel, we will isolate a second set of OMVs from the stool samples and their activity type will be evaluated using our previously described assays from subaims 1.1.2 and 1.1.3.

We expect this experimental outline will provide a foundational framework for our understanding of what proteases are present in *B. vulgatus*, alongside the conditions eliciting *B. vulgatus* protease secretion and the most common type of protease activity. However, as shown by our preliminary data, we fully expect the conditions listed in subaim 1.1.1 to reveal protease activity. It is possible that the protease activity we identify *in vitro* will not match the protease activity of UC samples. In this case, we will consider other factors that influence protease expression, such as the presence of other microbes or input from a host-like environment. To this end we can consider multiple-microbe cultures, as well as alternative mammalian culturing systems such as intestinal organoids, which our collaborators have extensive experience with. Additionally, while we fully expect subaim 1.1.3 will allow us to determine the cleavage patterns of proteases expressed by *B. vulgatus* as well as potential inhibitors, one potential confounder is that *B. vulgatus* proteases may not abide by canonically predicted cleavage patterns and thus our designed peptides could be cleaved multiple times or cleavage patterns could be attributed to the incorrect protease(s). In order to control for this, we would test them individually to ascertain their proteolytic cleavage patterns. To ensure experimental rigor, we will perform the above described experiments in Aim 1 with at minimal two other *B. vulgatus* strain types (additional ATCC strains or clinical isolates banked by our team).

Determining functional outputs of *B. vulgatus* protease mutants *in vitro*.

A combination of quantitative measures obtained from our multi-omics data sets and *in vitro* investigations into protease inhibitor classes has pinpointed a set of proteases that we predict are responsible for *B. vulgatus* intestinal epithelium penetration. Based on our strong preliminary data, we leverage our expertise in *Bacteriodes* genetics to now generate deletion mutants and their associated complemented strains to determine how specific *B. vulgatus* proteases contribute to UC pathology. Several fundamental questions will be addressed: Is one protease sufficient for inducing a disease phenotype or are several needed? What is the specific pathology associated with each *B. vulgatus* protease? Through the generation of *B. vulgatus* mutant strains, as well as strains of *B. thetaiotaomicron* containing expression vectors encoding *B. vulgatus* proteases, we can start to answer these important mechanistic questions in a controlled and rigorous fashion.

Experiment 1.2.1 – Generation of *B. vulgatus* protease mutant strains and controls: To further evaluate the correlations between *B. vulgatus* proteases and UC disease activity, we have performed a metaproteomic analysis on a second cohort of UC pateint fecal samples and summarized the overlapping correlations to disease activity in Figure 6.2. In this experiment we have selected five proteases for mutagenesis based upon the preliminary results and a thorough literature review. First, we chose dipeptidyl peptidase IV, our top ranked peptidase, consitently found correlated to UC activity (Fig.

