Shotgun Metaproteomics of the Human Distal Gut Microbiota

N. C. VerBerkmoes[1‡], A. L. Russell[1‡], M. Shah[1], A. Godzik[2], M. Rosenquist[3†], J. Halfvarsson[4], M. G. Lefsrud[1*], J. Apajalahti[5], C. Tysk[4], R. L. Hettich[1], J. K. Jansson[3, 6#]

[1] Oak Ridge National Laboratory, Oak Ridge, TN

[2] Burnham Institute for Medical Research, La Jolla, CA

[3] Swedish University of Agricultural Sciences, Uppsala, Sweden

[4] Örebro University Hospital, Örebro, Sweden

[5] Alimetrics Ltd, Helsinki, Finland

[6] Lawrence Berkeley National Laboratory, Berkeley, CA

*McGill University, Ste-Anne-de-Bellevue, QC, Canada.

†Uppsala University Hospital, Uppsala, Sweden

[‡]These authors contributed equally to this research

#Corresponding Author: jrjansson@lbl.gov

Abstract

The human gut contains a dense, complex, and diverse microbial community, comprising the gut microbiome. Metagenomics has recently revealed the composition of genes in the gut microbiome, but provides no direct information about which genes are expressed or functioning. Therefore, our goal was to develop a novel approach to directly identify microbial proteins in fecal samples to gain information about what genes were expressed and about key microbial functions in the human gut. We used a non-targeted, shotgun mass spectrometry-based whole community proteomics, or metaproteomics, approach for the first deep proteome measurements of thousands of proteins in human fecal samples, thus demonstrating this approach on the most complex sample type to date. The resulting metaproteomes had a skewed distribution relative to the metagenome, with more proteins for translation, energy production, and carbohydrate metabolism compared to what was earlier predicted from metagenomics. Human proteins, including antimicrobial peptides, were also identified, providing a non-targeted glimpse of the host response to the microbiota. Several unknown proteins represented previously undescribed microbial pathways or host immune responses, revealing a novel complex interplay between the human host and its associated microbes.

Introduction

The human gastrointestinal (GI) tract is host for myriads of microorganisms (approximately 10^{11}/gram feces) that carry out vital processes for normal digestive functions of the host and play an important, although not yet not fully understood, role in maturation of human immunity and defense against pathogens. Recent findings suggest that each human has a unique and relatively stable gut microbiota, unless disrupted by external factors such as antibiotic treatment (Jernberg *et al*., 2007). Increasing evidence suggests that the composition of the GI microbiota is linked to inflammatory bowel diseases (Peterson *et al*., 2008), such as Crohn's disease (Dicksved *et al*., 2008), and can even influence the propensity for obesity (Ley *et al*., 2006). Current estimates based on sequencing of 16S rRNA genes in DNA extracted from feces, are that 800-1000 different microbial species and >7000 different strains inhabit the GI tract (Bäckhead *et al*., 2005) and that the majority of these (> 80%) have not yet been isolated or characterized (Eckburg *et al*., 2005). Therefore, there is a vast microbial diversity with largely unknown function that is waiting to be explored.

Recently, metagenomic sequencing has revealed information about the complement of genes in the gut microbiota of two healthy individuals (Gill *et al*., 2006). Although this data set did not represent the entire GI microbiota, analysis of identified genes revealed that the GI microbiome has significantly enriched capacities for glycan, amino acid, and xenobiotic metabolism, methanogenesis, and synthesis of vitamins and isoprenoids. This indirect evidence suggested that there are unique microbial functions carried out in the gut environment.

A major limitation of DNA based approaches is that they predict potential functions, but it is not known if the predicted genes are expressed at all or if so, under what conditions and to what extent. In addition, it is not possible to determine whether the DNA is from active viable cells, dormant inactive cells, or even dead cells. These limitations can be overcome by directly assessing proteins, because the genes must have been transcribed and translated to produce a protein product. However, to date only a couple of microbial proteins have been identified from the human gut and these were obtained by 2 dimensional polyacrylamide gel electrophoresis (2D PAGE) (Klaassens *et al*., 2007), followed by excision and *de novo* sequencing of targeted spots on the gel.

Here, our aim was to develop a novel high throughput, non-targeted mass spectrometry (MS) approach to determine the identities of thousands of microbial proteins in the most complex sample type to date (i.e. feces) and to test the feasibility of using a non-matched metagenome data set for protein identification. This MS-based shotgun proteomics approach relies on detection and identification of all proteins in a lysed cell mixture without the need for gel based separation or *de novo* sequencing. Instead, the resulting peptides from an enzymatic digest of the entire proteome are separated by liquid chromatography and infused directly into rapidly scanning tandem mass spectrometers (2D-LC-MS/MS) via electrospray ionization. The resulting peptide mass information and tandem mass spectra are used to search against protein databases generated from genome sequences. To date, the shotgun metaproteomics approach has only been demonstrated in a limited number of studies and only for microbial communities with low diversity, such as acid mine drainage systems (Ram *et al*., 2005; Lo *et al*., 2007), endosymbionts (Markert *et al*., 2007), and sewage sludge water (Wilmes

et al., 2008). It remains a technical challenge to apply this shotgun approach to more complex microbial communities, such as those inhabiting the human gut.

For this study, it was first necessary to develop the shotgun proteomics approach to work with fecal samples containing large amounts of particulate matter and undigested food and a large diversity of microbial cells. Figure 1 provides an overview of the experimental approach developed. Fecal samples were chosen because sampling is non-invasive and feces have been shown to provide material that is representative of an individual's colonic microbiota (Eckburg *et al*., 2005). Our goal was the qualitative identification of the range and types of proteins that can be confidently and reproducibly measured (i.e. with high specificity and low false positive rates; 1-5% maximum) from gut microorganisms by comparing to available metagenome databases (Gill *et al*., 2006) and available gut isolate genomes and to determine if unmatched data sets could suffice for accurate protein identifications. An additional goal was to apply a novel bioinformatics approach to assign putative functions to unknown proteins not covered by standard analysis of clusters of orthologous groups (COGs). Ultimately, our aim was to use the protein data to provide direct evidence of dominant and key microbial functions in the human gut for the first time, some of which could serve as indicators of a healthy or diseased state. In addition, this non-targeted approach enables identification of human proteins associated with the gut microbiota, thus illustrating potential interactions between the human microbiome and host.

Materials and methods

Fecal sample collection

A female healthy monozygotic twin pair born in 1951 was invited to take part in a larger double blinded study, and details of these individuals with respect to diet, antibiotic usage, etc. are previously described: individuals numbered 6a and 6b (Dicksved *et al*., 2008), that provided Samples 7 and 8, respectively, thus were the focus of this study. The only differences between the individuals according to the submitted questionnaire data were that Individual 6a had gastroenteritis and Individual 6b had taken NSAIDs the last 12 months. Fecal samples were collected in 20 ml colonic tubes by the twins and immediately sent to Örebro University Hospital on the day of collection, where they were placed at –70°C and stored. The Uppsala County Ethics Committee and the ORNL human study review panel approved the study.

Microbial cell extraction from fecal samples

Fecal samples were thawed at +4°C and microbial cells were extracted from the bulk fecal material by differential centrifugation, as previously described (Apajalahti *et al*., 1998). This cell extraction method has previously been found to result in a highly enriched bacterial fraction from complex samples, such as soil and chicken feces, with negligible bacterial cell loss and a good representation of fecal microbiota (Apajalahti *et al*., 1998). The resulting bacterial cell pellets were immediately frozen at –70°C and stored until use.

Cell lysis and protein extraction from cell pellets

The microbial cell pellets (~100 mg) were processed via single tube cell lysis and protein digestion. Briefly, the cell pellet was resuspended in 6M Guanidine/10mM DTT to lyse cells and denature proteins. The guanidine concentration was diluted to 1M with 50 mM Tris buffer/10mM $CaCl_2$ and sequencing grade trypsin (Promega, Madison, WI) was added to digest proteins to peptides. The complex peptide solution was desalted via C18 solid phase extraction, concentrated and filtered (0.45um filter). For each LC-MS/MS analyses below, ~1/4 of the total sample was used.

2D-LC-MS/MS

Both samples were analyzed in technical duplicates via two-dimensional (2D) nano-LC MS/MS system with a split-phase column (RP-SCX-RP) (McDonald *et al*., 2002) on a LTQ Orbitrap (Thermo Fisher Scientific) with 22 h runs per sample (LC as previously described (Ram *et al*., 2005; Lo *et al*., 2007). The Orbitrap settings were as follows: 30K resolution on full scans in Orbitrap, all data-dependent MS/MS in LTQ (top five), 2 microscans for both Full and MS/MS scans, centroid data for all scans and 2 microscans averaged for each spectra, dynamic exclusion set at 1.

Proteome informatics

All MS/MS spectra were searched with the SEQUEST algorithm (Eng *et al*., 1994) and filtered with DTASelect/Contrast (Tabb *et al*., 2002) at the peptide level [Xcorrs of at least 1.8 (+1), 2.5 (+2), 3.5 (+3)]. Only proteins identified with two fully tryptic peptides from a 22 h run were considered for further biological study. Tandem MS/MS spectra

were searched against four databases, the first database (db1) contained two human subject metagenomes (Gill *et al*., 2006), a human database, and common contaminants. The existing metagenome databases (Gill *et al*., 2006) were deficient in *Bacteroides* sequences and since *Bacteroides* are known to be common and abundant in the human intestine (Eckburg *et al*., 2005) were also included *Bacteroides* genome sequences in a second database (metadb), plus other sequences from representatives of the normal gut microbiota deposited and available at the Joint Genome Institute (JGI) IMG database (http//img.jgi.doe.gov/). In addition, we included distracters that one would not commonly expect in the healthy gut. The third and fourth database were made by reversing or randomizing the DB1 and appending it on the end of DB1; these databases were used primarily for determining false positive rates, as previously described (Lo *et al*., 2007; Peng *et al*., 2003). Further descriptions of the databases, searching methods, and false positive rates can be found in supplementary information. All databases, peptide and protein results, MS/MS spectra and supplementary tables for all database searches are archived and made available as open access via the following link: http://compbio.ornl.gov/human_gut_microbial_metaproteome/

All MS .raw files or other extracted formats are available upon request.

Hypothetical Protein Prediction

Hypothetical proteins were submitted to the distant homology recognition server FFAS03 (Jaroszewski *et al*., 2005). The list of hypothetical proteins and predicted functions can be found in Supplementary Table S11. For 80% of the hypothetical proteins, a statistically significant match (Z-score below 9.5) to one of the proteins in the reference databases

was obtained. Functions of the matching proteins were used to assign a provisional function for the hypothetical proteins identified in this study. All the FFAS03 results are available from the FFAS03 server at http://ffas.burnham.org/ffascgi/ cgi/login.pl (Login: Janet_new, password: Janet_new). Links provided on the site can be followed to obtain detailed alignments, three dimensional models and other information.

Results

Metaproteomics of fecal samples

Our results present the first large-scale investigation of the human gut microbial metaproteome. The metaproteomes were obtained from two fecal samples (Samples 7 and 8) collected from two healthy female identical twins (Subjects 6a and 6b, respectively, see Dicksved *et al.* (2008) for a description of the individuals). The shotgun approach used enabled us to identify thousands of proteins by matching peptide mass data to available isolate genome and metagenome sequence databases (Supplementary Table S1). The total number of proteins identified from searching the first database (db1) that contained all predicted human proteins and the gut metagenomes, were 1822 redundant and 1534 non-redundant proteins, with approximately 600 to 900 proteins identified per sample and replicate (Table 1). From the entire non-redundant dataset, 446 proteins (~1/3) matched human proteins while 1368 (~2/3) matched predicted proteins from the microbial metagenome sequence data (Supplementary Table S2 for a complete list).

