Spatial Variability of the microbiota along the GI tract in OligoMM mice and Quality Assessment of the Taxonomic Binning after the infection with salmonella

551-1119-00L Microbial Community Genomics and Transcriptomics

Abstract

The gut microbiota consists of numerous bacteria species which co-exist and have an influence on many metabolic functions of the gastrointestinal tract and other organs of the host. The microbial community composition is being affected through dietary factors, health and disease and is thus of great interest in microbial community research. However, the distribution of the bacteria species along the different GI tract sections is less understood. Here we show the abundance and the activity of 12 OligoMM strains in gnotobiotic mice and the quality of the taxonomic binning step in metagenomics analysis. Our findings contain the comparison of the number of DNA to cDNA reads in the 7 sections of the GI tract and the feces for 9 of the 12 OligoMM strains with p-values below 1.1%. They show the diversity and difference in activity of the induced OligoMM strains in the 8 analysed sections in mice. These insights imply the possibility to assign the different bacteria species present in the gut to their functionality and to further analyse the host-microbiota interactions. Additionally, we achieved to benchmark the quality of the taxonomic binning step in metagenomics analysis of salmonella infected OligoMM mice and found co-occuring strains which were shown to be more closely related than others. Lachnoclostridium and Blautia and furthermore Bacteroides, Lachnoclostridium and Erysipelotrichaceae were found to often co-occur in one bin. We determined the percentage of all bins, which were assigned to one single species by at least 95%, to be 75.8%. These findings can be used to further improve the binning step in metagenomics analyses and to iteratively assess a quality score for the binning.

Introduction

Microbial communities consist of numerous bacteria species which include species that cannot be cultivated solely and thus often remain unclassified. The gut microbiota is known to comprise the largest population of bacteria in an organism's body and is thus of big interest in host-microbiota interactions research [1]. The microbial community composition is not stable during the lifetime and is known to be affected through e.g. dietary factors [2], health issues [3] and antibiotic treatments and vice versa. It is therefore of big importance to further analyse the microbial community composition along the different sections of the gut and to determine the functional context of the present microbiota. The introduction of 12 well-known and common bacteria strains (OligoMM strains) in gnotobiotic mice provides a simplified model of the gut microbiome which is easily modifiable and reproducible.
I speculated that the abundance and activity of the present bacteria vary along the gastrointestinal tract and differ greatly among the different bacteria species. Through a 16S rRNA gene analysis the microbiota was taxonomically classified and with DNA and RNA amount measurements, the abundance and activity of the different bacteria species in the different GI tract sections was determined. Furthermore, the change of the microbial community composition due to a salmonella enterica serovar infection was of interest. Therefore, five OligoMM mice were orally administered either virulent or avirulent salmonella and fecal samples were analysed with metagenomics. The aim of this study was to benchmark the quality of the taxonomic binning step to further improve the binning and the downstream analysis of the metagenomic data. Additionally, I was interested in finding patterns of species that tend to co-occur in the created bins to receive a greater insight into the influencing factors of the binning.

Materials and Methods

Six gnotobiotic mice were inoculated with 12 known bacterial species (OligoMM strains, Suppl. 1) that represent a simplified model for the gut microbiota of mice. DNA and RNA samples were taken from the six sections of the gastrointestinal tract (Duodenum, Jejunum, Ileum, Caecum, Proximal Colon, Distal Colon) and the feces. The RNA was converted into cDNA with a reverse transcriptase. The following steps were performed separately for the DNA and the cDNA content. The variable V4 region of the 16S rRNA gene was amplified with a forward and reverse primer through PCR. The amplicons were merged and low-quality reads were discarded through quality control containing the definition of a threshold for the maximum of expected errors and primer match filtering with cutadapt. During dereplication with usearch, all reads were aligned against each other to receive unique sequences with a corresponding number of identical reads. The unique sequences were clustered into operational taxonomic units (OTUs) with at least 97% sequence similarity. The resulting OTU tables containing the number of reads in each OTU per sample were further analysed with the programming language R. Through running a principal component analysis (PCA) the OTUs of interest regarding the abundance in the different sections or the activity were determined. The absolute number of reads of DNA and cDNA and the ratio cDNA/DNA was determined for all OTUs of interest. The OTUs were annotated with the taxonomy table resulting from the processing of the sequenced data and with a local BLAST against the OligoMM strains.

