Quantitative Metaproteomics and Activity-Based Probe Enrichment Reveals Significant Alterations in Protein Expression from a Mouse Model of Inflammatory Bowel Disease

Michael D. Mayers,† Clara Moon,† Gregory S. Stupp,‡ Andrew I. Su,*†‡ and Dennis W. Wolan,*†§

†Department of Molecular and Experimental Medicine, ‡Department of Integrative Structural and Computational Biology, and §Department of Chemical Physiology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

Supporting Information

ABSTRACT: Tandem mass spectrometry based shotgun proteomics of distal gut microbiomes is exceedingly difficult due to the inherent complexity and taxonomic diversity of the samples. We introduce two new methodologies to improve metaproteomic studies of microbiome samples. These methods include the stable isotope labeling in mammals to permit protein quantitation across two mouse cohorts as well as the application of activity-based probes to enrich and analyze both host and microbial proteins with specific functionalities. We used these technologies to study the microbiota from the adoptive T cell transfer mouse model of inflammatory bowel disease (IBD) and compare these samples to an isogenic control, thereby limiting genetic and environmental variables that influence microbiome composition. The data generated highlight quantitative alterations in both host and microbial proteins due to intestinal inflammation and corroborates the observed phylogenetic changes in bacteria that accompany IBD in humans and mouse models. The combination of isotope labeling with shotgun proteomics resulted in the total identification of 4434 protein clusters expressed in the microbial proteomic environment, 276 of which demonstrated differential abundance between control and IBD mice. Notably, application of a novel cysteine-reactive probe uncovered several microbial proteases and hydrolases overrepresented in the IBD mice. Implementation of these methods demonstrated that substantial insights into the identity and dysregulation of host and microbial proteins altered in IBD can be accomplished and can be used in the interrogation of other microbiome-related diseases.

KEYWORDS: metaproteomics, SILAM, activity-based probes, microbiome, MudPIT, quantitative proteomics, ComPIL, GO term enrichment

INTRODUCTION

Liquid chromatography coupled with tandem mass spectrometry (LC−MS/MS) is a powerful technique now being employed by researchers to determine the functional makeup of the highly complex proteomic contents of intestinal microbiota.1,2 Such studies are providing key information into the kinds of proteins most abundantly expressed by the gut microbiome. For example, nontargeted shotgun metaproteomics on samples prepared from healthy human distal gut microbiota identified many microbial proteins primarily involved housekeeping functions including translation, carbohydrate metabolism, and energy production.1 More recently, integration of metagenomics with proteomics was performed to elucidate the phylogenetic alterations and accompanying functional changes in microbial proteins from the gut microbiota of patients suffering from Crohn’s disease.2 These results and others demonstrate that the application of proteomics to the study of highly complex microbial proteomes yields compelling insights into proteins expressed and functional characterization.3–5

Despite the increasing application of LC−MS/MS analyses on gut microbiomes, two key limitations in metaproteomics remain. First, rigorous quantitation of microbiome proteomics data obtained through shotgun-based methods has not been performed to date. Differential protein expression has been...
measured exclusively by spectral counting, but this methodology is only semiquantitative due to limitations in data-dependent acquisition and low numbers of spectral counts for low-abundance proteins. Furthermore, significant differences in MS1 retention times and intensities for candidate peptides, used to perform spectral-based quantitation, are typically observed across samples and are due to variability in liquid chromatography.6−8 Metabolic labeling techniques help to overcome the many issues associated with spectral-based quantitation and limit systematic errors in sample preparation and LC−MS/MS proteomic data collection.

Quantitative proteomics between two biological states is readily accomplished in tissue-cultured cells and whole animals via isotope labeling.9 For example, stable isotopic labeling in mammals (SILAM) is a reliable method to accomplish measuring and quantitating proteomic differences in both mice and rats, whereby animals are restricted to a diet of isotopically 15N-labeled spirulina as the only source of nitrogen.10−14 Importantly, SILAM has shown utility in labeling and quantifying the whole organism including long-lived proteins of the brain.15 Despite these previous validations of SILAM, the isotope incorporation efficiency and quantification capabilities of microbiome constituents has yet to be determined. We wanted to build on the SILAM methodology and measure isotope incorporation in fecal samples of animals on a 15N diet by metaproteomics. Our primary goal was to determine if limiting environmental and genetic variables associated with gut microbial diversity would increase the likelihood of the identification and quantitation of differentially expressed proteins between isogenic control and diseased murine cohorts.

The second limitation in metaproteomics profiling is the high degree of complexity in microbiome samples. Recent proteomics-based advances have attempted to address the issue of the microbiome complexity by improving methods in sample preparation6,17 database search strategies and algorithms,16,19 and implementation of proteogenomics.20,21 Despite these extensive improvements to metaproteomics, the
sensitivity of LC–MS/MS instrumentation constrains data collection to a small percentage of the most abundant peptides in a proteomic sample. We and others have begun to apply prefractonization "enrichment" steps to circumvent instrument sensitivity issues and access low-abundant biologically important proteins within a complex sample.

One of the most efficient methods of "enrichment" includes the application of activity-based probes (ABP) that target specific protein families. Addition of small-molecule ABPs permits systematic quantitation of individual classes of proteins potentially lower in abundance than the detectable limit and therefore missed by LC–MS/MS of whole microbiome-derived proteomes. Such chemical probes have already been designed to target more than a dozen protein classes (i.e., hydrolases, proteases, kinases, phosphatases, and glycosidases) and have been successfully employed in identification of dysregulated proteins in cancerous tumors, parasitic infections, and fatty livers. To address the limitations in MS-based identification of low-abundance proteins in highly complex and concentrated proteomic samples, we applied an ABP that covalently labels proteins with nucleophilic cysteine residues.

Here, we combine these two strategies, SILAM-based quantitation and application of ABP probes, to address these two key bottlenecks in metaproteomic analysis. We applied this integrated system to the study of a mouse model of IBD. Using this model, we could identify and quantitate differences in host and microbiome protein functionalities between mice with intestinal inflammation and isogenic controls fed identical diets. Our data show that both host and microbiome proteins can be identified and quantitated between two cohorts. These results also demonstrate new techniques that can be extrapolated to the metaproteomic interrogation of other animal model and human microbiomes. These findings provide a deeper perspective of the microbiome proteome and of the proteins that are altered in expression between the control and IBD mouse groups.

