Insights into metabolic pathway representation from analysis of human metagenomic and metatranscriptomic samples

Koen van den Berg[a,1], **Victoria Pascal Andreu**[b], **and Marnix Medema**[c]

[a]BSc Plant Sciences, 970430053060; [b]PhD student Bioinformatics Department; [c]Assistant Professor Bioinformatics Department

June 4, 2018

Although some of the molecular mechanisms behind the function of the human microbiome are now well studied, most of the microbiota's primary metabolic pathways and their related compounds remain uncharacterized. In a complex ecosystem like the gut microbiome it is crucial to understand their functional capabilities that contribute to both microbial and host physiology. The current gap in knowledge is in part due to lack of tools that can find representation of pathways in microbiotic sequence data. Here a novel pipeline for finding such pathway representation is developed and probed on oral and gut metagenomic and metatranscriptomic data. In total, eight oral and stool samples were mined for representation of 15 pathways from anaerobic primary metabolism. The pipeline combines different existing tools with in-home scripts, including bowtie2 used for the mapping step. Our results confirmed that bowtie2 sensitive-local alignment is preferred among the other running modes and is able to find the representation of pathway sequences in bacterial metagenomes and metatranscriptomes. We found the pathways to be differently represented and expressed in the human oral and gut samples. In conclusion, this pipeline aims to metabolically profile the human metagenomic and metatranscriptomic samples given a set of pathways and will be extremely powerful to identify metabolic signatures for certain diseases and control and manipulate the bacteria contribution to overall host state.

Human Microbiome | Transcriptomics | Metagenomics | Primary Metabolic Pathways

Since 2005, the human microbiome and their relationship with human health and disease has been widely studied. It has become more obvious that the human digestive system goes beyond nutrient's breakdown function and it has been linked with disease ranging from allergies to bowel inflammation including autism spectrum disorders (1)(2). Recent studies have shown that the human gut and oral microbiomes appear to be unique for each individual, as no taxa were observed to be universally present among all body habitats (3). Oral and stool bacterial communities were especially diverse in terms of community membership and species composition. Although the human microbiome is not consistent among humans, research considering the general function of the human microbiome has shown that a general "core" function of the microbiome can be identified.

The human microbial community consists of thousands of distinct bacterial species of which 98% are anaerobic (4). These bacteria hold great potential, as they have a combined genetic potential that is 100 times greater than that of the human genome alone (5). These microbial species play an important role in food digestion and vitamin production and consequently form an symbiotic relationship with the human host. Important anaerobic phyla that contribute to overall diversity are the Firmicutes, Bacteroicides, Proteobacteria and Actinobacteria. The oral microbial community houses predominantly Proteobacteria while the gut microbial community mostly consists of Bacteroidetes. It has been shown that these microbial symbiotic communities provide traits that humans were not able to evolve on their own (3), which are important in the microbe-host interaction. Many of these traits consist of biochemical pathways that are not present in humans (6).

Although the underlying mechanisms by which these microbes interact with the human host are not fully understood, it has been shown that small molecules, such as short chain fatty acids (SCFA's) and aromatic amino acid derivatives, play a key role in host-microbe interactions (7). Three major SCFA's are acetate, propionate and butyrate. Acetate can cross the blood-brain barrier and can reduce appetite through a central homeostatic mechanism (8). Despite relatively low concentrations, propionate and butyrate can affect peripheral organs indirectly by activation of hormonal and nervous systems and can consequently help protect the host against cancers and inflammation (9). Amino acids are used by anaerobic bacteria as electron acceptors to drive fermentation, resulting in high levels of reductive amino acid metabolites, such as ρ-cresol and indolepropionic acid, which are not found in most other habitats. These derivative compounds are highly permeable and can thus accumulate persistently in the host, thereby impacting host physiology.

Efforts combining metagenomes or metatranscriptomes with metabolic gene clusters have been made to elucidate the metabolic pathways by which the gut microbes can interact with the human host (13). However, only a few studies have established a molecular connection between metabolites of the gut microbiota and human health. For example, Dodd et al. combined genetics with metabolic profiling and identified pathways that produce aromatic amino acid metabolites (11). Koh et al. combined research considering SCFA's and linked them to human host physiology (9). The issue remains that a great segment of studies is only associative, meaning that correlations between the composition of the gut microbiome and human health have been established while not being understood at the molecular level. The main reason is that the behaviour of the gut microbiota cannot be imitated *in vitro* due to their long-list growing requirements, which complicates the progress (14). Currently, it is a costly process to keep sam-

Victoria Pascal Andreu helped with supervising the project and with the writing process by providing feedback

[1] Correspondence should be addressed to koen.vandenberg@wur.nl

ples that are taken from the human gut alive and in the same conditions. Nevertheless, samples have been taken by some researches and have been analyzed to great extent. Despite a few studies that targeted their analysis in specific anaerobic pathways (of Dodd et al. and Koh et al.), the elucidation of the pathways involved in the production of key metabolites remains a challenge. In particular, the field requires novel approaches for finding bacterial pathway presence and expression in the human microbiome.

In this work, we conceived an innovative strategy to find the representation of bacterial pathways in human metagenomic and metatranscriptomic samples. Recent advances considering pathway analysis in microbial metagenomic or metatranscriptomic data have been made with the development of the HUMAnN (15) and MetaPath pipelines (16). There are however two application difficulties with these approaches in the context of finding unknown pathway representations. The first is that both approaches strongly rely on the KEGG database for pathway annotation. This means that only already annotated pathways can be analyzed for representation within samples. Second, both approaches use the sequence information to derive pathway information from the data. This complicates the process of inserting individual pathways of which the representation intends to be analyzed. In contrast to these approaches, we constructed a pipeline that can analyze not yet annotated pathways and enables individual pathway analysis.

