MCP RESEARCH

Quantitative Metaproteomics and Activity-based Protein Profiling of Patient Fecal Microbiome Identifies Host and Microbial Serine-type Endopeptidase Activity Associated With Ulcerative Colitis

Authors

Peter S. Thuy-Boun, Ana Y. Wang, Ana Crissien-Martinez, Janice H. Xu, Sandip Chatterjee, Gregory S. Stupp, Andrew I. Su, Walter J. Coyle, and Dennis W. Wolan

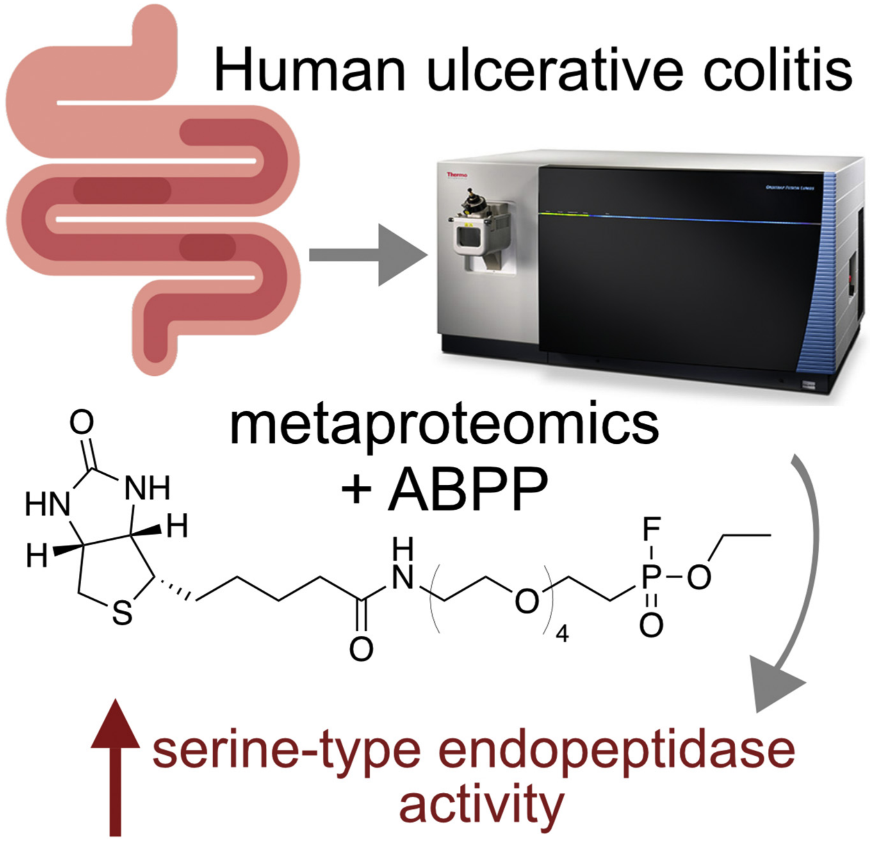
Correspondence

wolan@scripps.edu

In Brief

Thuy-Boun *et al.* quantitatively compare the stool microbiomes of healthy and ulcerative colitis patients with label-free data-dependent LC-MS/MS proteomics. Their analyses identified 176 significantly enriched protein groups between the two cohorts, and serine-type endopeptidase activity was one such functionality overrepresented in UC patients. Pre-enrichment of the clinical samples with a biotinylated fluorophosphonate probe further demonstrated that serine endopeptidases are active within the patient fecal samples and that additional putative serine hydrolases were identified by this approach compared with unenriched profiling.

Graphical Abstract



Highlights

- Identified 176 significantly altered protein groups between healthy and UC patients.
- Serine-type endopeptidase activity is overrepresented in UC patients.
- Fluorophosphonate ABPP shows that endopeptidases are active in fecal samples.
- ABPP enrichment helps identify additional putative serine hydrolases in samples.
- *De novo* sequencing used to estimate number of MS2 spectra unidentified by ComPIL.

2022, Mol Cell Proteomics *21*(*3*), 100197

https://doi.org/10.1016/j.mcpro.2022.100197

Quantitative Metaproteomics and Activity-based Protein Profiling of Patient Fecal Microbiome Identifies Host and Microbial Serine-type Endopeptidase Activity Associated With Ulcerative Colitis

Peter S. Thuy-Boun[1], Ana Y. Wang[1], Ana Crissien-Martinez[2], Janice H. Xu[1], Sandip Chatterjee[1], Gregory S. Stupp[3], Andrew I. Su[3], Walter J. Coyle[2], and Dennis W. Wolan[1,3,*]

The gut microbiota plays an important yet incompletely understood role in the induction and propagation of ulcerative colitis (UC). Organism-level efforts to identify UC-associated microbes have revealed the importance of community structure, but less is known about the molecular effectors of disease. We performed 16S rRNA gene sequencing in parallel with label-free data-dependent LC-MS/MS proteomics to characterize the stool microbiomes of healthy (n = 8) and UC (n = 10) patients. Comparisons of taxonomic composition between techniques revealed major differences in community structure partially attributable to the additional detection of host, fungal, viral, and food peptides by metaproteomics. Differential expression analysis of metaproteomic data identified 176 significantly enriched protein groups between healthy and UC patients. Gene ontology analysis revealed several enriched functions with serine-type endopeptidase activity overrepresented in UC patients. Using a biotinylated fluorophosphonate probe and streptavidin-based enrichment, we show that serine endopeptidases are active in patient fecal samples and that additional putative serine hydrolases are detectable by this approach compared with unenriched profiling. Finally, as metaproteomic databases expand, they are expected to asymptotically approach completeness. Using ComPIL and *de novo* peptide sequencing, we estimate the size of the probable peptide space unidentified ("dark peptidome") by our large database approach to establish a rough benchmark for database sufficiency. Despite high variability inherent in patient samples, our analysis yielded a catalog of differentially enriched proteins between healthy and UC fecal proteomes. This catalog provides a clinically relevant jumping-off point for further molecular-level studies aimed at identifying the microbial underpinnings of UC.

Inflammatory bowel disease (IBD) is a chronic medical condition characterized by relapsing inflammation of the gastrointestinal (GI) tract. This disease is broadly divisible into two categories based on where inflammation occurs. In ulcerative colitis (UC), inflammation is restricted to the large intestine, while in Crohn's disease (CD), inflammation can occur anywhere along the GI tract (1, 2). In addition to reduced life expectancy, IBD patients can suffer dramatic quality-of-life reductions and are at increased risk to develop gastrointestinal tract malignancy (3). The incidence and prevalence of IBD in developed countries have steadily increased in the last few decades, making this disease a public health concern with a potentially heavy cost burden due to a requirement for long-term management (4). Targeted cures for UC and CD are highly desirable, but the search for such treatments is hampered by our incomplete understanding of disease development. Genome-wide association studies (GWAS) have identified over 200 genetic loci associated with UC and CD, but the polygenic nature of these conditions explains only a minor portion of disease incidence (5–7). Concordance rates of about 30% for CD and 15% for UC among monozygotic twins suggest a significant nongenetic contribution to disease development (8). Because our gut microbes are in perpetual contact with our GI tracts, they comprise important but ill-defined environmental variables that many studies have implicated in IBD development. IBD triggers are unknown, but its progression is hypothesized to be amplified by

From the [1]Department of Molecular Medicine, The Scripps Research Institute, La Jolla, California, USA; [2]Scripps Green Hospital Division of Gastroenterology, La Jolla, California, USA; [3]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California, USA

[*]For correspondence: Dennis W. Wolan, wolan@scripps.edu.

https://doi.org/10.1016/j.mcpro.2022.100197

inappropriate host–microbe interactions that lead to dysbiosis and, eventually, observable gross pathology (9–13).

Efforts to identify potential microbial drivers of IBD are ongoing but stymied by the immense taxonomic complexity of the gut microbiota. The gut harbors hundreds of distinct species per individual, which can change over time and with perturbations to host lifestyle or xenobiotic exposure (14–16). Campaigns to characterize and monitor the gut microbiota frequently utilize amplicon and metagenomic sequencing technologies, which can provide information about microbial community structure, genetic potential, and transcriptional activities (17, 18). As the collective size of shotgun metagenomic sequence space has expanded from these efforts, so too have the opportunities for liquid chromatography tandem mass spectrometry (LC-MS/MS)-based metaproteomics (19), which rely on protein reference databases constructed from translated genome sequences (20). Host protein profiling is a straightforward process given the relative completeness of host (*e.g.*, human, mouse, *etc.*) genome assemblies, but microbiome protein profiling is more difficult due to the high strain diversity and presence of unculturable microbes in the gut (21). Sample-matched MAGs (metagenome assembled genomes) delivering good spectrum match rates have become an effective solution but can be cost-limiting and require expertise (15, 22–25). In addition, manual reference database curation has proven to be an important consideration in metaproteomics but becomes computationally burdensome as community diversity expands (26). To address this problem, we developed the Comprehensive Protein Identification Library (ComPIL), a large and scalable proteomics database generally intended for metaproteomics studies (27). In its current iteration (ComPIL 2.0), it houses >4.8 billion unique, tryptic peptides derived from >113 million bacterial, archaeal, viral, and eukaryotic parent protein sequences assembled from public sequencing repositories (28). With periodic incorporation of new sequences from shotgun metagenomics repositories, we envision that ComPIL will help enable interlaboratory consonance in the global interpretation and communication of bottom-up metaproteomics results. In addition to enabling the direct, large-scale observation of proteins in a complex mixture, LC-MS/MS-based metaproteomics techniques obviate a requirement for intact cells, facilitate the observation of posttranslational protein modifications, and enable functional interrogation of new or incompletely annotated proteins through such cognate techniques as activity- and affinity-based protein profiling (ABPP) (29, 30).

Relative to metagenomics, LC-MS/MS-based metaproteomics are less commonly applied and more rarely employed in IBD studies. In fact, the first large-scale endeavor to identify proteins from a microbial biofilm community was only disclosed by Banfield, *et al.* in 2005 (31–33). In 2009, Jansson, *et al.* leveraged high-resolution LC-MS/MS to demonstrate the viability of large-scale metaproteomics in fecal samples collected from a twin pair (34). The aforementioned study demonstrated for the first time that bottom-up LC-MS/MS-based proteomics technology is suitable for such a complex environment and that it could generate a model of the gut microbiome that is orthogonal to that produced by metagenomics. Since then, several groups have deployed bottom-up metaproteomics to investigate the etiology of IBD in humans by examining patient fecal extracts, intestinal biopsy tissue, and/or blood samples (35–44). Additionally, several groups including our own have paired traditional proteomic profiling with ABPP techniques in microbiome samples to detect and annotate key protein functionalities often undetectable without preenrichment (45–48).

The insights garnered from previous metaproteomics studies are valuable for forming a consensus about the constellation of IBD-related environmental factors. We aim to contribute to this nascent pool by presenting a combined 16S rRNA gene amplicon sequencing and metaproteomics analysis of fecal samples from healthy volunteers and ulcerative colitis patients to identify novel proteins associated with health or disease. Using a pipeline that incorporates a novel, strong-acid sample preparation procedure (49), label-free high-resolution LC-MS/MS, and the ComPIL database coupled to the ProLuCID/SEQUEST search engine (20, 50, 51), we identify 176 protein groups and several gene ontology (GO) terms enriched in either cohort (52). We show that proteomics can provide a more complete picture of gut microbiome taxonomy that includes host, microbial, and even dietary proteins. Using ABPP, we demonstrate that not only are microbiome proteins enzymatically active after collection, but additional proteomic depth can be achieved using ABPP enrichment strategies. Finally, using *de novo* peptide sequencing tools, we provide a means for estimating the size of database-elusive peptide space in our LC-MS/MS data, enabling a rough estimation of metaproteome completeness. This measure can help shape future decision-making processes regarding the need for additional shotgun metagenomic work to support a given metaproteomics study.

EXPERIMENTAL PROCEDURES

Patient Sample Collection

Collection and use of all patient samples were approved by the Office for the Protection of Research Subjects at Scripps Research and Scripps Green Hospital (IRB protocol IRB-14-6352). The written informed consent was obtained from all subjects in accordance with the Declaration of Helsinki.

Volunteers self-collected their own stool samples using administered standardized in-home sample collection kits and were instructed to immediately freeze specimens at −20 °C. Samples were stored in provided consumer-grade −20 °C minifreezers immediately after collection, transported by courier services on dry ice, and stored in laboratory-grade freezers at −20 °C until microbial extraction.

