Sphingolipid Exploration: Understanding the Role of Sphingolipid Metabolism in Cancer

Isabella R. Posey[a]

[a] Biomedical Engineering Fourth-Year Undergraduate Student
[1] Correspondence: irp3ac@virginia.edu (email)

Abstract

The medical burden of cancer is expected to continue to be an issue due to expected rises in cases. Colorectal cancer presents as a particular issue warranting attention due to its ranking as the second most common cause of cancer death. Previous research within the field has started to investigate the role that certain gut bacteria play in sphingolipid perturbations and ultimately colorectal cancer development. Certain sphingolipid-producing bacteria within the gut are vital for proper gut health. However, other research has investigated bacterial strains such as enterotoxigenic *Bacteroides fragilis* (ETBF) and have found that this strain of bacteria may be associated with colorectal cancer development. Therefore, this project aimed to understand certain genetic components of ETBF that may be responsible for impacting sphingolipid metabolism, ultimately leading to the development of colorectal cancer. This in turn would provide a better understanding of the functional role of sphingolipids and characteristics of bacteria that may impact sphingolipid metabolism. Through a combination of bioinformatics and computational approaches, nucleotide sequences of ETBF were compared with other sphingolipid-producing bacteria commonly found within the gut to generate hypotheses that could be used for wet lab application. Through utilization of the Basic Local Alignment Search Tool, it was found that the most statistically significant nucleotide alignment occurred between ETBF and *Bacteroides fragilis* with an E-Value of 0.0 and percent identity of 98.78%. Further comparisons between these two bacteria were carried out using MATLAB, and it was found that ETBF yielded higher adenosine and thymine densities compared to *Bacteroides fragilis*. This information was then utilized for ETBF specific analysis, which found the presence of a gene, bft, that was not present in the *Bacteroides fragilis* sequence. The gene, bft, was further concluded to serve as a potential pathogenicity island involved in impacting the production of the sphingolipid, spingosine-1-phosphate.

Keywords: sphingolipids, bacteria, enterotoxigenic *Bacteroides fragilis*, colorectal cancer

Introduction

Cases of cancer and cancer deaths have increased in recent years and are expected to remain a medical burden.[1,2] In the United States alone, there were estimated to be almost two million newly reported cancer cases and over 500,000 cancer deaths in the year 2020.[1] Colorectal cancer has been presented as a particularly challenging disease due to its ranking as the second most common cause of cancer death and an expected increase in the number of cases in upcoming years. In the year 2020 alone, there were over 147,000 colorectal cancer diagnoses and 53,000 deaths.[3] Therefore, colorectal cancer is an area within the medical field that requires further exploration and research, especially since the causes are unknown and there is no cure.[4]

Previous research within the field of colorectal cancer has started to explore sphingolipid perturbations with regards to certain bacteria as possible reasons for colorectal cancer development. Sphingolipids are a subclass of membrane-bound lipids and are distinguishable by their sphingoid long-chain base.[5] The general roles of sphingolipids include involvement in cell growth, cellular structural integrity, apoptosis, and proliferation.[6,7] Sphingolipid-producing bacteria exist in much smaller numbers compared to widespread mammalian cells, and a majority are restricted to the Bacteroidetes phylum and select members of the Chlorobi phylum.[8] Additionally, sphingolipid-producing bacteria commonly associate with a eukaryotic host and mediate specific immune responses that assist in maintaining host health through reducing inflammatory responses, decreasing host ceramide levels, and increasing host-microbe symbiosis.[8,9,10] Therefore, this project aims to understand the relationship between sphingolipid metabolism and common gut bacteria, as well as the impact of dysregulation of this metabolism on potential colorectal cancer development and treatment approaches.

A majority of previous experimentation and therapeutic solutions involving sphingolipids have dealt with the general class of sphingolipids rather than the division of bacterial sphingolipids.[11,12] Various studies have made advances with possible therapeutic options such as modifying current chemotherapeutic approaches with sphingosine kinase inhibitors, but these approaches are not necessarily applicable to the bacterial classification of sphingolipids.[13] Furthermore, few studies have investigated the impact of bacterial sphingolipid metabolism and its involvement in potential therapeutic treatments. Jang et al. examined mRNA expression levels of ceramide synthase genes, which impact sphingolipid metabolism, and found statistically significant overexpression of the genes CERS2, CERS5, and CERS6 in colorectal cancer patients.[14] However, the mechanisms impacting the dysregulation of sphingolipids were not fully explained. Other studies have also shown that the bacteria, enterotoxigenic *Bacteroides fragilis* (ETBF), is commonly associated with colorectal cancer development.[15] Patterson et al. explored ETBF, a toxin-producing strain of *Bacteroides fragilis*, and found that inhibition of glucosylceramide synthase, the enzyme involved in producing the sphingolipid, glucosylceramide, reduced integrity of colon models.[16] However, further exploration of genetic components involved in bacterial sphingolipid production is needed to fully understand factors impacting sphingolipid metabolism.