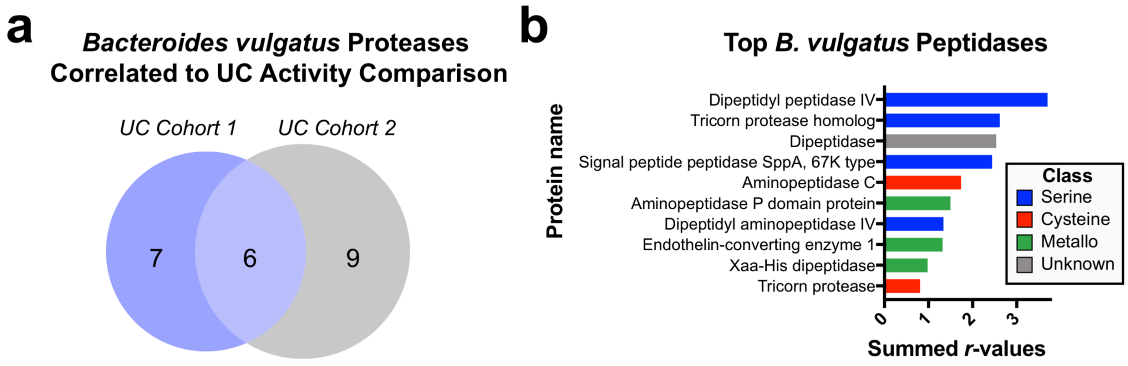


Figure 6.2 Prioritizing *B. vulgatus* proteases by their associations to UC disease activity. a, *B. vulgatus* or *B. dorei* enzymes or peptidases correlated with UC disease activity ($r > 0.3$), were compared by name between two separately collected UC patient cohorts. **b,** The correlation values for each protein from (a) were summed and the top 10 peptidases are shown according to the class of peptidase.

6.2). Further, it is an ortholog to a virulence factor in *P. gingivalis* and cleaves dipeptides ending in proline or alanine from N-terminal polypeptides(Kumagai et al. 2000). Supporting the activity of this protease were four dipeptides which significantly correlated with disease severity, including two X-Pro species. Conversely, given the specificity of previous results implicating *B. vulgatus* in UC pathology, we aim to create deletion mutants of a protease within the M28 family as the MEROPS database (https://www.ebi.ac.uk/merops/) identifies these proteins to be lineage specific. Mutants will be created using the ATCC *B. vulgatus* strain. We will isolate genomic DNA using manufacture recommended methods (Promega Wizard Genomic DNA Purification kit). Gibson assembly will be used (NEB HiFi DNA assembly master mix) to create vectors. The *E. coli* DC10B strain will be used to generate plasmids to introduce into *B. vulgatus* via electroporation. Mutant strains will be sequenced to ensure no off-target effects occurred during the strain generation process. As *B. thetaiotaomicron* did not induce any pathogenic phenotypes in our preliminary studies, we will additionally generate strains of *B. thetaiotaomicron* containing or not containing one of five *B. vulgatus* protease expression vectors to determine the individual effects of each protease.

Experiment 1.2.2 – Determine the effect of protease mutants in Caco-2 transwell assays: In order to test the role of *B. vulgatus* proteases in intestinal barrier penetration, we will use a classic model of epithelial integrity, the Caco-2 transwell assay. We will first collect *B. vulgatus* cells or supernatant grown under conditions tested in experimental section 1.1 and incubate Caco-2 monolayer cell cultures with either the supernatant or the microbes. We will then test for Caco-2 monolayer barrier integrity using transepithelial electrical resistance. In parallel, we will test whether *B. vulgatus*-protease-expressing *B. thetaiotaomicron* strains engineered in experiment 1.2.1 can reduce epithelial integrity similar to that of *B. vulgatus*. All measurements will be done in triplicate using methodology established in our preliminary results (Chapter 4).

Experiment 1.2.3 – Determine the effect of *B. vulgatus* protease mutants in patient derived colonic organoid models: In order to better understand the role of *B. vulgatus* proteases on a more complex and physiologically relevant experimental system, we will co-culture *B. vulgatus* WT, protease mutants and previously described *B. thetaiotamicron* engineered strains with patient derived enteroids.

Our team has ample experience in bacterial genetics and therefore we do not anticipate any major hurdles with making the deletion mutants. We anticipate that these experiments will provide a library of *Bacteroides* mutants that will guide our understanding of the mechanistic effects of specific proteases in the context of UC models. Genetic modifications of *B. vulgatus* will be rigorously investigated to define any off-target effects and experimental reagents will be authenticated prior to use. If positive results are obtained, we will make mutants in two other *Bacteriodes* strains to validate our results. In the case of potential issues with patient derived organoid experiments, our collaborative team also has established mouse-derived enteroids which provide an alternative approach.

Elucidating phenotypes of *Bacteroides vulgatus* proteases *in vivo*.