Collected stool samples were highly heterogeneous in color, texture, and viscosity both before and after microbial extraction.

Microbe Extraction

Stool samples were thawed to room temperature, diluted in PBS (pH 7.4), vortexed thoroughly to yield slurries, and then centrifuged at 100*g* for 1 min. The flocculent upper layer was extracted, filtered over 70 μm nylon mesh cell strainers to remove large, recalcitrant masses, and then centrifuged at 8000*g* for 5 min to pellet. Pellets were rinsed twice with PBS, then resuspended in PBS to a density of 100 mg wet microbial pellet per 500 μl of suspension.

DNA Extraction, 16S rRNA Gene Sequencing, and Data Processing

Microbial DNA was extracted from thawed fecal microbe aliquots using a fecal/soil extraction kit (Zymo Research, Irvine, CA, USA). In total, 50 to 100 ng of DNA per patient sample was submitted to the Scripps Research genomics core for next-generation sequencing, which was performed using the MiSeq platform (Illumina Inc). For taxonomy based on 16S rRNA gene amplicon sequencing, we targeted the bacterial 16S V4 region using a 300 bp paired-end approach aiming for 100,000 reads per sample. Reads were taxonomically mapped using QIIME2 (and associated plug-ins) and classifiers created from the SILVA 132 database (53, 54). For access to raw data, see Zenodo repository (55). For detailed methods, see Additional File 1.

Protein Extraction, Protein Digestion, Proteomics Data Collection

Protein was extracted according to a previously described protocol (49). Extracted protein was resuspended in H_2O, and concentration was measured by BCA assay (ThermoFisher). Extracted microbiome protein (100 μg) was reduced, alkylated, trypsinized, and desalted (ZipTip C18, MilliporeSigma) prior to LC-MS/MS analysis. Desalted peptides (1 μg) were separated using a 4-h C18 gradient by nano-flow liquid chromatography coupled to an Orbitrap Fusion Tribrid (Thermo Fisher) operating in data dependent mode. Both MS1 and MS2 spectra were recorded in the Orbitrap at 120K and 30K resolution, respectively. For detailed methods, see Additional File 1.

Proteomics Data Analysis

We collected a total of 2,829,920 MS2 spectra between all 18 patient samples. These spectra were searched against the ComPIL 2.0 database (contains 4.8 billion unique tryptic peptides from >225 million forward and reverse protein sequences) (27, 28) using the ProLuCID/SEQUEST search engine (20, 50, 51). In total, 523,155 (18.5%) MS2 spectra were mapped to 54,378 distinct peptides at a 1% peptide false discovery rate (two peptide per protein minimum) using a target-decoy strategy (56). In total, 576,625 protein sequences were identified and clustered into 95,000 protein groups at a 95% sequence similarity cutoff using CD-HIT (57, 58). Quantification at the MS1 level was performed using FlashLFQ with a match-between-runs strategy enabled (10 ppm precursor tolerance, 15 min window) (59). Peptide MS1 area-under-the-curve intensities were mapped to protein groups. Intensity belonging to peptides that mapped to >1 protein group were excluded. After removing protein groups with too many missing values (protein groups were removed if: (1) both conditions contained only null values, (2) one condition contained null values and the other contained <4 non-null values, or (3) both conditions contained <4 nonnull values each), 4622 protein groups remained for differential expression analysis, which was performed using Limma as part of the DEP package in the R statistical environment (60, 61). Protein groups were annotated with GO terms using InterProScan; these annotations were used for GO enrichment analysis in the GOstats package. GO relative abundance analysis was performed before removal of protein groups with too many missing values (62–64). Peptides were mapped to their respective taxa of origin using Unipept (65–67). Note that ideally, metaproteomics should be performed in conjunction with metagenomic sequencing to generate matched customized proteome databases. Protein provenance could more confidently be traced in this scenario enabling greater precision during taxonomy analysis. In the absence of matched metagenomic data, Unipept can provide peptide-based taxonomic mapping information but may do so with less accuracy. Peptides were then mapped to taxa to construct relative abundance tables and plots. Finally, *de novo* peptide sequencing was performed in Novor (68) and database-*de novo* peptide comparisons were performed using the Scikit-bio Python library. For more detailed methods, see Additional File 1. For access to LC-MS/MS data, see PRIDE repository PXD022433 (69). For *de novo* datasets, protein fasta files, and protein group files (CD-HIT clusters), see Zenodo repository (55) 10.5281/zenodo.5717460.

Experimental Design and Statistical Rationale

Single time-point stool samples from 18 patient volunteers were collected and analyzed. Healthy (n = 8, M/F ratio = 5:3, mean age = 44 years) and UC (n = 10, M/F ratio = 8:2, mean age = 46 years) volunteers consisted of a mixture of males and females between the ages of 21 to 76 years (global mean age = 45 years, global median age = 39 years) at the time of enrollment. UC patients presented with mild to severe symptoms and a range of Mayo scores (0–9) during enrollment. Individuals with BMI values >60, as well as any recent antibiotics usage (<3 months prior to sample collection), severe diarrheal illnesses, or *Clostridium difficile* infections were excluded.

For unenriched proteomics experiments, eight healthy biological replicates were compared against ten UC biological replicates (one technical replicate per biological replicate). Analyses did not rely on isotopic labels or internal standards. Instead, total protein and peptide concentrations were measured and normalized by BCA assay prior to LC-MS/MS. We chose to normalize samples by protein concentration to simplify comparisons, as collected patient stool samples were highly variable in volume, hydration, and consistency. This implies that in this study, proteins/protein group compositional fractions rather than absolute protein concentrations were compared between patients. ProLuCID/SEQUEST and DTASelect were used for peptide identification at a peptide-level FDR setting of 1% using a target-decoy strategy (20, 50, 51, 56, 70). Quantification was performed by MS1 peak intensity using a match-between-runs strategy. All protein sequences were grouped/clustered at a 95% sequence similarity cutoff using CD-HIT, then peptide intensities were mapped to protein groups/clusters (58). Peptides mapping to >1 group/cluster were excluded. Peptide intensities were summed within protein groups/clusters unless otherwise noted. Protein group/cluster intensities were normalized using the Limma/DEP package function “normalize_vsn” (variance stabilizing normalization) within the R statistical computing environment (60, 61). Protein group differential enrichment testing was performed using the “test_diff” function within the Limma/DEP package within the R statistical computing environment (60, 61). Differential testing *q*-values were calculated using the “qvalue” package in the R statistical environment, and a threshold of $q < 0.1$ was chosen as a relevant cutoff value for further analysis and discussion (71). Note that at more strict *q*-value thresholds, several protein groups were found to be significant (68 protein groups significant at $q < 0.01$, 55 protein groups at $q < 0.001$), but the modest $q < 0.1$ threshold was chosen to enable a more broad view of potential disease-associated protein functions with the acknowledged caveat that approximately 18 of 176 significantly enriched protein groups are false positives. Please see Additional File 1 for more details.

For FP-probe-enriched proteomics experiments, one healthy biological replicate and two UC biological replicates were analyzed by LC-MS/MS (one technical replicate for each sample). The stool samples used for these FP-enriched experiments were identical to patient stool used earlier for unenriched experiments. These FP-enriched experiments were qualitative and not statistically powered.

RESULTS

Metaproteomics Yields a More Comprehensive Taxonomic Diversity Than 16S rRNA Gene Analysis

At the kingdom level, LC-MS/MS-based proteomics identified peptides mapping to bacteria (38.3%), eukaryotes (43.5%) (including host), archaea (<0.1%), and viruses as well as a significant proportion of unassigned peptide (18.2%) [Additional File 1, supplemental Fig. S1]. In contrast, a majority of 16S reads (>99%) were attributable to bacteria with a much smaller proportion attributable to archaea (<0.1%) and <0.1% remaining unassigned [Additional File 1, supplemental Fig. S1].

Large discrepancies in taxonomic resolution begin to emerge at the phylum level. By LC-MS/MS proteomics, we identified 46 phyla, including Ascomycota, Basidiomycota, Spirochaetes, Chordata, and Streptophyta in addition to the eight identified by 16S rRNA gene sequencing alone (Euryarchaeota, Actinobacteria, Bacteroidetes, Cyanobacteria, Firmicutes, Fusobacteria, Proteobacteria, and Verrucomicrobia) (Fig. 1). Firmicutes account for only 22.7% of the sample composition according to the LC-MS/MS; however, this phylum dominates the composition of patient microbiota according to 16S amplicon sequencing and accounts for 86.2% of all reads. Interestingly, the abundance of Bacteroidetes was relatively minimal by 16S rRNA gene sequencing (except for samples H5, UC9, UC15, UC23) yielding a Bacteroidetes:Firmicutes ratio of approximately 1:100 (Fig. 1). In contrast, Bacteroidetes account for an average of 4.2% of the microbiota peptide content by LC-MS/MS-based proteomics yielding a Bacteroidetes:Firmicutes ratio of approximately 1:5. The majority of identified peptides identified *via* metaproteomics predominantly originate from the Chordata phylum and are presumably host-derived. Additionally, a significant number of peptides are derived from the Streptophyta and are attributable to a variety of dietary plants, including *Solanum tubersum* (potato), *Seasum indicum* (sesame), *Theobroma cacao* (chocolate), *Zea mays* (corn), and *Oryza sativa* (rice) among many others [see Additional File 2]. Where comparable, we posit that differences in DNA extraction efficiencies (*e.g.*, Gram$^+$ *versus*. Gram$^-$), differences in metabolic/secretory activities, and shared tryptic peptides between microbes likely contribute to the discrepancy in taxonomic compositions between 16S gene sequencing and proteomics methods.

The trend of increased taxonomic diversity across microbiota samples observed by LC-MS/MS proteomics over 16S amplicon sequencing is conserved at each classification tier [Additional File 1, supplemental Figs. S1–S7]. At the species

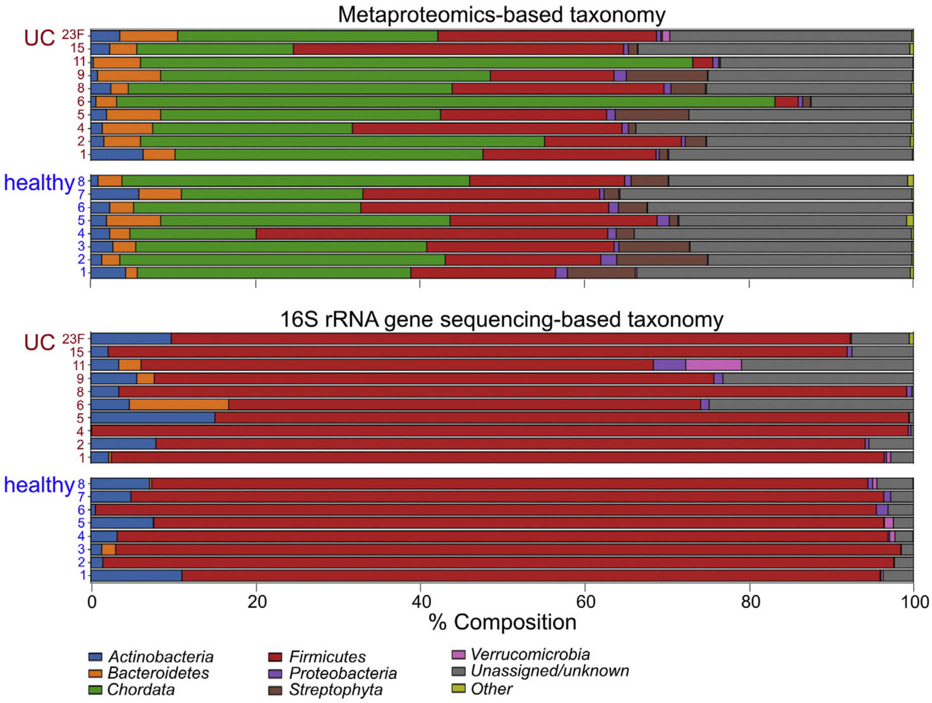


Fig. 1. **Relative abundance plots for all 18 patient samples.** Comparison of microbiome phylum-level taxonomy by 16S rRNA gene amplicon sequencing (*lower* 18 bars; based on sequence counts) against LC-MS/MS bottom-up proteomics (*upper* 18 bars; based on peptide intensities, note that shared peptides were grouped in the "Unassigned/unknown" category). LC-MS/MS, liquid chromatography tandem mass spectrometry.

level, we detected a total of 848 species/strains by LC-MS/MS and only 38 by amplicon sequencing. Despite the predicted increase in diversity by LC-MS/MS, an average of 85.9% of the identifiable peptides are not mappable to a particular species. This observation is due to the redundancy and conservation of microbial proteins across distinct and divergent species. Thus, taxonomic predictions based on the peptide composition in samples become increasingly difficult with more granular classification levels.

