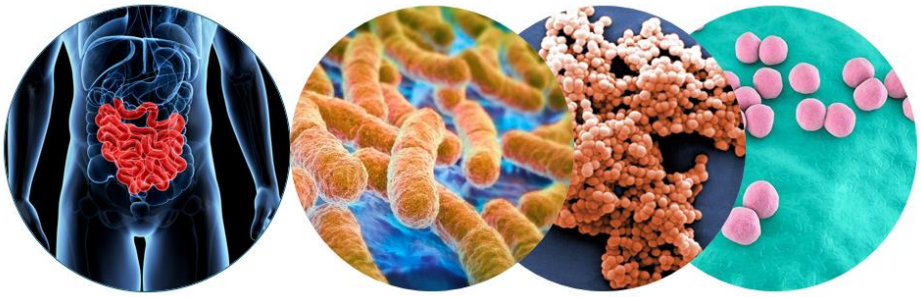
Meta-omics of the human small intestines to explore functional and compositional diversity

MSc. minor thesis report



Written by:

Lotte Witjes, BSc.[1]

registration no. 950405966120

Supervised by:

Jasper Koehorst, MSc.[1]

dr. Peter Schaap[1]

[1] Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Helix, Stippeneng 4, 6708 WE Wageningen, The Netherlands

Contents

Abstract

The human gut microbiota, in particular that of the colon, have previously been shown to play a significant role in the host's health and homeostasis. Compositional imbalances are known to participate in disease onset and development. Knowing the importance of the colonic microbiome, it would be of higher interest to examine the small intestinal microbiome considering the organ's crucial act in metabolism and immunity.

Here, the ileal microbiomes of four ileostomy subjects across four timepoints have been explored in terms of function and composition. Their metatranscriptomes were functionally analyzed by comparing protein domain and enzyme abundances and, besides functionality, the metatranscriptomic reads provided insight in the phage content. The metagenomes of the ileal samples were compared using genera and species counts.

The core domainome and enzymome were enriched for pathways and processes involved in growth, maintenance and carbohydrate metabolism and uptake. With regard to the composition, the following genera were present in all subjects: *Clostridium*, *Escherichia*, *Streptococcus*, *Ruminococcus*, *Romboutsia*, *Haemophilus*, *Eubacterium*, *Lachnospira* and *Mycobacterium*. Remarkably, the genus *Veillonella* was absent in the core. Phages known to infect *Escherichia*, other *Enterobacteria* and *Streptococcus* were found in all subjects. Species, phage, and functional diversity varied greatly between and within subjects. An increasing functional diversity could partly be explained by an increasing species diversity. Furthermore, a negative correlation was found between species and phage diversity. Differences between subjects in terms of functionality and composition were thought to be mainly due to dietary and other lifestyle-concerning differences. Still, the relative contribution of species of unknown/uncultured genera was strikingly high, expressing the need to further research the small intestinal microbiota.

Concludingly, it was shown here that the small intestinal microbiota differed between subjects and even fluctuated within subjects both functionally and compositionally. These observations could be explained by e.g. the rapid small intestinal luminal flow and the effect of diet and lifestyle. Changes in functional diversity could be partly explained by changes in species diversity. The commensal relationship concerning lactate production and consumption might actually involve other genera than previously thought.

Introduction

The composition of the colonic microbiota has already been associated with a wide variety of diseases. Compositional alterations have been shown to have profound roles in the onset, development and severity of the following inflammatory bowel diseases (IBD): Crohn's disease, ulcerative colitis, and inflammatory bowel syndrome (IBS) (Becker et al., 2015). The intestinal microbial community is greatly affected by our diet (Patman, G., 2015) and lifestyle, and this is also reflected by the fact that the intestinal microbiota has been linked with so-called lifestyle diseases e.g. colorectal cancer (Irrazábal et al., 2014), obesity, and diabetes (Miele et al., 2015). Even the gut-brain axis is influenced by gut microbiota as links were found with autism and depression (Dinan et al., 2015).

Since most of these studies are based on phylogenetic analyses of colonic or fecal samples, little is known about the small intestinal microbial composition or its functional landscape. Three regions can be distinguished in the small intestines in this order: the duodenum, jejunum, and ileum. The duodenum is connected with the stomach separated by the pyloric sphincter. Bile and pancreatic enzymes enter the duodenum via the ampulla of Vater and therefore the duodenum is the main place of enzymatic breakdown of food. The jejunum has a very large surface area through the presence of (micro)villi, making it the ideal location for absorption of nutrients. The ileum is he small intestinal terminal part and its main function is inducing immune responses reflected by the presence of gut-associated lymphoid tissue (GALT) and Peyer's patches (Aidy et al., 2015).

The small intestinal bacterial density is less, gradually increasing from 10^3 to 10^8 bacteria g^{-1}, compared to the colonic density (10^{12} bacteria g^{-1}). This is caused by rapid luminal flow, high acidity (originating from the stomach), secretion of bile and production of antimicrobial peptides. The latter serve to protect the highly permeable intestinal epithelium from putative pathogens. The small intestinal microbiota consists predominantly of *Streptococcus*, *Veillonella*, *Escherichia*, *Clostridium* and *Turicibacter* species, where the first two always coexist in a commensal relationship with, respectively, lactate production and consumption (Aidy et al., 2015).

Considering the significance of the small intestines to the host's metabolism and immune system, it might be of much more interest to explore the small intestinal microbiota functionally and phylogenetically. However, sampling of the small intestines is still an obstacle to overcome. Fasting, which affects the microbial composition, is required before sampling using naso-ileal catheters. Moreover, placement of these catheters needs supervision of gastro-enterologists further complicating sampling. Another means of obtaining small intestinal samples is to cooperate with ileostomy subjects enabling sampling over time (Aidy et al., 2015).

Metagenomic analysis on ileal samples in patients with Crohn's disease has already shown associations between depletion of specific Firmicutes and Bacteroidetes taxa, and expansion of Proteobacteria and the severity of the disease (Haberman et al., 2014). Furthermore, in the study of Zoetendal et al., from 2012 it was shown that *Streptococcus*, *Escherichia*, *Clostridium* and Gram-positive organisms with high GC-content are most abundant in the small intestines. Genes related to uptake (phosphotransferase system) and metabolism of simple carbohydrates were highly expressed in the small intestinal samples. It was concluded that the ileal microbial community is depending on the ability of fast uptake and metabolism of simple carbohydrates, and adaptation to fast changing nutrient availability (Zoetendal et al., 2012).

A metatranscriptomics analysis pipeline has already been proposed and tested on a part of the data that was used in this study (Leimena et al., 2013), a BLASTX-based method was used successfully to touch upon the functionality of the small intestinal microbiome. Recently, (Meta)SAPP was developed for de novo-based functional annotation of metatranscriptomes. SAPP (Semantic Annotation Platform with Provenance) was designed to support FAIR *de novo* computational genomics but can also be used to process and analyze existing genome annotations. Modules are available for prediction of genetic elements and protein annotation. Besides annotation information, it also stores associated dataset- and element-wise provenance by using a HDT (Header, Dictionary, Triples) format. SPARQL querying can be used to interrogate the SAPP databases (Koehorst et al., 2018).

This study aimed to further explore functional and phylogenetic diversity in the human small intestines by making use of ileostoma effluent samples. Therefore, here, BLASTX-, BLASTN- and *de novo*-based methods were compared for functional analysis of metatranscriptomics data of the human ileum. Hereafter, the method considered most suitable was used to identify the core domainome and enzymome across subjects, to identify the differences in domainome and enzymome between subjects, to compare functional and compositional diversity over time and to identify bacterial phages affecting compositional and/or functional diversity.

Methodology

The methodology encompasses sections on: data description and preprocessing, comparison of three methods, and both phylogenetic and functional analyses to analyze metatranscriptomics data. Whenever settings are not specified for the tools used, default settings were used. The workflow, specific commands and SPARQL queries can be found below in **appendices 1, 6 and 7**. Most graphs were made with ggplot2 (Wickham, 2009) in the R programming language (R Core Team, 2017), whereas the rest of the methods have been implemented in the Python programming language (Python Software Foundation,

http://www.python.org/). Other than ggplot2, the following R packages have been used: reshape2 (Wickham, 2007), ggfortify (Tang et al., 2016), VennDiagram (Chen, 2018) and vegan (Oksanen et al., 2018).

Data description and preprocessing

The dataset included 16S rDNA, 16S rRNA and mRNA profiling where the first two were sequenced with pyrosequencing and the latter with Illumina sequencing. All profiling was done for two male and two female ileostomy subjects (age 66 ± 9.0 years) for four timepoints. The subjects were colectomized at least five years prior to the sampling period, are clinically considered to be healthy, and have normal functioning small intestines. Ileostoma effluent was sampled on day one and day three both in the morning and afternoon. As to the Illumina sequencing, all samples were sequenced in single end, however paired end sequencing was applied for female subject four. Replicates for two samples (male subject one, day three morning and female subject 2, day one morning) were sequenced at a lower coverage. During the sampling period, no standardized diet was prescribed to the subjects. The datasets of male subject one and female subject four were already analyzed and published in a study in which the effect of lower coverage sequencing and single-end versus paired-end sequencing was investigated to propose a metatranscriptome analysis pipeline (Leimena et al., 2013). Here, only the 16S rRNA and mRNA profiling datasets were used. An overview of the data can be found in **appendix 2**.

All raw FASTQ files of the mRNA profiling were subjected to FastQC v.0.11.7 (Andrews S., 2010) to assess quality before and after trimming. Adaptors and low quality reads were removed or trimmed with Trimmomatic v0.36 (Bolger et al., 2014) with default settings for both single end and paired end reads with the Illumina TruSeq3 adaptors file. Reads originating from rRNA were removed with SortMeRNA v2.1b (Kopylova et al., 2012) by using the default settings and the following included SILVA rRNA databases (Quast et al., 2013): bacterial 23S, archaeal 16S, archaeal 23S, eukaryotic 18S, eukaryotic 28S, rfam 5S and rfam 5.8S. Reads originating from human mRNA were removed with the latter software program as well by using the default settings and a database of transcripts of the human genome version GRCh38.p12.

All raw FASTQ files of the 16S rRNA profiling were subjected to FastQC v.0.11.7 (Andrews S., 2010) to assess quality before and after trimming. Low quality reads were removed or trimmed with QTrim v1.1 (Shrestha et al., 2014) with default settings.

Comparison of BLASTX-, BLASTN- and *de novo*-based methods

The mRNA profiling datasets were analyzed with three different approaches followed by comparison. Hereafter, the three different approaches will be explained. All approaches led to Pfam domain and EC number counts per sample.

The trimmed non-human mRNA reads were queried with DIAMOND v0.9.21.122 (Buchfink et al., 2015) against the proteins (a total of 56,263,754 unique protein sequences) in the database containing approximately 100,000 bacterial genomes annotated with SAPP (Koehorst et al., 2018). The genetic code specific for Bacteria, Archaea and plant plastids was used for translation and only one target sequence was kept. Thresholds were applied for the DIAMOND results, an e-value lower than 10^{-6} and a bit score higher than 74, adapted from what was found in the study of Leimena et al., in 2013. Protein counts were converted to Pfam domain (database v31.0) (Finn et al., 2014) and EC number (Bairoch A., 2000) counts by using the functional annotation information predicted by InterProScan v5.25-64.0 (Jones et al., 2014) and EnzDP v1.0 (Nguyen et al., 2015) of the SAPP-based annotated bacterial genomes. Thresholds were set for the likelihood score and max bit score of EnzDP, 0.1 and 74 respectively. Furthermore, only full EC number were kept. The default thresholds for InterProScan were used.

For the BLASTN-based method, trimmed non-human mRNA reads were queried with MegaBLAST enclosed within the NCBI BLAST+ package v2.7.1 (Camacho et al., 2009) against the genes (a total of 150,492,747 unique gene sequences) in the database containing approximately 100,000 bacterial genomes annotated

with SAPP (Koehorst et al., 2018). Again, only one target sequence was kept per query read and gene counts were converted to Pfam domain and EC number counts as described for the BLASTX-based method. The same BLAST thresholds were applied for e-value and bit score, as well as the same EnzDP and InterProScan thresholds.
At last, trimmed non-human mRNA reads were functionally annotated with MetaSAPP (Koehorst et al., 2018). The following SAPP modules were applied in this order: MEGAHIT v1.1.2 (Li et al., 2014) for metatranscriptomic assembly, Prodigal v2.6.3 (Hyatt et al., 2010) for bacterial gene prediction, InterProScan v5.25-64.0 (Jones et al., 2014) for Pfam domain prediction and EnzDP v1.0 (Nguyen et al., 2015) for enzyme prediction. Prodigal was run using the genetic code for Bacteria, Archaea and plant plastids with the option `-meta` specific for metatranscriptomes (and metagenomes). All other modules were run with default settings. The predicted genes and their sequences were extracted by a SPARQL query, against which the trimmed non-human mRNA reads were blasted with BLASTN contained within the NCBI BLAST+ package v2.7.1 (Camacho et al., 2009) to extract gene counts. The latter were converted to Pfam domain and EC number counts as previously described. Again, the same BLAST thresholds were applied for e-value and bit score, as well as the same EnzDP and InterProScan thresholds.
Pfam domain and EC number count matrices resulting from the three different methods were compared with a principal component analysis (PCA) on the count matrices and visualized with Venn-diagrams for the presence/absence matrices. Rarefaction analyses provided insight in sequencing depth. The most suitable method for this dataset was chosen based on the following rules. The method should not be an outlier in the PCA, and the highest percentage of Pfams and ECs must have been found with the other two methods as well.

Functional analysis: protein domains and enzymes

After choosing the most suitable method for this dataset, the core domainome and enzymome (those Pfam domains and EC number present at least once in all subjects) were determined. The core Pfam domains were extracted and mapped against Gene Ontology (GO) biological process terms (The Gene Ontology Consortium, 2017) with the Pfam2GO (May 2018) mapping generated from data supplied by InterPro (Hunter et al., 2009). The GO biological process terms with the most Pfam domains (with a minimum of 10 Pfam domains) mapped were considered the core domainome. The core EC numbers were extracted and mapped against KEGG pathways. A hypergeometric test with Benjamini-Hochberg (BH) multiple testing correction was performed to determine significantly (adjusted p-value smaller than 0.05) enriched KEGG pathways (Kanehisa et al., 2000) in the core enzymome.
The functional differences between the subjects' small intestinal microbiota were identified as follows. The Pfam domains and EC numbers present in only one of the subjects were extracted and sorted based on the number of times the domains and enzymes were present in the subject. Those Pfam domains and EC numbers, that are uniquely present within a subject and were having the highest counts, were further investigated. Finally, the functional diversity within and between the subjects were analyzed by calculating and visualizing the domain and enzyme Shannon index and richness on Pfam domain and EC number counts. A Pearson's correlation coefficient (PCC) was calculated for the Shannon indexes and richness.

