The ISME Journal (2009) 3, 179–189

www.nature.com/ismej

ORIGINAL ARTICLE

Shotgun metaproteomics of the human distal gut microbiota

Nathan C Verberkmoes[1,7], Alison L Russell[1,7], Manesh Shah[1], Adam Godzik[2], Magnus Rosenquist[3,8], Jonas Halfvarson[4], Mark G Lefsrud[1,9], Juha Apajalahti[5], Curt Tysk[4], Robert L Hettich[1] and Janet K Jansson[3,6]

[1]*Oak Ridge National Laboratory, Chemical and Life Sciences Divisions, Oak Ridge, TN, USA;* [2]*Department of Bioinformatics and Systems Biology, Burnham Institute for Medical Research, La Jolla, CA, USA;* [3]*Department of Microbiology, Swedish University of Agricultural Sciences, Uppsala, Sweden;* [4]*Department of Internal Medicine, Division of Gastroenterology, Örebro University Hospital, Örebro, Sweden;* [5]*Alimetrics Ltd, Helsinki, Finland and* [6]*Department of Ecology, Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*

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 the genes 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 when 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.

The ISME Journal (2009) **3,** 179–189; doi:10.1038/ismej.2008.108; published online 30 October 2008
Subject Category: microbe–microbe and microbe–host interactions
Keywords: microbiome; human gut; metaproteome; shotgun proteomics; metagenomics; antimicrobial peptide

Introduction

The human gastrointestinal (GI) tract is host for myriads of microorganisms (approximately 10^{11} per g 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 dataset did not represent the

Correspondence: JK Jansson, Earth Science Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA.
E-mail: jrjansson@lbl.gov
[7]These authors contributed equally to this work.
[8]Current address: Department of Oncology, Radiology and Clinical Immunology, Uppsala University Hospital, Uppsala, Sweden.
[9]Current address: Bioresource Engineering, McGill University, Ste-Anne-de-Bellevue, Quebec, Canada.
Received 2 September 2008; revised 7 October 2008; accepted 8 October 2008; published online 30 October 2008

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 whether 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 cells that are active and viable, dormant or even dead. 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 two-dimensional (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 (that is, feces) and to test the feasibility of using a non-matched metagenome dataset 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) through 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 (that is, with high specificity and low false-positive rates; 1–5% maximum) from gut microorganisms by comparing with available metagenome databases (Gill *et al.*, 2006) and available gut isolate genomes and to determine whether unmatched datasets 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, and so on are described earlier: individuals numbered 6a and 6b (Dicksved *et al.*, 2008), who provided samples 7 and 8, respectively, 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 non-steroidal anti-inflammatory drugs during the past 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 Oak Ridge National Laboratory (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 described earlier (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 wet weight) were processed through single tube cell lysis and protein digestion. Briefly, the cell pellet was

