Programa de Estudios de Honor
Universidad de Puerto Rico
Recinto de Río Piedras

Honors Program
University of Puerto Rico
Rio Piedras Campus

TESIS O PROYECTO DE CREACIÓN

Aprobado como requisito parcial del
Programa de Estudios de Honor

Comité de Tesis o Proyecto de Creación — Nombre

Mentor Dra. Rosa E. Guzzardo Tamargo

Director de Estudios Dra. Mayra Vélez Serrano

Lector Dr. Giovanni Tirado Santiago

Lector Dra. Alma Simounet Bey

Lector ______________________

Visto Bueno Dra. Elaine Alfonso
Director PREH o su Representante

12 de diciembre de 2021
Fecha

University of Puerto Rico

Río Piedras Campus

Honor Studies Program

The Hemolysin-Coregulated Protein (Hcp) from the Bacterial Genotoxin Usp as a Risk Factor for Adenoma and Colorectal Cancer

By

Gabriela S. Montalvo Noguera

November 22, 2021

Thesis Committee:

Abel Baerga Ortiz, Ph.D.
Professor, Department of Biochemistry
School of Medicine, University of Puerto Rico, Medical Science Campus

Abiel Roche-Lima, Ph.D
Director, Integrated Informatics Services core IIS-RCMI
University of Puerto Rico, Medical Science Campus, RCMI Program

José A. Rodríguez Martínez, Ph.D.
Associate Professor, Department of Biology
University of Puerto Rico, Rio Piedras Campus

Abstract

Gut bacteria secrete genotoxic and pro-inflammatory factors that contribute to the development of colorectal cancer (CRC). The genotoxin Uropathogenic Specific Protein (Usp), made by some *E. coli* strains, is of particular interest since previous studies from our group identified a higher frequency of the *usp* sequence in Adenoma and CRC stool samples from US and PR cohorts, compared to healthy individuals. By using DNA sequencing and bioinformatic approaches, we also found *usp* genetic variants directly associated with CRC. It was observed that the *usp* sequence contained a conserved domain resembling the hemolysin-coregulated protein (Hcp), a component of the Type VI Secretion System (T6SS), that is involved in the secretion of toxic molecules. Due to its ring-like structure, Hcp functions as a channel that connects to the VgrG spike protein, which penetrates the target cells under this secretory system. However, the role of Hcp as a domain embedded in the N-terminal of Usp, and its relationship with inflammatory conditions such as adenomas and CRC, remains unclear. Therefore, this project aims to investigate the frequency of Hcp domain sequences of Usp, along with other Hcp variants from *E. coli*, in the metagenomic datasets of healthy, adenoma, and CRC patients. To our knowledge, this will be the first attempt to elucidate an association between Hcp and cancer progression using publicly available metagenomic data.

Acknowledgements

Throughout this project, I have been encouraged, helped, and supported by many people. Of these individuals, I would first like to thank my mentor and personal investigator, Dr. Abel Baerga-Ortiz, who accepted me into his laboratory group and introduced me to the project. His expertise and professional insight served as a trustful guide that continuously steered me in the right direction. I would also like to thank Dr. Abiel Roche-Lima and José A. Rodríguez Martínez. They also formed part of the thesis committee and thus provided helpful suggestions and made thoughtful questions that contributed to the project's success and refinement.

I would also like to thank my graduate student, Rachell Martínez, and RCMI Research Assistant, Kelvin Carrasquillo. They both served an essential role as they patiently taught me everything I needed to know about the project and how I could carry it out. In addition, I would like to thank the Honors Program of the University of Puerto Rico, who inspired me to be a part of this project and guided me through the process. Specifically, I am grateful for Dr. Ivelisse Rubio and Dr. Elaine Alfonso, who were always available to answer any questions or concerns that arose along the way. Moreover, I would like to thank Dr. Abel Baerga Ortiz's laboratory group, who supported me throughout the process and constantly made helpful recommendations that were incorporated into the study. Finally, I would like to thank my family for always supporting and motivating me as I take on new projects such as this one; I would not be where I am if it wasn't for them.

List of Figures and Tables

List of Abbreviations

CRC	Colorectal cancer
E. coli	*Escherichia coli*
Usp	Uropathogenic specific protein
Imu	Immunity protein
IBD	Inflammatory Bowel Disease
Hcp	Hemolysin-coregulated protein
T6SS	Type VI Secretion System
E-I	Effector-immunity
APEC	Avian pathogenic *Escherichia coli*
ExPEC	Extraintestinal porcine pathogenic *E. coli*
BLAST	Basic Local Alignment Search Tool
vgrG-1	Valine-glycine repeat protein G-1
tssk	Testis Specific Serine Kinase
AMmurB	*Akkermansia muciniphila* murB
ECmurB	*Escherichia coli* murB
EFmurB	*Enterococcus faecalis* murB

tcpC	TIR domain-containing protein
gelE	Gelatinase E
cnf-1	Cytotoxic necrotizing factor
WGS	Whole genome sequencing

Table of Contents

Chapter 1. Introduction

The American Cancer Society predicts that in 2021, approximately 52,980 Americans will perish from colorectal cancer (CRC). In the United States of America, CRC is the third most common cause of death due to cancer for both men and women separately and, when grouped together, it becomes the second leading cause of cancer-related death. The Puerto Rican Association of Gastroenterology states that CRC is the second most common cancer type, in both men and women, in Puerto Rico. The association also mentions that CRC is the second cause of death due to cancer in women and, in the case of men, it is the third cause.

After numerous years of investigation, researchers have concluded that the development of CRC is dependent on both genetic and environmental factors. As the PDQ Cancer Information Summaries states, approximately 75% of individuals with colorectal cancer have a sporadic disease, that is, their condition appears to not have been inherited. These evidence-based summaries also inform how the remaining 10% to 30% of cases do have a family history of colorectal cancer, therefore their condition is more likely to have emerged from a hereditary contribution on the genetic level (PDQ Cancer Genetics Editorial Board, 2020). The CDC states that the 75% of cases that did not inherit the condition genetically can be due to different factors such as low fiber and high-fat diets, high consumption of processed meats, inflammatory bowel diseases, obesity, tobacco use, alcohol consumption, and other environmental cues (CDC, 2021).

An environmental factor that has recently been given much attention is the gut microbiota, or the diverse community of microorganisms that lives in the human gut. It is thought that gut microbiota plays an essential role in cancer development. An intrinsic relationship between gut microbiota and colorectal cancer has been discovered since these gut bacterial communities can promote or suppress tumor development and growth. It has been found that the microbiota in the intestines uses different mechanisms to manipulate the environment, which may lead to tumorigenesis (Brennan & Garett, 2016). A vital member of these microbial communities is *Escherichia coli*. This bacterium inhabits the intestines of healthy animals and humans, but there are numerous strains of *E. coli* that could be opportunistic, which explains why its presence may trigger different effects. For example, it has been found that *E. coli* that produce colibactin, such as B2 *E. coli*, are associated with CRC development (Raisch, et al., 2014). Specifically, *E. coli* has been found with greater frequency in the mucosa of patients with CRC than in that of healthy individuals (Swidsinski, et al., 1998; Martin, et al., 2004).

The Uropathogenic specific protein, or Usp, is a bacteriocin-like genotoxin found in some *E. coli* strains that has demonstrated to cause DNA damage and cytoskeleton rearrangement in endothelial cells as it is shown in **Figure 1.1** (Nipič, et al., 2013). The research paper *"Escherichia coli* Uropathogenic-Specific Protein, Usp, is a Bacteriocin-Like Genotoxin" (Nipič, et al., 2013) explored the biochemical and biological function of Usp as well as its ability to cause DNA damage in the host cell, and it was discovered that "Usp possesses DNase activity and, particularly when coapplied with Imu2, exhibits genotoxic activity in mammalian cells" (Nipič, et al., 2013, p. 1545). Further work determined that infection with *E. coli* strains harboring Usp together with its imu1-3 proteins, affected mammal cells' viability and induced apoptosis.

Figure 1.1

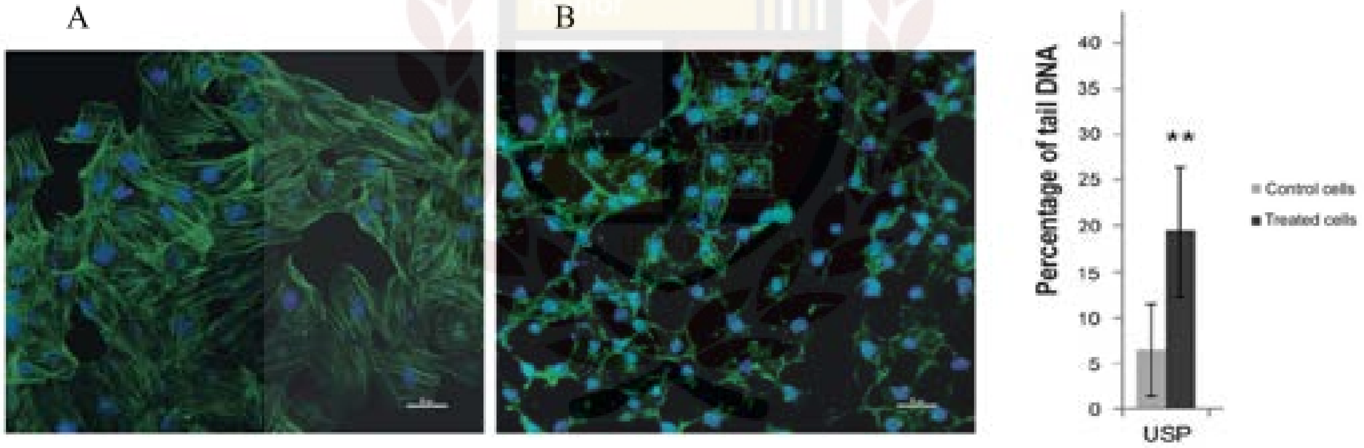


Figure 1.1 *Usp DNA damage in HUVEC cells using Fluorescence microscopy and Comet Assay (Nipic, et al., 2013).* The images display the control group of human umbilical vein endothelial cells (HUVEC) infected with *E. coli usp*$^-$ *imu1-3*$^-$ (A) and HUVECs infected with E. coli usp$^+$ imu1-3$^+$ (B). The graph included reflects the results from the Comet Assay, which shows that treated *usp* cells displays a statistically significant increase in genotoxic activity when compared to the control cells, which is seen in the greater percentage of tail DNA, an indicator of DNA damage. Standard deviation error bars are included. Asterisks (**) indicate $P < 0.001$when compared to control cells.

Previous research from our lab has established a relationship between the presence of specific genes, in this case the gene encoding Usp in *E. coli,* and CRC. This topic is explored in "The Presence of Gut Microbial Genes Encoding Bacterial Genotoxins or Pro-Inflammatory Factors in Stool Samples from Individuals with Colorectal Neoplasia" (Gómez-Moreno, et al., 2019, which evaluates the presence of certain genes linked to genotoxicity and pro-inflammatory factors in patients with adenomas and/ or CRC from US and Puerto Rico. The results obtained after completing the PCR analysis of the stool samples revealed a higher presence of *usp* in CRC and adenomas patients when compared to healthy individuals in the US cohort, while in PR this presence is higher in adenomas cohort (**Figure 1.2**), indicating a possible trend that links genotoxic genes such as *usp* to these diseases.

The paper, "Hotspots of Sequence Variability in Gut Microbial Genes Encoding Pro-Inflammatory Factors Revealed by Oligotyping" (Gómez-Moreno, et al., 2019) furthers this analysis. In said report, the sequence variability for some of these bacterial genes was explored, using a combination of deep sequencing and the process known as oligotyping, which is an application for data analysis that identifies the "hotspots" of mutations in short segments of DNA. The research found that the amplicons for *usp* from stool samples revealed the presence of five distinct oligotypes in two different regions (Gómez-Moreno, et al., 2019, p. 1). The GT sequence oligotype was present in a single isolate of uropathogenic bacteria and was also detected in fecal samples of individuals with CRC. Variants of the *usp* gene sequence in metagenomic databases were also found, which raised the possibility that some variants may have different activities and different toxicity profiles.

Figure 1.2

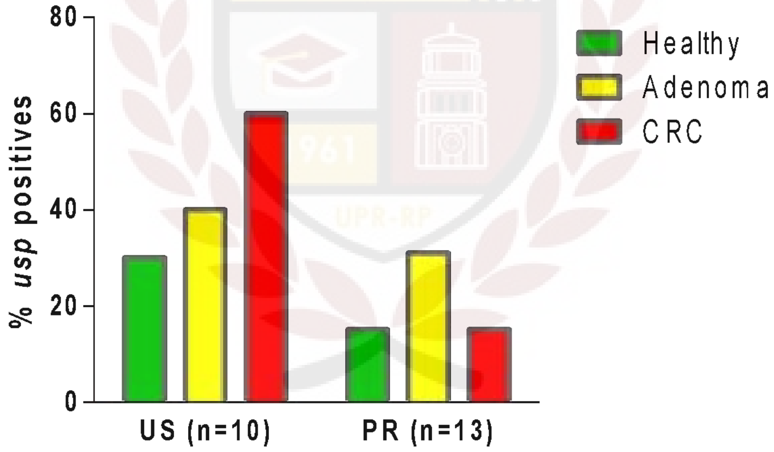


Figure 1.2. *Frequency of the presence of usp gene sequences in US and PR clinical stool samples (Edited figure from Gomez-Moreno, et al., 2019).* This graph was obtained from a previous study that investigated the presence of gut microbial genes that encode bacterial genotoxins or pro-inflammatory factors, including the *usp* gene sequence. The data for the US patients (N = 10 per condition) was obtained from the Early Detection Research Network (EDRN) and that of the PR sample, from the University of Puerto Rico Comprehensive Cancer Center (UPR CCC).