The second database (metadb) contained all of the sequences in the db1 database above, in addition to sequences from representatives of the normal gut microbiota,

including strains of *Bacteroides*, *Bifidobacteria*, *Clostridia*, and *Lactobacilli*, plus human pathogens and distracters that one would not commonly expect in the healthy gut, such as environmental isolates. The rice (*Oryza Sativa*) genome was included to help identify plant (food) related proteins. From the metadb, the total number of proteins identified were 2911 redundant and 2214 non-redundant; between 970 and 1340 proteins were identified per sample and replicate (Table 1). The categorical breakdown of identified proteins from each major database type and the complete list is shown in Supplementary Table S3. In three out of four runs, the highest percentage of protein identifications corresponded to the bacterial genome sequences that were screened. In the fourth run (i.e. run 2, Sample 8), most protein identifications matched to one of the metagenomes. By contrast, 30-35% of spectra matched to the human protein database, most likely due to a few highly abundant human proteins in the samples with a large number of spectral counts. The proteins matching to both rice and environmental isolate distracters were low, between 2-9%, indicating that the majority of the sequences matched to bacterial types and human sequences that one would expect in the human gut environment.

Among the microbial genomes screened, the highest protein matches were to expected sequences from gut isolates. Of the ~10,000-13,000 total spectra observed from each run, ~2,000 matched *Bacteriodes* or *Bifidobacterium* species, with the *Bacteriodes* species always having slightly more spectra, emphasizing the dominance of these groups and their functional significance in the human distal intestine. This data correlates well with our previously published microbial fingerprint data showing an abundance of *Bacteroides* spp. in both of the individuals studied here (Dicksved *et al*., 2008).

By using established methods of reverse database searching (Lo *et al*., 2007; Peng *et al*., 2003); we estimated a false positive rate at the peptide level of 1-5% for all identified peptides depending on the method. If only those peptides with corresponding high mass accuracy measurements (<10 ppm) were considered (80-85% of all identified peptides per run), then the rate dropped to 0.05-0.23% (See supplementary material for a complete description of false positive rate determinations and associated tables: Supplementary tables S4-8, Supplementary Figures S1-2)

COG categories in the gut metaproteome

The proteins identified from the db1 search were classified into COG categories and when compared between the two samples and the two technical runs, the data were highly reproducible and consistent (Figure 2). By comparison to the average metagenomes previously published from other individuals (Gill *et al*., 2006), we found that several COG categories were more highly represented in the average microbial metaproteomes of the individuals in the present study (Figure 3). The metaproteomes were significantly skewed, with a more uneven distribution of COG categories than those represented in the average metagenomes. The majority of detected proteins were involved in translation, carbohydrate metabolism, or energy production; together representing more than 50% of the total proteins in the metaproteome. In addition, more proteins in the metaproteomes were representative of COG categories for post-translational modifications, protein folding, and turnover. By contrast, other COG categories were under represented in the metaproteomes compared to the metagenomes, including

proteins involved in inorganic ion metabolism, cell wall and membrane biogenesis, cell division and secondary metabolite biosynthesis.

Label free estimation of relative protein abundance by NSAF

We estimated the relative abundances of the thousands of proteins that were detected in each sample by calculating normalized spectral abundance factors (NSAF) (Florens *et al*., 2006; Zybailov *et al*., 2006). The entire list of proteins sorted by averaged NSAF across all samples and technical runs is shown in Supplementary Tables S2-3. By comparing the NSAF data from each sample and technical run to each other, it was clear that the technical runs were highly reproducible for a given sample; R^2 values of 0.77 and 0.85 for Samples 7 and 8 respectively (Supplementary Figs. S3-4).

The most abundant proteins based on this prediction were common abundant human derived digestive proteins such as elastase, chymotrypsin C, and salivary amylases. The most abundant microbial proteins included those for expected processes, such as enzymes involved in glycolysis (e.g. Glyceraldehyde-3-phosphate dehydrogenase). Ribosomal proteins (in particular for *Bifidobacterium*) were also relatively abundant, as were DNA binding proteins, electron transfer flavoproteins, and Chaperonin GroEL/GroES (HP60 family).

The gut microbiomes previously published (Gill *et al*., 2006) were enriched for many COGs representing key genes in the methanogenic pathway, consistent with H_2 removal from the distal gut ecosystem via methanogenesis. By contrast, we found very few proteins represented by methanogens. One example is a hypothetical protein from *Methanobrevibacterium* found in Sample 8. Instead, analysis of the list of proteins based

on the NSAF ranking in our study revealed a high relative abundance of formyltetrahydrofolate synthetase (FTHFS), a key enzyme in the acetyl-CoA pathway of acetogens (Drake *et al*., 2008). Acetogenic bacteria utilize H_2 to reduce CO_2 and form acetate. Although methanogenesis is an important H_2 disposal route in about 30-50% of people in Western countries, in the remainder H_2 is consumed by sulfate reduction or reductive acetogenesis, and this seems to be the situation for the samples we have studied here.

Similar to the published metagenomes that reported several COGs responsible for host-derived fucose utilization that were enriched in the human gut microbiome relative to all microbial genomes (Gill *et al*., 2006), we also found several proteins involved in fucose metabolism, including fucose isomerase and propanediol fermentation (later steps in the pathway). In particular, we detected proteins corresponding to polyhedral bodies that are assumed to protect the cell by sequestering the toxic propionaldehyde intermediate of this pathway (Havemann and Bobik, 2003).

Butyrate kinase was the most highly enriched COG (odds ratio of 9.30) in the previous metagenomic study by Gill *et al*. (2006). This enzyme is the final step in butyrate fermentation. Although we did not identify butyrate kinase, we did find that butyryl-CoA dehydrogenase had a relatively high abundance based on the NSAF analyses. This enzyme catalyzes one of the previous steps in the same pathway; interestingly this protein was strongly expressed in Sample 8 but was not detected in Sample 7. Additional proteins of interest that were relatively abundant included NifU-like homologs and rubrerythrin. The role of NifU has been proposed as a scaffold protein for Fe-S cluster assembly (Ayala-Castro *et al*., 2008). Rubrerythrin is found in anaerobic

sulfate reducing bacteria and is a fusion protein containing an N-terminal iron binding domain and a C-terminal domain homologous to rubredoxin. The physiological role of rubrerythrin has not been identified, but it has been shown to protect against oxidative stress in *D. vulgaris* and other anaerobic microorganisms (Mukhopadhyay *et al*., 2007).

Average NSAF values were compared to determine unique and shared proteins in Samples 7 and 8 (Figure 4, metadb database; Supplementary Figure S5, db1 database). The scatter plot reveals five distinct areas: proteins found in similar abundances in both samples along the diagonal (listed in Supplementary Tables S9-10, 1st tabs), proteins found in only one sample on the respective axis, and two distinct lobes that are overexpressed in one sample or the other but present in both (Figure 4; data for proteins showing significant deviation from central line found in Supplementary Tables S9-10, 2nd tabs). We suggest that the group of approximately equally abundant proteins (747 total) represent core gut populations and functions, supported by the finding that a high proportion of these proteins were from common gut bacteria (i.e. *Bacteroides*, *Bifidobacterium* and *Clostridium*) and represented housekeeping functions: translation (19%), energy production (14%), post-translational modification and protein turnover (12%) and carbohydrate metabolism (16%) (Supplementary Table S10, 1st tab). By contrast, the proteins found in only one sample contained proportionately fewer in COG categories for housekeeping functions and from common gut species, but a higher proportion with unknown functions (28% compared to 11% found in both). These results suggest that the proteins present or over represented in only one sample could represent bacterial populations and functions that change according to environmental influences, such as immediate diet. For example, 33% of the unique proteins only found in Sample

7, are prolamin proteins, i.e. plant storage proteins having a high proline content found in seeds of cereals, suggesting recent ingestion of cereal grains by that individual. Although these individuals did not specify any particular dietary habits in the questionnaire data that accompanied the samples (Dicksved *et al*., 2008), we do not have any detailed information about their specific dietary intake immediately prior to sampling that would enable us to verify this finding.

Analysis of unknown-hypothetical proteins

We performed detailed analyses of the unknown proteins (116 from the published metagenomes (Gill *et al*., 2006) and 89 from bacterial isolate genomes) that could not be classified into COG families. The majority of these proteins belong to novel protein families that are overrepresented in genomes of gut microbes (Figure 5a). Five of the ten most abundant hypothetical proteins in the metaproteome belong to the novel protein family represented by hypothetical protein CAC2564 that was previously identified in human metagenomes (Gill *et al*., 2006), while four out of top ten belong to another novel protein family represented by a hypothetical protein BF3045 from *Bacteroides fragilis*. Members of both families are present in several *Bacteroides*, *Clostridium*, and *Vibrio* species, where they are always associated with each other (see the red and green arrows in Figure 5b) and various metabolic enzymes and transport systems. The neighborhood of these two proteins resembles a typical amino acid metabolic pathway, and we hypothesize that they are involved in amino acid metabolism, most likely cysteine or methionine.

Another interesting example is the CPE0573 family of hypothetical proteins, originally identified in the human gut metagenome (Gill *et al*., 2006). A distant homolog from this family was recently shown to belong to a novel Lacto/galacto-N-biose metabolic pathway, identified in *Bifidobacterium bifidum* (Derensy-Dron *et al*. 1999) and *Bifidobacterium longum* (Nishimoto and Kitaoka, 2007). Other proteins from this pathway were also found in the metaproteome samples, suggesting that it was active in our subjects who apparently ingested lactose in their diet. Additionally, an operon formed by a hypothetical protein BT2437 from *Bacteroides thetaiotaomicron* VPI-5482 was found which codes for a putative lipoprotein (Cheng *et al*., 1999). Proteins from this family are always associated with channel forming 8-stranded beta-barrel proteins from the OprF family (Saint *et al*., 2000) (Figure 4c). The list of hypothetical proteins and predicted functions can be found in Supplementary Table S11.

Identification of human proteins

Almost 30% of all identified proteins were human. The two largest groups of human proteins identified in our study were digestive enzymes and structural cell adhesion and cell-cell interaction proteins. However, the third largest category was comprised of human innate immunity proteins, including antimicrobial peptides, scavenger receptor cysteine-rich (SRCR) proteins (represented by the DMBT1 (deleted in malignant brain tumors) protein), and many other proteins linked to innate immunity and inflammation response (intellectin, resistin, and others). Most of the abundant human proteins were similar in the two individuals, but some differences were found in less abundant proteins (Supplementary Table S9, DB1_differential tab).

We were particularly interested in further investigation of DMBT1 (also called salivary agglutinin and glycoprotein-340) that is predominantly expressed in epithelial cells and secreted to the lumen. This protein has several proposed beneficial functions including tumor suppression, bacterial binding, and anti-inflammatory effects (Ligtenberg *et al*., 2007; Rosensteil *et al*., 2007). Detailed analysis of the distribution of DBMT1 peptides shows that they had fairly uniform distribution along the protein, including hits from all 17 domains present in the DBMT1 protein (Fig 6), suggesting that the DBMT1 protein was present in our samples as a complete, intact protein, that we postulate is indicative of a healthy gut environment.