By further exploring the genetic composition of ETBF, a strain of bacteria associated with colorectal cancer development, through a combination of bioinformatics and computational approaches, insight may be gained into possible reasons ETBF may be involved in colorectal cancer development, whereas other sphingolipid-producing bacteria are vital for proper gut health. Therefore, to explore this topic further, the specific aims for this project include:

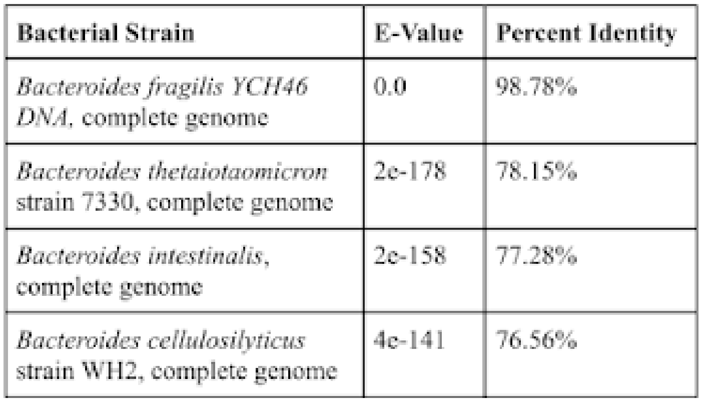
1. To characterize and better understand the functional role of sphingolipids and the role sphingolipid metabolism plays in colorectal cancer.
2. To develop models to analyze bacterial sphingolipid characteristics and potential points of interest for cell line development.
3. To develop a sphingolipid-deficient cell line to test hypotheses on sphingolipid characteristics to gain greater insight into the functional role of bacterial sphingolipids.

Through the exploration of these specific aims, it is hypothesized that certain genetic differences between ETBF and other sphingolipid-producing bacteria may play a role in ETBF being associated with colorectal cancer development. Development and testing of these hypotheses in turn will provide a better understanding of the functional role of bacterial sphingolipids, which may then be used for future experimentation and development of future treatment options for colorectal cancer in order to make progress towards decreasing the burden of cancer across the population.

Results

Basic Local Alignment Search Tool (BLAST) Results

The first section of results and analysis involved utilization of the Basic Local Alignment Search Tool (BLAST). During this section of data collection, candidate bacteria discovered during literature review were analyzed and compared using BLAST. Nucleotide alignment comparisons were taken between Enterotoxigenic *Bacteroides fragilis* (ETBF) and other bacteria known to be involved in sphingolipid production.



Bacterial Strain	E-Value	Percent Identity
Bacteroides fragilis YCH46 DNA, complete genome	0.0	98.78%
Bacteroides thetaiotaomicron strain 7330, complete genome	2e-178	78.15%
Bacteroides intestinalis, complete genome	2e-158	77.28%
Bacteroides cellulosilyticus strain WH2, complete genome	4e-141	76.56%

Table 1. BLAST Results. Alignment matches for ETBF were searched using BLAST. Select bacteria, corresponding E-Values, and percent identities are listed above. Lower E-Values and higher percent identities indicate a more statistically significant match.

As may be seen in Table 1, the most statistically significant alignment occurred between ETBF and *Bacteroides fragilis* (*B. fragilis*). The alignment between ETBF and *B. fragilis* (*Bacteroides fragilis YCH46 DNA*, complete genome as listed in Table 1) yielded an Expect Value (E-Value) of 0.0 and a percent identity of 98.78%. A lower E-Value indicates a more statistically significant alignment or significant match, and a higher percent identity indicates a higher similarity of genetic composition.[17] Additionally, other statistically significant alignments occurred between ETBF and other *Bacteroides* derived bacteria such as *Bacteroides thetaiotaomicron*, *Bacteroides intestinalis*, and *Bacteroides cellulosilyticus*. The alignment results yielded E-Values of 2e-178, 2e-158, and 4e-141, as well as percent identities of 78.15%, 77.28%, and 76.56%, respectively. Even though these alignments were not as statistically significant as the alignment between ETBF and *B. fragilis*, these results still indicated a significant match in nucleotide sequences.

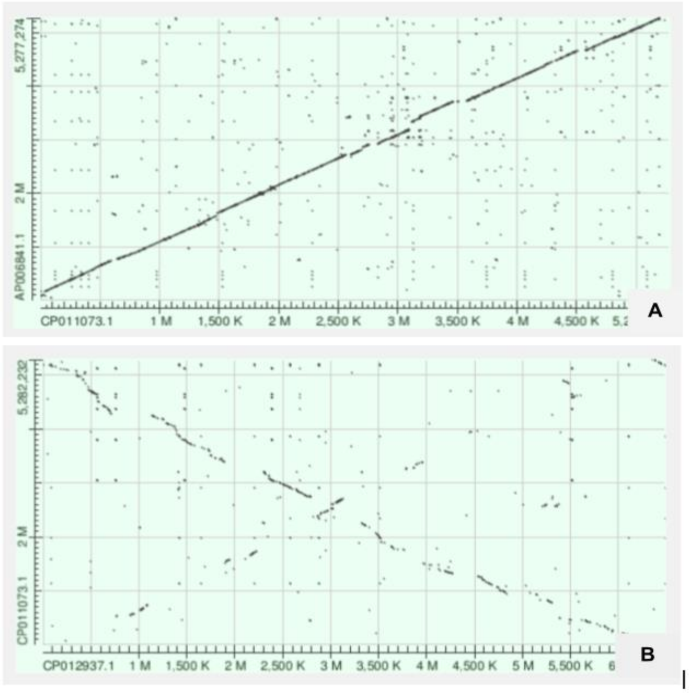


Figure 1. Dot Plots. Image A represents the dot plot for *B. fragilis* vs ETBF. Image B represents the dot plot for ETBF vs. *B. thetaiotaomicron*. The x and y axes represent the residues of the given bacteria, and a solid diagonal line indicates a significant match.