In a second step, five gnotobiotic OligoMM mice were orally administered either virulent or avirulent salmonella enterica serovar and one fecal sample was taken per mouse per day until the mice were sacrificed after 4 days. The samples were processed with metagenomics and assembly. The metagenomes were shotgun sequenced and the low-quality reads were discarded through quality control such as removing adapter sequences, PhiX reads and host contamination and cutting off low-quality tails. The high-quality reads were assembled to longer contigs. The contigs were binned based on the tetranucleotide frequency with metabat and the contigs in each bin were aligned to a local database including the 12 bacteria strains and salmonella with the aligner nucmer. The total number of contigs assigned to each species and the percentage of contigs assigned to each species were calculated for each bin. Further analysis was performed to determine the percentage of bins that were assigned to one single species and to find patterns of species that often co-occur in bins.

Results

Comparison of the activity and abundance of OligoMM bacterial strains along the gastrointestinal tract of mice. The variable V4 region of the 16S rRNA gene of the DNA and RNA samples from the six sections of the GI tract and the feces were amplified through PCR. The amplicons were merged, reviewed through quality control and clustered into OTUs. Four samples with a sequencing depth below 50000 reads were discarded for both DNA and cDNA. The OTU tables of the DNA and cDNA, containing the number of reads in each OTU for each sample, were merged and analysed. The samples were clustered with the Bray-Curtis similarity measure (Fig. 1) but there was no clear grouping into samples from different GI tract sections visible. Ten OTUs were found to be interesting for further downstream analysis by means of a Principal Component Analysis (Fig. 2). The ten OTUs of interest were annotated with the alignment against a local reference database including the 12 OligoMM strains and Salmonella and the annotations were verified with a local blast (Table 3). Nine of the ten OTUs were assigned to the OligoMM strains and OTU 17 was determined to originate from mitochondrial DNA of mice through an online blast against the nr/nt reference database. Additionally, OTU 7 was assigned to one of the OligoMM strains with the local blast but it was not expected to be interesting by means of the PCA and was not further analysed. For the OTUs of the PCA, the numbers of reads of DNA and cDNA in each section (Fig. 4) and the ratio between cDNA and DNA (Fig. 5) were calculated and plotted. To verify the statistical significance of the findings a Kruskal-Wallis test was performed for each OTU of interest along the sections (Table 6).

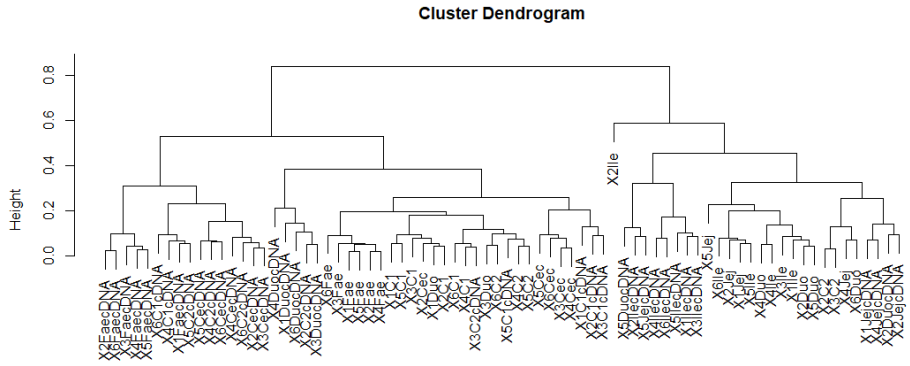

Fig. 1. Clustering of all remaining samples after preprocessing the data. The Bray-Curtis dissimilarity index was used as a similarity measure for the clustering.