Discussion

This is the first demonstration of an overall method for obtaining metaproteomics datasets from complex material, in this case human feces, and successful demonstration of the deepest coverage of a complex metaproteome to date. By comparison to previous work on natural environmental samples with only a few dominant species (Ram *et al*., 2005; Lo *et al*., 2005; Wilmes *et al*., 2008), the gut microbiota represents a highly diverse community with thousands of species and strain variants. Therefore, we are testing the technical limit of the use of a shotgun proteomics approach in this study. We were encouraged that the sample extraction and preparation methods worked well for fecal samples. Although there remain experimental and computational challenges, the results presented here indicate that this general approach will be applicable to other complex environments, such as marine and soil microbial communities.

We also successfully demonstrated for the first time that it was feasible to use an unmatched metagenome dataset to obtain valid protein identifications. It is currently more rapid and less expensive to obtain metaproteome data, as we have demonstrated here, than metagenome data. Therefore, this finding is promising for future metaproteomics studies of other environments that do not have matched metagenomics sequence data available.

One of the challenges we addressed was that of estimating protein abundances in these complex samples. Here we used label free methods based on spectral counting and normalized spectral abundance factors (NSAF) (Florens *et al*., 2006; Zybailov *et al*., 2006). NSAF is based on spectral counts but also takes into account protein size and the total number of spectra from a run, thus normalizing the relative protein abundance between samples. Efforts are still underway to develop better tools and statistics for label free methods, such as the absolute protein expression (APEX) method recently developed by Lu *et al.* (2007) that may allow for better statistical comparisons of two data sets. However, the APEX method was derived specifically for isolate data and is not applicable to complex microbial communities because it requires an estimate of the number of expressed proteins in the system and this is not known, for example, in our case.

Although our results present the largest coverage of the human gut microbial metaproteome to date, increasing the dynamic range beyond this initial study will be necessary in the future to more fully understand the function of the human gut microbiota and its interactions with the human host. Based on results from previous studies (Ram *et al*., 2005) and (VerBerkmoes, unpublished results with artificial mixtures) we are

confident that proteins can be detected from populations that represent at least 1% of a mixed community. However, the number of proteins detected (dynamic range) dramatically decreases from 1000s to 100s of proteins for those populations that are present at lower abundances. One possibility to increase the dynamic range of detection would be to enhance the protein separation steps prior to analysis. The trade off for increasing the number of separation steps would be the requirement for a greater amount of starting material and instrument time. Enrichment or depletion techniques could also be attempted to increase the coverage of community members present at low levels, but care must be taken to not effect the proteome during any manipulations. Increasing dynamic range is a clear challenge for all proteomic applications, but particularly so for complex microbial communities such as that found in the human gut, and this will be a pressing area for research and method development in the future.

We made several comparisons of our metaproteome data to the existing metagenome data (Gill *et al*., 2006). Some matches could be made between pathways predicted to be functioning based on abundant genes detected in the metagenome data to abundant proteins we found, such as those involved in fucose and butyrate fermentation. There were also some interesting discrepancies, such as the implication of methanogenesis in the former study and the apparent lack of methanogenesis in the samples we analyzed. Instead, our data suggest that acetogenesis was occurring in our samples, implicating different hydrogen scavenging routes in the subjects in the two studies.

Although about the same percentage of proteins with “unknown function” was found in both the metagenomes and the metaproteomes, the metaproteome data provide direct proof that such proteins are actually expressed. Overall, 67% of hypothetical proteins

identified in this study could be recognized as distant homologs of already characterized families, allowing putative function assignments, with most of them further enriching the amino acid and carbohydrate metabolism categories, but also including proteins involved in cell-cell signaling and active transport of nutrients across bacterial membranes. Also, fold recognition level structure predictions are possible for 55% of them, opening doors for modeling and more detailed function analysis.

There were additional discrepancies between some proteins predicted in the metagenomes that were not detected in the metaproteomes and reasons for this include all or some of the following: 1) the microbial community compositions and proteins produced were different in the different individuals, 2) the proteins were produced, but below the dynamic range of detection, 3) they might not have been expressed at significant levels at the time of sampling, or 4) the proteins may have mutated to a point that they are no longer detected by screening an unmatched metagenome (Denef *et al*., 2007). Therefore, although we successfully identified thousands of proteins using an unmatched dataset, it would still be very valuable to have matching metagenome and metaproteome data from the same samples and this will certainly be achieved via ongoing and future initiatives, such as the NIH Human Microbiome Project (http://nihroadmap. nih.gov/hmp/) and the European Union Meta-HIT project (http://www.international.inra.fr/ press/metahit). Recently, 13 additional human metagenome sequences were published from Japan (Kurokawa *et al*., 2007) and more representative genome sequences from commensal gut isolates are currently being sequenced (Peterson *et al*., 2008). Together these represent valuable resources that should eventually aid in identification of more proteins from the human gut.

A large proportion of the proteins detected in the samples (approximately 30%) were human proteins. This finding can be explained by the method we used to obtain a bacterial cell fraction. Differential centrifugation does not result in a pure bacterial fraction, but instead one that is highly enriched in bacterial cells compared to human cells and particulate matter in the original fecal sample. Any human proteins that adhered to the microbial cells would have been collected in the bacterial pellet. Also there are many more proteins in human cells than in bacterial cells. Therefore, even a minor contamination of the bacterial fraction with human cells could represent a significant number of human proteins. In hindsight this was advantageous because it enabled us to detect and identify human proteins, such as antimicrobial peptides, that reflect interaction between the host and the microbiota. Furthermore, this highlights the power of this technology to distinctly identify both microbial and human proteins in a combined mixture.

In summary, while it is evident that this massive dataset would require substantial effort to completely define and characterize, our goal was to develop an approach to obtain a first large-scale glimpse of the functional activities of the microbial community residing in the human gut. A wealth of information about functional pathways and microbial activities could be gleaned from this data, thereby providing one of the first views into the complex interplay of human and microbial species in the human gut microenvironment. It is clear that proteomics allows us to directly see potential host-commensal bacterial interactions. While the human immune response is usually described in terms of response to infection, it is clear that innate immunity proteins are part of a normal gut environment, shaping the gut microflora to the desired shape.

Finally, we would also like to point out that all data is freely accessible to the scientific community for future analyses and some proteins that we identified can have implications as potential biomarkers for human health.

Acknowledgements

We thank Dr. David Tabb and the Yates Proteomics Laboratory at Scripps Research Institute for DTASelect/Contrast software, the Institute for Systems Biology for proteome bioinformatics tools used in analysis of the MS data, and M. Land of the ORNL Genome Analysis and System Modeling Group for computational resources for proteomic analysis. We thank Patricia Carey (ORNL) for computational assistance with proteome informatics. Becky R. Maggard (ORNL) is thanked for secretarial assistance in the preparation of this manuscript. The ORNL part of this research was sponsored in part by U.S. Department of Energy under Contract DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed and operated by UT-Battelle, LLC. The SLU research was sponsored by the SLU Faculty for Natural Resources and Landscape Management, by the MICPROF grant funded by Uppsala Bio-X (www.uppsalabio.se/) and in part by U. S. Department of Energy Contract DE-AC02-05CH11231 with Lawrence Berkeley National Laboratory. The BIMR research was sponsored in part by the NIH grant P20 GM076221. The human sampling was sponsored by Örebro University Hospital Research Foundation and the Örebro County Research Foundation.

Supplementary information

All databases, datasets, and full supplementary tables can be found at http://compbio.ornl.gov/human_gut_microbial_metaproteome/.

References

Apajalahti JH, *et al*. (1998) Effective recovery of bacterial DNA and percent-guanine-plus-cytosine-based analysis of community structure in the gastrointestinal tract of broiler chickens. *Appl Environ Microbiol* **64**:4084-4088.

Ayala-Castro C, Saini A, Outten, W (2008) F. Fe-S Cluster assembly pathways in bacteria. *Microb Mol Biol Rev* **72**:110-125.

Bäckhead F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host-bacterial mutualism in the human intestine. *Science* **307**:1915-1920.

Chang HJ, Sheu SY, Lo SJ (1999) Expression of foreign antigens on the surface of *Escherichia coli* by fusion to the outer membrane protein traT. *J Biomed Sci* **6**:64-70.

Denef VJ, Shah MB, VerBerkmoes NC, Hettich RL, Banfield JF (2007) Implications of strain- and species-level sequence divergence for community and isolate shotgun proteomic analysis. *J Proteome Res* **6**:3152-3161.

Derensy-Dron D, Krzewinski F, Brassart C, Bouquelet S (1999) β-1,3- Galactosyl-N-acetylhexosamine phosphorylase from *Bifidobacterium bifidum* DSM 20082: characterization, partial purification and relation to mucin degradation. *Biotechnol Appl Biochem* **29**:3–10.

Dicksved J, *et al*. (2008) Molecular analysis of the gut microbiota of identical twins with Crohn's disease. *ISME J* **2**:716-727.

Drake HL, Gössner AS, Daniel SL (2008) Old Acetogens, New Light. *Ann NY Acad Sci* **1125**:100-128.

Eckburg PB, (2005) *et al*. Diversity of the human intestinal microbial flora. *Science* **308**:1635-1638.

Eng JK, McCormack AL, Yates III JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Mass Spectrom* **5**:976-989.

Florens L, *et al*. (2006) Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40**:303-311.

Gill SR, *et al*. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312**:1355-1359.

Havemann GD, Bobik TA (2003) Protein content of polyhedral organelles involved in coenzyme B12-dependent degradation of 1,2-propanediol in *Salmonella enterica* serovar Typhimurium LT2. *J Bacteriol* **185**:5086-5095.

Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. (2005) FFAS03: a server for profile–profile sequence alignments. (Web Server Issue) *Nucleic Acids Res.* **33**:W284-W288.

Jernberg C, Löfmark S, Edlund C, Jansson JK (2007) Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *Internat Soc Microbial Ecol (ISME) J* **1**:56-66.

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

Kurokawa K., *et al*. (2007) Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes. *DNA Res* **14**:169-181.

Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Human gut microbes linked to obesity. *Nature* **444**:1022–1023.

Ligtenberg AJ, Veerman EC, Nieuw Amerongen AV, Mollenhauer J (2007) Salivary agglutinin/glycoprotein-340/DMBT1: a single molecule with variable composition and with different functions in infection, inflammation and cancer. *Biol Chem* **12**:1275-1289.

Lo I, *et al.* (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**:537-541.

Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25**:117-123.

Markert S, *et al*. (2007) Physiological proteomics of the uncultured endosymbiont of *Riftia pachyptila*. *Science* **315:**247-250.

McDonald WH, Ohi R, Miyamoto DT, Mitchison TJ, Yates III JR (2002) Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *Int J Mass Spectrom* **219**:245-251.

Mukhopadhyay A, *et al*. (2007) Cell-wide responses to low-oxygen exposure in Desulfovibrio vulgaris Hildenborough. *J Bacteriol* **189**:5996-6010.

Nishimoto M, Kitaoka M (2007) Identification of N-acetylhexosamine 1-kinase in the complete lacto-N-biose I/galacto-N-biose metabolic pathway in *Bifidobacterium longum*. *Appl Environ Microbiol* **73**:6444-6449.