Additionally, visual representations of the alignment matches between each of the top hits were developed in order to better visualize the BLAST alignment results. As may be seen in Figure 1, the alignment between ETBF and *B. fragilis* appeared to be more significant and organized compared to the other sphingolipid-producing bacteria of interest such as *Bacteroides thetaiotaomicron*. The *B. fragilis* versus ETBF dot plot yielded a distinct diagonal line with few areas of dissimilarity, indicating a better genetic match (Figure 1, Image A). However, the same distinct diagonal line was not observed in the dot plot between ETBF and *Bacteroides thetaiotaomicron* (Figure 1, Image B). The lines that did appear on the dot plot were much more separated and not organized into a continuous diagonal formation, indicating a less significant match between the residues of the two bacterial strains. Therefore, moving forward, ETBF and *B. fragilis* were determined to be the two main bacteria of interest for further analysis.

MATLAB Sequence Results

Utilizing the initial BLAST results, further sequence analysis was performed using MATLAB to gain a better understanding of the genetic similarities and differences between ETBF and *B. fragilis*. Nucleotide density and A-T C-G density graphs were produced to determine apparent genetic differences between the two bacterial strains. ETBF yielded a higher distinction between the adenosine (A) and thymine (T) densities versus guanine (G) and cytosine (C) compared to *B. fragilis* (Figure 2, Images A and B). The nucleotide densities for *B. fragilis* were much more clustered together compared to ETBF (Figure 2, Image C). Also, the A-T C-G density plots showed differences between the densities of these base pairings between the two bacterial strains. As may be seen in Figure 2, Image B, ETBF yielded a consistent higher density of A-T base pairings compared to G-C base pairings. However, the *B. fragilis* A-T C-G density plot alternated points of higher densities for either A-T or G-C pairings (Figure 2, Image D).

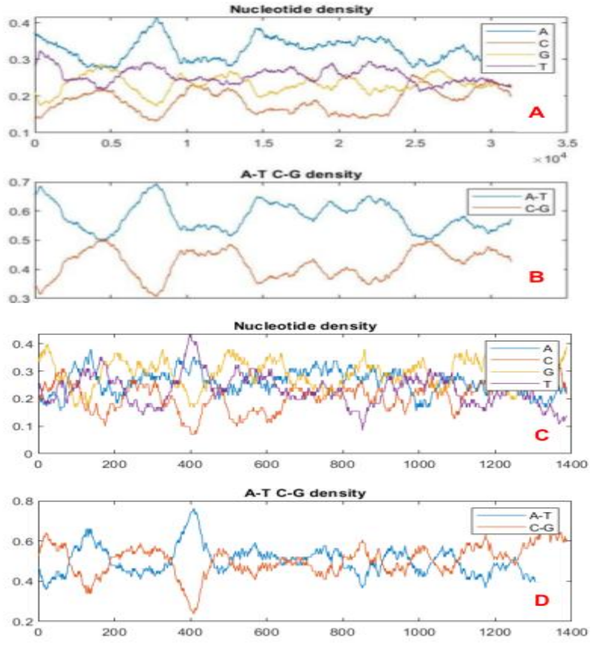


Figure 2. Density Graphs. Image A represents nucleotide density for ETBF. Image B represents A-T C-G density for ETBF. Image C represents nucleotide density for *B. fragilis*. Image D represents A-T C-G for *B. fragilis*. The x axis represents base location, and the y axis represents frequency.

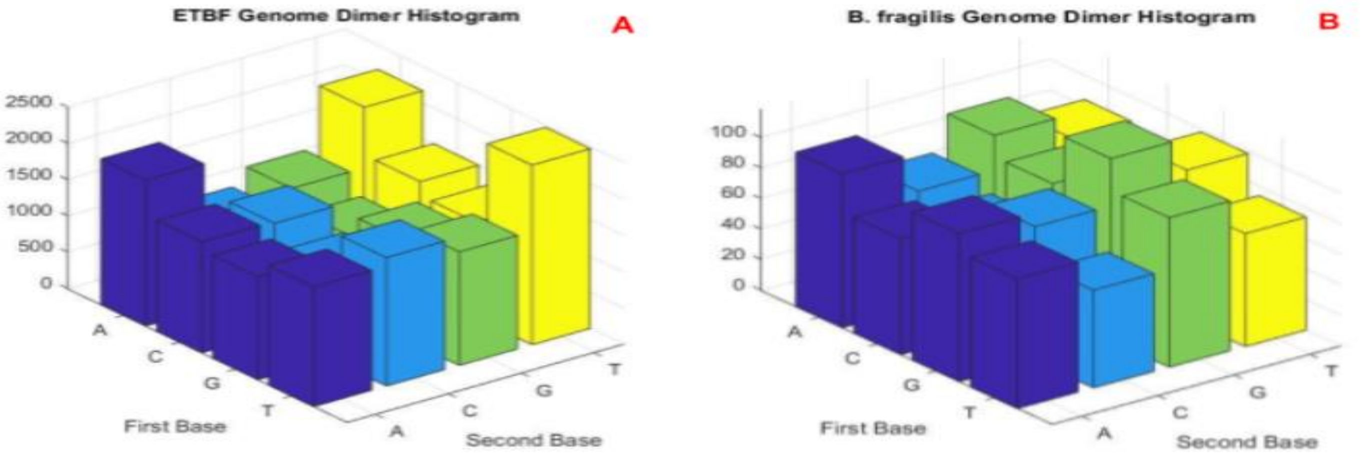


Figure 3. Dimer Histograms. Image A represents the dimer histogram for ETBF, whereas Image B represents the dimer histogram for *B. fragilis.* The x and y axes represent the first and second bases and the z axis represents the frequency of the dimer pairs.