Many experiments have also been conducted to find the prevalence of *usp* gene sequences in inflammatory conditions such as Inflammatory Bowel Disease (IBD). The article "The Presence of Genotoxic and Pro-Inflammatory Bacterial Genes in Gut Metagenomic Databases and Their Possible Link With Inflammatory Bowel Diseases" (Roche-Lima, et al., 2018) focuses on studying how the presence of bacterial genes that are genotoxic and pro-inflammatory affect gut microbiota and various inflammatory diseases such as Crohn's Disease and ulcerative colitis. The most relevant result of this experiment was that the murB Enterobacteriaceae and Enterococci gene sequences were found more frequently in IBD cohorts than in the healthy cohort. Surprisingly, the frequency of *usp* sequences was sharply increased in these IBD conditions, particularly in Ulcerative colitis, while *gelE* sequences (another bacterial toxin from *E. coli*) were found more frequently in Crohn's disease, suggesting "a significant association between the presence of some of these genotoxic or pro-inflammatory gene sequences and IBDs" (Roche-Lima, et al., 2018, p. 2). In the article "Characterization of *Escherichia coli* Isolated from Gut Biopsies of Newly Diagnosed Patients with Inflammatory Bowel Disease" (Sepehri, et al., 2011), which focused on how mucosa-associated strains of *E. coli* from intestinal biopsies may play a role in the pathogenesis of inflammatory bowel diseases (IBDs), *usp* sequences were also found in Ulcerative colitis patients. However, no association was found between the isolated strains of *E. coli* and inflammation in IBD tissues. On the other hand, it was found that "*E. coli* isolated from IBD patients were significantly more invasive than *E. coli* isolated from HC (Healthy Cohort)" (Sepehri, et al., 2011, p. 1456).

In terms of the structure composition, even though Usp was originally described as a 346 amino acid sequence with principal domains, genotoxicity studies have found that Usp is composed of three different domains that add up to a 593 amino acid protein (**Figure 1.3**). An analysis of the amino acid sequence conservation in Usp showed that it contained (1) an N-terminal Hemolysin-coregulated protein (Hcp) domain, (2) an extended C-terminal S-type pyocin domain, and (3) a colicin nuclease domain (Rihtar, et al., 2020). Of these three, the nuclease domain is best characterized, which is one of the reasons Usp has been found to have nuclease activity. Despite this, the role of the N-terminal Hcp domain is still unclear and will thus be this studies' focus. Hcp is a protein family comprised of different genetic sequences that can be found integrated into other proteins or by itself. Due to their hexameric structure seen in **Figure 1.4**, Hcps, along with Valine glycine repeat protein G (VgrG), have been found to play an essential role in the Type 6 Secretion System (T6SS) (Ma, et al., 2017). There are also isoforms of Hcp that vary in their functions; for example, Hcps in some organisms has been related to bacterial competition and colonization, and it could thus contribute to the pathogenicity of specific bacterial strains, such as in *E. coli* (Ma et al., 2017).

Figure 1.3

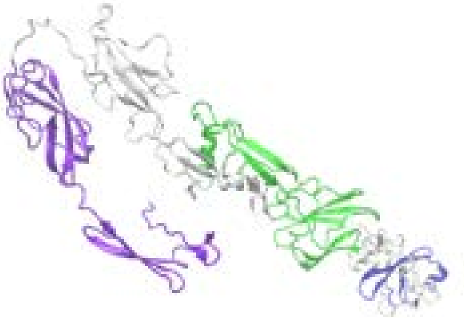


Figure 1.3. *Homology-based structure model of Usp.* This figure demonstrates the three domains that make up the Usp protein: the N-terminal Hcp domain (purple), the extended C-terminal S-type pyocin domain (green), and the colicin nuclease domain (blue). Constructed using homology-based structure prediction (I-TASSER, 1996 & VMD, 2015).

Figure 1.4

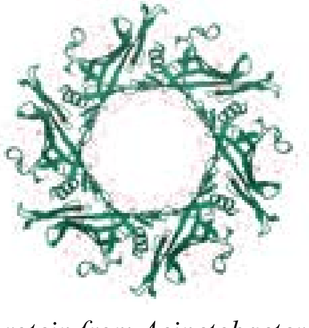


Figure 1.4. *Crystal structure of Hcp-1 protein from Acinetobacter baumannii AB0057 (Ruiz, et al., 2015).* The figure depicts the hexameric ring structure that Hcp proteins found in species harboring T6SS possess.

The research paper "Haemolysin Coregulated Protein is an exported receptor and chaperone of Type VI Secretion substrates" (Silverman, et al., 2013) investigates and analyzes the function of Hcp. They reported that Hcp is both a chaperone and a receptor protein of the T6SS. By focusing on the Tse2 protein, it was found that "a direct and highly specific interaction with the pore of its cognate Hcp, Hcp1 of the H1-T6SS, is required for the stability and export of the protein" (Silverman, et al., 2013, p. 585). The study shows that effectors with unrelated sequences, such as Tse1 and Tse2, also require direct interactions with the Hcp1 pore for secretion.

The article "Three Hcp homologs with divergent extended loop exhibit regions different functions in avian pathogenic *Escherichia coli"* (Ma, et al., 2018), focuses on T6SS's contribution to the pathogenicity of the avian pathogenic *Escherichia coli* (APEC), one of the causal agents of sepsis and meningitis in birds. The Hcp protein was found to play a crucial role in the system as "the variant region Vs2 (Loop L2, 3) in Hcp1 and Hcp2B was essential for the delivery of antibacterial effectors and the inhibition of macrophage phagocytosis" (Ma, et al., 2018, p. 1). Similarly, the researchers used a duck-specific Hcp that indicated that these Hcp proteins had different functions in APEC's pathogenic process and in immunoprotection. The research paper "Roles of Hcp family proteins in the pathogenesis of the porcine extraintestinal pathogenic *Escherichia coli* type VI secretion system" (Peng, et al., 2016) explored the role of Hcp family proteins, specifically in the pathogenesis of the T6SS in the extraintestinal porcine pathogenic *E. coli*, or ExPEC. Research also showed that the Hcp protein family was involved in bacterial competition and contributed to organ colonization in the PCN033 strain of the porcine ExPEC but was not involved in bacterial adhesion. Interestingly, it was also discovered that the three Hcp proteins found, Hcp1, Hcp2, and Hcp3, had different functions. The article mentions how "Hcp2 functioned predominantly in bacterial competition; all three proteins were involved in the colonization of mice; and Hcp1 and Hcp3 were predominantly contributed to bacterial-eukaryotic cell interactions" (Peng, et al., 2016, p. 1).

In turn, the review, "Effector–Immunity Pairs Provide the T6SS Nanomachine its Offensive and Defensive Capabilities" (Yang, et al., 2018), describes the cylindrical structure of T6SS, which occurs due to the hexameric structure of Hcp proteins. The reading explains how "the antibacterial functions of T6SSs are attributable to secreted antibacterial effector toxins, which are neutralized in secretor strains by corresponding antagonistic immunity proteins that prevent self-killing or sibling-intoxication" (Yang, et al., 2018, p. 2). The review concludes that E-I pairs constitute a new module of toxins-antitoxins, thus protecting sister cells against effector toxins. Additionally, the review "Type VI Secretion System in Pathogenic *Escherichia coli*: Structure, Role in Virulence, and Acquisition" (Navarro, et al., 2019) studied the T6SS's structure, its role in virulence, and the acquisition of this system in pathogenic *E. coli.* The authors mention that the central role of T6SS is in bacterial competition, since it can kill neighboring bacteria that are not immune. However, some T6SS have been directly associated with pathogenesis. The article mentions how pathogenic *E. coli* strains “have acquired a variety of sophisticated protein exporting nanomachines to secrete an arsenal of virulence factors implicated in their virulence. Both secretion machineries and effector proteins have been acquired by horizontal gene transfer through several kinds of mobile genetic elements" (Navarro, et al., 2019, p. 13).

Since little is known about the role of Hcp as an N-terminal domain of effectors proteins, in this study we examined the relationship between the presence of DNA sequences encoding Hemolysin-coregulated protein (Hcp) in publicly available human microbiome datasets, and the incidence of adenoma and colorectal cancer. To do so, we employed the Basic Local Alignment Search Tool (BLAST) algorithm to count the hit frequencies of the *hcp* DNA sequences, along with other genes of interest related with T6SS, such as *vgrG1* and *tssk*, among the metagenomic datasets of healthy, small adenoma, large adenoma, and CRC patients. These results were quantified and analyzed. Our research focused on searching for the following gene sequences: the Hcp domain embedded within Usp from *E. coli* strains (Hcp) as well as, *Hcp-1, Hcp-2,* and *Hcp-3,* which are Hcp isoforms that are found individually in the PCN033 *E. coli* strain.

Therefore, our main research question asked if the *hcp* gene sequences were more frequently found in the patients from Adenoma and CRC cohorts than in datasets from healthy controls? Beyond answering this direct question concerning the involvement of *hcp* in CRC and Adenoma, a reusable automated computational method was begun to facilitate the analysis of multiple bacterial genes in the etiology of CRC. The proposed computational platform may, in the future, fulfill the Human Microbiome Project's primary goal: to simplify the characterization of the microbiota found in the human body to obtain greater knowledge on the impact that these communities and their products secretion can have on human health. Therefore, this experiment aims to quickly obtain information that could be useful for the medical field and society, since the topic has not been explored yet. Thus, after analyzing the experimental data, our goal is to lay the foundation that serves as a starting point from which one can understand and study microbial involvement in these diseases.

To the research problem, our hypothesis states that we will find a higher number of Hcp DNA sequences in the datasets obtained shotgun metagenomics of stool samples from adenoma and CRC patients than in those from healthy individuals. This positive association that is hypothesized stems from the findings of previous experiments that explored Hcp and Usp proteins and their roles in colorectal health. Studies have found that bacterial genotoxic and/ or pro-inflammatory factors in the digestive tract can contribute to different diseases. For example, genetic *usp* polymorphisms associated with CRC have been found. Given these discoveries, the efforts of our research group will result in the creation of a pathway designed to investigate the prevalence of the Hcp domain and its genetic variants, along with other genes of interest, like *tssk*, *vgrG1*, and the *usp* sequence without the Hcp domain, in the databases of patients with adenoma or colorectal cancer and healthy individuals.

Chapter 2. Quantify and evaluate the presence of the Hcp and related sequences in the metagenomic datasets from the healthy, adenoma, and CRC cohorts.

2.1 Introduction

Previous studies have found that some genotoxic and pro-inflammatory factors of bacteria in the digestive tract contribute to the development of various diseases, including colorectal cancer (CRC) (Gómez-Moreno, 2019). One of these pro-inflammatory factors made by bacteria is the uropathogenic specific protein or Usp, a multidomain protein originally found in *E. coli* from urinary tract infections, but later found in gut-specific strains. In previous work, our research group showed that the *usp* gene was more frequently found in stool samples from patients with adenomas and CRC compared to healthy individuals (Gómez-Moreno, 2019). Moreover, genetic *usp* polymorphisms associated with CRC have also been found in stool through the DNA sequencing of patient samples.

The N-terminus of Usp is occupied by a domain with homology to the hemolysin-correlated protein (Hcp), a 247 amino acid protein that has been found both as a stand-alone protein and as an embedded domain within a multi-domain protein. The stand-alone Hcp has been found to serve as an essential hallmark of the Type VI Secretion System (T6SS) and seems to be involved in bacterial competition and organ colonization in the PCN033 strain of the porcine ExPEC (Peng, et al., 2016). *E. coli* can harbor some or all of three *hcp* genes: *hcp-1*, *hcp-2*, and *hcp-3*, three independent stand-alone Hcp that are associated with T6SS. These individual sequences were originally identified in the *E. coli* PCN033 bacterial strain and varied in function and in sequence length: *hcp-1* was 161 amino acids long, *hcp-2*, 173 amino acids, and *hcp-3*, with 383 amino acids, was the longest protein. Specifically, Hcp-2 seemed to function in bacterial competition, while Hcp1 and Hcp3 were associated to bacterial-eukaryotic cell interactions (Peng, et al., 2016).

Another protein found in the T6SS is vgrG-1 a key member of the spike that enables the system to penetrate and dispense the effectors into the target cell. This *vgrg* sequence, along with the *hcp* sequences, is thought of as an indicator of the presence of T6SS, which has been associated to infection and inflammation (Hersch, et al., 2020). Another gene of interest in *tssK*, a structural component of the T6SS that is another indicator of this system in the individuals studied. Specifically, this 449 amino acid protein sequence from the *Escherichia coli* O2:H6 strain is a baseplate component of the T6SS and is the factor responsible for mediating the baseplate's docking to the membrane of the target cell (Navarro, et al., 2019).