176 Protein Groups Are Significantly Altered Between the Healthy and UC Cohorts

Differential expression analysis yielded 176 protein groups (from host, microbes, and diet) significantly altered between healthy and UC patients ($p \leq 0.005$ and $q < 0.1$), with 65 groups enriched in healthy volunteers and 111 protein groups enriched in the UC volunteers (Fig. 2*A* and Additional File 3). Principal components analysis (PCA) of the dataset revealed a modest but distinguishable separation between the proteomic composition of healthy and UC fecal samples (Fig. 2*B*), which is in agreement with the Euclidean distance matrix generated for the same dataset (Fig. 2*C*).

Twenty nine of the 111 protein groups significantly enriched in UC fecal samples were host-derived; however, no host protein groups were enriched (q-value < 0.1) in the healthy individuals. STRING analysis of significantly enriched UC host proteins yielded 26 edges among 25 nodes and a highly significant protein–protein interaction (PPI) enrichment p-value of 1.84×10^{-12} at medium confidence (0.400) suggesting a strong association between tested proteins (Fig. 2*D* and Additional File 4) (72). At highest confidence (0.900), 13 edges between 25 nodes were found reinforcing a significant PPI enrichment p value of 2.0×10^{-11}. At medium confidence, the most significant reactome pathways (fdr <0.001) were neutrophil degranulation, innate immune system, antimicrobial peptides, immune system, and metal sequestration by antimicrobial proteins. GO biological process terms (fdr $<3.3 \times 10–^{8}$; regulated exocytosis, neutrophil degranulation, secretion by cell, transport, leukocyte mediated immunity, and antimicrobial humoral response) and GO cellular component

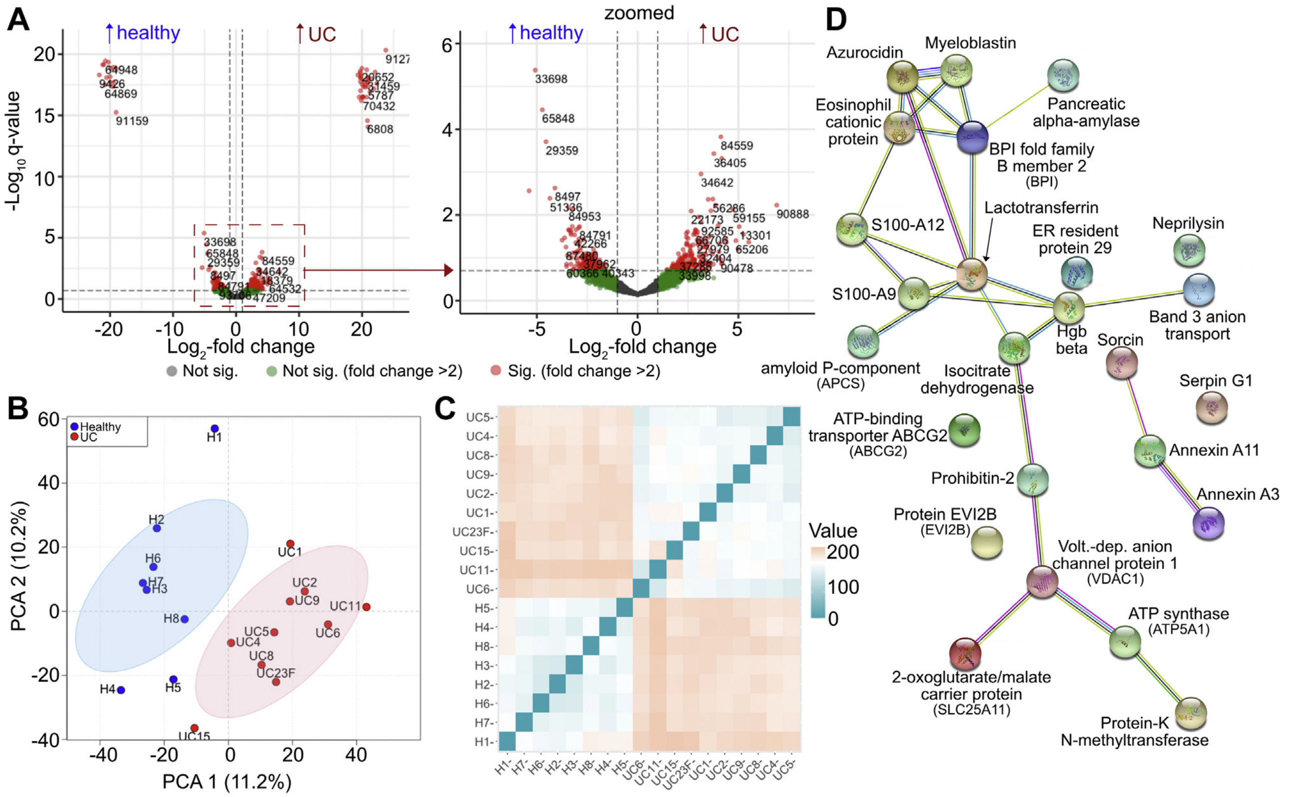


FIG. 2. **Differential expression and protein network analysis of host and microbial proteins.** *A*, volcano plot depicting differentially detected protein groups from patient fecal samples; whole plot (*left*), zoomed-in plot (*right*), *red* = significant ($q < 0.1$, foldchange > 2), *green* = not significant ($q > 0.1$, foldchange > 2), *gray* = N.S. ($q > 0.1$, foldchange ≤ 2), number adjacent red points represent unique "ClusterID" values corresponding to Additional File 4 labels. *B*, principal component analysis of all 18 patient samples across all differentially tested contrasts; *red* = healthy individuals, *blue* = UC patients. *C*, euclidean distance plot of all 18 patient samples across all differentially tested contrasts, *blue* = more similarity, *orange/red* = less similarity. *D*, STRING protein network analysis of host protein groups differentially enriched in UC patients' fecal samples at medium confidence (0.400); edges: known interactions (*aqua*: from curated database, *magenta*: experimentally determined), predicted interactions (*green*: gene neighborhood, *red*: gene fusions, *blue*: gene co-occurrence), others (*chartreuse*: text mining, *black*: coexpression, *lavender*: protein homology). UC, ulcerative colitis

terms (fdr $< 2 \times 10^{-7}$; cytoplasmic vesicle lumen, secretory granule, secretory granule lumen, vesicle, cytoplasmic vesicle part, and cytoplasmic vesicle) associated with host proteins all support the assertion that host immune-related secretory events are prevalent in the gastrointestinal tracts of UC patients. Interestingly, because no host proteins were significantly enriched across healthy fecal samples, we posit that host-centric biological pathways associated with colitis occur in addition to rather than in lieu of processes associated with homeostasis.

More than half of the nonhost protein groups have limited to no annotations despite being significantly altered. For example, 37 of 65 and 48 of 82 nonhost protein groups enriched in the healthy and UC cohorts, respectively ($p \leq 0.005$ and $q < 0.1$), were poorly annotated in the ComPIL database (*i.e.*, no annotation, annotated as hypothetical proteins or as domain of unknown function-containing proteins). While additional BLAST homology searches and InterProScan analyses were performed on each poorly annotated protein group, we were unable to make significant additional annotations for 11 of 37 protein groups enriched in the healthy cohort and 15 of 48 protein groups enriched in the ulcerative colitis cohort [see Additional File 4] (62, 63, 73). Such protein groups represent interesting targets for structural and biochemical validation as further inquiry could elucidate their possible roles, in the propagation of inflammatory or anti-inflammatory processes.

Significantly enriched and annotated nonhost protein groups among both healthy and UC cohorts had predicted and/or biochemically verified functions ranging from metabolism (glyceraldehyde-3-phosphate dehydrogenase and translation elongation factor Tu) to defense (type II secretion system protein). Notable nonhost entries that were increased in healthy volunteers include an acid-soluble spore protein (WP_071120403.1), methylene tetrahydrofolate reductase (SRS064276.159392-T1-C), and fruit bromelain (BROM1_ANACO). The enriched small spore protein is the only entry in its protein group and originates from the recently described bacterium *Romboutsia timonensis*, whose depletion has been associated with colorectal cancer incidence (74, 75). The methylene tetrahydrofolate reductase protein group enriched in the healthy cohort contains 29 members possessing similarity scores in the range of 98.6 to 100%. By BLAST analysis, these reductases likely originate from the Lachnospiraceae family of bacteria, and their examination could provide a glimpse into the microbial B-vitamin economy that importantly underpins host homeostasis, as humans are unable to *de novo* synthesize many essential B vitamins (76–78). Interestingly, fruit bromelain detected in four of eight healthy patient fecal extracts is a pineapple-derived cysteine protease we did not expect to encounter (79, 80). This protease is commonly sold as an over-the-counter supplement or as a component of meat tenderizers, and its detection may be an artifact introduced through patients' diets.

With respect to protein groups enriched in UC patients, notable annotated nonhost entries include hyaluronan glucosaminidase (CLONEX_02,131), a transglycosylase SLT domain-containing protein (HMPREF0462_0704), and a metallohydrolase (WP_081140786.1). The enriched hyaluronan glucosaminidase group contains four members with almost identical sequences (99.95–100% identity). These proteins likely originate from the Lachnospiraceae family members *Tyzzerlla* or *Coprococcus*. Hyaluronan is a high-molecular-weight carbohydrate component of the human extracellular matrix that can serve as an inflammatory/injury signal for host immune receptors when degraded by host hyaluronidases (81). Hyaluronan glucosaminidase activity may exacerbate host inflammatory processes and contribute to UC-related inflammation, as well as afford microbes the ability to infiltrate host barriers. The transglycosylase SLT (lytic) domain-containing protein group includes 70 members with sequence identities of 95.71 to 100% compared with the *Helicobacter pylori* (*H. pylori*) enzyme. These enzymes catalyze the nonhydrolytic intramolecular cyclization of *N*-acetylmuramyl residues on bacterial cell walls propagating the cell wall remodeling process (82, 83). For *H. pylori*, previous studies demonstrate the importance of transglycosylase-generated cell wall muropeptide fragments to inducing host inflammation, which can in turn promote gut colonization (84, 85). Finally, the enriched metallohydrolase originates from *Pantoea latae* and is annotated as a nonpeptide amide C–N bond hydrolase that may inactivate amide-containing molecules such as lactams, which are contained in an important class of antibiotics (86, 87). A complete list of protein groups found differentially expressed in this patient cohort can be found in the supplement [see Additional File 4].

