Metaproteomics reveals functional shifts in microbial and human proteins during a preterm infant gut colonization case

Jacque C. Young[1,2]*, Chongle Pan[2], Rachel M. Adams[1,2], Brandon Brooks[3], Jillian F. Banfield[3], Michael J. Morowitz[4] and Robert L. Hettich[2]

[1] Genome Sciences and Technology Graduate School, University of Tennessee, Knoxville, TN, USA [2] Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA [3] Department of Earth and Planetary Sciences, University of California, Berkeley, CA, USA [4] Division of Pediatric General & Thoracic Surgery, University of Pittsburgh, Pittsburgh, PA, USA

Correspondence: Dr. Robert Hettich, Oak Ridge National Laboratory, PO Box 2008 MS-6131, Oak Ridge, TN 37831-6131, USA E-mail: hettichrl@ornl.gov

Abstract

Microbial colonization of the human gastrointestinal tract plays an important role in establishing health and homeostasis. However, the time-dependent functional signatures of microbial and human proteins during early colonization of the gut have yet to be determined. To this end, we employed shotgun proteomics to simultaneously monitor microbial and human proteins in fecal samples from a preterm infant during the first month of life. Microbial community complexity increased over time, with compositional changes that were consistent with previous metagenomic and rRNA gene data. More specifically, the function of the microbial community initially involved biomass growth, protein production, and lipid metabolism, and then switched to more complex metabolic functions, such as carbohydrate metabolism, once the community stabilized and matured. Human proteins detected included those responsible for epithelial barrier function and antimicrobial activity. Some neutrophil-derived proteins increased in abundance early in the study period, suggesting activation of the innate immune system. Likewise, abundances of cytoskeletal and mucin proteins increased later in the time course, suggestive of subsequent adjustment to the increased microbial load. This study provides the first snapshot of coordinated human and microbial protein expression in a preterm infant's gut during early development.

Keywords: Infant gut / Metaproteomics / Microbial colonization / Microbiome / Systems biology

1 Introduction

Microbial communities in the gastrointestinal tract play important roles in human health by processing essential nutrients, protecting against pathogenic bacteria, promoting angiogenesis, and regulating host immune responses [1-5]. Initially near sterile, the infant gastrointestinal tract assembles a microbial community of diverse species composition in the first 2.5 years of life [5-10]. This symbiotic relationship requires a careful balance; it has been

postulated that disruption of the host–microbe relationship in the gut can lead to diseases such as inflammatory bowel disease and neonatal necrotizing enterocolitis (NEC) [11, 12]. Initial temporal colonization patterns and species distributions vary between individual infants and may be influenced by environmental exposures, delivery mode, diet, and health [6, 7, 9]. In preterm infants, the microbial community compositions in the gastrointestinal tract are low in diversity, highly variable between infants, and change over time at the species and strain level [13-19]. The lack of microbial community complexity in preterm infants provides a powerful opportunity to study the microbiome development at high resolution.

Previously, the gut microbial compositional patterns of a preterm infant during the first month of life were characterized in an rRNA gene and metagenomics-based study [14]. Through 16S rRNA gene-based analysis of fecal samples, the dominant taxa were identified, and community compositional changes revealed three distinct colonization phases. Specifically, days 5 through 9 of the infant's life were dominated by *Leuconostoc*, *Weisella*, and *Lactococcus* species, while days 10 through 14 consisted primarily of *Pseudomonas* and *Staphylococcus*. The third phase (days of life 15–21) was primarily composed of members of the *Enterobacteriaceae* family, including *Citrobacter* and *Serratia* species. This pattern coincided with dietary adjustments at days 9 and 15, and was similar to premature infants from other studies [14, 16, 19]. Sequencing of whole community DNA, followed by reconstruction and intensive curation of population genomic datasets of the dominant microbial members from days 10, 16, 18, and 21, revealed three major species in samples from these days: a *Serratia* strain (*UC1SER*), an *Enterococcus* strain (*UC1ENC*), and two closely related *Citrobacter* strains (*UC1CITi* and *UC1CITii*). Also present were plasmids *UC1CITp, UC1ENCp*, and bacteriophage *UC1ENCv* (as well as many other incompletely resolved phage/plasmids). While critically important, genomic information provides the *inventory* of possible gene products, but does not reveal the actual metabolic *activities*. Thus, here we employed shotgun proteomics via nano-2D-LC-MS/MS to elucidate *functional signatures* of translated gene products (i.e., proteins) from samples of the same preterm infant for which metagenomic data were available.

The use of MS-based proteomics allows characterization and quantification of thousands of proteins within a microbial community [20-25]. While early attempts using metaproteomics for the characterization of the infant microbiome demonstrated feasibility, protein identifications were limited due to insufficient genome information [26]. More recently, due to the availability of highly resolved genomes and the advancement in MS instrumentation, proteins can be confidently identified at the species and strain level allowing deep proteomic analysis of microbial communities [20, 21, 23, 24, 27-29]. While prior studies have focused on characterizing microbial genes and proteins, most current methodologies prohibit global analyses of microbial proteins in conjunction with human proteins. Here, we report results of the first

proteomics-based investigation of the coordinated expression of human and microbial proteins during initial microbial colonization of a preterm infant's gut microbiome.

2 Materials and methods

2.1 Description of preterm infant

A female infant born at 28 weeks gestation due to premature rupture of membranes was delivered by cesarean section and given antibiotics for the first 7 days of life [14]. Enteral feedings with breast milk were given on days 4–9, and then on days 9–13, feedings were withheld due to abdominal distension. After day 13, enteral feedings were readministered in the form of artificial infant formula. The baby also received supplemental parenteral nutrition until day 28. The baby had no major anomalies or comorbidities and was discharged to home on day of life 64. Fecal material was collected on days 5–21 as available, with institutional approval, and was immediately stored at −80°C until analysis. Metagenomic and 16S rRNA data were analyzed in a companion study from matched samples [14]. Based upon sample availability, proteomic measurements were performed on fecal samples from days 7, 13, 15, 16, 17, 18, 20, and 21 after birth.