Compositional analysis: species and phages

The 16S rRNA profiling datasets were queried with MegaBLAST embedded in the NCBI BLAST+ package v2.7.1 (Camacho et al., 2009) against the SILVA v132 SSU Ref NR 99 database (Quast et al., 2013). Only the best hit per read was kept and a threshold of 95% and 97% for the sequence similarity was applied to filter the MegaBLAST results, for respectively genera and species. Species diversity was analyzed by calculating the Shannon index and richness over time on species counts. Microbial composition on genus level was visualized in a relative stacked bar plot with each sample represented by a different bar.

The trimmed non-human mRNA datasets were queried as described above with MegaBLAST against a bacterial phages database. The database was made by collecting all phages with Bacteria as hosts from NCBI (link) and extracting their RefSeq coding nucleotide sequences in FASTA format (link). Again only the best hit per read was kept and thresholds for e-value and bit score were applied, respectively, lower than 10^{-6} and higher than 110. Phage diversity was analyzed and visualized by calculating the Shannon index and richness over time on phage counts.
For both species and phage count data, PCAs and Venn-diagrams were calculated and constructed for comparison between subjects. A PCC was calculated for the Shannon indexes and richness. Furthermore, the top five most abundant species and phages uniquely present per subject were analyzed.

Source code

All written code can be found on https://github.com/lottewitjes/MSc_minor_thesis.

Results

Data preprocessing

A total of 666,717,901 raw RNA reads were generated during the metatranscriptomic profiling. After adaptor and low quality trimming 554,799,063 reads were left. However, after the removal of reads originating from ribosomal RNA and human mRNA, a total number of 61,785,193 reads turned out to be non-human mRNA reads of sufficient quality. More information, regarding read numbers, can be found in **supplemental table 1 of appendix 2**. **Figure 1** shows a visual representation of the read counts per sample. It can be concluded that the number of mRNA reads greatly varied between the different subjects as well as within subjects, furthermore a substantial number of reads was discarded due to low quality and rRNA origin.
The number of trimmed 16S rRNA reads ranged from 4,539 to 24,467 per sample with a total of 222,231 reads. Here, again, the total number of reads varied significantly between samples and more information can be found in **supplemental table 2 of appendix 2**.

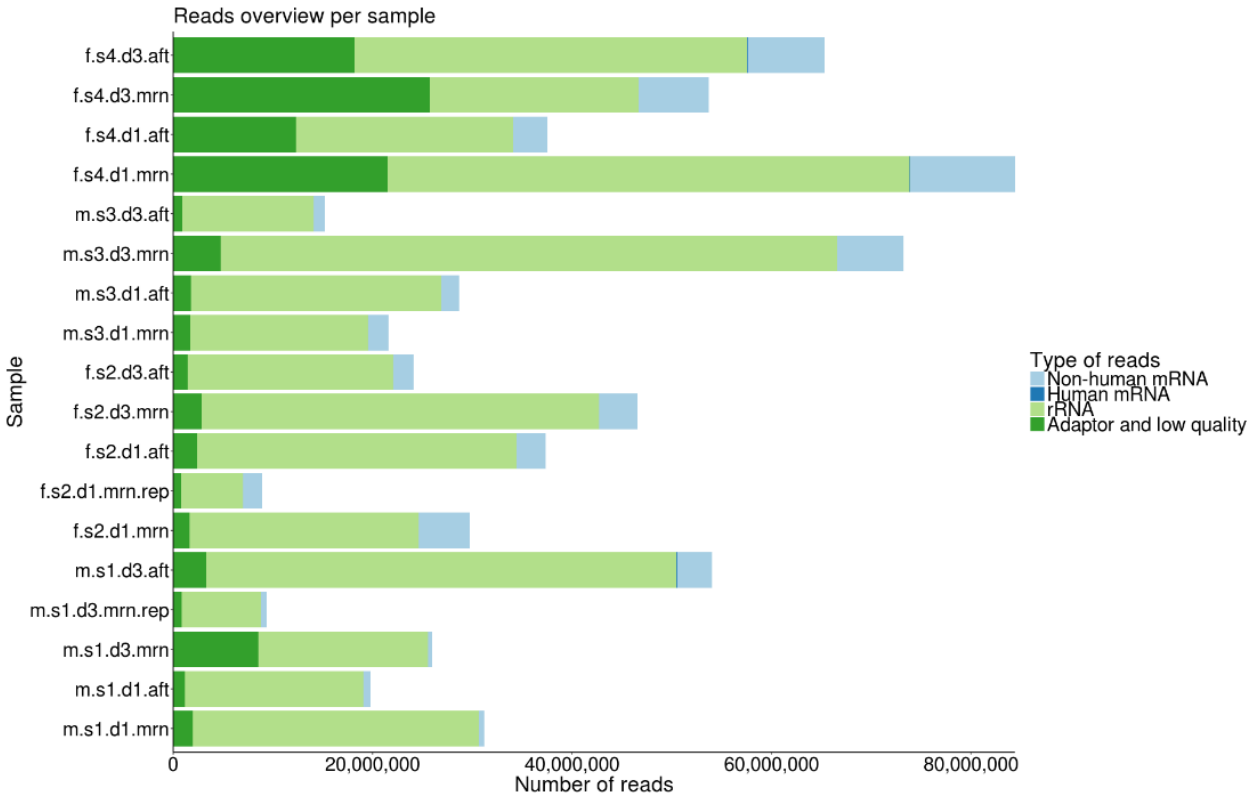


Figure 1 An overview of the number of reads per sample after adaptor and low quality trimming (dark green), after removal of rRNA (light green), and after removal of human mRNA (dark blue). The light blue coded part of the bar are the reads that are effectively used. The coding of the samples is as follows: gender, subject number, day number and day part. Those samples containing "rep" are technical replicates sequenced at lower depth.

Comparison of BLASTX-, BLASTN- and *de novo*-based methods

Three different methods were compared for functional analysis of the mRNA profiling datasets: a BLASTN-based, a BLASTX-based and a *de novo*-based method. The methods were compared based on the resulting Pfam domain and EC number count matrices.

Figure 2 shows the results of the PCAs on both count matrices (**figure 2A**: Pfam domains, **figure 2B**: EC numbers). Individual samples analyzed with the different methods were plotted on the first and second principal component (PC1 and PC2). For both PCAs and their first two PCs, the accompanying proportions of variance explained (PVE) were 10.77% and 7.39% for Pfam counts and 13.98% and 8.5% for EC number counts. This suggested that the variance between the methods could be better explained by EC numbers, and that the datasets showed variability. Clustering of the samples analyzed with the same method can be observed in both figures, but is more prominent in **figure 2B**. The latter graph implied that the results of the BLASTN- and BLASTX-based methods were more alike than the MetaSAPP results, indicating that the latter *de novo*-based method might be an outlier.

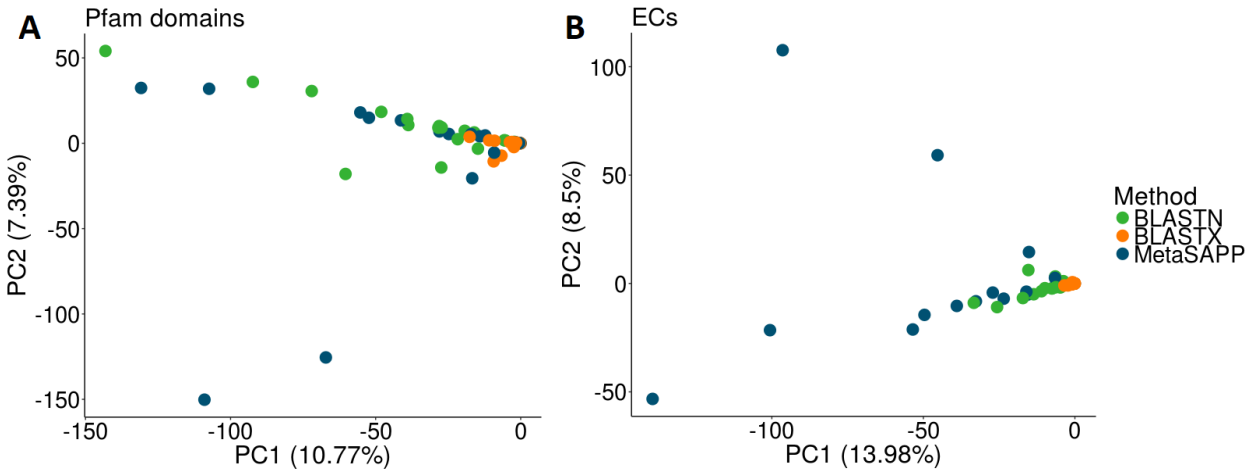


Figure 2 PCAs on Pfam domains (**A**) and EC numbers (**B**) visualized by plotting the second principal component (PC2) against the first principal component (PC1) with the accompanying proportions of variance explained (PVE). The green dots represent the samples analyzed with BLASTN, the orange dots with BLASTX, and the dark blue dots with MetaSAPP.

When zooming in on the unique Pfam domains and EC numbers that were found by each method, again dissimilarities can be noticed. **Figure 3** shows the Venn-diagrams representing the Pfam domains (**figure 3A**) and EC numbers (**figure 3B**) found with each method. Ample Pfams and ECs are found by all methods, respectively 3,658 and 1,184. Still, significant numbers of Pfams and ECs are only found with one method. BLASTX had the highest percentage of Pfam domains found also covered by other methods, namely 92.75%, compared to 90.58% and 87.13% for BLASTN and MetaSAPP, respectively. In addition, BLASTX had the highest percentage of EC numbers found also covered by other methods, namely 99.31%, compared to 93.37% and 63.59%. This might suggest that the BLASTX-based method had the least false positives in comparison with the other methods. The latter and the fact that BLASTX was not an outlier in the previous PCAs, made BLASTX appear the most suitable method to functionally analyze the metatranscriptomes in this study.

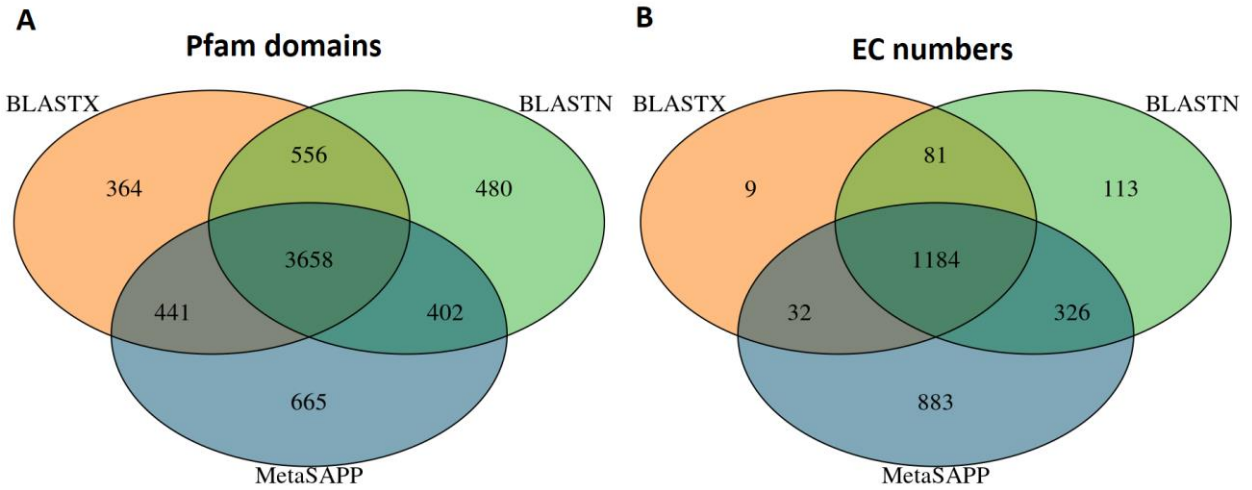


Figure 3 Venn-diagrams representing the Pfam domains (**A**) and EC numbers (**B**) found with the three methods. The numbers in overlapping circles represent Pfam domains or EC numbers found with both or all methods. The orange circle represents BLASTX, the green circle BLASTN, and the blue circle MetaSAPP.

Finally, rarefaction analyses were carried out for BLASTX results to assess sequencing depth. **Figure 4** shows plots of the number of unique Pfams (**A**) and EC numbers (**B**) plotted against the number of usable

mRNA reads. The black dots represent the samples and an asymptote is expected, meaning that eventually a platform will be reached where no new unique Pfam domains and EC numbers are found with increasing sequencing depth (number of reads). Still, for both graphs the plotted exponential functions (blue line with gray area) have not yet reached the expected platform. Therefore, it can be concluded that the sequencing depth was not sufficient in this study. A significant number of Pfam domains and EC numbers might have been missed.

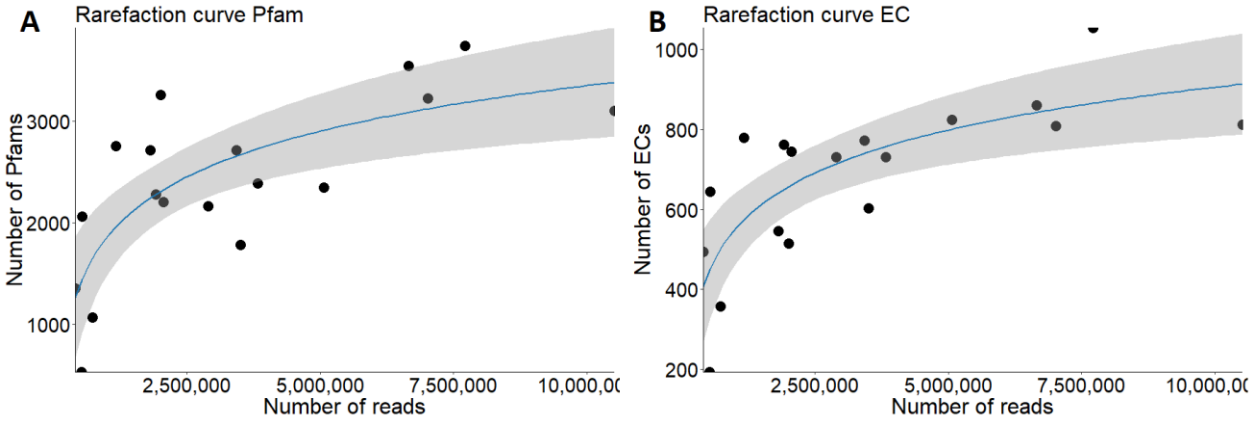


Figure 4 Rarefaction analyses of Pfam domains (**A**) and EC numbers (**B**) analyzed with BLASTX. For both figures, the number of unique Pfam domains and EC numbers were plotted against the number of mRNA reads. The black dots represent the different samples on which an exponential function is fitted (blue line with grey area).

