Metaproteomic analysis of the infant fecal microbiome

Laetitia Cortes[1], Aude Tartière[1], Julie Piquenot[1], Sebastian Tims[2], Joost W. Gouw[2], Jan Knol[2,3], Harm Wopereis[2,3] and Daniel Chelsky[1]

[1]Caprion Biosciences Inc, Montreal, Canada; [2]Nutricia Research, Utrecht, the Netherlands;
[3]Wageningen University, Laboratory of Microbiology, Wageningen, the Netherlands

Abstract and Background

Abstract

A metaproteomics analysis was conducted on the infant fecal microbiome to characterize global protein expression in 8 samples obtained from infants with a range of early-life experiences. Samples included breast-, formula- or mixed-fed, mode of delivery, and antibiotic treatment and one set of monozygotic twins. Although label-free mass spectrometry-based proteomics is routinely used for the identification and quantification of thousands of proteins in complex samples, the metaproteomic analysis of the gut microbiome presents particular technical challenges. Among them: the extreme complexity and dynamic range of member taxa/species, the need for matched, well-annotated metagenomics databases, and the high inter-protein sequence redundancy/similarity between related members. In this study, a metaproteomic approach was developed for assessment of the biological phenotype and functioning, as a complement to 16S rRNA sequencing analysis to identify constituent taxa. A sample preparation method was developed for recovery and lysis of bacterial cells, followed by trypsin digestion, and pre-fractionation using Strong Cation Exchange chromatography. Samples were then subjected to high performance LC-MS/MS. Data was searched against the Human Microbiome Project database, and a homology-based meta-clustering strategy was used to combine peptides from multiple species into representative proteins. Bacterial taxonomies were also identified, based on species-specific protein sequences, and protein metaclusters were assigned to pathways and functional groups. The results obtained demonstrate the applicability of this approach for performing qualitative comparisons of human fecal microbiome composition, physiology and metabolism, and also provided a more detailed assessment of microbial composition in comparison to 16S rRNA.

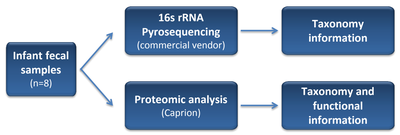
Background

→ The goal of this exploratory discovery study was to extract gut microbiome taxonomy and functional data from a small number of infant fecal samples using a multi-OMICS approach. The long-term objective is to understand the development of the microbiota of a breast-fed infant, its' influence on health and disease later in life, and how the microbiota can be modulated through nutrition.

→ 16s rRNA amplicon sequencing is a method frequently used for microbiome analysis, which allows for low-resolution identification of bacteria present in a sample.

→ A proteomics approach was also used to assess the state of the microbiome at the functional level and provide coverage of active biochemical pathways.

Figure 1: Overall study workflow



Materials and Methods

Sample Collection

Fecal samples were obtained from 8 infants between 2 to 4 months of age and living in the Netherlands. Parental written informed consent was obtained. Samples were collected into 10 ml stool containers (Greiner Bio-One) kept at 4°C, and homogenized and aliquotted within 1 hour. Aliquots were stored at −80°C until analysis by either 16S rRNA sequencing or proteomics.

Table 1: Infant Fecal Sample Characteristics

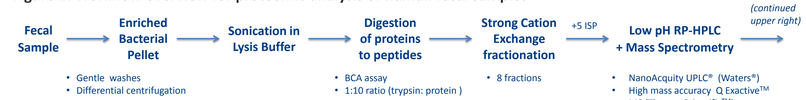
Subject ID	B	C	E	M	N	P	R	T
Gestational age (wks)	40	39	40	34	34	41	41	39
Gender	M	M	F	M	M	F	M	M
Age (days)	119	64	140	96	96	61	132	134
Delivery Mode	V	V	V	V	V	CS	V	V
Feeding Mode	BF	BF	FF	Mixed	Mixed	Mixed	BF	BF
Twin	No	No	No	Yes	Yes	No	No	No
Remarks				MZ twin#1	MZ twin#2	Vaccinated on sampling day (DKTP/HIN/HEB)	Antibiotic treatment (Amoxicillin)	
Sample weight (g)	0.56	0.23	0.52	0.51	0.58	0.55	0.53	0.51