2.2 Protein extraction and enzymatic digestion of fecal samples

Approximately 250 µg of fecal material was boiled for 5 min in 1 mL 100 mM Tris-Cl containing 4% w/v SDS and 10 mM DTT, and then underwent continuous bead beating on high setting for 30 min, in order to lyse cells and denature/reduce proteins. The supernatant was collected, boiled again, spun down (14 000 g), and precipitated with 20% trichloroacetic acid at −80°C overnight. Protein pellets were washed in ice-cold acetone, resolubilized in 8 M urea diluted in 100 mM Tris-HCl pH 8, and then sonicated using a Branson sonic disruptor in order to break up the pellet (5 min at 20%; 10 s on, 10 s off). Iodoacetamide was added to block disulfide bond reformation. Proteins were quantified using bicinchronic assay and between 1 and 3 mg protein was diluted to 4 M urea in 100 mM Tris-HCl pH 8, and enzymatically digested into peptides using sequencing grade trypsin (Promega) for 4 h at room temperature. Peptides were diluted to 2 M urea, a second dose of trypsin added, and digestion continued overnight. An acidic salt solution (200 mM NaCl, 0.1% formic acid), was used to clean up the peptides, which were then spun through a 10 kDa cutoff spin column filter (VWR). Peptides were quantified by bicinchronic assay assay and stored at −80°C until further use.

2.3 Nano-2D-LC-MS/MS

A 150 µg peptide mixture was loaded via a pressure cell onto a 150 µm inner diameter split-phase fused silica back column (Polymicro Technologies) hand-packed with reverse phase (C18) and SCX resin (Luna, Phenomenex). The back column was washed offline with 100% solvent A (95% H_2O, 5% CH_3CN, and 0.1% formic acid) for 15 min at ~140 bar to desalt the column. Peptides were placed in line with a nanospray emitter (New Objective)

packed with reverse phase material and then separated online using high-performance 2D-LC [30-32]. Peptides were eluted from the SCX resin by increasing ammonium acetate salt pulses followed by reverse-phase resolution over 2-h organic gradients as described previously [20, 21, 28], ionized via nanospray (200 nL/min) (Proxeon, Cambridge, MA, USA), and analyzed using an LTQ Orbitrap Velos mass spectrometer (ThermoFisher Scientific, San Jose, CA, USA). The mass spectrometer was run in data-dependent mode with the top ten most abundant peptides in full MS selected for MS/MS, and dynamic exclusion enabled (repeat count = 1, 60 s exclusion duration). Full MS scans were collected in the Orbitrap at 30 K resolution. Two microscans were collected in centroid mode for both full and MS/MS scans. Technical duplicates were run for all samples. The MS proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD000114.

2.4 Database construction and searching

Metagenomic sequences from matched samples collected on days 10, 16, 18, and 21 were translated, and the predicted protein sequences of the dominant community members were used to generate a search database. These included a *Serratia* species *UC1SER*, two closely related *Citrobacter* strains, *UC1CITi* and *UC1CITii*, an *Enterococcus* species *UC1ENC*, and associated virus and plasmids *UC1ENCp*, *UC1ENCv*, and *UC1CITp* [14]. Since microbial species from early time points were not represented in the metagenomic sequences, 20 additional isolate sequences were selected, based on 16S rRNA information from matched samples, and added to the search database (acquired from JGI: http://www.hmpdacc-resources.org/cgi-bin/img_hmp/main.cgi in January of 2011). For example, according to the 16S data, *Leuconostoc* was prevalent on days 5–9 [14]. But since metagenomic sequencing was not performed on days 5–9, *Leuconostoc* would have been absent from the analysis. 16S sequencing could only identify *Leuconostoc* at the genus level, so we chose a representative species of this genus, *Leuconostoc mesenteroides cremoris*, to be included in the database. In addition, human protein sequences (NCBI RefSeq_2011) and common contaminants (i.e., trypsin) were appended to the database. All MS/MS spectra were searched against the concatenated database with the SEQUEST algorithm v.27 (rev.9) [33], and filtered with DTASelect version 1.9 [34] to assemble the identified peptides into their corresponding protein sequences (Supporting Information Table 1). Due to carbamidomethylation effects of iodoacetamide, a static cysteine modification (+57) was included in all searches. Only proteins identified with two fully tryptic peptides were considered for further biological study. Reversed protein sequences were appended to the database in order to calculate FDRs. The final concatenated search database contained 214 520 protein sequences, including forward and reversed sequences (The search database, Isolate_UC1_HRefSeq2011_IgAM_20120113_CFR_anno3.fasta, is available to

download from the ProteomeXchange repository using identifier PXD00011). Conservative cross-correlation filters were used to achieve FDRs between 0.5 and 2.4% at the peptide level [35].

2.5 Database clustering and spectral balancing

Since MS-based proteomics identifies proteins by their corresponding peptide sequences, data analysis must take into consideration the high levels of protein redundancy within and between species to avoid inflating the total number of proteins identified or misinterpretation of the biological conclusions by overrepresenting proteins with the same function. Therefore, we applied a bioinformatic clustering algorithm to the database in order improve confidence in protein identification and quantification. Specifically, using the publically available software, USEARCH v.5.0 [36], microbial proteins were clustered into a protein group if they shared 100% amino acid identity, and human proteins were clustered into a protein group if they contained ≥90% amino acid similarity. These differing similarity thresholds were chosen based on the higher numbers of paralogous proteins present within the human genome, and were supported by plotting similarity thresholds ranging from 0.5 to 1 against the percent proteome reduction via clustering [37]. Each protein group contained at least one unique peptide. Spectral counts were assigned, balanced, normalized, and adjusted according to methods previously described, yielding adjusted normalized spectral abundance factor (NSAF) values [37-39]. In total, 4413 microbial and 3062 human protein groups were detected across the dataset (Supporting Information Table 2). Protein groups ranged from singletons to groups that contain multiple protein isoforms. All peptides identified throughout the time course are provided in Supporting Information Table 3.

2.6 Data analyses

Clusters of orthologous group (COG) assignments for each microbial protein sequence were determined by running rpsblast against the COG database from NCBI, using an *E*-value threshold of 0.00001 and the top hit used for the assignment [40]. Adjusted NSAFs from all microbial protein groups were summed and grouped into their respective (COG) categories. A linear regression analysis, computed using the basic statistical package in *R*, was used to model the relationship between the proportions of proteins within each COG category and increasing time points [41]. The major canonical pathways for human proteins detected across the dataset were determined using Ingenuity Pathway Analysis software (Ingenuity Systems, www.ingenuity.com). The significance of the association was measured by calculating the ratio of number of *detected* proteins that map to the pathway divided by the *total* number of proteins from that pathway (orange boxes). The Fisher's exact test was used to calculate a *p*-value determining the probability that the association between the proteins in the dataset and the canonical pathway is explained by chance alone (*y*-axis). Hierarchical clustering of individual human proteins was carried out using JMP Genomics

using the log transformed NSAFs values above and below the median across all time points for each protein.