## MATERIALS AND METHODS

### Murine Adoptive T Cell Transfer Chronic IBD Model and Fecal Sample Collection

Animal protocols were approved by The Institutional Animal Care and Use Committee (IACUC) at The Scripps Research Institute (TSRI). All mice were purchased from Jackson Laboratories and maintained in a specific pathogen-free barrier facility at TSRI for the duration of the study.

We employed the well-established T cell transfer model of colitis to induce intestinal inflammation, as this method promoted rapid and reproducible intestinal inflammatory pathophysiology within an approximate 8-week time frame. Twenty 10-week-old female B6.129S7-Rag1tm1Mom/J (Rag1−/−) mice were purchased from Jackson Laboratories (stock #002216), separated into four cages (5 mice per cage), and cohoused for 1 week to normalize the microbiota. CD3+CD4+CD8+ Foxp3+EGFP naïve T cells collected via fluorescence-activated cell sorting from the spleens of donor Foxp3-EGFP reporter mice (Jackson Laboratories, stock #006772) were transferred retro-orbitally to 10 Rag1−/− mice (approximately 5 × 10^7 T cells per mouse) (hereafter referred to as "IBD mice"). The control Rag1+/− mice (hereafter referred to as "RAG−/− mice") were injected retro-orbitally with a similar volume of sterile PBS and subsequently separated for the remainder of the time course from the IBD cohort.

Fourteen days post T cell transfer, the mice were switched from standard chow (Harlan Teklad) to a spirulina-based diet, with the RAG−/− mice receiving 15N labeled spirulina and the IBD mice receiving a 14N control spirulina diet (Cambridge Isotope Laboratories) (Figure 1A). The "heavy" 15N-labeled diet has no effect on food consumption, weight gain, normal physiology, reproduction, or development. Mice fecal samples were collected daily, pooled by cage and week, and immediately frozen at −80 °C for further use (Supporting Information, Table S1). Five IBD mice were sacrificed prior to the day 56 end point due to severe signs of disease and a decrease in 20% body weight with respect to day 0 weight (Supporting Information, Figure S1). The remaining five IBD mice exhibited signs of disease including reduced weight gain (Supporting Information, Figure S1), diminished grooming habits, and hunched posture. All IBD mice were confirmed to have intestinal inflammation by histology (Supporting Information, Figure S2) and were included in the experiment.

### Fecal Microbiome Sample Preparation for LC–MS/MS Analysis

Pooled fecal samples were thawed in 3 mL of cold PBS and vigorously vortexed to break up fecal pellets. Suspended samples were subjected to low-speed centrifugation (100g, 2 min, 4 °C) to separate insoluble fecal matter. The fecal supernatant containing suspended bacterial cells was aliquoted into 1 mL fractions. Fractions were pelleted (6500g, 15 min, 4 °C), washed twice with cold PBS, and suspended into 450 μL of lysis buffer consisting of 0.1% Triton-X 100 in PBS and Roche Complete protease inhibitor cocktail. Samples were lysed via sonication in a Qsonica Q700 sonicator with Cup Horn attachment at 4 °C for 15 min. Insoluble cellular material was removed via centrifugation (10,000g, 5 min, 4 °C) and the remaining soluble protein concentration was measured (Pierce BCA Protein Assay Kit). Prepared soluble microbiome samples from RAG−/− control and IBD mice were mixed 1:1 in aliquots containing 50 μg each 15N and 14N labeled proteome (Figure 1B). The proteomic content was precipitated with 30% w/v of trichloroacetic acid (TCA) and stored at 4 °C overnight. The microbiome proteome was pelleted (21,000g, 20 min, 4 °C) and washed twice with −80 °C acetone prior to trypsin digestion to generate peptides for LC–MS/MS data collection (Figure 1B).

### Fecal Microbiome Preparation with BioGlyCMK Probe Labeling for “Enriched” Samples

Fifty milligrams of RAG−/− (15N) and IBD (14N) mouse bacterial pellets derived from fecal matter (isolated and washed as described above) was resuspended in PBS and mixed in a 1:1 ratio in a total volume of 500 μL. The combined samples were subjected to labeling with 100 μM of a biotinylated glycine-chloromethyl ketone (BioGlyCMK) probe (1% DMSO for unlabeled controls) and incubated overnight at 4 °C under light agitation (see the Supporting Information for BioGlyCMK synthesis). Bacteria were pelleted (6500g, 15 min, 4 °C), washed twice with 1 mL of PBS to remove unreacted probe, and resuspended in 450 μL of PBS with 0.1% SDS. Samples were lysed via sonication in a Qsonica Q700 sonicator with Cup Horn attachment at 4 °C for 15 min. All lysed microbial samples were assessed for BioGlyCMK labeling by streptavidin blot and concentrations determined with colorimetric BCA assays. Proteins were denatured via dilution with 500 μL of 2% SDS in PBS followed by heating at 95 °C for 15 min. Denatured samples were diluted with 4 mL of PBS (0.2% SDS).
and incubated with 100 μL of high capacity streptavidin agarose beads (Pierce) overnight at room temperature. Beads were pelleted (500g, 2 min) and subjected to extensive washing with 0.2% SDS in PBS (1×), PBS (3×), and ultrapure water (3×) prior to trypsin digestion.

