MURAJ
Minnesota Undergraduate Research & Academic Journal

Volume 1 • Issue 1
z.umn.edu/MURAJ

Detecting Microbiome in Human Colorectal Cancers from High Throughput Sequencing Data

Shiyao Huang, Ce Yuan, and Subayya Subramanian

Abstract:

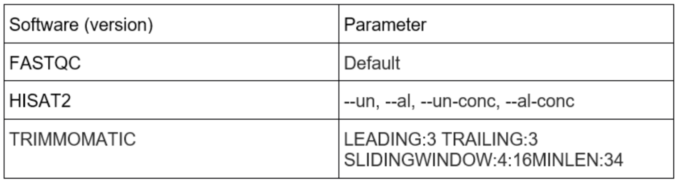
The microbiome is widely distributed on the human body and has a complex relationship with the host, through active metabolic interactions. The gut microbiome has shown to influence the risk and progression of colorectal cancer (CRC). Several tools are available for detecting microbiome from high throughput sequencing data. However, little is studied using the Cancer Genome Atlas (TCGA) CRC RNA-seq datasets. Here, we use the MOCAT2 software to detect bacterial sequences from the TCGA-CRC RNA-seq data. We analyzed a total of 55 cases, with 41 tumor samples and 14 control samples. The RNAseq data were first aligned to the human genome (hg19) and the reads that were not aligned to the human genome were used to detect bacterial sequences using MOCAT2. We found that *Escherichia coli* and *Propionibacterium* acnes were two bacterial species most commonly detected in the tumor samples (61% and 55% respectively), however, they were only detected in 14.29% and 7.14% of control samples, respectively. The *Cupriavidus necator* bacteria were most commonly detected in control samples (71.43%), while only present in 2.44% of tumor samples. *Cupriavidus necator* is the only bacteria species show significantly different presence between the tumor and control samples (chi-square test, p = 2.16E-07). Due to the limited sample size and reads detected by MOCAT2, future studies will use other software options such as PathSeq to detect the bacterial sequences.

Introduction

Colorectal cancer (CRC) is the 2nd cause of cancer-related death in the United States. In 2017, it is predicted that approximately 135,430 people will be diagnosed, and 50,260 will die from CRC. A recent study show more people between the age of 20 to 49 are being diagnosed with CRC.[1] It is a worrying trend and thus important to develop novel diagnostics and treatments to improve the survival of CRC patients. The high-throughput sequencing technologies are widely being used to study CRC as well as the bacterial community (microbiome) attached to it.[2] Several microbiome species, such as *Fusobacterium nucleatum* has been indicated as a potential biomarker for colon cancer (CRC). [3]

The average human intestine harbors >1014 microorganisms.17 In a healthy colon, the microbiome produces various metabolites. The gut microbiome in healthy humans produce about 70% of energy required by the colon epithelium to keep the function. Without the microbiome, the colon epithelium undergoes autophagy and fails to maintain its structure and normal function.[4] The gut microbiome has a complex relationship with the colon cancer (CRC). For example, the microbiome can influence the host through the metabolites produced, leading to diseases. The host, on the other hand, can influence the composition of the microbiome due to metabolic or behavior changes caused by genetic alterations.[2] Evidence indicates that altered gut microbiota composition is related with colorectal cancer (CRC)[5]. Several studies have shown inhibited colon tumor growth in mouse models without microbiome.[6]

We hypothesize that CRC microenvironment modulates the surrounding microbiome metabolism to compensate for the nutrient needs of the tumor from the metabolic pathway. A previous research analyzed microbiome taxonomy and metabolic pathway using MOCAT2 from 2 populations. The study shows such as *Fusobacteria nucleatum*(p=0.003) was detected in 76.9% CRC patients and only 48.1% in healthy people. In addition, the study found several metabolic pathway such as leucine degradation, citrate cycle, and methionine biosynthesis were significant correlation with colorectal cancer.[7] However, the bacterial transcripts information have not been extensively studied in the the Cancer Genome Atlas CRC datasets.8 Because TCGA also include data such as gene expression, we would like to combine the gene expression data



Software (version)	Parameter
FASTQC	Default
HISAT2	--un, --al, --un-conc, --al-conc
TRIMMOMATIC	LEADING:3 TRAILING:3 SLIDINGWINDOW:4:16MINLEN:34

Table 1. Parameters used for RNA-seq reads processing. The table shows the parameters for FASTQC, HISAT2, and TRIMMOMATIC in Gopher-pipeline in RNA reads processing.

with bacterial sequences identified by MOCAT2 from the RNA-seq data to study host-microbiome metabolic interactions. Here, the focus of this research project is to use the MOCAT2 program to identify bacterial sequences from TCGA-CRC RNA-seq data.