3 Results

3.1 Overall proteome characterization

Proteome extracts of fecal samples from a preterm infant on days 7, 13, 15, 16, 17, 18, 20, and 21 after birth were examined via nano-2D-LC-MS/MS. Up to 73 257 mass spectra, 16 605 peptides, and 4031 proteins were identified per run (Supporting Information Table 1), providing deep proteomic coverage of both microbial and human components. Technical duplicates were run for each sample, with comparable reproducibility between replicates (Supporting Information Fig. 1).

By measuring both microbial and human proteins *simultaneously* in each run, we observed an increased complexity of the microbial composition and a decrease in the ratio of total human/microbial proteins with time (Fig. 1). At the earliest time point, when the initial microbial communities were being established, human proteins comprised ~96% of all proteins identified (day 7). The low microbial load may be a consequence of antibiotic administration during the first week of life for this particular infant. Human proteins comprised ~72% of the identified protein dataset on day 13, and by day 15 the percent of human proteins decreased to ~30%, with a concomitant increase in the number of microbial proteins detected. The ratio of human to microbial proteins remained at this level for the remainder of the times measured, with the exception of day 20, when an unexpected rise in human proteins was detected (to be discussed in more detail below). The number of total spectra collected on day 20 was comparable to adjacent days (Supporting Information Fig. 2), so the variance was likely not due to a technical issue related to the MS measurement.

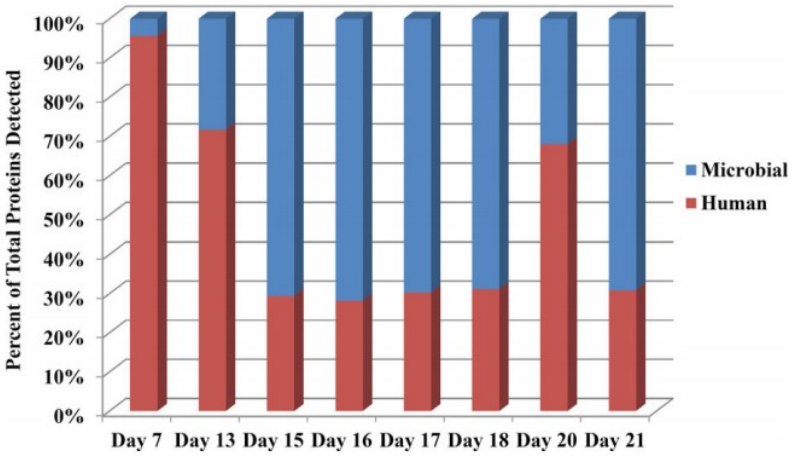

Figure 1. Distribution of human and microbial proteins. Adjusted NSAFs from microbial (blue) and human (red) protein groups were averaged between two technical replicates, summed for each time point (*x*-axis), and plotted as a percent of the total proteins detected for each day (*y*-axis). Technical variability between the two replicates was low as shown by the SDs of the microbial and human protein groups: day 7 = 0.02%, day 13 = 0.82%, day 15 = 4.19%, day 16 = 7.74%, day17 = 7.99%, day 18 = 1.71%, day 20 = 0.03%, and day 21 = 0.57%.

3.2 Microbial protein distribution and functional categorization

When microbial protein groups from different species were compared across time, the distribution of species was similar to that seen in 16S rRNA and metagenomic data (Supporting Information Fig. 3) [14]. At day 7, microbial proteins were very low in abundance, but increased by day 13. This time point was dominated by *Pseudomonas* and *Staphylococcus* proteins. However, by day 15 we began to see the emergence of *Serratia* (*UC1SER*) and *Citrobacter* (*UC1CIT*) proteins, which persisted in days 16–21. In total, this corresponds closely with previous metagenomic data from matched samples, which showed distinct community memberships in colonization phase I (days 5–9), phase II (days 10–15), and phase III (days 16–21) [14]. Proteomic data also suggest *UC1SER* and *UC1CIT* were the functionally dominant members of the community during the third colonization phase, as demonstrated by the highest contribution of microbial proteins from these species during these time points.

Microbial community functions were analyzed by grouping proteins into COG categories, and then measuring the relative changes in the proportions of proteins in each category over time (Fig. 2 and Supporting Information Fig. 4). Because detailed metabolic characterization of microbial membership functions at the strain level was beyond the scope of this paper, and since the main objective of this study was to more broadly compare and link microbial/human host protein signatures over temporal development, we opted to use COG category representation to identify the range of metabolic activities of the microbial community and assess the functional changes over time. At day 7, although there was a very limited level of microbial peptide abundance measured, most of the signal originated from an aspartokinase I–homoserine dehydrogenase protein belonging to *Bacteriodes fragilis*. This enzyme, from the amino acid metabolism and transport COG category, catalyzes a reaction in the aspartate pathway and may aid in providing essential amino acids from dietary sources to the human host (infant) at this early stage of development [42]. Since there were so few microbial proteins detected from day 7 samples in comparison to other days, we excluded this day from the remainder of the COG analysis.