For designing this pipeline, we have used gut and oral metagenomic and metatranscriptomic data to assess the representation of 15 primary metabolic pathways. These 15 pathways have been characterized in *Clostridium* and *Salmonella* species and are known to play essential roles in human metabolism. All pathways produce high abundant metabolites while the bacteria that express these pathways are often in low abundance. By assessing the representation of these pathways, we address an important knowledge gap, as their presence has not been quantified yet. The main goal of this research is to profile and investigate the representation of the aforementioned 15 metabolic pathways in the gut and oral microbiome. In this way, we will not only gain insights into pathway representation within the human microbiome, but also evaluate the proposed pipeline with our results.

Materials and Methods

In summary, this research used the metagenomic and metatranscriptomic stool and oral samples acquired by *Fransoza et al.* (14) for analysis of the representation of 15 metabolic pathways in the human microbiome. The 72 samples were downloaded using the NCBI SRA toolkit (17). Fastq files were extracted from the SRA-files using fastq-dump 2.3.5 (18) and checked for quality using FastQC v0.11.4 (19). 15 Metabolic pathways were provided by the Fischbach group (table 1). In order to capture all the variants that may exist of these pathways in other bacteria, results from MultiGeneBlast (20) using the complete Genbank database and each pathway as query were provided. This allowed us to manually choose 3 homologs per pathway. In order to evaluate the expression level of the pathways we included in the analysis 5 housekeeping genes from the Uniprot database. In the pipeline, bowtie2 is used to map the samples against the reference pathways and homologs with four local alignment methods. The resulting read mappings were counted using SAMtools version 0.1.19-96b5f2294a (21). Reads were normalized for read length and transcript length (TPM) during the counting (figure 1). The resulting raw read counts were normalized for library size using MetagenomeSeq (22). The results were visualized in R using the ggplot2 and pheatmap libraries. Significance was tested by performing a multiple hypothesis testing analysis with Bonferroni false positive correction. Details of these steps are described in the following sections and the performed analysis including code is deposited on Github (link).

Data description. The data originated from eight male subjects who provided stool and oral samples via three transportation modes: frozen, fixed in RNAlater and fixed in ethanol. DNA was extracted from all the samples and sequenced by Illumina HiSeq on one lane. In addition, RNA was extracted and converted to cDNA, which was sequenced using Illumina HiSeq on two separate lanes. The acquired reads were filtered to identify high-quality microbial DNA and RNA sequences (14). The samples are stored at the NCBI SRA run selector repository under SRP019038 project id. (link). The samples were downloaded to the server using the SRA-download toolkit from the NCBI. Using the fastq-dump program (version 2.3.5)

Table 1. Here the 15 primary metabolic pathways are shown of several *Clostridium* and *Salmonella* bacteria present in the human gut. The end products for each pathway are shown together with the reference.

	Metabolic Pathway	End Product(s)	Species
1	TMA cluster	Ethylamine (7)	*C. sporogenes ATCC15579*
2	Putative glutamate to butyric acid	Butyric acid (9)	*C. symbosium*
3	Propanediol utilization (pdu) Operon	Propionic acid (7) (9)	*S. enterica*
4	Proline to 5-aminovalerate cluster	5-aminovalerate (7)	*C. sticklandii*
5	ρ-cresol cluster	ρ-cresol (7)	*C. difficile*
6	Lysine degradation pathway	Butyric aci (9)	*C. sticklandii*
7	Leucine reductive branch	(E)-4-methyl-2-pentenoate (10)	*C. sticklandii*
8	Leucine reductive branch	(E)-4-methyl-2-pentenoate (10)	*C. sporogenes ATCC15579*
9	Ethylamine pathway (EUT pathway)	Ethylamine (7)	*S. typhimurium*
10	Aromatic Amino Acid reductive branch	IPA, 4-OH-PPA and PPA (11)	*C. sporogenes ATCC15579*
11	Butyrate pathway from acetate	Butyric acid (9)	*C. sporogenes ATCC15579*
12	A putative arginine degradation pathway	IPA, 4-OH-PPA and PPA (11)	*C. sticklandii*
13	Bai operon vpi	deoxycholic acid and lithocholic acid (12)	*C. scindens VPI12708*
14	Bai operon atcc	deoxycholic acid and lithocholic acid (12)	*C. scindens ATCC35707*
15	porA	phenylacetic acid (11)	*C. sporogenes ATCC15579*

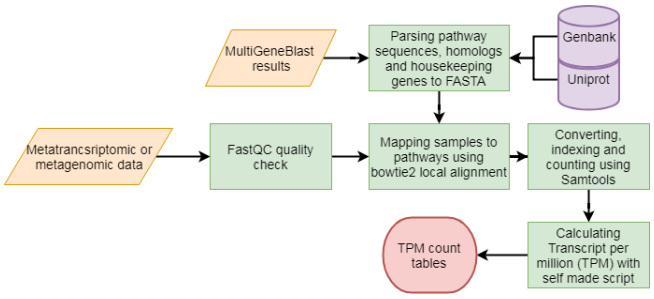

Fig. 1. Outline of the approach for finding pathway representation within metagenomic or metatranscriptomic samples. First the quality of the samples is evaluated. Using the MultiGeneBlast results, related homologs are identified and their sequences are parsed to FASTA using the GenBank whole genome sequences and pathway coordinates. Housekeeping genes are acquired from the Uniprot database. Next, bowtie2 local alignment is used to map the samples to the pathways. The read mappings are then counted and normalized by TPM calculation. This produces TPM count tables which are used as input data in the analysis.

the fastq-files were extracted from the SRA-files. In total, 72 fastq-files were collected from NCBI run selector of which 24 metagenomic singular DNA and 48 metatranscriptomic paired cDNA sequences.

Acquisition of pathway, homolog and housekeeping gene sequences. The 15 pathways that are used in this research originally stemmed from the Fischbach group located in the University of Stanford, America (unpublished data). They were provided as 15 separate files in GenBank format. What's more, the 15 pathways were run on a MultigeneBlast analysis which provided the most likely known homologs for these pathways. For each pathway three homologs have been collected manually by taking the hits with highest similarity and species divergence. In this way, a more generic representation for each of the 15 pathway sequences could be found. The homolog pathway FASTA sequences were acquired by parsing the whole genome sequence of the particular organism, which was first downloaded from GenBank using the Entrez module from Biopython, using the coordinates of the starting and end gene of the pathway (script link).