Materials and Methods:

TCGA data

The TCGA-CRC data contains a total of 650 cases with RNA-seq and clinical data. Here we downloaded the fastq files which contain RNA-seq reads and quality information of 108 samples. The fastq data were based on the biospecimen samples that collected from colon or rectum adenocarcinoma patients. The patients were newly diagnosed and did not receive any treatment for the diseases previously. Every sample maintained average 60% nucleotides in tumor cells less than 20% of cell necrosis for the protocol requirements. RNA and DNA from the tissue were extracted by using modified a DNA/RNA AllPrep kit (Qiagen), and the microRNA was purified by a mirVana miRNA Isolation Kit (Ambion). Therefore, only the samples had at least 6.9 ug of tumor DNA, 5.3 ug of total purified RNA, and 4.9 ug of germline DNA was involved in the recorded data. Specifically, the RNAseq data was used for mRNA expression profiling in this experiment was sequenced by microarray (Agilent) and RNA-Seq (Illumina). On the other hands, the metadata followed personal information of individual patients, such as age, date of diagnosis, and date of death (if available) etc. Finally, the biospecimen and clinical metadata along with the identifiers was stored in a customer relational database in the TCGA Data Coordinating Center (DCC).[8]

Gopher rnaseq-pipeline:

To eliminate the human genome reads from each of the whole RNAseq fastq data, we used the Gopher RNAseq-pipeline from Minnesota Supercomputing institution to process each fastq file. The FastQC screened each file for reporting quality control. The Trimmomatic trimmed low-quality bases and removed the sequence adapters for each reads base on the quality control plot. After a second quality control checking by The FastQC, the HISAT2 (Tophat2) performed alignment for each sample into a reference (human) genome database, We then used the featureCounts program and Cuffquant software program to generat transcript abundance files or each fastq files. This gives us the amount of gene expression in the tumor tissues. We then saved the reads that were not aligned to the human genome for MOCAT2 analysis. The parameters for the pipeline shows on table 1.[9]

MOCAT 2:

Mocat 2 was developed for assembling metagenomics and gene prediction from metagenomic data. We processed the "paired_not_concordat" files using Mocat 2 to find the bacterial gene in each case. The reads in each file were trimmed and filtered by the same method as Gopher RNAseq-pipeline used. To ensure eliminating more integrated human read and reduce the time for mapping bacterial genome, the reads, therefore, mapped into hg19 (Genome Reference Consortium Human Reference) database and used SOAPAligner2 to remove the human origin. The rest of reads was aligned to multiple bacterial genome databases and base (read) coverage and taxonomic composition was calculated. The SOAPdenovo found the sequence that mapped to the reference sequence by reckoned the insert size of each sequencing library. And the gap-tolerant BWA aligner revised base error and append indels to the mapped sequence that would be matched a complete gene from the reference database. Either the Prodigal or MetageneMark finally predicted the bacterial protein-coding gene based on the revised sequences.[10]