**Microbiome Protein Trypsin Digestion**

Both “unenriched” and “enriched” BioGlyCMK probe-labeled washed lysate pellets were subjected to trypsin digestion to generate peptides for MudPIT shotgun proteomics analysis.28 Proteins were resuspended and denatured in 60 μL of 8 M urea, 100 mM Tris-HCl, pH 8.0, introduced to 1 μL of 300 mM TCEP to reduce all disulfide bonds, and agitated for 20 min at 25 °C. The reduced thiols were then alkylated via addition of 7 μL of 500 mM 2-chloroacetamide and incubated with gentle agitation for 15 min at 25 °C while protected from light. Following alklylation, samples were diluted with 180 μL of 100 mM Tris-HCl, pH 8.0 to reduce the urea concentration to less than 2 M. General proteomic digestion was performed by addition of 2.5 μg of trypsin to each sample in the presence of 1 mM CaCl2 and incubation overnight at 37 °C. “Unenriched” samples were quenched with 13 μL of formic acid, centrifuged (21 000g, 20 min), and the supernatant was stored at −20 °C until LC–MS/MS analysis. BioGlyCMK-enriched samples were filtered via spin column (100 μm, 1 min) and the beads were washed twice with 50 μL of ultrapure H2O. All filtrates were combined and acidified with 17 μL of formic acid, centrifuged, and the supernatant was stored at −20 °C until LC–MS/MS analysis.

**Preparation of MudPIT LC Column**

Trypsin-digested peptides were loaded onto a biphasic MudPIT column (250 μm fused silica (Agilent), paced with 2.5 cm of 5 μm Aqua C18 resin followed by 2.5 cm of Partisphere strong cation exchange resin (SCX)). An analytical column was prepared from 100 μm fused silica pulled to a 5 μm tip by a micropipette puller (Sutter Instrument Company, Model P-2000). This column was then pressure loaded with 12 cm of 3 μm Aqua C18 resin.

**LC–MS/MS MudPIT Data Collection**

Standard MudPIT tandem mass spectrometry was performed using a Thermo LTQ-Orbitrap XL mass spectrometer. The sample and analytical columns were joined by a zero dead volume union (Waters). Peptides were eluted at 300 nL/min using an 11-step MudPIT program. Each step began with 1 min of 100% Buffer A (95% H2O, 5% acetonitrile, 0.1% formic acid), a 4 min salt pulse with 8% buffer C (500 mM ammonium acetate, 95% H2O, 5% acetonitrile, 0.1% formic acid), then 5 min 100% buffer A, followed by a 105 min gradient from 5 to 65% buffer B (20% H2O, 80% acetonitrile, 0.1% formic acid), and finally 5 min of 100% buffer A. The 4 min buffer C salt pulses (x) were as follows: 10%, 15%, 20%, 25%, 30%, 35%, 40%, 50%, 60%, 80%, 100% with the final pulse consisting of 90% buffer C and 10% buffer B. Precursor ions were recorded by scanning in the range of m/z 400.00–1800.00 with the FTMS analyzer and a resolution of 60 000. The top eight peaks were selected for fragmentation using HCD with normalized collision energy set to 35.0. Dynamic exclusion was enabled with exclusion duration set to 60.0 s.

**Peptide Identification with ComPIL**

Precursor and fragmentation ion data were extracted from the Xcalibur RAW files via rawXtract 1.9.9.2 (http://fields.scripps.edu/yates/wp/?page_id=17) in the MS1 and MS2 file formats. The MS2 spectra were scored with Blazmass 0.9993 against peptides of the Comprehensive Protein Identification Library (ComPIL) database, containing over 80 million proteins from multiple microbial database sources as well as human, mouse, and plant proteins.18 Both Blazmass and ComPIL source code are open source (https://github.com/sandipchatterjee/blazmass_compil). Settings for peptide scoring included: (1) a variable modification of oxidized methionine (+15.9949 Da), (2) a static modification for alkylated cysteine residues (+57.02146 Da), and (3) a precursor mass tolerance of 10 and 50 ppm tolerance for fragmentation ions. Filtering was performed using DTASelect 2.1.3 (http://fields.scripps.edu/yates/wp/?page_id=17), requiring two peptides per protein and a false discovery rate (FDR) of 1% with respect to proteins. The following parameters were used for filtering when run from the command line: “--quiet --brief --trypstat --modstat -y 2 -DM 10 --extra --dm --ppf 0.01 -p 2”. Samples containing a mixture of 14N and 15N peptides for ratio quantification were searched against a ComPIL database consisting of 14N and then 15N-labeled peptides. DTASelect filtering was performed on each search individually as well as the combined results, which produced three outputs for several downstream bioinformatic analyses.

**Peptide and Protein Quantification**

The 14N/15N ratio of each peptide was quantified using the program Census (available on the Integrated Proteomics Pipeline (IP2), http://goldfish.scripps.edu/). Census uses the results of DTASelect filtering along with the extracted MS1 spectral data to both determine the isotopic enrichment within a sample and to calculate the 14N/15N isotopic ratio for each identified peptide within one LC–MS/MS data collection experiment. Enrichment calculations were performed on filtered results searched only against the 14N database as to not bias the results toward 15N peptides.29 Census quantification on the combined 14N/15N filtering results file was performed with the default settings for 15N ratio quantification in high-resolution, including a 15 ppm tolerance for isotope extraction. Census filtering parameters were modified to disable iterative outlier analysis and exclude outliers with a p-value > 0.05. The median of the peptide ratios was centered to 1 via division by the inverse log of the median of the log-transformed ratios. Protein loci ratios were determined by computing a weighted average of the associated peptides, weighted by the regression factor (r), a value determined by Census to describe the confidence in ratio quantification.

**LC–MS/MS Data Analysis**

Code from the Microbiome Metaproteomics package (available at https://bitbucket.org/sulab/metaproteomics) was modified to incorporate DTASelect results for the individual filtering of the 14N and 15N database searches as well as the combined filter result and Census ratio quantification into each sample. The source code for this process, as well as all analyses is available in Python 3.5.2 (https://github.com/mmayers12/n15_mice/). Protein clustering, cluster taxonomy, and gene ontology (GO) term annotations were stored within ComPIL and assigned as previously described.18 Samples were grouped together by type to determine the average fold change for each protein locus. Significance values were determined by a Student’s t test, adjusted by the Benjamini-Hochberg correction.30
GO Term Enrichment Analysis

For determination of GO term enrichment between the sample preparation conditions, BioGlyCMK-enriched and “unenriched,” a Fisher’s Exact Test was performed. This was dependent on the number of annotation occurrences of a given term in the set of protein clusters (i.e., proteins with 70% sequence identity and common functionality), unified across experimental replicates for the sample preparation condition. To compare GO terms enriched in the microbial proteins between different biological conditions, IBD and RAG\(^{−/−}\), gseapy 0.7.0 (https://github.com/BioNinja/gseapy), a Python implementation of the Broad Institute’s Gene Set Enrichment Analysis (GSEA) algorithm was used.\(^{31}\) GO gene sets were generated from all identified protein clusters in a given sample type (BioGlyCMK-enriched or “unenriched”). Terms were subsequently filtered according to the msigdb guidelines: (1) large sets, defined as those containing more than half the total number of protein clusters identified, were removed; (2) sets with less than five members were removed; (3) child terms with the exact same protein cluster members as their parent were removed; and (4) sibling terms with the exact same protein cluster members as other siblings were removed such that only one sibling remained.