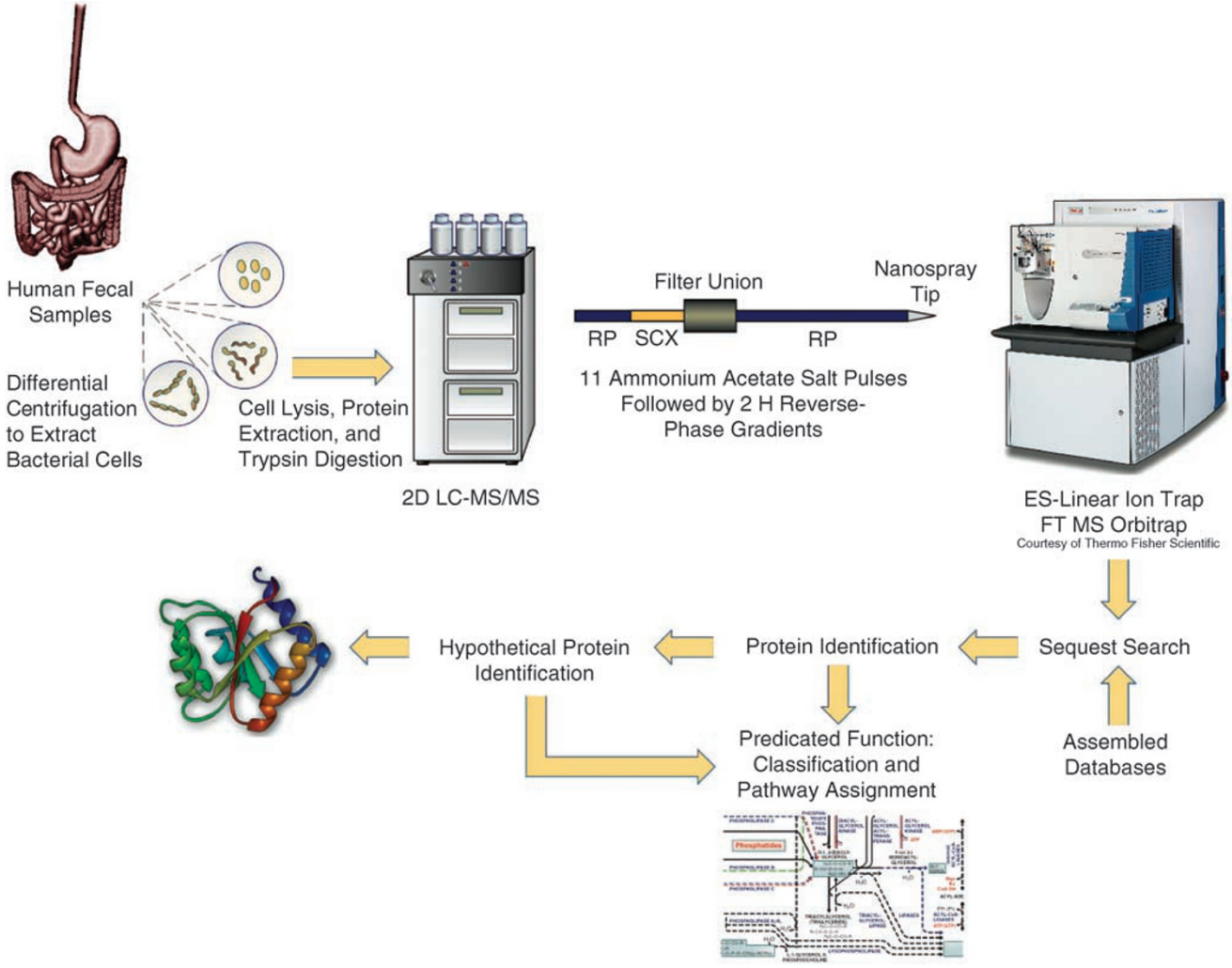

Figure 1 Shotgun metaproteomics approach used to identify microbial proteins in human fecal samples.

resuspended in 6 M guanidine/10 mM dithiothreitol to lyse cells and denature proteins. The guanidine concentration was diluted to 1 M with 50 mM Tris buffer/10 mM $CaCl_2$ and sequencing grade trypsin (Promega, Madison, WI, USA) was added to digest proteins to peptides. The complex peptide solution was desalted through C18 solid-phase extraction, concentrated and filtered (0.45 μm filter). For each LC-MS/MS analysis below, ~1/4 of the total sample was used.

2D-LC-MS/MS

Both samples were analyzed in technical duplicates through a 2D nano-LC MS/MS system with a split-phase column (RP-SCX-RP) (McDonald *et al.*, 2002) on a LTQ Orbitrap (ThermoFisher Scientific, San Jose, CA, USA) with 22 h runs per sample (LC as described earlier; 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), two microscans for both full and MS/MS scans, centroid data for all scans and two microscans averaged for each spectrum, 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 as *Bacteroides* are known to be common and abundant in the human intestine (Eckburg *et al.*, 2005), *Bacteroides* genome sequences were also included 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/cgi-bin/pub/main.cgi). 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 described earlier (Peng *et al*., 2003; Lo *et al*., 2007). 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 information for all database searches are archived and made available as open access through the following link: http://compbio.ornl.gov/human_gut_microbial_metaproteome/.

All MS raw files or other extracted formats are available on 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/gut.metaproteome (login: Janet_new, password: Janet_new). Links provided on the site can be followed to obtain detailed alignments, 3D 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 600–900 proteins identified per sample and replicate (Table 1). From the entire non-redundant dataset, ~1/3 matched human proteins, ~2/3 matched predicted proteins from the microbial metagenome sequence data (see 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 are 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 (that is, 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

Table 1 Number of protein, peptide and spectra identifications for samples 7 and 8 (two technical runs each) using the db1 and metadb databases (see Supplementary information)

Sample ID	*Protein identifications*[a]	*Peptide identifications*	*MS/MS spectra*	*Peptides between 10 and −10 p.p.m.*[b]
db1 database				
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

[a]Numbers given are non-redundant identifications.
[b]Mass accuracy.

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 and 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, ~2000 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. These data correlate 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 (Peng *et al.*, 2003; Lo *et al.*, 2007), 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 p.p.m.) were considered (80–85% of all identified peptides per run), then the rate dropped to 0.05–0.23% (see Supplementary information for a complete description of false-positive rate determinations and associated tables: Supplementary Tables S4–S8, Supplementary Figures S1 and S2).

