|  |
| --- |
| Practical Introduction to Metagenomics  Community and Functional Profiling of Healthy Individuals and Colorectal Cancer Patients from Separate Environments  Zach Young1  1Johns Hopkins University AAP |

[[1]](#endnote-1)\*abstract

Colorectal cancer (CRC) is one of the leading causes of deaths worldwide and projected to increase substantially over the next decade with over 2.2 million new cases expected to arise by 20302. While much of the research into CRC has looked into the genetic mutations and genomic alterations that develop within a patient’s own cells, it is becoming increasingly more apparent that alterations in the gut microbiota composition can also influence the development, progression and response to therapy of this disease. By examining a patient’s own gut microbiome, it may be possible to gather information in a less invasive way than current methods and this information could potentially be used for therapeutic and prognostic purposes. This project represents the beginning of a meta-analysis study that aims to find distinct differences between healthy individuals and CRC patients as well as differences between CRC patients from different environments. Community profiling revealed an elevation of the Firmicutes phylum and Clostridia class in CRC patients versus healthy individuals and the presence of the Gammaproteobacteria class among Japanese CRC patients which was absent from the other two groups. Diversity at the phylum and class levels was much greater in the CRC patients versus the healthy individuals. Functional profiling revealed less distinctions overall, with many of the most abundant gene families identified being involved in essential cellular processes such as replication and energy production.

# introduction

Metagenomics is the study of genetic material, often arising from prokaryotic organisms like bacteria, from environmental samples. Since its beginnings, this field has given us a much greater understanding and appreciation for the vast diversity within microbial ecosystems, most of which was previously unknown due to our inability to culture many of these microbes directly. This is especially true in regards to the bacterial species that reside on and within our own bodies, together known as the human microbiome. In the past decade, thanks to advances in sequencing technologies and analysis tools, our understanding of the human microbiome and its impact on health has increased dramatically. In particular, the many bacterial species found throughout the digestive system, collectively known as the gut microbiome, has been found to play a role in a number of human disorders. The gut microbiome has functions in many important human processes including metabolism, host protection and immune-system development1. Imbalances in the normal microbial makeup of the gut microbiome have been associated with a variety of disorders including neurological conditions such as Parkinson’s and depression, Crohn’s disease and even some cancers, including CRC1,4.

A number of studies have examined the link between individual microbes as well as entire microbial communities and their ability to potentiate CRC incidence. Several of these have used 16S rRNA profiling to directly examine the community makeup of patients with CRC versus control groups. In 2012, Chen et al. found an enrichment of Fusobacterium, Peptostreptococcus and Lactobacillales in cancerous tissue from CRC patients along with a reduction in Faecalibacterium2. A study by Wu et al. found 16 genera that were significantly more abundant in CRC samples versus controls, including the potentially pathogenic Fusobacterium and Campylobacter species6. While the actual species present within CRC patients is informative, it is also important to consider what functional activities these microbes are carrying out. Zeller et al. showed a total of 24 KEGG modules to be differentially abundant in CRC patients which ultimately resulted in a global metabolic shift from bacteria predominantly utilizing host dietary fiber to utilizing more host-derived energy sources8. Furthermore, it has been shown that microbes and microbial products can enter the tumor microenvironment and contribute to carcinogenesis in a several ways, including modulating host metabolism, shaping the immune response and even influencing therapy4.

Given that CRC is already the third most commonly diagnosed cancer and expected to increase in prevalence by 60% by the year 2030, being able to accurately profile the gut microbiome in a patient may help with preventative measures, as well as be used for prognostic and therapeutic purposes3. To truly understand the interaction between the gut microbiome and CRC development and progression, as well as come up with meaningful tools to act on this knowledge, it will be imperative to start performing meta-analyses on a wide range of samples from varying environments. To this end, Wirbel et al. conducted an expansive study of CRC patients from different regions to see if the same patterns of community imbalance existed across different environmental conditions7. They identified a core set of 29 microbes enriched across CRC patients from varying environments and functional analysis revealed an increase in both protein and mucin catabolism genes along with a depletion in carbohydrate degradation genes7. In this project, a similar approach will be taken, looking for differences in community and functional profiles from three separate groups: healthy individuals, CRC patients from Japan and CRC patients from Germany. This project is the basis for what could become a larger meta-analysis study down the road.