Serine-type Endopeptidase Activity is Significantly Enriched in UC Samples

For GO term relative abundance analysis, we used mean peptide intensities for peptides within the same protein group to account for comparisons between proteins of different lengths. Of the 8538 quantifiable protein groups selected for relative abundance plotting, we identified 575, 394, and 85 terms for the molecular function, biological process, and cellular component GO namespaces, respectively. In general, GO relative abundance breakdowns between all samples for each namespace appear similar by unweighted (count-based) assembly (Fig. 3*B* and Additional File 1, supplemental Figs. S8–S10). However, when weighted by corresponding ion intensities, GO term relative abundances between samples differ dramatically (Fig. 3*A* and Additional File 1, supplemental Figs. S8–S10). The "None" and "Other" categories occupy the largest areas both by unweighted and weighted assembly for all three GO namespaces. With respect to molecular function, global relative abundances (when averaged over all samples) for the terms glutamate-cysteine ligase activity (GO: 0004357), aminopeptidase activity (GO: 0004177), and serine-type

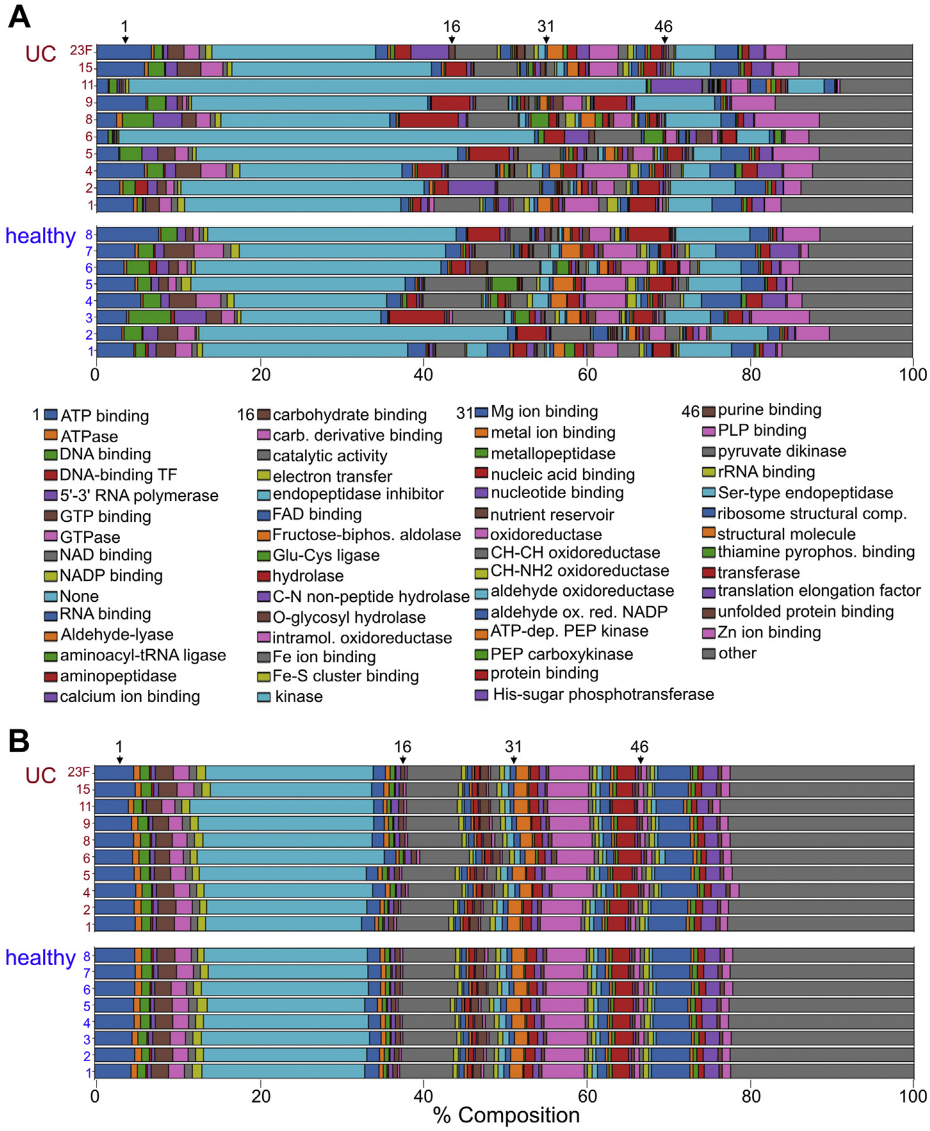


Fig. 3. **GO molecular function relative abundance plots for all 18 patient samples**. Comparison of microbiota GO molecular function breakdown by LC-MS/MS using either (*A*) weighted measures (*upper* 18 bars; peptide intensity-based; to control for protein length, each GO term's constituent protein group/cluster contributes the mean intensity of its constituent peptides, see Additional File 1 for more detail) or (*B*) unweighted measures (*lower* 18 bars; count-based; each GO term's constituent protein group/cluster contributes one count). Bar segments represent the proportion of each GO term's intensity or count relative to the total intensity or count (respectively) for each patient sample. Loosely, (*A*) represents GO terms as a function of protein copy number and (*B*) represents GO terms as a function of protein sequence diversity. GO, gene ontology; LC-MS/MS, liquid chromatography tandem mass spectrometry.

endopeptidase activity (GO: 0004252) expand 31-, 18-, and 14-fold respectively going from an unweighted to weighted assembly. Conversely, global relative abundance for the terms enoyl-[acyl-carrier-protein] reductase (NADH) activity (GO: 0004318) and mismatched DNA binding (GO: 0030983) contract >50-fold going from unweighted to weighted assembly. Similar relative abundance expansions and contractions going from unweighted to weighted assemblies were observed for the biological process and cellular component namespaces [see Additional File 1, supplemental Figs. S9 and S10].

The large 31-fold unweighted-to-weighted relative abundance expansion of glutamate-cysteine ligase activity could be attributed to one protein group with one member (WP_027345637.1). This enzyme originates from *Hamadaea tsunoensis* and catalyzes a key step in the synthesis of glutathione, a key antioxidant for the microbiota (88). The 18-fold unweighted-to-weighted relative abundance expansion observed for the aminopeptidase activity term originated from 19 protein groups with one very dominant protein group (WP_027209280.1) representing 96.2% of the GO term namespace's relative abundance. This predicted M18 family protease originates from *Butyrivibrio hungatei* but currently has no structural or biochemical annotation. Finally, the 14-fold expansion observed for the serine-type endopeptidase activity GO term is attributable to 38 protein groups with the majority share originating from host (85.7%) and minor shares originating from pig (13.5%) and microbes (0.8%). The serine-type endopeptidase from pig is an artifact, as it originates from the sequencing grade porcine trypsin used to generate peptides for LC-MS/MS analysis. Interestingly, human chymotrypsin-like elastase 3A (CEL3A_Human) and chymotrypsin-C (CTRC_human) are more abundant than porcine trypsin, resulting in 35.8% and 14.9% of the GO term share compared with porcine trypsin. Other prominent protein groups (less abundant than porcine trypsin) include chymotrypsin-like elastase 3B (CEL3B_human), cathepsin G (CATG_human), and trypsin-1 (TRY1_human) and comprise 11.3%, 5.1%, and 3.0% of the serine-type endopeptidase activity GO term, respectively.

We performed GO enrichment analysis in the R GOstats package with GO terms corresponding to differentially expressed protein groups serving as the enriched GO term set and GO terms mapping to all nondifferentially expressed protein groups as the "universe" set (64). GO terms over-represented in either healthy or UC cohorts for each GO namespace ($p < 0.01$) are listed in the supplement [see Additional File 1, supplemental Figs. S11–S18].

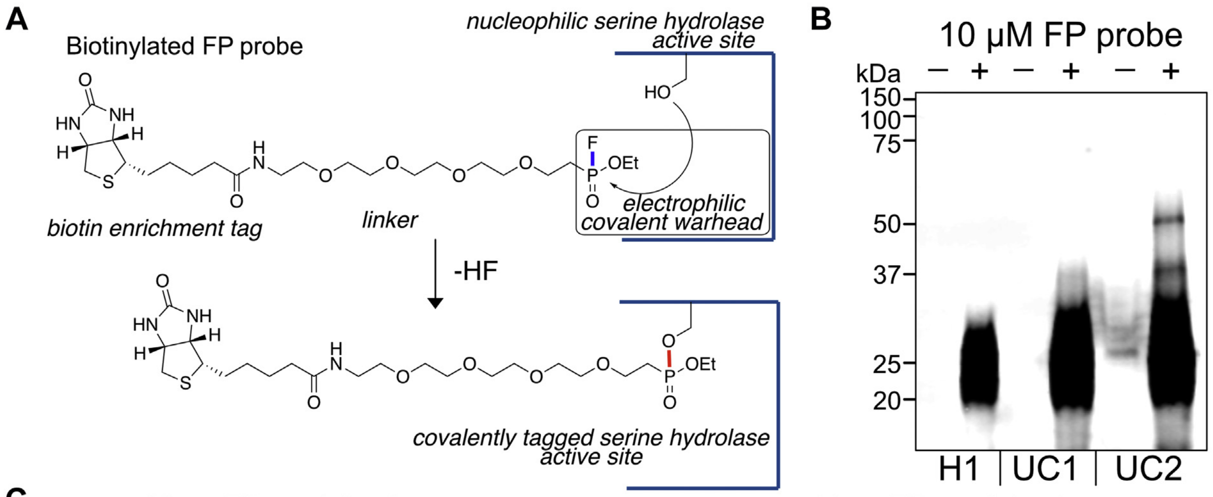
For the healthy patient cohort, nine molecular function GO terms were enriched including hydrolase activity (hydrolyzing O-glycosyl compounds) (GO:0016787), cysteine-type peptidase activity (GO:0008234), rRNA binding (GO:0019843), and oxidoreductase activity (acting on iron–sulfur proteins as donors) (GO:0016730) among the most significantly enriched ($p < 0.001$, odds ratio >10). The biological process namespace contained nine GO terms enriched in healthy patients with polysaccharide catabolic process (GO:0000272), homeostatic process (GO:0042592), sulfur amino acid metabolic process (GO:0000096), nitrogen fixation (GO:0009399), and asexual sporulation (GO:0030436) among the most significantly enriched terms ($p < 0.001$, odds ratio >20). Only three GO cellular component terms were found to be enriched in healthy patients including oxidoreductase complex (GO:1990204), vanadium–iron nitrogenase complex (GO:0016613), and endospore-forming forespore (GO:0042601) ($p < 0.001$, odds ratio >100). Together the three GO namespaces strongly associate carbohydrate processing and microbial sporulation activities with healthy patients.

For the ulcerative colitis cohort, 21 terms were enriched in the molecular function GO namespace including calcium ion binding (GO:0005509), peptidase activity (acting on L-amino acid peptides) (GO:0008233), lipid binding (GO:0008289), catalytic activity (acting on a protein) (GO:0140096), serine-type endopeptidase activity (GO:0004252), serine hydrolase activity (GO:0017171), and calcium-dependent phospholipid binding (GO:0005544) ($p < 0.001$, odds ratio >10) [see Additional File 5]. The biological process namespace contains 18 entries including proteolysis (GO:0006508), aromatic amino acid family metabolic processes (GO:0009072), propionate metabolic process (methylcitrate cycle) (GO:0019679), acetate metabolic process (GO:0006083), short-chain fatty acid metabolic process (GO:0046459), antibacterial humoral response (GO:0019731), antifungal humoral response (GO:0019732), regulation of cytokine production (GO:0001817), and response to fungus (GO:0009620) among the most enriched terms ($p < 0.003$, odds ratio >10). The enriched cellular component term list was much shorter with only five entries, including extracellular space (GO:0005615), organelle lumen (GO:0043233), endoplasmic reticulum lumen (GO:0005788), intracellular membrane-bound organelle (GO:0043231), and extracellular region (GO:0005576) ($p < 0.008$, odds ratio >10). Together, these enriched GO terms associate the extracellular space/secretome, antimicrobial host responses and serine protease activity with UC patients in our study [see Additional File 1, supplemental Figs. S11 and S17]. Interestingly, serine proteases, and more broadly serine hydrolases, are an intensely studied class of enzymes with exquisitely selective mechanism-based fluorophosphonate probes described in the literature (89). These probes present an opportunity to better examine this class of enzyme in the context of human UC.

ABPP Confirms the Presence of Active Serine-type Endopeptidases and Identifies Previously Undetected Serine-type Endopeptidases

We treated three patient fecal samples with a biotinylated fluorophosphonate-based (FP-based) serine-reactive probe to label and thus establish whether serine hydrolases were active in patient fecal samples (Fig. 4*A*) (89). Biotinylated fluorophosphonate probe (FP probe)-labeled proteins were enriched with streptavidin-agarose beads, visualized by Western blot (Fig. 4*B*) and identified by LC-MS/MS analysis using a previously described two-step large-to-focused database search strategy (90). Analysis of the LC-MS/MS data revealed several hundred noncontaminant protein sequences. These sequences were clustered into 104 distinct protein groups (95% similarity cutoff using CD-HIT) and further reduced to 63 protein groups with highly homologous host proteins condensed together. Of note, 27 and 35 protein groups derived from host (Fig. 4*C*) and nonhost (Fig. 4*D*) were identified, respectively.