Results from our lab previously revealed germ-free (GF) mice inoculated with fecal material from UC patients with high levels of *B. vulgatus* proteases exhibited colitis. While cell culture models (monolayers and organoid) are often useful to test initial conditions, they do not capture the complexity of an entire host. This aim is designed to determine a direct correlation between how specific *B. vulgatus* proteases contribute to animal models of colitis. We will use a combination of outputs related to tight-junction integrity, cytokine profiles and histological disease scoring, which allow for a detailed investigation of pathological phenotypes.

Experiment 1.3.1 – Determine the role of *Bacteroides* protease mutants in a monocolonization mouse model. Previous work has shown that strains of *B. vulgatus* can induce colitis phenotypes in genetically susceptible mice(Bloom et al. 2011). Using this mouse model, we will test colitis phenotypes of our genetically engineered strains of *B. vulgatus* containing deletions in proteases of interest. Common metrics such as macro and microscopic damage scores, colon thickness, granulocyte infiltration will be collected alongside stool pellets for confirmation of protease expression. We will also perform FITC-Dextran-based gut permeability assays to evaluate intestinal epithelial integrity. This experiment will help evaluate if proteases of interest have *in vivo* relevance.

Experiment 1.3.2 – Determine the role of *Bacteroides* protease mutants in an IL10$^{-/-}$ monocolonization model: The role of IL-10, a canonical anti-inflammatory cytokine, in UC has been established in the literature. In order to test the role of *B.*

vulgatus proteases in an IL-10 knockout mouse model, we will inoculate GF IL10$^{-/-}$ mice (n=3) with protease mutants. To confirm the specificity of these findings, we will introduce gain of function *B. thetaiotaomicron* strains expressing or not expressing *B. vulgatus* mutants into a separate set of GF IL10$^{-/-}$ mice and perform the same tests as described above. Similar to the previous aim, we will record physiological metrics as well as FITC-Dextran-based gut permeability assays.

We anticipate this subaim will be a step forward in determining the host immune response against *B. vulgatus* proteases. We have experience in animal models and have the proper infrastructure to perform experiments in the discussed mouse models. However, it is possible that a deletion of one protease will not be sufficient to induce significant results *in vivo*. In this instance, we will thoroughly characterize the *Bacteroides* and protease content of UC patient fecal samples and apply fecal transplantation studies as previously shown (Chapter 4).

<u>Determine how nutrient availability impacts the expression patterns of *B. vulgatus* proteases.</u>

One critical gap in our knowledge of the role that *B. vulgatus* proteases play in UC pathology is the environmental signal that promotes increased protease production in severe cases of UC. While *Bacteroides* are known to play roles in digesting a large number of carbohydrates, their role in the breakdown of proteins is largely unexplored. Mounting evidence suggests that there is a strong link between the fluctuation in gut microbiome communities and diet. Therefore, we hypothesize *B. vulgatus* increases the secretion/production of proteases due to changes in nutrient (microbe accessible carbohydrates, fats, or proteins) availability. As such, the goal of this aim is to determine whether a change in environmental availability of nutrients induces increased expression or activity of *B. vulgatus* protease production.

Experiment 1.4.1 – Monitoring protease expression related to broth protein content *in vitro*: While *Bacteroides* are traditionally thought to rely heavily on carbohydrates *in vivo*, traditional BHI-S broth contains a significant amount of protein. Here we aim to characterize the effect that peptone content in BHI-S broth has on protease activity in *B. vulgatus*. While traditional broth contains 14.5 g/L casein peptone, we will additionally test the protease activity in supernatant of *B. vulgatus* cultures when a BHI-S broth contains 0, 5, 20, and 25 g/L peptone. Cells will be normalized by optical density before testing protease activity as previously shown (Fig. 6.1).

Experiment 1.4.2 – Identify how protein availability influences *Bacteroides* protease expression *in vivo*: Given the increase in protease abundance we noted in our previous research, we will also compare mice on standard rodent diets to a separate cohort (n=8) on a high-protein content diet. Similar to the previous aim, mice will be kept on a high protein diet for 2 weeks and, similar to the previous aim, stool pellets will be collected throughout the experimental timeline. From these samples, we will monitor for expression changes in *B. vulgatus* proteases through metatranscriptomics.

