Microbial Analysis of Metagenomic Samples from Mouse Gut Microbiome

By Kiarash Rastegar

Introduction

The gut microbiome plays an important part in the process of digestion, absorption, and many other essential functions. It has been seen that the gut microbiome can play an important role in managing weight (Valdes, Walters, et..al). Considering the fact that we do not know all of the microbes in the gut environment we can not determine all of the beneficial effects that it has for us. Historically, researchers have relied on Next Generation Sequencing technologies and 16SrRNA to determine the composition of microbial communities. These technologies offer a fast and efficient way to not only determine the identities of microbes but to understand the mechanisms of how they work by analyzing different patterns and gene expressions shown in their genome. The fastest most cost-efficient way to understand the composition of an entire community is to use a method called shotgun metagenomics. This is a process where researchers extract the DNA from an unknown sample and end up fragmenting the DNA using restriction enzymes. These fragments are then attached with PCR primers called adapters which are used to bind to a Illumina's patent microarray and form clusters. These clusters then produce fluorescent signals based on the different types of nucleotides that are present inside the clusters. These fluorescent patterns are then used to determine the reads that were produced by the shotgun process.

The issue with shotgun metagenomic analysis is that it produces an enormous amount of data, which by itself does not make any sense. To figure out the composition of communities based on this data we need to be able to reconstruct the metagenome that was initially fragmented. There are several different methods available to do this. One of the more popular methods to do this is to use a de bruijn graph assembler which uses a concept in computer science called Graph Theory. The de bruijn graph treats each read as a node on the graph and the edges are the overlapping kmers between each read. Kmers are sequences of length K that behave like a scanning window but overlap with the previous window by n-1 elements. Meaning that the window moves one space and creates a new overlapping window and repeats this process throughout the entire sequence. These reads and kmers form the nodes and edges of the graph respectively and follow an eulerian path which can produce an approximation of the original DNA sequence. Usually, after the assembly process, it is customary to determine the quality of the assembly by comparing the assembled reads to a reference genome to determine the quality of assembly. Once the genomic fragments are reassembled they can be used for multiple downstream processes, including taxonomic identification. In this paper I am going to investigate if it is possible to identify the contents of a metagenome, without having a reference genome. To do this I will be using two unknown metagenomes that were sampled from the mouse gut environment and will perform different metagenomic analysis to see if it will be possible to determine the composition of these metagenomes based on the assembly of the shotgun DNA sequences.

Methods

The data that was obtained for analysis was provided by Dr. Scott Kelley and Kelley Labs at SDSU Department of Biological and Medical informatics. Data used for this experiment were metagenomes sequenced from a mouse's gut and were received as fastq files. Metagenomes were paired-end sequences, containing both a forward and reverse sequence. There were a total of two metagenomic samples that were analyzed for this analysis. Fastq files were trimmed using Fastp (version 0.12.4), which produced a summary report shown in (Table 1). Metagenomic reassembly was done using SPAdes (version 3.13.1) metaspades function. Icarus viewer and Krona chart (figure 3) are both outputs produced from metaspades. Quality of contigs produced by metaspades was quantified by using Quast (v5.0.2).

Results

The first step taken for any metagenome analysis is to clean up your reads and make sure there is no contamination coming from adapter sequences, as well as, keeping all the high quality reads. To do this I used the FastP software which produces a report after it completes the sequence quality control process. Table 1 shows a summary report of the state of the sequences before and after the filtering process. From the report we can see that 5% of reads were flagged as low quality, resulting in 95% percent of the original reads being kept. After the filtering process we do not see a drastic change in the GC content (47% to 46%). Before filtering we see that the Q20 and Q30 scores are 88% and 95% respectively. After the filtering process we do see a slight increase in these (90% and 96%). After trimming the sequences, assembly of shotgun sequences was done via metaSPAdes function of the software tool SPAdes to reassemble the metagenome producing longer fragmented reads called contigs. From those contigs SPAdes uses the paired-end sequences to fill in the gaps between contigs to make scaffolds which can be used for downstream analysis. To assess the quality of the reassembly done by metaSPAdes I used another tool called Quast, which produces a report with different quality metrics. Since I did not have a reference genome to use to assess the quality of the assembly, the metaQuast function in Quast downloads a set of reference genomes from NCBI's Genbank and uses these samples as references to assess the quality of the assembly.