In this chapter, we verified the frequency of ten genes to establish their presence among the three patient cohorts: cancer, adenomas, and healthy (Table S1); seven possible components of T6SS and three control genes. The T6SS genes of interest included (1) *usp* without the Hcp domain (Usp), (2) the *hcp* portion of the *usp* gene (Hcp), (3) *hcp1*, (4) *hcp2*, (5) *hcp3*, (6) *vgrg-1* and (7) *tssk*. The *ECmurB* from *E. coli* and *EFmurB* from *Enterococcus faecalis* gene sequences were used as controls, since both are housekeeping genes and are thus expected to be present similarly in all three cohorts. Secondly, the *AMmurB* gene sequence from *Akkermansia muciniphila* was included since it has been previously linked to CRC and other inflammatory conditions (Weir, et al., 2013).

2.2 Methods

2.2.1 Study site and metagenomic data obtainment

The research was carried out at the Medical Sciences Campus, specifically at Dr. Abiel Roche-Lima's laboratory, from August 2020 until November 2021. This location provided appropriate access to the genomic sequences necessary to carry out the analysis. The whole metagenomic DNA sequence data was obtained from the European Nucleotide Archive (https://www.ebi.ac.uk/ena) database (Accession No. PRJEB12449). The selected data was originally obtained as part of a case-control study conducted in Washington DC (Vogtmann, et al., 2016). This dataset has been used in the past to identify bacterial genes associated with CRC, but it also includes individuals diagnosed with small adenoma and large adenoma. These databases also included sequences from healthy individuals who were not diagnosed with any of the diseases mentioned before.

2.2.2 Genes of interest

The DNA sequences for the genes of interest, which include *hcp1*, *hcp2*, and *hcp3*, along with *vgrG1*, the Hcp domain from Usp (*hcp*), and *tssk*, were acquired from the PCN033 and *E. coli* O2:H6 strains (Table S1). Specifically, we used the PCN033 strain for the *hcp1-3* sequences and for *vgrG1*. The *E. coli* O2:H6 strain was used for the Hcp domain from Usp (*hcp*) and *tssk*. Additional genes, such as *clbB*, *gelE*, and *tcpC* were also analyzed.

2.2.3 BLAST analysis

The sequences were all used as queries in independent BLAST searches among the healthy (N = 61), adenomas (N = 42: small adenoma N = 27; large adenoma N = 15), and CRC (N = 53) patient-derived sequences. The number of times that these specific sequences aligned with those of the patients from each cohort was quantified to make an assessment on whether a statistically higher number of "hits" is observed in any of the disease groups. It should be noted that patients' personal information always remained anonymous, and individuals voluntarily shared their genomic information during the collection of DNA samples.

To carry out this project, the gene sequences of the patients from each of the four cohorts were downloaded and decompressed in Fastq format. The sequences were then filtered and separated based on the specific disease, that is, small adenoma, large adenoma, cancer, and healthy. These files were then merged and converted to Fasta format for easier processing in the proceeding steps. Once the sequences are ready, our metagenomic analysis pipeline, which mimicked the NCBI-BLAST, or "Basic Local Alignment Search Tool" database, was executed using the Linux command line. This useful tool used to study specific gene sequences was employed to analyze the genes of interest against the patient databases with the WGS sequences, created for each of the four groups. This process was carried out two times since all patients were evaluated twice to produce two replica groups to ensure validity.

2.2.3 RStudio analysis

After running the BLAST algorithm, the gene sequence alignment results for each set of replicas within each cohort was evaluated using the RStudio program. Firstly, all patients who did not meet the definition of a "hit" were filtered out. That is, only the patients who presented at least one hit in each of the two replicas for a specific gene were considered and studied. Afterwards, the hit results from the two replica groups of these individuals were merged and averaged. After carrying out this data retrieval three times, the sequence hit results obtained were organized and analyzed through RStudio's statistical programming language. For this next stage, given how the sample sizes of the small adenoma (N = 27) and large adenoma (N = 15) are individually inferior to those of the healthy (N = 61) and CRC (N = 53) cohorts, we analyzed the two adenoma groups as one, where N = 42.

Having grouped the adenomas and once all gene sequence hits were quantified, the frequency percentage of the each gene within each cohort was calculated using formula 1, where "Positives" equals the number of patients that were positive for the specific gene within each individual cohort and "N" is the total number of individuals in each cohort.

$$Frequency\,\% = \left(\frac{Positives}{N}\right) x\ 100 \qquad (1)$$

The frequency results obtained were included in a table, along with the total number of hits and the number of positive patients per gene per cohort.

Using the RStudio program, we conducted Fisher's Exact Test to evaluate the statistical significance between the number of positive patients found in each gene for the adenomas and CRC groups against those found in the healthy cohort. Therefore, by obtaining these specific p-values, we assessed if there was a stronger correlation between the genes of interest and a particular group. Specifically, all $P \leq 0.05$ were considered significant. These P-values were also included in the previously mentioned table, alongside the frequency values.

To verify how the hits for each gene were distributed among the patients, we used the RStudio program to quantify the number of hits each patient had per gene. Once this information was obtained, we created a table where the average number of hits for each gene between the positive patients, as well as the minimum and maximum values, median, mode, and the standard deviation for these values in each cohort, is presented.

2.2.4 Data visualization: Tables and Figures

The tabulated data was used to generate figures using the GraphPad Prism 6 program. Specifically, two clustered column graphs were generated: one which compares the total gene hit results for each gene in each cohort and one that displays the frequency percentages per gene per cohort.

2.2.5 Hcp-Usp gene Hits Alignment

To visualize the distribution of the Usp and Hcp domain sequences obtained in the study, using the NCBI web page, six NCBI BLAST Alignment Graphs were generated. Specifically, these hit sequences were aligned against the complete Usp sequence, that is, the Usp protein with its three domains. Thus, for each cohort, two graphs were generated: one for the hit sequences for the Hcp domain (Hcp) and another for the hits obtained for Usp without the Hcp domain (Usp).

2.3 Results

The genes of interest were aligned with the metagenomic data from each cohort, that is, Healthy, adenoma, and CRC groups. The sequence hit results, alongside their respective frequencies and the number of patients positive for a specific hit, were tabulated. The data for the small and large adenoma patients was grouped to obtain a more similar group number (N) to that of the cancer and healthy cohorts. Our general results showed that VgrG1 is present at a significantly increased frequency (43%; P-value = 0.05) in the adenomas group in comparison to the healthy cohort (23%) (Table 2.1). However, this increase in frequency was not seen in the cancer group (21%). Surprisingly, the Hcp domain and Usp sequences were found less frequently in the adenomas group, so much so that the Usp sequence results for this group was close to showing a significant downward difference (0.07) when compared to the healthy individuals. In comparison, when evaluating the large and small adenoma groups individually (Table S2), no statistically significant differences were obtained, including those pertaining to the Usp sequence (P-value = 0.10 and 0.44 for small and large adenoma, respectively).

After using the RStudio program to find the specific amount of hits each patient had for each gene in their specific cohort (Table S3), the average number of hits between all the patients of a cohort for each specific gene was calculated. Furthermore, the standard deviation, median, mode, minimum, and maximum values between these patients were also calculated to better observe how the hit results were distributed (Tables 2.1, A-J). In many cases, when comparing the average hit values with the minimum, median, and maximum values for each gene in each cohort, it was noticeable that there were some outliers present that increased the average value obtained. Additional analysis was performed without considering these outliers (Table S4), but the general results did not change.

Table 2.1. Presence and frequency of Hcp isoforms and T6SS- related genes in the selected metagenomic databases cohorts, i.e., Healthy, Adenomas, and CRC

A. AMmurB Gene

Cohort	Total Hits	Mean Hits ± σ	Positives	Frequency % (P-value)	Minimum	Median	Maximum	Mode
Healthy	168	8 ± 10	21	34	1	5	45	1
Adenomas	132	8 ± 7	16	38 (0.83)	1	6	28	7
Cancer	367	15 ± 26	24	45 (0.26)	1	5	123	2

B. ECmurB Gene

Cohort	Total Hits	Mean Hits ± σ	Positives	Frequency % (P-value)	Minimum	Median	Maximum	Mode
Healthy	243	12 ± 16	21	34	1	5	69	1
Adenomas	90	6 ± 8	14	33 (1.00)	1	4	33	5
Cancer	386	18 ± 29	21	40 (0.70)	2	7	102	4

C. EFmurB Gene

Cohort	Total Hits	Mean Hits ± σ	Positives	Frequency % (P-value)	Minimum	Median	Maximum	Mode

Healthy	22	3 ± 3	7	11	1	3	10	1
Adenomas	21	5 ± 4	4	10 (1.00)	2	4	12	-
Cancer	17	4 ± 3	4	8 (0.54)	1	4	8	-

D. Hcp Gene

Cohort	Total Hits	Mean Hits ± σ	Positives	Frequency % (P-value)	Minimum	Median	Maximum	Mode
Healthy	66	7 ± 8	9	15	2	4	23	2
Adenomas	35	9 ± 11	4	10 (0.55)	2	4	26	4
Cancer	46	7 ± 5	7	13 (1.00)	2	4	13	3

E. Hcp-1 Gene

Cohort	Total Hits	Mean Hits ± σ	Positives	Frequency % (P-value)	Minimum	Median	Maximum	Mode
Healthy	59	8 ± 13	7	11	1	2	36	2
Adenomas	8	4 ± 3	2	5 (0.30)	2	4	6	-
Cancer	90	11 ± 13	8	15 (0.59)	1	6	39	3

F. Hcp-2 Gene

Cohort	Total Hits	Mean Hits ± σ	Positives	Frequency % (P-value)	Minimum	Median	Maximum	Mode
Healthy	38	8 ± 11	5	8	2	3	27	-
Adenomas	16	3 ± 2	5	12 (0.74)	1	3	6	-
Cancer	93	16 ± 18	6	11 (0.75)	3	5	41	3

G. Hcp-3 Gene

Cohort	Total Hits	Mean Hits ± σ	Positives	Frequency % (P-value)	Minimum	Median	Maximum	Mode

Healthy	80	6 ± 5	14	23	1	3	17	3
Adenomas	8	3 ± 1	3	7 (0.06)	2	3	4	-
Cancer	36	4 ± 3	9	17 (0.49)	1	4	9	2

H. tssK Gene

Cohort	**Total Hits**	**Mean Hits ± σ**	**Positives**	**Frequency % (P-value)**	**Minimum**	**Median**	**Maximum**	**Mode**
Healthy	29	3 ± 1	11	18	1	3	5	2
Adenomas	17	6 ± 4	3	7 (0.15)	1	7	9	-
Cancer	138	14 ± 26	10	19 (1.00)	2	5	86	2

I. Usp Gene

Cohort	**Total Hits**	**Mean Hits ± σ**	**Positives**	**Frequency % (P-value)**	**Minimum**	**Median**	**Maximum**	**Mode**
Healthy	88	8 ± 8	11	18	2	5	24	3
Adenomas	36	18 ± 20	2	5 (0.07)	4	18	32	-
Cancer	68	6 ± 5	11	21 (0.81)	2	4	18	4

J. VgrG1 Gene

Cohort	**Total Hits**	**Mean Hits ± σ**	**Positives**	**Frequency % (P-value)**	**Minimum**	**Median**	**Maximum**	**Mode**
Healthy	274	20 ± 40	14	23	1	7	152	1
Adenomas	154	9 ± 10	18	43 (0.05)	1	3	30	3
Cancer	489	33 ± 62	15	28 (0.53)	1	6	224	1

Table 2.1. The table is divided into the three cohorts (healthy, adenomas and cancer). The columns present the total hits, average number of hits with the standard deviation (σ), positive patients, frequency

percentage with the corresponding P-value*, minimum, median, maximum value, and mode found for each cohort. The adenomas cohort results are obtained by grouping the small adenoma and large adenoma cohort results. *P-values obtained using Fisher's Exact Test when comparing the Adenomas and Cancer positive patients' frequencies to the ones from the Healthy cohort. P-values are statistically significant when $P \leq 0.05$. Statistically significant values are placed in bold.

Using the values from each "Total Hits" column from Tables 2.1 A-J, the gene hit results for each cohort were graphed to observe their distribution among each diagnostic group (Figure 2.1). The cancer cohort displayed a higher gene hit total for most genes of interest (AMmurB, ECmurB, Hcp-1, Hcp-2, tssk, and vgrG1), while the adenomas cohort always had the lowest total in each category.

Figure 2.1

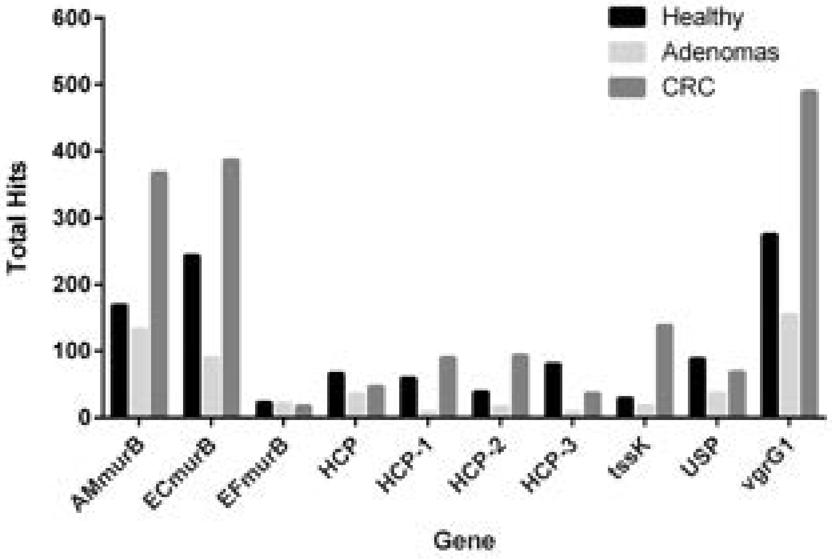


Figure 2.1. *Hits per gene of interest among positive patients from Healthy, Adenoma and CRC metagenomic databases*. In the preceding graph, the y-axis corresponds to the number of total hits found for each gene in the three separate cohorts and the x-axis presents each gene in question. As presented in the legend, the blue line is representative of the hits obtained by the healthy controls, the green line corresponds to those of the adenomas group, and the grey line pertains to the hits from the cancer patients.