16s rRNA Sequencing and Bioinformatics

Fecal aliquots were thawed on ice and 20 to 60 mg of each aliquot was mixed with 450 µl DNA extraction buffer (100 mM Tris-HCl, 40 mM EDTA, pH 9.0) and 50 µl of 10% sodium dodecyl sulfate. Phenol-chloroform extractions combined with bead-beating were subsequently performed as described by Matsuki, et al. [1] except that extracted DNA was re-suspended in 0.1 ml of TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). The V3-V5 regions of the 16S rRNA gene were amplified using forward primer 357°F, and a 'bifidobacteria-optimised' reverse primer 926Rb[2]. The reverse primers included a 12 base-pair error-correcting Golay barcode. PCR was carried out in quadruplicate as previously described[2]. Replicate amplicons were pooled and purified and pyrosequencing was carried out on a 454 GS FLX (Roche, Branford, CT, USA) following the Roche Amplicon Lib-L protocol. The 'Quantitative Insights Into Microbial Ecology' (QIIME) v1.9.0 package was used to analyse sequence data[3]. Sequence alignment was carried out using the SILVA rRNA database (SSU_REF111)[4] as reference. Chimera filtering, clustering at 97% sequence identity into operational taxonomic units (OTUs) and taxonomic assignment were performed using the USEARCH and UCLUST algorithms[5,6]. Rarefaction was performed and singletons were removed. The resulting taxonomic compositions (read counts and relative abundances) were summarized at the genus level.

Proteomic Analysis

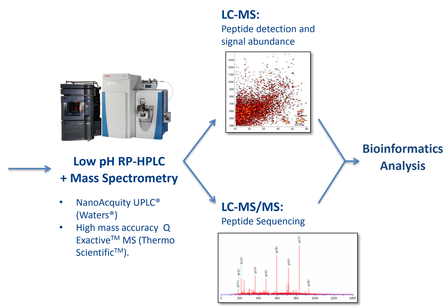
Figure 2: Workflow overview for proteomic analysis of human fecal samples



REFERENCES

1. Matsuki T, Watanabe K, Fujimoto J, et al. Quantitative PCR with 16S rRNA-Gene-Targeted Species-Specific Primers for Analysis of Human Intestinal Bifidobacteria. Applied and Environmental Microbiology. 2004;70(1):167-173. **2**. Sim K, Cox MJ, Wopereis H, Martin R, Knol J, Li M-S, et al. Improved Detection of Bifidobacteria with Optimised 16S rRNA-Gene Based Pyrosequencing. PLoS One. 2012;7(3):e32543. **3.** Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335-6. **4.** Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 2007;35(21):7188-96. **5.** Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013;10(10):996-8. **6.** Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460-1. **7.** Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. 2016;44(Database issue):D457-D462. **8.** Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. 2000;28(1):27-30.

Figure 2 (continued): Workflow overview for proteomic analysis



Bioinformatics

Protein ID and protein metaclustering: LC-MS/MS spectra were submitted to database search using Mascot software v2.2.06 (Matrix Science), with search parameters: enzyme= trypsin, allowed missed cleavages=2 , peptide tolerance= 20ppm, MS/MS tolerance= 0.05Da, variable modifications= Deamidation (N), Oxidation (M). A custom database was used which included all bacteria from the Human Microbiome Project, plus additional taxonomies identified via 16S rRNA analysis (downloaded 20140605), and Uniprot Human (reviewed entries only, downloaded 20140120). A decoy reverse database was used to evaluate false positive error rate, and Peptide/Protein Teller was used to derive the simplest list of proteins to explain observed peptides with FDR=1.6%. Identified proteins were grouped into metaclusters if they shared the same gene name and/or peptide evidence and/or the same Uniref50 identifier. For the subsequent analyses, only organisms/ genus/ metaclusters/ proteins with unique peptide evidence were considered, wherein, the 'uniqueness' of a peptide was determined independently at the corresponding protein, genus, metacluster or organism level.

Pathway analysis: Protein metaclusters with peptide redundancy removed at the genus level, were submitted to pathway analysis using the 'KEGG Mapper – Search&Color Pathway' mapping tool[7,8], and comparative analysis was performed on the different infant categories (e.g. treated with antibiotic, feeding mode, etc), (Figure 6). Pathways with obvious differences between infant categories were evaluated further by extracting the corresponding pathway genes and spectral count data (Figure 7).