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

Peterson DA, Frank DN, Pace NR, Gordon JI (2008) Metagenomic approaches for defining the pathogenesis of inflammatory bowel disease. *Cell Host & Microbe* **3**:417-427.

Ram RJ, *et al*. (2005) Community proteomics identifies key activities in a natural microbial biofilm. *Science* **308**:1915-1920.

Rosenstiel P, *et al*. (2007) Regulation of DMBT1 via NOD2 and TLR4 in intestinal epithelial cells modulates bacterial recognition and invasion. *J Immunol* **15**:8203-8211.

Saint N, El Hamel C, De E, Molle G (2000) Ion channel formation by N-terminal domain: a common feature of OprFs of *Pseudomonas* and OmpA of *Escherichia coli*. *FEMS Microbiol Lett* **190**:261-265.

Tabb DL, McDonald WH, Yates III JR (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **1**:21-26.

Wilmes P, *et al*. (2008) Community proteogenomics highlights strain-variant microbial protein expression within activated sludge performing enhanced biological phosphorus removal. Published online, *ISME J* **2**:853-64.

Zybailov B, *et al*. (2006) Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisia*. *J Proteome Res* **5**:2339-2347.

Table 1

Number of protein, peptide, and spectra identifications for Samples 7 and 8 (2 technical runs each) using the db1 and metadb databases (see supplementary material).

db1 database				
Sample ID	Protein identifications*	Peptide identifications	MS/MS Spectra	Peptides between 10 and -10 ppm**
Sample 7, Run 1	634	1886	4069	81.70
Sample 7, Run 2	722	2253	4440	80.42
Sample 8, Run 1	974	3021	5829	83.41
Sample 8, Run 2	983	2948	6131	81.47
metadb database				
Sample 7, Run 1	970	2441	4829	84.47
Sample 7, Run 2	1098	2977	5364	81.67
Sample 8, Run 1	1341	3586	6509	84.71
Sample 8, Run 2	1275	3374	6635	82.92

*Numbers given are non-redundant identifications

** Mass accuracy

Figure Legends:

Figure 1. Shotgun metaproteomics approach used to identify microbial proteins in human fecal samples.

Figure 2. Microbial proteins identified from fecal Samples 7 (blue bars) and 8 (yellow bars) according to COG functions. Bars represent technical proteome runs 1 and 2.

Figure 3. Comparison of average COG categories for available human metagenomes and metaproteomes. (A) Average COG categories of the two *metagenomes* from the gut microbiota of two individuals from a previous study (Gill *et al*., 2006) (B) compared to average COG categories of the *metaproteomes* from the gut microbiota of two individuals in the present study.

Figure 4. Comparison of relative abundances (NSAF values) of proteins detected in Samples 7 and 8. NSAF values for Samples 7 and 8 were averaged amongst their individual technical runs and plotted on a log scale. The dark blue squares represent all of the proteins identified in each sample from screening the metadb database. The straight diagonal line represents the location of all proteins that had approximately equal expression in both samples.

Figure 5. Detailed analysis of hypothetical proteins identified in human gut metaproteome. (A) Protein representation in the genomes of human gut associated microbes; scale changes from 1 (only found in human gut microbes) to -1 (never found there), 0 represents even distribution. Conserved genomic neighborhoods of the

CAC2564 (B) and BT2437 (C) families. Detailed functions of other proteins, identified by numbers in the figure, are provided in the supplementary material.

Figure 6. Positions of DMBT1 peptide fragments along the length of the DMBT1 protein are shown as blue boxes (figure is not to scale). DBMT1 has a length of 1785 amino acids. PFAM domain names: SRCR (Scavenger receptor cysteine-rich domain); CUB (from complement C1r/C1s, Uegf, Bmp1) is a domain found in many in extracellular and plasma membrane-associated proteins; Zona pellucida, a large, cysteine rich domain distantly related to integrins, found in a variety of mosaic eukaryotic glycoproteins, usually acting as receptors.

Figure 1.

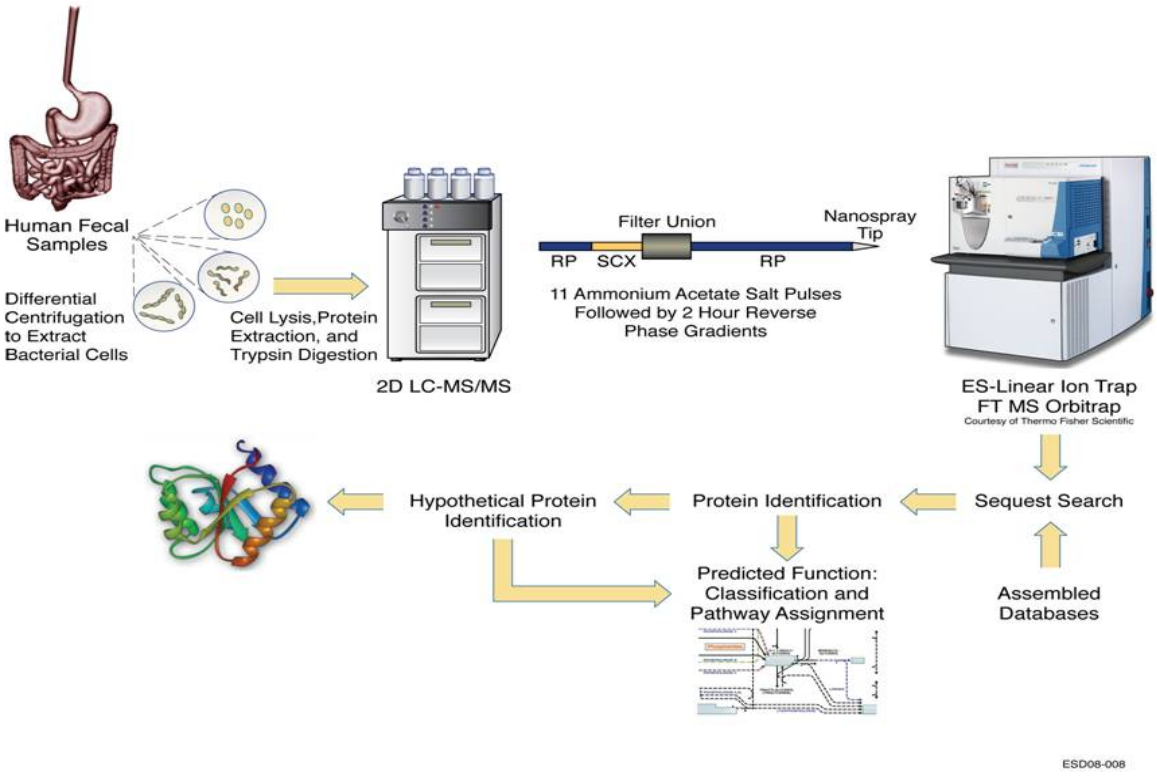

Human Fecal Samples
Differential Centrifugation to Extract Bacterial Cells
Cell Lysis, Protein Extraction, and Trypsin Digestion
2D LC-MS/MS
Filter Union
Nanospray Tip
RP
SCX
RP
11 Ammonium Acetate Salt Pulses Followed by 2 Hour Reverse Phase Gradients
ES-Linear Ion Trap FT MS Orbitrap
Courtesy of Thermo Fisher Scientific
Sequest Search
Assembled Databases
Protein Identification
Predicted Function: Classification and Pathway Assignment
Hypothetical Protein Identification
ESD08-008

Figure 2

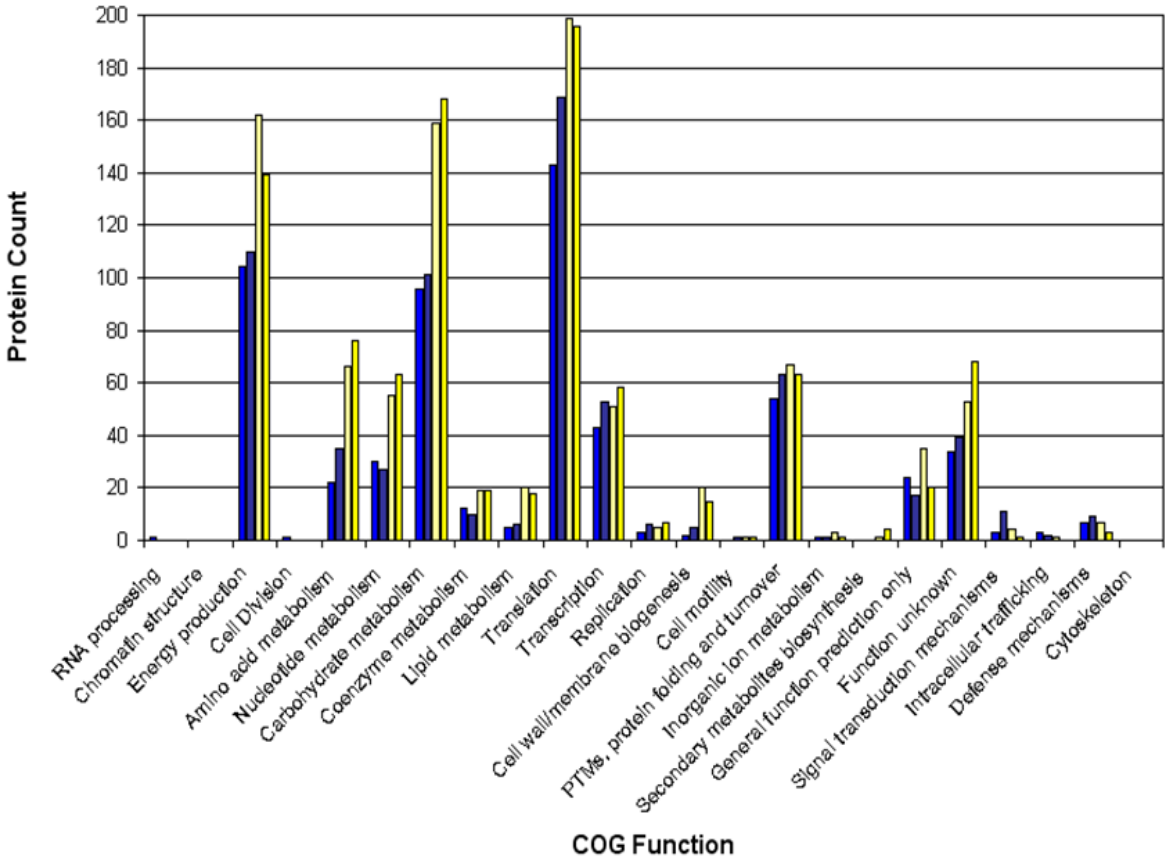

Protein Count
200
180
160
140
120
100
80
60
40
20
0
RNA processing
Chromatin structure
Energy production
Cell Division
Amino acid metabolism
Nucleotide metabolism
Carbohydrate metabolism
Coenzyme metabolism
Lipid metabolism
Translation
Transcription
Replication
Cell motility
Cell wall/membrane biogenesis
PTMs, protein folding and turnover
Inorganic ion metabolism
Secondary metabolites biosynthesis
General function prediction only
Function unknown
Signal transduction mechanisms
Intracellular trafficking
Defense mechanisms
Cytoskeleton
COG Function