Further comparative analysis was then performed between the dimer compositions of ETBF and *B. fragilis*. As may be seen in Figure 3, Images A and B, there were noticeable differences between the dimer compositions of both strains of bacteria. The dimer histogram for ETBF revealed a higher overall prevalence of dimer pairings involving thymine compared to the other bases (Figure 3, Image A).

Dimer Pair	ETBF	Percentage	*B. fragilis*	Percentage
AA	2009	**8.04%**	101	**7.33%**
AC	1109	4.44%	76	5.52%
AG	1328	5.32%	99	**7.19%**
AT	2161	**8.65%**	84	6.10%
CA	1530	6.12%	76	5.52%
CC	1502	6.01%	67	4.87%
CG	1064	4.26%	85	6.17%
CT	1507	6.03%	66	4.79%
GA	1419	5.68%	97	7.04%
GC	1222	4.89%	88	6.39%
GG	1358	5.44%	119	**8.64%**
GT	1316	5.27%	98	7.12%
TA	1649	6.60%	85	6.17%
TC	1769	7.08%	64	4.65%
TG	1564	6.26%	98	7.12%
TT	2475	**9.91%**	74	5.37%

Table 2. Dimer Counts. The total number of dimers for ETBF and *B. fragilis* are under each of their respective columns. The percentages of each dimer pair were calculated for both bacterial strains. The red dimer percentages represent the top three dimers for each bacterial strain.

For *B. fragilis*, the dimer histogram revealed a more even distribution of pairings among the bases, with the highest prevalence of dimer pairings involving guanine (Figure 3, Image B). In addition to attaining these general results from the dimer histograms of both bacterial strains, the total number of dimer pairs was found and dimer percentages for each bacterial strain were calculated. Percentages of each dimer were calculated to provide a means of comparison that accounted for differences in sequence lengths. As may be seen in Table 2, the highest dimer percentages for ETBF were found with TT (9.91%), AT (8.65%), and AA (8.04%). The highest dimer percentages for *B. fragilis* were GG (8.64%), AA (7.33%), and AG (7.19%). Both bacterial strains had AA as one of the highest percentage dimers. None of the dimer percentages exceeded 9% with *B. fragilis*, but ETBF had one dimer (TT) that was almost 10%.

Enterotoxigenic Bacteroides fragilis (ETBF) Specific Analysis

Initial comparisons between ETBF and *B. fragilis* were used as a basis for more thorough analysis of ETBF sequence data utilizing various Bioinformatics Toolbox functions within MATLAB and the Sequence Viewer app. The first part of analysis involved the creation of a heat map of the codon frequencies of ETBF. As may be seen in Figure 4, the codons of AAA, ATA, AAG, GAA, and ATG presented as some of the higher frequency codons within the ETBF genetic sequence. The lighter regions on the heat map indicate a higher frequency, whereas the darker regions indicate a lower frequency. This information was then explored further through locating and analyzing opening reading frames, which are sequences of DNA or RNA that have the potential to be translated into a protein.[18]

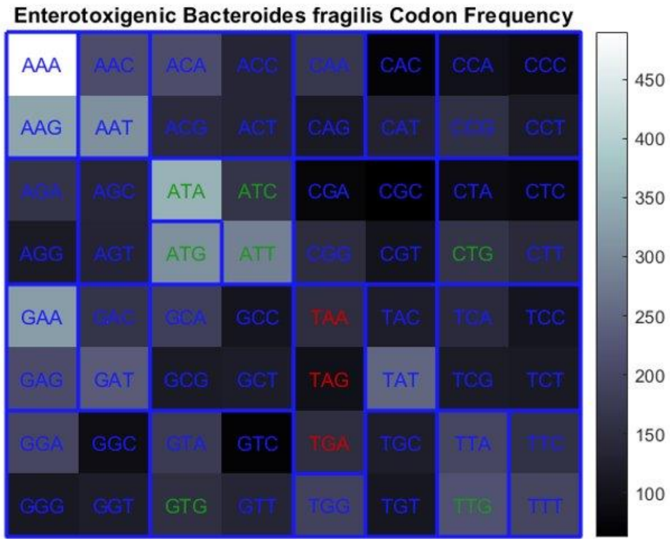


Figure 4. ETBF Heat Map. The heat map above shows the codon frequencies for ETBF. The lighter regions represent a higher frequency, whereas the darker regions represent lower frequencies.

Initial MATLAB data collected through exploring opening reading frames included the presence of multiple larger genes located within all three reading frames. Thus, another approach was required in order to locate potential genes within ETBF that may impact sphingolipid metabolism. Feature extraction occurred, which extracted DNA sequences and paired them with their given features. From this process, the gene, bft, was extracted from the ETBF sequence and further analyzed. The matched feature information indicated that the protein associated with bft was bfmC. The location of this gene within the genetic sequence of ETBF was determined to start at the 10,117 base pair and end at the 11,286 base pair. Also, this gene was not found within the *B. fragilis* sequence.