Furthermore, utilizing the frequency percentage values presented in Table 1.1, a clustered column graph that plotted the frequency values obtained for each gene in each cohort was generated (Figure 2.2). Like in Figure 2.1, the cancer cohort usually possessed the highest frequency values (AMmurB, ECMurB, Hcp-1, tssk, and Usp). Contrary to Figure 2.1, the adenomas group showed the highest frequency value for the Hcp-2 and vgrG1 genes.

Figure 2.2

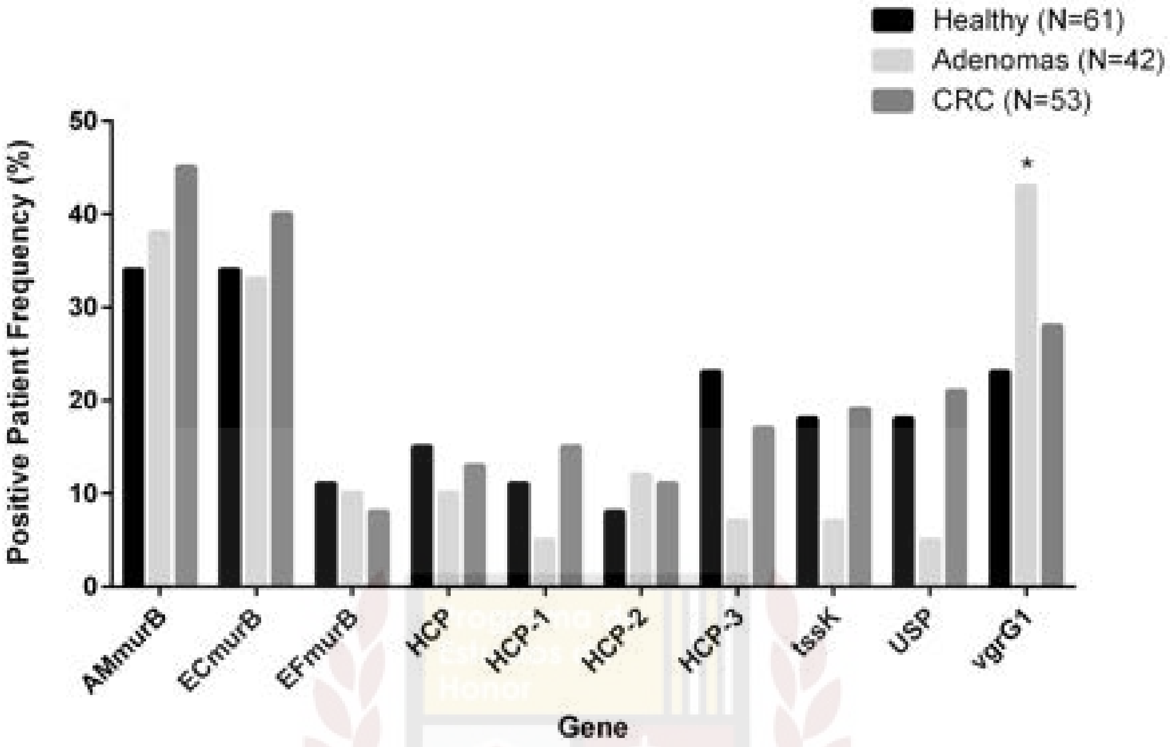
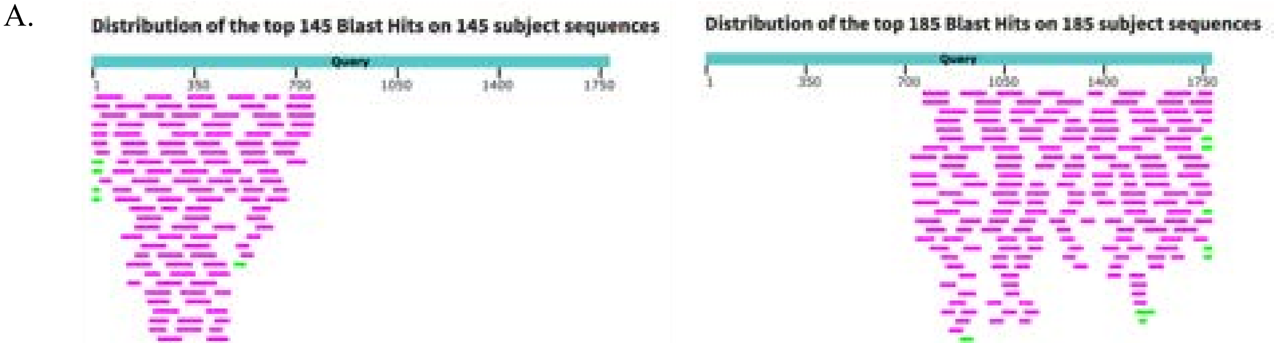


Figure 2.2. *Frequency of positive patients among genes of interest from Healthy, Adenoma and CRC metagenomic databases*. In this second graph, the y-axis presents the frequency percentage values calculated by dividing the number of positive patients per gene per cohort by each cohort's total number of patients (N). The y-axis presents each gene in question. Once more, the blue bars correspond to of the frequency percentages from the healthy cohort, the green bars represent the percentages of the adenomas cohort, and the grey bars depict those from the CRC cohort.

As an additional step in the study, the hits obtained for the Hcp domain from Usp and the Usp sequence without the Hcp domain used in the study were individually aligned against the complete Usp sequence using the NCBI BLAST website. Six alignment graphs were generated: one for the Hcp domain sequence and one for the Hcp-less Usp sequence within each of the three cohorts (Figure 2.3). When comparing the six graphs, it is evident that each sequence had conserved regions within the complete Usp sequence. Additionally, the graphs included a legend corresponding to the alignment scores of each sequence hit.

Figure 2.3



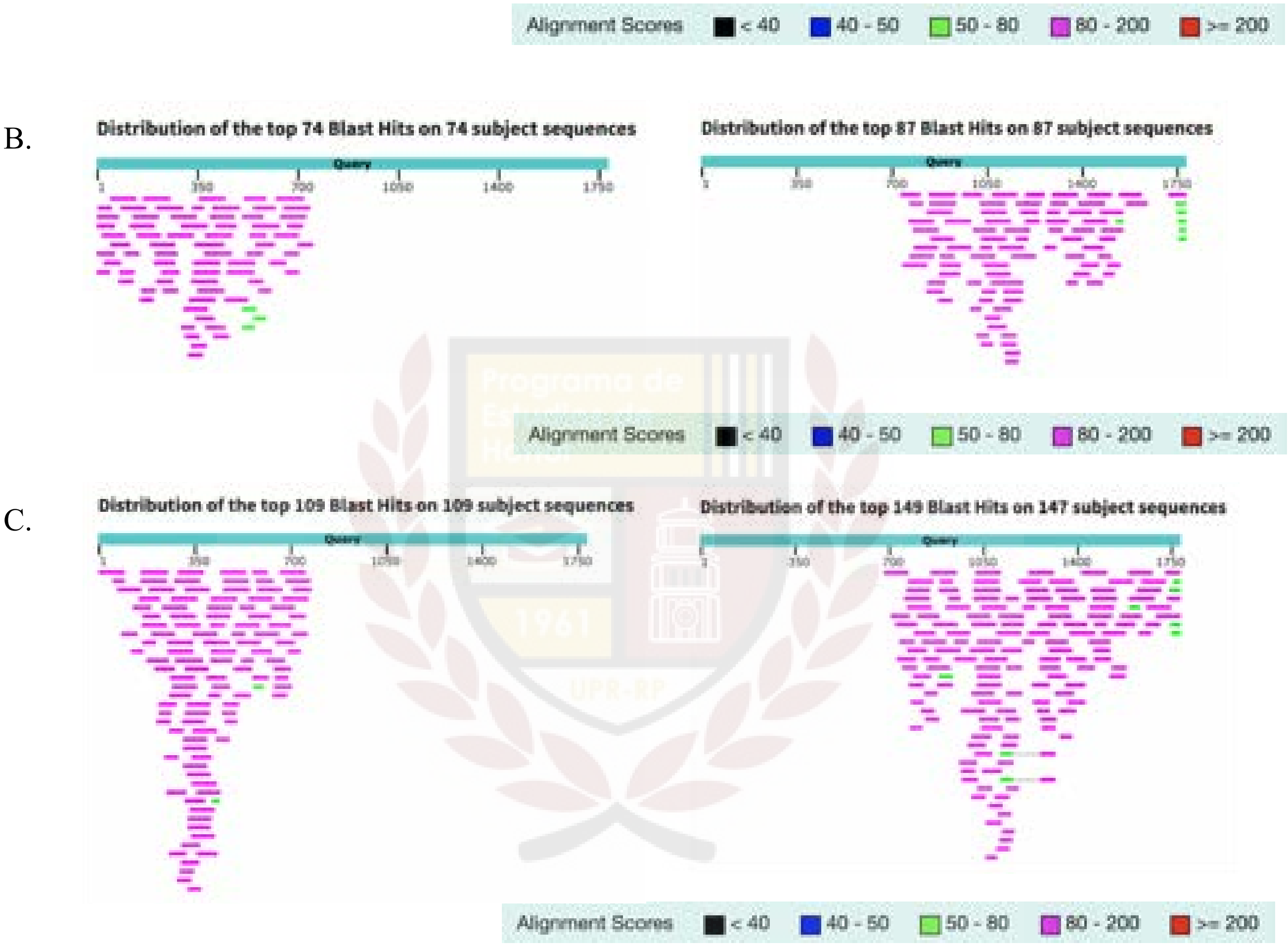


Figure 2.3. *NCBI BLAST alignment graphs for Hcp and Usp sequences in the Healthy (A), Adenomas (B), and CRC (C) cohorts*. In these six figures, the Hcp and Usp hit sequences obtained from the project (Hcp=145 & Usp= 185 for Healthy; Hcp=74 & Usp=87 for Adenomas; Hcp=109 & Usp= 149 for CRC) were blasted against the complete Usp sequence to observe how these were arranged.

2.4 Discussion

In recent years, much research has been conducted on the involvement of the gut microbiota in diseases. This line of work has been expanded by the growth and reach of deep sequencing or next-generation sequencing platforms, which has generated mounds of useful data. While many efforts have focused on the generation of sequence data, fewer efforts have attempted to derive plausible hypotheses or mechanisms that explain how the microbiota modulates host tissue to result in disease. In this project, we used the metagenomic data provided by the European Nucleotide Archive (https://www.ebi.ac.uk/ena) database (Accession No. PRJEB12449) to study the frequency of several genes in three different cohorts (CRC, adenomas, and healthy controls). To study these, we made use of the Basic Alignment Search Tool (BLAST) and RStudio programs, which were essential in this study and serve as clear examples of how these tools contribute to the in-depth investigation of metagenomic sequences in disease-specific cohorts. The procedure we utilized to obtain, filter, and analyze the data, which included a constant verification of the procedure, ensured validity in the results. All the metagenomic data from each of the cohorts was processed and analyzed using a computational pathway created using the BLAST and RStudio platforms.

While there are many *E. coli* genes that could have been included, we focused on a specific subset of gene sequences (Supplementary Table S1). Most of these were previously found to possess genotoxic or pro-inflammatory behavior (Gomez-Moreno, et al., 2014; Roche-Lima, et al., 2018) but we also included new sequences, such as *vgrG1* and the different *hcp* sequences. The *ECmurB* from *Escherichia coli* and *EFmurB* from *Enterococcus faecalis* gene sequences were included as controls, since both are housekeeping genes. The *AMmurB* gene sequence from *Akkermansia muciniphila* was also included since it has been previously linked to CRC and other inflammatory conditions (Weir, et al., 2013). As expected, in Table 1, the AMmurB gene had most total hits and greatest frequency in the cancer cohort, but there was no statistical significance found (P-value = 0.26). The frequency in the healthy and adenomas cohorts for this gene was very similar. This difference explains why the resulting P-value for the CRC cohort was more significant than that of the adenomas cohort, though neither value was statistically significant.