Figure 3

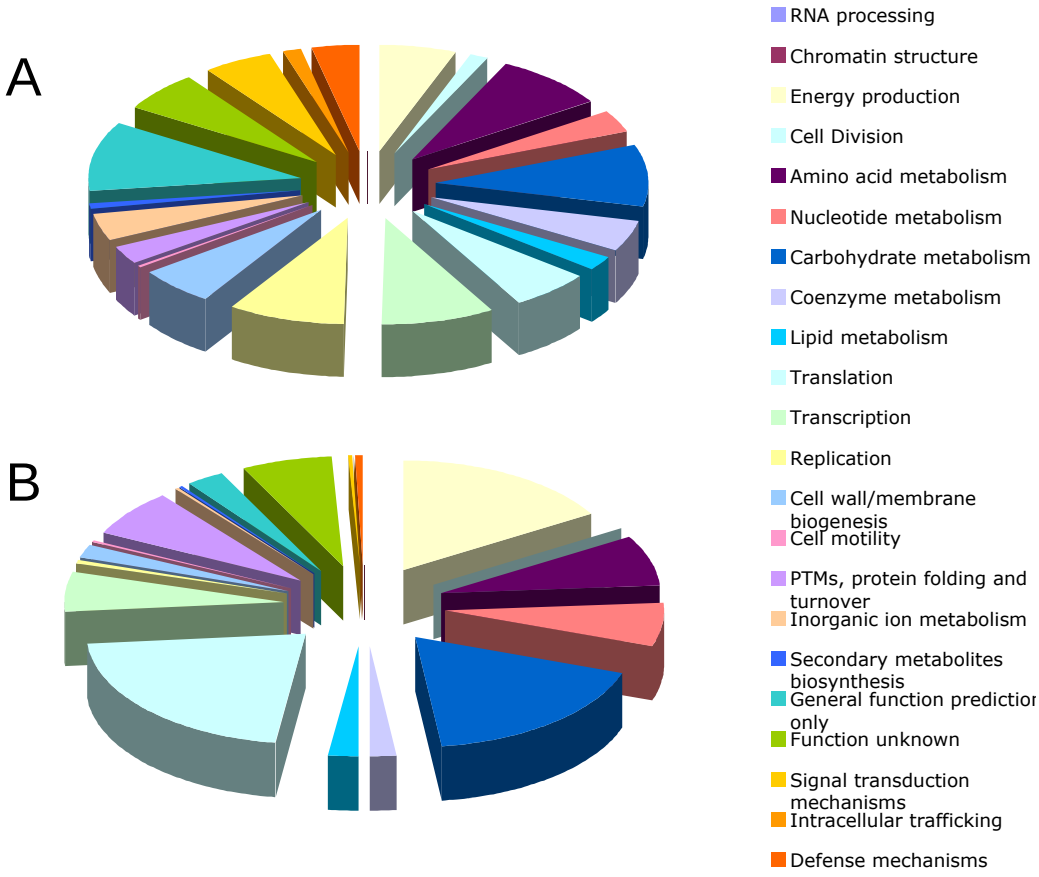

A
B
RNA processing
Chromatin structure
Energy production
Cell Division
Amino acid metabolism
Nucleotide metabolism
Carbohydrate metabolism
Coenzyme metabolism
Lipid metabolism
Translation
Transcription
Replication
Cell wall/membrane biogenesis
Cell motility
PTMs, protein folding and turnover
Inorganic ion metabolism
Secondary metabolites biosynthesis
General function predictior only
Function unknown
Signal transduction mechanisms
Intracellular trafficking
Defense mechanisms

Figure 4

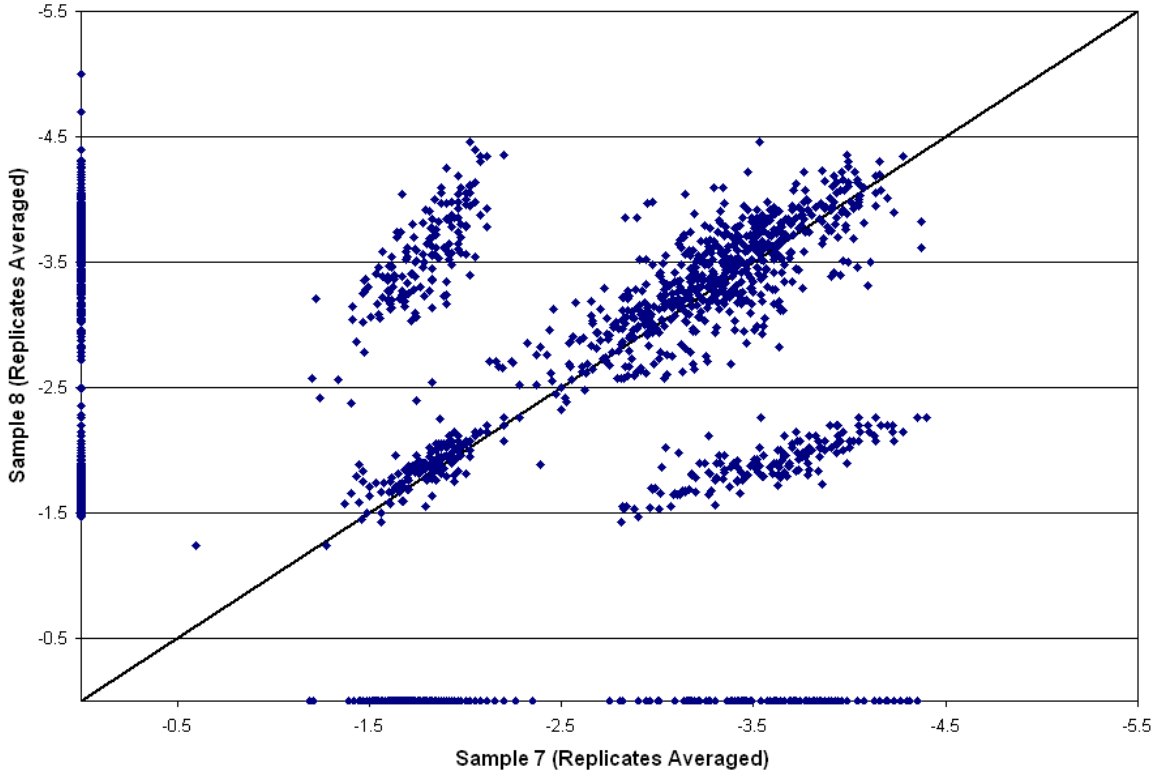
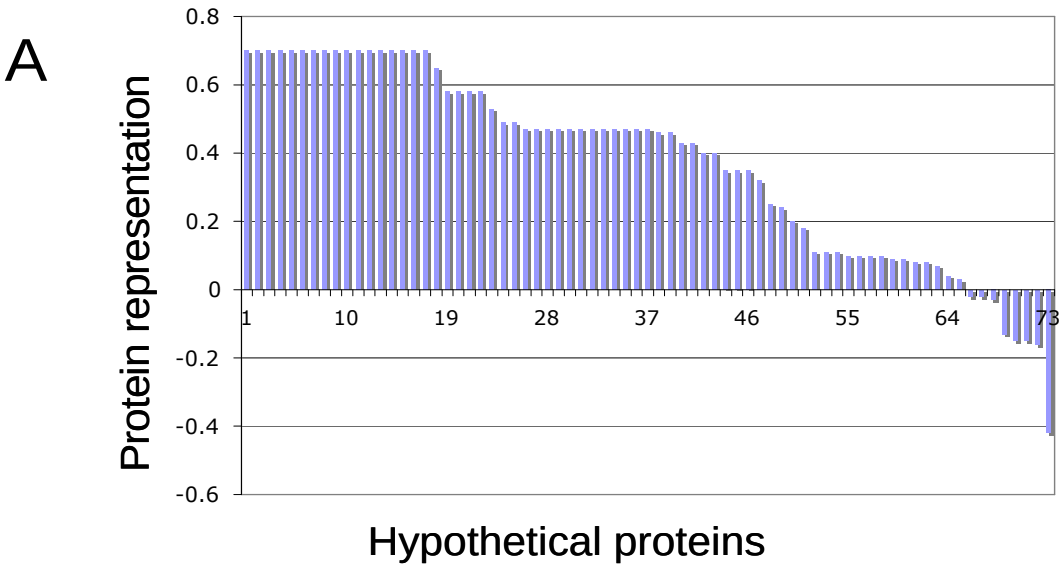
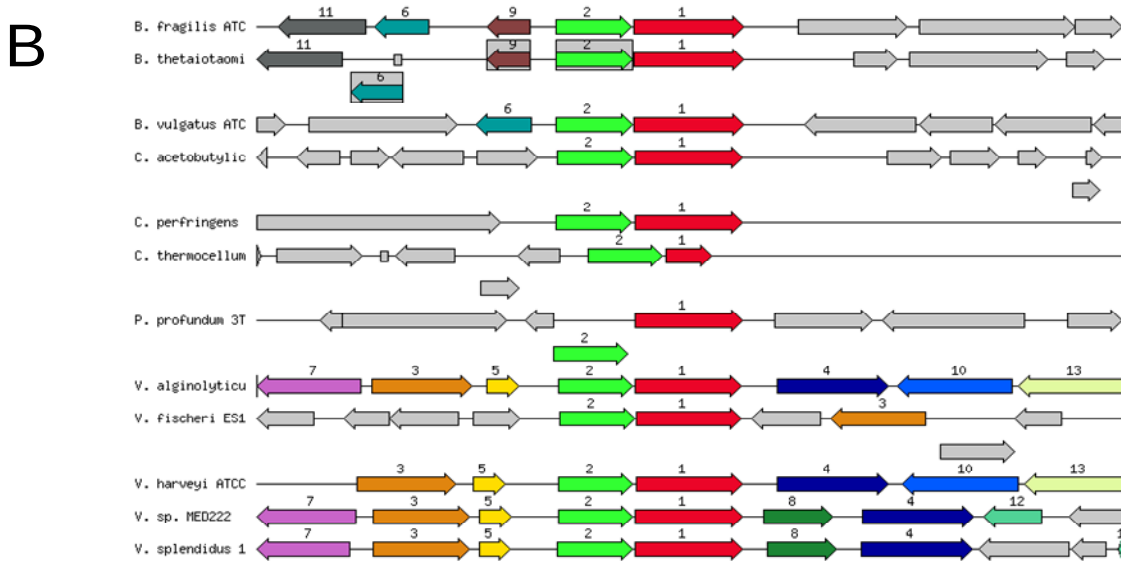

-5.5
-4.5
-3.5
-2.5
-1.5
-0.5
Sample 8 (Replicates Averaged)
-0.5
-1.5
-2.5
-3.5
-4.5
-5.5
Sample 7 (Replicates Averaged)

Figure 5


A
Protein representation
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
1
10
19
28
37
46
55
64
73
Hypothetical proteins


B
B. fragilis ATC
B. thetaiotaomi
B. vulgatus ATC
C. acetobutylic
C. perfringens
C. thermocellum
P. profundum 3T
V. alginolyticu
V. fischeri ES1
V. harveyi ATCC
V. sp. MED222
V. splendidus 1