Functional analysis: protein domains and enzymes

Having chosen the most suitable method to functionally analyze the subjects' metatranscriptomes, namely BLASTX, it is now time to dig deeper into the (dis)similarities between the subjects and timepoints in terms of function. The metatranscriptomes were compared based on the resulting Pfam domain and EC number count matrices.

Figure 5 shows the results of the PCAs on both count matrices (**figure 5A**: Pfam domains, **figure 5B**: EC numbers). The individual samples colored per subject were plotted on the first and second principal component (PC1 and PC2). For both PCAs and their first two PCs, the accompanying proportions of variance explained (PVE) were 21.05% and 12.79% for Pfam counts and 22.06% and 10.14% for EC number counts. Contrary to the PCA on the different methods, this suggested that the variance between the subjects could be better explained by Pfam domains, and that the datasets were less variable between subjects compared to between methods. Clustering of the samples of the same subject can be observed in both figures, but is more prominent in **figure 5A**. The latter graph implied that the metatranscriptomes of subject one, two and four were more similar to each other than to subject three, with subject three being an outlier. However, this observation is less obvious in **figure 5B**, where clustering appeared to a lesser extent.

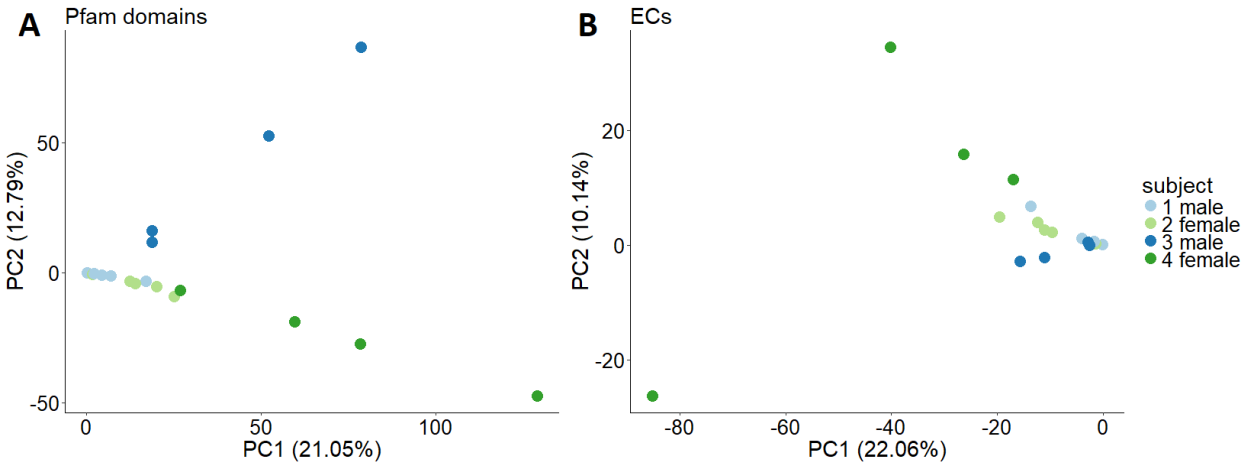


Figure 5 PCAs on Pfam domains (**A**) and EC numbers (**B**) visualized by plotting the second principal component (PC2) against the first principal component (PC1) with the accompanying proportions of variance explained (PVEs). The differently colored dots represent the different subjects: light blue for male 1, light green for female 2, dark blue for male 3, and dark green for female 4.

When zooming in on the unique Pfam domains and EC numbers that were found in each subject, again dissimilarities can be noticed. **Figure 6** shows the Venn-diagrams representing the Pfam domains (**figure 6A**) and EC numbers (**figure 6B**) found within each subject. Total numbers of 5019 Pfam domains and 1306 EC numbers were found, of which a generous number of Pfams and ECs were found in all subjects' metatranscriptomes, respectively 2,290 and 734. These can be seen as the core domainome and core enzymome, and will be analyzed and described later. More Pfam domains are found in comparison to EC numbers due to the possible presence of multiple domains in one enzyme. Apart from what is similar amongst the subjects, it can be noticed in **figure 6** that a lot of Pfam domains and EC numbers were uniquely present in different subjects. A total number of 66, 55, 485 and 434 Pfam domains were found uniquely in respectively subject one, two, three and four. For the EC numbers, a total number of 21, 27, 54 and 129 enzymes could be found uniquely in respectively subject one, two, three and four. These observations indicated that the metatranscriptomes of the different subjects varied.

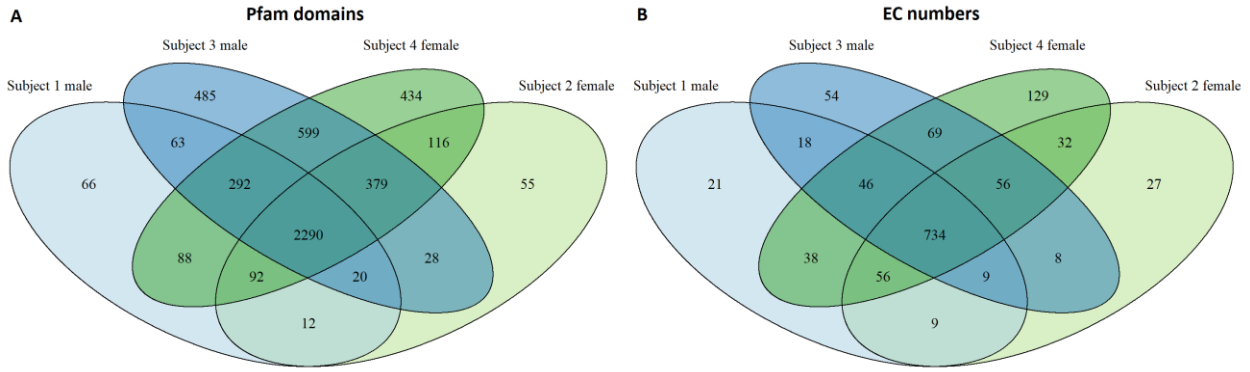


Figure 6 Venn-diagrams representing the Pfam domains (**A**) and EC numbers (**B**) found within the subjects' metatranscriptomes. The numbers in overlapping circles represent Pfam domains or EC numbers found within two, three or all subjects. The differently colored circles represent the different subjects: light blue for male 1, light green for female 2, dark blue for male 3, and dark green for female 4.

Knowing that the small intestines have a rapid luminal flow and therefore might have a highly fluctuating microbial composition, it is interesting to see if this was also reflected in the functional diversity over time. **Figure 7** shows graphs of the Shannon diversity and richness of Pfam domains and EC numbers over time. It can immediately be noted that the functional diversity and richness was varying in all subjects, and

interestingly varied more in subject one, three and four. The trends observed for Pfam domains (for both Shannon diversity and richness) matched to some extent the trends observed for EC numbers. No clear patterns regarding morning and afternoon rhythm could be seen. The previous observation that there are more Pfam domains than EC numbers is reflected in these graphs too. **Supplemental figure 2 and 3 in appendix 5** show networks of PCCs calculated from Shannon indexes and richness for Pfam domains, EC numbers, species and phages. The PCC for correlation between Shannon Pfam domain diversity and Shannon EC number diversity was 0.7996, indicating that an increase in domain diversity resulted in an increase in enzyme diversity. The PCC for correlation between Pfam and EC richness was 0.8020, reflecting the same relationship.

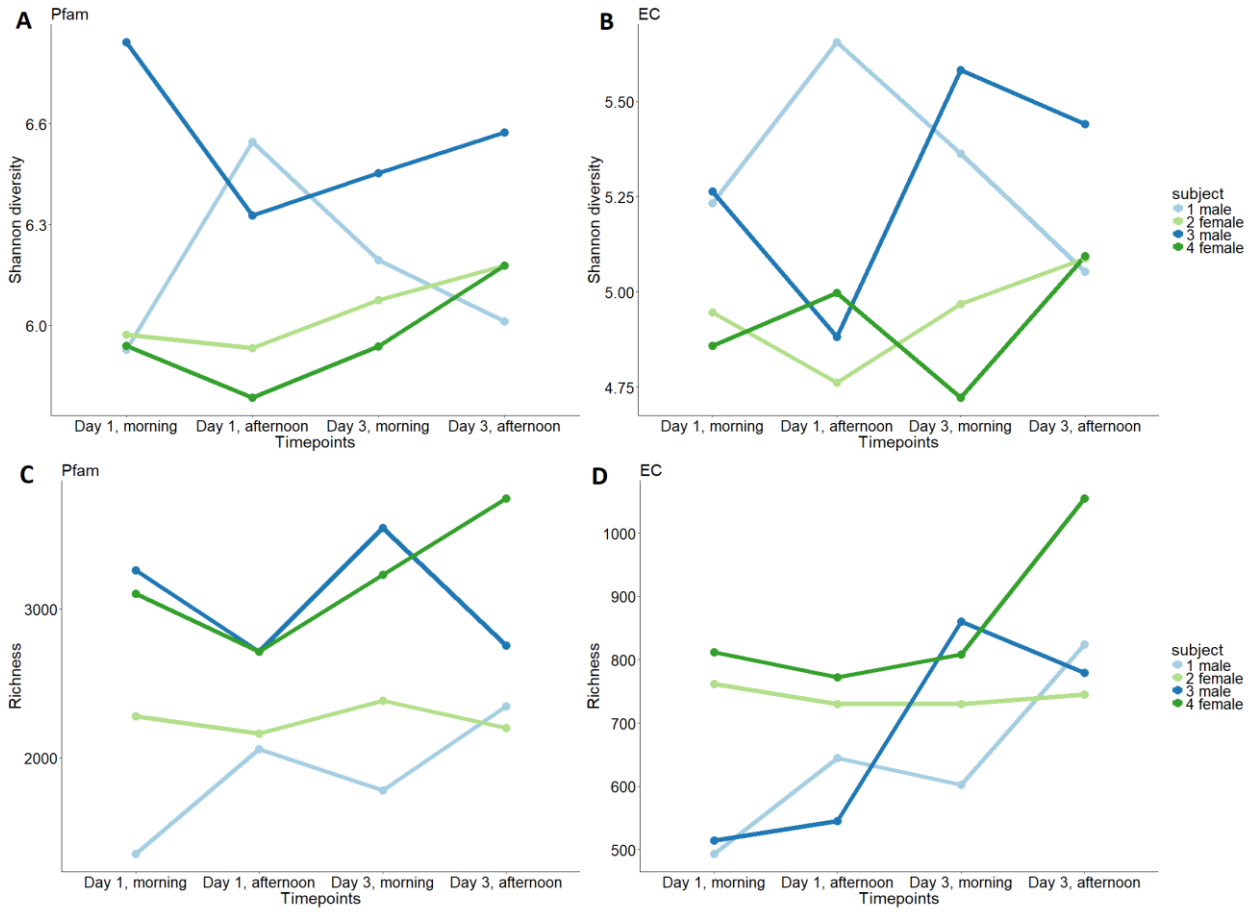


Figure 7 Shannon diversity and richness over time of the different subjects' metatranscriptomes based on Pfam domain and EC number counts. Figure **A** and **B** visualize the changes in Shannon diversity for Pfam domains and EC numbers over time, whereas figure **C** and **D** visualized the changes in richness. The differently colored lines represent the different subjects: light blue for male 1, light green for female 2, dark blue for male 3, and dark green for female 4.

The core domainome and enzymome consisted of respectively 2,290 and 734 Pfam domains and EC numbers. The core Pfam domains were mapped against GO biological process terms and those terms with a number of mapped Pfam domains higher than ten are shown in **table 1**. The core enzymes were mapped on KEGG pathways and tested for enrichment with a Benjamini-Hochberg corrected hypergeometric test, the significantly enriched pathways are shown in **table 2**. As was shown in the study of Zoetendal et al., from 2012, the core was enriched for processes and pathways related to carbohydrate uptake and metabolism. Moreover, processes and pathways related to growth and maintenance were enriched as

well. More specialized pathways were detected in the core too, e.g.: streptomycin biosynthesis, selenocompound metabolism, and pantothenate and CoA biosynthesis.

Table 1 The core domainome across all four subjects mapped against GO biological process terms, those GO terms with a minimal of ten mapped Pfam domains are shown.

GO biological process	Mapped Pfam domains
Oxidation-reduction process	120
Carbohydrate metabolic process	47
Metabolic process	42
Regulation of transcription, DNA-templated	34
Transmembrane transport	33
Proteolysis	30
Translation	28
DNA replication	19
DNA repair	15
Transcription, DNA-templated	15
Biosynthetic process	14
Phosphoenolpyruvate-dependent sugar phosphotransferase system	13

Table 2 The core enzymome across all four subjects mapped against KEGG pathways and tested for enrichment with a Benjamini-Hochberg corrected hypergeometric test. Only significantly enriched KEGG pathways are shown.