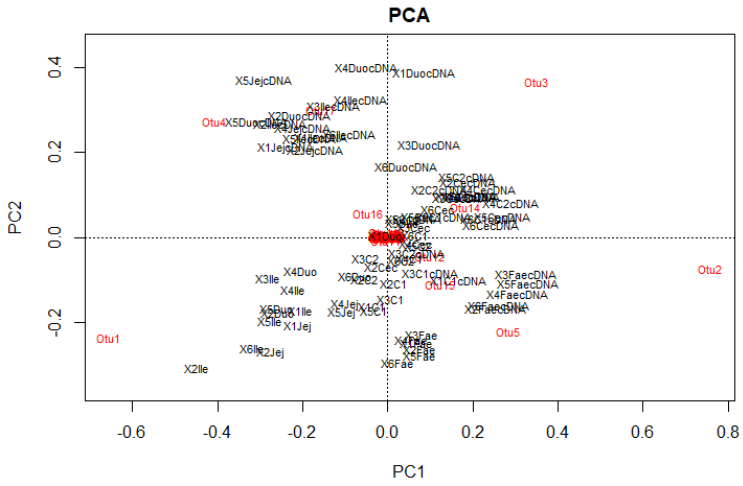

Fig. 2. Principal Component Analysis (PCA) showing the variability of the bacterial abundances in the 7 analysed sections along PC1 and the variability of the activity of the bacteria along PC2. The ten following OTUs were expected to be interesting regarding the abundance in different sections or the activity: 1, 2, 3, 4, 5, 12, 14, 15, 16, 17.

Table 3. Taxonomic annotation of the OTUs of interest with the alignment against a local reference database including the 12 OligoMM bacteria strains and Salmonella. The annotations were verified with a local blast. Through this step OTU 7 was also annotated.

OTU 1	Akkermansia muciniphila YL44
OTU 2	Bacteroides caecimuris I48
OTU 3	Lachnoclostridium YL32
OTU 4	Parabacteroides YL27
OTU 5	Blautia YL58
OTU 7	Lactobacillus reuteri I49
OTU 12	Enterococcus faecalis KB1
OTU 14	Flavonifractor plautii YL31
OTU 15	Erysipelotrichaceae bacterium I46
OTU 16	Bifidobacterium animalis YL2
OTU 17	Mus musculus mitochondrial DNA