To assess the relative representation of the 15 primary metabolic pathways bacterial housekeeping genes were included in the analysis. The sequences of the bacterial housekeeping genes were obtained from the Uniprot database and are based on the findings of Rocha et al. (23). This included DNA gyrase A, DNA gyrase B, Recombinase A, RNA polymerase A and RNA polymerase B. The housekeeping genes were obtained from the *Salmonella* and *Clostridium* strains from which the 15 initial pathways were derived (table 1), which is 30 housekeeping genes in total. The inter-specific similarity of the sequences is tested to evaluate whether the sequences should be combined using the sum or the mean further down the pipeline. This multiple sequence alignment has been performed with ClustalW 2.1 (24). From the resulting alignment file a distance matrix is computed using distmat 6.6.0.0 (25) to obtain similarity values. Hereafter the 15 original pathways and the corresponding 45 homolog sequences were joined together with the 30 housekeeping genes as mapping reference.

Fastq data quality assessment. The quality of each sample was checked using the java-based program FastQC v0.11.4.

Table 2. Four local alignment methods used with varying seeding parameters: seed length (L), seed interval (i), seed mismatch (N), consecutive seeding fails (D) and reseeds (R)

method	L	i	N	D	R
very-fast	25	2.00	0	5	1
fast	22	1.75	0	10	2
sensitive	20	0.75	0	15	2
very-sensitive	20	0.50	0	20	3

This program considers the quality scores for each sequence in a sample and produces for each a quality control report. Each sample is checked separately and each quality control report is loaded into memory and parsed for the total number of sequences flagged as poor quality. Sequences are flagged as poor quality if the average phred score is lower than 30 for that sequence. The results were summarized and manually checked for number of poor quality sequences.

Read mapping and counting. The read mapping has been performed with the use of bowtie2 2.2.6 (26). First a bowtie2 index file is produced of the pathway, homolog and housekeeping gene sequences with default settings. Then the metagenomic and metatranscriptomic oral and stool samples were each individually mapped against the reference. The bowtie2 run has been performed with four local alignment settings: very-fast-local, fast-local, sensitive-local and very-sensitive-local (table 2). These methods enable bowtie2 to use four different seeding strategies during the initial alignment calculation of which very-fast-local is the most efficient method time-wise. The variable seeding parameters consist of seeding length (L), extracted seed interval (i), number of mismatches (N), number of consecutive seed extension fails (D) and number of reseeds (R). For the sensitive and very sensitive local alignment R and D are high while L and i are small, enabling higher sensitivity and accuracy. The resulting SAM files were converted to BAM format using SAMtools. Each of the metatranscriptomic samples is merged with its pair, which is found using the SRA runtable from the NCBI SRA run selector repository. The merging is performed with the merge function from SAMtools. The BAM files were sorted, indexed and counted using the sort, index and idxstats functions from SAMtools respectively.

Using the values of the raw read counts (X) and the length of the reads (l) the transcript per million (TPM) values are calculated with

$$TPM_i = \frac{X_i}{\bar{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_i}{l_i}} \right) \cdot 10^6,$$

where 10^6 is used to obtain values that are not too far below zero. TPM normalization allows for comparison of relative representation among samples, as the counts are normalized for read length and sequencing depth. Hereafter the TPM counts were written to a tabular format which was used in the result processing.

TPM count analysis in R. Normalization of the TPM count data for library size has been performed using MetagenomeSeq using default values. Log transformation of the data is performed to obtain interpretable numbers. Hereafter each pathway and their related homolog is aggregated over their sum into one conclusive pathway ($\sum_{i=4} TPM_i$). Next the

housekeeping genes are aggregated over sum and corrected for TPM count relative to the homologs by dividing over n housekeeping genes and multiplying with p pathways: $\frac{TPM}{n} \cdot p$. The samples, which consisted of 3 different collection methods, were combined using mean aggregation.

Visualisation and Multiple hypothesis testing. Heatmaps are produced using the pheatmap package (27). Ggplot2 (28) is used for the bar chart and line plots visualisations and PCA plots are generated with the help of the factoextra package (29). Significance is tested using multiple hypothesis testing using two-sided two-sample t-tests. False positive correction has been conducted using Bonferroni false positive correction. Plots of the adjusted p-values are used to determine the best correction method. The significance threshold is set to $p < 0.01$.

Results

In total, 72 samples have been checked for sequence quality and zero of those have been flagged as poor quality. For each sequence the length was 202bps. Multiple sequence alignment of the bacterial housekeeping genes shows similarity ranging from 30.81 up to 91.29 percent for each bacterial housekeeping gene.

Mapping of oral and stool samples to reference pathways and genes. After establishing the data quality of the samples, the fastq-files were mapped using four different local alignment methods. Alignment has been performed nine times in total, generating 584 sam files. Count parsing and TPM calculation resulted in nine count tables of which four contained metagenomic samples, four contained metatranscriptomic samples and one oral samples. Mapping was successful for every pathway, homolog and bacterial housekeeping gene.

Evaluation local alignment methods. In line with the expectations, the very-fast-local seeding strategy showed to be the least sensitive method yielding on average the lowest TPM count for each pathway and bacterial housekeeping gene (figure 2). However, this method was able to run the samples

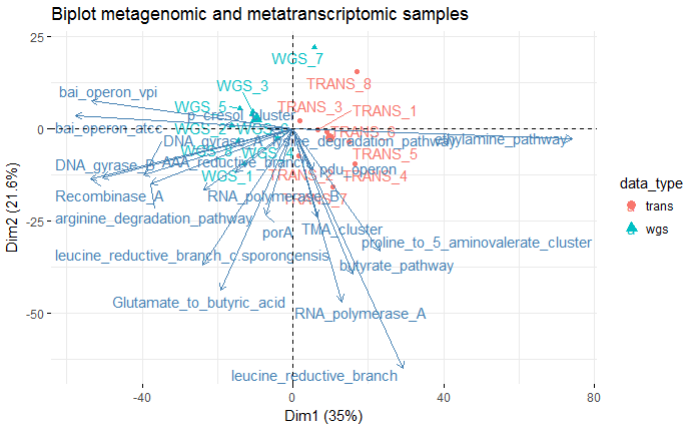

Fig. 3. Principal component biplot based on the stool TPM counts of the sensitive-local alignment method plotted for the first and second principal component. Arrows indicate the loading vectors for the pathways and bacterial housekeeping genes. Samples are indicated by either a dot (transcriptomic data) or a triangle (metagenomic data).