Experiment 1.4.3 – Establish the effect of dietary fat content on protease abundance *in vivo*: Prior research has suggested that a high fat diets (HFD) are a risk factor for the development of UC, and can alter protease activity. To test how HFD affects *B. vulgatus* protease expression, we will subject one group of GF mice colonized with a moderate complexity community (n=8) to a HFD and another group of mice (n=8) to standard chow (same group from 1.4.2). Stool pellets will be collected over the course of two weeks. Changes in the stool proteome of all these conditions, including *B. vulgatus* protease abundance will be characterized via quantitative multiplexed mass spectrometry.

We expect that in at least one of these experiments we will observe differential expression of protease production and/or activity, leading to conditions that will impact disease outcomes. However, it may be that the experimental design chosen does not elicit the robust protease response from *B. vulgatus*. If this is the case, we will further investigate the proteome and transcriptome of *B. vulgatus* to identify any putative transcriptional regulators that are co-expressed with our targeted proteases. If we find highly correlated genes that are putative transcriptional regulators, we will use a genetic approach to delete the genes of interest and functionally test the mutants for protease expression.

6.3 Interrogate mechanisms of host response to *B. vulgatus* proteases

Identify *B. vulgatus* proteases with mucin or collagen degrading activity.

Our prior results identified several *B. vulgatus* proteases that may act on host protein substrates. However, the host targets and downstream effects of specific *B. vulgatus* proteases remain unknown. This aim will reveal, in detail, how these proteases act on epithelial barriers including tight-junction proteins, mucin-family proteins and collagens. We will do so by leveraging readily available commercial kits that can characterize *B. vulgatus* protease activity against substrates we previously identified to be degraded, and by the use of innovative mass-spectrometry based peptidomic approaches developed by our team.

Experiment 2.1.1 – Determine *B. vulgatus* proteases with collagenase activity: We hypothesize that *B. vulgatus* contains proteases that degrade collagen present in the colon, potentially leading to increased intestinal permeability and immune cell infiltration. To test this, supernatent of *B. vulgatus* mutants from subaim 1.2 will be collected from liquid cultures. The group of mutants includes all *B. vulgatus* protease mutants (n=5) as well as *B. thetaiotaomicron* supernatant with and without plasmid-based expression of all *B. vulgatus* proteases (n=5). These cultures will be grown in triplicate and sufficient supernatant will be collected to run each condition in triplicate. This supernatant will be used to screen for proteolytic activity against collagen using the EnzCheck Collagenase assay kit (Molecular Probes)(Popov et al. 2005).

Experiment 2.1.2 – Determine *B. vulgatus* proteases with activity against mucins: Similar to the previous subaim, we hypothesize *B. vulgatus* contains proteases that degrade mucus present in the colon, potentially leading to increased intestinal permeability and immune cell infiltration. To test this, supernatant of *B. vulgatus* mutants from subaim 1.2.1 will be collected from liquid cultures. This group of mutants includes all *B. vulgatus* protease mutants as well as *B. thetaiotaomicron* supernatant with and without plasmid-based expression of all *B. vulgatus* proteases. Sufficient supernatant will be collected to run each condition in triplicate. This supernatant will be used to screen for proteolytic activity against mucin family proteins commonly expressed in the colon including MUC-2, 13, 20, and 21. This using purified mucins and previously established protocols(Desai et al. 2016). To account for non-specific proteolytic activity and protein degradation, we will also include a media only incubation condition.

We expect this subaim will reveal how *B. vulgatus* proteases interact with host proteins critical for epithelial defense from microbial invasion. One potential issue we may encounter is whether or not *B. vulgatus* mutant strains will produce sufficient levels of the protease of interest in order to make a detectable impact on host targets. In this case, molecular cloning will be undertaken for expression and purification of proteins to be used in these assays.

Identification of host targets of *B. vulgatus* proteases through a peptidomic approach.