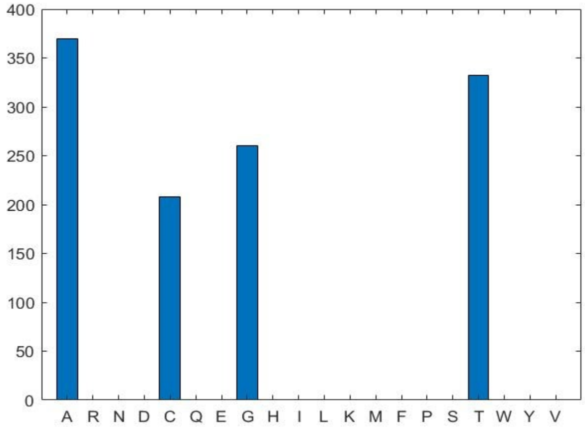


Figure 5. Amino Acid Histogram for bfmC Protein. The histogram above includes the amino acid labels on the x axis and frequency on the y axis. The four amino acids found were alanine, glycine, cysteine, and threonine.

The amino acid content of the bfmC protein was then explored through the creation of an amino acid histogram. As may be seen in Figure 5, the four amino acids that were present included alanine (A), glycine (G), cysteine (C), and threonine (T). Alanine occurred at the highest amount, whereas cysteine occurred at the lowest. This combination of information was then utilized to determine the potential impact of the presence of this gene within ETBF.

Discussion

The results obtained from the combined bioinformatics and computational approaches may be utilized to gain a better understanding of the role gut bacteria, such as ETBF, may play in sphingolipid metabolism. BLAST results yielded statistically significant matches to multiple bacteria that both exist in the gut and have been determined to be involved in sphingolipid metabolism. The most statistically significant match, which occurred between ETBF and *B. fragilis*, was expected due to ETBF being a strain of *B. fragilis*. The high percentage identity and low E-Value led to the genetic differences between the two bacteria to be explored further to try and understand the genetic differences that cause ETBF to be associated with colorectal cancer development.

The genetic differences that were observed between ETBF and *B. fragilis* provided further direction for understanding the role bacterial genetics play in sphingolipid metabolism. From initial MATLAB results, the ETBF sequence appeared to contain higher A-T pairings and densities compared to the *B. fragilis* sequence (Figure 2). Furthermore, the dimer data collected on both bacterial strains supported the higher A-T richness found in ETBF compared to the more evenly distributed A, G, C, and T base densities found in *B. fragilis* (Figure 2 and Table 3). This information was then utilized to look for A-T rich areas within the nucleotide sequence of ETBF, since the higher density A and T regions deviate from the makeup of *B. fragilis* and potentially indicate a genetic region that may play a role in impacting sphingolipid metabolism and colorectal cancer development.

Ultimately, through additional codon and protein analysis of ETBF, it was found that the presence of a certain gene, bft, within the ETBF sequence started at base pair 10,117 and ended at the 11,286 base pair. This specific sequence section contained a high amount of the amino acids of alanine (A), glycine (G), cysteine (C), and threonine (T), and encodes for the protein bfmC. The product of this specific sequence is associated with metalloprotease-toxin-

2, which is also known as the *Bacteroides fragilis* toxin (BFT).[15,16] This same sequence was not found within the *B. fragilis* strain, which indicates a possible genetic difference responsible for ETBF's involvement in sphingolipid perturbations. Also, previous experimentation has shown that the presence of this toxin (BFT) can induce colitis and cause colon tumor formation within mice.[16] Therefore, it may be concluded that the base pair locations on the ETBF strain between 10,117 and 11,286 may be responsible for encoding for the BFT toxin, which is involved in colorectal cancer development.

Additionally, the specific section encoding for BFT may be determined to be a pathogenicity island, which would be responsible for virulent factors such as toxin-production. Previous research identified a larger pathogenicity island (BfPAI) within ETBF strains that was associated with subjects that developed diarrheal disease. This research also identified two mobilization genes, bfmA and bfmB, that were located near this pathogenicity island.[15] Both bfmA and bfmB were present near the bft gene that was found during this project, which would further support the bft gene found between the base pairs of 10,117 and 11,286 on the ETBF strain being a pathogenicity island.

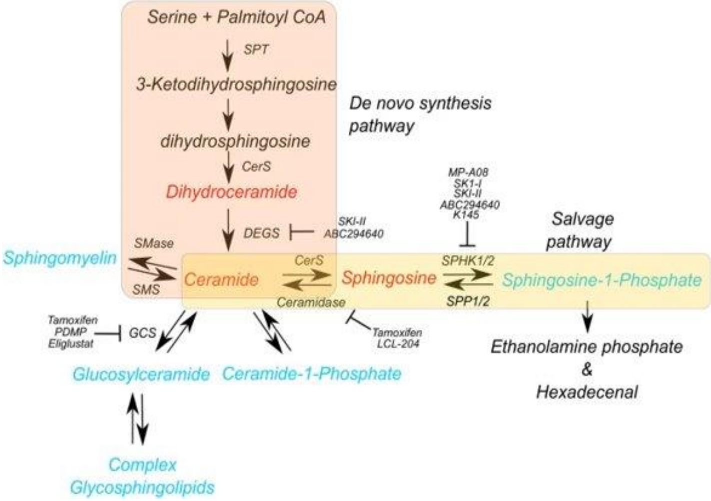


***Figure 6. Sphingolipid Metabolism Pathway*.** The model above shows the pathways involved in sphingolipid metabolism. *B. fragilis* impacts the SPT enzyme, whereas as ETBF is hypothesized to impact the production of sphingosine-1-phosphate.