# methods

**2.1 Data**

The data for this project came from three separate sources and consists of samples from healthy individuals, CRC patients from Japan and CRC patients from Germany. The healthy samples came from participants in the Human Microbiome Project. Data for CRC patients from Japan was gathered from the DNA Data Bank of Japan database under study identification no. DRP004793. The raw sequencing data for the samples for German CRC patients was gathered from the European Nucleotide Archive under study no. PRJEB27928. All data was initially collected from stool samples and sequenced using the Illumina HiSeq 2500 platform. The raw sequencing data was all contained in paired, FASTQ formatted files.

**2.2 Tools**

The tools used for this project were MetaPhlAn for the community profiling and HUMAnN2 for functional profiling. MetaPhlAn is computational tool used for determining the composition of a microbial community from shotgun sequencing data down to species level resolution9. To determine the relative abundances within a sample, this tool aligns metagenomic reads to a database of over one million unique, clade-specific marker genes taken from roughly 100,000 reference genomes9. The output from this process is two columns. The first column lists clades starting at the kingdom level and progressing down to the species level. The second column lists that clade's percent abundance in the sample. Each level will sum up to 100%; that is all kingdom-level, phylum-level, class-level, etc. clades will add up to 100%.

The functional profiling was done with the HUMAnN2 tool. HUMAnN2 is a method for determining the presence, absence, and abundance of metabolic pathways within a microbial community based off of metagenomic shotgun reads10. Essentially, this tool works by mapping the metagenomic data to various databases to identify potential genes, and then compiling those genes into pathways and gene families. It also uses MetaPhlAn to identify which species the genes are coming from. The final output from this process is typically three files: a gene family abundance file, a pathway abundance file and a pathway coverage file, all of which are stratified by organism. For this project the main file used was the gene family abundance file. After this was created for each sample, the files were merged together and normalized to account for the varying depths of coverage from the original files.

**2.3 Workflow**

The workflow for this project consisted of first downloading the raw sequencing data and setting up the necessary bioinformatics tools through their own Docker images. Chiron was used to implement these tools as this has specific pipelines for both the community and functional profiling that was performed in this project. All of this was done through the Google Cloud Platform. Once the community and functional profiling was complete, the necessary files were transferred to a local environment for further analysis. The files were loaded into data frames using the Pandas library and then further visualization and analysis tasks were performed. The code to perform these tasks can be found in a supplementary information file.

# results

The first feature of these samples that was analyzed was the community profiles generated by MetaPhlAn. At the phylum level, Bacteroidetes was the most common phylum for two of the three healthy patients in terms of relative abundance percentage. One patient displayed a slightly higher percentage of Firmicutes, but the abundances were close (53% to 44%). This individual also had a small percentage of the Actinobacteria phylum present that was not seen in the other two healthy individuals. Overall, at the phylum level the CRC patients showed elevated abundances of both Actinobacteria and Proteobacteria as compared to the healthy individuals. Japanese patients appeared to have larger numbers of Proteobacteria than the German patients and German patients had larger numbers of Actinobacteria than the Japanese patients. Interestingly, one German patient had no present of either these phyla and appeared to be more closely profiled to the healthy individuals.

Similar patterns were found at the class level as well. The most predominant class for all three healthy patients was Bacteroidia, followed by Clostridia. This was opposite for the CRC patients which all had higher levels of Clostridia than Bacteroidia. One healthy individual had a high abundance of Negativicutes at close to 40%, but besides this and the Clostridia and Bacteroidia classes, no other class showed significant abundance in the healthy individuals. CRC patients on the other hand had much more diversity at the class level overall. All CRC patients showed some abundance of both the Actinobacteria and Negativicutes classes. All Japanese CRC patients also displayed varying levels of Gammaproteobacteria which was not seen in either of the other groups. One German CRC patient had a relative abundance of 5.25% for the Deltaproteobacteria class, while no other sample exhibited notable levels of this class.

Examining the functional profiling data, many of the most abundant gene families were the same between all of the samples. These mostly included gene families involved in essential cellular process such as replication and energy production. Peptidylprolyl isomerase, NADH dehydrogenase, DNA helicase, NADH ubiquinone reductase, H(+)-transporting two-sector ATPase, DNA-directed DNA polymerase, Signal peptidase I, Tryptophan synthase, Aldose l-epimerase, 6-phosphofructokinase and Methionyl aminopeptidase were all among the most abundant gene families on average for all samples. Beta-galactosidase was an abundant gene family in both healthy individuals and the German CRC patients but wasn’t one of the most abundant gene families in the Japanese patients. These patients instead had high abundances of Phosphoglycerate mutase and 3-isopropylmalate dehydratase which the other two groups did not exhibit. For the German CRC patients, the gene families 3-oxoacyl reductase and Sulfate adenylyl transferase were both among the most abundant gene families which was not the case in the other two groups.