Our previous results using UC patient fecal samples identified several potential host targets of *B. vulgatus* proteases. These included barrier proteins like mucins and collagens. Additionally, our *in vitro* results with Caco-2 cells suggested an effect of *B. vulgatus* proteases on tight-junction proteins. In this subaim, we will further elucidate the proteolytic landscape associated with *B. vulgatus* proteases through our previously established peptidomic approach(Quinn et al. 2019a).

Experiment 2.2.1 – Determine the mammalian proteolytic fragments within *in vitro* studies: Our lab has repeatedly found utility in adopting novel mass-spectrometry based peptidomic approaches, either for the identification of novel peptide virulence factors(Gonzalez et al. 2012; Gonzalez et al. 2014; O'Neill et al. 2020), or for inferences regarding proteolysis(Quinn et al. 2019b). While our results from UC patients provided several interesting potential targets, a more direct profiling of the proteolytic targets of *B. vulgatus* is still needed. Here, we will adapt our metapeptidomic methods utilized to identify proteolysis in UC patient samples (Fig. 4.2) to our *in vitro* studies with Caco-2 cells. The supernatant from five replicates of Caco-2 transwells co-cultured with *B. vulgatus*, *B. thetaiotaomicron*, and media controls will be collected for peptide extraction and mass-spectrometry analysis. Host-derived peptides will be compared with UC patient results to find overlapping and novel targets of *B. vulgatus* derived proteolysis.

Experiment 2.2.2 – Elucidating time-resolved dynamics of *B. vulgatus* mediated proteolysis in *vitro*: Building upon Experiment 2.2.1 we aim to further our understanding of *B. vulgatus* mediated proteolysis of Caco-2 proteins in a time-dependent manner. Our preliminary results showed that 36-hours post inoculation, *B. vulgatus* exerts significant proteolysis on Caco-2 cells. Here we will expand this result and profile peptide fragment dynamics present in *B. vulgatus*-Caco-2 transwell supernatant at 12, 24, 28, 32, 36, and 40 hours post-innoculation using our previously described peptidomic approach (Experiment 2.2.1). With this approach, we aim to determine the order and speed at which *B. vulgatus* can degrade host barriers.

Experiment 2.2.3 – Determine the mammalian proteolytic fragments within *in vivo* studies: This experiment will use metapeptidomic methodology used in UC patient fecal samples in our proposed *in vivo* studies. Here we aim to tie together the peptide results of *in vitro* experiments (2.2.1, 2.2.2) and our preliminary results in UC patient samples to our proposed *in vivo* experiments (1.3, 3.2). Peptide extraction and mass-spectrometry analysis will be performed alongside PEAKS de-novo peptide identification(Zhang et al. 2012), lowest common ancestor analysis(Mesuere et al. 2015), and cleavage site analysis. We expect these approaches to provide a benchmark for how well the proteolysis of our experimental models fits the proteolysis occurring in UC patients.

Our experimental design aims to answer key questions regarding how well our models of *B. vulgatus* induced proteolysis mimic what is actually occurring in patients. Additionally, the untargeted nature of these experiments may identify additional host-targets of proteolysis for *B. vulgatus* proteases. However, the untargeted nature of the peptidomic approaches proposed may result in challenges in the consistent identification of important peptide fragments. An alternative approach would be to develop a targeted mass-spectrometry approach for detecting known peptide fragments of interest. Further, the limit of detection and peptides present in media may present further obstacles to the proposed techniques that might be overcome by enrichment, depletion or targeted mass-spectrometry techniques.

6.4 Determine the impact of protease inhibition in models of IBD

Identify protease inhibitors that prevent *B. vulgatus* epithelial monolayer disruption/penetration.

Our preliminary research suggests that protease inhibition can prevent *B. vulgatus* induced epithelial monolayer penetration. However, the protease inhibitor used in these studies contained a proprietary blend of serine and cysteine protease inhibitors, making it less useful for identifying specific inhibitors with therapeutic potential or isolating a protease's unintended consequences to host-health. As such, the goal of this subaim is to further characterize classes of protease inhibitors that inhibit *B. vulgatus* proteases, with the hope that a novel therapeutic approach of protease inhibition may one day be used for the treatment of UC.