Summary

General

fastp version:	0.12.4
sequencing:	paired end (101 cycles + 101 cycles)

Before filtering

total reads:	63.518890 M
total bases:	6.415408 G
Q20 bases:	6.101644 G (95.109210%)
Q30 bases:	5.708433 G (88.980053%)
GC content:	47.025497%

After filtering

total reads:	60.649286 M
total bases:	6.107289 G
Q20 bases:	5.880720 G (96.290193%)
Q30 bases:	5.528467 G (90.522452%)
GC content:	46.749231%

Filtering result

reads passed filters:	60.649286 M (95.482283%)
reads with low quality:	2.857768 M (4.499084%)
reads with too many N:	7.468000 K (0.011757%)
reads too short:	4.368000 K (0.006877%)

Table 1. Fastp report summary table showing filtering results, GC content, and initial reads that passed the filtering process.

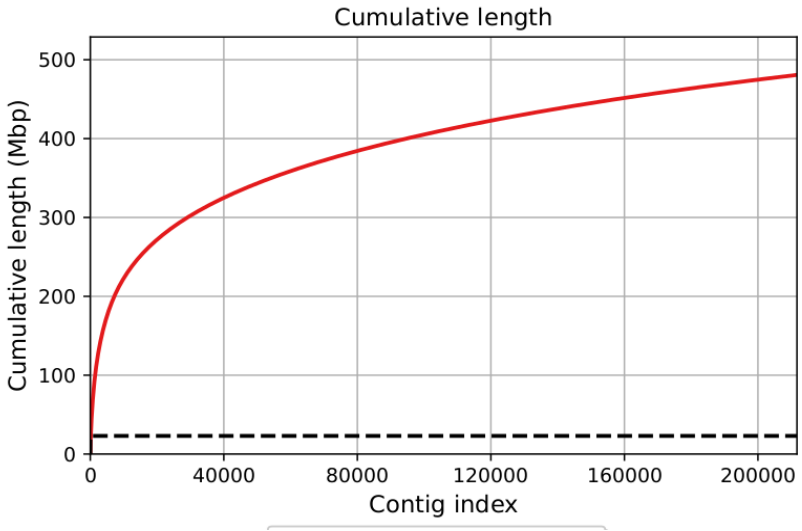


Figure 1. Cumulative length plot produced by metaQuast. The x-axis is sorted indexed contigs, going from smallest to largest in length. The Contig indexed at position 211,641 has a length of > 480Mbp.

Report

	contigs
# contigs (>= 0 bp)	881596
# contigs (>= 1000 bp)	93965
# contigs (>= 5000 bp)	14163
# contigs (>= 10000 bp)	6902
# contigs (>= 25000 bp)	2354
# contigs (>= 50000 bp)	895
Total length (>= 0 bp)	652443252
Total length (>= 1000 bp)	399355432
Total length (>= 5000 bp)	247168672
Total length (>= 10000 bp)	197167536
Total length (>= 25000 bp)	127454099
Total length (>= 50000 bp)	77091326
# contigs	211687
Largest contig	508016
Total length	480787953
Reference length	23014681
N50	5467
N90	750