the fastest compared to all the other tested methods. What's more, very-sensitive-local alignment yielded overall the most counts for each pathway and bacterial housekeeping genes, although the difference with the sensitive-local method is not notable. The sensitive-local alignment method was not able to capture DNA gyrase A and DNA gyrase B the same as the very-sensitive-local alignment method in both the metagenomic and metatranscriptomic samples. For further analysis of the representation of the primary metabolic pathways in the stool and oral samples the sensitive method is used as this method shows the best sensitivity in the smallest loading time.

The bacterial housekeeping genes show overall the largest average TPM count ranging from 33 to 44. In comparison with the primary metabolic pathways, the bacterial housekeeping genes are most represented in both the metagenomic and metatranscriptomic samples. Representation of the bile acid induced operon (bai operon) is relatively large in both the metagenomic and metatranscriptomic samples. The ethy-

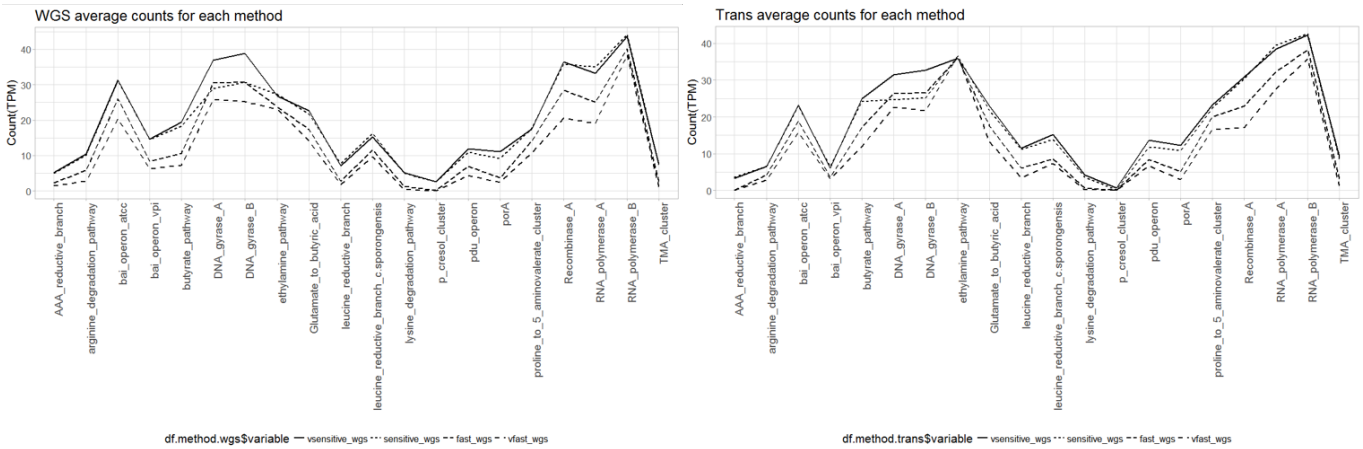

Fig. 2. Four different local alignment methods have been executed for bowtie2 mapping evaluation using the stool metagenomic and metatranscriptomic data. Homologs have been aggregated for each pathway and bacterial housekeeping gene to obtain the average TPM counts. Plots indicate the average TPM count per pathway for each method. Left the whole genomic sequences (WGS) or metagenomes and right the metatranscriptomic sequences are shown. In both plots the solid line represents the very-sensitive method, the dotted line the sensitive method, the dashed line the fast method and the dotdash line the very-fast method.

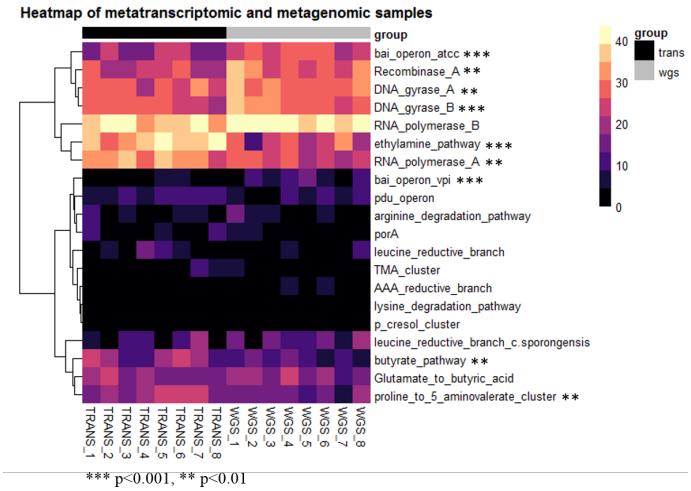

Fig. 4. Heatmap of the metagenomic and metatranscriptomic stool samples based on the normalized TPM count data. The transcriptomic samples are grouped towards the left (black bar) and the metagenomic samples towards the right (grey bar). Colours represent different TPM count values. The heatmap has been produced using the Pheatmap package in R. The asterisks next to the pathway names are indicative for the significance values that were computed during the multiple hypothesis testing using two-sided t-tests and Bonferroni false positive correction.