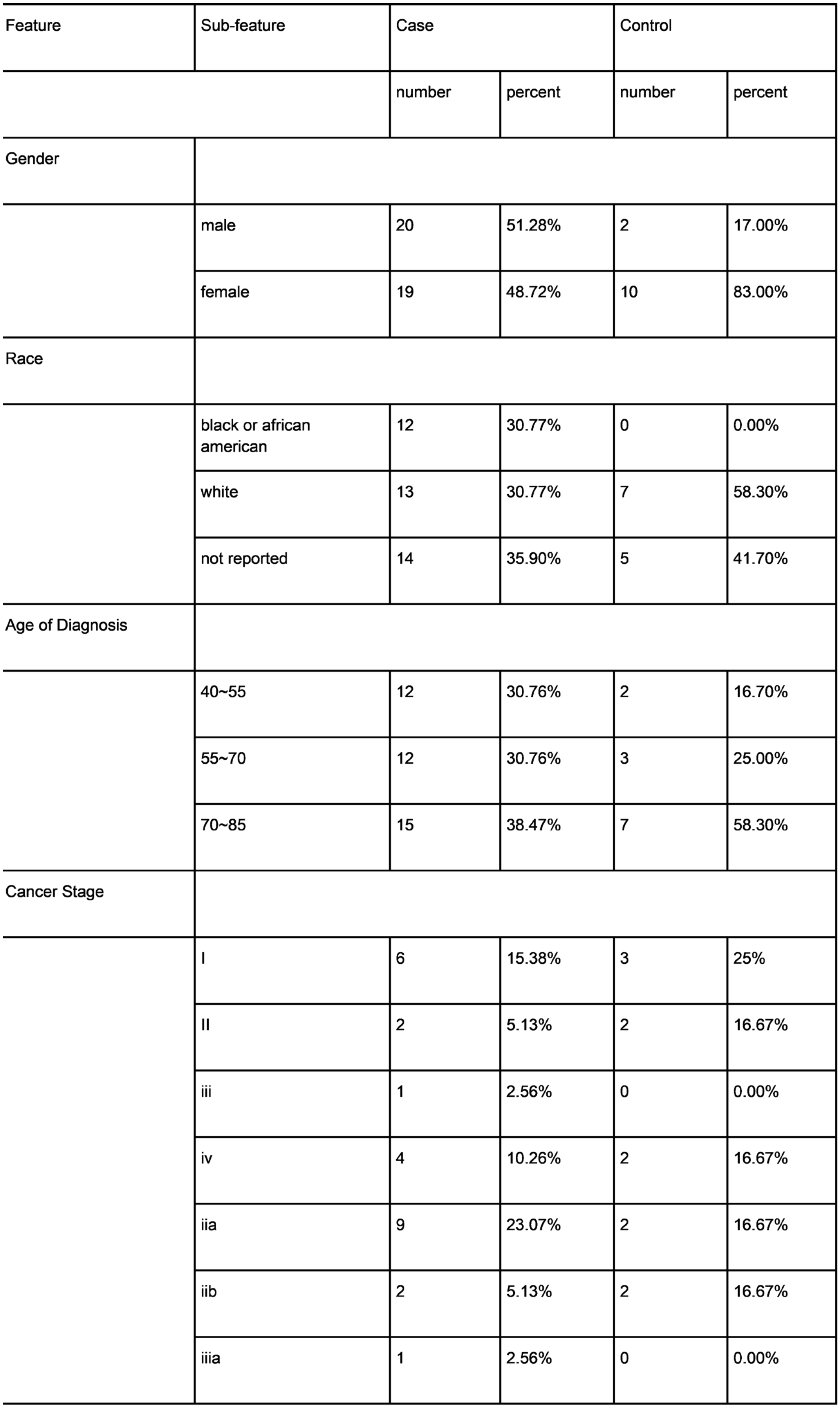
Statistical analysis

We collected the number of reads in each sample to determine whether the reads are enough to predict metabolic pathways. the richness of bacterial species in both cases and control samples was calculated in the percentage of detected samples using the results from Mocat 2. Therefore, the difference of bacterial species richness between cases and controls was statistically calculated using Chi-square and adjusted Chi-square in R.

Result and Data:

Bacterial Genome Read:

We processed 108 sample data, and 55 samples received bacterial reads which include 41 cases and 14 controls. The reads aligned to the microbiome genome database by Mocat 2 in each case as shown in Figure 1. Only five cases collected more than a hundred reads, and most of the cases received less than 10 reads. The inefficiency of reads could not guarantee the coverage to map the genome, which highly reduces the accuracy of the metabolic



Feature	Sub-feature	Case		Control	
		number	percent	number	percent
Gender					
	male	20	51.28%	2	17.00%
	female	19	48.72%	10	83.00%
Race					
	black or african american	12	30.77%	0	0.00%
	white	13	30.77%	7	58.30%
	not reported	14	35.90%	5	41.70%
Age of Diagnosis					
	40~55	12	30.76%	2	16.70%
	55~70	12	30.76%	3	25.00%
	70~85	15	38.47%	7	58.30%
Cancer Stage					
	I	6	15.38%	3	25%
	II	2	5.13%	2	16.67%
	iii	1	2.56%	0	0.00%
	iv	4	10.26%	2	16.67%
	iia	9	23.07%	2	16.67%
	iib	2	5.13%	2	16.67%
	iiia	1	2.56%	0	0.00%

Table 2. Metadata for TCGA colorectal cancer patients. The table presents the metadata such as gender, race, the age of diagnosis and cancer stage of the tumor and control samples.