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

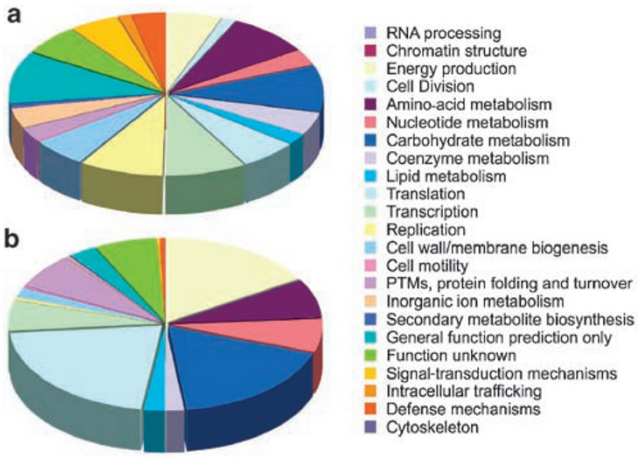

Figure 3 Comparison of average clusters of orthologous group (COG) categories for available human metagenomes and metaproteomes. (**a**) Average COG categories of the two *metagenomes* from the gut microbiota of two individuals from a previous study (Gill *et al.*, 2006), (**b**) compared to average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

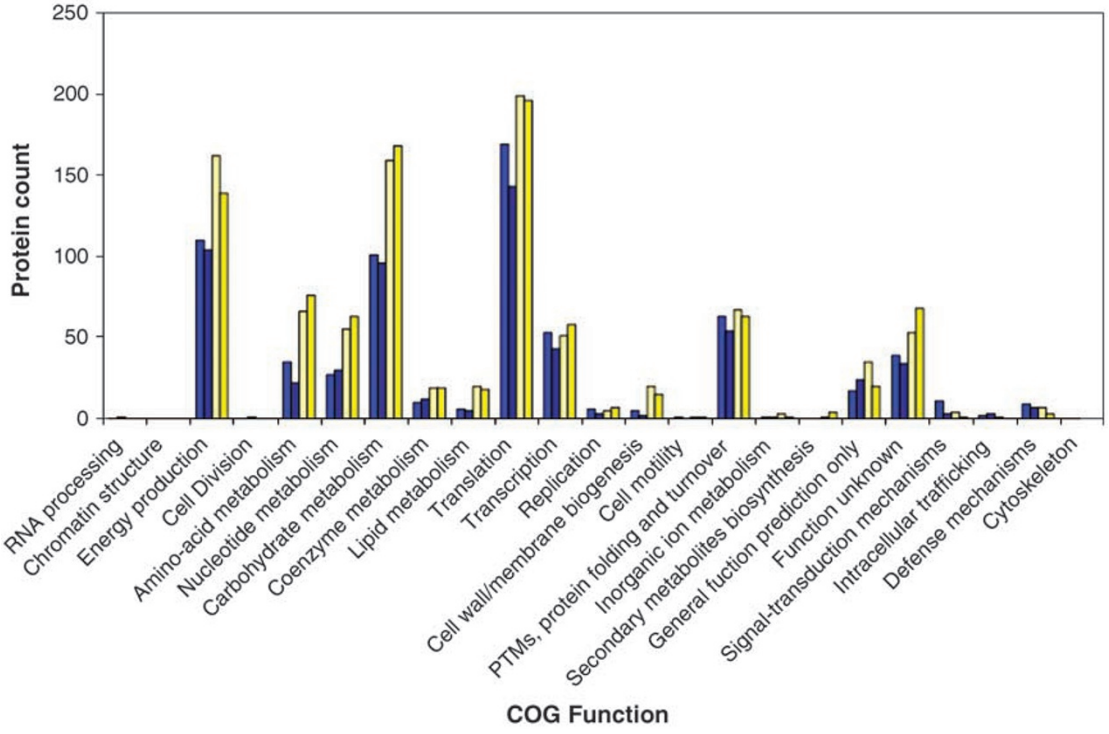

Figure 2 Microbial proteins identified from fecal samples 7 (blue bars) and 8 (yellow bars) according to clusters of orthologous group (COG) functions. Bars represent technical proteome runs 1 and 2.

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 when compared with 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 normalized spectral abundance factor

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 and S3. By comparing the NSAF data from each sample and technical run with 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 Figures S3 and S4).

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 (for example, 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 through 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, 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 finding of COGs responsible for host-derived fucose utilization that were enriched in the human gut microbiome (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 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 *Desulfovibrio vulgaris* and other anaerobic microorganisms (Mukhopadhyay *et al.*, 2007).