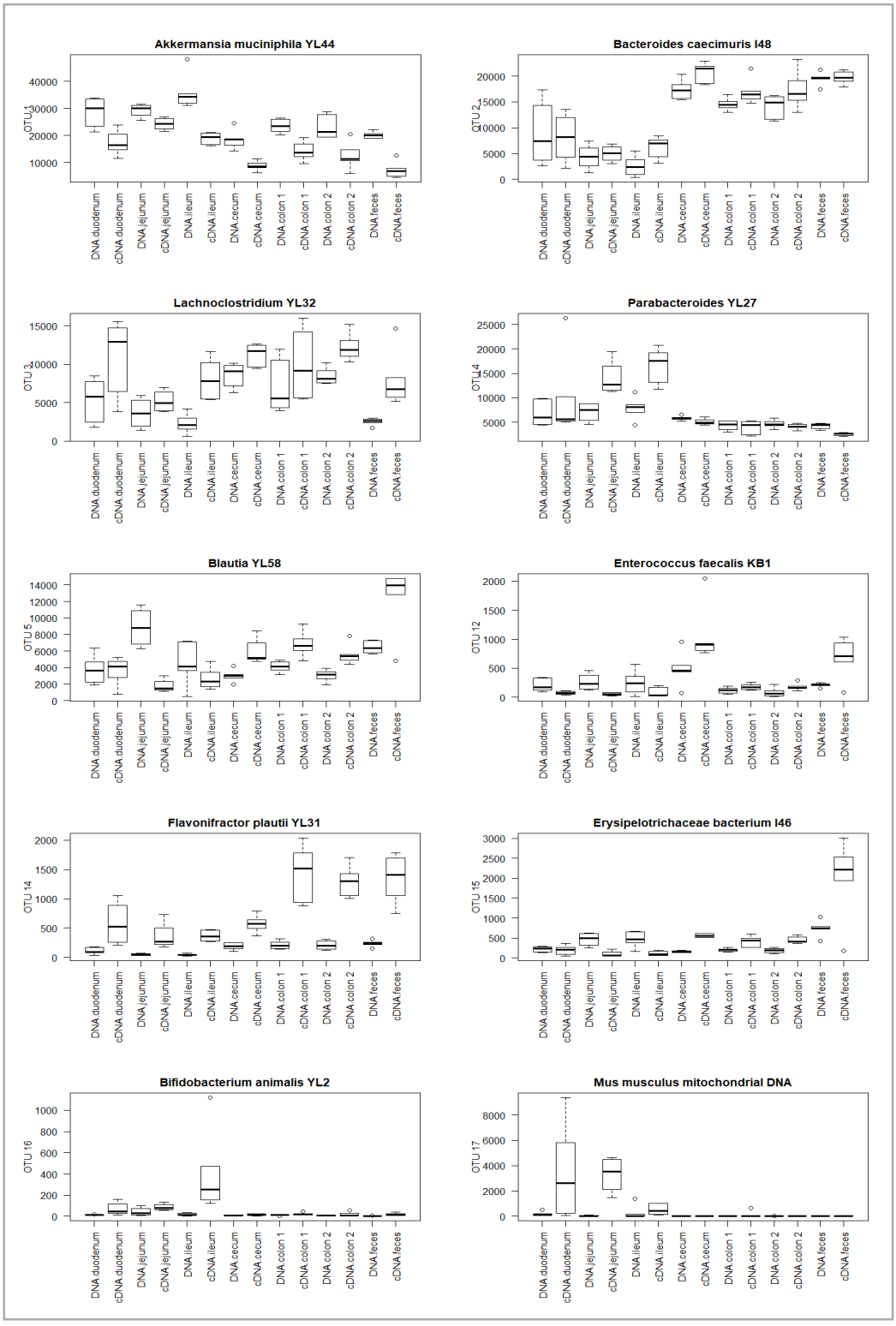

Fig. 4. Total number of DNA and cDNA reads in the ten OTUs that were expected to show a significant difference in abundance or activity from the PCA. They are ordered along the 7 sections of the GI tract and feces.