I also examined which gene families had the largest differences on average between the three groups. Some of the largest differences between the groups were also from families that were the most abundant within individual groups such as Peptidylprolyl isomerase, NADH dehydrogenase, DNA helicase, NADH ubiquinone reductase, H(+)-transporting two-sector ATPase, and Signal peptidase I. Families that had a large difference between healthy individuals and German CRC patients included 3-oxoacyl reductase, 7-alpha-hydroxysteroid dehydrogenase, 5-dehyro-4-deoxy-D-glucouronate isomerase and dihydrolipoyl dehydrogenase. Between healthy and Japanese CRC patients Beta-galactosidase, N-acetylglucosamine-6-phosphate deacetylase and 3-isopropylmalate dehydratase were among families with a large difference in abundance. Between the CRC patient groups, 3-oxoacyl reductase, 7-alpha-hydroxysteroid dehydrogenase, Beta-galactosidase and Lysozyme showed large differences.

4. Discussion

Overall, the analysis revealed differences between the three groups examined here at both the community level and at a functional level. This project did not examine enough samples to determine if any of these differences were statistically significant, but there were still interesting findings here that appear to coincide with previous studies looking at the microbial makeup between healthy individuals and CRC patients, especially at the community level. At both the phylum and class levels, there was more diversity present in the CRC patients as compared to the healthy individuals. At both these levels, the healthy samples were predominantly made up of just two groups of bacteria while the CRC patients had multiple types present. At the phylum level, Firmicutes were most abundant among CRC patients while Bacteroidetes were most abundant among the healthy patients. Similarly, the Clostridia class was most abundant in CRC patients while Bacteroidia was most abundant in the healthy individuals. Previous studies have found an enrichment of Peptostreptococcus and Lactobacillales in cancerous tissue, both of which belong to the Firmicutes phylum and may be responsible for some of the differences observed here. Fusobacterium has also been observed to be enriched in CRC patients and there was one Japanese patient that displayed this pattern. The Gammaproteobacteria class was the third most abundant class overall for the Japanese patients, but was not present in either of the other two groups indicating this class to possibly be a unique marker for this population. It would be interesting to look at healthy Japanese patients to see if this class is present in all individuals from this environment or if this class in only enriched in CRC patients from this environment for some reason. A number of different bacterial species have been shown to be differentially abundant in CRC patients versus healthy controls and this project, albeit with a small sample size, seemed to back this fact.

While there were differences observed at the functional level, it is difficult to tell how meaningful these differences really are just by examining the gene families present. Previous studies have shown protein and mucin catabolism to be enriched in CRC patients along with a depletion of carbohydrate degradation, but neither of these patterns were explicitly evident just by examining gene family abundances. The HUMAnN2 method can also calculate metabolic pathway abundances which would probably have been more telling of the true functional potential of the various samples examined here instead of looking only at individual gene families. This information would be beneficial to examine more thoroughly going forward. It is also worth noting that when dealing with only metagenomic data, we can only get an estimate on the functional potential of a community and not necessarily the actual functional activity. Including metatranscriptomic and metaproteomic data in a study such as this would be the best way to truly examine functional differences between a healthy and diseased state. Nonetheless, I believe that this project has provided a framework method for which I could expand upon by looking into more samples with the goal of truly uncovering meaningful differences between healthy individuals and CRC patients. Given that previous studies have already shown how bacteria can play a role in tumorigenesis and influence therapy, I believe that it may be possible that the differences that exist within CRC patients could one day be used as a target to help better diagnose and even treat this disease.