The majority of host-derived protein groups labeled and enriched with the FP probe were also identified within the unenriched LC-MS/MS datasets [see Additional File 4]. Fourteen of 27 FP probe-enriched host proteins are known serine hydrolases including the chymotrypsin-like elastase family (2A, 2B, 3A, 3B), cathepsin G, dipeptidyl peptidase 4, neutrophil elastase, myeloblastin, trypsin 1, and phospholipase A2 (Fig. 4*C*). Enrichment of these particular hydrolases



C

Host FP-enriched

Cluster Representative	H1	UC1	UC2	Description
A0A286YEY1	🟦	🟧	🟧	Ig heavy constant alpha 1
ATAD2	🟧	🟧	🟦	ATPase family AAA protein 2
BD1L1	🟧	🟧	🟦	Biorientation of chromosomes protein 1
CATG	🟧	🟧	🟦	Cathepsin G
CEL2A	🟦	🟧	🟦	Chymotrypsin-like elastase 2A
CEL2B	🟦	🟧	🟦	Chymotrypsin-like elastase 2B
CEL3A	🟦	🟦	🟦	Chymotrypsin-like elastase 3A
CEL3B	🟦	🟦	🟦	Chymotrypsin-like elastase 3B
CHKB	🟧	🟧	🟦	Choline/ethanolamine kinase
CTRB1	🟧	🟧	🟦	Chymotrypsinogen B
CTRB2	🟦	🟧	🟦	Chymotrypsinogen B2
CTRC	🟦	🟦	🟦	Chymotrypsin C
DEF3	🟧	🟧	🟦	Neutrophil defensin 3
DPP4	🟧	🟧	🟦	Dipeptidyl peptidase 4
E9PS66	🟧	🟧	🟦	Zinc finger protein 69 homolog B
ELNE	🟧	🟧	🟦	Neutrophil elastase
Keratin	🟦	🟦	🟦	Keratin
KLH42	🟧	🟧	🟦	Kelch-like protein 42
LYSC	🟧	🟧	🟦	Lysozyme C
MGAL	🟧	🟧	🟦	Probable maltase-glucoamylase 2
PA21B	🟧	🟧	🟦	Phospholipase A2
PRTN3	🟧	🟧	🟦	Myeloblastin
S10A8	🟧	🟧	🟦	Protein S100-A8
S10A9	🟧	🟧	🟦	Protein S100-A9
TRY1	🟧	🟧	🟦	Trypsin 1
UTP20	🟧	🟧	🟦	Small subunit processome 20

🟧 not detected

🟦 identified in sample

non-Host FP-enriched

Cluster Representative	H1	UC1	UC2	Description
364106.GI:91209495	🟦	🟧	🟦	hypothetical protein, yajQ, DUF520
HMPREF0986_02828	🟧	🟧	🟦	Protease Do
HMPREF9347_00782	🟧	🟧	🟦	50S ribosomal protein L11
HMPREF9545_01601	🟦	🟧	🟦	DNA starvation/stationary phase protection Dps
NUSG_BUCAP	🟦	🟧	🟧	Transcription termination/antitermination NusG
Streptavidin	🟦	🟦	🟦	Streptavidin
SRS015369.41495-T1-C	🟦	🟧	🟧	S9 family peptidase
SRS016095.25851-T1-C	🟦	🟧	🟧	S9 family peptidase
SRS016495.2528-T1-C	🟦	🟦	🟦	hypothetical protein
SRS019161.273588-T1-C	🟦	🟧	🟧	possible esterase, serine hydrolase, beta-lactamase
SRS019397.59782-T1-C	🟦	🟧	🟦	S9 family peptidase
SRS023176.73473-T1-C	🟧	🟦	🟧	biotin attachment protein
SRS064276.13938-T1-C	🟧	🟧	🟦	acetyl-CoA carboxylase biotin carboxyl carrier
TRYP_PIG	🟦	🟦	🟦	Trypsin (Pig)
WP_000940350.1	🟧	🟧	🟦	30S ribosomal protein S5
WP_001054895.1	🟧	🟧	🟦	serine endoprotease DegP
WP_007563239.1	🟦	🟧	🟧	S9 family peptidase
WP_012904828.1	🟧	🟧	🟦	YajQ family cyclic di-GMP-binding protein
WP_013200502.1	🟧	🟧	🟦	50S ribosomal protein L9
WP_013975808.1	🟧	🟧	🟦	transcription termination/antitermination protein NusG
WP_024994469.1	🟦	🟧	🟧	S9 family peptidase
WP_026823563.1	🟦	🟧	🟦	transcription termination/antitermination NusG
WP_032167473.1	🟦	🟧	🟦	transcription termination/antitermination NusG
WP_034750989.1	🟦	🟧	🟦	DUF1015 domain-containing protein
WP_036222587.1	🟦	🟧	🟧	winged helix DNA-binding protein
WP_044948116.1	🟦	🟧	🟧	biotin/lipoyl-binding carrier protein
WP_062747128.1	🟧	🟧	🟦	serine endoprotease DegP
WP_067704128.1	🟧	🟧	🟦	50S ribosomal protein L9
WP_072107813.1	🟦	🟧	🟦	50S ribosomal protein L6
WP_074680184.1	🟦	🟧	🟧	biotin/lipoyl-binding carrier protein
WP_082715169.1	🟧	🟦	🟧	nitroreductase family protein, partial
WP_086076416.1	🟧	🟧	🟦	IS5-like element ISKpn26 family transposase
WP_087262956.1	🟧	🟧	🟦	M23 family metallopeptidase
WP_090115612.1	🟦	🟦	🟧	Holliday junction migration DNA helicase RuvB

FIG. 4. **Targeting serine hydrolases *via* activity-based protein profiling.** *A*, FP probe structure and reaction schematic for covalent attachment to nucleophilic active-site serine in hydrolases. *B*, Western blot of three patient fecal lysates treated and enriched with FP probe followed by streptavidin bead enrichment visualized with fluorophore-conjugated streptavidin. Host (*C*) and nonhost (*D*) proteins from patient fecal samples enriched by FP probe and detected by LC-MS/MS (*blue* = protein detected in corresponding patient sample, *orange* = protein not detected in corresponding patient sample, *green highlights* = annotated proteases). FP, fluorophosphonate; LC-MS/MS, liquid chromatography tandem mass spectrometry.

over all other host proteins provides confidence that the FP probe is selective for nucleophilic serine hydrolases in the tremendously complex fecal protein matrix. Aside from demonstrating that the hydrolases are active, these results also suggest that this fraction of proteases remain uninhibited by antiproteolytic proteins often found in the gut (91–93).

Most nonhost proteins are microbial in origin with the exception of streptavidin and porcine trypsin introduced during sample preparation (Fig. 4D). The most promising FP probe-susceptible proteins include protease Do entries (DegP), S9 family peptidases, and a beta-lactamase, as determined by sequence analysis. Interestingly, of the ten identified nonhost serine hydrolase protein groups, only one (SRS019397.59782-T1-C) was detected without FP-probe demonstrating the utility of chemical-based enrichment strategies for the identification of novel proteins in a complex environment. Of the 167,554 MS2 spectra collected for all FP-enriched LC-MS/MS data sets, only 5352 (3.2%) were assigned by ComPIL database searches. There is a strong likelihood that other serine hydrolase-derived peptides are present in our microbiota samples, but they remain unidentified due to limitations imposed by the incompleteness problem associated with metaproteomics database searching (26). Unfortunately, the techniques for database-independent, high-confidence identification of these peptides and their parent protein sequences are currently not well established.

De Novo *Peptide Sequencing Enables Glimpses into the Dark Peptidome*

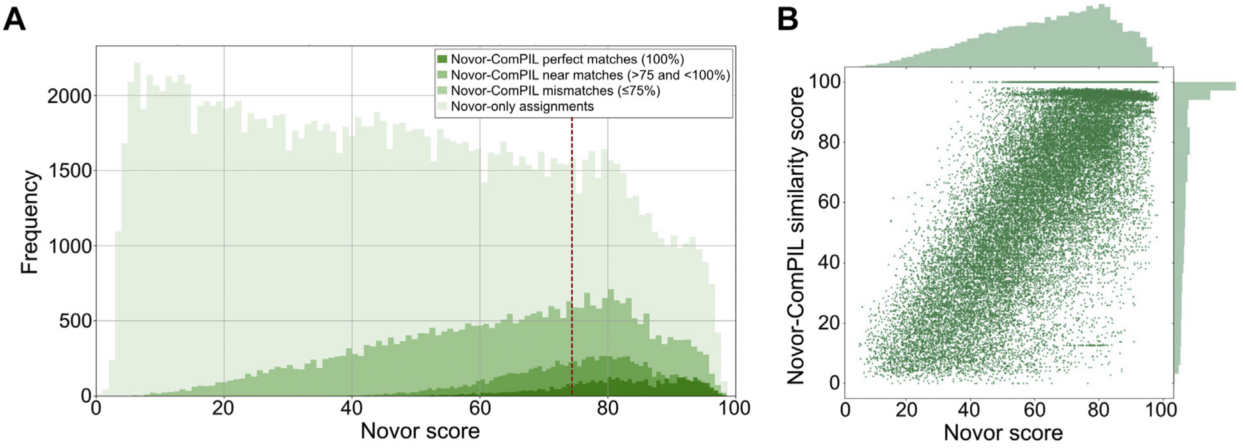
Of the 2,829,920 MS2 fragmentation spectra we collected overall, 523,155 (18.5%) were matched to a corresponding peptide with a 1% peptide false positive rate using a target-decoy search strategy paired with the ComPIL database. The modest number of peptide spectrum matches we observed is likely attributable to [1] a loss in filtering sensitivity that often accompanies database expansion (90, 94) and [2] a never-complete database that is perennially associated with metaproteomics. We posit that a nontrivial portion of unmatched MS2 spectra map to either known or unknown peptide sequences, and we aim to estimate the size of unmatched MS2 spectrum space or “the dark peptidome,” using a complimentary *de novo* peptide sequencing approach (95).

We subjected MS2 spectra from all patient fecal sample LC-MS/MS datasets to *de novo* peptide sequencing using the Novor algorithm (96). Novor attempts to deduce peptide sequence from MS2 fragmentation spectra, generating a *de novo* peptide spectrum match (PSM) and an accompanying confidence score (Novor score; higher scores indicated better predicted matches). Where available, we paired *de novo* PSMs with their corresponding database PSMs (ComPIL2-assigned) and calculated an additional Novor-ComPIL2 similarity score (*de novo* database similarity score) based on the Needleman–Wunsch comparison algorithm (raw scores were scaled to 100, where 100 represents a perfect match) (68). We used these values to construct overlapping histograms [Fig. 5*A* and Additional File 1, supplemental Figs. S19 and S20] and joint plots [Fig. 5*B* and Additional File 1, supplemental Figs. S21 and S22] depicting possible unidentified peptide space in patient fecal microbiota samples.

ComPIL database-assigned PSMs were nonuniformly distributed along the Novor score axis with a larger proportion of database PSMs grouped near the high-confidence *de novo* sequencing Novor scores (69–99) (Fig. 5*A*). While perfect *de novo* database agreements were rare, a large proportion of database PSMs possessed strong similarity to *de novo* PSMs, a relationship best depicted by a Novor score *versus* Novor-ComPIL2 similarity score joint plot (Fig. 5*B*). Thus, it is reasonable to expect that above a conservative cutoff value (Novor score = 75), significant numbers of MS2 spectra without database assignments correspond to peptides that either are not contained in the ComPIL2 database or were rejected due to our high search-filter stringency. Based on this assessment, it is estimated that an average of 14,075 MS2 spectra with a Novor score of 75 or greater remain unidentified per sample (Fig. 5*C*). This corresponds to approximately 9% of all MS2 spectra per patient sample, which could increase global identifications by approximately 50%. Note that 9% is a lower limit estimate for global unidentified MS2 spectra as this value is based only on MS2 spectra with Novor scores >75. A significant fraction of MS2 with Novor scores <75 have been identified by ComPIL and further suggest that the upper limit of unidentified MS2 is much larger than 9% (Fig. 5*A*, left of red vertical dashed line). BLAST analysis of several *de novo* PSMs, which do not have corresponding database PSMs, returned reasonable, high-similarity score matches to microbial peptides from the NCBI nonredundant database supporting this assertion. Below a Novor score of 75, there are likely many unidentified peptide-MS2; however, these peptides are also likely intermixed with many non-peptide-MS2.