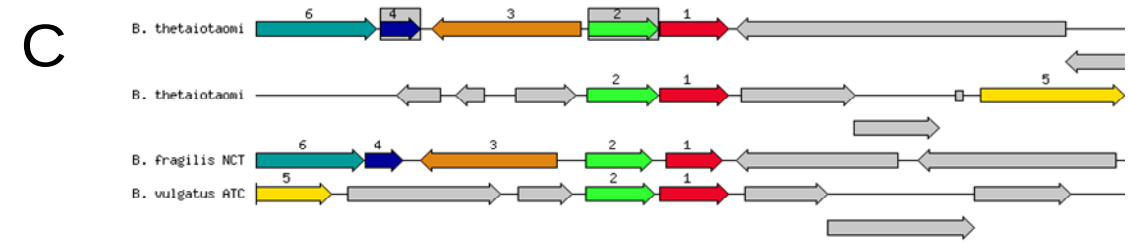
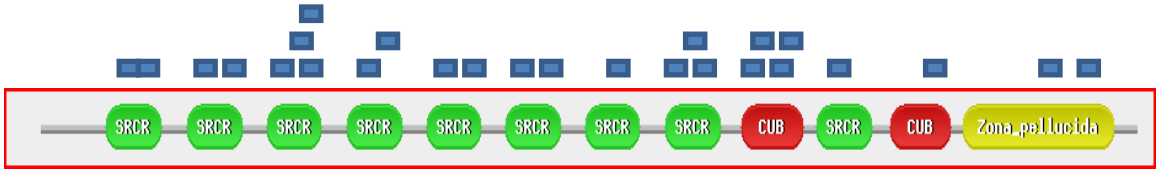

C
B. thetaiotaomi
B. thetaiotaomi
B. fragilis NCT
B. vulgatus ATC

Figure 6


SRCR
SRCR
SRCR
SRCR
SRCR
SRCR
SRCR
SRCR
CUB
SRCR
CUB
Zona_pellucida

Shotgun Metaproteomics of the Human Distal Gut Microbiota

VerBerkmoes et al. **Supplementary Online Information**

All datasets, databases and supplementary data files (spreadsheets in .xls format) can be found at http://compbio.ornl.gov/human_gut_microbial_metaproteome

Proteome informatics

All MS/MS spectra were searched with the SEQUEST algorithm (Eng *et al*., 1994 [(enzyme type, trypsin; Parent Mass Tolerance, 3.0; Fragment Ion Tolerance, 0.5; up to 4 missed cleavages allowed (internal lysine and arginine residues), and fully tryptic peptides only (both ends of the peptide must have arisen from a trypsin specific cut, except N and C-termini of proteins)] and filtered with DTASelect/Contrast (Tabb *et al*., 2002) at the peptide level [Xcorrs of at least 1.8 (+1), 2.5 (+2) 3.5 (+3)]. Only proteins identified with two fully tryptic peptides from a 22 h run were considered for further biological study. Monoisotopic theoretical masses for all peptides identified by SEQUEST were generated and compared to observed masses. Observed high resolution masses were extracted from .raw files from the full scan preceding best identified spectra; parts per million (ppm) calculations were made comparing each identified peptides observed and theoretical mass. When quality MS/MS spectra didn't have an observed mass (low intensity) a mass of 0 was reported and ppm was calculated as infinity.

Four database searches were performed with the above settings. The databases are outlined in Supplemental Table S1. The first database (db1) contained two human subject's metagenomes (Gill *et al*., 2006) a human database, and common contaminants such as trypsin, human keratins, etc. The existing metagenome databases were deficient in *Bacteroides* sequences and since *Bacteroides* are known to be common and abundant in the human intestine (Eckburg *et al*., 2005) we also included *Bacteroides* genome sequences in a second database (metadb), plus other sequences from representatives of the normal gut microbiota deposited and available at the Joint Genome Institute (JGI) IMG database (http//img.jgi.doe.gov/), including representatives of *Bacteroides*, *Bifidobacteria, Clostridia* and *Lactobacilli*. In addition, we included representative human pathogens and included distracters that one would not commonly expect in the gut; i.e. environmental bacterial isolates, plus the rice (*Oryza Sativa*) genome (to help identify food-related proteins). Distracters were uniquely numbered so they could be easily extracted and compared with identifications from proteins thought to be associated with the human microbiome. While this is not actually a false positive rate or false discovery rate as properly defined it gives an indication of how well the method uniquely identifies gut related proteins vs. other protein databases. See discussion below on false discovery rates.

The third and fourth databases were used for estimating false discovery rates on the db1 search as previously described (Lo *et al*., 2007; Peng *et al*., 2003). For the third

database (db3), we took the db1 database and precisely reversed each protein entry (i.e., n-terminus became c-terminus in each case) and then appended these reversed sequences onto the original database. The same was done for the fourth database (db4), but in this case proteins were not reversed but were randomized (Elias *et al*., 2007). Proteins with the reversed or randomized orientations were given a unique identifier for easy extraction. All databases, peptide and protein results, MS/MS spectra and supplementary tables for all database searches are archived and made available as open access via the following link: http://compbio.ornl.gov/human_gut_microbial_metaproteome/
Raw files are available on request.

False positives

Currently, there are many ways of estimating error associated with peptide identifications. Until the field of proteomics comes to a conclusion on the proper way of reporting proteomic data, different versions will exist (Tabb *et al*., 2008). Even the semantics of calling it false discovery rates or false positive rates are under debate. Below we refer to them as false positive rates based on the publications the formulas were derived from. For this large scale study, false-positive rates were used in order to differentiate between true and false peptide identifications rather than false discovery rates (FDR). The overall false-positive rate (FPR) was estimated using the formula: false-positive rate = $2[n_{rev}/(n_{rev} + n_{real})]*100$ where n_{rev} is the number of peptides identified from the reverse database and n_{real} is the number of peptides identified from the real database (Peng *et al*., 2003). A false-positive rate (FPR) was calculated using three different database searches. First, a composite target-decoy database was created with the db1 database. Data analysis consisted of only the forward peptide identification except in the calculation of the FPR where both forward (correct) and reverse (false) identifications were required (Peng *et al*., 2003) (Elias *et al*., 2007). The data was separated based on ppm values that were between +10 and -10 ppm (<±10ppm), values that were able to define a charge state for the peptide but were not between + 10 and -10 ppm (>±10ppm), and values that were unable to properly identify the charge state of the full scan peptide mass spectra (unresolved values). This +10 and -10 ppm division is based on earlier research (Lefsrud *et al*., 2007), with this metaproteome data summarized in Supplementary Figures S1 and S2. The majority of identified peptides have <±10 ppm values, accounting for an average of 83.0% of the total peptides (Table S4). Peptides with >±10 ppm accounted for an average of 8.8% over the data set with 8.2% resulting in unresolved values (Table S4).

The average ppm for the total identified peptides from the human microbial metaproteome data was 33.7 ppm, however when the data was filtered for <±10 ppm this value dropped to -4.1 ppm (Table S5). Also, the average delta amu (atomic mass units) for the total identified peptides for the human microbial metaproteome data was 0.054 amu, but this value dropped to -0.006 amu when filtered for <±10 ppm (Table S5). The FPR for the total identified peptides from the human microbial metaproteome data set was between 3.17% and 1.18% for both samples and duplicate runs (Table S6) when all peptides were considered. However, when the data was filtered for only those peptides with <±10 ppm these values dropped significantly, the highest FPR for filtered data <±10 ppm was 0.21% and the lowest was 0.05%.

A second approach to estimate the error associated with peptide identifications also involved using a composite target-decoy database with db1, except in this case each protein sequences was randomly shuffled creating a "decoy" database. The purpose of this decoy database was to create a more randomized database by shuffling the amino acids of each protein rather than simply reversing the n-terminus and c-terminus. A shuffled database creates more nonsense, thus, reducing the overall chances of making false identifications. Any proteins that are identified with the decoy database indicate that the forward peptide is in fact illegitimate. A FPR can be estimated using a similar formula as previously described. The number of peptides identified from the shuffled database is multiplied by 2 and divided by the sum of all shuffled peptides plus forward peptides identified from the target database. A FPR was estimated for both samples and runs (Table S7) and as described (Elias *et al*., 2007) was similar to the rate determined by the reverse database method.

We also estimated the false discovery rate in the metadb database search by different method. Here we were interested in seeing the number of unique and total peptides identified to known gut isolates, metagenomes, human proteins and rice proteins vs. distracter sequences including the genomes of *Leptospirillum ferrooxidans, Shewanella oneidensis* MR-1, *Rhodopseudomonas palustris* and others. For the entire list of database entries, please visit the website url: http://compbio.ornl.gov/human_gut_microbial_metaproteome/databases/ .

The majority of peptides that matched to the distracter database were in fact non-unique peptides. These shouldn't be counted as false peptides since they overlap with peptides and proteins from isolates, metagenomes etc that could be in the gut. Thus we only counted unique peptides matching the distracter sequences. A FPR was estimated for both samples per run by comparing the number of total unique peptides from the distracter database to both the total unique peptides from the rest of the database and the total peptides from the rest of the database by the same equation given above for the other two methods of determining FPR. When only unique peptides were considered a false positive rate of 3-5% was found. When all peptides were considered then a false positive rate of ~1% was found. These results are very similar to the false positive rates determined for db1 with the reverse and shuffled methods.

Assigning proteins to COG groups

To create Figure 2 in the manuscript the JGI IMG/M database was used (http://imgweb.jgi-psf.org/cgi-bin/m/main.cgi). This database contains COG information for all proteins which we used for COG assignment for bacterial isolates and the human gut metagenome sequences. We found that 37.9% of the proteins could not be assigned to COGs when screening the metagenome databases, similar to 34% estimated for the existing metagenome data deposited at JGI. Most of these proteins were hypothetical or conserved hypotheticals, therefore, they were assigned "S" which is function unknown. On the other hand, several known proteins, such as DNA-directed RNA polymerase, did not have an assigned COG function. In these cases, we assigned them based on our own knowledge to a COG category and if we did not know, we assigned them to "R" which is

general function prediction only. For example, we assigned the DNA-directed RNA polymerase protein to "K", for transcription. Supplemental Table S2 and S3 have COG entries for all detected microbial proteins.

Proteins found in all replicates and runs

We extracted the list of all proteins found in each technical replicate and in both biological samples, i.e. the "conserved proteins". The list for db1 and associated NSAF values for each run can be found in supplemental table 9 first tab (474 proteins total), the list for metadb and associated NSAF values for each run can be found in supplemental table 10 first tab (749 proteins total).

Label free Quantitation methods.