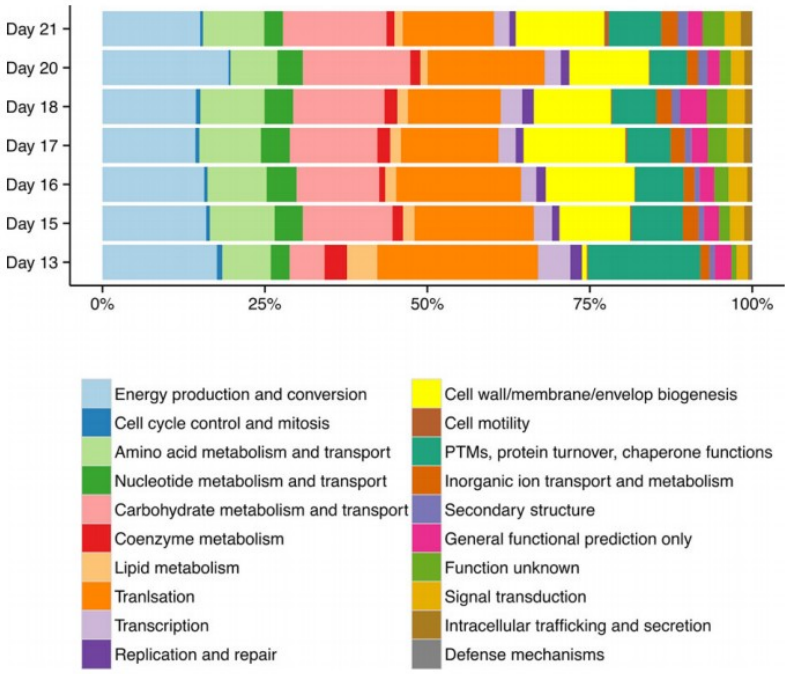

Figure 2. Analysis of microbial proteins by COG category classifications. Microbial proteins were assigned to COG categories and adjusted NSAFs for each group summed and plotted as percent of total NSAFs for each time point [40]. Day 7 is removed from the analysis due to low abundance values of microbial proteins from that time point. Categories that significantly increased in proportions of proteins over time included: carbohydrate transport and metabolism ($p = 0.015$), secondary structure ($p = 0.002$), and intracellular trafficking and secretion ($p = 0.025$). Proportions of proteins in the categories of cell cycle control and mitosis ($p = 0.040$), lipid metabolism ($p = 0.039$), and translation ($p = 0.046$) decreased over time.

Changes in the proportions of proteins belonging to each COG category over time were assessed using a linear regression analysis (Supporting Information Fig. 4). The statistical analysis revealed that proportions of proteins belonging to the categories of lipid transport and metabolism, cell cycle control and mitosis, and translation ribosomal structure and biogenesis categories were abundant early, but decreased in relative abundance over time. In contrast, carbohydrate transport and metabolism, secondary metabolites biosynthesis, transport, and catabolism, membrane biogenesis, and intracellular trafficking and secretion, while lower in proportion initially, increased with time. In general, this information indicates that the microbial community initially focused its resources on biomass growth, protein production, and lipid metabolism (presumably to establish the stable microbiome), and then switched to more complex metabolic functions, such as carbohydrate metabolism, once the community stabilized and matured (around day 15). Interestingly, the functional distribution of the microbiome after about 3 weeks is very similar to what is observed in the stable adult human gut [20].

3.3 Functional distributions of human proteins

Human proteins detected across all time points were categorized into canonical pathways using Ingenuity Pathway Analysis software. The topmost abundant categories were determined based on the number of proteins belonging to that category, and included those related to basic cellular functions such as glycolysis, oxidative phosphorylation, and elongation factor 2 signaling (Supporting Information Fig. 5). Other categories, such as inflammatory response, were not displayed since the total number of

proteins detected in this category was not among the top 20 overall. However, we detected over 30 inflammatory proteins, and some of these individual proteins were among the most abundantly detected in the samples (Calprotectin [18 385; this is the adjusted NSAFs collected across all time points], ALPI [13 650], SERPINA [37 140] and ANPEP [9415]) (Supporting Information Table 4).

Some of the most abundant human proteins detected are involved in host–microbe interactions (Supporting Information Table 4). In particular, the most abundant protein detected in our samples, the calcium-activated chloride-channel 1 (CLCA1) protein [17 002], is involved in mucus secretion by goblet cells [43]. Likewise, Fc fragment of the IgG-binding protein (FCRPB/Fcgbp) [23 096] is expressed by placental and colonic epithelial cells, and has been reported to bind mucin 2 (MUC2), and play a key role in immune protection and inflammation [44]. In addition, antimicrobial and innate immune proteins including lactoferrin (LTF), intelectin (ITLN1), and olfactomedin (OLFM4) were among the most abundant proteins detected [8298, 11 214, and 4875 NSAFs, respectively]. Lactoferrin (aka lactotransferrin), an iron-binding glycoprotein, is a key player in the innate immune system and is abundant and ubiquitous in human secretions such as breast milk. It has been shown to attenuate pathogenic bacteria, interfering with colonization and biofilm formation [45-47].

3.4 Intestinal barrier proteins

Throughout our proteome datasets, we identified numerous human proteins involved in intestinal barrier formation and function (Supporting Information Table 5). The intestinal barrier is composed of enterocytes, absorptive epithelial cells held together by tight junctions, which serve as a physical barrier, and the mucus layer. We detected numerous tight junction proteins including occludin (OCLN), claudins (CLDN18, CLDN23, CLDN3, CLDN7), and tight junction proteins 1, 2, and 3 (TJP1, TJP2, TJP3, or zona occludens 1, 2, and 3). In addition, proteins involved in the tight junction-signaling pathway were identified (Fig. 3). Also detected were numerous mucin proteins, including both secretory gel-forming mucins (MUC2, MUC5AC, MUC5B, and MUC6) and membrane-bound mucins (MUC1, MUC3B, and MUC4) (Supporting Information Table 5). Several enzymes in the *o*-glycan biosynthesis pathway also were detected, including those involved in synthesizing core 3 type glycans, the major type associated with MUC2 [48, 49]. In addition, all three trefoil factor family peptides TFF1, TFF2, and TFF3, a family of proteins, which play an important role in maintenance and repair of the intestinal mucosa, were detected [50].