DISCUSSION

16S rRNA gene amplicon sequencing has been a workhorse technique for microbiota studies over the last decade due in large part to the simplicity of material extraction and the broad availability of resources needed to generate meaningful data. In contrast, the use of LC-MS/MS-based metaproteomics profiling has been more sparse for the opposite reasons. While functional proteomic interrogation of the microbiome by LC-MS/MS was our main goal, we considered it important to contrast our proteomics-based taxonomy findings with those generated by 16S sequencing, a technique more familiar to the microbiome research community. Gut microbiome taxonomy through the lens of 16S gene analysis *versus* LC-MS/MS-based proteomics is expectedly different, but in some unexpected ways (34). At the phylum level, we expected to see a similar configuration by both techniques with the exception that proteomics would include a small concession for host proteins. Unexpectedly, we observed both host and diet-derived (potato, rice, corn, *etc.*) proteins in great relative abundance to microbial proteins. Our analyzed samples were pre-enriched for microbes by filtration, differential centrifugation, and several washing steps, yet nearly half of the detected proteome was mapped back to host or dietary plant proteins. Another unexpected finding was a discrepancy between the relative abundance ratios of Bacteroidetes and Firmicutes. This discrepancy could be artifactual and originate from differences in DNA extraction efficiencies between microbes, but it could also be the result of biologically relevant phenomena. For example, overrepresentation of Firmicutes by 16S



C

Sample	detected proteins	total MS2 collected	total MS2 assigned by ComPIL (forward matches)		MS2 with Novor score >75									MS2 without ComPIL ID that are likely peptides as % of all collected MS2
					MS2	MS2 with ComPIL ID		MS2 with ComPIL ID and 100% Novor-ComPIL similarity score		MS2 with ComPIL ID and >75% Novor-ComPIL similarity score		MS2 without ComPIL ID that are likely peptides		
	#	#	#	%	#	#	%	#	%	#	%	#	%	%
H1	131463	153453	32451	21.1	20565	8636	42.0	1554	18.0	6558	75.9	**9059**	44.1	6
H2	152714	158254	32726	20.7	27715	11270	40.7	3340	29.6	9097	80.7	**13274**	47.9	8
H3	107712	155810	22031	14.1	28924	7619	26.3	2293	30.1	6074	79.7	**16985**	58.7	11
H4	132014	156004	28838	18.5	31488	9904	31.5	2577	26.0	7679	77.5	**16735**	53.1	11
H5	144506	161284	24143	15.0	29247	8652	29.6	2352	27.2	6751	78.0	**16070**	54.9	10
H6	123522	155593	25899	16.6	29533	8853	30.0	2833	32.0	7053	79.7	**16475**	55.8	11
H7	252892	156731	46855	29.9	31946	14422	45.1	4286	29.7	11356	78.7	**13799**	43.2	9
H8	152904	157453	30294	19.2	28575	10365	36.3	1550	15.0	7690	74.2	**13510**	47.3	9
UC1	180806	152184	43253	28.4	24068	11394	47.3	2327	20.4	8719	76.5	**9698**	40.3	6
UC2	197005	154920	40945	26.4	25683	11715	45.6	3030	25.9	9140	78.0	**10898**	42.4	7
UC4	109317	156887	20619	13.1	29566	7252	24.5	1757	24.2	5424	74.8	**16689**	56.4	11
UC5	58849	154458	15151	9.8	25759	5413	21.0	1367	25.3	4136	76.4	**15546**	60.4	10
UC6	6092	157954	7999	5.1	16458	2705	16.4	767	28.4	2264	83.7	**11511**	69.9	7
UC8	73672	153576	19648	12.8	26349	6732	25.5	1566	23.3	5312	78.9	**15479**	58.7	10
UC9	148887	164245	34881	21.2	30443	12461	40.9	3578	28.7	10049	80.6	**14501**	47.6	9
UC11	5877	158549	21033	13.3	21715	7703	35.5	2302	29.9	6451	83.7	**11735**	54.0	7
UC15	244952	160626	44974	28.0	35872	15738	43.9	4256	27.0	12233	77.7	**15650**	43.6	10
UC23F	160325	161939	31415	19.4	30943	11117	35.9	3077	27.7	8824	79.4	**15737**	50.9	10
Mean	132417	157218	29064	**18.5**	27492	9553	34.3	2490	26.0	7489	78.6	**14075**	51.6	9

Fig. 5. **Estimating the size of database-elusive peptide ("dark peptidome") space *via de novo* peptide sequencing.** *A*, overlapping histograms of all H2 patient sample MS2 spectra by Novor score (0–100, x-axis); *darkest green area* represents MS2 correctly assigned by Novor determined by comparison to ComPIL (database) result; *lightest green area* represents MS2 without ComPIL peptide assignments; *dashed vertical red line* represents Novor score = 75 cutoff and MS2 to the *right* of this line were used to estimate the size of unassigned peptide space. *B*, joint plot of H2 patient sample MS2 spectra depicting correlation between Novor score (0–100) and Novor-ComPIL similarity score (0–100). *C*, number of MS2 spectra from all patient samples with Novor scores >75 that likely represent peptides but do not have ComPIL peptide matches ("dark peptidome"). With duplicates removed, the total number of proteins detected between all 18 samples is 576,625. ComPIL, comprehensive protein identification library.

sequencing could stem from an abundance of Firmicutes cells that are metabolically inactive (spores) relative to Bacteroidetes cells (97). This would be in agreement with our finding that the asexual sporulation GO term is enriched in healthy patient samples. Finally, we found that at more granular taxonomic strata [see Additional File 1, supplemental Figs. S1–S7], the number of identifiable organisms was unexpectedly greater for proteomics (using unique peptides as a proxy) than for 16S sequencing. By proteomics, we observed peptides originating from hundreds of organisms at the species level *versus* several dozen by 16S. Note that in our case, a standard 16S V4 analysis by paired-end short-read sequencing was performed; however, other higher-resolution techniques are becoming more accessible (98–100). Expectedly, the proportion of uniquely mappable peptides progressively decreased at more granular taxonomic levels such that at the species level, about 75% of all peptide intensity could not be mapped to a particular species. Because peptides are proxies for both taxonomy and function, this observation hints at a functional redundancy among microbes in the gut, which could be better examined by differential expression and GO term analysis of proteomics data.

One of the leading motivations for performing differential expression analyses on microbiome samples is to identify specific biomarkers or disease-associated microbial proteins for further examination. Toward this goal, we identified 176 protein groups significantly enriched ($q < 0.1$) in either healthy or UC volunteers, with a major share originating from microbes. Interestingly, no host proteins were identified as significantly enriched in the fecal extracts of healthy volunteers while several were found enriched in the fecal extracts of UC patients. Among the host proteins enriched in UC patients, we identified the calibrating entry, protein S100-A9 ($p < 0.004$, $q < 0.07$), a component of fecal calprotectin and established IBD biomarker (101, 102). According to STRING and reactome analyses, many host proteins enriched in UC patients are also inflammation-aligned lending more credibility to the prospect that the enriched proteins we have identified are truly UC-associated. For a comprehensive list of enriched protein groups, see Additional File 4. While most enriched protein groups had some annotation, a significant portion had little to none. This finding presents an exciting opportunity for the structural and biochemical study of novel sequences. Given the enormous number of domains of unknown function (DUF) and unknown function-type proteins catalogued from microbiome metagenomic sequencing efforts, we are faced with a prioritization problem wherein the most disease-relevant sequences are obscured by less impactful ones (103). LC-MS/MS-based proteomics appears in this context to be an important tool for identifying sequences that are both expressed and biologically relevant, which will help focus our future studies. Lastly, it is important to point out that poorly annotated proteins (*i.e.* proteins without GO assignments) factor weakly or not at all into broader functional analyses like GO enrichment simply due to the nature of enrichment testing (*i.e.* hypergeometric). Therefore, novel sequences without known biological- or disease-relevance are important to eventually characterize.

Within the detected microbiome proteome, known functional diversity is high with several hundred molecular function GO terms represented. A flat depiction of molecular function wherein a 1-sequence-1-count paradigm is applied reveals a consistent relative abundance configuration between all samples. We reasoned that count-based GO term depictions effectively reveal molecular diversity as each GO term's constituent protein group/cluster equally contributes to a term's size. However, this measure alone fails to capture material abundance. To depict material abundance, we have instead weighted GO terms by the peptide intensities (a very loose proxy for protein copy number) of their constituent protein groups/clusters (see Additional File 1 for details regarding this calculation procedure). Interestingly, when this paradigm is applied, a very different picture emerges. The relative abundance of many molecular function GO terms shifts, sometimes dramatically. One of the most conspicuous terms to us was "serine-type endopeptidase activity" that expanded an average of 14-fold among all patient samples going from count-based to intensity-based representation. Additionally, we found this same term enriched in UC patient fecal samples by hypergeometric testing, warranting a closer inspection of the protein groups that contribute to this term. We found that the major contributors were host-derived serine proteases (85.7%) such as chymotrypsin-C and the chymotrypsin-like elastase family with minor contributions from porcine trypsin (13.5%) (an artifact of sample preparation) and microbial serine proteases (0.8%). The high relative abundance of serine proteases in fecal samples is not surprising given that they are important components of host digestive enzyme cocktails secreted into the gut lumen. We were, however, surprised to find both host and microbial serpins, which are known active-site directed suicide inhibitors for serine/cysteine proteases, in fecal samples. This observation suggests that there might be important regulatory host–microbe cross talk with respect to proteolytic activity that occurs in the gut. By comparing the abundance of proteases or serpins, it would still be difficult to determine which and what fraction of serine proteases remained active upon fecal sample collection. To identify active serine proteases, we treated fecal samples with an active-site directed serine-hydrolase selective chemical probe (FP probe) for labeling, enrichment, and target identification *via* LC-MS/MS (ABPP) (29, 30, 89). We examined three patient fecal samples (one healthy, two UC) and qualitatively found human chymotrypsin-like elastases 3A and 3B and chymotrypsin-C enriched and therefore active in all samples. For one UC sample (UC2), we identified additional FP probe-enriched host proteases including cathepsin G, chymotrypsin-like elastase 2A and 2B, dipeptidyl peptidase 4, neutrophil elastase, and trypsin 1, lending support to the idea that aberrantly increased protease activity is associated with IBD (92, 104). In addition to host proteases, we identified several microbial proteases from all three patient samples upon FP probe enrichment. Surprisingly, we found that nine of ten microbial proteases were not detected by LC-MS/MS at all without FP probe enrichment. This finding suggests that there are likely many microbial proteases expressed in the gut microbiota and that they are likely below the limit of detection by most current sampling and LC-MS/MS profiling strategies. We speculate that this sentiment also holds true for other low-abundance, high-impact protein functionalities, underscoring the importance of pre-enrichment strategies for future proteomics studies.

In addition to sampling limits, many microbial peptides that are sampled by LC-MS/MS are liable to go undetected due in large part to database-completeness limitations. We attempted to estimate the number of peptide-likely fragmentation spectra per LC-MS/MS experiment using a *de novo* sequencing tool (Novor) in order to define a rough boundary around the amount of unassigned peptide space captured by the mass spectrometer but unidentified by our database workflow. Homology searching of high-confidence peptide-like fragmentation spectra revealed high numbers of exact and near-matches to peptide sequences in the NCBI nonredundant database. Though *de novo* peptide sequencing coupled to homology searching can help capture database-elusive peptides in more defined contexts (95), we were reluctant to rely more heavily on this strategy without a stringent methodology for distinguishing between sequencing errors and truly homologous peptides, especially in a context as taxonomically diverse as the gut microbiota. An additional layer of difficulty rests in determining how to treat *de novo* only peptides that are constituents of completely unknown parent protein sequences. Deep genome sequencing and the use of custom MAGs are an obvious path forward especially as long-read technologies become more accurate and accessible (15, 22–25, 105–108). Notably, long-read technologies are expected to yield more contiguous genome assemblies, thus accelerating the functional annotation process for novel sequences by enhancing our ability to contextualize these novel sequences within their respective genomes. For peptides/proteins that elude this approach, however (low abundance microbes, heavily posttranslationally modified peptides/proteins, nonribosomal peptides/proteins, *etc.*), perhaps genomics-agnostic, middle- and top-down proteomics sequencing could be applied (109, 110). We anticipate that the expanded use of ABPP techniques in the microbiota will enrich for many protein sequences not contained in large databases such as ComPIL, and robust high-throughput methods for identifying these whole novel protein sequences will be needed.

In summary, we identified 176 discrete host and microbial protein groups differentially enriched between healthy and UC patients. Our analysis revealed several protein functions associated with ulcerative colitis, with the function “serine-type endopeptidase activity” featuring prominently. We also identified host and microbial serine protease inhibitors in concert with serine proteases. Using an activity-based chemical tagging strategy, we were able to enrich for serine hydrolases/proteases and showed that these enzymes are still active in the gut despite the presence of active-site directed protease inhibitors. This strategy also revealed the presence of previously undetected serine proteases demonstrating the utility of activity-based tagging for the amplification of low-abundance proteins. Finally, we paired our database metaproteomics strategy with *de novo* peptide sequencing to estimate the size of high-confidence peptide space in our samples that remains unidentified despite the use of a large comprehensive database. Our data suggests that at the lower bound, at least an average of 9% of all our collected fragmentation spectra per run likely correspond to peptides, but remain unmatched.