The label free methods rely on intrinsic values obtained in the course of the experiment such as peak intensities or areas of peptides (Old *et al*., 2005), spectral counts (Liu *et al*., 2004) and normalized spectral abundance factors (Florens *et al*., 2006; Zybailov *et al*., 2006) to quantify peptides and thus proteins. They have grown in popularity due to simplicity, cost considerations and the fact they can be used on any sample assuming proper experimental design is implemented. There is strong effort in the proteome informatics community to develop better tools and statistics for label free methods (Zhang *et al*., 2006), (Lu *et al*., 2007). The absolute protein expression (APEX) method recently developed by Lu *et al.* (2007) may allow for a better statistical comparison of two data sets but was derived specifically for isolate data such as *E. coli* and yeast. The APEX method will not be applicable to complex microbial communities because it requires an estimate on the number of proteins being expressed in the system. This is not possible with a complex microbial community from the gut where it's impossible to estimate the number of different cell types, species or total proteins. Thus, we applied simpler methods for protein quantitation based on spectral counts and normalized spectral abundance factors (NSAFs). Unlike spectral counting, NSAF is based on spectral counts but takes into account protein size and the total number of spectra from a run, thus normalizing the relative protein abundance between samples. A spectral abundance factor (SAF) is first calculated by dividing the number total number of spectral counts for each protein by its mass or length. The NSAFs are then calculated by normalizing each SAF to one by dividing by the sum of all SAFs for all protein (Florens *et al*., 2006; Zybailov *et al*., 2006). We first compared the NSAF results from all proteins found in Samples 7 and 8 using db1 and metadb, but limited ourselves to those that were only found in both technical replicates. As can be seen in Figures S3 and S4 the reproducibility of technical runs, based on NSAFs is high with an R^2 of 0.77 for Sample 7 and 0.85 for Sample 8 with the metadb (similar results with db1 not shown). We then averaged the NSAF values for Samples 7 and 8, but left all proteins in the graph to determine what was found uniquely in one sample and not the other. The results from this comparison are found in Figure S5 and suggest some proteins that differed significantly in expression between the two samples. The figure indicates four major clusters; two clusters are located on each extreme where proteins were found only in one individual but not the other. The other two intermediate clusters were found where proteins were present, but expressed in different amounts in comparison with the other individual. We found the cutoff of these clusters to be around a log ratio difference

between 1.1 and 2.4. Thus, we created a sub-table of those proteins showing large differences in expression between the two samples in db1 via those boundaries. We further manually curated the data to only include only those proteins represented by identification in the both runs were they were considered "higher protein abundance" as well as 2x increase in average spectral counts over the other sample (Supplementary Table S9, second tab). In total 225 proteins were found differentially expressed between Sample 7 and Sample 8. This same process was repeated for the metadb dataset. Again there was similar reproducibility in the technical replicates for the two samples (data not shown) and a similar trend in the comparison of Samples 7 and 8 (Figure 4 manuscript). In total, 308 proteins were found differentially expressed between Sample 7 and Sample 8 (for metadb Supplementary Table S10, second tab).

Hypothetical Protein Prediction

Sequences of all hypothetical proteins identified above (116 from the Gill metagenomes and 89 from bacterial isolate genomes), were submitted to the distant homology recognition server FFAS03 (Jaroszewski *et al*., 2005). This server automatically builds a sequence profile for the submitted sequences and compares it against a curated library of sequence profiles, encompassing several sets of annotated proteins (COG, PDB, PFAM and structure determination targets from the JCSG structural genomics center). In independent tests, FFAS03 was shown to consistently outperform PSI-BLAST and other distant recognition algorithms. In Supplementary Table S11 we have summarized results of the analysis. For 80% of the hypothetical proteins a statistically significant match (Z-score below 9.5) to one of the proteins in the reference databases can be obtained. Functions of the matching proteins were used to assign a provisional function for the hypothetical proteins identified in this study. It is important to such analysis can narrow down the possible function of the analyzed protein but, because of the distant homology, detailed function may have diverged from that of the homolog identified in this analysis. More detailed analysis of active site residue conservation and other features is necessary for more detailed function assignment. All the FFAS03 results are available from the FFAS03 server at http://ffas.burnham.org/ffas-cgi/cgi/login.pl (Login: Janet_new, password: Janet_new). Links provided on the site can be followed to obtain detailed alignments, three dimensional models and other information.

Figure S4 (manuscript). Supplementary information:

Genome neighborhood analysis was performed and figures were prepared using the SEED environment for genome annotations as implemented at the National Microbial Pathogen Data Resource project website (http://www.nmpdr.org/cur/FIG/wiki/view.cgi/Main/WebHome). Sequences of BF3046 and BT2437 genes, representing the two families discussed in the text, were compared against the Bacteroides fragilis ATCC genome (NCBI Taxonomy Id: 272559).

B) BF3046 conserved genomic neighborhood

(1)	red	CAC2564 family, as discussed in the text
(2)	green	new family of hypothetical proteins, as discussed in the text
(3)	light brown	LysR family transcriptional regulator
(4)	blue	N-succinyltransferase
(5)	yellow	DNA damage inducible protein
(6)	aquamarine	new family of hypothetical proteins
(7)	violet	telluride resistance protein
(8)	dark green	multiple antibiotic resistance protein
(9)	dark brown	new family of hypothetical proteins
(10)	light blue	Glycerophosphoryl phosphodiesterase

C) BT2437 conserved genomic neighborhood

(1)	red	BT2437 family, as discussed in the text
(2)	green	new family of hypothetical proteins, as discussed in the text
(3)	brown	tripeptidyl aminopeptidase
(4)	blue	MarR family transcriptional regulator
(5)	yellow	Aspartate decarboxylase
(6)	aquamarine	Coenzyme A disulphate reductase

Supplementary Tables

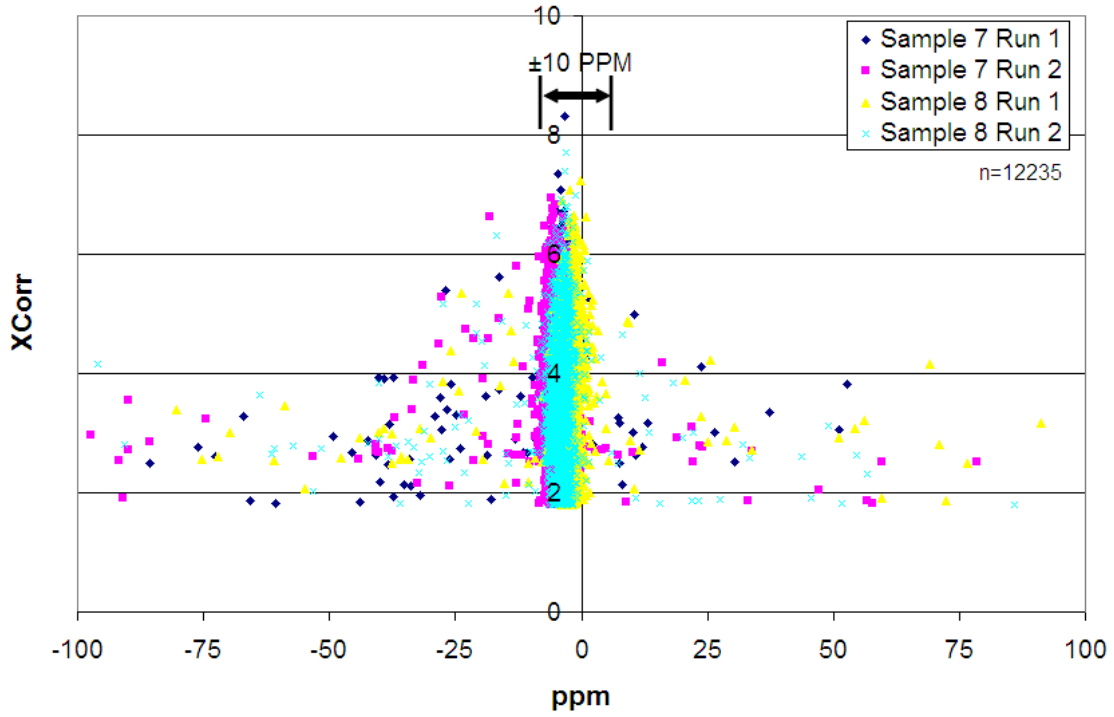
Supplementary Table S1. Description of databases that were screened, see website for complete breakdown.

Database	Sequences included	References
db1	Metagenome, Individual 7 Metagenome, Individual 8 Human proteins	Gill et al. 2006 Gill et al. 2006
metadb	db1 Human commensals and pathogens *Bacteroides* *Bifidobacterium* Etc. Environmental isolates *Leptospirillum* Etc. Rice (*Oryza Sativa*)	JGI/IMG
db3 and db4	db1 in reverse (db3) or random (db4) orientation and appended to db1	

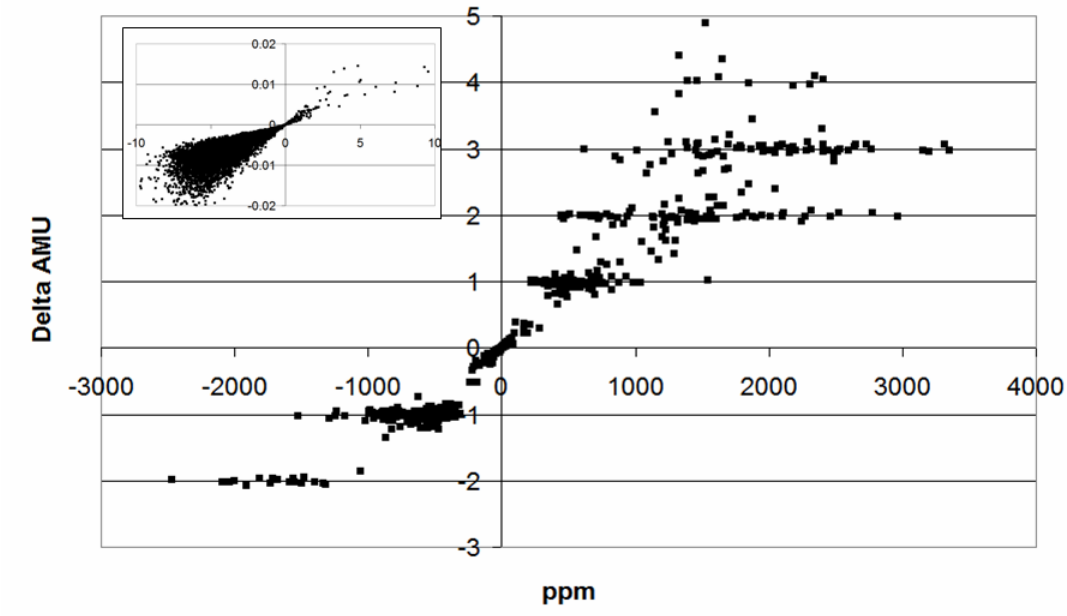
Supplementary Table S2. Protein Identification with NSAF counting from all runs from db1. The results from each individual run can be found on the website.

Supplementary Table S3. Tab 1: Categorical breakdown of identifications to each database type. Tab 2: Protein Identification with NSAF counting from all runs from Metadb. The results from each individual run can be found on the website.

Supplementary Figure S1. PPM Variability Verse XCorr on Forward Peptide Distribution from Human Microbial Metaproteome Data (db1).



Supplementary Figure S2. Delta Atomic Mass Units Verse PPM on Forward Peptide Distribution from Human Microbial Metaproteome Data.



Supplementary Table S4. Identified Forward Peptides from db1 database search.

	Forward Identified Peptides			% of Forward Identified Peptides		
Sample ID	**Values <±10ppm**	**Values >±10ppm**	**Unresolved values**	**Values <±10ppm**	**Values >±10ppm**	**Unresolved values**
Sample 7 Run1	2316	279	195	83.0%	10.0%	7.0%
Sample 7 Run2	2592	275	330	81.1%	8.6%	10.3%
Sample 8 Run1	3664	357	309	84.6%	8.2%	7.1%
Sample 8 Run2	3397	355	350	82.8%	8.7%	8.5%

Supplementary Table S5. Average Delta AMU and Average ppm.