Results

Pyrosequencing vs Proteomics: Identified Taxa and Relative Abundance

- Pyrosequencing identified ~30 species, proteomic analysis identified 446 species.
- The most abundant genera were readily detected by both methods and show similar rank order in terms of abundance.

Table 2: Comparison of Pyrosequencing and Proteomics results across the most abundant taxa

Low — High

	PYROSEQUENCING (Read abundance)								PROTEOMICS (Spectral counts)							
Taxonomy	**E**	**T**	**P**	**N**	**M**	**C**	**B**	**R**	**E**	**T**	**P**	**N**	**M**	**C**	**B**	**R**
Bifidobacterium	4682	3536	2014	12255	11985	13814	7943	4	4345	2394	2522	3751	3093	2182	6916	28
Veillonella	107	16	8	96	5	0	0	0	19	20	67	194	3	0	2	5
Lactobacillus	0	3	4	33	5	1	0	0	3	9	11	116	4	7	8	4
Enterobacteriaceae*	10	12	18	17	482	4	125	4400	22	22	41	110	291	4	170	3013
Enterococcus	0	0	13	5	1	0	0	0	3	7	62	42	12	0	4	2
Streptococcus	6	1	3	4	40	4	0	3	11	1	64	88	189	38	10	30
Clostridium	4	0	0	2	0	0	0	0	4	5	10	8	6	5	10	4
Parabacteroides	1	493	2396	0	0	5	0	0	2	135	758	2	2	0	2	6
Bacteroides	246	1935	0	0	9	0	0	0	230	341	213	13	24	10	19	14
Actinomyces	0	42	0	0	0	0	0	0	1	23	11	2	3	2	2	3
Haemophilus	0	1	0	0	0	0	1	0	0	0	0	0	1	0	1	4
Collinsella	13	0	0	0	0	0	0	0	318	120	131	2	2	0	3	1
Rothia	0	0	0	0	0	5	0	0	1	2	6	6	7	55	0	2
Unknown/Others	101	38	15	2	0	0	0	0	663	734	1579	530	300	145	134	132

Metaclustering of Proteins

- Proteomic analysis of the enriched bacterial pellet identified mostly bacterial sequences, and some human proteins.
- Bacteria account for ~84% of SC with useful annotation and ~88% of overall protein metaclusters.
- 50% of SC cluster to 456 known bacterial proteins.
- 39% of SC map to uncharacterized proteins.

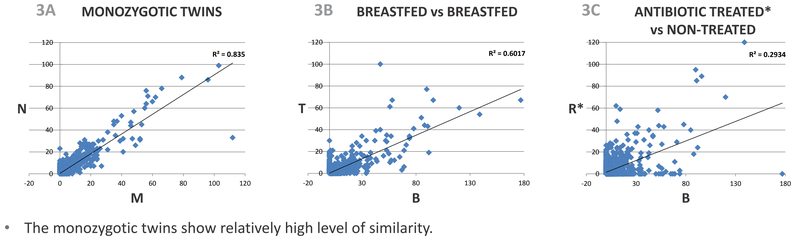
Table 3: Assignment of Spectral Counts (SC) to Organisms and Metaclusters

		# SC	# Unique peptide sequences	# Metaclusters
Bacterial	With useful annotation	9391	9145	456
	Without useful annotation	7172	6948	1289
Human	Well annotated	1800	1774	239

→ For future studies, improvements to the metaclustering algorithm will provide a higher rate of assignment of useful annotations.

Correlation of Relative Abundance of Protein Metaclusters

Figure 3: Pairwise comparisons of protein metacluster spectral counts of individual infants



- The monozygotic twins show relatively high level of similarity.
- Unrelated individuals retain some similarity for most metaclusters.
- Antibiotic treated infant is significantly different from a non-treated infant.

Genus and *Bifidobacteria* Species Distribution

Figure 4: Relative genus abundance per infant

- *Bifidobacterium* dominates all Subjects except R (+antibiotic).
- Subject R is dominated by *Klebsiella* (linked to anti-biotic associated diarrheas and inflammation).
- Mixed-fed/formula-fed infants have a more diverse microbiome.