DATA AVAILABILITY

The datasets generated and analyzed during the current study are available in the Proteomics Identification Database (PRIDE) repository (69) under project PXD022433 or the Zenodo data repository (55) with https://doi.org/10.5281/zenodo.5717460.

Supplemental data—This article contains supplemental data.

Acknowledgments—We thank G. Tsaprailis and C. Sharager-Tapia for assistance with mass spectrometry data collection. We thank S. K. R. Park, T. Jung, and J. R. Yates for assistance with mass spectrometry data analysis. We thank H. Rosen, R. L. Wiseman, and L. L. Lairson for access to instrumentation.

Funding and additional information—This work was supported by NIH R21AI139744 (to D. W. W.) and Boehringer Ingelheim (to D. W. W., A. I. S., and W. J. C.) The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions—P. S. T.-B., A. Y. W., and J. H. X. data curation; P. S. T.-B., S. C., G. S. S., A. I. S., and D. W. W. formal analysis; P. S. T.-B., A. Y. W., and J. H. X. investigation; P. S. T.-B., W. J. C., and D. W. W. methodology; A. C.-M. and W. J. C. resources, P. S. T.-B. and D. W. W. writing—original draft; P. S. T.-B., A. Y. W., A. C.-M., J. H. X., S. C., G. S. S., A. S., W. J. C., and D. W. W. writing—review and editing.

Conflict of interest—The authors declare no competing interests.

Abbreviations—The abbreviations used are: ABPP, activity/affinity-based protein profiling; BLAST, basic local alignment search tool; CD, Crohn’s disease; ComPIL, comprehensive protein identification library; DNA, deoxyribonucleic acid; FP, fluorophosphonate; GI, gastrointestinal tract; GO, gene ontology; GWAS, genome-wide association studies; IBD, inflammatory bowel disease; LC-MS/MS, liquid chromatography tandem mass spectrometry; MAG, metagenome assembled genome; MS, mass spectrometry; NCBI, National Center for Biotechnology Information; PBS, phosphate buffered saline; PCA, principal component analysis; PPI, protein–protein interaction; PSM, peptide-spectrum match; QIIME, quantitative insights into microbial ecology; rRNA, ribosomal ribonucleic acid; STRING, search tool for recurring instances of neighboring genes; UC, ulcerative colitis.

Received June 4, 2021, and in revised from, January 10, 2022
Published, MCPRO Papers in Press, January 13, 2022, https://doi.org/10.1016/j.mcpro.2022.100197

REFERENCES

1. Baumgart, D. C., and Carding, S. R. (2007) Inflammatory bowel disease: Cause and immunobiology. *Lancet* **369**, 1627–1640
2. Baumgart, D. C., and Sandborn, W. J. (2007) Inflammatory bowel disease: Clinical aspects and established and evolving therapies. *Lancet* **369**, 1641–1657
3. Bernstein, C. N., Blanchard, J. F., Kliewer, E., and Wajda, A. (2001) Cancer risk in patients with inflammatory bowel disease: A population-based study. *Cancer* **91**, 854–862
4. Molodecky, N. A., Soon, I. S., Rabi, D. M., Ghali, W. A., Ferris, M., Chernoff, G., Benchimol, E. I., Panaccione, R., Ghosh, S., Barkema, H. W., and Kaplan, G. G. (2012) Increasing incidence and prevalence of inflammatory bowel diseases with time, based on systematic review. *Gastroenterology* **142**, 46–54
5. Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., Essers, J., Mitrovic, M., Ning, K., Cleynen, I., Theatre, E., *et al*. (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124
6. Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., Ripke, S., Lee, J. C., Jostins, L., Shah, T., Abedian, S., Cheon, J. H., Cho, J., Dayani, N. E., Franke, L., *et al*. (2015) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986
7. de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S. G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., *et al*. (2017) Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261
8. Brant, S. R. (2011) Update on the heritability of inflammatory bowel disease: The importance of twin studies. *Inflamm. Bowel Dis.* **17**, 1–5
9. Dalal, S. R., and Chang, E. B. (2014) The microbial basis of inflammatory bowel diseases. *J. Clin. Invest.* **124**, 4190–4196
10. Nishida, A., Inoue, R., Inatomi, O., Bamba, S., Naito, Y., and Andoh, A. (2018) Gut microbiota in the pathogenesis of inflammatory bowel disease. *Clin. J. Gastroenterol.* **11**, 1–10
11. Ni, J., Wu, G. D., Albenberg, L., and Tomov, V. T. (2017) Gut microbiota and IBD: Causation or correlation? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 573–584
12. Manichanh, C., Borruel, N., Casellas, F., and Guarner, F. (2012) The gut microbiota in IBD. *Nat. Rev. Gastroenterol. Hepatol.* **9**, 599–608
13. Caruso, R., Lo, B. C., and Nunez, G. (2020) Host-microbiota interactions in inflammatory bowel disease. *Nat. Rev. Immunol.* **20**, 411–426
14. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., *et al*. (2010) A human gut microbial gene catalog established by metagenomic sequencing. *Nature* **464**, 59–65
15. Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., Lawley, T. D., and Finn, R. D. (2019) A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504
16. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012) Diversitiy, stability, and resilience of the human gut microbiota. *Nature* **489**, 220–230
17. Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355–1359
18. Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., and Huttenhower, C. (2015) Sequencing and beyond: Integrating molecular 'omics' for microbial community profiling. *Nat. Rev. Microbiol.* **13**, 360–372
19. Wilmes, P., and Bond, P. L. (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ. Microbiol.* **6**, 911–920
20. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989
21. Stewart, E. J. (2012) Growing unculturable bacteria. *J. Bacteriol.* **194**, 4151–4160
22. Dang, D. D., Froula, J., Egan, R., and Wang, Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ.* **3**, e1165
23. Chen, L.-X., Anantharaman, K., Shaiber, A., Muran Eren, A., and Banfield, J. F. (2020) Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333
24. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., *et al*. (2019) Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662
25. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., and Kyrpides, N. C. (2019) New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510
26. Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., Muth, T., Rapp, E., Martens, L., Addis, M. F., and Uzzau, S. (2016) The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome* **4**, 51
27. Chatterjee, S., Stupp, G. S., Park, S. K. R., Ducom, J.-C., Yates, J. R., Su, A. I., and Wolan, D. W. (2016) A comprehensive and scalable database search system for metaproteomics. *BMC Genomics* **17**, 642
28. Park, S. K. R., Jung, T., Thuy-Boun, P. S., Wang, A. Y., Yates, J. R., and Wolan, D. W. (2019) ComPIL 2.0: An updated comprehensive metaproteomics database. *J. Proteome Res.* **18**, 616–622
29. Jessani, N., and Cravatt, B. F. (2004) The development and application of methods for activity-based protein profiling. *Curr. Opin. Chem. Biol.* **8**, 54–59
30. Cravatt, B. F., Wright, A. T., and Kozarich, J. W. (2008) Activity-based protein profiling: From enzyme chemistry to proteomic chemistry. *Annu. Rev. Biochem.* **77**, 383–414
31. Ram, R. J., Verberkmoes, N. C., Thelen, M. P., Tyson, G. W., Baker, B. J., Blake, R. C., 2nd, Shah, M., Hettich, R. L., and Banfield, J. F. (2005) Community proteomics of a natural microbial biofilm. *Science* **308**, 1915–1920
32. Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43
33. VerBerkmoes, N. C., Denef, V. J., Hettich, R. L., and Banfield, J. F. (2009) Functional analysis of natural microbial consortia using community proteomics. *Nat. Rev. Microbiol.* **7**, 196–205
34. Verberkmoes, N. C., Russell, A. L., Shah, M., Godzik, A., Rosenquist, M., Halfvarson, J., Lefsrud, M. G., Apajalahti, J., Tysk, C., Hettich, R. L., and Jansson, J. K. (2009) Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* **3**, 179–189
35. Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., Raes, J., Verberkmoes, N. C., Fraser, C. M., Hettich, R. L., and Jansson, J. K. (2012) Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* **7**, e49138
36. Juste, C., Kreil, D. P., Beauvallet, C., Guillot, A., Vaca, S., Carapito, C., Mondot, S., Sykacek, P., Sokol, H., Blon, F., Lepercq, P., Levenez, F., Valot, B., Carré, W., Loux, V., *et al*. (2016) Bacterial protein signals are associated with Crohn's disease. *Gut* **63**, 11566–11577
37. Zhang, X., Ning, Z., Mayne, J., Deeke, S. A., Li, J., Starr, A. E., Chen, R., Singleton, R., Butcher, J., Mack, D. R., Stintzi, A., and Figeys, D. (2016) *In vitro* metabolic labeling of intestinal microbiota for quantitative metaproteomics. *Anal. Chem.* **88**, 6120–6125
38. Heintz-Buschart, A., May, P., Laczny, C. C., Lebrun, L. A., Bellora, C., Krishna, A., Wampach, L., Schneider, J. G., Hogan, A., de Beaufort, C., and Wilmes, P. (2016) Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* **2**, 16180
39. Zhang, X., Chen, W., Ning, Z., Mayne, J., Mack, D., Stintzi, A., Tian, R., and Figeys, D. (2017) Deep metaproteomics approach for the study of human microbiomes. *Anal. Chem.* **89**, 9407–9415
40. Zhang, X., Deeke, S. A., Ning, Z., Starr, A. E., Butcher, J., Li, J., Mayne, J., Cheng, K., Liao, B., Li, L., Singleton, R., Mack, D., Stintzi, A., and Figeys, D. (2018) Metaproteomics reveals associations between

microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nat. Commun.* **9**, 2873