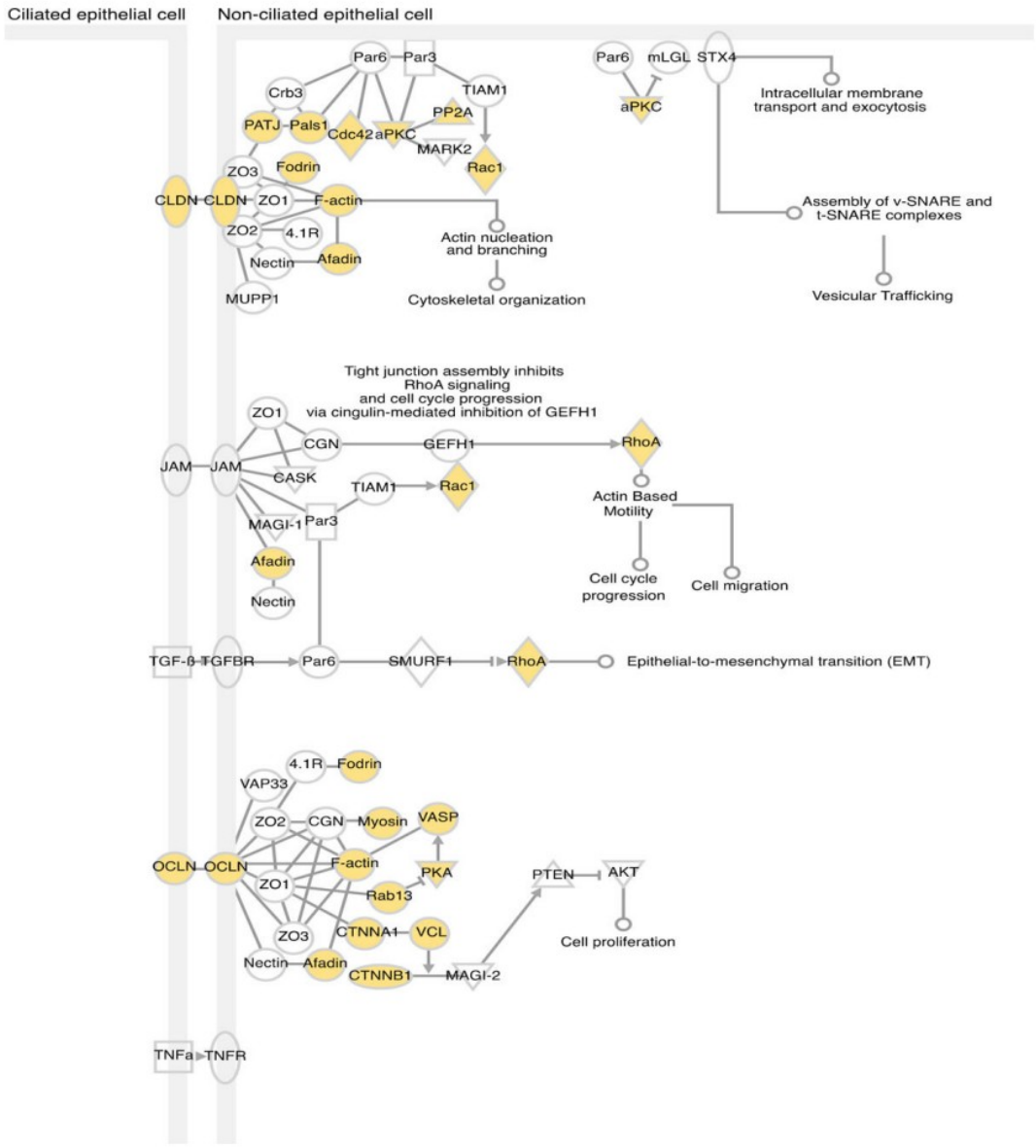

Figure 3. Tight junction signaling pathway. Proteins in the tight junction signaling pathway as determined by Ingenuity Pathway Analysis software. Proteins colored in yellow are those detected by proteomics.

Secretory IgA is an important component of the intestinal barrier that specifically binds bacteria, limiting their association with the epithelial cell surface and restricting penetration across the gut epithelia [51-54]. We detected components of secretory Immunoglobulin A, including the two IgA heavy chain constant regions (IgA1 and IgA2), the J chain (15 kDa polypeptide), and the secretory component of the polymeric immunoglobulin receptor (pIgR: 130 kDa). The poly-Ig receptor is expressed by epithelial cells, binds to the IgA oligomers, and allows transport across the mucosal epithelium.

In addition, we detected several antimicrobial proteins, including alpha defensins (DEFA1, DEFA5), lysozyme (LYZ), and phospholipase A2 (PLA2). These antimicrobial proteins are secreted by a subset of gut epithelial cells,

Paneth cells, which directly sense gut commensal bacteria through MyD88-dependent toll-like receptor signaling that triggers expression of certain antimicrobial factors, limits bacterial penetration of host tissues, and maintains microbial-host homeostasis in the intestine [55]. Thus, detection of these proteins might suggest that the premature infant's gut, even at early stages of development, could have been responding to the introduction of microbial inhabitants and exerted pressure on the community to maintain homeostasis.

3.5 Differential protein expression across time

Overall, human proteins, when summed across all samples, contributed mostly to generalized maintenance functions (Supporting Information Fig. 4). However, when human proteins were clustered based on shared trends in spectral count abundance changes (Fig. 4), time shifts were apparent. Several neutrophil derived proteins such as neutrophil elastase (ELANE), calprotectin (S100-A8/S100-A9), and myeloperoxidase were most abundant at day 7 (Fig. 4, cluster #6) suggesting that activation of the innate immune system occurs early in correspondence with the initial arrival and ensuing establishment of the microbiome. Human cytoskeletal proteins (KRT8, KRT13, KRT18, KRT19, and KRT20) and mucins (MUC2, MUC5B) were more predominant in later time points (days 20–21) (Fig. 4, clusters 7 and 10), suggesting structural and epithelial barrier proteins are compensating for the increased microbial load.

4 Discussion

In this study, we simultaneously monitored both human and microbial proteomes over a time course in early microbiome development of a preterm infant. Microbial proteins detected in this time course are consistent with metagenomic inference of distinct colonization phases with vastly different species composition [14]. The functions of the microbial community shifted over time, with early resources focusing on cell division, protein production, and lipid metabolism. As time progressed, the microbial community increased in size and diversity, stabilized, and switched its focus to breaking down carbohydrates, making secondary metabolites, and secreting and trafficking proteins.