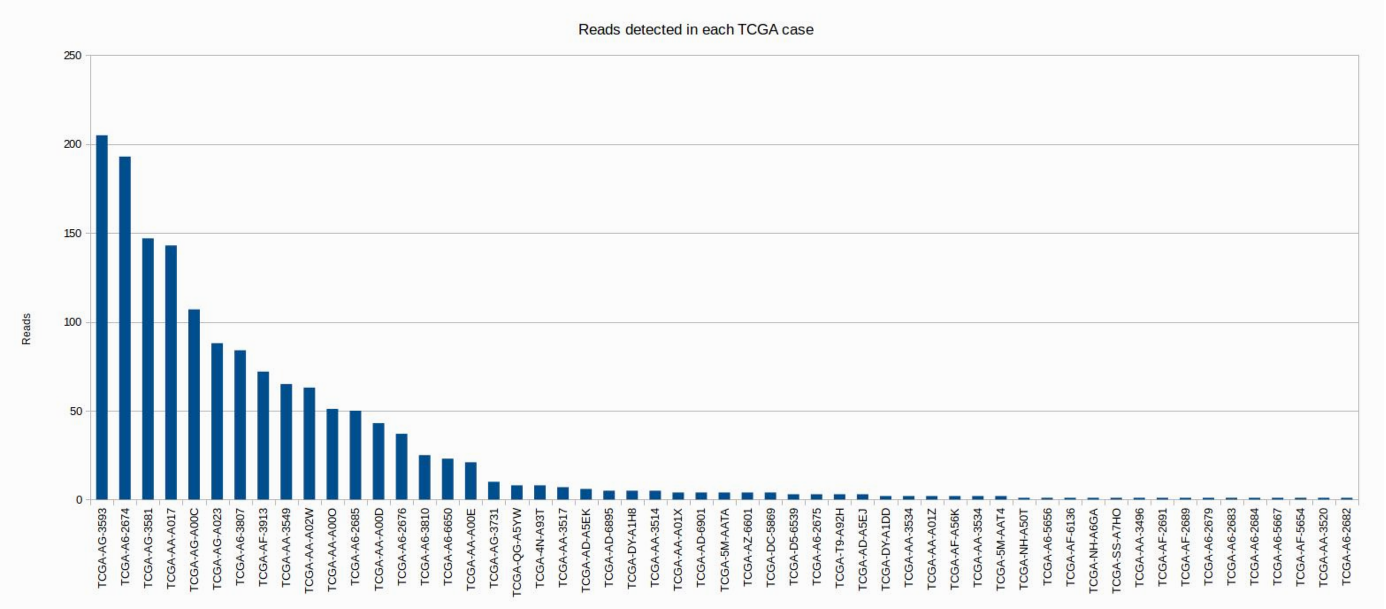


Figure 1. The bacterial read distribution of each TCGA case. The figure shows the number of bacterial reads detected using MOCAT2 in each sample. Five samples contains more than 100 reads, and 10 samples have more than 50 reads. Most of the samples have less than 10 reads.

pathways prediction. Therefore, we decided to focus on how the bacterial taxonomy related to colon cancer.

Characteristics of Metadata:

The features of 39 CRC cases and 12 controls from TCGA metadata (2 cases and 2 control did not record on the metadata) are shown in Table 2. There are approximately equal number of males and females involved in these cases, and most of the controls are from the females. Each of one-third of patients is either white or black, and another 35% of patients' race was not reported. However, none of the controls are a race of black. Therefore, more than one third of patients were diagnosed between the ages of 70 to 85, and also 58% of controls are from this range. In addition, 15.38% and 23.07% of cases were diagnosed in cancer stage i and iia, which has the highest portion in the cases. And 3 of 12 control sample was cut from the patients with cancer stage i.

Cases Taxonomy:

The bacterial taxonomy from patient samples was conducted by Mocat 2, shown in Figure 2, which indicates the bacterial species was detected in more than 10% of cases. *Escherichia coli* (61%), *Propionibacterium acnes* (55%), and *Propionibacterium sp.* 434-HC2 (43%) show the most significant association to CRC tissues. Therefore, *Escherichia sp.* 4_1_40B (31.91%), *Propionibacterium sp.* 409-HC1 (31.91%), *Propionibacterium sp.* CC003-HC2 (29.79%), *Shigella dysenteriae* (27.66%), *Shigella boydii* (27.66%), *Escherichia sp.* 1_1_43 (27.66%), and *Shigella sonnei* (21.28%) was detected in more than 20% of cases.