Taxonomy Analysis

Peptide spectral counts were normalized across all samples by a normalization factor of the total number of counts for one experiment divided by the median across all LC–MS/MS experiments. Peptide taxonomy search space was restricted to the proteins identifiable in a given sample. Analysis was performed at the phylum level. Each peptide was traced back to a phylum and if uniquely classifiable, the peptide was classified with a weight of normalized counts. Peptides without a discernible phylum (e.g., could belong to more than one) were discarded from analysis. The normalized counts were then used to determine an approximate fractional taxonomic makeup of the sample.

■ RESULTS AND DISCUSSION

Isotope-Labeled SILAM Mice Rapidly Incorporate \(^{15}\)N into Their Microbiomes

We first established the rate and extent of \(^{15}\)N isotope incorporation within the microbiomes of both RAG\(^{−/−}\) and IBD mice. Mice were placed on an isotopically enriched spirulina-based chow 14 days post T cell transfer, whereby the RAG\(^{−/−}\) control group (10 mice) received the \(^{15}\)N isotopically labeled spirulina and the IBD group (10 mice) were fed the corresponding unlabeled (natural abundance) \(^{15}\)N spirulina.\(^{10,11}\) LC–MS/MS MudPIT data collection and quantitative analysis were performed on the unenriched fecal bacterial samples within the first 24 h on the spirulina-based chow. Data collection and analysis of the 24-h time point demonstrated that the murine microbiomes incorporated a relatively high level of \(^{15}\)N with a median peptide isotope enrichment of approximately 84% (Figure 1C). However, this value is an estimate, as previous studies suggested a minimum of 500 quantifiable peptides for accurate assessment of enrichment. We identified 181 and 80 quantifiable peptides from the \(^{15}\)N IBD and \(^{15}\)N RAG\(^{−/−}\) cohorts, respectively. Despite the initial rapid rate of \(^{15}\)N incorporation, the rate slowed over the first week with 86% enrichment 4 days after the dietary change and 88% at the end of 7 days, as based on 569 and 923 peptides, respectively. Importantly, the acceptable level of enrichment required for quantification of 95% was attained within 4 weeks on the \(^{15}\)N spirulina diet. By the day 56 end point (or 6 weeks on the isotopically labeled spirulina chow), the mice reached 96% enrichment.

Previous metaproteomic-based microbiome studies have detected the presence of dietary proteins in fecal samples; however, these diet-related proteins are usually found in low abundance.\(^{3}\) For our experiment, the source of dietary protein originates entirely from a single prokaroyotic organism (e.g., spirulina) and we needed to verify that dietary proteins would remain in low abundance in the fecal samples. As such, our initial concern was that dietary proteins may be in such abundance as to dominate LC–MS/MS collection and significantly limit detection of any host- and microbiome-derived peptides. Mass spectrometry data obtained from the \(^{14}\)N IBD mice demonstrated these fecal samples to have a relatively low number of spirulina proteins with respect to all host and bacterial proteins (Supporting Information, Figure S3). The attributable presence of spirulina proteins in the IBD mice samples reached a 1.3% maximum of all identifiable proteins. Conversely, the abundance of spirulina proteins as a fraction of spectral counts in the \(^{14}\)N labeled RAG\(^{−/−}\) control samples were significantly more prominent than the IBD group. Day 1 samples reached as high as 72% of all LC–MS/MS measured spectral counts and reduced to approximately 20% throughout the remainder of the experiment (Figure 1E). Despite the high spectral count signals derived from spirulina proteins, the fraction of identifiable protein loci originating from spirulina samples remained fairly low, with the day 1 sample containing 18% spirulina loci and the remainder of collected fecal samples between 3 and 7% with respect to all host and microbial proteins (Figure 1D). We expected spirulina protein signals to dissipate to levels demonstrated for other dietary experiments allowing for detectable levels of microbial proteins. Despite these persistent proteins from the SILAM spirulina diet, we were able to quantitate many host and microbial proteins that are differntially present between control and IBD mice.

IBD Mice Overexpress Host Protease Inhibitors and Inflammatory Proteins

We focused metaproteomic data collection and analysis on week-8 end point samples when inflammation was most severe, and the differences in protein expression between the RAG\(^{−/−}\) control and IBD mice would be greatest. Samples were pooled and prepared for experimental replicates as described Supporting Information, Table S1. After MudPIT LC–MS/MS and data analysis with ComPIL, the average replicate contained 1989 ± 180 peptides with 1040 ± 136 corresponding matched proteins, and 1424 ± 86 peptides matching 800 ± 62 proteins from the IBD and RAG\(^{−/−}\) samples, respectively. This amounted to 3277 protein clusters identified across all replicates and conditions (Supporting Information, Table S2). To assess reproducibility among experimental replicates, samples were visualized by a hierarchical clustering dendogram based on the Jaccard distance calculated from presence or absence of peptides in a sample (Supporting Information, Figure S4). The samples clustered well with some evident batch effects. Close clustering was observed among the RAG\(^{−/−}\) control experimental replicates as well as for the IBD mouse samples suggesting differences in the proteomic makeup between control and IBD mice (Supporting Information, Figure S4).
Our increased detection of Akkermansia sp. and Lactobacillus sp. proteins in IBD microbiota samples correlates with several previous studies on humans and mice. Unfortunately, there are conflicting reports as to the prominence and importance of both genuses with respect to microbiomes of healthy and inflamed states. Akkermansia muciniphila has been determined to decrease in abundance in IBD patients, as assessed by 16S rRNA sequencing; however, the bacterium was demonstrated to promote gut inflammation in mice infected with Salmonella enterica. Similarly, despite the use of members of the Lactobacillus sp. as probiotics to fight intestinal inflammation, increases in the genus as well as Bifidobacterium were measured by 16S rRNA gene rtPCR on the biopsies of Crohn’s disease patients with active inflammation. Despite the inconsistencies among reports, alterations in the microbiome phylotype have been shown to accompany the changes in the mucosal layer of the distal gut during inflammation. Those bacteria that degrade mucus as a food source, such as Akkermansia sp. may have improved ability to survive over food-dependent luminal bacteria in stressful gastrointestinal conditions. Our results demonstrate that application of quantitative metaproteomics to microbiome studies will complement sequencing efforts and help shed light on the phylogenetic alterations associated with microbiome-associated intestinal diseases.