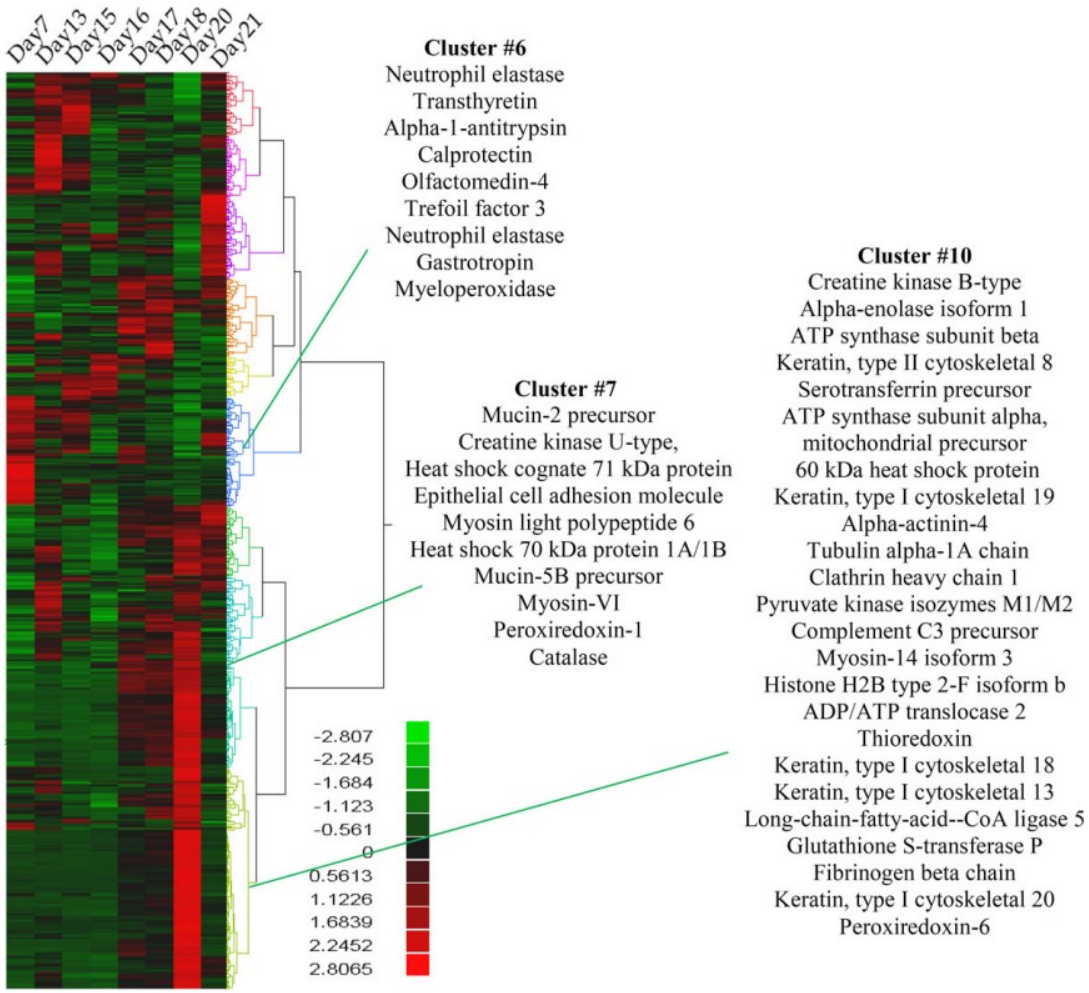

Figure 4. Human proteins changing across time. Proteins were clustered based on abundance changes across time. The mean of the normalized spectral counts across all time points for each protein was taken. The scale reflects the log transformed value above and below the median.

Predominant throughout our measurements were human proteins involved in intestinal barrier function. Breakdown of the intestinal barrier or incomplete formation, as seen in premature infants, can contribute to bacterial translocation and disease states such as NEC [11]. The mucus layer is a major component of the intestinal barrier that helps maintain homeostasis between the gut microbiota and their host by minimizing physical contact between the microbes and intestinal epithelial cells [2]. In the colon, the outer mucus layer harbors commensal bacteria while the thicker, impenetrable inner layer offers protection by providing a physical barrier as well as containing antimicrobial compounds and secretory IgA [48, 56]. The small intestine is composed of only one mucus layer, but still provides a physical barrier with a 50 µm area separating the bacteria from the epithelia [57]. The mucus layer is composed of mucins, glycoconjugates of a polypeptide core covered in O-linked carbohydrate side chains that are secreted by goblet cells. The O-linked glycans provide an energy source for bacteria in the outer mucus layer [4, 58]. In our proteomic analyses, we detected numerous mucin proteins.

Most of these were detected at relatively constant abundances throughout all the time points. However, some like the mucin 2 precursor, increased in abundance during the third colonization phase. Mucin 2 (MUC2) is the most abundant mucin in the intestine and has been directly linked to protecting the colonic epithelium from enteric pathogens [48]. It is downregulated in patients with ulcerative colitis and Crohn's disease [59]. Detection and quantification of the numerous intestinal barrier proteins in this study suggest that comprehensive proteomic analysis of easily obtained fecal samples may represent an effective yet noninvasive means to evaluate gut barrier function in human patients.

As noted, there was a dramatic increase in the numbers of human proteins identified on day 20. Since many of these proteins were keratins, which are important components of both skin cells and gut epithelial cells, there are two possible explanations: (1) human contamination during sample handling, or (2) an increased sloughing event in the GI tract at this time. The 134 human proteins that were solely detected at this time point contribute to a wide range of biological functions (Supporting Information Fig. 6A). Thus, we propose that they are more likely to represent sloughed epithelial cells rather than skin cells, and thus conclude that contamination during sample handling was not the most likely explanation. The most highly expressed canonical pathways on this day were those of basic metabolic functions including: EIF2 signaling, pyruvate metabolism, glycolysis (Supporting Information Fig. 6A). Additional proteins from the pathways for pyruvate metabolism, glycolysis, and granzyme A signaling were detected on day 20 (Supporting Information Fig. 6B and C). Note that the ratio of microbial/human proteins and the range of microbial activities revert back to expected values on day 21, indicating that the microbiome was not highly perturbed by whatever event happened on day 20.

Initial microbial colonization of the gastrointestinal tract is a crucial process in a healthy human infant. This process educates the innate immune system and initiates the establishment of a delicate homeostasis between human host and resident microbes. In premature infants, the host–microbe relationship is undoubtedly impacted significantly by underdevelopment of the intestinal barrier, an immature innate immune system, antibiotic administration, and exposure to pathogenic organisms in the intensive care unit [60]. While prior studies have investigated the succession of gut microbiota primarily at the gene level, the functional signatures of microbial and human proteins early in life can provide detailed metabolic activity information [61]. Thus, this study provides detailed information about the microbial and human proteins in fecal samples from a newborn premature infant during the first month of life, and reveals the complex but synergistic host adaptation to microbiome establishment.