Control Taxonomy:

There were 20 bacterial species was constructed from control samples (Figure 3). The *Cupriavidus necator* was abundant in the healthy tissues, which found in 71.43% healthy samples; and Bacteroides families include *Bacteroides dorei* (21.43%), *Bacteroides sp.* 3_1_33FAA (21.43%), and *Bacteroides sp.* 9_1_42FAA (21.43%), *Bacteroides vulgatus* (14.29%), *Bacteroides sp.* 3_1_40A(14.29%) and *Bacteroides sp.* 4_3_47FAA (14.29%) was highly detected from the control samples, but rarely found from the case samples. On the other hands, *Propionibacterium acnes* (14.29%), *Propionibacterium sp.* 434-HC2 (14.29%), *Escherichia coli* (7.14%), *Propionibacterium sp.* 409-HC1 (7.14%), *Propionibacterium sp.* CC003-HC2 (7.14%), *Escherichia sp.* 4_1_40B (7.14%), *Shigella dysenteriae* (7.14%), *Shigella flexneri* (7.14%), *Shigella boydii* (7.14%), *Escherichia sp.* 1_1_43 (7.14%), and *Escherichia fergusonii* (7.14%), was found in both cases and control samples, but more likely exits from the CRC tissue. In addition, species such as [*Ruminococcus*] *torques, Roseburia inulinivorans, Leptothrix cholodnii, Pseudomonas mendocina, Xanthomonas campestris, Faecalibacterium prausnitzii, Rhodoferax ferrireducens, Pseudomonas stutzeri, Aeromonas caviae, Parabacteroides merdae, Mesorhizobium opportunistic* and *Cronobacter sakazaki*i were found in 7.14% of the control samples, but no detection in the case samples.

Statistical Analysis

In order to determine the significant difference in the bacterial species abundance between CRC tissues and healthy tissues. We tested chi-square in each species based on the percentage exists in the cases and control samples (Table 3). Five species *Cupriavidus necator* (p=2.16E-07), *Bacteroides dorei* (p=0.01794), *Bacteroides sp.* 3_1_33FAA (p=0.01794), *Bacteroides sp.* 9_1_42FAA (p=0.01794), and *Propionibacterium acnes* (p=0.02428) was statistically significant different between cases and controls. Since most of the samples are unmatched individuals in the two groups, we applied adjusted Chi-square to test the data. Therefore, only Cupriavidus necator (p=4.06E-05) received p-value lower than 0.05, which indicates it is statistically significantly correlated with CRC.

Discussion

The primary goal of this study was to determine if MOCAT2 can be used to identify microbial transcripts from TCGA-CRC RNA-seq data. Here, we found that MOCAT2 is not an effective tool to identify the bacterial transcripts from the TCGA RNA-seq data. There were not enough to cover the bacterial genome which would affect the accuracy of the prediction in the metabolic pathways.

Despite lack of efficiency to identify metabolic pathways, we were able to detect bacterial species in the datasets. We found *Cupriavidus necator* in 71.43% of control samples and it is statistically significantly different compared to the CRC (chi-square test; FDR-adjusted p-value=2.16E-07). Research shows *Cupriavidus necator* have functions in generating ATP from organic compounds through cell respiration, which is important components of glycolysis, and it can generate Ribose 5-phosphate though pentose phosphate pathway under autotrophic conditions, which is a precursor of nucleotides synthesis.[11,12] However, the relationship between *Cupriavidus necator* and colon cancer have not been determined yet. *Escherichia coli* was present in 61% tumor samples, research shows the pks gene expressed from *Escherichia coli* could damage host DNA and develops colon cancer, which explains why we found *Escherichia coli* in the colon cancer tissue.[13] A clinical research sequenced DNA from CRC patient blood, but they did not found *Propionibacterium acnes* DNA reads from the blood sample, which its