Proteins Increased in Control Mice Are Primarily from Microbes

The majority of significantly increased proteins in the RAG\(^{-/-}\) mice (or depleted in the IBD mice) are of microbial origin. We observed several enriched peptides that correspond to housekeeping proteins from Firmicutes, such as lipases, transferases, flagellar proteins, and acyl transferases (Supporting Information, Table S4). One genus with an over-representation of proteins in RAG\(^{-/-}\) includes the Arthrobacter sp. and likely represents an artifactual enrichment due to peptides from this source having a higher level of \(^{15}\)N incorporation than those from mouse and microbiome sources (Figure 2B, Supporting Information, Table S4). Our proteomic data also identified host proteins decreased in diseased mice, including murine pentraxin (Mptx1), intelectin (Ilt1n1), and alpha defensin (Defa7) (Figure 2A). Importantly, these proteins have been previously identified to be altered in gut inflammation, and further validates the utility of our MS-based microbiome interrogation. Murine pentraxin was previously found to have diminished expression during intestinal oxidative stress. Similarly, the glycolipid barrier protection protein Ilt1n1 and the antimicrobial peptide Defa7 are secreted by specialized intestinal goblet and Paneth
cells and have been identified to be downregulated in inflammatory states.48−51

**GO Term Analysis Shows Alterations in Microbial Protein Functionalities of IBD Mice**

Our LC−MS/MS data collection and ComPIL analysis of the SILAM samples identified a total of 2893 unique microbiome protein clusters among both groups of mice that were not produced by the host or spirulina diet. Out of a total of 3277 protein clusters in our SILAM data sets (contributed by host, diet, microbes), 201 were found to significantly change between healthy and disease samples (Figure 2A,B). From the identified clusters, 83.6% of the 3277 clusters have at least one GO term annotation (i.e., molecular function, biological process, cellular component) in ComPIL (Tables S5, S6, and S7). To take advantage of this high level of annotation coverage, we applied a Gene Set Enrichment Analysis (GSEA) algorithm to uncover any statistically significant functional alterations between the control and IBD mice. In terms of the annotated molecular function GO terms, the upregulated microbial proteins in IBD mice were dominated by oxidoreductase and lyase activities (Supporting Information, Table S5). Those GO term functions that describe biological processes revealed an upregulation in IBD mice of many metabolic processes, including carbohydrate catabolic processes (Supporting Information, Table S6). To take advantage of this high level of annotation coverage, we applied a Gene Set Enrichment Analysis (GSEA) algorithm to uncover any statistically significant functional alterations between the control and IBD mice. In terms of the annotated molecular function GO terms, the upregulated microbial proteins in IBD mice were dominated by oxidoreductase and lyase activities (Supporting Information, Table S5). Those GO term functions that describe biological processes revealed an upregulation in IBD mice of many metabolic processes, including carbohydrate catabolic processes (Supporting Information, Table S6). Conversely, our metaproteomic results suggested that RAG−/− controls have higher levels of housekeeping functionalities including RNA Pol activity and many biosynthetic processes (Tables S5 and S6). While the RAG−/−-associated functions are indicative of normal metabolic pathways, the depletion of these microbial proteins in IBD mice may suggest a limited abundance of resources for biosynthesis.

**Differences in Taxonomic Composition between Control and IBD Mice**

Aside from the identification and quantitation of microbial proteins, we wanted to evaluate the ability of our metaproteomics data to estimate the microbial composition of the samples analyzed. All peptides from the SILAM data sets were traced to the lowest common bacterial ancestor from which they were uniquely derived and used to generate insights into the phylogenetic composition of the control and IBD microbiome samples. The approximate bacterial composition was determined by weighting all peptides by normalized spectral counts. With respect to phylum-level composition, our metaproteomics data are in strong agreement with published 16S sequencing whereby Firmicutes and Bacteroidetes dominate the overall population of the microbial content and Bacteroidetes decrease significantly in an inflammatory state (Figure 3A,B).52−54 Importantly, our data demonstrated a statistically significant increase in Proteobacteria and Verrucomicrobia (e.g., Akkermansia sp.) in the IBD mice and these results are strongly correlated with the phylogenetic composition as assessed by metagenomics sequencing (Figure 3C,D). The overabundances of both phyla have been observed in humans with ulcerative colitis as well as Crohn’s disease. While Proteobacteria has more commonly been observed to bloom in IBD, several publications have demonstrated Verrucomicrobia also increase in number.51,60,61 The correlation of genomic sequencing and our proteomics data suggest that our microbiome sample preparation, LC−MS/MS data collection, and analytic methods are not biased toward any particular microbial components with respect to phylum level and can be used to corroborate sequencing results.
Labeling of Reactive Cysteines Allows for Interrogation of a New Subset of Proteins

Our results verified the overabundance/expression of host antiproteolytic proteins already observed in IBD mice. However, we did not detect any host and/or microbial proteases also previously established to accompany intestinal inflammation.62−64 The extreme complexity of the proteome derived from intestinal contents in combination with current limitations in tandem mass spectrometer sensitivity ensures that a significant number of host and microbial proteins will be missed with MudPIT shotgun metaproteomics. We therefore wanted to determine if an additional "enrichment" step that would target a subset of proteins, including proteases, from within our fecal samples would provide further insight into the presence of host and microbial proteases. To accomplish this, we applied an additional ABP enrichment step to the SILAM control and IBD fecal samples to isolate proteins with cysteine nucleophilic reactivity. Use of an activity-based probe enrichment step to label specific functionalities of interest may be able to reduce the complexity of the proteomic environment while simultaneously highlighting new differences between two biological groups. Here, our goal was to determine if the enrichment process resulted in the LC−MS/MS identification of a different set of host and microbial proteins in comparison to the unenriched data sets, and if any biologically relevant information on aberrant protease functionalities can be quantitated between the control and IBD mice.