KEGG pathway	Mapped ECs	ECs in pathway	BH-corrected p-value
Peptidoglycan biosynthesis	17	19	$1.345e^{-09}$
Aminoacyl-tRNA biosynthesis	23	31	$2.019e^{-09}$
Pyrimidine metabolism	31	65	$1.893e^{-05}$
Arginine biosynthesis	17	28	$3.197e^{-05}$
Purine metabolism	43	110	$1.226e^{-04}$
Glycolysis/gluconeogenesis	23	48	$1.900e^{-04}$
Pyruvate metabolism	29	67	$2.175e^{-04}$
Alanine, aspartate and glutamate metabolism	23	50	$3.067e^{-04}$
Carbon fixation pathways in prokaryotes	23	50	$3.067e^{-04}$
Other glycan degradation	7	9	$3.778e^{-04}$
Starch and sucrose metabolism	30	77	$1.239e^{-03}$
Lysine biosynthesis	14	28	$1.621e^{-03}$
Valine, leucine and isoleucine biosynthesis	8	14	$5.278e^{-03}$
Drug metabolism – other enzymes	12	25	$5.278e^{-03}$
Citrate cycle (TCA cycle)	12	25	$5.278e^{-03}$
Fatty acid biosynthesis	9	17	$6.004e^{-03}$
D-Alanine metabolism	4	6	$1.489e^{-02}$
Galactose metabolism	18	48	$1.925e^{-02}$
Metabolic pathways	384	1649	$2.030e^{-02}$
Phenylalanine, tyrosine and tryptophan biosynthesis	15	39	$2.410e^{-02}$
Streptomycin biosynthesis	8	18	$3.384e^{-02}$
One carbon pool by folate	10	24	$3.384e^{-02}$
Selenocompound metabolism	8	18	$3.384e^{-02}$
Pantothenate and CoA biosynthesis	12	31	$3.735e^{-02}$

Supplemental table 3 and 4 of appendix 3 show the top five uniquely found Pfam domains and EC numbers per subject, sorted on average count. The unique Pfam domains and EC numbers of subject one were not considered due to low average abundance. The unique Pfam domains will be analyzed first. A sialidase enzyme penultimate C terminal domain (PF12135), uniquely observed in subject two, is found in

the toxin-producing *Clostridium perfringens* (Adams et al., 2008). Furthermore, a tetracycline repressor C-terminal all-alpha domain (PF16295) was observed, matching the fact that the most abundant genera in the small intestines are resistant to tetracyclines (Grossman, 2016). Lastly for subject two, an ADP-ribosyltransferase exoenzyme (PF03496) was found which is known to be involved in the production of the iota-toxins in *C. perfringens* (Tsuge et al., 2003). Subject three had some unique Pfam domains related to carbohydrate metabolism. A glycogen synthesis protein domain (PF089761) which is known to be expressed in *Escherichia coli* in response to starvation (Kozlov et al., 2004). A maltose transport system permease protein MalF P2 domain (PF14785), and an oligogalacturonate-specific porin protein (PF06178) involved in the uptake of pectin derivatives (Blot et al., 2002). The metatranscriptome of subject four uniquely contained the domain of CIA30 (PF08547), this domain is present in the mitochondrial complex I in human and mouse. Interestingly, the domain family is also present in *Schizosaccharomyces pombe* (Janssen et al., 2002). A bacteriocin domain related to lactococcin 972 (PF09683) was found as well as a CRISPR-associated protein Csn2 subfamily domain (PF16813). Lactococcin 972 is known to be produced by *Lactococcus lactis* and is bactericidal to sensitive strains (Martínez et al., 1996), whereas Csn2 subfamily domains are found in e.g. *Streptococcus* and *Enterococcus* species and play a role in prokaryotic immunity (Lee et al., 2012).

Having analyzed the unique Pfam domains, let's now focus on the unique EC numbers per subject. In the metatranscriptome of subject two appeared a cyanophycin synthase (EC 6.3.2.30) which functions in nitrogen storage in Cyanobacteria, however recent studies found this enzyme to be expressed in *C. perfringens* and might there play a role in spore assembly (Lui et al., 2016). Lastly, subject three had CDP-abequose synthases (EC 1.1.1.341) which are expressed in *Yersinia pseudotuberculosis* (Kessler et al., 1991) and *Salmonella enterica* (Wyk et al., 1989) conferring antigen specificity.

Compositional analysis: species and phages

After functionally analysis of the subjects' metatranscriptomes with BLASTX, it is now time to dig deeper into the (dis)similarities between the subjects and timepoints in terms of composition. The metagenomes and metatranscriptomes were analyzed with BLASTN against the latest version of the SILVA SSU rRNA database and a phage database, and were compared based on the resulting genera, species and phage count matrices.

Figure 8 shows the relative contribution of the genera per sample. The coding of the sample names is as follows: gender, subject number, day number, and part of day. The fraction of "Uncultured" bacteria was remarkable, reflecting the limited knowledge on the small intestinal microbiota. The genera, that were present in all subjects at some timepoint being the core genera, were: *Clostridium*, *Escherichia*, *Streptococcus*, *Ruminococcus*, *Romboutsia*, *Haemophilus*, *Eubacterium*, *Lachnospira* and *Mycobacterium* (albeit the latter in very low numbers). Contrary to what was found by previous studies, the genus *Veillonella* turned out to be not as abundant. Besides the commonalities, **figure 8** shows a striking drop in the relative contribution of the *Clostridium* genus in the last two samples from subject four (f.s4.d3.mrn and f.s4.d3.aft), in comparison with the first two samples of subject four (f.s4.d1.mrn and f.s4.d1.aft). Furthermore, it seemed that a distinction could be made between an *Escherichia*-based and a *Clostridium*-based microbiota. In conclusion, the genera composition varied notably between subjects and even within subjects.

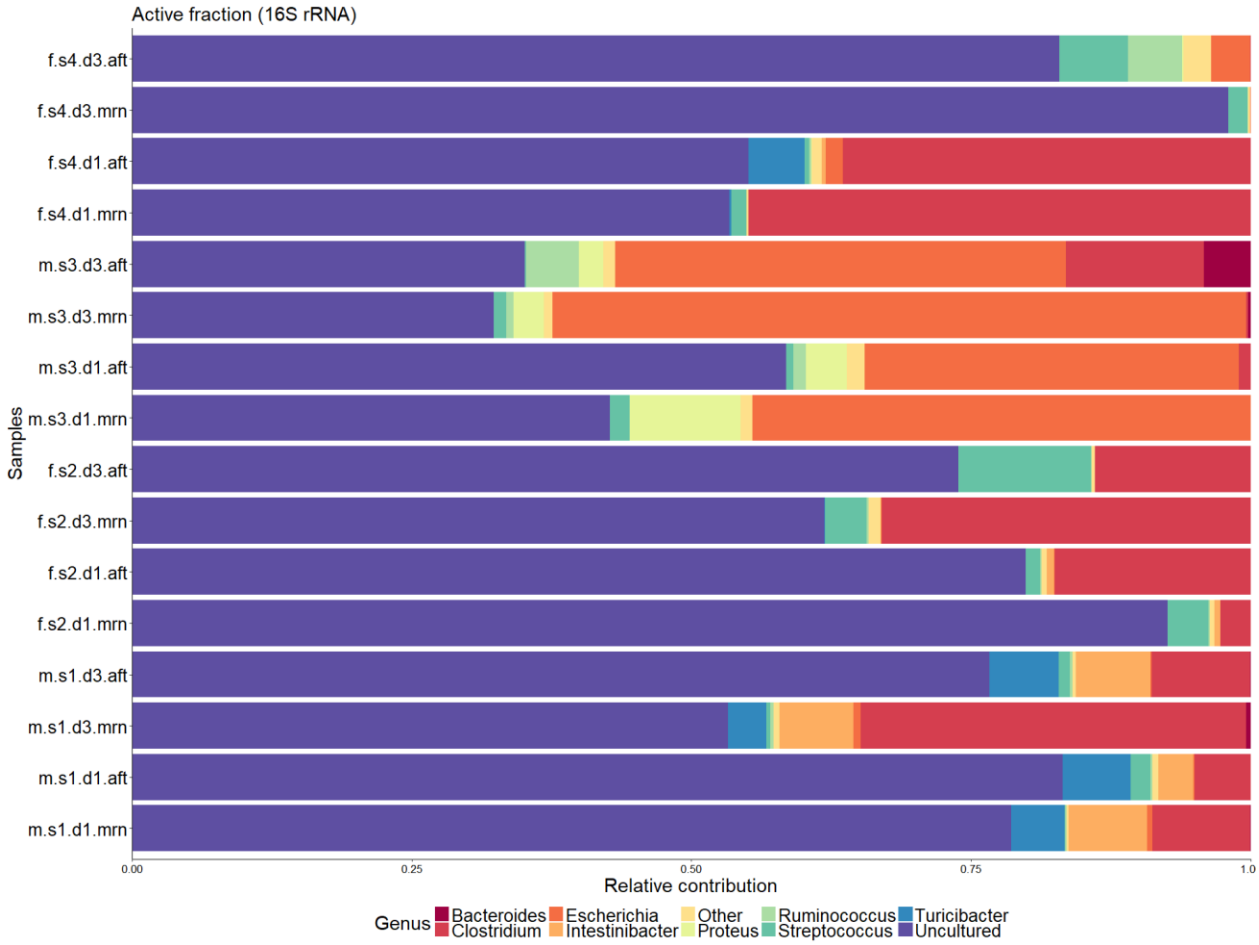


Figure 8 Relative contribution of genera per sample based on the 16S rRNA reads. The coding of the samples is as follows: gender, subject number, day number, and part of day. The category "Other" represents all genera together that were not in the top ten most abundant genera.

Figure 9 shows the results of the PCAs on both count matrices (**figure 9A**: species, **figure 9B**: EC phages). The individual samples colored per subject were plotted on the first and second principal component (PC1 and PC2). For both PCAs and their first two PCs, the accompanying proportions of variance explained (PVE) were of considerable amount, being 13.04% and 10.83% for species counts and 15.42% and 12.28% for phage counts. This suggested that the variance between the subjects could be better explained by phage counts. Clustering of the samples of the same subject can be observed in both figures. Both graphs implied that of subject one and two were more similar to each other than to the other subjects, with subject three and some samples of subject four being outliers. Interestingly, subject three and four being outliers was also observed in the functional PCAs.

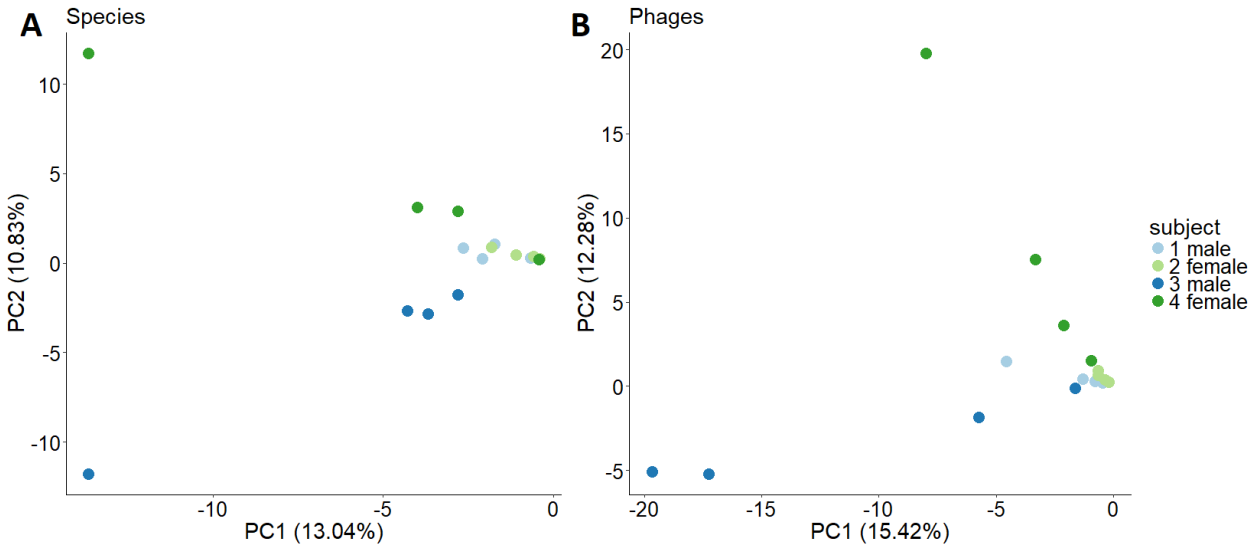


Figure 9 PCAs on species (**A**) and phages (**B**) visualized by plotting the second principal component (PC2) against the first principal component (PC1) with the accompanying proportions of variance explained (PVE). The differently colored dots represent the different subjects: light blue for male 1, light green for female 2, dark blue for male 3, and dark green for female 4.

When zooming in on the unique species and phages that were found in each subject, again dissimilarities can be noticed. **Figure 10** shows the Venn-diagrams representing the species (**figure 10A**) and phages (**figure 10B**) found within each subject. Total numbers of 138 species and 204 phages were found, of which a number of species and phages were found in all subjects' meta- genomes and transcriptomes, respectively 13 and 19. These can be seen as the core species and core phages, and will be analyzed and described later. Apart from the similarities, it can be noticed that quite some species and phages were uniquely present in different subjects. A total number of 13, 31, 25 and 7 species were found uniquely in respectively subject one, two, three and four. A total number of 16, 51, 27 and 5 phages could be found uniquely in respectively subject one, two, three and four. These observations indicated that the meta-genomes and transcriptomes of the different subjects varied compositionally.

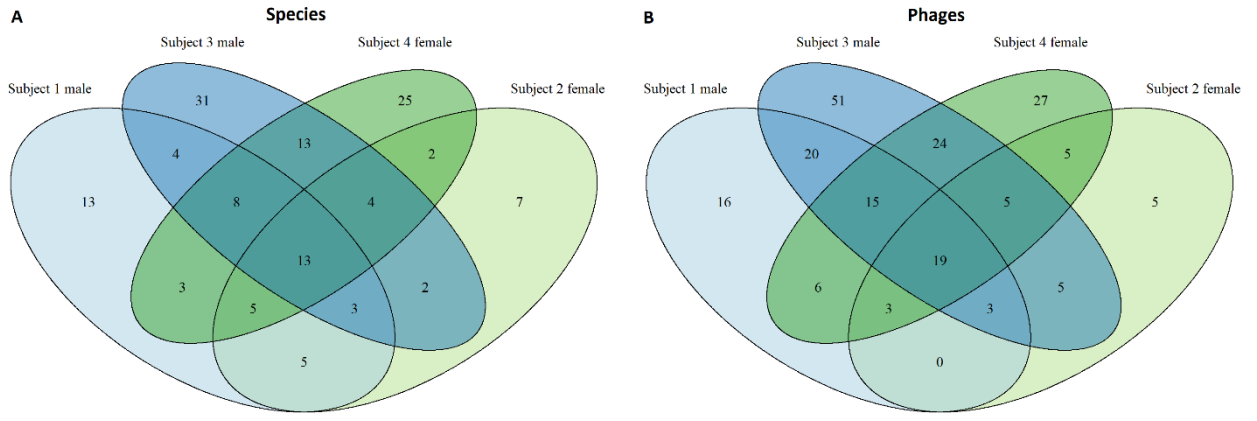


Figure 10 Venn-diagrams representing the species (**A**) and phages (**B**) found within the subjects' meta- genomes and transcriptomes. The numbers in overlapping circles represent species or phages found within two, three or all subjects. The differently colored circles represent the different subjects: light blue for male 1, light green for female 2, dark blue for male 3, and dark green for female 4.