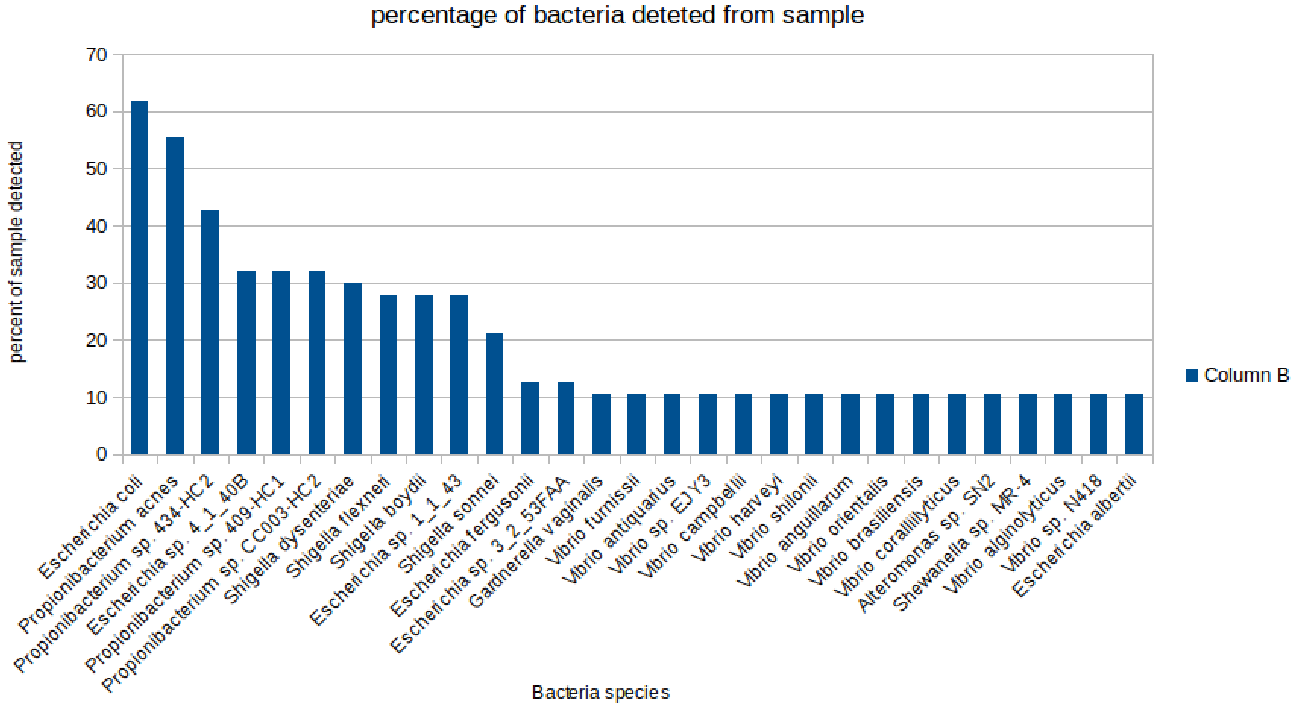


Figure 2. Bacterial species detected in more than 10% tumor sample. The figure shows the percentage of bacterial species detected in tumor sample. Twenty-nine species were detected in more than 10% of tumor samples. Species such as *Escherichia coli* (61%), *Propionibacterium acnes* (55%), and *Propionibacterium sp.* 434-HC2 (43%) was detected from more than 40% of samples.

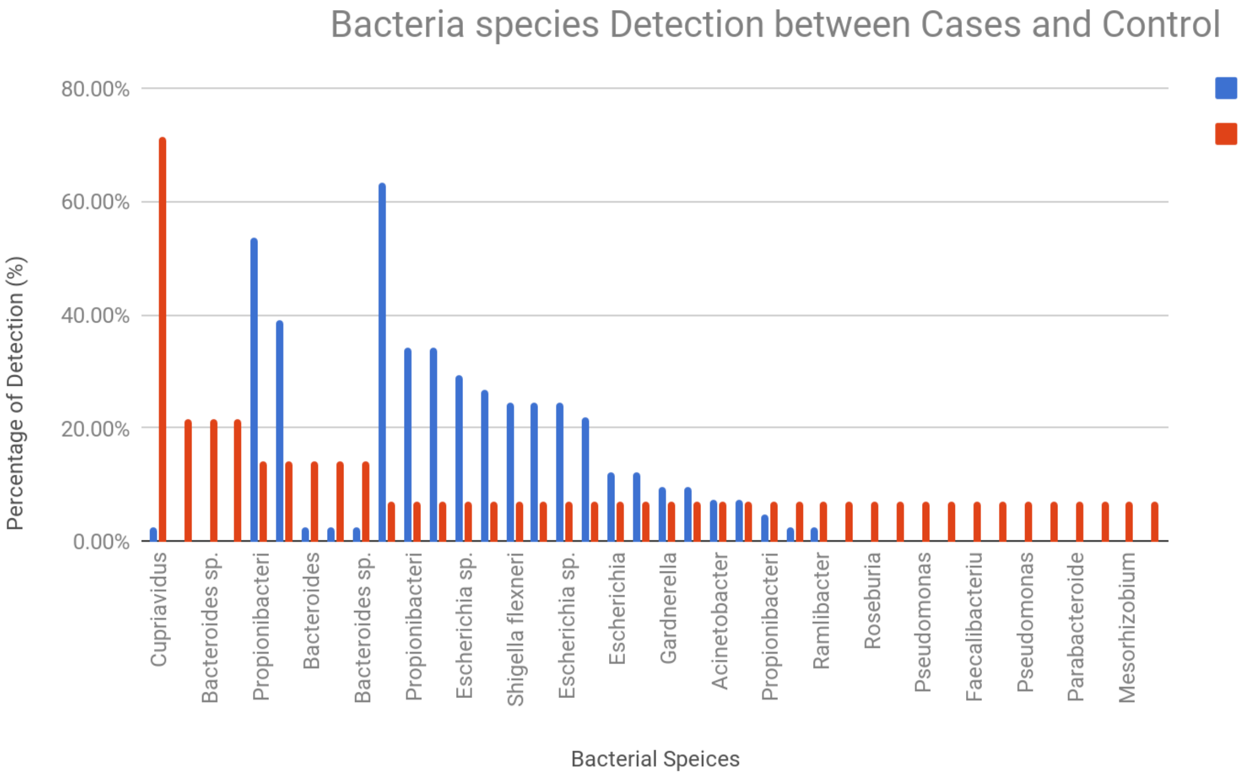
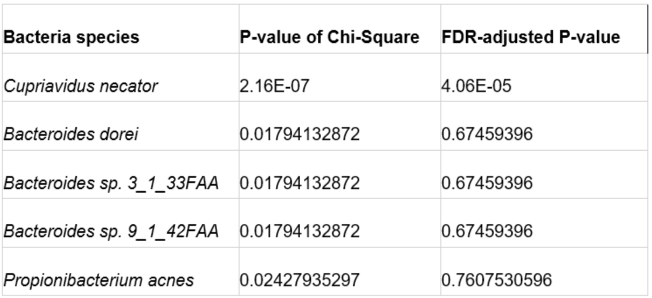