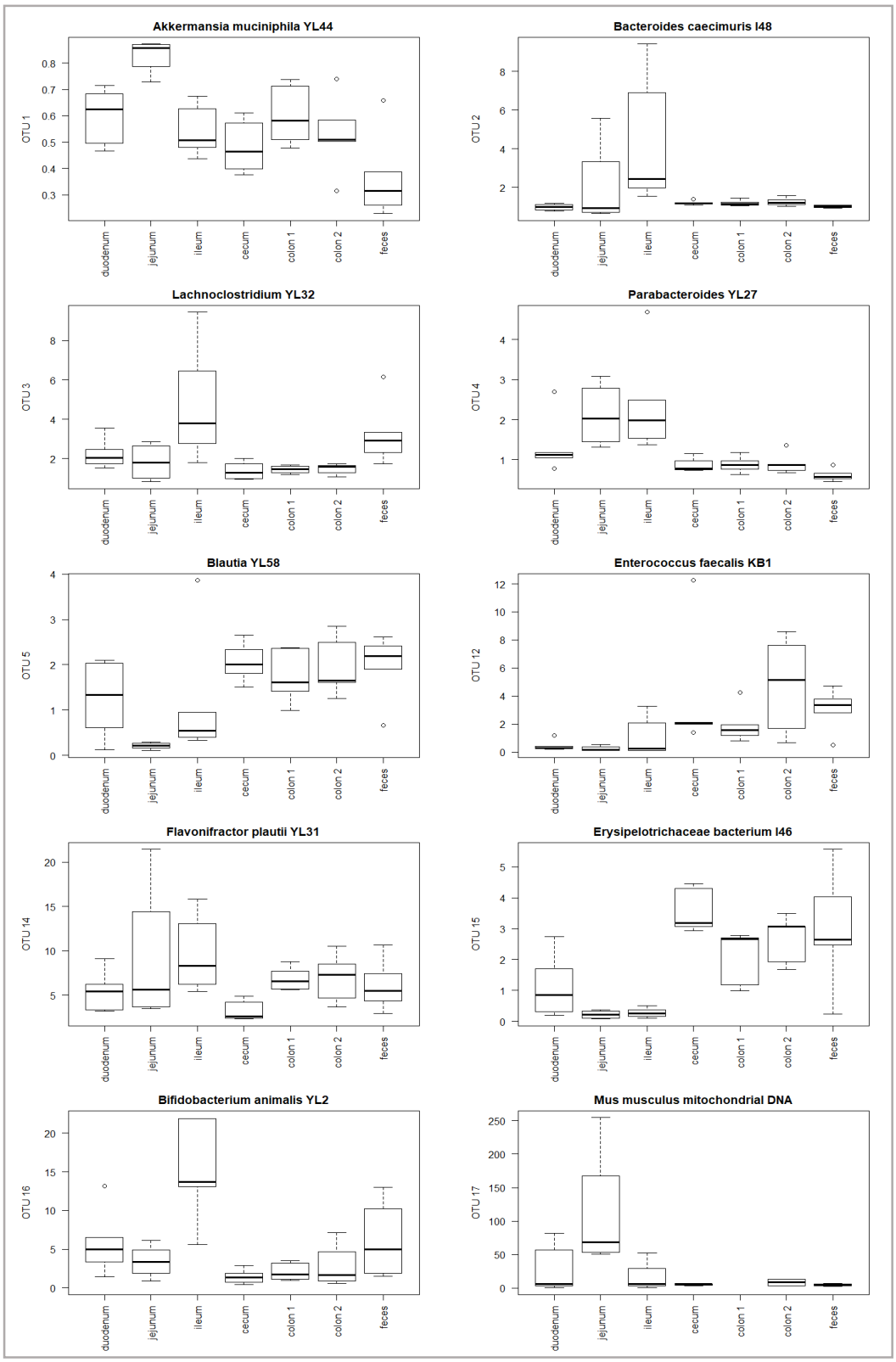

Fig. 5. Ratio cDNA/DNA of the ten OTUs of interest along the 7 sections of the GI tract and feces.

Table 6. P-values resulting from the Kruskal-Wallis test for the ten OTUs of interest.

OTU 1	0.004786
OTU 2	0.003013
OTU 3	0.001174
OTU 4	0.0001389
OTU 5	0.01009
OTU 12	0.0009459
OTU 14	0.01038
OTU 15	0.000303
OTU 16	0.002596
OTU 17	0.06366

Discussion

The aim of this part of the study was to compare the amount of DNA and RNA of the 12 OligoMM strains in the different GI tract and feces samples. Thus, it could be determined which of the bacteria are active or inactive in which sections of the gut. Our findings contain the DNA to cDNA comparison in the 7 sections of the GI tract and the feces for 9 of the 12 OligoMM strains. The p-values resulting from the Kruskal-Wallis test for the 9 OTUs of interest are all below or slightly above 0.01. OTU 17 was determined to originate from mitochondrial DNA and RNA of mice thus represents host contamination [4]. The given results may suggest that most of the bacteria that are usually present in the gut microbiota in mice are unequally distributed in the different gut sections. Additionally, the bacteria strains are active/inactive in specific sections of the gut which might be a determining characteristic for each of the species. With the given study, the aim was only attained for 9 of the 12 OligoMM strains. In the future, the following questions would be of interest: How could the experimental setup or the clustering into the OTUs be improved to receive an OTU for each of the 12 OligoMM strains? Are there bacterial strains which are equally distributed and equally active in all sections of the gut? Could the study be addressed to a more complex and accurate model of the gut microbiota in mice?