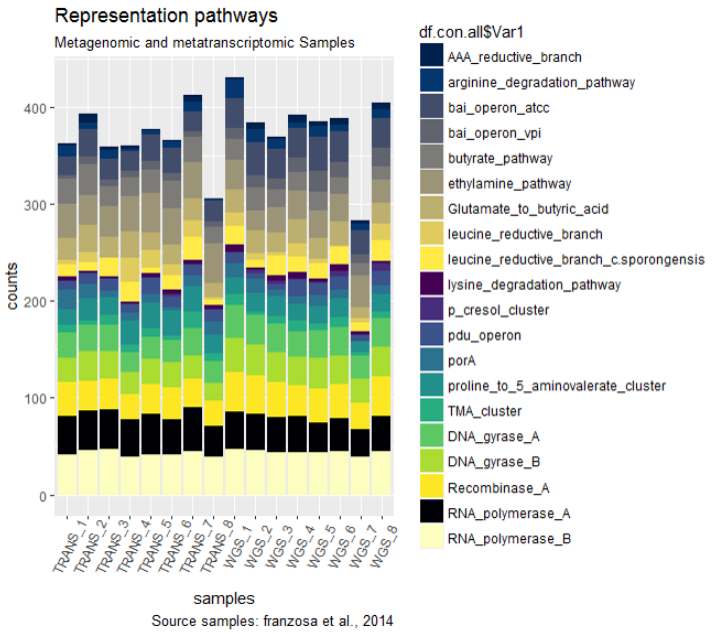

Fig. 5. Stacked barplot of the representation of the 15 primary metabolic pathways and five bacterial housekeeping genes based on the normalized TPM counts. Pathways are represented by differing colours and metatranscriptomic samples are grouped towards the left (TRANS) and metagenomic samples are grouped towards the right (WGS).

lamine pathway shows high representation in only the metatranscriptomic samples with an average TPM count of 30.7 for all methods.

Comparison between pathway encoding on the metagenome and pathway expression in the gut microbiome. For the comparison between the representation of the 15 primary metabolic pathways between the metagenomic and metatranscriptomic samples biplots, heatmaps and stacked barcharts have been generated. Again, pathways have been aggregated with the homolog data over sum as well as the bacterial housekeeping gene data. A clear distinction is found between metagenomic and metatranscriptomic samples using principal component analysis. Principal components are based on the TPM count data for all eight samples. The first and second principal component included 35% and 21.6% of the total variance respectively (figure 3). In total, 16 principal components have been computed and more than 90% of the variance was captured within the first 7 PC's. Large within group variation is found for the eight men. Based on the loading vectors in the biplot, which are represented by the arrows, the ethylamine pathway is uncorrelated with both bai operon pathways. Also, metagenomic sample 7 (WGS_7) is an outlier relative to the other metagenomic samples, which was unexpected as samples are from individuals from the same cohort.

The bai operon from *C. scindens ATCC35707*, the bai operon from *C. scindens VPI12708* and DNA gyrase B are significantly differently expressed with p-values of $1.507e-09$, $4.73e-07$ and $3.21e-04$ respectively (figure 4). The heatmap indicates that they have an higher representation in the metagenomic samples compared to the metatranscriptomic samples. The ethylamine pathway is significantly differently expressed in the metatranscriptomic samples compared to the metagenomic samples with a p-value of $1.58e-07$. Also, the butyrate pathway and the proline to 5-aminovalerate cluster are significantly differently expressed in favour of the metatranscriptomic samples with p-values of $3.20e-03$ and $2.21e-03$ respectively. Significance is not found for the other primary metabolic pathways. Moreover, in comparison with the five bacterial housekeeping genes both the ethylamine pathway and the bai operon from *C. scindens ATCC35707* are steadily expressed within the gut microbiome. For the bai operon from *C. scindens VPI12708* this is not the case as expression is notably lower than that of the bacterial housekeeping genes.

Analysis of the representation of the pathways in the stool microbiome samples showed again that sample 7 is an outlier compared to the other samples (figure 5). Bacterial housekeeping genes are approximately evenly distributed within both the metagenomic and metatranscriptomic samples. What's more, both sample types retain a total TPM count value of around 375. The porA pathway, ρ-cresol cluster, lysine degradation pathway and the aromatic amino acid (AAA) reductive branch are not highly represented in both sample types while the ethylamine pathway is highly represented.

Comparison of stool and oral metagenomic samples. Representation of the 15 primary metabolic pathways in stool metagenomic and oral metagenomic samples have been compared based on the normalized TPM counts. A notable separation has been observed between both samples and 71.4% of the total variance is captured within the firs PC and 8.4% within the second PC (figure S2). Separation between oral and stool sample is predominantly found on the first PC while within sample variation is found on the second PC. Both the oral and stool metagenomic samples show approximately the same within sample distribution.

Stool metagenomic samples have an overall higher representation of the 15 primary metabolic pathways than the oral metagenomic samples (figure 6). The plot with the adjusted

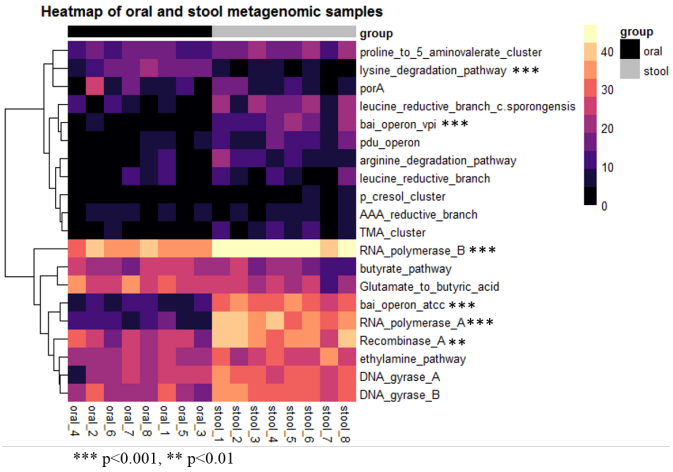

Fig. 6. Heatmap of the oral and stool metagenomic sampels based on the normalized TPM count data. The oral metagenomic samples are grouped towards the left (black bar) and the stool metagenomic samples towards the right (grey bar). Colours represent different TPM count values. Heatmap has been produced using the Pheatmap package in R. The asterisks next to the pathway names are indicative for the significance values that were computed during the multiple hypothesis testing using two-sided t-tests and Bonferroni false positive correction.