41. Mills, R. H., Vázquez-Baeza, Y., Zhu, Q., Jiang, L., Gaffney, J., Humphrey, G., Smarr, L., Knight, R., and Gonzalez, D. J. (2019) Evaluating metagenomic prediction of the metaproteome in a 4.5-year study of a patient with Crohn's disease. *mSystems* **4**, e00337-18
42. Blakely-Ruiz, J. A., Erickson, A. R., Cantarel, B. L., Xiong, W., Adams, R., Jansson, J. K., Fraser, C. M., and Hettich, R. L. (2019) Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes. *Microbiome* **7**, 18
43. Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., *et al*. (2019) Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662
44. Lehmann, T., Schallert, K., Vilchez-Vargas, R., Benndorf, D., Puttker, S., Sydor, S., Schulz, C., Bechmann, L., Canbay, A., Heidrich, B., Reichl, U., Link, A., and Heyer, R. (2019) Metaproteomics of fecal samples of Crohn's disease and ulcerative colitis. *J. Proteomics* **201**, 93–103
45. Mayers, M. D., Moon, C., Stupp, G. S., Su, A. I., and Wolan, D. W. (2017) Quantitative metaproteomics and activity-based probe enrichment reveals significant alterations in protein expression from a mouse model of inflammatory bowel disease. *J. Proteome Res.* **16**, 1014–1026
46. Whidbey, C., Sadler, N. C., Nair, R. N., Volk, R. F., DeLeon, A. J., Bramer, L. M., Fansler, S. J., Hansen, J. R., Shukla, A. K., Jansson, J. K., Thrall, B. D., and Wright, A. T. (2019) A probe-enabled approach for the selective isolation and characterization of functionally active subpopulations in the gut microbiome. *J. Am. Chem. Soc.* **141**, 42–47
47. Parasar, B., Zhou, H., Xiao, X., Shi, Q., Brito, I. L., and Chang, P. V. (2019) Chemoproteomic profiling of gut microbiota-associated bile salt hydrolase activity. *ACS Cent. Sci.* **5**, 867–873
48. Jariwala, P. B., Pellock, S. J., Goldfarb, D., Cloer, E. W., Artola, M., Simpson, J. B., Bhatt, A. P., Walton, W. G., Roberts, L. R., Major, M. B., Davies, G. J., Overkleeft, H. S., and Redinbo, M. R. (2020) Discovering the microbial enzymes driving drug toxicity with activity-based protein profiling. *ACS Chem. Biol.* **15**, 217–225
49. Wang, A. Y., Thuy-Boun, P. S., Stupp, G. S., Su, A. I., and Wolan, D. W. (2018) Triflic acid treatment enables LC-MS/MS analysis of insoluble bacterial biomass. *J. Proteome Res.* **17**, 2978–2986
50. Xu, T., Venable, J. D., Park, S. K., Cociorva, D., Lu, B., Liao, L., Wohlschlegel, J., Hewel, J., and Yates, J. R. (2006) ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. *Mol. Cell. Proteomics* **5**, S174
51. Xu, T., Park, S. K., Venable, J. D., Wohlschlegel, J. A., Diedrich, J. K., Cociorva, D., Lu, B., Liao, L., Hewel, J., Han, X., Wong, C., Fonslow, B., Delahunty, C., Gao, Y., Shah, H., *et al*. (2015) ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. *J. Proteomics* **129**, 16–24
52. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., *et al*. (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29
53. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., *et al*. (2019) Reproducible, interactive, scalable, and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857
54. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596
55. Thuy-Boun, P. S., and Wolan, D. W. (2020) Quantitative metaproteomics and activity-based protein profiling of patient fecal microbiota identifies host and microbial proteins associated with ulcerative colitis. *Zenodo*. https://doi.org/10.5281/zenodo.5717460
56. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identification by mass spectrometry. *Nat. Methods* **4**, 207–214
57. Li, W., Jaroszewski, L., and Godzik, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics* **17**, 282–283
58. Li, W., and Godzik, A. (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659
59. Millikin, R. J., Solntsev, S. K., Shortreed, M. R., and Smith, L. M. (2018) Ultrafast peptide label-free quantification with FlashLFQ. *J. Proteome Res.* **17**, 386–391
60. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47
61. Zhang, X., Smits, A. H., van Tilburg, G. B., Ovaa, H., Huber, W., and Vermeulen, M. (2018) Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nat. Protoc.* **13**, 530–550
62. Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., *et al*. (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240
63. Burge, S., Kelly, E., Lonsdale, D., Mutowo-Muellenet, P., McAnulla, C., Mitchell, A., Sangrador-Vegas, A., Mulder, N., and Hunter, S. (2012) Manual GO annotation of predictive protein signatures: The InterPro approach to GO curation. *Database*. https://doi.org/10.1093/database/bar068
64. Falcon, S., and Gentleman, R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258
65. Mesuere, B., Devreese, B., Debyser, G., Aerts, M., Vandamme, P., and Dawyndt, P. (2012) Unipept: Tryptic peptide-based biodiversity analysis of metaproteome samples. *J. Proteome Res.* **11**, 5773–5780
66. Mesuere, B., Debyser, G., Aerts, M., Devreese, B., Vandamme, P., and Dawyndt, P. (2015) The Unipept metaproteomics analysis pipeline. *Proteomics* **15**, 1437–1442
67. Singh, R. G., Tanca, A., Palomba, A., Van der Jeugt, F., Verschaffelt, P., Uzzau, S., Martens, L., Dawyndt, P., and Mesuere, B. (2019) Unipept 4.0: Functional analysis of metaproteome data. *J. Proteome Res.* **8**, 606–615
68. Needleman, S. B., and Wunsch, C. D. (1970) A general method applicable to search for similarities in amino acid sequence of 2 proteins. *J. Mol. Biol.* **48**, 443–453
69. Thuy-Boun, P. S., and Wolan, D. W. (2020) Ulcerative colitis human gut microbiome. *Proteomics Identification Database (PRIDE)*. PXD022433
70. Tabb, D. L., McDonald, W. H., and Yates, J. R. (2002) DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1**, 21–26
71. Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9440–9445
72. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., and von Mering, C. (2019) STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613
73. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410
74. Ricaboni, D., Maihe, M., Khelaifia, S., Raoult, D., and Million, M. (2016) Romboutsia timonensis, a new species isolated from the human gut. *New Microbes New Infect.* **12**, 6–7
75. Mangifesta, M., Mancabelli, L., Milani, C., Gaiani, F., de'Angelis, N., de'Angelis, G. L., van Sinderen, D., Ventura, M., and Turroni, F. (2018) Mucosal microbiota of intestinal polyps reveals putative biomarkers of colorectal cancer. *Sci. Rep.* **8**, 13974
76. Engevik, M. A., Morra, C. N., Röth, D., Engevik, K., Spinler, J. K., Devaraj, S., Crawford, S. E., Estes, M. K., Kalkum, M., and Versalovic, J. (2019) Microbial metabolic capacity for intestinal folate production and modulation of host folate receptors. *Front. Microbiol.* **10**, 2305
77. Sharma, V., Rodionov, D. A., Leyn, S. A., Tran, D., Iablokov, S. N., Ding, H., Peterson, D. A., Osterman, A. L., and Peterson, S. N. (2019) B-vitamin sharing promotes stability of gut microbial communities. *Front. Microbiol.* **10**, 1485
78. Rossi, M., Amaretti, A., and Raimondi, S. (2011) Folate production by probiotic bacteria. *Nutrients* **3**, 118–134

79. Hale, L. P., Chichlowski, M., Trinh, C. T., and Greer, P. K. (2010) Dietary supplementation with fresh pineapple juice decreases inflammation and colonic neoplasia in IL-10-deficient mice with colitis. *Inflamm. Bowel Dis.* **16**, 2012–2021
80. Fitzhugh, D. J., Shan, S. Q., Dewhirst, M. W., and Hale, L. P. (2008) Bromelain treatment decreases neutrophil migration to sites of inflammation. *Clin. Immunol.* **128**, 66–74
81. de la Motte, C. A. (2011) Hyaluronan in intestinal homeostasis and inflammation: Implications for fibrosis. *Am. J. Physiol. Gastrointest. Liver Physiol.* **301**, G945–G949
82. Chaput, C., Labigne, A., and Boneca, I. G. (2007) Characterization of Helicobacter pylori lytic transglycosylases Slt and MltD. *J. Bacteriol.* **189**, 422
83. Dik, D. A., Marous, D. R., Fisher, J. F., and Mobashery, S. (2017) Lytic transglycosylases: Concinnity in concision of the bacterial cell wall. *Crit. Rev. Biochem. Mol. Biol.* **52**, 503–542
84. Viala, J., Chaput, C., Boneca, I. G., Cardona, A., Girardin, S. E., Moran, A. P., Athman, R., Mémet, S., Huerre, M. R., Coyle, A. J., DiStefano, P. S., Sansonetti, P. J., Labigne, A., Bertin, J., Philpott, D. J., *et al*. (2004) Nod1 responds to peptidoglycan delivered by the Helicobacter pylori cag pathogenicity island. *Nat. Immunol.* **5**, 1166–1174
85. Wyckoff, T. J., Taylor, J. A., and Salama, N. R. (2012) Beyond growth: Novel functions for bacterial cell wall hydrolases. *Trends Microbiol.* **20**, 540–547
86. Wright, G. D. (2005) Bacterial resistance to antibiotics: Enzymatic degradation and modification. *Adv. Drug Deliv. Rev.* **57**, 1451–1470
87. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O., and Piddock, L. J. V. (2015) Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.* **13**, 42–51
88. Million, M., Armstrong, N., Khelaifia, S., Guilhot, E., Richez, M., Lagier, J. C., Dubourg, G., Chabriere, E., and Raoult, D. (2020) The antioxidants glutathione, ascorbic acid, and uric acid maintain butyrate production by human gut clostridia in the presence of oxygen *in vitro*. *Sci. Rep.* **10**, 7705
89. Liu, Y., Patricelli, M. P., and Cravatt, B. F. (1999) Activity-based protein protein profiling: The serine hydrolases. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 14694–14699
90. Jagtap, P., Goslinga, J., Kooren, J. A., McGowan, T., Wroblewski, M. S., Seymour, S. L., and Griffin, T. J. (2013) A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **13**, 1352–1357
91. Silverman, G. A., Bird, P. I., Carrell, R. W., Church, F. C., Coughlin, P. B., Gettins, P. G., Irving, J. A., Lomas, D. A., Luke, C. J., Moyer, R. W., Pemberton, P. A., Remold-O'Donnell, E., Salvesen, G. S., Travis, J., and Whisstock, J. C. (2001) The serpins are an expanding superfamily of structurally similar but functionally diverse proteins. *J. Biol. Chem.* **276**, 33293–33296
92. Kriaa, A., Jablaoui, A., Mkaouar, H., Akermi, N., Maguin, E., and Rhimi, M. (2020) Serine proteases at the cutting edge of IBD: Focus on gastrointestinal inflammation. *FASEB J.* **34**, 7270–7282
93. Uchiyama, K., Naito, Y., Takagi, T., Mizushima, K., Hirai, Y., Hayashi, N., Harusato, A., Inoue, K., Fukumoto, K., Yamada, S., Handa, O., Ishikawa, T., Yagi, N., Kokura, S., and Yoshikawa, T. (2012) Serpin B1 protects colonic epithelial cell *via* blockage of neutrophil elastase activity and its expression is enhanced in patients with ulcerative colitis. *Am. J. Physiol. Gastrointest. Liver Physiol.* **302**, G1163–G1170
94. Kumar, P., Johnson, J. E., Easterly, C., Mehta, S., Sajulga, R., Nunn, B., Jagtap, P. D., and Griffin, T. J. (2020) A sectioning and database enrichment approach for improved peptide spectrum matching in large, genome-guided protein sequence databases. *J. Proteome Res.* **19**, 2722–2785
95. Ma, B., and Johnson, R. (2012) *De novo* sequencing and homology searching. *Mol. Cell. Proteomics* **11**. O111.014902
96. Ma, B. (2015) Novor: Real-time peptide de novo sequencing software. *J. Am. Soc. Mass Spectrom.* **26**, 1885–1894
97. Browne, H. P., Forster, S. C., Anonye, B. O., Kumar, N., Neville, B. A., Stares, M. D., Goulding, D., and Lawley, T. D. (2016) Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543–546
98. Fuks, G., Elgart, M., Amir, A., Zeisel, A., Turnbaugh, P. J., Soen, Y., and Shental, N. (2018) Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* **6**, 17
99. Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., and Weinstock, G. M. (2019) Evaluation of 16S rRNA gene sequences for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029
100. Callahan, B. J., Wong, J., Heiner, C., Oh, S., Theriot, C. M., Gulati, A. S., McGill, S. K., and Dougherty, M. K. (2019) High-throughput amplicon sequencing of the full-length 16S rRNA gene and single-nucleotide resolution. *Nucleic Acids Res.* **47**, e103
101. Tibble, J. A., Sigthorsson, G., Bridger, S., Fagerhol, M. K., and Bjarnason, I. (2000) Surrogate markers of intestinal inflammation are predictive of relapse in patients with inflammatory bowel disease. *Gastroenterology* **119**, 15–22
102. D'Haens, G., Ferrante, M., Vermeire, S., Baert, F., Noman, M., Moortgat, L., Geens, P., Iwens, D., Aerden, I., Van Assche, G., Van Olmen, G., and Rutgeerts, P. (2012) Fecal calprotectin is a surrogate marker for endoscopic lesions in inflammatory bowel disease. *Inflamm. Bowel Dis.* **18**, 2218–2224
103. Jaroszewski, L., Li, Z., Krishna, S. S., Bakolitsa, C., Wooley, J., Deacon, A. M., Wilson, I. A., and Godzik, A. (2009) Exploration of uncharted regions of the protein universe. *PLoS Biol.* **7**, e1000205
104. Baugh, M. D., Perry, M. J., Hollander, A. P., Davies, D. R., Cross, S. S., Lobo, A. J., Taylor, C. J., and Evans, G. S. (1999) Matrix metalloproteinase levels are elevated in inflammatory bowel disease. *Gastroenterology* **117**, 814–822
105. Frank, J. A., Pan, Y., Tooming-Klunderud, A., Eijsink, V. G. H., McHardy, A. C., Nederbragt, A. J., and Pope, P. B. (2016) Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci. Rep.* **6**, 25373
106. Bishara, A., Moss, E. L., Kolmogorov, M., Parada, A. E., Weng, Z., Sidow, A., Dekas, A. E., Batzoglou, S., and Bhatt, A. S. (2018) High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol.* **36**, 1067–1075
107. Moss, E. L., Maghini, D. G., and Bhatt, A. S. (2020) Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* **38**, 701–707
108. Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T. P. L., and Pevzner, P. A. (2020) metaFlye: Scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110
109. Catherman, A. D., Skinner, O. S., and Kelleher, N. L. (2014) Top down proteomics: Facts and perspectives. *Biochem. Biophys. Res. Commun.* **445**, 683–693
110. Cristobal, A., Marino, F., Post, H., van den Toorn, H. W. P., Mohammed, S., and Heck, A. J. R. (2017) Toward an optimized workflow for middle-down proteomics. *Anal. Chem.* **89**, 3318–3325