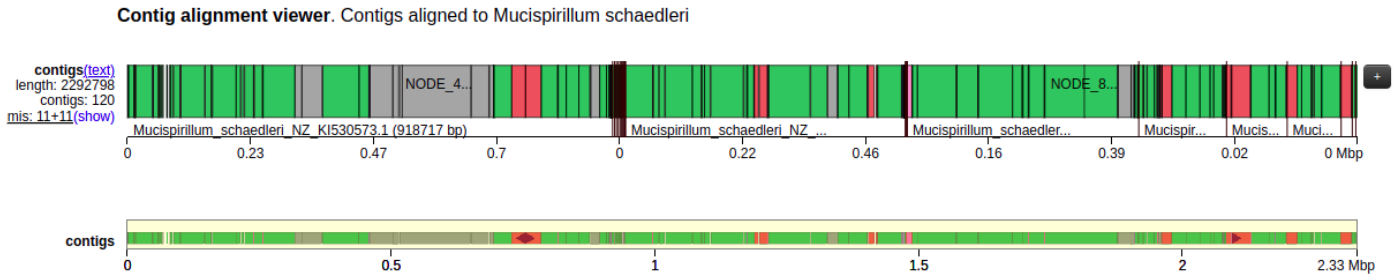
Table 2. Summary report from quast showing most important assembly quality metrics.

From Figure 1 we can see that the largest assembled contig is over 450 Mpb long, which is almost an entire genome practically by itself. The dotted line is considered a reference line, but since we did not have one specific reference, metaQuast just reports the reference genome length as 0. To summarize the major metrics that we have seen in the assembly we look at Table 2. The total number of contigs produced from our assembly is 211,687 and the number of contigs that are longer than 50,000bp is 895. Our N50 and N90 scores are 5467 and 750 contigs respectively. From Table 3 we can see that the genomes that had the most coverage from the contigs produced are Mucispirillum schaedleri and Muribaculum intestinale with 97.8% and 91.8%. While the genomes that had the least coverage from the contigs were Oscillibacter sp. 1 3 and Parabacteroides sp. CT06 with 48.1% and 65.1% respectively. To better visualize the data shown in Table 3, I looked at the contig browser for the highest covered reference genome (Mucispirillum schaedleri) and the least covered reference genome (Oscillibacter sp. 1 3) shown in Figure 2.

Genome	#fragments	Length, bp	Mean genome fraction, %	# misassembled blocks
Lactobacillus murinus	2	2 722 946	33.036	270
Mucispirillum schaedleri	22	2 332 248	97.772	38
Muribaculum intestinale	1	3 307 069	91.847	385
Oscillibacter sp. 13	12	4 467 686	48.119	805
Parabacteroides distasonis	1	4 812 066	67.189	203
Parabacteroides sp. CT06	1	5 372 666	65.122	225

Table 3. Results table produced by metaQuast. Table shows the amount of contigs and fraction of the genome covered by contigs that was produced by metaSPAdes.

a)



b)

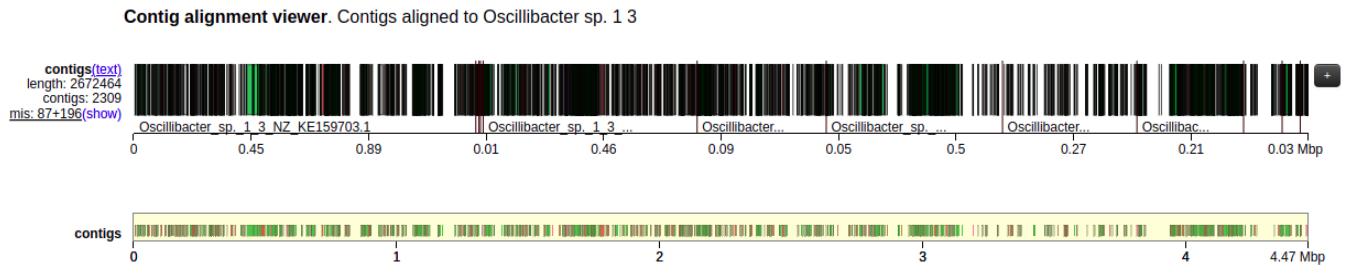


Figure 2. Contig alignment browser produced from metaQuast. **a)** Contig alignment browser for Mucispirillum schaedleri **b)** Contig alignment browser for Oscillibacter sp. 13