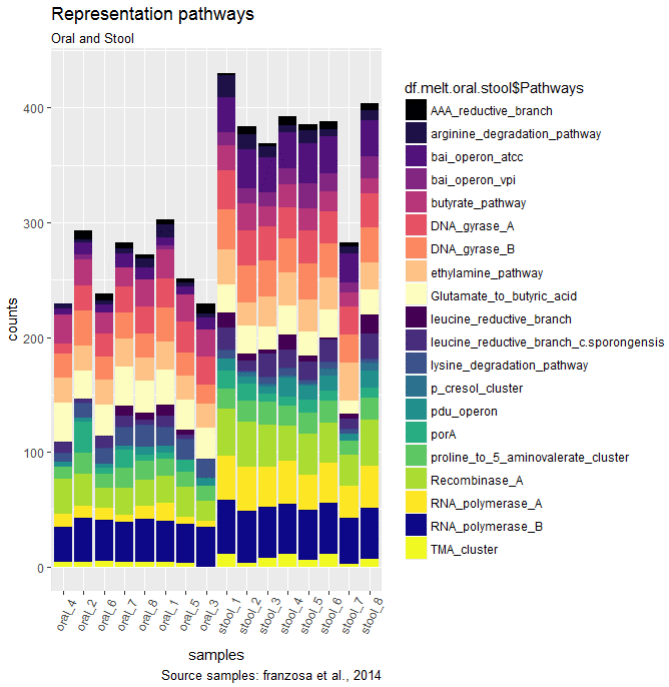

Fig. 7. Stacked bar-plot of the representation in the oral and stool metagenome of the 15 primary metabolic pathways and five bacterial housekeeping genes based on the normalized TPM counts. Pathways are represented by differing colours and oral samples are grouped towards the left and stool samples are grouped towards the right.

p-values shows that Bonferroni is the strictest method for false positive correction (figure S3). Recombinase A, RNA polymerase A, the bai operon from *C. scindens ATCC35707*, RNA polymerase B and the bai operon from *C. scindens VPI12708* are significantly more represented in the stool metagenomic samples with p-values of $2.58e-03$, $1.59e-08$, $2.32e-09$, $4.08e-05$ and $2.02-04$ respectively. The lysine degradation pathway is the only pathway that is significantly more represented in the oral metagenome with a p-value of $3.76e-04$ which is well below the threshold of $p = 0.01$. In comparison with bacterial housekeeping genes the representation of the lysine degradation pathway does not show high occurrence.

The 15 primary metabolic pathways and the five bacterial housekeeping genes are notably higher represented in the stool samples than in the oral samples (figure 7). Of the bacterial housekeeping genes, RNA polymerase B, recombinase A and DNA gyrase A & B are evenly distributed in both the oral and stool metagenomes. RNA polymerase B is significantly higher represented in the stool samples than in the oral samples ($p = 4.08e-05$). The TMA cluster is present in all oral samples except for sample three. Also the arginine degradation pathway is missing in the fourth oral sample.

Discussion and conclusions

In the current research a novel approach is established to determine the representation of bacterial metabolic gene clusters in bacterial metagenomes and metatranscriptomes. This method extends previous approaches, such as the HUMAnN or MetaPath pipeline, by enabling analysis of not yet annotated bacterial gene clusters. The HUMAnN pipeline makes use of the KEGG database to annotate the genes of interest and uses translated BLAST and MinPath to in the end dissect pathway representation, making the pipeline dependent on pathway annotations. This process makes it difficult to input individual pathway sequences of interest. Conversely, in the current approach pathways sequences that are not yet annotated can be individually evaluated in the pipeline and mapped against metagenomic or metatranscriptomic samples. This in the end provides new means to research any pathway representation in any bacterial metagenomic or metatranscriptomic sample, given that the DNA or cDNA pathway sequence is available. It is expected that this approach will have a marked effect on studies considering the functionality of the human microbiome, increasing both the accuracy and scope of conclusions. Although in this research only oral and stool samples are taken into account, this approach is highly extensible and applicable to other microbial communities in the human body as well.

In addition to producing a new approach for finding pathway representation within bacterial metagenomes and metatranscriptomes, this work also evaluated the mapping procedure using 4 different seeding strategies. In this way, the different parameters of the bowtie2 local alignment could be evaluated for sensitivity and time-efficiency. The current findings suggest that the very-sensitive-local method is not that different from the sensitive-local method. These results are in line with the findings of Langmead & Salzberg as they concluded that bowtie2 very-sensitive-local is 2 times slower than sensitive-local while only increasing 0.34% in sensitivity (26). This suggests that very-sensitive-local alignment is not required for most experiments, unless when mapping against very big references where a 0.34% difference could cause great changes in the results. For the current pipeline it is not expected that the reference will reach that size, as the number of pathways should be kept rather low to aid experiment overview. Therefore, the bowtie2 sensitive-local alignment method is the default running mode.

Furthermore, placed in the context of this new approach,

we also combined results from gut metagenomes and metatranscriptomes to probe primary metabolic pathway representation in the human gut microbiome and to evaluate the conceived pipeline. Intriguingly, the results indicate that the bai operons from two species are both higher represented on the metagenome while showing different expression levels, indicating that expression levels are strain dependent. The main function of the bai operon is 7α-dehydroxylation of primary bile acids to deoxycholic acid (DCA) and lithocholic acid (LCA). It has been observed that bacteria that perform this transformation are present in low abundance in the human gut, yet they can dehydroxylate an extensive amount of bile acids, producing high concentrations of DCA and LCA that can reach pharmaceutical concentrations (12) (data not published). Also, a distinction between "high" and "low" activity 7α-dehydroxylating strains has been made and *C. scindens VPI12708* has been characterized as a "high" activity strain. This means that *C. scindens VPI12708* is one of the strains that can dehydroxylate great amounts of bile acids. Remarkably, our results contradict this in that we find low representation and expression of the bai operon from *C. scindens VPI12708* which is in contrast with its characterization as a "high" activity strain. Perhaps this difference is caused by the absence of *C. scindens VPI12708* within the analyzed samples. On the contrary, we find that *C. scindens ATCC35707* shows high representation within the metagenome in comparison with the bacterial housekeeping genes and might thus be one of the strains that plays a role in the high-throughput bile acid dehydroxylation which can be characterized as "high" activity. To evaluate the current finding, further research could be focused on finding connections between meta-data, such as diet and smoking habits, and bai operon representation to be able to confirm the absence of *C. scindens VPI12708* within stool samples.