In terms of the three individual isoforms of Hcp from the PCN033 strain, that is, *hcp-1*, *hcp-2*, and *hcp-3*, the healthy and CRC cohorts presented similar frequencies for each individual sequence, though there was not a uniform difference between the two groups. That is, for *hcp-3*, the healthy cohort had a greater total amount of hits and a higher frequency while, for *hcp-1* and *hcp-2*, the opposite occurred. In terms of the adenomas cohort, the results showed a much lower number of hits for all the hcp sequences when compared to the other two cohorts but showed a slightly higher frequency for the *hcp-2* gene when compared to the CRC cohort. These tendencies can possibly be explained by examining the known functions of each Hcp. For example, though all three individual Hcps aid in colonization, the Hcp-1 effector induces apoptosis and the release of cytokine effectors into the extracellular environment (Zhou, et al., 2012), the Hcp-2 effector is involved in inter-bacterial competition, and Hcp-3 primarily functions in the interaction with eukaryotic host cells (Peng, et al., 2016). Given how apoptosis is intrinsically related to cancer it is used to restraint the out-of-control cell division in these mutated cells (Wong, 2011), this could be a possible explanation as to why Hcp-1 has more gene hits and a higher frequency in the CRC cohort. Also, the higher frequency of the Hcp-2 gene in both the adenomas and CRC patients could be explained by the fact that both conditions have been directly associated to different bacterial factors which contribute to the inflammation. This behavior would explain a higher amount of bacterial competition and, thus, more Hcp-2.

When studying the results for the hcp sequence embedded in Usp, we focused on how these matched up with those of the C-terminal portion of Usp, that is, Usp without its Hcp domain. We expected that, since these separate sequences were supposed to add up to one functional protein, we would have obtained matching results across samples. However, in the healthy and CRC cohorts, we observed a greater amount of hits and a higher frequency for Usp when compared to Hcp. These results are interesting since, given the fact that the hcp domain is shared between different effector proteins (Ma, et al, 2017), one would expect a higher presence of Hcp, which is evidently not the case in the healthy and CRC groups. When we visualized the individual hits of both the Hcp domain and the C-terminal Usp sequence, we observed that the hits were distributed evenly throughout the complete HCP-Usp gene sequence (Figure 2.3). Thus, the difference in the number of hits at the Hcp (N-terminal portion) and Usp (C-terminal portion) could be explained by the fact that the Usp sequence is longer than that of the Hcp domain. That is, the S-pyocin domain and the Nuclease domain in Usp's C-terminal is 346 amino acids long, while Usp's N-terminal Hcp domain has 247 amino acids. Thus, since there is a higher probability for hits in the Usp sequence, it is possible that it is being detected more easily than the Hcp portion. However, the adenomas cohort presented the opposite case since the Hcp domain sequence registered higher frequency values than those obtained for *Usp*. Thus, this could suggest that the environment present in adenoma patients could allow the presence of different types of effector proteins, all of which would contain a similar Hcp domain, in response to the inflammation and bacterial competition levels found in this condition.

Among all the genes studied, the only one whose difference showed a statistically significant P-value, was the vgrG1 sequence in the adenomas group (P-value = 0.05 by Fisher's exact test). Just as with the different Hcp sequences, there are numerous *vgrG* sequences that have varying functions. Apart from the known structural functions these proteins carry out in the Type VI Secretion System (T6SS), it has been shown that the VgrG1 protein found in Vibrio Cholerae's T6SS, covalently cross-links actin and results in the formation of toxic actin oligomers that lead to cell rounding (Durand, et al, 2012). VgrG sequences within the T6SS also display a variety of functions. In the T6SS present in *V. cholerae*, it has been found that the VgrG-1 is different than other VgrG proteins because its effector domain has been translocated into the infected eukaryotic cells' cytosol along with an additional actin cross-linking domain (ACD) at the C-terminus (Dutta, et al., 2019). Thus, the VgrG protein family's functional variability could possibly explain the difference in total number of hits and the frequencies in the cohorts. It could also be possible that the inflammatory environments surrounding the host cells can influence the selection for the gene or specific strains around this tissue. That is, it is possible that in the adenomas group, where the inflammation present is not as severe as that of a cancer patient, there is some selective pressure to increase the proportion of VgrG. Given how there is little to no information on this specific subject, it is crucial to study this phenomenon in order to arrive at conclusions that are more certain, thus potentially leading to a better understanding of how the T6SS and its components effect conditions such as the ones studied.

Chapter 3. Develop a computational platform that can facilitate the study of specific gene sequences from gut microbiota in disease-specific datasets.

3.1 Introduction

Analyzing the prevalence or frequency of genotoxic and/ or proinflammatory genes from gut bacteria among the metagenomic databases is currently a very challenging process. Despite the importance of analyzing such large datasets, there are few “user-friendly” tools for parsing the information and generating new hypotheses. Current protocols involve a computational manual and code-intensive workflow that involves downloading large datasets into local servers and running BLAST routines for sequence comparison followed by Rstudio or Python scripts that require a level of programming knowledge above that of your typical biomedical scientist. Clearly there is a need for an automated workflow that can be used to search for specific sequences in disease-specific datasets. In this aim, we offer the stepwise protocol as a sort of “standard operating procedure” for such sequence searches. To the best of our knowledge, there are no computational tools, web-based or local, to easily perform the kind of analysis presented in this document.

As a test case, we have chosen four additional genotoxin-encoding genes; *clbB*, *cnf-1*, *gelE*, and *tcpC*, for analysis using the computational pathway created. These genes were chosen due to their specific functions and roles evidenced in previous studies (Gomez-Moreno, et al., 2014). For example*,* *clbB* is highly associated with CRC as it is a non-ribosomal peptide that is part of the colibactin polyketide synthase multienzyme (pks island), which induces mammalian chromosome instability in mouse and human colon cells in culture and contributes to the production of colibactin (Gomez-Moreno, et al., 2014). Additionally, cytotoxic necrotizing factor-1, or *cnf-1,* is a bacterial toxin that is produced by some pathogenic *E. coli* strains and has been widely associated with cancer proliferation previously since it's an activator of oncogenic Rho GTPases (Gomez-Moreno, et al., 2014). Moreover, *gelE* is a metalloprotease that hydrolyzes insoluble hydrophobic substances and has shown to act as a virulence factor and induces inflammation (Gomez-Moreno, et al., 2014). Finally, *tcpC* is a bacterial toxin that modulates the host's immune response and, therefore, is associated to inflammation. We chose to analyze these genes using the exact number and order of steps taken for the analysis of Usp and Hcp, to validate our protocol as the seed for an eventual computational pathway for the search of any microbial gene in disease-specific cohorts.

3.2 Methods

The same computational steps applied in Chapter 2 (2.2.1-2.2.3) were employed for the additional set of genes, specifically *clbB* (AM229678.1), *cnf-1* (U42629.1), *gelE* (D85393.1), and *tcpC* (GQ902994.1), in order to validate the metagenomic pathway that will seed a computational application or platform for gene analysis. Thus, the procedure conducted for both chapters is presented in the following BLAST and RStudio scripts which, taken together, form the computational procedure for the project.

3.2.1 BLAST analysis

The following set of commands detail the initial steps necessary to download the sequence data and to obtain the unprocessed hit results. The healthy cohort was used as the example presented but all cohort replicas were processed in the following manner. Briefly, using the Linux command line, a script was written to download the sequences which were available in the Fastq format. The files were then converted from the Fastq format to the Fasta format, which presents the sequences in a more processable form. Once the data was cleaned and the replica groups were merged, the number of times that our genes of interest were found within these resulting datasets was quantified using a command line version of BLAST. Below is the stepwise procedure employed:

Step 1: Download fastq files for each cohort using python script

```
{
  import os

  file = open('PRJEB-----.txt', 'r')
  for line in file:
    print line
  os.system('wget -r '+line)
  file.close()
}
```

Step 2: Group sample fastq files into a single fastq file for each replica to create the database

```
cat *.fastq > Controls_PRJEB6070merged-1.fastq

cat *.fastq > Controls_PRJEB6070merged-2.fastq
```

Step 3: Convert files from fastq to fasta format

```
cat Controls_PRJEB6070merged-1.fastq | paste - - - - | sed 's/^@/>/g'| cut
-f1-2 | tr '\t' '\n' > Controls_PRJEB6070merged-1.fasta

cat Controls_PRJEB6070merged-2.fastq | paste - - - - | sed 's/^@/>/g'| cut
-f1-2 | tr '\t' '\n' > Controls_PRJEB6070merged-2.fasta
```

Step 4: Use stand-alone NCBI BLAST software ncbi-blast-2.6.0+ to create a BLAST database for each replica within each cohort

```
makeblastdb -in Controls_PRJEB6070merged-1.fasta -input_type fasta -dbtype
nucl -parse_seqids -out Controls_F-Population-1

makeblastdb -in Controls_PRJEB6070merged-2.fasta -input_type fasta -dbtype
nucl -parse_seqids -out Controls_F-Population-2
```

Step 5: Use stand-alone NCBI BLAST software ncbi-blast-2.6.0+ to run BLAST

```
blastn -db Controls_F-Population-1 -query ProInflammatorygenes.fasta -out
Controls_F-Population-1-blastn-output.csv -outfmt "6 std qcovs qseqid sseqid
slen qstart qend length mismatch gapopen gaps sseq"

blastn -db Controls_F-Population-2 -query ProInflammatorygenes.fasta -out
Controls_F-Population-2-blastn-output.csv -outfmt "6 std qcovs qseqid sseqid
slen qstart qend length mismatch gapopen gaps sseq"
```

3.2.2 RStudio analysis

The following R Markdown Script details the steps taken throughout the generated pathway in order to obtain the study results. The following example pertains to the Healthy cohort, but the same pathway detailed was applied to all cohorts.

Step 6: Choose Working Directory

The first step is to indicate the location of the data you will be utilizing using the “setwd” function.

```
setwd("~/Documents/GabrielasProject2020/CNR_&_Healthy_outputs_04-23-2021/Outp
uts_PRJEB6070")
```

Step 7: Call the data.table packages required

The following packages were used to generate the project's results.

```
if(!require(data.table)){
  install.packages("data.table")
  library(data.table)
}
```

```
if(!require(stringi)){
  install.packages("stringi")
  library(stringi)
}

if(!require(dplyr)){
  install.packages("dplyr")
  library(dplyr)
}
```

Step 8: Open the tsv files of interest

The two .tsv files, one of each replica, generated during the BLAST analysis for each cohort were inputted with the “read.table” function and saved within the R script using the arrow symbol. These individual files contained the patients who displayed hits for the genes evaluated.

```
Healthy_blastn_1 <- read.table("Controls_1_MergedPIGenes_1-29-2021.tsv",
header = F, sep = "\t")
Healthy_blastn_2 <- read.table("Controls_2_MergedPIGenes_1-29-2021.tsv",
header = F, sep = "\t")
```

Step 9: Load metadata file

The metadata file, which contains the clinical data for all the patients present in the metagenomic data used, was called for with the “read.csv” function. The column originally called "Run_s" was renamed "run_accession" to facilitate future data processing.

```
metadata <- read.csv("Clinical_SRAPRJEB6070_02-5-2021.csv", header = T, sep =
",")
names(metadata)[names(metadata) == "Run_s"] <- "run_accession"
```

Step 10: Filter metadata for diagnosis of interest

In order to only analyze the healthy individuals, the metadata file was filtered using the "Diagnosis_s" column. Additionally, to remove potential duplicates, only patients whose "LibraryLayout_s" equaled "SINGLE" were used.

```
metadata_Healthy <- metadata[ which(metadata$Diagnosis_s=="Normal"), ]
metadata_Healthy <- metadata_Healthy[
which(metadata_Healthy$LibraryLayout_s=="SINGLE"), ]
```

Step 11: Process the accession numbers

To facilitate the future processing and analysis of the patients' hit information, their run accession numbers presented per hits in each replica groups were simplified so they would only identify the patient instead of the specific hit.

```
Healthy_blastn_1$V2 <- sub('\\..*', '', Healthy_blastn_1$V2)
Healthy_blastn_2$V2 <- sub('\\..*', '', Healthy_blastn_2$V2)
```

Step 12: Convert replica data frames to tables

```
dt_1 <- data.table(Healthy_blastn_1)
dt_2 <- data.table(Healthy_blastn_2)
```

Step 13: Rename data table columns

The columns were renamed to resemble the default nomenclature of the BLAST program. The only exception was the “query coverage” column.

```
# For first replica group
names(dt_1)[names(dt_1)=="V1"] <- "queryseqid"
names(dt_1)[names(dt_1)=="V2"] <- "run_accession"
names(dt_1)[names(dt_1)=="V3"] <- "qident"
names(dt_1)[names(dt_1)=="V4"] <- "length"
names(dt_1)[names(dt_1)=="V5"] <- "mismatch"
names(dt_1)[names(dt_1)=="V6"] <- "gapopen"
names(dt_1)[names(dt_1)=="V7"] <- "qstart"
names(dt_1)[names(dt_1)=="V8"] <- "qend"
names(dt_1)[names(dt_1)=="V9"] <- "sstart"
names(dt_1)[names(dt_1)=="V10"] <- "send"
names(dt_1)[names(dt_1)=="V11"] <- "evalue"
names(dt_1)[names(dt_1)=="V12"] <- "bitscore"
names(dt_1)[names(dt_1)=="V13"] <- "querycover"

# For second replica group
names(dt_2)[names(dt_2)=="V1"] <- "queryseqid"
names(dt_2)[names(dt_2)=="V2"] <- "run_accession"
names(dt_2)[names(dt_2)=="V3"] <- "qident"
names(dt_2)[names(dt_2)=="V4"] <- "length"
names(dt_2)[names(dt_2)=="V5"] <- "mismatch"
names(dt_2)[names(dt_2)=="V6"] <- "gapopen"
names(dt_2)[names(dt_2)=="V7"] <- "qstart"
names(dt_2)[names(dt_2)=="V8"] <- "qend"
names(dt_2)[names(dt_2)=="V9"] <- "sstart"
names(dt_2)[names(dt_2)=="V10"] <- "send"
names(dt_2)[names(dt_2)=="V11"] <- "evalue"
names(dt_2)[names(dt_2)=="V12"] <- "bitscore"
names(dt_2)[names(dt_2)=="V13"] <- "querycover"
```