Comparing both contig browsers we see that Figure 2a has a lot more green regions and Figure 2b has a lot more black regions. Green regions are correctly aligned contigs to the reference genome, red regions are misassembled due to known reasons, gray regions are missambled due to unknown reasons, and black regions are unaligned gaps in the reference genome. To look at the entire composition of aligned contigs, metaQuast created a Krona chart which shows what percentage of the entire contig assembly belongs to specific species of bacteria (Figure 3). Here we see that 20% contigs belong to Muribaculum intestinale, 21% belong to Parabacteroides distasonis, 23% belong to Parabacteroides sp. CT06, 14% belong to Mucispirillum schaedleri, 5% belong to Lactobacillus murinus, and 16 % belong to Oscillibacter sp. 13.

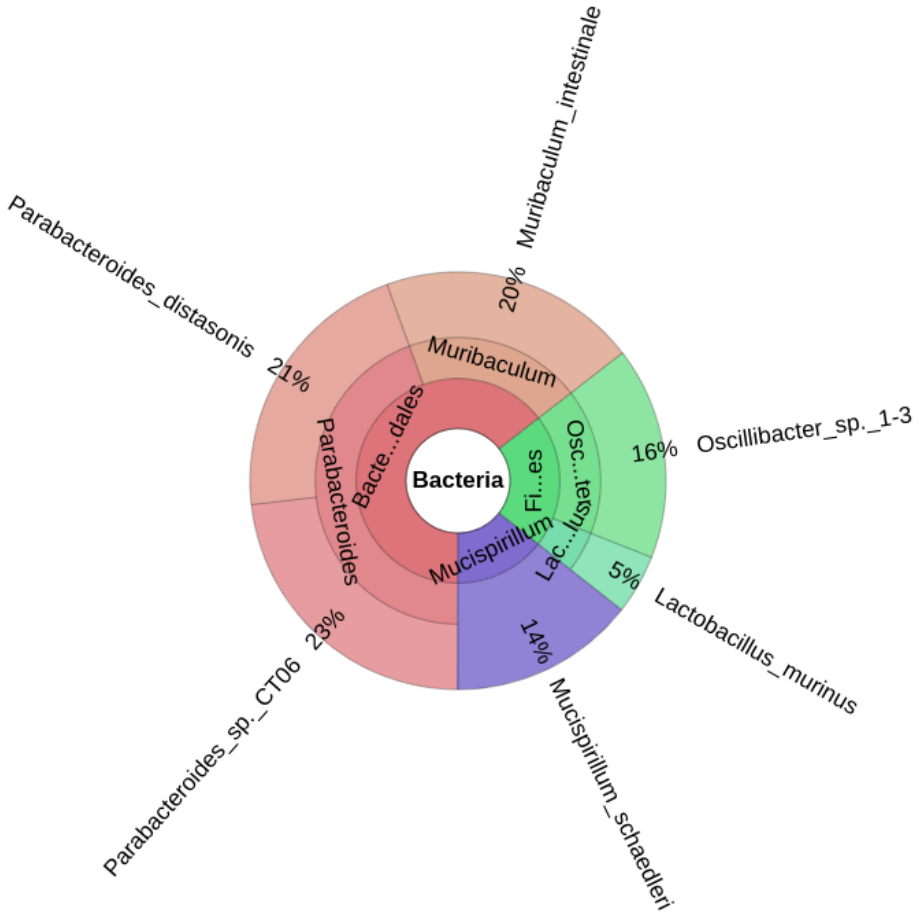


Figure 3. Krona chart showing composition of contigs to specific bacteria species.