Experiment 3.1.1 – Determine protease inhibitors most effective at preventing *B. vulgatus* induced epithelial penetration: We previously showed the efficacy of the Roche cOmplete EDTA-free protease inhibitor cocktail (Sigma) to prevent *B. vulgatus* epithelial monolayer penetration. We aim to expand our experiments to include a serine protease inhibitor, 4(2-Aminoethyl)benzenesulfonyl Fluoride (MP Biomedicals), an aspartic acid proteinase inhibitor, Pepstatin A (MP Biomedicals), a metalloprotease inhibitor, GM6001 (EMD Millipore) and a cysteine protease inhibitor, E-64 (Sigma). Using this preliminary data, we will further priortize and expand the use and testing of protease inhibitor classes identified as active in *B. vulgatus* supernatant (Experiment 1.1.2).

Experimental 3.1.2 – Evaluate the host effects of protease inhibitors *in vitro*: After identifying the most effecacious protease inhibitor class in experimental 3.1.1, we will test 5 additional protease inhibitors of the same class. These protease inhibitors will be selected based on potential toxicity in mammalian cells. To evaluate the potential toxicity to mammalian cells, we will monitor growth rates of mature polarized Caco-2 monolayer cells as well as transcriptional profiles with or without the addition of each selected protease inhibitor. Each experimental condition will be done in triplicate in order to increase measurement robustness.

We expect this aim will further elucidate the most efficacious inhibitor class as well as determine the inhibitor's ability to impact the *in vitro* system's viability. However, similar to prior aims, one potential issue is that the use of these inhibitors will likely only allow for a low-resolution understanding of what *B. vulgatus* proteases are active during the experiments, limiting their usefulness. In this case, after the general efficacy of inhibitor class is analyzed, *B. vulgatus* proteases that fall into this category can be either expressed in lab strain *E. coli* and collected, or synthesized and tested individually if *E. coli* expressed protein yields are low.

<u>Evaluate the efficacy of newly selected inhibitors to prevent colitis *in vivo*.</u>

Our preliminary research suggests that protease inhibition can prevent colitis phenotypes in UC patient fecal transplant studies in germ-free IL10$^{-/-}$ mice. However, these inhibitor cocktails were proprietary blends of serine and cysteine protease inhibitors, and may have unintended consequences to host-health. We aim to further characterize classes of protease inhibitors that inhibit *B. vulgatus* proteases, with the hope that a novel therapeutic approach of protease inhibition may one day be used for the treatment of UC.

Experiment 3.2.1 – Evaluation of protease inhibitors in a fecal transplant model: Our preliminary research showed the capacity of fecal samples from UC patients with high clinical severity scores to induce colitis in IL10$^{-/-}$ gnotobiotic mice. We aim to use these methods to evaluate the efficacy of two additional protease inhibitors selected from our *in vitro* experiments described in experimental 2.1.2. Our experiments will be performed as previously described with additional measurements taken to evaluate side-effects of each protease inhibitor.

Experimental 3.2.2 – Evaluate the host effects of protease inhibitors *in vivo*: Given that the expense of IL10$^{-/-}$ gnotobiotic mice is a limiting factor, we will evaluate the effect of protease inhibitors on host health using conventional mice. This will allow for a greatly expanded capacity to elucidate any host-related effects of a large number of protease inhibitors for their host effects.

We hypothesize that these experiments will help inform the design of future therapeutics targeting *B. vulgatus* proteases. Given that the methods have been previously established, we do not anticipate major problems. However, it is possible that protease inhibitors chosen will not be efficacious, in which case new inhibitors within the same family will be selected.

Chapter 6 contains preliminary ideas and writing to form the basis of a grant application. The dissertation author played a primary role in the conceptualization and writing of this section. This work also contains editing contributions from Carlos Gonzalez and David J. Gonzalez.