This information may be further applied to the sphingolipid pathway. *B. fragilis* contributes to sphingolipid metabolism through the presence of the gene that encodes for the enzyme, serine palmitoyltransferase.[7,8,19] This enzyme is needed for the conversion of serine and palmitoyl CoA into 3-Ketodihydrosphinosine in the de novo synthesis pathway of sphingolipid metabolism (Figure 6). Without this conversion step, sphingolipid production cannot take place, causing a sphingolipid deficiency. However, *B. fragilis* does not contain the proposed pathogenicity island found in ETBF. Due to the location of the proposed pathogenicity island, location of this toxin-producing gene, and amino-acid composition, it may be hypothesized that ETBF potentially affects sphingolipid metabolism by impacting the production of the signaling sphingolipid, sphingosine-1-phosphate.[15] Specifically, due to the previously mentioned components of ETBF, it may possible that ETBF and its toxin-producing capabilities impact the SPHK1/2 and SPP1/2 components of the Salvage pathway. However, further wet lab experimentation needs to be conducted in order to confirm or deny this. Overall, the findings from this project do support than certain genetic components within gut bacteria such as *B. fragilis* and ETBF may impact sphingolipid production and ultimately assist in determining whether colorectal cancer development occurs.

Limitations and Constraints

One of the major limitations within this project was the inability to utilize a wet lab setting due to COVID-19 restrictions. Originally, the project was to include a balance between computational and wet lab approaches. However, due to lab restrictions, the project was adapted to be more computationally driven. This in turn led to hypotheses being developed based on the results obtained through the given results, but also led to the inability to test these hypotheses in an *in vitro* or *in vivo* lab setting due to undergraduate research restrictions. Another limitation of this project involved the nucleotide sequence data that was utilized for analysis. All nucleotide sequence data was obtained from publicly available databases such as BLAST or GenBank. Since the data was accessed through these databases, there was no control over the exact sequence data that was given. The sequences available were assumed to be correctly labeled and accurate. Analysis had to be performed using the information that was accessible. This involved utilizing partial sequence data or incomplete bacterial strain information that in turn may have impacted the results obtained, formulated hypotheses for wet lab experimentation, and the holistic understanding of the involvement of bacterial sphingolipids in colorectal cancer development. BLAST searches also included certain limitations. Due to BLAST being a heuristic algorithm, all alignments may not be reported or determined to be significant. This means that not all results may be repeatable and completely objective if the database expands since reported alignments are determined based on database size and alignment length. Additionally, BLAST uses an algorithm shortcut that makes it difficult to search for shorter sequences.[20] Lastly, another design constraint included the inability of MATLAB to process very large nucleotide sequence information. This required either partial nucleotide sequences to utilized for analysis or sections of nucleotide sequences to be analyzed at different times, which may have impacted the results and conclusions of this project.

Alternatives

Possible alternatives for this project include the utilization of different genetic databases, as well as alterations to the timeline of the methodology to allow for consistent testing of formulated hypotheses. BLAST was used to gather nucleotide sequences for the given bacteria but included certain limitations that may have impacted the collected data. Possible alternatives include the utilization of different bioinformatics approaches and algorithms such as BLAT, a modified DNA alignment tool, and Ensembl Bacteria, a browser for bacterial genomes.[21,22] Additionally, an alternative approach may be to utilize either BLAST or MATLAB, instead of combining both approaches, since the combination of approaches may have led to discrepancies in results due to the different statistical assumptions and approaches of both systems. Lastly, an alternative approach would include incorporating wet lab applications throughout the data collection period instead of at the very end of the project, which would allow for different hypotheses to be tested throughout the duration of the experiment. This would allow more application-based data to be collected, which would provide insight into certain sphingolipid metabolism elements that may not be apparent through a bioinformatics or computational approach.

Future Work

Immediate steps towards future work would include testing the hypotheses that were formulated throughout this project in a wet lab setting. Testing within a wet lab setting will allow for the hypotheses concerning ETBF to either be supported or rejected and will provide for a better understanding of the role of bacterial sphingolipids in colorectal cancer progression. Specifically, the information found during this project regarding the presence of a possible pathogenicity island in ETBF would be tested through knocking out this specific gene sequence and then determining if the resulting bacteria is sphingolipid-deficient. Lipid assays would be conducted to determine if sphingosine-1-phosphate, which was the sphingolipid that was predicted to be impacted due to the genetic composition of ETBF, is present with the ETBF knockout. Additionally, these lipid results would also be run against lipid results obtained from *B. fragilis*, *B. fragilis* with a SPT knockout, and ETBF without the pathogenicity island knockout. *B. fragilis* with the SPT knockout and ETBF should yield sphingolipid deficiencies, whereas *B. fragilis* and ETBF with the pathogenicity island knockout should not yield sphingolipid deficiencies.[15] Additionally, future work may include investigating other elements of the sphingolipid metabolism pathways such as creating knockout bacteria of the sphingosine-1-phosphate generating enzyme in order to create a sphingolipid-deficient cell line that may be used for further testing.[7] Long-term future work would include utilizing a previously developed sphingolipid-deficient cell-line for testing novel drug techniques to advance therapeutic options for colorectal cancer. Development of a sphingolipid-deficient cell line would be useful in serving as a testing model for future experimentation and research. Once a sphingolipid-deficient cell line is developed, more information may be gained about the various factors that affect sphingolipid metabolism and thus colorectal cancer development. Furthermore, a sphingolipid-deficient cell line may be widely used throughout the research community for testing potential cancer drugs and treatments options.

Materials and Methods

A thorough literature review was performed and used to collect background information on sphingolipids, sphingolipid metabolism, and potential bacteria of interest. During this literature review, ETBF became a bacterium of interest and was explored further in the subsequent steps.