On the basis of the successes of previous probe-based research in identifying proteases from the lysates of animal tissues, we synthesized a biotinylated glycine containing a C-terminal chloromethyl ketone warhead to generate a general cysteine-reactive molecule termed BioGlyCMK (Figure 4A, Supporting Information for synthesis). Attack by nucleophilic cysteine residues on BioGlyCMK results in irreversible biotinylation of proteins that then permits the probe-reactive subset of proteins to be enriched and isolated from the general proteome.

Figure 4. Comparison of activity-based probe enriched samples. (A) Schematic demonstrating the workflow for the ABP enriched samples. Bacterial cells were mixed in a 1:1 ratio before activity-based labeling and lysed after the process. A BioGlyCMK probe was used to target reactive cysteines in the proteome. (B) Venn diagram showing the differences in protein clusters found via the unenriched protocol as compared to those found via BioGlyCMK enrichment. (C) Log−log plot of peptide ratios determined via corrected spectral count ratios versus ratios determined from MS1 precursor ion intensities. Run-to-run variance in the ratios of low abundance peptides in the (D) unenriched or BioGlyCMK (E) enriched samples. Ratios were generated either via spectral counts or via MS1 precursor ion intensities.
proteome by avidin-coated beads. Here, our fecal sample preparation was modified to accommodate the additional probe incubation step (Figure 4A). Our proof-of-principle study with labeling introduced the probe after the 1:1 mix of $^{14}$N/$^{15}$N BioGlyCMK probe.

Enrichment with BioGlyCMK resulted in identification of an average of 1837 ± 293 peptides corresponding to 1088 ± 160 proteins, and 968 ± 128 peptides with 895 ± 114 protein matches for the $^{14}$N IBD and $^{15}$N control samples, respectively. Despite the overall number of measured peptides and proteins matching the unenriched LC–MS/MS data, only 51% of the identified protein loci were common to both the BioGlyCMK-enriched and unenriched SILAM data sets (Figure 4B).

Importantly, 49% of the identified protein loci were unique to the BioGlyCMK data, supporting our hypothesis that probe-based enrichment isolates a unique subset of proteins in comparison to the unenriched data sets (Figure 4B, Supporting Information, Table S8). Microbial proteins with reactive active-site cysteines specifically identified in the BioGlyCMK-enriched data include, but are not limited to, protein clusters with GO terms related to peptidases, alcohol dehydrogenases, and acetaldehyde dehydrogenases (Table 1). These findings suggest that use of our probe reduced the presence of housekeeping functionalities lacking a reactive cysteine (Table 1, Supporting Information, Table S8). Of note, a Fisher’s Exact test used to compare the overall representation of GO terms annotated in the unenriched versus BioGlyCMK-enriched samples demonstrated that the most strongly enriched molecular function GO term was cysteine-type peptidase activity (GO:0008234) with a ratio of 14.9 ($p$-value $= 3.0 \times 10^{-3}$, Table 1).

Use of an ABP likely reduces the complexity of the microbiota sample in comparison to unenriched sample preparation methods, as the probe will preferentially isolate only those proteins with selective reactivity. Here, the number of identified proteins is larger in the BioGlyCMK data set than for the collection of proteins found in the unenriched data set.