The infant in this study was born prematurely, treated with antibiotics, and experienced gastrointestinal distention during which feedings were withheld. These and other factors could have influenced microbial colonization. Recent

studies have reported high intraindividual diversity in microbial species compositions among preterm infants [15, 62]. As more proteomic data are collected on preterm infants, it will be interesting to see how microbial community functions compare between individuals—between both healthy preterm infants and those who develop NEC. While there may be commonalities or correlations in proteomic responses that predict which preterm infants progress to NEC, there is a strong possibility that due to intraindividual variability, treatment will need to be specialized to each individual. Thus, further research in this area could support personalized medicine for neonatal care.

Acknowledgment

We thank Dr. David Tabb and the Yates Proteomics Laboratory at Scripps Research Institute for DTASelect/Contrast software, Langho Lee for bioinformatics assistance, and the PRIDE team. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy. J.Y. acknowledges stipend support from the Genome Science and Technology program at the University of Tennessee, Knoxville. This work was funded in part by March of Dimes Foundation research grant 5-FY10-103 (M.J.M), *NIH* grant *1R01-GM-103600*, *and an NSF* Grad*uate Fellowship to B.B*.

References

1 Stappenbeck, T. S., Hooper, L. V., Gordon, J. I., Developmental regulation of intestinal angiogenesis by indigenous microbes via Paneth cells. *Proc. Natl. Acad. Sci. USA* 2002, 99, 15451– 15455.

2 Hooper, L. V., Midtvedt, T., Gordon, J. I., How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr.* 2002, 22, 283– 307.

3 MacDonald, T. T., Pettersson, S., Bacterial regulation of intestinal immune responses. *Inflamm. Bowel Dis.* 2000, 6, 116– 122.

4 Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., Gordon, J. I., Host-bacterial mutualism in the human intestine. *Science* 2005, 307, 1915– 1920.

5 Putignani, L., Del Chierico, F., Petrucca, A., Vernocchi, P., Dallapiccola, B., The human gut microbiota: a dynamic interplay with the host from birth to senescence settled during childhood. *Pediatr. Res.* 2014, 76, 2– 10.

6 Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., Brown, P. O., Development of the human infant intestinal microbiota. *PLoS Biol.* 2007, 5, e177.

7 Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D. et al., Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. USA* 2011, 108, 4578– 4585.

8 Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I. et al., Human gut microbiome viewed across age and geography. *Nature* 2012, 486, 222– 227.

9 Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M. et al., Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. USA* 2010, 107, 11971– 11975.

10 Bergstrom, A., Skov, T. H., Bahl, M. I., Roager, H. M. et al., Establishment of intestinal microbiota during early life: a longitudinal, explorative study of a large cohort of Danish infants. *Appl. Environ. Microbiol.* 2014, 80, 2889– 2900.

11 Neu, J., Walker, W. A., Necrotizing enterocolitis. *N. Engl. J. Med.* 2011, 364, 255– 264.

12 Morowitz, M. J., Poroyko, V., Caplan, M., Alverdy, J., Liu, D. C., Redefining the role of intestinal microbes in the pathogenesis of necrotizing enterocolitis. *Pediatrics* 2010, 125, 777– 785.

13 Brown, C. T., Sharon, I., Thomas, B. C., Castelle, C. J., Morowitz, M. J., Banfield, J. F., Genome resolved analysis of a premature infant gut microbial community reveals a varibaculum cambriense genome and a shift towards fermentation-based metabolism during the third week of life. *Microbiome* 2013, 1, 30.

14 Morowitz, M. J., Denef, V. J., Costello, E. K., Thomas, B. C. et al., Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proc. Natl. Acad. Sci. USA* 2011, 108, 1128– 1133.

15 Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K. et al., Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* 2013, 23, 111– 120.

16 Wang, Y., Hoenig, J. D., Malin, K. J., Qamar, S. et al., 16S rRNA gene-based analysis of fecal microbiota from preterm infants with and without necrotizing enterocolitis. *ISME J.* 2009, 3, 944– 954.

17 Mai, V., Torrazza, R. M., Ukhanova, M., Wang, X. et al., Distortions in development of intestinal microbiota associated with late onset sepsis in preterm infants. *PLoS One* 2013, 8, e52876.

18 Mai, V., Young, C. M., Ukhanova, M., Wang, X. et al., Fecal microbiota in premature infants prior to necrotizing enterocolitis. *PLoS One* 2011, 6, e20647.

19 Mshvildadze, M., Neu, J., Shuster, J., Theriaque, D. et al., Intestinal microbial ecology in premature infants assessed with non–culture-based techniques. *J. Pediatr.* 2010, 156, 20– 25.

20 Verberkmoes, N. C., Russell, A. L., Shah, M., Godzik, A. et al., Shotgun metaproteomics of the human distal gut microbiota. *ISME J* 2009, 3, 179– 189.

21 Ram, R. J., Verberkmoes, N. C., Thelen, M. P., Tyson, G. W. et al., Community proteomics of a natural microbial biofilm. *Science* 2005, 308, 1915– 1920.

22 Knief, C., Delmotte, N., Chaffron, S., Stark, M. et al., Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. *ISME J* 2012, 6, 1378– 1390.

23 Hettich, R. L., Sharma, R., Chourey, K., Giannone, R. J., Microbial metaproteomics. Identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. *Curr Opin Microbiol*. 2012, 15, 373– 380.

24 Kolmeder, C. A., de Vos, W. M., Metaproteomics of our microbiome - developing insight in function and activity in man and model systems. *J. Proteomics* 2014, 97, 3– 16.

25 Hettich, R. L., Pan, C., Chourey, K., Giannone, R. J., Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. *Anal. Chem.* 2013, 85, 4203– 4214.

26 Klaassens, E. S., De Vos, W. M., Vaughan, E. E., Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. *Appl. Environ. Microbiol.* 2007, 73, 1388– 1392.

27 Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y. et al., Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* 2012, 7, e49138.

28 Lo, I., Denef, V. J., Verberkmoes, N. C., Shah, M. B. et al., Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* 2007, 446, 537– 541.

29 Wilmes, P., Bond, P. L., Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol.* 2006, 14, 92– 97.

30 McDonald, W. H., Ohi, R., Miyamoto, D. T., Mitchison, T. J., Yates, J.R., 3rd, 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.* 2002, 219, 245– 251.