Step 14: Rename data table rows

The rows for the "queryseqid" column in each replica table were renamed to represent the gene in question and, afterwards, the column name was changed to "Gene_Name".

```
#For first replica group
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="gi|112292700:41762-51382"]
<- "clbB"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="gi|112292700:41762-51382"]
<- "clbN"
```

```
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="gb|U42629.1|ECU42629:858-39
02"] <- "cnf-1"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="gi|1339878|dbj|D85393.1|ENE
GE1E"] <- "gelE"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="murB_E_coli_Baerga_gi|55650
3834:4172057-4173085"] <- "ECmurB"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="murB_E_faecalis"] <-
"EFmurB"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="murB_Akkermansia_muciniphil
a"] <- "AMmurB"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="gi|308387403|gb|GQ902994.1|
"] <- "tcpC"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="gi|7416047|dbj|AB027193.1|"
] <- "USP"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="gi|37547387|gb|AY128544.1|"
] <- "cif"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="CP043181.1:3517124-3518905"
] <- "HCP"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="CP043181.1:3647520-3648866"
] <- "tssK"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="CP006632.1:246100-246582"]
<- "HCP-1"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="CP006632.1:267403-267921"]
<- "HCP-2"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="CP006632.1:4258749-4259897"
] <- "HCP-3"
levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="CP006632.1:268131-269756"]
<- "vgrG1"

levels(dt_1$queryseqid)[levels(dt_1$queryseqid)=="queryseqid"]  <-
"Gene_Name"

# For second replica group
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="gi|112292700:41762-51382"]
<- "clbB"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="gi|112292700:41762-51382"]
<- "clbN"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="gb|U42629.1|ECU42629:858-39
02"] <- "cnf-1"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="gi|1339878|dbj|D85393.1|ENE
GE1E"] <- "gelE"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="murB_E_coli_Baerga_gi|55650
3834:4172057-4173085"] <- "ECmurB"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="murB_E_faecalis"] <-
"EFmurB"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="murB_Akkermansia_muciniphil
a"] <- "AMmurB"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="gi|308387403|gb|GQ902994.1|
"] <- "tcpC"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="gi|7416047|dbj|AB027193.1|"
] <- "USP"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="gi|37547387|gb|AY128544.1|"
] <- "cif"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="CP043181.1:3517124-3518905"
] <- "HCP"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="CP043181.1:3647520-3648866"
] <- "tssK"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="CP006632.1:246100-246582"]
<- "HCP-1"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="CP006632.1:267403-267921"]
<- "HCP-2"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="CP006632.1:4258749-4259897"
] <- "HCP-3"
levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="CP006632.1:268131-269756"]
<- "vgrG1"

levels(dt_2$queryseqid)[levels(dt_2$queryseqid)=="queryseqid"]  <-
"Gene_Name"
```

Step 15: Merge individual replica tables with metadata

Using the "run_accession" column as the common column, each replica table (dt_1 and dt_2) was merged with the metadata table.

```
samplehits_1 = merge(dt_1, metadata, by="run_accession")
samplehits_2 = merge(dt_2, metadata, by="run_accession")
```

Step 16: Number of hits per gene per patient in each replica

After merging the replica tables with the metadata information, the "queryseqid" and "BioSample_s" columns present in the two sample hits tables were used to group and count the number of times each gene was repeated among each of the patients evaluated in each replica. This gave the total number of hits accounted for each gene evaluated in each of patients present in the replica groups. The resulting two tables presented the gene in question, the patient or sample, and the number of hits. Because of this, these three columns were renamed accordingly.

```
# For first replica group
uniqsampleswhits_1 <- summarise(group_by(samplehits_1, queryseqid,
BioSample_s),length(queryseqid))

colnames(uniqsampleswhits_1)[1] <- "Gene"
colnames(uniqsampleswhits_1)[2] <- "Sample"
colnames(uniqsampleswhits_1)[3] <- "Hits"

# For second replica group
```

```
uniqsampleswhits_2 <- summarise(group_by(samplehits_2, queryseqid,
BioSample_s),length(queryseqid))

colnames(uniqsampleswhits_2)[1] <- "Gene"
colnames(uniqsampleswhits_2)[2] <- "Sample"
colnames(uniqsampleswhits_2)[3] <- "Hits"
```

Step 17: Total hits per patient per gene

The two separate tables generated in the previous step were merged using the three common columns. This step ensured that the resulting patients had at least one hit per gene and thus meet our definition of a "hit". An additional column was added where the hit values for each gene for each patient in each replica were averaged. Thus, this column presented the total number of hits that each patient had per gene.

```
# Merging the two replica tables
merged_gene_hits <- as.data.frame(merge(uniqsampleswhits_1,
uniqsampleswhits_2, queryseqid, by.x = c("Sample", "Gene"), by.y =
c("Sample", "Gene"), sort = TRUE))
head(merged_gene_hits)

##          Sample   Gene Hits.x Hits.y
## 1 SAMEA2466887 AMmurB     22     31
## 2 SAMEA2466887 ECmurB      2      3
## 3 SAMEA2466890 AMmurB      3      6
## 4 SAMEA2466891 AMmurB      7     13
## 5 SAMEA2466891 ECmurB     67     71
## 6 SAMEA2466891    HCP      6      2

# Mean hits column
Mean_hits <-data.frame(Sample=merged_gene_hits[,1:4],
Mean_Hits=rowMeans(merged_gene_hits[,c("Hits.x", "Hits.y")], na.rm=TRUE))
colnames(Mean_hits)[3] <- "Hits_replica1"
colnames(Mean_hits)[4] <- "Hits_replica2"
head(Mean_hits)

##   Sample.Sample Sample.Gene Hits_replica1 Hits_replica2 Mean_Hits
## 1  SAMEA2466887      AMmurB            22            31      26.5
## 2  SAMEA2466887      ECmurB             2             3       2.5
## 3  SAMEA2466890      AMmurB             3             6       4.5
## 4  SAMEA2466891      AMmurB             7            13      10.0
## 5  SAMEA2466891      ECmurB            67            71      69.0
## 6  SAMEA2466891         HCP             6             2       4.0
```

Step 18: Total hits per gene result

The "aggregate" function was applied to the "Sample.Gene" column in the Mean_Hits table created in the last step. This function resulted in the total number of hits that each individual gene had in the filtered pool of healthy individuals.

```
TotalHitsPerGene <- aggregate(cbind(Mean_Hits) ~ Sample.Gene, data=Mean_hits,
FUN=sum)
head(TotalHitsPerGene)

##   Sample.Gene Mean_Hits
## 1       HCP-1      59.0
## 2       HCP-2      38.0
## 3       vgrG1     274.0
## 4       HCP-3      79.5
## 5         HCP      66.0
## 6        tssK      29.0
```

Step 19: Total positives per gene result

The "aggregate" function was reapplied to the "Sample.Gene" column in the Mean_Hits table, but instead of using the "sum" function, "function(x){NROW(x)}" was used to count the number of times a specific Sample ID, representative of each patient, was repeated in each gene.

```
Total_Positives <- aggregate(cbind(Mean_Hits) ~ Sample.Gene, data=Mean_hits,
FUN = function(x){NROW(x)})
head(Total_Positives)

##   Sample.Gene Mean_Hits
## 1       HCP-1         7
## 2       HCP-2         5
## 3       vgrG1        14
## 4       HCP-3        14
## 5         HCP         9
## 6        tssK        11
```

Step 20: Merged table with positives frequency value

The TotalHitsPerGene and Total_Positives tables were merged using the "Sample.Gene" column. A fourth column with the frequency percentage value for the positives for each gene was added. These values were obtained by dividing the values from the "positives" column by the total number of patients present in the healthy cohort.

```
# Merging total hits per gene and total positives per gene tables
Positives_with_hits <- merge(TotalHitsPerGene, Total_Positives, Sample.Gene,
by.x = "Sample.Gene", by.y = "Sample.Gene", sort = TRUE)

# Rename columns
colnames(Positives_with_hits)[1] <- "Gene"
colnames(Positives_with_hits)[2] <- "Total Hits"
colnames(Positives_with_hits)[3] <- "Positives"

# Frequency of positives per gene
Positives_with_hits$Frequency <-
```

```
as.numeric(as.character(Positives_with_hits$Positives))/ 61 *100
head(Positives_with_hits)

##     Gene Total Hits Positives Frequency
## 1 AMmurB      168.0        21 34.426230
## 2   clbB        6.0         3  4.918033
## 3  cnf-1      158.0         8 13.114754
## 4 ECmurB      242.5        21 34.426230
## 5 EFmurB       22.0         7 11.475410
## 6   gelE       23.0         4  6.557377
```

Step 21: Fisher Test Example

The following section presents one of the Fisher Tests realized in the study. The following example corresponds to the comparison between the adenomas cohort (A) and the healthy individuals (H). It is important to note that the same procedure was employed for the remaining groups of the study.

```
# Import Healthy versus Adenomas file
Healthy_Adenomas <-
read.csv("/home/aroche/Documents/GabrielasProject2020/Data
Visualization/Healthy_vs_L&S.csv", header = T, sep = ",")

# Apply fisher test function for row values
row.names(Healthy_Adenomas) <- Healthy_Adenomas $Gene.Name
Healthy_Adenomas[1] <- NULL

row_fisher <- function(row, alt = 'two.sided', cnf = 0.95) {
  f <- fisher.test(matrix(row, nrow = 2), alternative = alt, conf.level =
cnf)
  return(c(row,
           p_val = f$p.value,
           or = f$estimate[[1]],
           or_ll = f$conf.int[1],
           or_ul = f$conf.int[2]))
}

Healthy_Adenomas_Fisher <- data.frame(t(apply(Healthy_Adenomas, 1,
row_fisher)))
head(Healthy_Adenomas_Fisher)

##         Mean.Positives (H) Mean.Negatives (H)   Mean.Positives (A)
## AMmurB                  19                 42                   16
## ECmurB                  20                 41                   13
## EFmurB                   6                 55                    4

            Mean.Negatives (A)   p_val        or      or_ll      or_ul
## AMmurB                  26 0.5279932 0.7373602 0.2979267  1.829111
## ECmurB                       29 1.0000000 1.0872701 0.4329705  2.789415
## EFmurB                       38 1.0000000 1.0360099 0.2277628  5.340826
```

Step 22: Saving a table example

Any time a table is to be saved to the computer's files, the "write.table" function must be employed.

```
write.table(Healthy_Adenomas_Fisher, file = "Healthy_Adenomas_Fisher.csv",
sep = ",", col.names = TRUE,
            qmethod = "double")
```

3.3 Results

Utilizing the metagenomic path generated for this aim, the results pertaining to the genes evaluated in Chapter 3, that is, *clbB*, *cnf-1*, *gelE*, and *tcpC*, were obtained. Firstly, the total number of hits, positive patients, and corresponding frequency values found for each gene among each individual cohort was tabulated (Table 3.1). Furthermore, the standard deviation, minimum value, median, maximum value and mode between these patients were also found to observe the distribution of said hit results. As in Chapter 2, the P-values pertaining to the difference between the positive patient frequency values in the adenoma and CRC cohorts compared to those of the healthy individuals was also found and included. There were no statistically significant values within the genes studied.

Table 3.1. Presence and frequency of proinflammatory and genotoxin genes in the selected metagenomic databases cohorts, i.e., Healthy, Adenomas, and CRC

Healthy Cohort				
Gene	**clbB**	**cnf-1**	**gelE**	**tcpC**
Total Hits	7	160	27	25
Mean Hits ± σ	2 ± 1	20 ± 25	6 ± 6	7 ± 6
Positives	4	10	7	6
Frequency %	7	16	11	10
Min	2	1	2	2
Median	2	7	3	7
Max	3	70	15	14
Mode	2	-	-	-

Adenomas Cohort				
Gene	**clbB**	**cnf-1**	**gelE**	**tcpC**
Total Hits	0	7	24	0
Mean Hits ± σ	-	4 ± 0	5 ± 4	-
Positives	0	2	8	0
Frequency % (P-value)[a]	0 (0.14)	5 (0.12)	19 (0.39)	0 (0.08)
Min	-	4	2	-
Median	-	4	4	-
Max	-	4	10	-
Mode	-	-	-	-

Cancer Cohort				
Gene	**clbB**	**cnf-1**	**gelE**	**tcpC**
Total Hits	5	67	32	12
Mean Hits ± σ	1 ± 0	13 ± 14	6 ± 5	2 ± 1
Positives	4	7	7	7
Frequency % (P-value)[a]	8 (1.00)	13 (0.79)	13 (0.78)	13 (0.77)
Min	1	1	1	1
Median	1	5	5	2
Max	2	29	14	3
Mode	1	-	-	1

Table 3.1. The table is divided into the three cohorts (healthy, adenomas and cancer). The column categories are the corresponding genes (clbB, cnf-1, gelE, and tcpC), and the row categories present the average number of hits with the standard deviation (σ), minimum value, median, maximum value, and mode found for each cohort. The adenomas cohort results are obtained by grouping the small adenoma and large adenoma cohort results. [a] P-values obtained using Fisher's Exact Test when comparing the Adenomas and Cancer positive patients to the ones from the Healthy cohort. P-values are statistically significant when $P \leq 0.05$. Statistically significant values are placed in bold.