Analysis of the quality of the Taxonomic Binning. The 20 fecal samples from the salmonella infected mice were analysed with metagenomics and assembly. The contigs were binned with metabat and then the contigs in each bin were aligned to a local reference database. The number of bins that were created for each sample were compared (Table 7). Sample 13 did not create any bin because it contained very few reads. For the samples of Day 4 from the three mice that were infected with virulent salmonella only 2 to 5 bins, consisting mainly of the Salmonella genome and Lachnoclostridium, were created. For all the other samples between 10 and 16 bins were created. Furthermore, the percentage of each bin assigned to each species was calculated and plotted in a heatmap (Fig. 8). The contigs that could not be assigned to any bin formed an unbinned file for each sample. The unbinned files were treated like the bins to receive a heatmap with the content of the unbinned contigs (Fig. 9). Enterococcus faecalis KB1 was often not assigned to any bin and especially in the samples 15, 17 and 18, which represent the three sickest mice, the unbinned file almost exclusively consists of its genome. Furthermore, the number of contigs per bin was plotted against the highest percentage of the contigs that was assigned to one single species in a bin (Fig. 10). There was no correlation between the lower number of contigs per bin and the higher quality of the taxonomic binning found. In addition, the percentage of all bins which were assigned to one single species by at least 95% was calculated to be 75.8%. The quality of the taxonomic binning thus seems to be sufficiently good. Lachnoclostridium and Blautia and furthermore Bacteroides, Lachnoclostridium and Erysipelotrichaceae were found to often co-occur in one bin which shows good correlation with the taxonomic degree of relatedness of the OligoMM strains (Suppl. 1).

Table 7. Experimental setup showing the sample number of the fecal samples at each day for the five mice. The number of bins for each of the 20 samples are shown in brackets. The progressing sickness of the three mice that were administered viral salmonella is highlighted in colour.

Mouse	**Day 1**	**(# Bins)**	**Day 2**		**Day 3**		**Day 4**	
Vir1	1	(12)	6	(13)	11	(11)	15	(5)
Vir2	2	(13)	7	(14)	12	(13)	17	(4)
Vir3	3	(13)	8	(11)	13	(-)	18	(2)
Avir1	4	(14)	9	(12)	14	(16)	19	(10)
Avir2	5	(13)	10	(15)	16	(10)	20	(10)

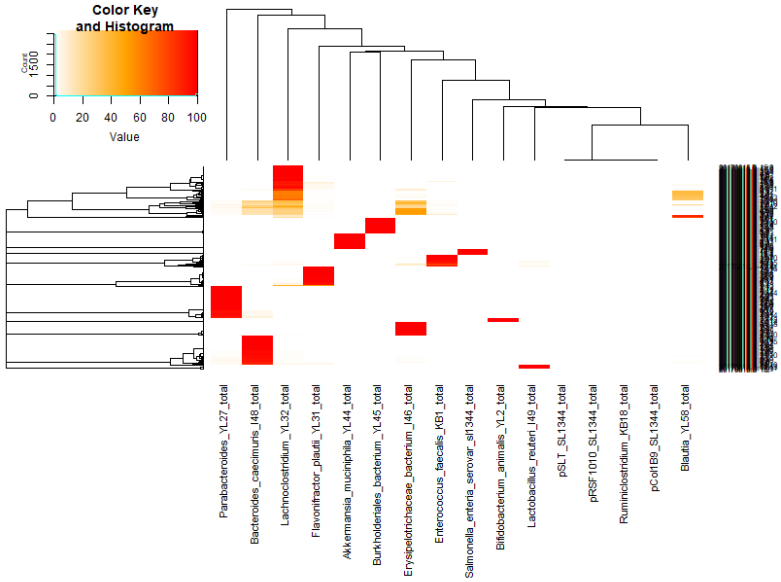

Fig. 8. Heatmap showing the percentage of each species represented in each bin. On the y axis, the 211 bins that were created from all samples were clustered by similarity means. On the x axis the 12 OligoMM strains and the salmonella including three plasmids were clustered. The colour indicates the percentage of the contigs which were aligned to the different species for each bin. The colour red means that 100% of the contigs in one bin were assigned to one species.