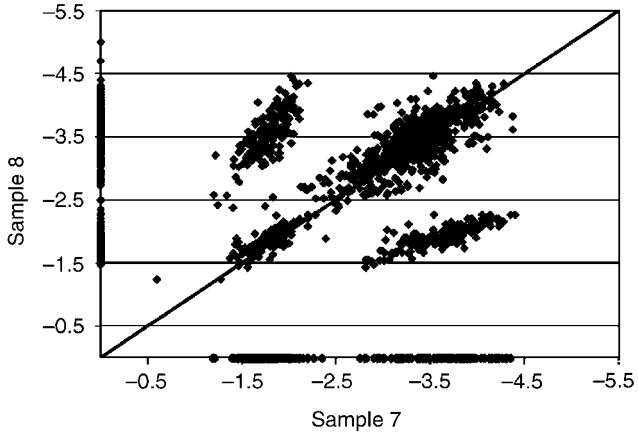

Figure 4 Comparison of relative abundances (normalized spectral abundance factor (NSAF) values) of proteins detected in samples 7 and 8. NSAF values for samples 7 and 8 were averaged among their individual technical runs and plotted on a log scale. The square black symbols represent all of the proteins identified in each sample from screening the metadb database. The diagonal line represents the location of all proteins that had approximately equal expression in both samples.

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 and S10, first 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 and S10, second 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 (*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, first tab). By contrast, the proteins found in only one sample contained proportionately fewer COG categories for housekeeping functions and from common gut species, but a higher proportion with unknown functions (28% compared with 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 found only in sample 7 are prolamin proteins, that is, 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 belonged to novel protein families that are over-represented 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, identified earlier in human metagenomes (Gill *et al.*, 2006), whereas four out of the top ten belong to another novel protein family represented by a hypothetical protein BF3045 from *Bacteroides fragilis*. Members of both families are

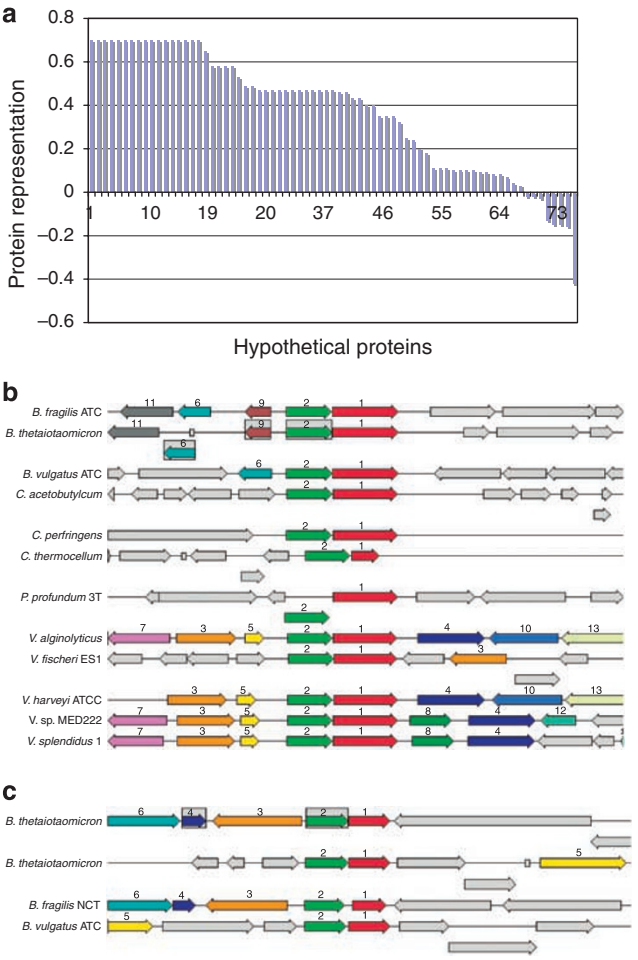

Figure 5 Detailed analysis of hypothetical proteins identified in human gut metaproteome. (**a**) Protein representation in the genomes of human gut-associated microbes; scale changes from 1 (found only in human gut microbes) to −1 (never found there), 0 represents even distribution. Conserved genomic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

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 that codes for a putative lipoprotein (Chang *et al.*, 1999). Proteins from this family are always associated with channel-forming eight-stranded beta-barrel proteins from the OprF family (Saint *et al.*, 2000) (Figure 5c). 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 into the lumen. This protein has several proposed beneficial functions, including tumor suppression, bacterial binding and anti-inflammatory effects (Ligtenberg *et al.*, 2007; Rosenstiel *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 (Figure 6), suggesting that the DBMT1 protein was present in our samples as a complete, intact protein, which we postulate is indicative of a healthy gut environment.