### Table 1. GO Terms Enriched in BioGlyCMK Labeled Samples over the Unenriched Samples

<table>
<thead>
<tr>
<th>GO term</th>
<th>description</th>
<th>odds ratio</th>
<th>$p$-value</th>
<th>no. annotated</th>
<th>no. in group</th>
<th>no. expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0008234</td>
<td>Cys-type peptidase</td>
<td>14.9</td>
<td>$3.9 \times 10^{-9}$</td>
<td>25</td>
<td>21</td>
<td>12</td>
</tr>
<tr>
<td>GO:0016153</td>
<td>urocarnate hydratase</td>
<td>14.8</td>
<td>$4.3 \times 10^{-9}$</td>
<td>13</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>GO:0004197</td>
<td>Cys-type endopeptidase</td>
<td>8.9</td>
<td>$4.4 \times 10^{-8}$</td>
<td>17</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>GO:0004022</td>
<td>alcohol dehydrogenase (NAD)</td>
<td>8.2</td>
<td>$1.3 \times 10^{-7}$</td>
<td>16</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>GO:0004825</td>
<td>methionine-tRNA ligase</td>
<td>7.9</td>
<td>$1.3 \times 10^{-7}$</td>
<td>12</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>GO:0008774</td>
<td>acetaldehyde dehydrogenase</td>
<td>7.4</td>
<td>$3.9 \times 10^{-6}$</td>
<td>15</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>GO:0016879</td>
<td>ligase, C–N bonds</td>
<td>3.8</td>
<td>$2.7 \times 10^{-5}$</td>
<td>33</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>GO:0016810</td>
<td>linear amide hydrolase, acting on C–N (but not peptide) bonds</td>
<td>3.4</td>
<td>$8.1 \times 10^{-5}$</td>
<td>46</td>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>GO:0004177</td>
<td>aminopeptidase</td>
<td>3.3</td>
<td>$9.3 \times 10^{-5}$</td>
<td>22</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>GO:0016620</td>
<td>oxidoreductase, acting on the aldehyde or oxo group of donors, NAD or NADP acceptor</td>
<td>3.2</td>
<td>$5.7 \times 10^{-11}$</td>
<td>140</td>
<td>72</td>
<td>49</td>
</tr>
<tr>
<td>GO:0016811</td>
<td>hydrolase, acting on C–N</td>
<td>3.2</td>
<td>$3.2 \times 10^{-5}$</td>
<td>28</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>GO:0016874</td>
<td>ligase</td>
<td>2.9</td>
<td>$1.3 \times 10^{-5}$</td>
<td>77</td>
<td>38</td>
<td>30</td>
</tr>
<tr>
<td>GO:0016638</td>
<td>oxidoreductase, acting on the CH-NH$_2$ group of donors</td>
<td>2.7</td>
<td>$3.0 \times 10^{-7}$</td>
<td>115</td>
<td>55</td>
<td>41</td>
</tr>
<tr>
<td>GO:0016639</td>
<td>oxidoreductase, acting on the CH-NH$_2$ group of donors, NAD or NADP acceptor</td>
<td>2.5</td>
<td>$7.4 \times 10^{-10}$</td>
<td>107</td>
<td>49</td>
<td>39</td>
</tr>
<tr>
<td>GO:0008233</td>
<td>peptidase</td>
<td>2.3</td>
<td>$9.1 \times 10^{-5}$</td>
<td>100</td>
<td>44</td>
<td>37</td>
</tr>
<tr>
<td>GO:0016746</td>
<td>acyl group transferase</td>
<td>2.1</td>
<td>$1.3 \times 10^{-5}$</td>
<td>91</td>
<td>38</td>
<td>34</td>
</tr>
<tr>
<td>GO:0016781</td>
<td>phosphotransferase, paired acceptors</td>
<td>2.0</td>
<td>$3.4 \times 10^{-5}$</td>
<td>173</td>
<td>69</td>
<td>59</td>
</tr>
<tr>
<td>GO:0016835</td>
<td>C–O lyase</td>
<td>1.9</td>
<td>$2.4 \times 10^{-5}$</td>
<td>110</td>
<td>43</td>
<td>40</td>
</tr>
<tr>
<td>GO:0016747</td>
<td>acyl group transferase activity, transferring acyl groups other than amino-acyl groups</td>
<td>1.9</td>
<td>$7.3 \times 10^{-5}$</td>
<td>84</td>
<td>33</td>
<td>32</td>
</tr>
<tr>
<td>GO:0016836</td>
<td>hydrolase</td>
<td>1.8</td>
<td>$6.6 \times 10^{-5}$</td>
<td>108</td>
<td>41</td>
<td>39</td>
</tr>
</tbody>
</table>

### Table 2. GSEA Showing Molecular Function GO Terms IBD versus RAG$^{-/-}$ Control Mice, Enriched via BioGlyCMK Probe

<table>
<thead>
<tr>
<th>term</th>
<th>description</th>
<th>enrichment score (ES)</th>
<th>normal ES</th>
<th>$p$-value</th>
<th>FDR</th>
<th>size</th>
<th>no. matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0008233</td>
<td>peptidase</td>
<td>0.76</td>
<td>2.28</td>
<td>$0.0^{*}$</td>
<td>6.2</td>
<td>$10^{-4}$</td>
<td>70</td>
</tr>
<tr>
<td>GO:0008234</td>
<td>Cys-type peptidase</td>
<td>0.93</td>
<td>2.27</td>
<td>0.0</td>
<td>3.1</td>
<td>$10^{-4}$</td>
<td>24</td>
</tr>
<tr>
<td>GO:0004197</td>
<td>Cys-type endopeptidase</td>
<td>0.93</td>
<td>2.21</td>
<td>0.0</td>
<td>2.1</td>
<td>$10^{-4}$</td>
<td>16</td>
</tr>
<tr>
<td>GO:0004175</td>
<td>endopeptidase</td>
<td>0.83</td>
<td>2.16</td>
<td>0.0</td>
<td>4.7</td>
<td>$10^{-4}$</td>
<td>29</td>
</tr>
<tr>
<td>GO:0016860</td>
<td>intramolecular oxidoreductase</td>
<td>0.80</td>
<td>2.00</td>
<td>$1.8 \times 10^{-5}$</td>
<td>4.4</td>
<td>$10^{-3}$</td>
<td>32</td>
</tr>
<tr>
<td>GO:0016797</td>
<td>hydrolase</td>
<td>0.42</td>
<td>1.91</td>
<td>0.0</td>
<td>1.0</td>
<td>$10^{-2}$</td>
<td>267</td>
</tr>
<tr>
<td>GO:0016810</td>
<td>hydrolase, acting on C–N (but not peptide) bonds</td>
<td>0.78</td>
<td>1.89</td>
<td>0.0</td>
<td>1.0</td>
<td>$10^{-2}$</td>
<td>35</td>
</tr>
<tr>
<td>GO:0016798</td>
<td>hydrolase, acting on glycosyl bonds</td>
<td>0.57</td>
<td>1.77</td>
<td>$8.6 \times 10^{-5}$</td>
<td>2.9</td>
<td>$10^{-2}$</td>
<td>33</td>
</tr>
<tr>
<td>GO:0004553</td>
<td>hydrolase, acting on O-glycosyl compounds</td>
<td>0.59</td>
<td>1.75</td>
<td>$1.8 \times 10^{-2}$</td>
<td>3.3</td>
<td>$10^{-2}$</td>
<td>31</td>
</tr>
<tr>
<td>GO:0003676</td>
<td>nucleic acid binding</td>
<td>0.56</td>
<td>1.54</td>
<td>$4.6 \times 10^{-2}$</td>
<td>1.3</td>
<td>$10^{-1}$</td>
<td>73</td>
</tr>
<tr>
<td>GO:0000287</td>
<td>magnesium ion binding</td>
<td>0.37</td>
<td>1.49</td>
<td>$4.4 \times 10^{-2}$</td>
<td>1.5</td>
<td>$10^{-1}$</td>
<td>72</td>
</tr>
</tbody>
</table>

*Nominal $p$-value of $< 1/1000$. DOI: 10.1021/acs.jproteome.6b00938 J. Proteome Res. 2017, 16, 1014–1026
These results suggest that the general promiscuity of our BioGlyCMK probe for proteins with nucleophile cysteine residues results in enriched samples that are still greater in diversity and complexity than the LC−MS/MS detection capabilities. Preparation of the samples in an aerobic environment may result in a fraction of the nucleophile cysteine-containing proteins to oxidize and limit the ability to bind BioGlyCMK. Use of ABPs with more specific spectra of target functionalities will likely generate enriched data sets that fall within the detection limit of tandem mass spectrometers and would further improve the reproducibility of both biological and technical proteomic data replicates.