As the small intestines have a highly fluctuating microbial composition in comparison to the colonic microbiome, it is interesting to see if this is also confirmed by this study. **Figure 11** shows graphs of the

Shannon diversity and richness of species and phages over time. It can immediately be noted that the compositional diversity and richness was varying in all subjects, and interestingly varied more, as was seen with the functional diversity and richness, in subject one, three and four. No clear patterns regarding morning and afternoon rhythm could be seen. The sudden drop in the relative contribution of the *Clostridium* genus in subject four as was previously described, is reflected in **figure 11A and 11C** as well. One could note a drop in Shannon species diversity and species richness here too for subject four (dark green line). The Shannon phages diversity stayed close to constant for the latter subject, but an increase in phage richness could be observed in **figure 11D**. **Supplemental figure 2 and 3 in appendix 5** show networks of PCCs calculated from Shannon indexes and richness for Pfam domains, EC numbers, species and phages. The PCC for correlation between Shannon species diversity and Shannon phage diversity was -0.3508, indicating that an increase in phage diversity resulted in a decrease in species diversity. The PCC for correlation between species and phage richness was 0.4865, indicating that an increase in species richness resulted in an increase in phage richness. This shows that the negative correlation between the Shannon diversity indexes could be explained by a change in evenness since the Shannon index is affected by both evenness and richness. The PCC for Shannon species diversity and Shannon Pfam diversity was 0.3914, whereas the PCC for Shannon species diversity and Shannon EC diversity was 0.4318. The PCC for species richness and Pfam richness was 0.4465, whereas the PCC for species richness and EC richness was 0.4278. The latter correlations indicated that an increase in compositional diversity and richness resulted in an increase in functional diversity and richness.

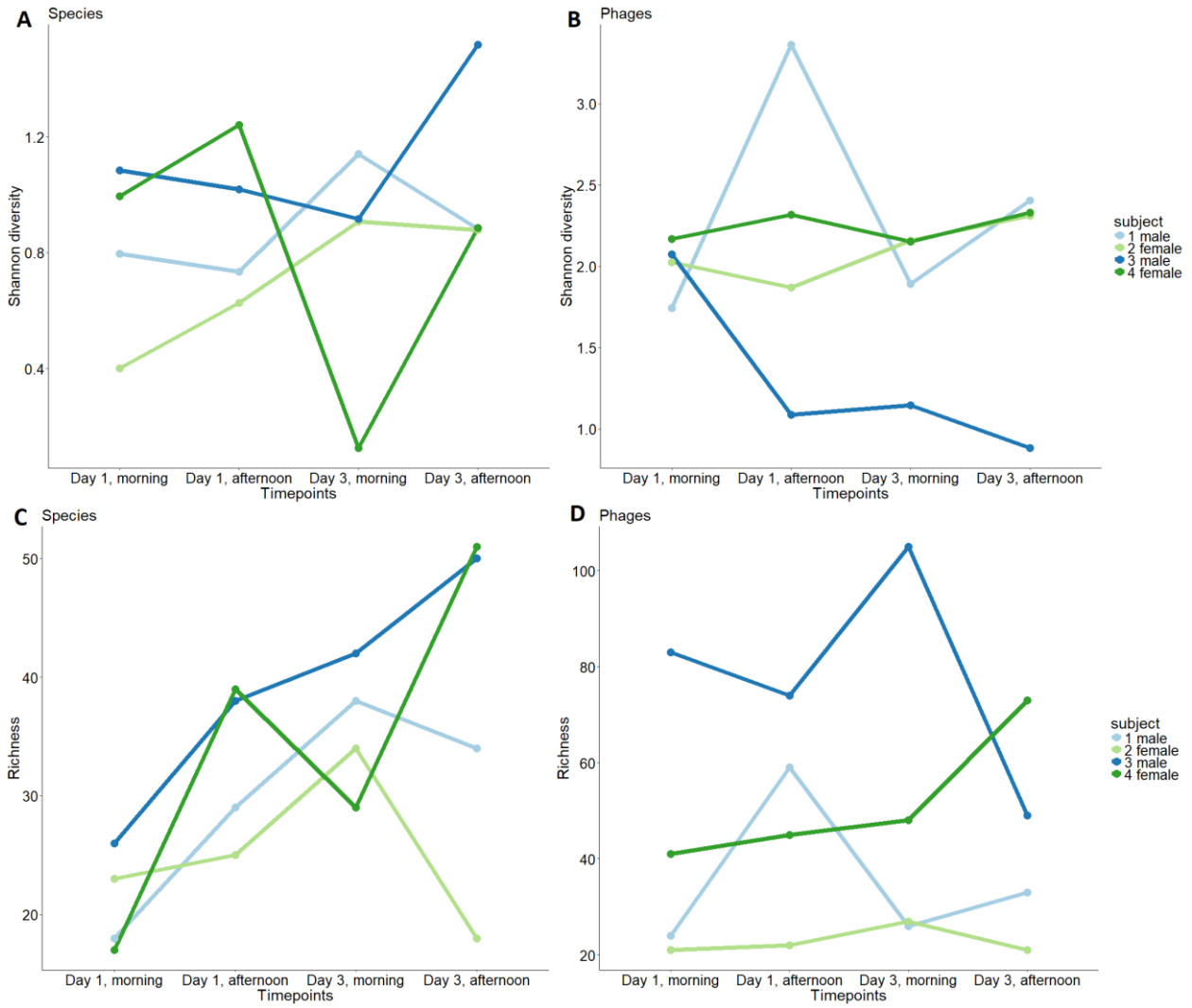


Figure 8 Shannon diversity and richness over time of the different subjects' meta- genomes and transcriptomes based on species and phage counts. Figure **A** and **B** visualize the changes in Shannon diversity for species and phages over time, whereas figure **C** and **D** visualized the changes in richness. The differently colored lines represent the different subjects: light blue for male 1, light green for female 2, dark blue for male 3, and dark green for female 4.

The compositional core consisted of respectively 13 and 19 species and phages and those are shown in **table 3 and 4**. As in the genera composition graph, species were present in the core of the following genera or families: *Escherichia*, *Clostridiaceae*, *Ruminococcus*, *Streptococcus*, *Lachnospiraceae*, *Haemophilus*, *Eubacterium* and *Mycobacterium* (albeit the latter in very low numbers). Again, a large part of the small intestinal core microbiome was unknown or "Uncultured" and no species of the genus *Veillonella* was present in the core. The presence of these genera and families was echoed in the core phages since various *Escherichia*, *Enterobacteria* and *Streptococcus* phages were found. The presence of various Stx2 converting phages was remarkable. These phages, when present in certain *Escherichia coli* strains, can result in pathogenic Shiga toxin 2 (Stx2)-producing *E. coli* that can cause severe illness (Beutin et al., 2012). The *Shigella* phage SfIV infecting *Shigella flexneri* is thought to be an important determinant for pathogenesis in shigellosis (Jakhetia et al., 2013). Overall, the most abundant core phages infect *E. coli* and *Streptococcus* species.

Table 3 The core species and their rounded average read counts across all samples. Whenever there is "bacterium" in the species name, only the genus or family name was known.

Core species	Average counts
Uncultured	7211
Escherichia coli	1365
Clostridiaceae bacterium	839
Ruminococcus bacterium	66
Streptococcus salivarius	48
Streptococcus bacterium	24
Streptococcus parasanguinis	16
Streptococcus sanguinis	11
Lachnospiraceae bacterium	8
Streptococcus infantis	4
Haemophilus parainfluenza	4
Eubacterium bacterium	2
Mycobacterium tuberculosis	1

Table 4 The core phages and their rounded average read counts across all samples.

Core phages	Average counts
Stx2 converting phage II	1918
Enterobacteria phage lambda	588
Streptococcus phage 7201	177
Streptococcus phage Abc2	127
Shigella phage SfIV	122
Streptococcus phage YMC-2011	103
Streptococcus phage 5093	59
Phage cdtI	52
Streptococcus virus 9872	42
Stx2 converting phage 1717	26
Enterobacteria phage WA13	16
Stx2 converting phage 86	15
Salmonella phage SJ46	6
Streptococcus phage Sfi21	5
Lactococcus phage CB13	3
Streptococcus phage PH10	3
Streptococcus phage SM1	2
Lactococcus phage P680	2
Streptococcus phage M102AD	2

Supplemental table 5 and 6 of appendix 4 show the top five uniquely found species and phages per subject, sorted on average count. The unique presence of *Clostridium beijerinckii* and *Clostridium diolis* in subject four immediately stood out. *C. beijerinckii*, formerly *Clostridium acetobutylicum*, is known to ferment dietary fibers to alcohols (Lütke-Eversloh, 2014), whereas *C. diolis* is known to produce 1,3-propanediol by fermentation of glycerol and lignocellulosic hydrolysates (Xin et al., 2016). It turned out that especially these two species were present in high numbers in the first two samples of subject four (f.s4.d1.mrn and f.s4.d1.aft) but totally disappeared in the last two samples (f.s4.d3.mrn and f.s4.d3.aft) causing the trends in Shannon diversity and richness as previously described, however no sudden increase in *Clostridium* phages could be found. The increase in phage diversity and richness for this subject could be caused by increasing abundances of several Streptococcus and Stx2 converting phages. In subject three, *Serratia marcescens* and *Pantoea agglomerans* appeared uniquely with considerable abundances. The first is shown to be injurious to intestinal epithelial cells causing diarrhea (Ochieng et al., 2014) and *P. agglomerans* is a plant pathogen that can cause disease in humans (Cruz et al., 2007). The functional analysis of subject two showed domains and enzymes unique for *Clostridium perfringens*, and the latter species was also found in the metagenome of subject two.

Concerning the uniquely present phages, the *Shigella* phage Ss-VASD that was uniquely present in subject three, is a Shiga toxin 1a converting phage of *Shigella sonnei* that resembles Stx2 converting phages of *E. coli* and might be horizontally transferred from *E. coli*. These *Shigella* phages were isolated from stool samples of subjects with diarrhea and abdominal discomfort (Carter et al., 2016). Subject one contained unique *Lactococcus* and *Lactobacillus* phages. The metatranscriptome of subject two had hits with coding sequences of Haemophilus influenzae phages HP1 and HP2. The top five most abundant unique phages in subject four were all phages of *Streptococcus* species. The *Streptococcus* phage ALQ13.2 phage had been shown to be a virulent phage of *Streptococcus thermophilus* (Guglielmotti et al., 2009), just as the *Streptococcus* phage 2972 (Lévesque et al., 2005).

In conclusion, all analyses indicated that the ileal microbiota and its functionality greatly varied between and within subjects. Some variation in functionality could be explained by compositional variation and vice versa. Differences between subjects in terms of functionality and composition might be a result of dietary differences. Still, there were some culprits and limitations as will be discussed in the next section.

Discussion

Finding more and more evidence for the profound role of the colonic microbiome in disease and host overall well-being, has risen questions with regard to the role of the small intestinal microbiome. The latter is of higher interest as the small intestines are mainly responsible for metabolism and gut-associated immunity. The goal of this study was to explore the similarities and differences in the ileal microbiota of four subjects across four timepoints in terms of functionality and composition. All in all, the results showed that there was ample variation between subjects and within subjects. However, core processes, pathways, genera and phages were identified as well. Furthermore, some of the functional variation could be explained by compositional variation and vice versa. This section will discuss some remarkable discoveries, limitations and propose future work.

The varying low mRNA sequencing depth might have been a limitation for the *de novo*-based assembly and annotation with MetaSAPP, that caused this method to be an outlier in comparison with the BLASTN- and BLASTX-based methods. In general, higher coverage is better for *de novo* assembly. The generated number of mRNA reads could have been higher if the rRNA removal was more efficient. Besides limitations for the method comparison, the varying low sequencing depth might also have partly caused the differences in the number of unique protein domains, enzymes, species and phages that were found in the functional and compositional analyses. As was seen in **figure 4**, the number of uniquely found Pfam domains and EC numbers increased with an increasing sequencing depth. Still, these differences could have also been caused by actual differences in compositional and functional diversity in the ileal microbiota. This culprit could have been solved by the use of biological replicates: multiple effluent samples at the same timepoint processed and sequenced independently. Trends in compositional and functional diversity (e.g. morning and afternoon rhythm) could have been discovered with the availability of biological replicates and the correlations could have been substantiated with more certainty. Finally, the additional application of a standardized diet could have provided more insight into the effect of the host and diet on the small intestinal microbiome. Yet, the core domainome, enzymome and genera were as expected except for the absence of *Veillonella* in the core genera. The latter was a remarkable discovery since it was previously known that *Veillonella* species are in commensalism with *Streptococcus* species in lactate consumption and production, respectively. It might be that this commensal relationship comprises different genera than *Veillonella*. From **figure 8**, it could be observed that the genera *Streptococcus-Clostridium* and *Turicibacter-Clostridium* coincided. *Streptococcus salivarius* is known to produce lactic acid (Aidy et al., 2015) and was present in all metagenomes. The putative lactate fermenter might be *Clostridium bartletti* which appeared in high abundance in those samples with *Streptococcus-Clostridium* coincidence and is able to produce acetic acid in fermentation processes (Song et al., 2004), like *Veillonella* species. This hypothesis, however, needs confirmation since still little is known about *C. bartletti*. This organism may also play a role in the putative commensal relationship with a *Turicibacter* species as the lactate provider. *Turicibacter sanguinis* might be a candidate for lactic acid provision by fermentation (Bosshard et al., 2002) and was present in high abundance in those samples with *Turicibacter-Clostridium* coincidence. Another observation was the seemingly distinction between *Clostridium*-based and *Escherichia*-based metagenomes but this might change radically when the "Uncultured" section is unraveled. Further research is needed to confirm this distinction and to resolve the underlying causes. With regard to subject-specific findings, it was remarkable that subject two's metatranscriptome had protein domains and enzymes associated with *Clostridium perfringens* and that this organism was also

present in the metagenome. Knowing that *C. perfringens* causes gastroenteritis (inflammation of the stomach and small intestines) in humans (Adams et al., 2008), it might actually be that subject two suffered from the latter disease caused by this toxin-producing organism. The metatranscriptome of subject four showed the presence of a CIA30 protein domain known to be part of the mitochondrial complex I in humans and mouse, but is also known to occur in *Schizosaccharomyces pombe* and therefore might originate from the latter probably due to consumption of fermented food (Janssen et al., 2002). This hypothesis could not be backed up by the metagenomic composition since *S. pombe* was not found there. Lastly, there was a coincidence of *Clostridium beijerinckii* and *Clostridium diolis* in high abundances in the first two samples of subject four where after in the last two samples these two species completely disappeared. Since *C. beijerinckii* is known to ferment dietary fibers to alcohols (Lütke-Eversloh, 2014) and *C. diolis* carries out fermentation of glycerol and lignocellulosic hydrolysates thereby producing 1,3-propanediol (Xin et al., 2016). It might be that the rapid decline of these putative commensals was caused by a sudden change in the consumption of dietary fibers (and lignocellulosic material) by subject four as no prominent clues were found in the phage analysis. However, for subject four some unique lytic or virulent phages were found for *Streptococcus thermophilus* and it might be that these also played a role in the sudden disappearance of the latter *Clostridium* species. Still, it can be concluded that subject four might have had some health-related problems due to the virulent phages. The latter examples showed that dietary habits could be inferred from the metatranscriptome and, vice versa, that diet affects the small intestinal microbiome.