Discussion

This is the first demonstration of an approach 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 with earlier work on environmental samples with only a few dominant species (Ram *et al.*, 2005; Lo *et al.*, 2007; Wilmes *et al.*, 2008), the gut microbiota represents a highly diverse community with thousands of species. Therefore, we tested the technical limit of the use of the shotgun proteomics approach. We were encouraged that the sample extraction and preparation methods worked well for fecal samples. Although there remain experimental and computational challenges, this general approach should be applicable to other complex environments, such as marine and soil microbial communities.

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

One particular challenge is to estimate protein abundances in complex samples. Here, we used label-free methods based on spectral counting and NSAFs (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 underway to develop better tools for label-free methods, such as the absolute protein expression (APEX) method recently developed by Lu *et al.* (2007). However, the APEX method was derived specifically for isolate data and is not directly applicable to complex 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. Previous studies (Ram *et al.*, 2005) and current work (NCV, unpublished results with artificial mixtures)

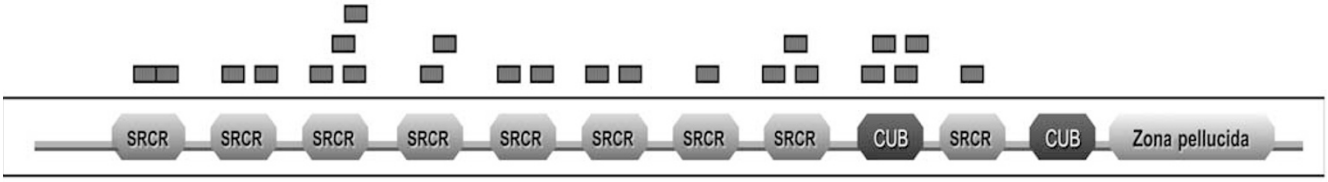

Figure 6 Positions of DMBT1 peptide fragments along the length of the DMBT1 protein are shown as grey boxes (figure is not to scale). DBMT1 has a length of 1785 amino acids. PFAM domain names: SRCR (scavenger receptor cysteine-rich domain); CUB (from complement C1r/C1s, Uegf, Bmp1) is a domain found 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.

suggest that proteins can be detected from populations representing at least 1% of the community. However, the number of proteins detected (dynamic range) dramatically decreases from thousands to hundreds 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 affect the proteome during any manipulations. Increasing the dynamic range is a clear challenge for all proteomic applications 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. The few, low-level, non-unique peptide hits to methanogens that we found were not sufficient to indicate that these organisms were present or functioning. 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 through 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). Taken together, these represent valuable resources that should eventually aid in the 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 differential centrifugation method that we used to obtain a bacterial cell fraction, which is not pure but highly enriched in bacterial cells when compared to human cells and particulate matter in the original fecal sample. Any human protein 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, although 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. Although 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 are 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 the 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 paper. The ORNL part of this research was sponsored in part by US 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 US Department of Energy contract DE-AC02-05CH11231 with Lawrence Berkeley National Laboratory. The Burnham Institute for Medical Research (BIMR) 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. Alison L Russell was funded by the Genome Science and Technology Program at the University of Tennessee and by Oak Ridge National Laboratory.

References

Apajalahti JH, Särkilahti LK, Mäki BR, Heikkinen JP, Nurminen PH, Holben WE. (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 FW. (2008). 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, Halfvarson J, Rosenquist M, Järnerot G, Tysk C, Apajalahti 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, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M *et al.* (2005). 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, Carozza MJ, Swanson SK, Fournier M, Coleman MK, Workman JL *et al.* (2006). Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40**: 303–311.

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS *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. *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, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A *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, Denef VJ, Verberkmoes NC, Shah MB, Goltsman D, DiBartolo G *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, Arndt C, Felbeck H, Becher D, Sievert SM, Hügler M *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, Redding AM, Joachimiak MP, Arkin AP, Borglin SE, Dehal PS *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, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake II RC *et al.* (2005). Community proteomics identifies key activities in a natural microbial biofilm. *Science* **308**: 1915–1920.

Rosenstiel P, Sina C, End C, Renner M, Lyer S, Till A *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, Andersson AF, Lefsrud MG, Wexler M, Shah M, Zhang B *et al.* (2008). Community proteogenomics highlights strain-variant microbial protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J* **2**: 853–864.

Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. (2006). Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J Proteome Res* **5**: 2339–2347.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)