Figure 3. Bacterial species detected in cases and controls samples. The figure shows the bacterial species detected in the control samples (red bar) and the tumor samples (blue bar). *Cupriavidus necator* was detected in 71.43% healthy samples. Species such as *Bacteroides dorei*, *Bacteroides sp.* 3_1_33FAA, and *Bacteroides sp.* 9_1_42FAA were detected in over 21.43% healthy sample, but none of them were detected from the tumor sample. *Propionibacterium acnes* (53.66% in tumor, 14.29% in healthy tissue), *Propionibacterium sp.* 434-HC2 (39.02% in tumor, 14.29% in healthy tissue), *Escherichia coli* (63.41% in tumor, 7.14% in healthy tissue), *Propionibacterium sp.* 409-HC1 (34.15% in tumor, 7.14% in healthy tissue), *Propionibacterium sp.* CC003-HC2 (34.15% in tumor, 7.14% in healthy tissue), *Escherichia sp.* 4_1_40B (29.27% in tumor, 7.14% in healthy tissue), *Shigella dysenteriae* (26.83% in tumor, 7.14% in healthy tissue), *Shigella flexneri* (26.83% in tumor, 7.14% in healthy tissue), *Shigella boydii* (24.39% in tumor, 7.14% in healthy tissue), *Escherichia sp.* 1_1_43 (24.39% in tumor, 7.14% in healthy tissue), and *Shigella sonnei* (21.95% in tumor, 7.14% in healthy tissue) are more abundant in tumor samples.

does not match to our result as found in 55% of case sample.[14] However, Gut *Bacteroides dorei* (found in 21.43% tumor samples) has been shown to be associated with autoimmunity and type 1 diabetes, but has not proved to associate with colon cancer.[15] In addition, the species of *Bacteroides* sp. 3_1_33FAA (found in 21.43% control sample), and *Bacteroides sp.* 9_1_42FAA (found in 21.43% control sample) has not been well characterized yet. Moreover, the evidence of whether these species, especially *Cupriavidus necator,* influence colon cancer, require proving through biological experiments.

Using different tools to analysis same dataset would cause variate result between studies.[7] The MOCAT2 generated enough reads to predicted metabolic pathways in one study, but the algorithm was not able to identify bacterial sequences from TCGA-CRC data in our case.



Bacteria species	P-value of Chi-Square	FDR-adjusted P-value
Cupriavidus necator	2.16E-07	4.06E-05
Bacteroides dorei	0.01794132872	0.67459396
Bacteroides sp. 3_1_33FAA	0.01794132872	0.67459396
Bacteroides sp. 9_1_42FAA	0.01794132872	0.67459396
Propionibacterium acnes	0.02427935297	0.7607530596

Table 3: Top bacteria species significantly differ between tumor and control sample. The *Cupriavidus necator* has the p-value less than 0.05 for both p-value and FDR -adjusted p-value (Chi-Square). *Bacteroides dorei, Bacteroides sp.* 3_1_33FAA, Bacteroides sp. 9_1_42FAA and *Propionibacterium acnes* has p-value lower than 0.05, however not significant after FDR adjustment.