	Average Delta AMU		**Average ppm**	
Sample ID	**Values excluding unresolved values**	**Values <±10ppm**	**Values excluding unresolved values**	**Values <±10ppm**
Sample 7 Run1	0.065	-0.007	46.9	-4.4
Sample 7 Run2	0.041	-0.003	20.5	-2.2
Sample 8 Run1	0.068	-0.008	43.6	-5.2
Sample 8 Run2	0.040	-0.006	24.2	-4.2

Supplementary Table S6. Total Identified Peptides, Identified Reverse Peptides and False Positive Rate from db1 determined by reverse database method (db3).

Sample ID	Total Identified Forward Peptides	Identified Reverse Peptides			False Positive Rate	
		Values <±10ppm	Values >±10ppm	Unresolved values	Total False Positive	False Positive <±10ppm
Sample 7 Run1	2790	3	40	2	3.17%	0.21%
Sample 7 Run2	3197	1	14	4	1.18%	0.06%
Sample 8 Run1	4330	4	37	11	2.37%	0.18%
Sample 8 Run2	4102	1	32	7	1.93%	0.05%

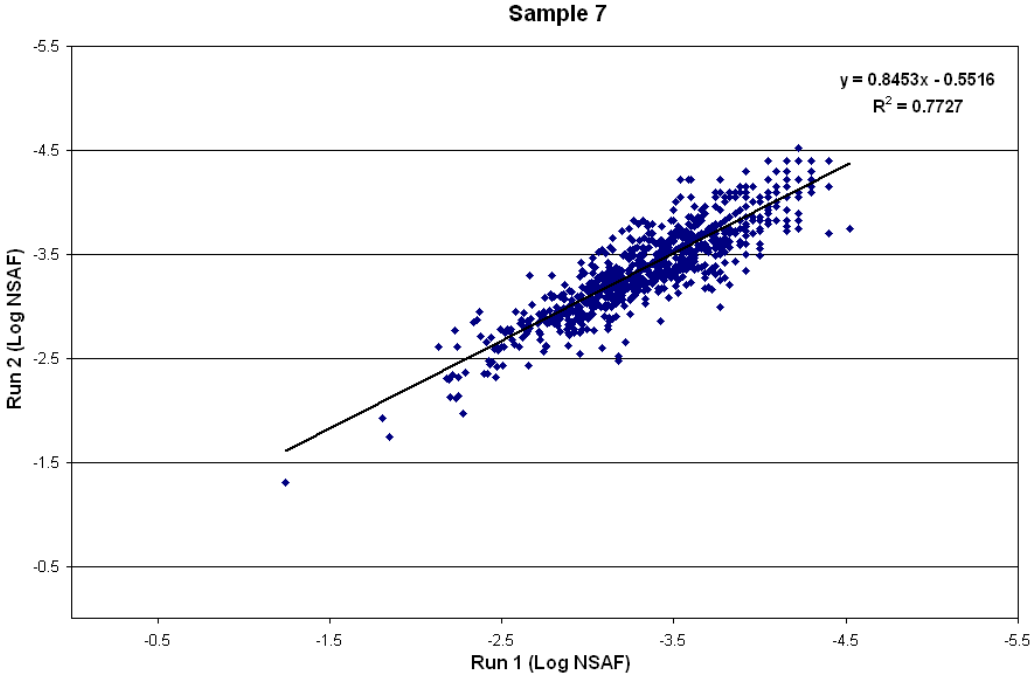
Supplementary Table S7. Total Identified Peptides, Identified Shuffled Peptides and False Positive Rate from Human Microbial Metaproteome Data (db1) determined by random database method (db4)

Sample ID	Total Identified Forward Peptides	Identified Shuffled Peptides			False Positive Rate	
		Values <±10ppm	Values >±10ppm	Unresolved values	Total False Positive	False Positive <±10ppm
Sample 7 Run1	2789	3	31	1	2.48%	0.21%
Sample 7 Run2	3279	3	51	3	3.42%	0.18%
Sample 8 Run1	4324	0	42	6	2.20%	0.00%
Sample 8 Run2	4230	5	40	3	2.24%	0.23%

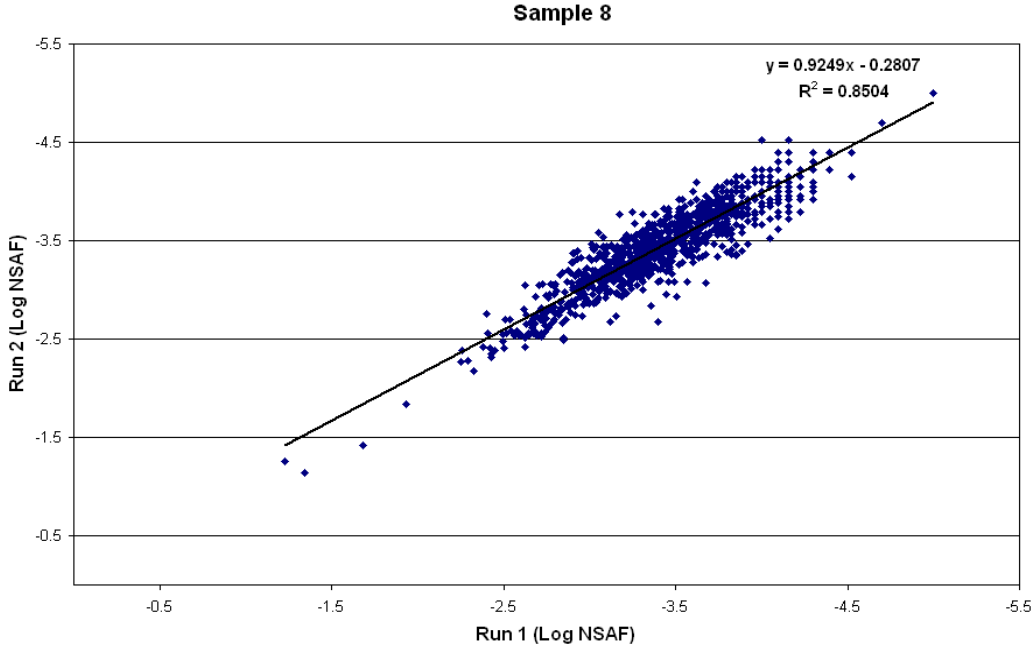
Supplementary Table S8. Total Identified Distracter Peptides, Identified Gut Peptides and False Positive Rate from Human Microbial Metaproteome Data (metadb)

Sample ID	**Unique Distracter Peptides**	**Non-Unique Distracter Peptides**	**Unique Gut Peptides**	**Total Gut Peptides**	**FPR (%) <u>unique</u> distracter & <u>unique</u> gut peptides**	**FPR (%) <u>unique</u> distracter & <u>total</u> gut peptides**
Sample 7 Run 1	30	272	1135	5080	5.15	1.17
Sample 7 Run 2	31	184	1436	6036	4.23	1.02
Sample 8 Run 1	31	205	1899	7115	3.21	0.87
Sample 8 Run 2	33	151	1808	6511	3.59	1.01

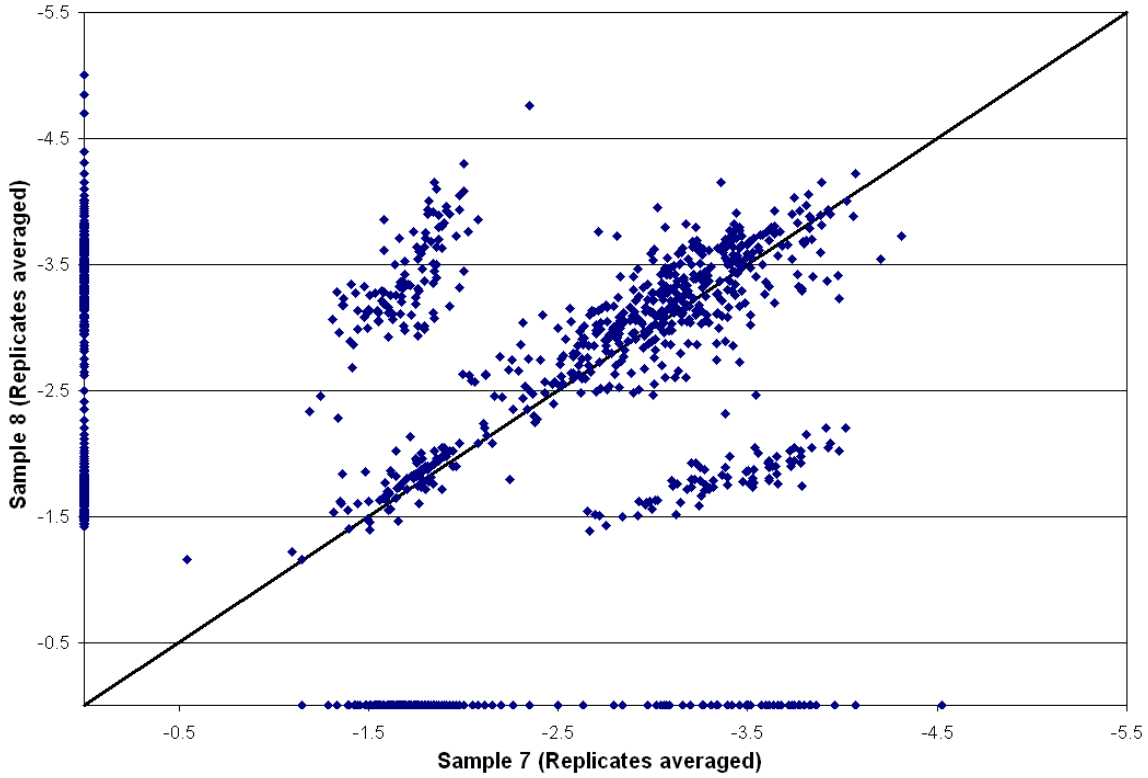
Supplementary Figure S3. Comparison of NSAF values. Sample 7, run 1 and run 2 NSAF values are plotted on a log scale. The dark blue squares represent all of the proteins that were identified in both runs from metadb.



Supplementary Figure S4. Comparison of NSAF values. Sample 8, run 1 and run 2 NSAF values are plotted on a log scale. The dark blue squares represent all of the proteins that were identified in both runs from metadb.



Supplementary Figure S5. Comparison of NSAF values for Samples 7 and 8. NSAF values were averaged amongst two individual technical runs pre sample and plotted on a log scale. The dark blue squares represent all of the proteins identified in each sample from db1. The straight diagonal line is for visualizing the location of all proteins that had approximately equal expression in both samples



Supplementary Table S9. Tab one Proteins found in both samples and replicates with db1. Tab two proteins showing abundance differences based on NSAF calculations Samples 7 and 8 with db1.

Supplementary Table S10. Tab one Proteins found in both samples and replicates with metadb. Tab two proteins showing abundance differences based on NSAF calculations Samples 7 and 8 with metadb.

Supplementary Table S11. All identified hypothetical proteins and predicted functions. Column B is original predicted function, column C is the new computational predicted function. More detailed listing can be found on website.

Supplementary References

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

Lefsrud MG, *et al*., (2007) Proteogenomics reveals Key Insight into the Microbial Activities of Enhanced Biological Phosphorus removal in Activated Sludge. ASMS Conference Proceedings, Indianapolis, IN. June 3-7, 2007.

Liu H, Sadygov RG, Yates III JR, (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76**:4193-4201.

Old WM, *et al*. (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* **4**:1487-1502.

Tabb DL (2008) What's driving false discovery rates? *J Proteome Res* **7**:45-46.

Zhang B. *et al*., (2006) Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* **5:** 2909-2918.