With regard to metagenomics based on 16S rRNA similarity, the cut-offs used can be found arbitrary. Here, cut-offs of 95% and 97% sequence similarity for genera and species assignment were used, respectively. In the study of Fournier et al., from 2015 a sequence similarity cut-off for 16S rRNA-based species assignment of 98.7% was proposed, indicating the uncertainty of this method with the rapidly increasing 16S rRNA databases. The latter culprit might have caused the introduction of false positive species in the compositional analysis. A potential false positive might be the *Mycobacterium tuberculosis* present in the core species as none of these subjects had been diagnosed with (abdominal) tuberculosis, the disease that can be caused by this species (Debi et al., 2014).

Despite limitations and possible improvements, it was shown here that the ileal microbiota varied greatly between and within subjects. The core genera turned out to have a different composition than was known before. Clues in the unique presence of protein domains, enzymes, species and phages in the subject's microbiomes could be used to hypothesize about putative infections and dietary habits.

Acknowledgements

I want to thank the Laboratory of Systems and Synthetic Biology from Wageningen University & Research for insights, supervision and data. Special thanks to Jasper Koehorst MSc., dr. Peter Schaap, Benoit Carreres MSc. and dr. Maria Suarez Diez.

References

Adams, J. J., Gregg, K., Bayer, E. A., Boraston, A. B., & Smith, S. P. (2008). Structural basis of Clostridium perfringens toxin complex formation. Proceedings of the National Academy of Sciences of the United States of America, 105(34), 12194–12199. http://doi.org/10.1073/pnas.0803154105

Aidy, S. E., van den Bogert, B., & Kleerebezem, M. (2015, April 1). The small intestine microbiota, nutritional modulation and relevance for health. Current Opinion in Biotechnology. Elsevier Ltd. https://doi.org/10.1016/j.copbio.2014.09.005

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bairoch, A. (2000). The ENZYME database in 2000. Nucleic Acids Research, 28(1), 304–305.

Becker, C., Neurath, M. F., & Wirtz, S. (2015). The intestinal microbiota in inflammatory bowel disease. ILAR Journal, 56(2), 192–204. https://doi.org/10.1093/ilar/ilv030

Beutin, L., Hammerl, J. A., Strauch, E., Reetz, J., Dieckmann, R., Kelner-Burgos, Y., … Reinhardt, R. (2012). Spread of a Distinct Stx2-Encoding Phage Prototype among Escherichia coli O104:H4 Strains from Outbreaks in Germany, Norway, and Georgia. Journal of Virology, 86(19), 10444–10455. https://doi.org/10.1128/JVI.00986-12

Blot, N., Berrier, C., Hugouvieux-Cotte-Pattat, N., Ghazi, A., & Condemine, G. (2002). The oligogalacturonate-specific porin KdgM of Erwinia chrysanthemi belongs to a new porin family. Journal of Biological Chemistry, 277(10), 7936–7944. https://doi.org/10.1074/jbc.M109193200

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120. http://doi.org/10.1093/bioinformatics/btu170

Bosshard, P. P., Zbinden, R., & Altwegg, M. (2002). Turicibacter sanguinis gen. nov., sp. nov., a novel anaerobic, Gram-positive bacterium. International Journal of Systematic and Evolutionary Microbiology, 52, 1263–1266. https://doi.org/10.1099/ijs.0.02056-0.A

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods, 12(1), 59–60. https://doi.org/10.1038/nmeth.3176

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. BMC Bioinformatics, 10, 421. http://doi.org/10.1186/1471-2105-10-421

Carter, C. C., Fierer, J., Chiu, W. W., Looney, D. J., Strain, M., & Mehta, S. R. (2016). A Novel Shiga Toxin 1a-Converting Bacteriophage of Shigella sonnei With Close Relationship to Shiga Toxin 2-Converting Pages of Escherichia coli. Open Forum Infectious Diseases, 3(2), ofw079. http://doi.org/10.1093/ofid/ofw079

Chen, H. (2018). VennDiagram: Generate high-resolution Venn and Euler plots. R package version 1.6.20. https://cran.r-project.org/package=VennDiagram

Cruz, A. T., Cazacu, A. C., & Allen, C. H. (2007). Pantoea agglomerans, a Plant Pathogen Causing Human Disease . Journal of Clinical Microbiology, 45(6), 1989–1992. http://doi.org/10.1128/JCM.00632-07

Debi, U., Ravisankar, V., Prasad, K. K., Sinha, S. K., & Sharma, A. K. (2014, October 28). Abdominal tuberculosis of the gastrointestinal tract: Revisited. World Journal of Gastroenterology. WJG Press. https://doi.org/10.3748/wjg.v20.i40.14831

Dinan, T. G., & Cryan, J. F. (2015, October 9). The impact of gut microbiota on brain and behaviour: Implications for psychiatry. Current Opinion in Clinical Nutrition and Metabolic Care. Lippincott Williams and Wilkins. https://doi.org/10.1097/MCO.0000000000000221

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: the protein families database. Nucleic Acids Research, 42(Database issue), D222–D230. http://doi.org/10.1093/nar/gkt1223

Fournier, P.-E., Rossi-Tamisier, M., Benamar, S., & Raoult, D. (2015). Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species. International Journal of Systematic and Evolutionary Microbiology, 65(6), 1929–1934. https://doi.org/10.1099/ijs.0.000161

Grossman, T. H. (2016). Tetracycline antibiotics and resistance. Cold Spring Harbor Perspectives in Medicine, 6(4). https://doi.org/10.1101/cshperspect.a025387

Guglielmotti, D. M., Deveau, H., Binetti, A. G., Reinheimer, J. A., Moineau, S., & Quiberoni, A. (2009). Genome analysis of two virulent Streptococcus thermophilus phages isolated in Argentina. International Journal of Food Microbiology, 136(1), 101–109. https://doi.org/10.1016/j.ijfoodmicro.2009.09.005

Haberman, Y., Tickle, T. L., Dexheimer, P. J., Kim, M.-O., Tang, D., Karns, R., … Denson, L. A. (2014). Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. The Journal of Clinical Investigation, 124(8), 3617–3633. http://doi.org/10.1172/JCI75436

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., … Yeats, C. (2009). InterPro: the integrative protein signature database. Nucleic Acids Research, 37(Database issue), D211–D215. http://doi.org/10.1093/nar/gkn785

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 11, 119. http://doi.org/10.1186/1471-2105-11-119

Irrazábal, T., Belcheva, A., Girardin, S. E., Martin, A., & Philpott, D. J. (2014, April 24). The multifaceted role of the intestinal microbiota in colon cancer. Molecular Cell. Cell Press. https://doi.org/10.1016/j.molcel.2014.03.039

Jakhetia, R., Talukder, K. A., & Verma, N. K. (2013). Isolation, characterization and comparative genomics of bacteriophage SfIV: A novel serotype converting phage from Shigella flexneri. BMC Genomics, 14(1). https://doi.org/10.1186/1471-2164-14-677

Janssen, R., Smeitink, J., Smeets, R., & van den Heuvel, L. (2002). CIA30 complex I assembly factor: a candidate for human complex I deficiency? Human Genetics, 110(3), 264–270. https://doi.org/10.1007/s00439-001-0673-3

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), 1236–1240. http://doi.org/10.1093/bioinformatics/btu031

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1), 27–30.

Kessler, A. C., Brown, P. K., Romana, L. K., & Reeves, P. R. (1991). Molecular cloning and genetic characterization of the rfb region from Yersinia pseudotuberculosis serogroup IIA, which determines the formation of the 3,6-dideoxyhexose abequose. J Gen Microbiol, 137(12), 2689–2695. https://doi.org/10.1099/00221287-137-12-2689

Koehorst, J. J., van Dam, J. C. J., Saccenti, E., Martins dos Santos, V. A. P., Suarez-Diez, M., & Schaap, P. J. (2018). SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles. Bioinformatics, 34(8), 1401–1403. http://doi.org/10.1093/bioinformatics/btx767

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics, 28(24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

Lee, K. H., Lee, S. G., Eun Lee, K., Jeon, H., Robinson, H., & Oh, B. H. (2012). Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. Proteins: Structure, Function and Bioinformatics, 80(11), 2573–2582. https://doi.org/10.1002/prot.24138

Leimena, M. M., Ramiro-Garcia, J., Davids, M., van den Bogert, B., Smidt, H., Smid, E. J., … Kleerebezem, M. (2013). A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. BMC Genomics, 14(1). https://doi.org/10.1186/1471-2164-14-530

Lévesque, C., Duplessis, M., Labonté, J., Labrie, S., Fremaux, C., Tremblay, D., & Moineau, S. (2005). Genomic organization and molecular analysis of virulent bacteriophage 2972 infecting an exopolysaccharide-producing Streptococcus thermophilus strain. Appl Environ Microbiol, 71(7), 4057–4068. https://doi.org/10.1128/AEM.71.7.4057-4068.2005

Li, D., Liu, C.-M., Luo, R., Lam, T.-W., & Sadakane, K. (2014). MEGAHIT: An ultra-fast single-node solution for large ad complex metagenomic assembly via succinct de Bruijn graph. Bioinformatics, 31(10), 11. https://doi.org/10.1093/bioinformatics/btv033

Liu, H., Ray, W. K., Helm, R. F., Popham, D. L., & Melville, S. B. (2016). Analysis of the Spore Membrane Proteome in Clostridium perfringens Implicates Cyanophycin in Spore Assembly. Journal of Bacteriology, 198(12), 1773–1782. http://doi.org/10.1128/JB.00212-16

Lütke-Eversloh, T. (2014). Application of new metabolic engineering tools for Clostridium acetobutylicum. Applied Microbiology and Biotechnology. Springer Verlag. https://doi.org/10.1007/s00253-014-5785-5

Martínez, B., Suárez, J. E., & Rodríguez, A. (1996). Lactococcin 972: A homodimeric lactococcal bacteriocin whose primary target is not the plasma membrane. Microbiology, 142(9), 2393–2398. https://doi.org/10.1099/00221287-142-9-2393

Miele, L., Giorgio, V., Alberelli, M. A., De Candia, E., Gasbarrini, A., & Grieco, A. (2015, December 1). Impact of Gut Microbiota on Obesity, Diabetes, and Cardiovascular Disease Risk. Current Cardiology Reports. Current Medicine Group LLC 1. https://doi.org/10.1007/s11886-015-0671-z

Nguyen, N.-N., Srihari, S., Leong, H. W., & Chong, K.-F. (2015). EnzDP: Improved enzyme annotation for metabolic network reconstruction based on domain composition profiles. Journal of Bioinformatics and Computational Biology, 13(05), 1543003. https://doi.org/10.1142/S0219720015430039

Ochieng, J. B., Boisen, N., Lindsay, B., Santiago, A., Ouma, C., Ombok, M., … Nataro, J. P. (2014). Serratia marcescens is injurious to intestinal epithelial cells. Gut Microbes, 5(6), 729–736. http://doi.org/10.4161/19490976.2014.972223

Oksanen, J., Guillaume Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H. (2018). Vegan: Community ecology package. R package version 2.5-2. https://cran.r-project.org/package=vegan

Patman, G. (2015, January 1). Gut microbiota: The difference diet makes to metabolites and microbiota. Nature Reviews Gastroenterology and Hepatology. Nature Publishing Group. https://doi.org/10.1038/nrgastro.2014.227

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., … Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 41(Database issue), D590–D596. http://doi.org/10.1093/nar/gks1219

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rodríguez-r, L. M., Castro, J. C., Rodriguez-r, L. M., Castro, J. C., Kyrpides, N. C., Cole, J. R., & Tiedje, J. M. (2018). How Much Do rRNA Gene Surveys Underestimate Extant Bacterial Diversity ? crossm How Much Do rRNA Gene Surveys Underestimate Extant, (January). https://doi.org/10.1128/AEM.00014-18

Shrestha, R. K., Lubinsky, B., Bansode, V. B., Moinz, M. B. J., McCormack, G. P., & Travers, S. A. (2014). QTrim: A novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. BMC Bioinformatics, 15(1). https://doi.org/10.1186/1471-2105-15-33

Song, Y. L., Liu, C. X., McTeague, M., Summanen, P., & Finegold, S. M. (2004). Clostridium bartlettii sp. nov., isolated from human faeces. Anaerobe, 10(3), 179–184. https://doi.org/10.1016/j.anaerobe.2004.04.004

Tang, Y., Horikoshi, M., Li, W. (2016). ggfortify: Unified interface to visualize statistical results of popular R packages. The R Journal, 8.2, 478-489.

The Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research, 45(Database issue), D331–D338. http://doi.org/10.1093/nar/gkw1108

Tsuge, H., Nagahama, M., Nishimura, H., Hisatsune, J., Sakaguchi, Y., Itogawa, Y., … Sakurai, J. (2003). Crystal structure and site-directed mutagenesis of enzymatic components from Clostridium perfringens Iota-toxin. Journal of Molecular Biology, 325(3), 471–483. https://doi.org/10.1016/S0022-2836(02)01247-0

Wickham, H. (2007). Reshaping Data with the reshape package. Journal of Statistical Software, 21(12), 1-20. http://www.jstatsoft.org/v21/i12/

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

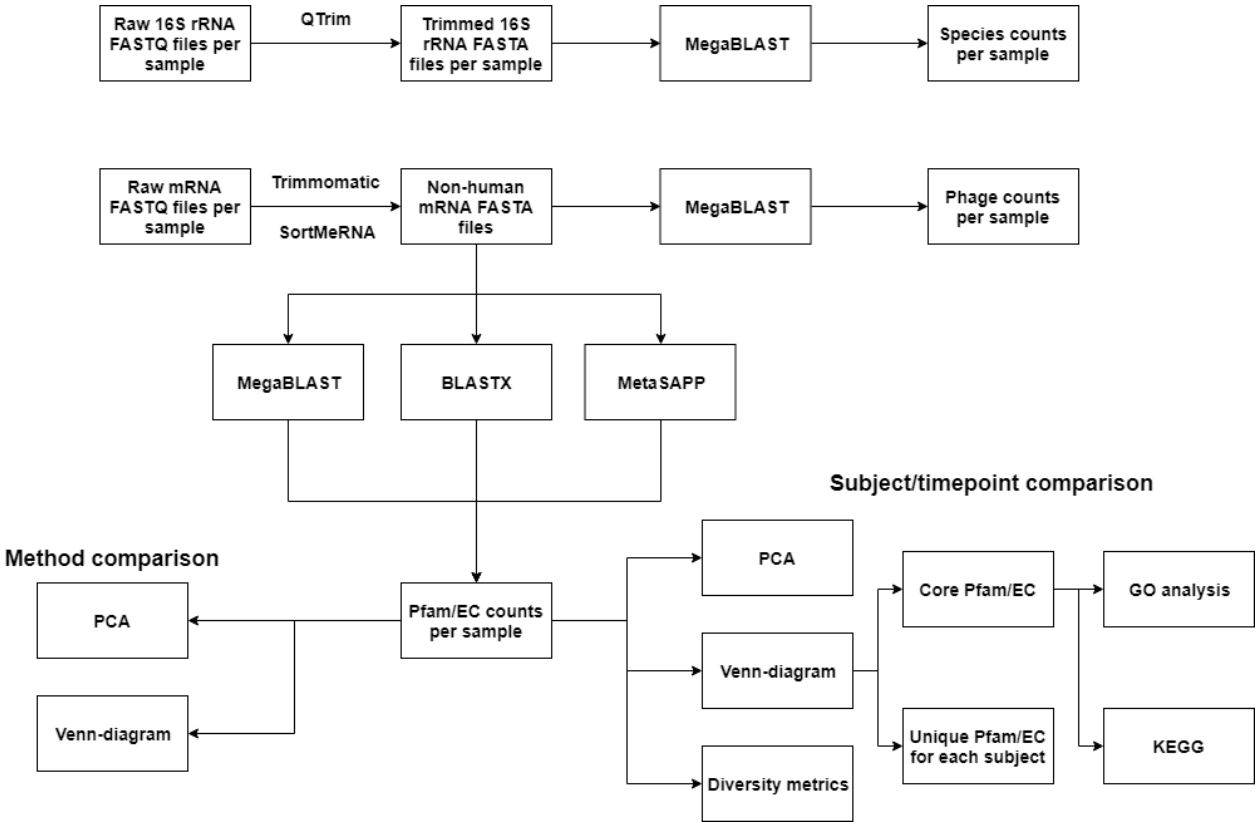
Wyk, P., & Reeves, P. (1989). Identification and sequence of the gene for abequose synthase, which confers antigenic specificity on group B salmonellae: homology with galactose epimerase. Journal of Bacteriology, 171(10), 5687–5693.

Xin, B., Wang, Y., Tao, F., Li, L., Ma, C., & Xu, P. (2016). Co-utilization of glycerol and lignocellulosic hydrolysates enhances anaerobic 1,3-propanediol production by Clostridium diolis. Scientific Reports, 6. https://doi.org/10.1038/srep19044

Zoetendal, E. G., Raes, J., van den Bogert, B., Arumugam, M., Booijink, C. C., Troost, F. J., … Kleerebezem, M. (2012). The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. The ISME Journal, 6(7), 1415–1426. http://doi.org/10.1038/ismej.2011.212

Appendices

1. Workflow



Supplemental figure 1 General workflow of this study.

2. Data overview

Supplemental table 1 Overview of mRNA FASTQ files and number of reads after each preprocessing step.

Filename	Sample name	Raw reads	Reads after trimming	Reads after removing rRNA	Reads after removing human mRNA
Subject 1					
NG-5593_1A.fastq	m.s1.d1.mrn	31,180,479	29,217,597	525,202	524,336
NG-5593_1B.fastq	m.s1.d1.aft	19,781,634	18,597,481	734,478	733,191
NG-5593_1C.fastq	m.s1.d3.mrn	25,962,452	17,420,062	402,456	399,740
NG-5593_1D.fastq	m.s1.d3.aft	54,015,995	50,698,152	3,515,667	3,505,943
Subject 2					
NG-5593_2A.fastq	f.s2.d1.mrn	29,709,279	28,095,191	5,075,173	5,072,197
NG-5593_2B.fastq	f.s2.d1.aft	37,354,030	34,968,648	2,896,022	2,893,578
NG-5593_2C.fastq	f.s2.d3.mrn	46,532,673	43,666,664	3,832,310	3,827,546
NG-5593_2D.fastq	f.s2.d3.aft	24,112,121	22,661,045	2,060,021	2,056,295
Subject 3					
NG-5593_3A.fastq	m.s3.d1.mrn	21,557,933	19,826,309	2,013,325	2,004,724
NG-5593_3B.fastq	m.s3.d1.aft	28,695,216	26,925,707	1,818,640	1,814,236
NG-5593_3C.fastq	m.s3.d3.mrn	73,247,937	68,509,930	6,661,618	6,651,942
NG-5593_3D.fastq	m.s3.d3.aft	15,237,944	14,329,235	1,163,205	1,162,144
Subject 4					
NG-5593_4A_read_1.fastq	f.s4.d1.mrn	42,211,887	31,448,423	5,278,308	5,261,776
NG-5593_4B_read_1.fastq	f.s4.d1.aft	18,762,398	12,621,142	1,720,827	1,714,002
NG-5593_4C_read_1.fastq	f.s4.d3.mrn	26,866,251	13,997,320	3,520,452	3,508,987
NG-5593_4D_read_1.fastq	f.s4.d3.aft	32,664,709	23,554,625	3,892,675	3,855,731
NG-5593_4A_read_2.fastq	f.s4.d1.mrn	42,211,887	31,448,423	5,278,308	5,261,776
NG-5593_4B_read_2.fastq	f.s4.d1.aft	18,762,398	12,621,142	1,720,827	1,714,002
NG-5593_4C_read_2.fastq	f.s4.d3.mrn	26,866,251	13,997,320	3,520,452	3,508,987
NG-5593_4D_read_2.fastq	f.s4.d3.aft	32,664,709	23,554,625	3,892,675	3,855,731
Low depth replicates					
NG-5450_A.fastq	m.s1.d3.mrn	9,368,635	8,499,689	547,599	538,786
NG-5450_B.fastq	f.s2.d1.mrn	8,951,083	8,140,333	1,929,310	1,919,543
Total		666,717,901	554,799,063	61,999,550	61,785,193

Supplemental table 2 Overview of 16S rRNA reads per sample after trimming.

Filename	Sample name	Reads after trimming
Subject 1		
SRR882257.fastq	m.s1.d1.mrn	6,597
SRR882256.fastq	m.s1.d1.aft	17,079
SRR882259.fastq	m.s1.d3.mrn	13,514
SRR882258.fastq	m.s1.d3.aft	24,467
Subject 2		
SRR882263.fastq	f.s2.d1.mrn	7,375
SRR882260.fastq	f.s2.d1.aft	6,514
SRR882265.fastq	f.s2.d3.mrn	9,054
SRR882264.fastq	f.s2.d3.aft	8,866
Subject 3		
SRR882267.fastq	m.s3.d1.mrn	9,966
SRR882266.fastq	m.s3.d1.aft	12,561
SRR882269.fastq	m.s3.d3.mrn	11,768
SRR882268.fastq	m.s3.d3.aft	16,258
Subject 4		
SRR882271.fastq	f.s4.d1.mrn	4,539
SRR882270.fastq	f.s4.d1.aft	23,606
SRR882273.fastq	f.s4.d3.mrn	13,901
SRR882272.fastq	f.s4.d3.aft	8,140
Total		194,205

3. Functional analysis: protein domains and enzymes

Supplemental table 3 The top five most abundant Pfam domains present uniquely in each subject. The average counts are between brackets. Protein domains named with "DUFXXXX" are Domains of Unknown Function.

Subject 1 male	Subject 2 female	Subject 3 male	Subject 4 female
PF14461: Prokaryotic E2 family B (1)	PF12135: Sialidase enzyme penultimate C terminal domain (16)	PF08971: Glycogen synthesis protein (98)	PF08547: Complex I intermediate-associated protein 30 (CIA30) (86)
PF07875: Coat F domain (1)	PF16295: Tetracycline repressor, C-terminal all-alpha domain (9.5)	PF14785: Maltose transport systems permease protein MalF P2 domain (91.75)	PF09683: Bacteriocin (Lactococcin_972) (21)
PF02686: Glu-tRNAGln amidotransferase C subunit (0.75)	PF03496: ADP-ribosyltransferase exoenzyme (3.25)	PF07351: DUF1480 (44.5)	PF16813: CRISPR-associated protein Csn2 subfamily St (17.75)
PF04422: Coenzyme F420 (de)hydrogenase, beta subunit N-term (0.75)	PF13282: DUF4070 (1.75)	PF13808: DDE_Tnp_1-associated (31.25)	PF10665: Minor capsid protein (8.75)
PF04432: Coenzyme F420 (de)hydrogenase, beta subunit C terminus (0.75)	PF04511: Der1-like family (1.75)	PF06178: Oligogalacturonate-specific porin protein (KdgM) (26.75)	PF15515: MvaI/BcnI restriction endonuclease family (8)
...	...	...	...

Supplemental table 4 The top five most abundant EC number present uniquely in each subject. The average counts are between brackets.

Subject 1 male	Subject 2 female	Subject 3 male	Subject 4 female
5.1.1.5: Lysine racemase (0.5)	6.3.2.30: Cyanophycin synthase (L-arginine-adding) (7.5)	3.2.1.49: Alpha-N-acetylgalactosaminidase (11.75)	1.1.1.47: Glucose 1-dehydrogenase (NAD(P)(+)) (13.5)
3.2.1.82: Exo-poly-alpha-galacturonosidase (0.5)	6.3.3.3: Dethiobiotin synthase (1.75)	2.8.3.19: CoA:oxalate CoA-transferase (11.5)	1.1.1.251: Galactitol-1-phosphate 5-dehydrogenase (5.5)
2.1.1.113: Site-specific DNA-methyltransferase (cytosine-N(4)-specific) (0.5)	6.2.1.14: 6-carboxyhexanoate-CoA ligase (1.5)	3.5.3.9: Allantoate deiminase (9.25)	2.4.2.45: Decaprenyl-phosphate phosphoribosyltransferase (5.25)
3.5.5.1: Nitrilase (0.5)	4.1.99.2: Tyrosine phenol-lyase (1.5)	1.1.1.341: CDP-abequose synthase (5.75)	1.10.2.2: Quinol-cytochrome-c reductase (4.25)
3.1.11.6: Exodeoxyribonuclease VII (0.5)	3.2.1.n1: Blood group B branched chain alpha-1,3-galactosidase (0.75)	1.3.8.9: Very-long-chain acyl-CoA dehydrogenase (3)	1.5.5.2: Proline dehydrogenase (4)
...	...	...	...

4. Compositional analysis: species and phages

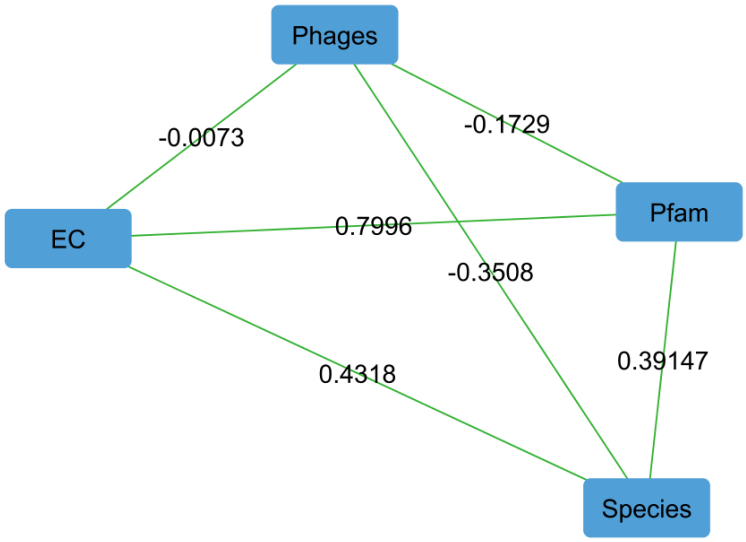
Supplemental table 5 The top five most abundant species present uniquely in each subject. The average read counts are between brackets.

Subject 1 male	Subject 2 female	Subject 3 male	Subject 4 female
Streptococcus dentirousetti (7.25)	*Veillonella atypica* (2.75)	*Serratia marcescens* (16.25)	*Clostridium beijerinckii* (1819.25)
Bifidobacterium lactis (4.25)	*Bifidobacterium pseudocatenulatum* (1.75)	*Pantoea agglomerans* (15.75)	*Clostridium diolis* (200.50)
Romboutsia lituseburensis (1)	*Turicibacter* bacterium (1)	*Streptococcus sobrinus* (4.50)	*Escherichia fergusonii* (3.75)
Erysipelotrichaceae bacterium (1)	*Eubacterium tenue* (0.50)	*Clostridium clostridioforme* (4.25)	*Haemophilus sputorum* (3.50)
Lactobacillus fermentum (0.75)	*Streptococcus cristatus* (0.25)	*Enterobacteriaceae* bacterium (3.75)	*Streptococcus australis* (1)
...	...	...	...

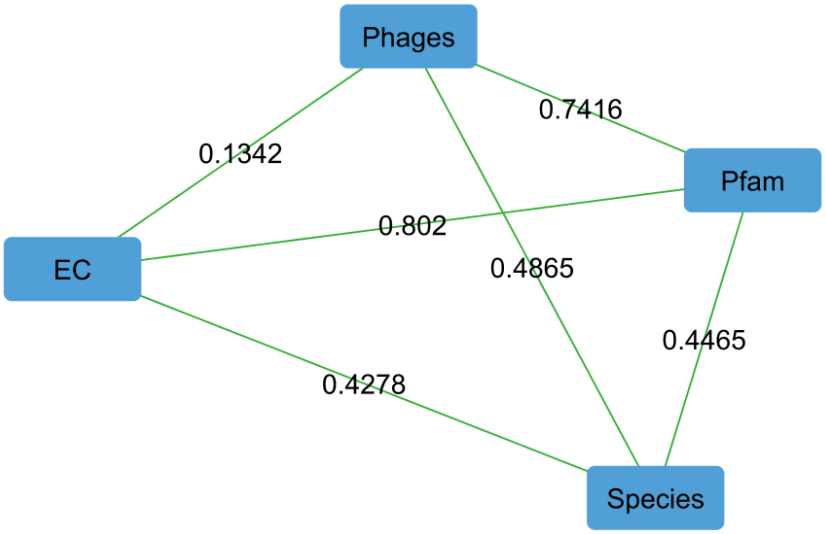
Supplemental table 6 The top five most abundant phages present uniquely in each subject. The average counts are between brackets.