3.4 Discussion

Despite the importance of metagenomic sequence repositories for the analysis of microbial genes in diseases, there is currently no available platform to facilitate the navigation through such complex data files to obtain information that could lead to novel hypotheses. The search through these data repositories continues to be the exclusive province of data scientists and programmers that can easily derive home-made scripts for the automation of searches and for the analysis of search results. Still unavailable is a simple platform that can easily navigate the data repositories to infer the involvement of specific microbial genes in disease. In this project, we have laid down the foundation for such a method by proposing a series of 22 commands that result in the search for specific genes in these databases. In this work, we focused on a specific subset of bacterial gene sequences (Table S1). Our results showed that no statistically significant difference was found between these additional genes in any of the cohorts when compared to the healthy individuals (Table 3.1). As seen in the table, not all the genes were present in all cohorts. Specifically, no *clbB* or *tcpC* was found in the Adenomas cohort.

This general lack of differences with statistical significance was expected. The microbiome is a complex and changing collection of genes and sequences in which individual genes are naturally dilute. For instance, experimentally it has been observed that colibactin genes (in this case exemplified by gene *clbB*), are only present in roughly 20% of healthy individuals. Thus, in a cohort of 50, we would only expect 10 individuals with clbB, if the shotgun approach was exhaustive with 100% coverage of the microbiome. However, we know that most *E. coli* isolates do not harbor these genotoxin-encoding genes and most individuals test negative for these genes by PCR (Shimpoh, et al., 2017). Another limitation of these studies is the lack of knowledge of the "true" health status of individuals in the healthy cohort. There is no information on the complete health status and medical history of the individuals evaluated, which could potentially affect the study. That is, we do not know how healthy the "healthy" group was or if any patient had other conditions that could have affected the results. This lack of knowledge on the true health status of individuals in the cohort, could have resulted in the presence of outliers in the data, i.e., individuals with an abnormally high number of gene hits in any cohort.

In a later analysis step, we removed outliers defined statistically as values outside of two standard deviations from the mean. The results with outlier values removed is shown in Table S4. Clearly the removal of these outliers did not make a difference from the comparison of these values vs. our original results. That is, even though the values of most total hits, mean hits, their respective standard deviations, and positive frequency percentages decreased, it was quite uniform among all cohorts, which led us to believe that the P-values would not change greatly. Furthermore, the vgrG1 group from the adenomas cohort was not affected since no outliers were found, thus further validating our original results.

Interestingly, by obtaining the amount of total hits, the number of positives, and the corresponding frequency percentage of each of the four genes, we were able to arrive at plausible results, thus further authenticating a newly established automatable workflow. As was done for the results reported in Chapter 2, the primary results obtained from our newly applied pathway were used to find the average value of hits for each gene within a cohort, the minimum and maximum values, median, mode, and the standard deviation for these values (Table 3.2). This work was carried out with the datasets resulting from a single study that analyzed stool samples from two populations (Washington DC, and France). However, with new metagenomic data populating the ever-growing sequence repositories, it will be possible to employ this strategy for the analysis of microbial genes in a larger and more global scale.

Chapter 4. Conclusion

Throughout this study we were able to perform an in-depth analysis of the metagenomic data from the European Nucleotide Archive. By thoroughly processing and analyzing the data obtained from the alignment of the qualified patients from each cohort with our specific genes of interest, we achieved a better understanding of the roles and effects that certain gene sequences may exert, as well as how these may be susceptible to the environment in which they are found. Though a positive correlation between the presence of the Hcp sequences and the adenomas and colorectal cancer was not found, other patterns of interest were viewed. For example, the varied presence of the four different Hcp sequences within each cohort and how these results compared to those of vgrG1. Moreover, because of the cautious procedures carried out throughout the study, we have constructed a reusable pathway that could be used to further study these sequences or be applied to other studies. Utilizing this metagenomic path, after frequently validating the process and filtering the data, we were able to obtain numerous hits with small E values for most of the genes related to the T6SS studied, which had not been done under one study. Thus, despite there was only one statistically significant difference obtained (vgrG1, Adenomas cohort), there were noticeable patterns or tendencies present. The most curious of these was that most of these genes studied reduced their frequency percentage in the Adenomas group. It was also found that Usp was found with higher frequencies than the Hcp domain alone, even though this protein domain is shared with numerous effectors (Ma, et al., 2017). Therefore, we can conclude that these DNases are well represented in these conditions. Related to this, it was also interesting to see how the Hcp isoforms had different frequencies, which could be related to their different roles and functions.

Regarding the limitations mentioned in the discussion of the second aim, it is important to find further ways to validate the study, such as applying Fisher’s Exact Test to the results obtained after removing the statistical outliers (Table S4). In addition to the limitations faced in the current project, in order to gain a more reliable outlook on the presence of these specific proteins within the disease groups, a study of the protein domains themselves could be incorporated. That is, by also searching for these amino acid sequences, a fuller understanding of the distribution of each gene within each cohort could be obtained. Following this framework, we could further investigate the presence of other emerging bacterial sequences for their involvement in disease. Future investigations should delve into the role of *vgrG* in adenomas, since this sequence was the only one that proved to be statistically different in adenomas vs. controls in the entire project. It would be interesting to see whether VgrG itself or a VgrG homolog has any gut-specific function against the host's cells. Regardless of its many limitations, we developed a computational pipeline that successfully enabled the study and search of a select group of genes with the adenomas and colorectal cancer conditions. This process can be implemented in future studies to better our understanding of genotoxic and proinflammatory factors in specific diseases.

Bibliographical References

Armstrong, G. (2013). Uropathogenic *Escherichia coli* colicin-like Usp and associated proteins: Their evolution and role in pathogenesis. *The Journal of Infectious Diseases*, 208(10), 1539-1541. doi:10.1093/infdis/jit482.

Brennan, C.A. & Garett, W.S. (2016). Gut microbiota, inflammation, and colorectal cancer. *Annu Rev Microbiol,* 70, 395–411. doi:10.1146/annurev-micro-102215-095513.

Centers for Disease Control and Prevention. (2021, February 8). *What Are the Risk Factors for Colorectal Cancer?* Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/colorectal/basic_info/risk_factors.htm

Črnigoj, M., Podlesek, Z., Budič, M., & Žgur-Bertok, D. (2014). The *Escherichia coli* uropathogenic-specific-protein-associated immunity protein 3 (Imu3) has nucleic acid – binding activity. *BMC Microbiol,* 14(16), 1-8. doi:10.1186/1471-2180-14-16.

Durand E, Derrez E, Audoly G, Spinelli S, Ortiz-Lombardia M, Raoult D, Cascales E, & Cambillau C. (2012). Crystal structure of the VgrG1 actin cross-linking domain of the Vibrio cholerae type VI secretion system. *J Biol Chem*, 287(45), 38190-9. doi: 10.1074/jbc.M112.390153.

Dutta, P., Jijumon, A. S., Mazumder, M., Dileep, D., Mukhopadhyay, A. K., Gourinath, S., & Maiti, S. (2019). Presence of actin binding motif in VgrG-1 toxin of Vibrio cholerae reveals the molecular mechanism of actin cross-linking. *International journal of biological macromolecules*, *133*, 775–785. https://doi.org/10.1016/j.ijbiomac.2019.04.026

Gómez-Moreno R, González-Pons M, Soto-Salgado M, Cruz-Correa M, & Baerga-Ortiz A. (2019) The presence of gut microbial genes encoding bacterial genotoxins or pro-inflammatory factors in stool samples from individuals with colorectal neoplasia. *Diseases*, 7(1),16. doi:10.3390/diseases7010016.

Gómez-Moreno, R., Martínez-Ramírez, R., Roche-Lima, A., Carrasquillo-Carrión, K., Pérez-Santiago, J., & Baerga-Ortiz, A. (2019). Hotspots of sequence variability in gut microbial genes encoding pro-inflammatory factors revealed by oligotyping. *Frontiers,* 10(631), 1-9. doi:10.3389/fgene.2019.00631.

Gomez-Moreno, R., Robledo, I. E., & Baerga-Ortiz, A. (2014). Direct detection and quantification of bacterial genes associated with inflammation in DNA isolated from stool. *Adv. Microbiol.* 4, 1065–1075. doi: 10.4236/aim.2014.415117.

Hersch, S. J., Manera, K., & Dong, T. G. (2020). Defending against the type six secretion system: Beyond immunity genes. *Cell Reports, 33*(2), 108259. doi:10.1016/j.celrep.2020.108259

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD - Visual Molecular Dynamics. J. Molec. Graphics 14, 33–38. doi:10.1016/0263-7855(96)00018-5.

Ma, J., Sun, M., Pan, Z., Song, W., Lu, C., & Yao, H. (2018). Three Hcp homologs with divergent extended loop regions exhibit different functions in avian pathogenic *Escherichia coli*. *Taylor and Francis Online,* 143(1), 1-13. doi:10.1038/s41426-018-0042-0.

Ma, J., Pan, Z., Huang, J., Sun, M., Lu, C., & Yao, H. (2017). The Hcp proteins fused with diverse extended-toxin domains represent a novel pattern of antibacterial effectors in type VI secretion systems. *Virulence*, *8*(7), 1189–1202. https://doi.org/10.1080/21505594.2017.1279374

Martin, H. M., Campbell, B. J., Hart, C. A., Mpofu, C., Nayar, M., Singh, R., et al. (2004). Enhanced Escherichia coli adherence and invasion in Crohn's disease and colon cancer 1. *Gastroenterology* 127, 80–93. doi: 10.1053/j. gastro.2004.03.054

Mougous, J.D., Cuff, M.E., Raunser, S., Shen, A., Zhou, M., Gifford, C.A., Goodman, A.L., Joachimiak, G., Ordoñez, C.L., Lory, S. Walz, T., Joachimiak, A., & Mekalanos, J.J. (2006). A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus. *Science,* 312(5779), 1526–1530. doi:10.1126/science.1128393.

Navarro-Garcia, F., Ruiz-Perez, F., Cataldi, A., & Larzábal, M. (2019). Type VI secretion system in pathogenic *Escherichia coli*: Structure, role in virulence, and acquisition. *Frontiers*, 10(1965), 1-17. doi:10.3389/fmicb.2019.01965.

Nipič, D., Podlesek, Z., Budič, M., Black goi, M., & Žgur-Bertok, D. (2013). *Escherichia coli* Uropathogenic-Specific Protein, Usp, is a bacteriocin-like genotoxin. *The Journal of Infectious Diseases,* 208(10), 1545-1552. doi:10.1093/infdis/jit480.

PDQ Cancer Genetics Editorial Board. (2020, December 17). *Genetics of Colorectal Cancer*

(PDQ®). PDQ Cancer Information Summaries [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK126744/.

Peng, Y., Wang, X., Shou, J., Zong, B., Zhang, Y., Tan, J., Chen, J., Hu, L., Zhu, Y., Chen, H., & Tan, C. (2016). Roles of Hcp family proteins in the pathogenesis of the porcine extraintestinal pathogenic *Escherichia coli* type VI secretion system. *Scientific Report*, 6, 1-9. doi:10.1038/srep26816.

Raisch, J., Buc, E., Bonnet, M., Sauvanet, P., Vazeille, E., de Vallée, A., et al. (2014). Colon cancer-associated B2 Escherichia coli colonize gut mucosa and promote cell proliferation. *World J. Gastroenterol.* 20, 6560–6572. doi: 10.3748/wjg.v20.i21.6560.

Rihtar, E., Žgur Bertok, D., & Podlesek, Z. (2020). The Uropathogenic Specific Protein Gene *usp* from Escherichia coli and Salmonella bongori is a Novel Member of the TyrR and H-NS Regulons. *Microorganisms*, *8*(3), 330. MDPI AG. Doi: 10.3390/microorganisms8030330.

Roche-Lima, A., Carrasquillo-Carrión, K., Gómez-Moreno, R., Cruz, J., Velázquez-Morales, D., Rogozin, I., & Baerga-Ortiz, A. (2018). The presence of genotoxic and/or pro-inflammatory bacterial genes in gut metagenomic databases and their possible link with inflammatory bowel diseases. *Frontiers,* 9(116), 1-8. doi: 10.3389/fgene.2018.00116.

Ruiz, F. M., Santillana, E., Spínola-Amilibia, M., Torreira, E., Culebras, E., & Romero, A.

(2015). Crystal Structure of Hcp from Acinetobacter baumannii: A Component of the Type VI Secretion System. *PloS one*, 10(6), e0129691. doi: 10.1371/journal.pone.0129691.

Sepehri, S., Khafipour, E., Bernstein, C., Coombes, B., Pilar, A., Karmali, M., Ziebell, K., & Krause, D. (2011). Characterization of Escherichia coli isolated from gut biopsies of newly diagnosed patients with inflammatory bowel disease. *Inflammatory Bowel Diseases*, 17(7), 1451–1463. doi:10.1002/ibd.21509.