Basic Local Alignment Search Tool Analysis

The Basic Local Alignment Search Tool (BLAST), which is a program that examines similarities between biological sequences and determines statistical significance between nucleotide and protein sequences, was used to search for nucleotide sequences that closely aligned with ETBF.[23] The accession number for ETBF was found using the BLAST Genomes search feature provided by BLAST. The resulting accession numbers were then cross-referenced with literature values. The resulting accession number for ETBF, CP011073.1, was chosen due to the completeness of the sequence and its ability to be run within the nucleotide BLAST (blastn) feature completely. The accession number was entered in the query search of the blastn feature, and the top 20 most statistically significant alignment matches appeared. Each alignment match was referenced with the list of sphingolipid-producing bacteria found within the gut, which was created during the literature review portion of the project. The ETBF alignment matches that were considered gut bacteria involved in sphingolipid metabolism were then pulled for further analysis. Further analysis involved collecting Expect Values (E-Values) and percent identities. The E-Value represents the number of hits that are expected to be seen by chance when searching a database of a particular size. A E-Value less than 0.05 was determined to be statistically significant, with lower E-Values or E-Values closer to zero being deemed more statistically significant.[17] Additionally, percent identities were used to determine the similarity between the ETBF sequence and the bacteria sequence of interest. The percent identity provides a percentage of the characters in each sequence that are identical. Therefore, a higher percent identity was interpreted as a more statistically significant match.[24] The blastn feature was then used to create dot plots of the top alignment matches to ETBF. This feature allowed for sequences of interest to be chosen and used to create dot plots, which provided a visual representation of the alignment matches to ETBF. A distinct diagonal line formed on a dot plot represents a more significant match.[25] After these steps were completed, the information gained from BLAST analysis was explored further utilizing MATLAB.

MATLAB Analysis

The Bioinformatics Toolbox within MATLAB was downloaded and utilized for sequence analysis. Sequence information was collected from GenBank, which is a genetic sequence database created by the NIH that contains annotated collections of publicly available DNA sequences.[26] Accession numbers for ETBF and the other significant alignment matches were searched using the Nucleotide search feature. Each accession number attained was referenced with previous literature, and the location, published year, base pair counts, and other annotations were taken into consideration. Additionally, sequences containing contigs, which are series of overlapping DNA sequences, had to be excluded since the utilized MATLAB functions did not have the ability to account for contigs in analysis.[27] Therefore, the most complete sequences that did not contain contigs and were labeled properly were utilized for MATLAB analysis. The accession number used for ETBF was AY372755, and the accession number used for *B. fragilis* was HE608160. The latest versions of both annotated sequences were loaded into MATLAB. Initial exploration of the sequences involved collecting the total length of the sequences and accessing different parts of the genetic sequences through MATLAB indexing commands. The compositions of both ETBF and *B. fragilis* were explored using the *ntdensity* function within MATLAB and differences between the base pairs of the two bacterial strains were compared using the resulting nucleotide density graphs. Furthermore, the base and dimer counts were collected using the *basecount* and *dimercount* functions. Dimer counts were compared between ETBF and *B. fragilis* and total percentages of each dimer were taken to provide a means of comparison that accounted for differences in sequence lengths. Open reading frames (ORFs), which are sequences of DNA or RNA that can potentially be translated into proteins, were explored in ETBF and *B. fragilis* using the *seqshoworfs* function.[18]

ETBF-specific analysis involved inspection of annotated features provided by GenBank. Annotated coding sequences (CDS) were explored and cross-referenced with the previously identified ORFs.[18] The yielded CDS were gathered for ETBF and the location, gene, indices, products, protein IDs, sequence, and any notes were collected for each sequence section. The proteins collected were extracted and analyzed using the *nt2aa* function, and the amino acid content of each protein was analyzed using the *aacount* function. A codon frequency map of ETBF was produced, and the codon counts for ETBF as well as each CDS feature were calculated. This information was used for comparison with *B. fragilis* and was cross-referenced with previous experimentations involving ETBF. Differences in codon counts, base pair densities, genes, locations of genes, and amino acid compositions were compiled and compared. This information was then utilized and applied to the sphingolipid metabolism model to make informed hypotheses about the genetic differences between ETBF and *B. fragilis* that may contribute to ETBF being associated with colorectal cancer development through means of sphingolipid perturbations.

<u>End Matter</u>

Author Contributions and Notes

The author declares no conflict of interest.

Acknowledgments

I would like to acknowledge Dr. Todd Fox in the Department of Pharmacology for serving as my technical advisor for these past two semesters. His support and guidance allowed me to explore this topic more thoroughly. I would also like to thank Dr. Kester and the rest of the Kester Lab for their support of my undergraduate research experience and ultimately this project. I would also like to acknowledge Dr. Logan Patterson, who served as my research mentor starting my second year and led me to get involved in this project. Lastly, I would like to thank the Biomedical Engineering Capstone teaching team, which was led by Dr. Barker and Dr. Allen and included Delaney Fisher, Kareem El-Ghazawi, and Taylor Eggersten.