Application of BioGlyCMK Probe Highlights Additional Differences in Control and IBD Mice

We focused our analysis to quantitate differences among the BioGlyCMK-enriched RAG−/− control and IBD proteomic data sets to those peptides and proteins uniquely identified in comparison to the unenriched data sets. Most significant is the overall reduction in ubiquitous and highly conserved proteins, including ribonuclease activity, RNA polymerase activity, and DNA binding proteins. GO term analysis with the GSEA algorithm shows a statistically significant increase in peptides from microbial peptidases and hydrolases from the IBD microbiome samples, and proteolysis ranked as the most highly enriched biological process (Table 2). While our probe only targeted the subset of cysteine-based proteases, it is clear from these results, as well as the unenriched data set where host-produced protease inhibitors are tremendously increased, that proteolytic activity is a critical component of the IBD mouse model. Use of probes that target serine, metallo-, and aspartate proteases will likely provide additional insights and help generate a compendium of potential microbially produced proteins that can be further assessed for their importance in propagating IBD.

Isotopic Labeling with 15N Increases Quantification Reliability in Enriched Samples

Although isotopic labeling is considered to produce more reliable relative quantitation in proteomic studies, many label-free methods of quantification, including those that rely heavily on spectral counts, have been developed and are currently in use.57 To attempt to address this question and the utility of isotopic labeling in samples as complex as the microbiome, we compared the isotopic ratios generated by precursor ions to the ratio of the normalized spectral counts between a 14N and 15N sample of a given mass spectrum. Examination of the relationship between the precursor ion intensity ratio with that of the spectral counts for each identified peptide yielded a moderately strong correlation of 0.67 (Figure 4C). While there are many peptides that are discordant between these two measures, the accuracy of the two cannot readily be determined. One area where 15N isotopic labeling may confer some advantage over spectral count quantification is in peptides of low abundance that have five or fewer counts among the 14N and 15N samples. In the unenriched data set, 578 peptides fall under this category, which accounted for 9% of all peptides identified. Examination of the MS data collection variance between replicates for these low abundance peptides by either SILAM ratios or relative spectral counts yielded almost identical distributions for the unenriched SILAM data (Figure 4D). Notwithstanding, performing this analysis on the BioGlyCMK-enriched samples shows much tighter variance when quantifying ratios via the 15N method (Figure 4E). Our analyses demonstrate that isotopic 15N labeling may confer a quantitative advantage over spectral count-based methods in metaproteomic studies when used in combination with an ABP enrichment step. We find that SILAM isotope quantitation on the highly complex unenriched microbial samples produces a similar level of variance to use of spectral count-based quantitation and use of SILAM should be considered as a potential approach for future metaproteomic studies.

CONCLUSIONS

Our study introduced SILAM and ABP enrichment to identify and quantitate differences in the highly complex protein mixture of host and microbial proteomes of RAG−/− control and IBD mice. MudPIT shotgun proteomics on SILAM isotopic labeled murine fecal samples in combination with our previously described ComPIL database permitted the identification of 4434 protein clusters in the microbial proteomic environment. Of these clusters, 276 were found to be in differential abundance between control and IBD mice, many of which are microbial proteins of unknown function. In addition, incorporation of an ABP enrichment step into the sample preparation process allowed for a unique subset of the microbial proteins to be identified by LC−MS/MS relative to the unenriched sample collection. Together, SILAM and our cysteine-reactive BioGlyCMK ABP identified several peptidases and hydrolases to be overly abundant in IBD. Addition of these methodologies to metaproteomics of human and disease model systems will help to identify and measure microbial proteins dysregulated in disease and begin to create a novel list of drug discovery targets to combat microbiome-related intestinal diseases.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00938.

BioGlyCMK synthesis, mass spectrometry, NMR; graph of weight loss; IBD mice histology; species sources of proteins in 15N samples; graph/plot of experimental replicate reproducibility and IBD versus control; figure legends and descriptions of Excel tables; collected and pooled SILAM samples (PDF)

All proteins identified in all LC−MS/MS data sets (XLSX)

IBD cohort microbial proteins overexpressed (XLSX)

Control cohort overexpressed proteins (XLSX)

Molecular GO terms (XLSX)

Biological processes GO terms (XLSX)

Cellular component GO terms (XLSX)

BioGLYCMK-enriched proteins uniquely identified with respect to unenriched data set (XLSX)

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD005667.

AUTHOR INFORMATION

Corresponding Authors

E-mail: wolan@scripps.edu.
Journal of Proteome Research

*E-mail: asu@scripps.edu.

ORCID

Dennis W. Wolan: 0000-0001-9879-8353

Author Contributions

D.W.W. and A.I.S. conceived of the project. M.D.M. performed wet lab experiments and collected tandem MS data. C.M. helped generate the IBD mice and advised on sample and data collection. G.S.S. designed and maintained ComPIL. M.D.M. and G.S.S. analyzed tandem LC–MS/MS data. A.I.S., M.D.M., and G.S.S. contributed to peptide mapping and functional analysis. All authors contributed to the preparation and editing of the manuscript.

Notes

The authors declare no competing financial interest.

Acknowledgments

We thank D. McClatchy for consultation on SILAM techniques. A. Wang for assistance with mass spectrometry and probe labeling protocols. S. Chatterjee for technical assistance with sample preparation and mass spectrometry instrumentation. This work was supported by The Scripps Research Institute and the U.S. National Institutes of Health 1R21CA181027 (to D.W.W.) and NIH Training Grant No. T32AI007244 (to C.M.).

References