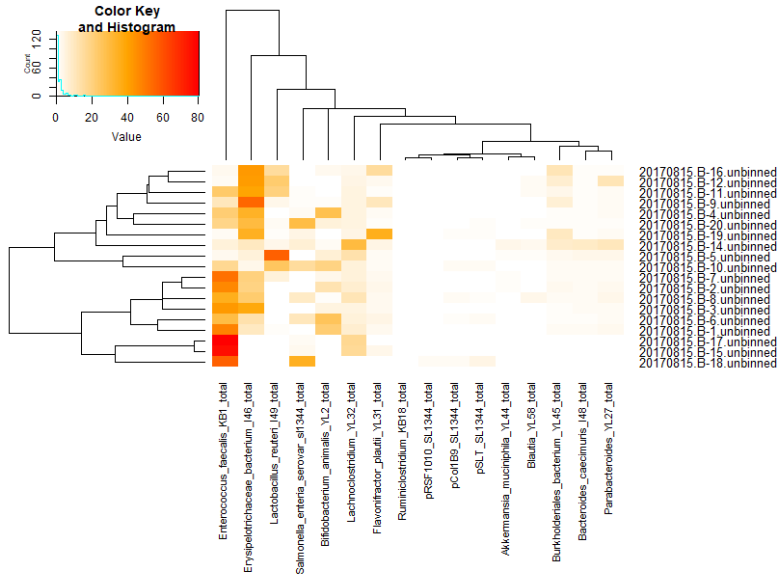

Fig. 9. Heatmap showing the percentage of each species represented in each of the 19 unbinned files.

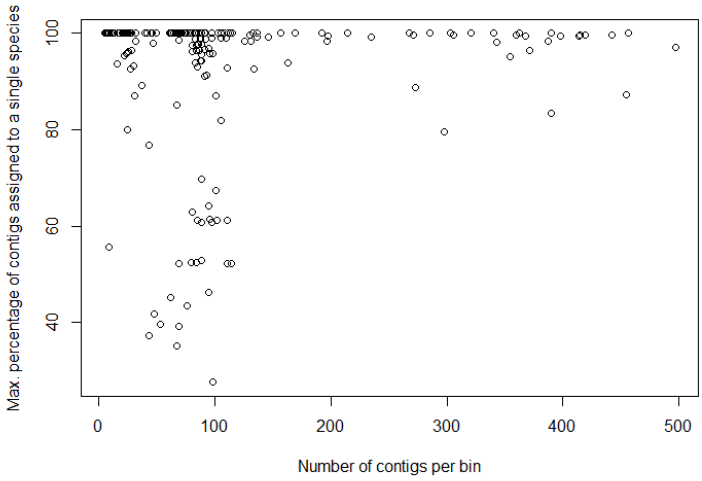

Fig. 10. Number of contigs per bin on the x axis plotted against the highest percentage of the contigs in one bin assigned to one species on the y axis. Each circle represents one bin.

Discussion

The aim of the second part of the study was to find a possibility to benchmark the quality of the taxonomic binning step. This aim was accomplished through the alignment of the contigs in each bin against the local reference database of the 12 OligoMM strains and the salmonella. In addition, certain patterns of co-occuring species were found which were taxonomically closely related. This can be explained by the fact that the binning method with metabat uses the tetranucleotide frequencies to assign the contigs to the bins. Because the closely related species should have a more similar genome sequences compared with other species they would more often be assigned to the same bin. A possible explanation for the high amount of Enterococcus contigs in the unbinned files could be that the Enterococcus genome contains genomic islands which affect the tetranucleotide frequency which metabat uses for the assembly. Consequently, the contigs containing the genomic island do not lead to a tetranucleotide frequency comparable with any of the reference genomes. There was no correlation found between the number of contigs per bin and the quality of the taxonomic binning. If we would have less contigs per bin and if these contigs were longer, we would expect the binning to work better. The graph (Fig. 10) shows no decrease in the highest percentage described by one species per bin with increasing count of contigs per bin. Though, it should be mentioned that the lengths of the contigs were not considered. To receive further insights regarding this question, an additional analysis could include plotting the N50 or L50