Discussion

To achieve the overall objective of understanding what is in the metagenome sample and being able to reconstruct it properly it is important to follow a few key steps. The first step that was taken was trimming the fastq files so that our results would not be skewed by any contamination or low quality reads. To determine the quality of the reads it was important to pay attention to the Q30 score which represents the percentage of reads that had a quality score of 30. Looking at the fastp results it seems that the metagenomes that I was given were already trimmed due to the fact that the Q30 score did not seem to change (too much) after the filtering process. This is further supported by the quality score graph of the sequences before filtering (not shown in this report). All positions of the sequence seemed to have a score of over 35 out of 40, which is considered a high score. Although the Q30 scores were already high before filtering, fastp trimmed approximately 5% of the overall sequence, which resulted in a slight increase in the Q30.

After trimming the sequence, the next step was to assemble the shotgun sequences so that they can be used for further downstream analysis. This was done using the metaSPAdes function in the SPAdes software. MetaSPAdes is a de Bruijn graph assembly method which treats each shotgun sequence as a read and the overlapping kmers as the edges. It uses graph

theory and follows an Eulerian path to assemble the contigs that will be used for taxonomic classification. To determine the quality of the contigs, I used a function called metaQuast from the Quast software. Since there were no reference genomes to compare the assemblies to, metaQuast downloads several reference genomes from NCBI Genbank database which are most similar to the contigs. It does this by aligning the contigs to different reference genomes and chooses which best fit with the ensemble of contigs that are present. Looking at the cumulative length of the contigs in Figure 1, it is safe to assume that metaSPAdes was able to produce fairly large contigs (largest is over 480Mbp). This is further supported by looking at Table 2, where we see that the N90 score is 750. This means that there are 750 contigs that are longer than 90% of the cumulative length of all contigs. Essentially representing the quality of the assembly.

After producing a report with all of the quality metrics of the contigs assembly, metaQuast also produces a report for taxonomic classification. The taxonomic classification comes from aligning the contigs to the downloaded reference genomes and binning the contigs to specific reference genomes based on their alignment. From Table 3 we see that there are 6 major bacterial species that all the contigs are binned in. Looking at the fraction of genome coverage for each species, it is safe to assume that there are Mucispirillum schaedleri and Muribaculum intestinale in the metagenome samples that I was given. This is because both bacteria genomes have over 90% of their genome overlapping with contigs from the metaSPAdes assembly. Upon further research on these two bacteria species, it is known that both Mucispirillum schaedleri and Muribaculum intestinale have been found in the gut of rodents. After researching all of the bacterial genomes that were listed in the metaQuast taxonomic report (Table 3), It seems that the listed bacteria in the table are most likely what are in my metagenome. The composition of the metagenomic samples can be best summarized by Figure 3.

References

1. Prjibelski, Andrey, et al. “Using Spades De Novo Assembler.” *Saint Petersburg State University*, Wiley-Blackwell, 19 June 2020, https://pureportal.spbu.ru/en/publications/using-spades-de-novo-assembler.
2. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072-5. doi: 10.1093/bioinformatics/btt086. Epub 2013 Feb 19. PMID: 23422339; PMCID: PMC3624806.
3. Valdes, Ana M, et al. “Role of the Gut Microbiota in Nutrition and Health.” *The BMJ*, British Medical Journal Publishing Group, 13 June 2018, https://www.bmj.com/content/361/bmj.k2179.
4. Loy A, et al.“Lifestyle and Horizontal Gene Transfer-Mediated Evolution of Mucispirillum Schaedleri, a Core Member of the Murine Gut Microbiota.” *MSystems*, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/28168224/.
5. Lagkouvardos, Ilias, et al. “The Mouse Intestinal Bacterial Collection (MIBC) Provides Host-Specific Insight into Cultured Diversity and Functional Potential of the Gut

Microbiota.” Nature News, Nature Publishing Group, 8 Aug. 2016, https://www.nature.com/articles/nmicrobiol2016131.
6. Chen, Shifu, et al. “Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor.” *OUP Academic*, Oxford University Press, 8 Sept. 2018, https://doi.org/10.1093/bioinformatics/bty560.