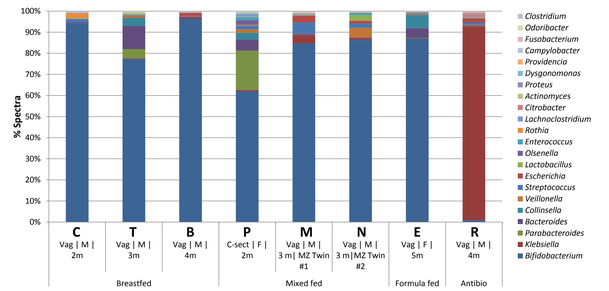
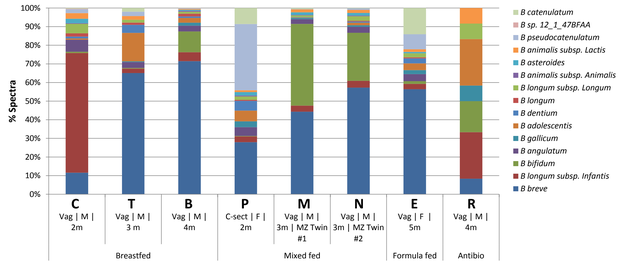


Figure 5: Relative *Bifidobacteria* species abundance per infant

- *B breve* abundance increases with infant age.
- *B. longum sub. Infantis* is favored in infants due to its ability to catabolize Human Milk Oligosaccharides non-digestible by the infant, and is associated with immune protection.
- Subject P (C-section) has a difference in the overall proportion and distribution of *Bifidobacteria.*



Comparison of Functional Activities in Antibiotic-treated vs All Non-treated

Figure 6 (right):

Kegg Pathway analysis using protein metaclusters and corresponding spectral counts

More abundant/unique to non-treated
Equivalent
More abundant/unique to treated

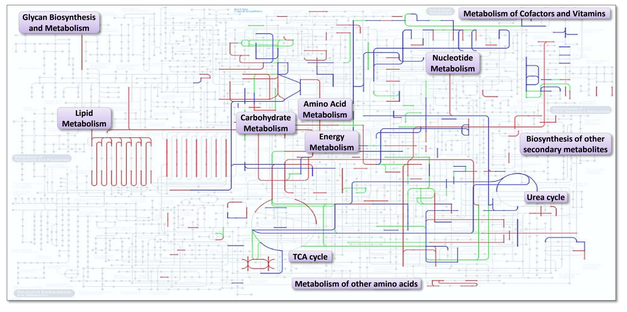


Figure 7 (below right): Examples of pathways with differential coverage and/or abundance

Non-treated
Antibiotic-treated

A) Urea cycle

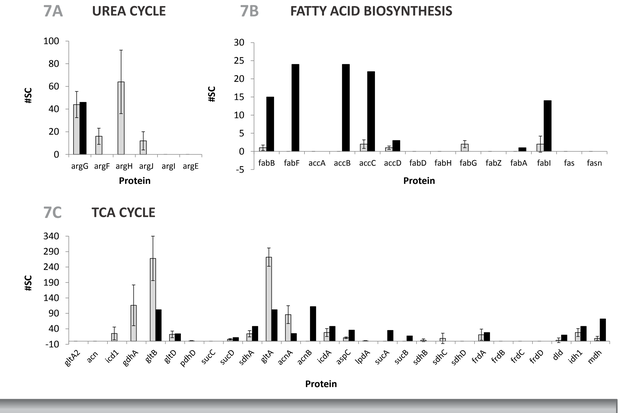
- *Bifidobacterium* well adapted to use urea, abundantly present in human milk.
- Urea is used as a source of nitrogen and is transferred to the host for protein synthesis.

B) Fatty acid biosynthesis

- *Klebsiella* not adapted to use sugars present in milk.
- Relies on fat instead?

C) TCA cycle

- At the crossroad of two previous pathways.
- Adjusted to reflect differences in main source of energy? In ability to fix nitrogen?



Conclusions

→ Proteomics allows insights into functional changes at a given point in time.
→ Proteomics analysis allows combining species composition and protein level status information.
→ Bacterial taxonomy identification and relative quantification.
  - Based on species-specific protein sequences.
  - Comparable results to pyrosequencing.
→ Bacterial protein identification and relative quantification.
  - Meta-clustering creates cross-species assessment of bacterial metabolic status.
  - Provides insights into functional response to diet, disease and therapy.