In addition, the butyrate pathway and ethylamine pathway are higher represented in the metatranscriptome of gut microbiota. In comparison with the bacterial housekeeping genes, both pathways show high expression indicating that both pathways are highly active in the human gut microbiome. The butyrate pathway is known to synthesize butyrate, which is a short chain fatty acid (SCFA), from lactate or acetate. Butyrate can act as a inducer of hormonal and neural activity and may act as modulator of cancer. In line with the current findings, Koh et al. states that butyrate is highly present in the gut lumen. Here we find that the butyrate pathway is expressed within the gut microbiota, and thus its end product is butyrate present in healthy human individuals. Although this finding is only a confirmation of previous work, it remains relevant as butyrate is a major contributor to human health (9). Additionally, previous observations indicate that the ethylamine pathway is highly expressed in the gut environment (unpublished data). Our results confirm this observation, as we find high representation of this pathway on the metatranscriptome, meaning that this pathway is indeed highly expressed within the human gut microbiome. This result might add to future research considering the ethylamine pathway and its role in microbe and host physiology.

Consequently, we speculate based on the oral data analysis that the lysine degradation pathway has been revealed to play a notable role in the oral human microbiome. Our findings list the lysine degradation pathway as the only pathway that is more represented in the oral metagenome than in the gut metagenome. This is made even more notable by the fact that the oral metagenome is relatively smaller in general compared to the stool metagenome. The lysine pathway is known to produce butyrate from other proteins in the gut (9) when other nutritional resources are not available. However, further research is needed to be able to connect the current findings to the general function of the oral microbiome.

It is however, important to note the limitations of the current research. Here we used bowtie2 local alignment for the alignment of the sequence data, but it is not the only available package for performing local alignment. Meta-aligner could also have been used as this software is particularly developed for longer read alignments (>300bps). The reads in this research were 202bps long and were aligned with four bowtie2 local alignment settings. However, for sequences of 300bps long, Meta-aligner performed with a precision of 99.8% where bowtie2 had a precision of 97.7% within the same time-frame (30). Therefore, this specific analysis could be improved by implementing Meta-aligner instead of bowtie2 local alignment. It is however likely that future reads will not exceed 202bps in length, thus making bowtie2 local alignment the best general tool for this pipeline.

An important factor in this research was the inclusion of homolog sequences for each of the 15 pathways in the analysis. The homologs have been selected using the MultiGeneBlast results and have been used to extend the representation for each of the 15 pathways. This was done to obtain a more generic idea of the pathway representation in the human microbiome. This had a notable influence on the final observed representation for each pathway, as the TPM counts of the homologs were added to the TPM counts of the pathways resulting in a higher overall TPM count for each pathway. However, many of the homologs showed different expression patterns compared to the related pathway and to each other. These differences can be attributed to the species from which the homolog sequence was derived, as some species have a high abundance in the human microbiome and are thus expected to have an higher representation than low abundance species. So by only taking three homolog sequences for each pathway, we did not accurately measure the pathway-level representation, because due to the mentioned species dependency each pathway will get different TPM count additions. To correct for this we would need to identify and include as many homolog sequences as possible for each pathway, so that all the species in which a particular pathway can be found are included. Further research could consider all the homolog sequences for the pathway by using all the hits of the MultiGeneBlast results and using them in a comparable analysis.

Another limitation in the results analysis is that only 8 samples are analyzed for metabolic pathway representation and that these samples were from the same cohort. More reliable conclusions could be drawn by increasing the sample population size, so that a bigger part of the human population could be analyzed and more significant conclusions could be drawn. Also, including samples acquired from patients who are diseased, for example inflammatory bowels disease or obese patients, will enable further more profound conclusions considering the representation of primary metabolic pathways in healthy humans. Also missing in this research was the metadata of the 8 samples. By also including the metadata

in the research the results could have been linked to patient behaviour or other factors that could play a role.

In the future, these novel insights pave the path towards the understanding of the human microbiome and its function in human physiology. We have shown that the representation of primary metabolic pathways might be used to identify pathways which play important roles in this metabolic interaction. Also, by developing the novel pipeline, further research for finding representation of any pathway in any metagenomic or metatranscriptomic samples has never been easier.

ACKNOWLEDGMENTS. I thank Victoria Pascal Andreu for supervising the project and giving powerful tips on improving the research as a whole. Marnix Medema, for giving some tips and feedback at the halfway mark of the project and on the research paper.