31 Washburn, M. P., Ulaszek, R., Deciu, C., Schieltz, D. M., Yates, J. R., 3rd, Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal. Chem.* 2002, 74, 1650– 1657.

32 Washburn, M. P., Wolters, D., Yates, J. R., 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 2001, 19, 242– 247.

33 Eng, J. K., McCormack, A. L., Yates, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am. Soc. Mass Spectrom.* 1994, 5, 976– 989.

34 Tabb, D. L., McDonald, W. H., Yates, J. R., 3rd, DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* 2002, 1, 21– 26.

35 Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., Gygi, S. P., Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* 2003, 2, 43– 50.

36 Edgar, R. C., Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010, 26, 2460– 2461.

37 Abraham, P., Adams, R., Giannone, R. J., Kalluri, U. et al., Defining the boundaries and characterizing the landscape of functional genome expression in vascular tissues of Populus using shotgun proteomics. *J. Proteome Res.* 2011, 11, 449– 460.

38 Giannone, R. J., Huber, H., Karpinets, T., Heimerl, T. et al., Proteomic characterization of cellular and molecular processes that enable the Nanoarchaeum equitans-Ignicoccus hospitalis relationship. *PLoS One* 2011, 6, e22942.

39 Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K. et al., Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.* 2006, 5, 2339– 2347.

40 Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A. et al., The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 2001, 29, 22– 28.

41 R Development Core Team, *R Foundation for Statistical Computing*, Vienna, Austria 2011.

42 Viola, R. E., The central enzymes of the aspartate family of amino acid biosynthesis. *Acc. Chem. Res.* 2001, 34, 339– 349.

43 Loewen, M. E., Forsyth, G. W., Structure and function of CLCA proteins. *Physiol. Rev.* 2005, 85, 1061– 1092.

44 Johansson, M. E. V., Thomsson, K. A., Hansson, G. C., Proteomic analyses of the two mucus layers of the colon barrier reveal that their main component, the Muc2 mucin, is strongly bound to the Fcgbp protein. *J Proteome Res.* 2009, 8, 3549– 3557.

45 Qiu, J., Hendrixson, D. R., Baker, E. N., Murphy, T. F. et al., Human milk lactoferrin inactivates two putative colonization factors expressed by Haemophilus influenzae. *Proc. Natl. Acad. Sci. USA* 1998, 95, 12641– 12646.

46 Singh, P. K., Parsek, M. R., Greenberg, E. P., Welsh, M. J., A component of innate immunity prevents bacterial biofilm development. *Nature* 2002, 417, 552– 555.

47 Legrand, D., Pierce, A., Elass, E., Carpentier, M., Mariller, C., Mazurier, J., Lactoferrin structure and functions. In *Bioactive Components of Milk* (pp. 163– 194). Springer New York. 2008.

48 Johansson, M. E. V., Larsson, J. M. H., Hansson, G. C., The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host–microbial interactions. *Proc. Natl. Acad. Sci. USA* 2011, 108, 4659– 4665.

49 Larsson, J. M. H., Karlsson, H., Sjövall, H., Hansson, G. C., A complex, but uniform O-glycosylation of the human MUC2 mucin from colonic biopsies analyzed by nanoLC/MSn. *Glycobiology* 2009, 19, 756– 766.

50 Albert, T. K., Laubinger, W., Müller, S., Hanisch, F. G. et al., Human intestinal TFF3 forms disulfide-linked heteromers with the mucus-associated FCGBP protein and is released by hydrogen sulfide. *J. Proteome Res.* 2010, 9, 3108– 3117.

51 Hooper, L. V., Macpherson, A. J., Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nature Rev. Immunol.* 2010, 10, 159– 169.

52 Macpherson, A. J., Uhr, T., Induction of protective IgA by intestinal dendritic cells carrying commensal bacteria. *Science* 2004, 303, 1662– 1665.

53 Macpherson, A. J., Gatto, D., Sainsbury, E., Harriman, G. R. et al., A primitive T cell-independent mechanism of intestinal mucosal IgA responses to commensal bacteria. *Science* 2000, 288, 2222– 2226.

54 Suzuki, K., Meek, B., Doi, Y., Muramatsu, M. et al., Aberrant expansion of segmented filamentous bacteria in IgA-deficient gut. *Proc. Natl. Acad. Sci. USA* 2004, 101, 1981– 1986.

55 Vaishnava, S., Behrendt, C. L., Ismail, A. S., Eckmann, L., Hooper, L. V., Paneth cells directly sense gut commensals and maintain homeostasis at the intestinal host-microbial interface. *Proc. Natl. Acad. Sci. USA* 2008, 105, 20858– 20863.

56 Rodríguez-Piñeiro, A. M., Post, S. V. D., Johansson, M. E., Thomsson, K. A., Nesvizhskii, A. I., Hansson, G. C., Proteomic study of the mucin granulae in an intestinal goblet cell model. *J. Proteome Res.* 2012, 11, 1879– 1890.

57 Vaishnava, S., Yamamoto, M., Severson, K. M., Ruhn, K. A. et al., The antibacterial lectin RegIII {gamma} promotes the spatial segregation of microbiota and host in the intestine. *Science* 2011, 334, 255– 258.

58 Fu, J., Wei, B., Wen, T., Johansson, M. E. V. et al., Loss of intestinal core 1–derived O-glycans causes spontaneous colitis in mice. *J. Clin. Invest.* 2011, 121, 1657– 1666.

59 Moehle, C., Ackermann, N., Langmann, T., Aslanidis, C. et al., Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. *J. Mol. Med.* 2006, 84, 1055– 1066.

60 Cilieborg, M. S., Boye, M., Sangild, P. T., Bacterial colonization and gut development in preterm neonates. *Early human development* 2012, 88, S41–S49.

61 Lichtman, J. S., Marcobal, A., Sonnenburg, J. L., Elias, J. E., Host-centric proteomics of stool: a novel strategy focused on intestinal responses to the gut microbiota. *Mol. Cell. Proteomics* 2013, 12, 3310– 3318.

62 Costello, E. K., Carlisle, E. M., Bik, E. M., Morowitz, M. J., Relman, D. A., Microbiome assembly across multiple body sites in low-birthweight infants. *mBio* 2013, 4, e00782– e00713.