Subject 1 male	Subject 2 female	Subject 3 male	Subject 4 female
Lactococcus phage jj50 (4.25)	*Haemophilus* phage HP2 (4.75)	*Shigella* phage Ss-VASD (51.50)	*Streptococcus* phage 858 (26)
Lactococcus phage 712 (3.75)	*Haemophilus* phage HP1 (1)	*Escherichia* phage TL-2011c (17.75)	*Streptococcus* phage ALQ13.2 (22.50)
Lactobacillus phage J-1 (0.50)	*Streptococcus* prophage 315.4 (0.50)	*Enterobacteria* phage phiP27 (17.50)	*Streptococcus* phage 2972 (16)
Lactococcus phage phiLC3 (0.50)	*Gordonia* phage Bantam (0.25)	*Enterococcus* phage VB_EfaS_IME196 (11.75)	*Streptococcus* phage phiARI0468-2 (2.50)
Streptococcus phage DCC1738 (0.50)	*Streptococcus* phage phiARI0031 (0.25)	*Enterobacteria* phage T4 (6.25)	*Streptococcus* prophage 315.2 (2)
...		...	...

5. PCC networks of Pfam domain, EC number, species and phage Shannon indexes and richness



Supplemental figure 2 Pearson Correlation Coefficient (PCC) network of Shannon indexes between all samples in terms of Pfam domains, EC numbers, species and phages.



Supplemental figure 3 Pearson Correlation Coefficient (PCC) network of richness between all samples in terms of Pfam domains, EC numbers, species and phages.

6. Commands

Data preprocessing

```
fastqc NG-5450_A.fastq -o NG-5450_A

Trimmomatic-0.36.jar SE -threads 4 -phred33 NG-5450_A.fastq NG-
5450_A_trimmed.fastq ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36

Trimmomatic-0.36.jar PE -threads 4 -phred33 NG-5593_4A_read_1.fastq  NG-
5593_4A_read_2.fastq NG-5593_4A_read_1_paired.fastq NG-
```

```
5593_4A_read_1_unpaired.fastq NG-5593_4A_read_2_paired.fastq NG-
5593_4A_read_2_unpaired.fastq ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36

fastqc NG-5450_A_trimmed.fastq -o NG-5450_A_trimmed

indexdb_rna --ref ./rRNA_databases/silva-bac-16s-id90.fasta,./index/silva-bac-16s-db:\
./rRNA_databases/silva-bac-23s-id98.fasta,./index/silva-bac-23s-db:\
./rRNA_databases/silva-arc-16s-id95.fasta,./index/silva-arc-16s-db:\
./rRNA_databases/silva-arc-23s-id98.fasta,./index/silva-arc-23s-db:\
./rRNA_databases/silva-euk-18s-id95.fasta,./index/silva-euk-18s-db:\
./rRNA_databases/silva-euk-28s-id98.fasta,./index/silva-euk-28s:\
./rRNA_databases/rfam-5s-database-id98.fasta,./index/rfam-5s-db:\
./rRNA_databases/rfam-5.8s-database-id98.fasta,./index/rfam-5.8s-db

sortmerna --ref ./rRNA_databases/silva-bac-16s-id90.fasta,./index/silva-bac-16s-
db:./rRNA_databases/silva-bac-23s-id98.fasta,./index/silva-bac-23s-
db:./rRNA_databases/silva-arc-16s-id95.fasta,./index/silva-arc-16s-
db:./rRNA_databases/silva-arc-23s-id98.fasta,./index/silva-arc-23s-
db:./rRNA_databases/silva-euk-18s-id95.fasta,./index/silva-euk-18s-
db:./rRNA_databases/silva-euk-28s-id98.fasta,./index/silva-euk-
28s:./rRNA_databases/rfam-5s-database-id98.fasta,./index/rfam-5s-
db:./rRNA_databases/rfam-5.8s-database-id98.fasta,./index/rfam-5.8s-db --reads NG-
5450_A_trimmed.fastq --num_alignments 1 --fastx --aligned NG-5450_A_trimmed_rRNA --
other NG-5450_A_trimmed_mRNA --log -v -a 4

sortmerna --ref ./rRNA_databases/silva-bac-16s-id90.fasta,./index/silva-bac-16s-
db:./rRNA_databases/silva-bac-23s-id98.fasta,./index/silva-bac-23s-
db:./rRNA_databases/silva-arc-16s-id95.fasta,./index/silva-arc-16s-
db:./rRNA_databases/silva-arc-23s-id98.fasta,./index/silva-arc-23s-
db:./rRNA_databases/silva-euk-18s-id95.fasta,./index/silva-euk-18s-
db:./rRNA_databases/silva-euk-28s-id98.fasta,./index/silva-euk-
28s:./rRNA_databases/rfam-5s-database-id98.fasta,./index/rfam-5s-
db:./rRNA_databases/rfam-5.8s-database-id98.fasta,./index/rfam-5.8s-db --reads NG-
5593_4A_trimmed_merged.fastq --num_alignments 1 --fastx --aligned NG-
5593_4A_trimmed_merged_RNA --other NG-5593_4A_trimmed_merged_mRNA --paired-in --log -v
-a 4

indexdb_rna --ref GRCh38_latest_rna.fna,GRCh38_transcripts_db -v

sortmerna --ref
/scratch/lottewitjes/GRCh38_latest_rna.fna,/metagenomics/lottewitjes/programs/sortmerna
-2.1b/index/GRCh38_transcripts_db --reads NG-5450_A_trimmed_mRNA.fastq --
num_alignments 1 --fastx --aligned NG-5450_A_trimmed_human_mRNA --other NG-
5450_A_trimmed_nonhuman_mRNA --log -v -a 4

sortmerna --
ref /scratch/lottewitjes/GRCh38_latest_rna.fna,/metagenomics/lottewitjes/programs/sortm
erna-2.1b/index/GRCh38_transcripts_db --reads NG-5593_4A_merged_trimmed_mRNA.fastq --
num_alignments 1 --fastx --aligned NG-5593_4A_trimmed_merged_human_mRNA -other NG-
5593_4A_trimmed_merged_nonhuman_mRNA --paired-in --log -v -a 4
```

BLASTX-based against SAPP protein database

```
diamond makedb --in protein_unique.fasta -d protein_SAPP
diamond blastx --threads 16 --db SAPP_protein.dmnd --out NG-
5593_1A_trimmed__nonhuman_mRNA.tsv --outfmt 6 --query NG-
5593_1A_trimmed_nonhuman_mRNA.fastq --max-target-seqs 1 --query-gencode 11
```

BLASTN-based against SAPP gene database

```
makeblastdb -in SAPP_gene.fasta -dbtype nucl -parse_seqids -out SAPP_gene

makembindex -input SAPP_gene -iformat blastdb -output SAPP_gene
```

```
blastn -task megablast -db SAPP_gene -query NG-5593_1A_trimmed_nonhuman_mRNA.fasta -
out NG-5593_1A.tsv -outfmt 6 -max_target_seqs 1 -num_threads 12
```

MetaSAPP *de novo* assembly and prediction

```
java -jar /metagenomics/lottewitjes/programs/SAPP/assembly.jar -threads 4 -megahit -
identifier NG-5593_1A -output NG-5593_1A_assembly.hdt -read1 NG-
5593_1A_trimmed_nonhuman_mRNA.fastq

java -jar /metagenomics/lottewitjes/programs/SAPP/assembly.jar -threads 4 -megahit -
identifier NG-5593_4A -output NG-5593_4A_assembly.hdt -read1 NG-
5593_4A_read_1_trimmed_nonhuman_mRNA.fastq -read2 NG-
5593_4A_read_2_trimmed__nonhuman_mRNA.fastq

java -jar /metagenomics/lottewitjes/programs/SAPP/Conversion.jar -merge -input NG-
5593_1A_assembly.hdt,NG-5593_1B_assembly.hdt,NG-5593_1C_assembly.hdt,NG-
5593_1D_assembly.hdt -output NG-5593_1_assembly.hdt

java -jar /metagenomics/lottewitjes/programs/SAPP/genecaller.jar -prodigal -meta -codon
11 -input NG-5593_1_assembly.hdt -output NG-5593_1_prodigal.hdt

java -jar /metagenomics/lottewitjes/programs/SAPP/InterProScan.jar -input NG-
5593_1_prodigal.hdt -output NG-5593_1_interproscan.hdt

java -jar /metagenomics/lottewitjes/programs/SAPP/EnzDP.jar -input NG-
5593_1_interproscan.hdt -output NG-5593_1_enzdp.hdt

makeblastdb -in NG-5593_1A.fasta -dbtype nucl -parse_seqids -out NG-5593_1A

blastn -task blastn -db NG-5593_1A -query NG-5593_1A_trimmed_nonhuman_mRNA.fasta -
out NG-5593_1A.tsv -outfmt 6 -max_target_seqs 1 -num_threads 1
```

MegaBLAST against phage database

```
makeblastdb -in bacteria_phage_refseq_CDS.fasta -dbtype nucl -parse_seqids -
out bacteria_phage_refseq_CDS

makembindex -input bacteria_phage_refseq_CDS -iformat blastdb -
output bacteria_phage_refseq_CDS

blastn -task megablast -db bacteria_phage_refseq_CDS -query NG-
5593_1A_trimmed_nonhuman_mRNA.fasta -out NG-5593_1A.tsv -outfmt 6 -max_target_seqs 1 -
num_threads 1
```

7. SPARQL queries

BLASTX

SPARQL query to make the SAPP-based protein database

```
PREFIX gbol: <http://gbol.life/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?protein ?sequence WHERE {
    ?protein a gbol:protein .
    ?protein gbol:sequence ?sequence
}
```

SPARQL query to make SAPP-based Sha384-key – Pfam look-up table

```
PREFIX gbol: <http://gbol.life/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?protein ?acc WHERE {
    VALUES ?db {<http://gbol.life/0.1/db/pfam>} .
    ?protein a gbol:protein .
    ?protein gbol:xref ?xref .
    ?xref gbol:db ?db .
```

```
    ?xref gbol:accession ?acc .
}
```

SPARQL query to make SAPP-based Sha384-key – EC look-up table

```
PREFIX gbol: <http://gbol.life/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?protein ?acc WHERE {
    VALUES ?db {<http://gbol.life/0.1/db/ec>} .
    ?protein a gbol:protein .
    ?protein gbol:xref ?xref .
    ?xref gbol:db ?db .
    ?xref gbol:accession ?acc .
}
```

BLASTN

SPARQL query to make the SAPP-based gene database

```
PREFIX gbol: <http://gbol.life/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?gene ?lcsequence WHERE {
    ?gene a gbol:Gene .
    ?gene gbol:transcript ?transcript .
    ?transcript gbol:sequence ?sequence
    BIND (lcase(?sequence) as ?lcsequence
}
```

SPARQL query to make SAPP-based gene – protein sequence database

```
PREFIX gbol: <http://gbol.life/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?gene ?sequence WHERE {
    ?gene a gbol:Gene .
    ?gene gbol:transcript ?transcript .
    ?transcript gbol:feature ?cds .
    ?cds gbol:protein ?protein .
    ?protein a gbol:Protein .
    ?protein gbol:sequence ?sequence
}
```

SPARQL query to make SAPP-based gene – Pfam look-up table

```
SELECT DISTINCT ?gene ?acc WHERE {
    VALUES ?db {<http://gbol.life/0.1/db/pfam>} .
    ?gene a gbol:Gene .
    ?gene gbol:transcript/gbol:feature ?cds .
    ?cds gbol:protein ?protein .
    ?protein gbol:xref ?xref .
    ?xref gbol:db ?db .
    ?xref gbol:accession ?acc .
}
```

SPARQL query to make SAPP-based gene – EC look-up table

```
PREFIX gbol: <http://gbol.life/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?gene ?acc WHERE {
    VALUES ?db {<http://gbol.life/0.1/db/ec>} .
    ?gene a gbol:Gene .
    ?gene gbol:transcript/gbol:feature ?cds .
    ?cds gbol:protein ?protein .
    ?protein gbol:xref ?xref .
    ?xref gbol:db ?db .
    ?xref gbol:accession ?acc .
}
```

MetaSAPP

SPARQL query to extract Pfam counts per sample predicted with MetaSAPP

```
PREFIX gbol: <http://gbol.life/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?sample ?acc (COUNT(?acc) as ?acc_count)WHERE {
    VALUES ?db {<http://gbol.life/0.1/db/pfam>}
    ?sample a gbol:Sample .
    ?dnaobject gbol:sample ?sample .
    ?dnaobject gbol:feature ?gene .
    ?gene a gbol:Gene .
    ?gene gbol:transcript/gbol:feature ?cds .
    ?cds gbol:protein ?protein .
    ?protein gbol:xref ?xref .
    ?xref gbol:db ?db .
    ?xref gbol:accession ?acc .
}
GROUP BY ?sample ?acc
ORDER BY ?sample ?acc
```

SPARQL query to extract EC counts per sample predicted with MetaSAPP

```
PREFIX gbol: <http://gbol.life/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?sample ?acc (COUNT(?acc) as ?acc_count)WHERE {
    VALUES ?db { <http://gbol.life/0.1/db/ec> }
    ?sample a gbol:Sample .
    ?dnaobject gbol:sample ?sample .
    ?dnaobject gbol:feature ?gene .
    ?gene a gbol:Gene .
    ?gene gbol:transcript/gbol:feature ?cds .
    ?cds gbol:protein ?protein .
    ?protein gbol:xref ?xref .
    ?xref gbol:db ?db .
    ?xref gbol:accession ?acc .
}
GROUP BY ?sample ?acc
ORDER BY ?sample ?acc
```

SPARQL query to extract genes with sequences for every sample predicted with MetaSAPP

```
PREFIX gbol: <http://gbol.life/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?gene ?lcsequence WHERE {
    VALUES ?sample {<http://gbol.life/0.1/NG-5450_A>}
    ?sample a gbol:Sample .
    ?dnaobject gbol:sample ?sample .
    ?dnaobject gbol:feature ?gene .
    ?gene a gbol:Gene .
    ?gene gbol:transcript ?transcript .
    ?transcript gbol:sequence ?sequence .
    BIND (lcase(?sequence) as ?lcsequence) .
}
```