References

1. Guinane CM, Cotter PD (2006) Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ. *Therapeutic advances in gastroenterology* 6(4):295–308.
2. Lagier JC, Million M, Hugon P, Armougom F, Raoult (2012) Human gut microbiota: repertoire and variations. *Frontiers in cellular and infection microbiology* 2(136):1–19.
3. Huttenhower C, et al. (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–214.
4. Ley RE, Peterson DA, Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 124(4):1–19.
5. Fujimura KE, Slusher NA, Cabana MD, Lynch SV (2010) Role of the gut microbiota in defining human health. *Expert review of anti-infective therapy* 8(4):435–454.
6. Yadav M, Verma MK, Chauhan NS (2018) A review of metabolic potential of human gut microbiome in human nutrition. *Archives of microbiology* 200(2):203–217.
7. Donia MS, Fischbach MA (2015) Small molecules from the human microbiota. *Science* 349(6246):395.
8. Frost G, et al. (2014) The short-chain fatty acid acetate reduces appetite via a central homeostatic mechanism. *Nature communications* 5(3611).
9. Koh A, De Vadder F, Kovatcheva-Datchary P, Bäckhed F (2016) From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. *Cell* 165(6):1332–1345.
10. Dickert S, Pierik AJ, Buckel W (2002) Molecular characterization of phenyllactate dehydratase and its initiator from clostridium sporogenes. *Molecular microbiology* 44(1):49–60.
11. Dodd D, et al. (2017) A gut bacterial pathway metabolizes aromatic amino acids into nine circulating metabolites. *Nature* 551(7682):648.
12. Ridlon JM, Kang DJ, Hylemon PB (2010) Isolation and characterization of a bile acid inducible 7α-dehydroxylating operon in clostridium hylemonae tn271. *Anaerobe* 16(2):137–146.
13. Donia MS, et al. (2014) A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* 158(6):1402–1414.
14. Franzosa EA, et al. (2014) Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences* 111(22):E2329–E2338.
15. Abubucker S, et al. (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS computational biology* 8(6):1–17.
16. Liu B, Pop M (2011) Metapath: identifying differentially abundant metabolic pathways in metagenomic datasets in *BMC proceedings*. (BioMed Central), Vol. 5, p. S9.
17. Leinonen R, Sugawara H, Shumway M, Collaboration INSD (2010) The sequence read archive. *Nucleic acids research* 39(suppl_1):D19–D21.
18. NCBI (2018, access date: 2018-05-29) *SRA Toolkit Documentation*. https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_docf=fastq-dump.
19. Bioinformatics SAB (2018, access date: 2018-05-29) *FastQC*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
20. Medema MH, Takano E, Breitling R (2013) Detecting sequence homology at the gene cluster level with multigeneblast. *Molecular biology and evolution* 30(5):1218–1223.
21. Li H, et al. (2009) The sequence alignment/map format and samtools. *Bioinformatics* 25(16):2078–2079.
22. Paulson JN, Pop M, Bravo HC (2013) *metagenomeSeq: Statistical analysis for sparse high-throughput sequncing.* Bioconductor package.
23. Rocha DJ, Santos CS, Pacheco LG (2015) Bacterial reference genes for gene expression studies by rt-qpcr: survey and analysis. *Antonie Van Leeuwenhoek* 3(108):685–693.
24. Thompson JD, Higgins DG, Gibson TJ (1994) Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* 22(22):4673–4680.
25. Carver T (2001, access date: 2018-05-29) *Distmat Emboss*. http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/distmat.html.
26. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. *Nature methods* 9(4):357.
27. Kolde R (2015) *pheatmap: Pretty Heatmaps*. R package version 1.0.8.
28. Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York).
29. Kassambara A, Mundt F (2017) *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.5.
30. Nashta-ali D, et al. (2017) Meta-aligner: long-read alignment based on genome statistics. *BMC bioinformatics* 18(1):126.

Supplementary material

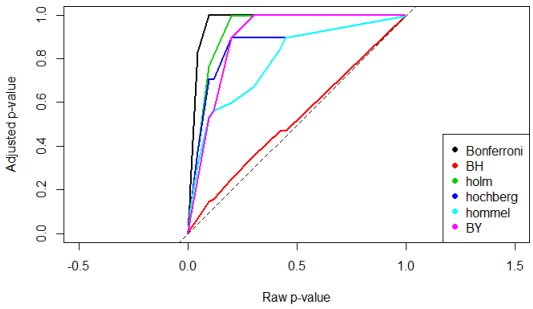

Fig. S1. plot that shows the raw p-values on the x-axis and the adjusted p-values on the y-axis for the comparison between stool metagenomic and metatrancsriptomic samples. adjustments methods used are bonferroni, benjamini-hochberg (bh), holm (bonferroni step-down), hochberg, hommel and benjamini-yekutieli (by).

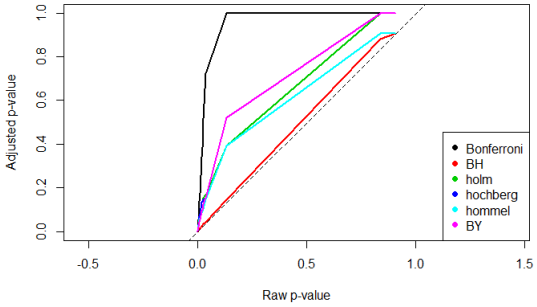

Fig. S3. plot that shows the raw p-values on the x-axis and the adjusted p-values on the y-axis of the multiple hypothesis testing for the oral versus stool. adjustments methods used are bonferroni, benjamini-hochberg (bh), holm (bonferroni step-down), hochberg, hommel and benjamini-yekutieli (by).

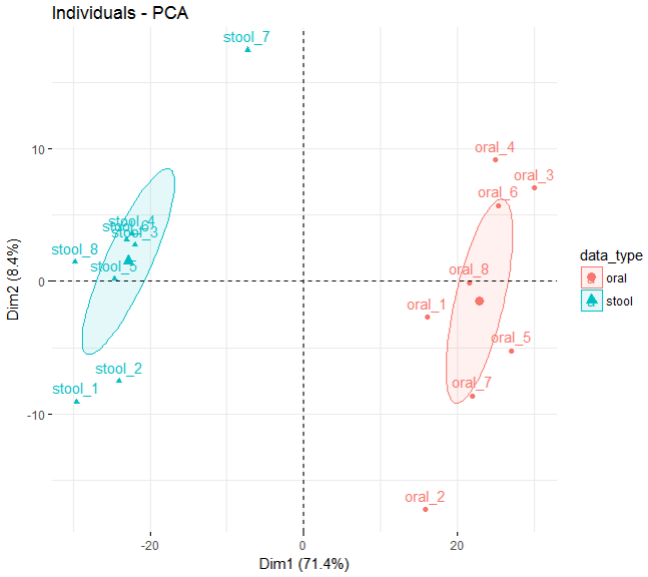

Fig. S2. Principal component biplot based on the oral TPM counts of the sensitive-local alignment method plotted for the first and second principal component. Arrows indicate the loading vectors for the pathways and bacterial housekeeping genes. Samples are indicated by either a dot (oral data) or a triangle (stool data).