References

1. Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. CA: A Cancer Journal for Clinicians, 70(1), 7-30. doi:10.3322/caac.21590
2. Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. CA: A Cancer Journal for Clinicians, 69(1), 7-34. doi:10.3322/caac.21551
3. Siegel, R. L., Miller, K. D., Sauer, A. G., Fedewa, S. A., Butterly, L. F., Anderson, J. C., . . . Jemal, A. (2020). Colorectal cancer statistics, 2020. CA: A Cancer Journal for Clinicians, 70(3), 145-164. doi:10.3322/caac.21601
4. National Cancer Institute. (2021, January 25). Colon cancer treatment (pdq®)–health professional version. https://www.cancer.gov/types/colorectal/hp/colon-treatment-pdq#:~:text=Cancer%20of%20the%20colon%20is,the%20ultimate%20cause%20of%20death.
5. .Lahiri, S., & Futerman, A. H. (2007). The metabolism and function of sphingolipids and glycosphingolipids. Cellular and Molecular Life Sciences, 64(17), 2270-2284. doi:10.1007/s00018-007-7076-0
6. Saddoughi, S. A., Song, P., & Ogretmen, B. (n.d.). Roles of Bioactive Sphingolipids in Cancer Biology and Therapeutics. Subcellular Biochemistry Lipids in Health and Disease, 413-440. doi:10.1007/978-1-4020-8831-5_16
7. Kitatani, K., Taniguchi, M., & Okazaki, T. (2015). Role of Sphingolipids and Metabolizing Enzymes in Hematological Malignancies. Molecules and Cells, 38(6), 482-495. doi:10.14348/molcells.2015.0118
8. Heaver, S., Johnson, E., & Ley, R. (2018). Sphingolipids in host–microbial interactions. Current Opinion In Microbiology, 43, 92-99. doi: 10.1016/j.mib.2017.12.011
9. Sphingolipid. (1980). https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/sphingolipid
10. Brown, E., Ke, X., Hitchcock, D., Jeanfavre, S., Avila-Pacheco, J., Nakate, T.,.Xavier, R. (2019). Bacteroides-derived sphingolipids are critical for maintaining intestinal homeostasis and symbiosis. Cell Host & Microbe, 25(5), 668-680. doi:10.1016/j.chom.2019.04.002.
11. Furuya, H., Shimizu, Y., & Kawamori, T. (2011). Sphingolipids in cancer. Cancer and Metastasis Reviews, 30(3-4), 567-576. doi:10.1007/s10555-011-9304-1
12. Ponnusamy, S., Meyers-Needham, M., Senkal, C. E., Saddoughi, S. A., Sentelle, D., Selvam, S., . . . Ogretmen, B. (2010). Sphingolipids and cancer: Ceramide and sphingosine-1-phosphate in the regulation of cell death and drug resistance. Future Oncology, 6(10), 1603-1624. doi:10.2217/fon.10.116
13. Truman, J., García-Barros, M., Obeid, L. M., & Hannun, Y. A. (2014). Evolving concepts in cancer therapy through targeting sphingolipid metabolism. Biochimica Et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids, 1841(8), 1174-1188. doi:10.1016/j.bbalip.2013.12.013
14. Jang, S., Park, W., Min, H., Kwon, T., Baek, S., Hwang, I., . . . Park, J. (2018). Altered mRNA expression levels of the major components of sphingolipid metabolism, ceramide synthases and their clinical implication in colorectal cancer. Oncology Reports. doi:10.3892/or.2018.6712
15. Franco AA, Cheng RK, Chung GT, Wu S, Oh HB, Sears CL. Molecular evolution of the pathogenicity island of enterotoxigenic Bacteroides fragilis strains. J Bacteriol. 1999 Nov;181(21):6623-33. doi: 10.1128/JB.181.21.6623-6633.1999. PMID: 10542162; PMCID: PMC94125.
16. Patterson, L., Allen, J., Posey, I., Shaw, J. J., Costa-Pinheiro, P., Walker, S. J., . . . Kester, M. (2020). Glucosylceramide production maintains colon integrity in response to Bacteroides fragilis toxin-induced colon epithelial cell signaling. The FASEB Journal. doi:10.1096/fj.202001669r
17. Madden, T. The BLAST Sequence Analysis Tool. https://unmc.edu/bsbc/docs/NCBI_blast.pdf
18. MathWords. (n.d.). Calculating and visualizing sequence statistics. Retrieved from https://www.mathworks.com/help/bioinfo/ug/calculating-and-visualizing-sequence-statistics.html
19. Johnson, E., Heaver, S., Waters, J., Kim, B., Bretin, A., Goodman, A.,...Ley, R. (2020). Sphingolipids produced by gut bacteria enter host metabolic pathways impacting ceramide levels. Nature Communications, 11(2471), https://doi.org/10.1038/s41467-020-16274-w
20. Resnik, R. (2015). 3 Problems with NCBI BLAST and Finding Sequence Alignments | GQ Life Sciences.https://www.gqlifesciences.com/3-problems-with-using-blast-for-sequence-alignments-in-ip-searching/
21. Kent, J. (2020). Human BLAT Search. https://genome.ucsc.edu/cgi-bin/hgBlat
22. Ensembl Bacteria. (2020). http://bacteria.ensembl.org/index.html
23. National Institutes of Health. (2020). BLAST: Basic Local Alignment Search Tool. Retrieved 11 December 2020, from https://blast.ncbi.nlm.nih.gov/Blast.cgi
24. Tufts University. (n.d.). Using BLAST. https://ase.tufts.edu/chemistry/walt/sepa/Activities/BLASTpractice.pdf
25. Piecewise sequence comparison. (n.d.). http://www.bioinfo.org.cn/lectures/index-4.htm#:~:text=Thus%2C%20when%20two%20sequenes%20share,example%20of%20a%20dot%20plot.
26. National Institute of Health. (n.d.). GenBank overview.https://www.ncbi.nlm.nih.gov/genbank/
27. National Human Genome Research Institute. (n.d.). Contig. https://www.genome.gov/genetics-glossary/Contig