Future studies will focus on applying or developing new tools to identify bacterial sequences from the TCGA-CRC datasets.

In conclusion, we used the MOCAT2 software to identify bacterial species in colorectal cancers. We found that the algorithm was not able to identify enough bacterial sequence from the data to analyze metabolic pathways. Future studies will focus on applying or developed new tools to identify bacterial sequence from the RNA-seq datasets. This will help study bacterial metabolic functions in colorectal cancer patients. In addition, we found that the *Cupriavidus necator* is highly enriched in healthy colon compared to colon cancer samples. This warrant functional validation to ascertain the function of this bacteria in colorectal cancer.

References

1. Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G., Barzi, A., & Jemal, A. (2017). Colorectal cancer statistics, 2017. *CA: a cancer journal for clinicians,* 67(3), 177-193.

2. Blekhman, R., Goodrich, J. K., Huang, K., Sun, Q., Bukowski, R., Bell, J. T., ... & Clark, A. G. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome biology*, 16(1), 191.

3. Lopez, J. P., Diallo, A., Cruceanu, C., Fiori, L. M., Laboissiere, S., Guillet, I., ... & Ernst, C. (2015). Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing. *BMC medical genomics,* 8, 35-35.

4. Donohoe, D. R., Garge, N., Zhang, X., Sun, W., O'Connell, T. M., Bunger, M. K., & Bultman, S. J. (2011). The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell metabolism*, 13(5), 517-526.

5. Zackular, J. P., Baxter, N. T., Iverson, K. D., Sadler, W. D., Petrosino, J. F., Chen, G. Y., & Schloss, P. D. (2013). The gut microbiome modulates colon tumorigenesis. *MBio*, 4(6), e00692-13.

6. Dove, W. F., Clipson, L., Gould, K. A., Luongo, C., Marshall, D. J., Moser, A. R., ... & Jacoby, R. F. (1997). Intestinal neoplasia in the ApcMin mouse: independence from the microbial and natural killer (beige locus) status. *Cancer research,* 57(5), 812-814.

7. Vogtmann, E., Hua, X., Zeller, G., Sunagawa, S., Voigt, A. Y., Hercog, R., ... & Sinha, R. (2016). Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PloS one*, 11(5), e0155362.

8. Cancer Genome Atlas Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), 330-337.

9. Wiki. (n.d.). Retrieved December 09, 2017, from https://bitbucket.org/jgarbe/gopher-pipelines/wiki/rna-seq-pipeline.rst

10. Kultima, J. R., Sunagawa, S., Li, J., Chen, W., Chen, H., Mende, D. R., ... & Wang, J. (2012). MOCAT: a metagenomics assembly and gene prediction toolkit. *PloS one,* 7(10), e47656.

11. Pohlmann, A., Fricke, W. F., Reinecke, F., Kusian, B., Liesegang, H., Cramm, R., ... & Strittmatter, A. (2006). Genome sequence of the bioplastic-producing "Knallgas" bacterium Ralstonia eutropha H16. *Nature biotechnology*, 24(10), 1257-1262.

12. Cramm, R. (2009). Genomic view of energy metabolism in *Ralstonia eutropha* H16. *Journal of molecular microbiology and biotechnology*, 16(1-2), 38-52.

13. Arthur, J. C., Perez-Chanona, E., Mühlbauer, M., Tomkovich, S., Uronis, J. M., Fan, T. J., ... & Rhodes, J. M. (2012). Intestinal inflammation targets cancer-inducing activity of the microbiota. *science*, 338(6103), 120-123.

14. Mollerup, S., Friis-Nielsen, J., Vinner, L., Hansen, T. A., Richter, S. R., Fridholm, H., ... & Mourier, T. (2016). *Propionibacterium acnes*: Disease-causing agent or common contaminant? Detection in diverse patient samples by next-generation sequencing. *Journal of clinical microbiology,* 54(4), 980-987.

15. Davis-Richardson, A. G., Ardissone, A. N., Dias, R., Simell, V., Leonard, M. T., Kemppainen, K. M., ... & Ilonen, J. (2014). *Bacteroides dorei* dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Frontiers in microbiology,* 5.