6.5 References

Bloom SM, Bijanki VN, Nava GM, Sun L, Malvin NP, Donermeyer DL, Dunne WM, Jr., Allen PM, Stappenbeck TS. 2011. Commensal Bacteroides species induce colitis in host-genotype-specific fashion in a mouse model of inflammatory bowel disease. *Cell Host Microbe* **9**: 390-403.

Desai MS, Seekatz AM, Koropatkin NM, Kamada N, Hickey CA, Wolter M, Pudlo NA, Kitamoto S, Terrapon N, Muller A et al. 2016. A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility. *Cell* **167**: 1339-1353 e1321.

Elhenawy W, Debelyy MO, Feldman MF. 2014. Preferential packing of acidic glycosidases and proteases into Bacteroides outer membrane vesicles. *MBio* **5**: e00909-00914.

Foley MH, Cockburn DW, Koropatkin NM. 2016. The Sus operon: a model system for starch uptake by the human gut Bacteroidetes. *Cell Mol Life Sci* **73**: 2603-2617.

Gonzalez DJ, Okumura CY, Hollands A, Kersten R, Akong-Moore K, Pence MA, Malone CL, Derieux J, Moore BS, Horswill AR et al. 2012. Novel phenol-soluble modulin derivatives in community-associated methicillin-resistant Staphylococcus aureus identified through imaging mass spectrometry. *J Biol Chem* **287**: 13889-13898.

Gonzalez DJ, Vuong L, Gonzalez IS, Keller N, McGrosso D, Hwang JH, Hung J, Zinkernagel A, Dixon JE, Dorrestein PC et al. 2014. Phenol soluble modulin (PSM) variants of community-associated methicillin-resistant Staphylococcus aureus (MRSA) captured using mass spectrometry-based molecular networking. *Mol Cell Proteomics* **13**: 1262-1272.

Kumagai Y, Konishi K, Gomi T, Yagishita H, Yajima A, Yoshikawa M. 2000. Enzymatic properties of dipeptidyl aminopeptidase IV produced by the periodontal pathogen Porphyromonas gingivalis and its participation in virulence. *Infect Immun* **68**: 716-724.

Lapek JD, Jr., Jiang Z, Wozniak JM, Arutyunova E, Wang SC, Lemieux MJ, Gonzalez DJ, O'Donoghue AJ. 2019. Quantitative Multiplex Substrate Profiling of Peptidases by Mass Spectrometry. *Mol Cell Proteomics* **18**: 968-981.

Mesuere B, Debyser G, Aerts M, Devreese B, Vandamme P, Dawyndt P. 2015. The Unipept metaproteomics analysis pipeline. *Proteomics* **15**: 1437-1442.

O'Neill AM, Nakatsuji T, Hayachi A, Williams MR, Mills RH, Gonzalez DJ, Gallo RL. 2020. Identification of a Human Skin Commensal Bacterium that Selectively Kills Cutibacterium acnes. *J Invest Dermatol* doi:10.1016/j.jid.2019.12.026.

Popov SG, Popova TG, Hopkins S, Weinstein RS, MacAfee R, Fryxell KJ, Chandhoke V, Bailey C, Alibek K. 2005. Effective antiprotease-antibiotic treatment of experimental anthrax. *BMC Infect Dis* **5**: 25.

Quinn RA, Adem S, Mills RH, Comstock W, DeRight Goldasich L, Humphrey G, Aksenov AA, Melnik AV, da Silva R, Ackermann G et al. 2019a. Neutrophilic proteolysis in the cystic fibrosis lung correlates with a pathogenic microbiome. *Microbiome* **7**: 23.

Quinn RA, Vrbanac A, Melnik AV, Patras KA, Christy M, Nelson AT, Aksenov A, Tripathi A, Humphrey G, da Silva R et al. 2019b. Chemical Impacts of the Microbiome Across Scales Reveal Novel Conjugated Bile Acids. *bioRxiv* doi:10.1101/654756: 654756.

Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B. 2012. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* **11**: M111 010587.