References

1. Cani P. D. (2018). Human gut microbiome: hopes, threats and promises. Gut, 67(9), 1716–1725. https://doi.org/10.1136/gutjnl-2018-316723

2. Chen, W., Liu, F., Ling, Z., Tong, X., & Xiang, C. (2012). Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. PloS one, 7(6), e39743. https://doi.org/10.1371/journal.pone.0039743

3. Gagnière, J., Raisch, J., Veziant, J., Barnich, N., Bonnet, R., Buc, E., Bringer, M. A., Pezet, D., & Bonnet, M. (2016). Gut microbiota imbalance and colorectal cancer. World journal of gastroenterology, 22(2), 501–518. https://doi.org/10.3748/wjg.v22.i2.501

4. Maisonneuve, C., Irrazabal, T., Martin, A., et al. (2018). The Impact of the Gut Microbiome on Colorectal Cancer. Annual Review of Cancer Biology. Vol. 2:229-249. doi:030617-050240.

5. Yachida, S., Mizutani, S., Shiroma, H. et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med 25, 968–976. https://doi.org/10.1038/s41591-019-0458-7

6. Wu, N., Yang, X., Zhang, R. et al. (2013). Dysbiosis Signature of Fecal Microbiota in Colorectal Cancer Patients. Microb Ecol 66, 462–470. https://doi.org/10.1007/s00248-013-0245-9

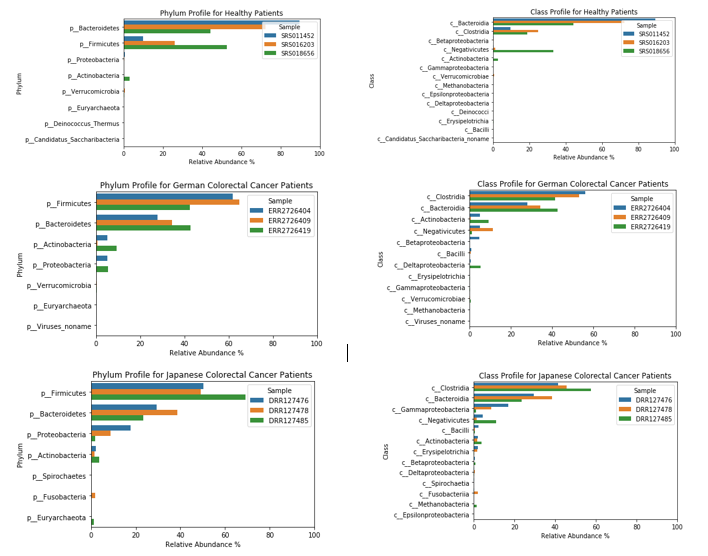
7. Wirbel, J., Pyl, P.T., Kartal, E. et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 25, 679–689. https://doi.org/10.1038/s41591-019-0406-6

8. Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular systems biology, 10(11), 766. https://doi.org/10.15252/msb.20145645

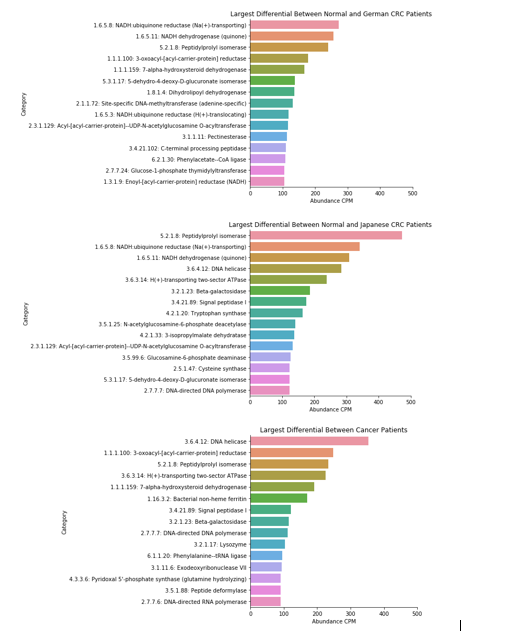
9. McIver LJ, Abu-Ali G, Franzosa EA, Schwager R, Morgan XC, Waldron L, Segata N, Huttenhower C. bioBakery: a meta'omic analysis environment. Bioinformatics. 2018 Apr 1;34(7):1235-1237. PMID: 29194469

10. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Schwarzberg Lipson K, Knight R, Caporaso JG, Segata N, Huttenhower C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods 15: 962-968.Alexandrescu,A. (2001) Modern C++ Design: Generic Programming and Design Patterens Applied. Addision Wesley Professional, Boston.

Supplemental Figure 1: Community Profiles for Healthy Individuals and CRC Patients



Supplemental Figure 2: Gene Family Differences Between Healthy Individuals and CRC Patients



1. [↑](#endnote-ref-1)