value of the most common species per bin against the percentage of the contigs assigned to the most common species per bin. N50 describes the length of the contig, which along with all longer contigs sums up to half of the genome length [5]. L50 describes the smallest number of contigs whose length sum equals half of the genome. Furthermore, it would be interesting to compare the percentage of all contigs that were binned with the ones that were unable to be binned per sample. The possibility to benchmark the quality of the taxonomic binning implies that the impact of different factors on the binning step can be determined and compared to further optimize the binning step. To improve the quality of the taxonomic binning, one could not only include the tetranucleotide frequency but also the abundance of each contig per library. The improvement of the data preprocessing and the assembly and the usage of other binning software could be means to further increase the binning quality.

References

[1] Kostic, Aleksandar D., Michael R. Howitt, and Wendy S. Garrett. "Exploring Host-microbiota Interactions in Animal Models and Humans." *Genes & Development* **27.7** (2013): 701–718. *PMC*. Web. 22 Oct. 2017.

[2] Roselli, Marianna et al. "Impact of Supplementation with a Food-Derived Microbial Community on Obesity-Associated Inflammation and Gut Microbiota Composition." *Genes & Nutrition* **12** (2017): 25. *PMC*. Web. 22 Oct. 2017.

[3] Hall, A. B. et al. "Human genetic variation and the gut microbiome in disease. " *Nature Publishing Group* **18** (2017): 690.

[4] Online BLAST against the nr/nt reference database:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

[5] N50, L50 and related statistics:

https://en.wikipedia.org/wiki/N50,_L50,_and_related_statistics

[6] Brugiroux, S. et al. "Genome-guided design of a defined mouse microbiota that confers colonization resistance against Salmonella enterica serovar Typhimurium." *Nature Microbiology* **2** (2016): 16215, Supplementary information: 18.

Supplementary Materials

Suppl. 1. Taxonomic classification of the 12 OligoMM strains [6].

Taxonomic Classification	Strain ID	DSM number	Classification	Eztaxon best hit
phylum Actinobacteria				
class Actinobacteria				
order Bifidobacteriales				
family Bifidobacteriaceae	**YL2**	26074	*Bifidobacterium longum subsp. animalis*	
phylum Bacteroidetes				
class Bacteroidia				
order Bacteroidales				
family Muribaculaceae	**YL27**	28989	'Muribaculum intestinale'	86.16% *B. intestinihominis* YIT11860(T) (ADLE01000001)
family Bacteroidaceae	**I48**	26085	'Bacteroides caecimuris'	96.86% *B. xylanisolvens* XB1A(T) (AM230650)
phylum Proteobacteria				
class Betaproteobacteria				
order Burkholderiales				
family Sutterellaceae_f	**YL45**	26109	'Turicimonas muris'	93.92% *Parasutterella excrementihominis* YIT 11859(T) (AFBP01000029)
phylum Verrucomicrobia				
class Verrucomicrobiae				
order Verrucomicrobiales				
family Verrucomicrobiaceae	**YL44**	26127	*Akkermansia muciniphila*	99.86% *A. muciniphila*_ATCC_BAA-835(T) (CP001071)
phylum Firmicutes				
class Bacilli				
order Lactobacillales				
family Enterococcaceae	**KB1**	32036	*Enterococcus faecalis*	
family Lactobacillaceae	**I49**	32035	*Lactobacillus reuteri*	
class Clostridia				
order Clostridiales				
family Lachnospiraceae	**YL32**	26114	*Clostridium clostridioforme*	
	YL58	26115	*Blautia coccoides*	
family Ruminococcaceae	**YL31**	26117	*Flavonifractor plautii*	
	KB18	26090	'Acutalibacter muris'	
class Erysipelotrichia				
order Erysipelotrichales				
family Allobaculum_f	**I46**	26113	*Clostridium innocuum*	92.09% *C. leptum* DSM 753T (ABCB02000019)