Shimpoh, T., Hirata, Y., Ihara, S., Suzuki, N., Kinoshita, H., Hayakawa, Y., Ota, Y., Narita, A., Yoshida, S., Yamada, A., & Koike, K. (2017). Prevalence of *pks*-positive *Escherichia coli* in Japanese patients with or without colorectal cancer. *Gut Pathog,* 9(35). doi: 10.1186/s13099-017-0185-x

Silverman, J.M., Agnello, D.M., Zheng, H., Andrews, B. T., Li, M., Catalano, C. E., Gonen, T., Mougous, J. D. (2013). Haemolysin Coregulated Protein is an exported receptor and chaperone of Type VI Secretion substrates. *Molecular Cell,* 51(5), 584-593. doi: 10.1016/j.molcel.2013.07.025.

Swidsinski, A., Khilkin, M., Kerjaschki, D., Schreiber, S., Ortner, M., Weber, J., et al. (1998). Association between intraepithelial Escherichia coli and colorectal cancer. *Gastroenterology* 115, 281–286. doi: 10.1016/S0016-5085(98)70194-5

The American Cancer Society. (2021, January 12). *Colorectal Cancer Statistics: How Common Is Colorectal Cancer?* American Cancer Society. https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html.

Vogtmann, E., Hua, X., Zeller, G., Sunagawa, S., Voigt, A.Y., Hercog, R., Goedert, J.J., Shi, J., Bork, P., & Sinha, R. (2016). Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PLoS One*, 12;11(5), e0155362. doi: 10.1371/journal.pone.0155362.

Weir, T. L., Manter, D. K., Sheflin, A. M., Barnett, B. A., Heuberger, A. L., & Ryan, E. P. (2013). Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS One* 8:e70803. doi: 10.1371/journal.pone.0070803.

Wong R. S. (2011). Apoptosis in cancer: from pathogenesis to treatment. *Journal of experimental & clinical cancer research : CR*, *30*(1), 87. https://doi.org/10.1186/1756-9966-30-87

Yang, X., Long, M., & Shen, X. (2018). Effector–immunity pairs provide the T6SS nanomachine its offensive and defensive capabilities. *Molecules,* 23(5), 1-14. doi:10.3390/molecules23051009.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature Methods,* 12, 7–8. doi:10.1038/nmeth.3213.

Zaw, M.T., Lai, Y.M., & Lin, Z. (2013). Uropathogenic Specific Protein: Epidemiologic marker of Uropathogenic Escherichia coli as well as Non-specific DNase. *International Journal of Collaborative Research on Internal Medicine & Public Health*, 5, 630-640. https://www.iomcworld.org/articles/uropathogenic-specific-protein-epidemiologic-marker-ofuropathogenic-escherichia-coli-as-well-as-nonspecificdnase.pdf

Zaw, M.T., Yamasaki, E., Yamamoto, S., Nair, G.B., Kawamoto, K., & Kurazono, H. (2013). Uropathogenic specific protein gene, highly distributed in extraintestinal uropathogenic *Escherichia coli*, encodes a new member of H-N-H nuclease superfamily. *Gut Pathog,* 5(13), 1-9. doi:10.1186/1757-4749-5-13.

Zhou, Y., Yu, H., Ni, J., Zeng, L., Teng, Q., Kim, K. S., Zhao, G. P., Guo, X., & Yao, Y. (2012). HCP Family Proteins Secreted via the Type VI Secretion System Coordinately Regulate *Escherichia Coli* K1 Interaction with Human Brain Microvascular Endothelial Cells. *Infection and Immunity*, 80(3), 1243–51. doi: 10.1128/iai.05994-11.

Supplementary Information

Table S1. Genes of interest with corresponding accession number

Gene	Accession Number
hcp (domain)	CP043181.1:3517124-3518905
hcp-1	CP006632.1:246100-246582
hcp-2	CP006632.1:267403-267921
hcp-3	CP006632.1:4258749-4259897
tssK	CP043181.1:3647520-3648866
usp	AB027193
vgrG1	CP006632.1:268131-269756

Table S1. The genes of interest of this study are listed in the first column, followed by their unique accession numbers, obtained from NCBI database.

Table S2. Presence of T6SS genes of interest in small and large adenoma cohorts

	Healthy cohort (N = 61)			**Small Adenoma cohort (N = 27)**			**Large Adenoma cohort (N = 15)**		
Gene	**Total hits**	**Positives**	**Frequency %**	**Total hits**	**Positives**	**Frequency % (P-value)**	**Total hits**	**Positives**	**Frequency % (P-value)**
AMmurB	168	21	34	57	8	30 (0.81)	75	8	53 (0.24)
ECmurB	243	21	34	71	11	41 (0.63)	20	3	20 (0.36)

EFmurB	22	7	11	5	2	7 (0.72)	16	2	13 (1.00)
HCP	66	9	15	29	2	7 (0.49)	6	2	13 (1.00)
HCP-1	59	7	11	8	2	7 (0.72)	0	0	0 (0.33)
HCP-2	38	5	8	16	5	19 (0.27)	0	0	0 (0.58)
HCP-3	80	14	23	6	2	7 (0.13)	2	1	7 (0.28)
tssK	29	11	18	1	1	4 (0.10)	16	2	13 (1.00)
USP	88	11	18	32	1	4 (0.10)	4	1	7 (0.44)
vgrG1	274	14	23	121	11	41 (0.12)	33	7	47 (0.10)

Table S2. The columns represent the Gene names, Total number of hits, Number of patients with hits (Positives), and the corresponding frequency of each gene within each cohort. The P-values displayed were obtained using Fisher's Exact Test when individually comparing the Small Adenoma and Large Adenoma positives to those of the Healthy cohort. P-values are statistically significant when P ≤ 0.05.

Table S3. Number of hits per positive patient per gene per cohort

	Healthy Cohort (N = 61)		Small Adenoma Cohort (N = 27)		Large Adenoma Cohort (N = 15)		Adenomas Cohort (N = 42)		Cancer Cohort (N = 53)	
Gene	**Patient ID**	**Number of hits**	**Patient ID**	**Number of hits**	**Patient ID**	**Number of hits**	**Patient ID**	**Number of hits**	**Patient ID**	**Number of hits**
AMmurB	SAMEA2466887	27	SAMEA2466943	7	SAMEA2466895	8	SAMEA2466943	7	SAMEA2466908	2
	SAMEA2466890	5	SAMEA2466959	3	SAMEA2466906	6	SAMEA2466959	3	SAMEA2466915	5
	SAMEA2466891	10	SAMEA2466960	2	SAMEA2466910	2	SAMEA2466960	2	SAMEA2466922	5
	SAMEA2466896	4	SAMEA2466974	10	SAMEA2466921	16	SAMEA2466974	10	SAMEA2466926	16
	SAMEA2466898	15	SAMEA2466975	3	SAMEA2466925	7	SAMEA2466975	3	SAMEA2466930	20
	SAMEA2466903	7	SAMEA2466994	1	SAMEA2466935	5	SAMEA2466994	1	SAMEA2466931	3
	SAMEA2466907	6	SAMEA2467003	28	SAMEA2466945	16	SAMEA2467003	28	SAMEA2466932	2
	SAMEA2466909	4	SAMEA2467006	4	SAMEA2466950	16	SAMEA2467006	4	SAMEA2466934	4
	SAMEA2466911	10					SAMEA2466895	8	SAMEA2466937	8
	SAMEA2466961	45					SAMEA2466906	6	SAMEA2466942	4
	SAMEA2466965	1					SAMEA2466910	2	SAMEA2466947	2
	SAMEA2466982	5					SAMEA2466921	16	SAMEA2466957	4
	SAMEA2467016	7					SAMEA2466925	7	SAMEA2466964	35
	SAMEA2467021	1					SAMEA2466935	5	SAMEA2466968	9
	SAMEA2467024	11					SAMEA2466945	16	SAMEA2466972	1

	SAMEA2467025	4					SAMEA2466950	16	SAMEA2466973	29
	SAMEA2467028	1							SAMEA2466983	10
	SAMEA2467033	3							SAMEA2466987	123
	SAMEA2467035	2							SAMEA2466989	48
	SAMEA2467039	4							SAMEA2466991	4
	SAMEA2467041	1							SAMEA2466993	2
									SAMEA2467001	20
									SAMEA2467002	12
									SAMEA2467004	4
clbB	SAMEA2467013	2							SAMEA2466915	2
	SAMEA2467018	2							SAMEA2467001	1
	SAMEA2467035	3							SAMEA2467004	1
cnf-1	SAMEA2466901	2	SAMEA2466975	4			SAMEA2466975	4	SAMEA2466915	29
	SAMEA2466911	1							SAMEA2466932	1
	SAMEA2466979	8							SAMEA2466948	27
	SAMEA2467012	6							SAMEA2466984	5
	SAMEA2467013	23							SAMEA2467004	3
	SAMEA2467018	47								
	SAMEA2467035	70								
	SAMEA2467041	3								
ECmurB	SAMEA2466887	3	SAMEA2466919	1	SAMEA2466921	5	SAMEA2466919	1	SAMEA2466915	21
	SAMEA2466891	69	SAMEA2466923	6	SAMEA2466949	12	SAMEA2466923	6	SAMEA2466917	19
	SAMEA2466897	2	SAMEA2466943	3	SAMEA2466962	4	SAMEA2466943	3	SAMEA2466922	3
	SAMEA2466911	9	SAMEA2466951	5			SAMEA2466951	5	SAMEA2466927	2
	SAMEA2466913	1	SAMEA2466952	8			SAMEA2466952	8	SAMEA2466937	3
	SAMEA2466940	32	SAMEA2466959	2			SAMEA2466959	2	SAMEA2466938	95
	SAMEA2466979	5	SAMEA2466960	33			SAMEA2466960	33	SAMEA2466941	3
	SAMEA2467012	3	SAMEA2466966	4			SAMEA2466966	4	SAMEA2466944	102
	SAMEA2467013	12	SAMEA2466971	2			SAMEA2466971	2	SAMEA2466946	9
	SAMEA2467016	26	SAMEA2466999	7			SAMEA2466999	7	SAMEA2466948	9
	SAMEA2467017	8	SAMEA2467006	2			SAMEA2467006	2	SAMEA2466957	4
	SAMEA2467018	17					SAMEA2466921	5	SAMEA2466984	2
	SAMEA2467019	1					SAMEA2466949	12	SAMEA2466986	14

	SAMEA2467021	4					SAMEA2466962	4	SAMEA2466989	56
	SAMEA2467024	2							SAMEA2466991	5
	SAMEA2467030	6							SAMEA2466993	4
	SAMEA2467032	4							SAMEA2466998	14
	SAMEA2467033	14							SAMEA2467001	7
	SAMEA2467035	26							SAMEA2467002	13
	SAMEA2467037	1							SAMEA2467004	4
	SAMEA2467042	4							SAMEA2467009	4
EFmurB	SAMEA2466897	1	SAMEA2466923	3	SAMEA2466935	5	SAMEA2466923	3	SAMEA2466938	1
	SAMEA2466903	1	SAMEA2466995	2	SAMEA2466949	12	SAMEA2466995	2	SAMEA2466942	8
	SAMEA2466911	3					SAMEA2466935	5	SAMEA2466948	2
	SAMEA2467018	3					SAMEA2466949	12	SAMEA2466968	6
	SAMEA2467030	10								
	SAMEA2467034	2								
	SAMEA2467035	4								
gelE	SAMEA2466911	4	SAMEA2466923	2	SAMEA2466935	5	SAMEA2466923	2	SAMEA2466938	1
	SAMEA2467025	2	SAMEA2466995	3	SAMEA2466949	10	SAMEA2466995	3	SAMEA2466942	14
	SAMEA2467029	3					SAMEA2466935	5	SAMEA2466948	6
	SAMEA2467030	15					SAMEA2466949	10	SAMEA2466968	5
									SAMEA2466986	4
HCP	SAMEA2466891	4	SAMEA2466960	26	SAMEA2466949	4	SAMEA2466960	26	SAMEA2466904	2
	SAMEA2466913	2	SAMEA2466966	4	SAMEA2466962	2	SAMEA2466966	4	SAMEA2466915	13
	SAMEA2466979	3					SAMEA2466949	4	SAMEA2466946	4
	SAMEA2467012	2					SAMEA2466962	2	SAMEA2466948	12
	SAMEA2467013	8							SAMEA2466957	3
	SAMEA2467018	19							SAMEA2467002	10
	SAMEA2467019	2							SAMEA2467004	3
	SAMEA2467033	5								
	SAMEA2467035	23								
HCP-1	SAMEA2466891	36	SAMEA2466923	6			SAMEA2466923	6	SAMEA2466915	12
	SAMEA2466911	2	SAMEA2466999	2			SAMEA2466999	2	SAMEA2466938	39
	SAMEA2466940	14							SAMEA2466944	21
	SAMEA2466980	1							SAMEA2466986	7

	SAMEA2467017	2							SAMEA2466989	4
	SAMEA2467030	3							SAMEA2466991	3
	SAMEA2467033	2							SAMEA2466993	1
									SAMEA2467002	3
HCP-2	SAMEA2466891	27	SAMEA2466923	5			SAMEA2466923	5	SAMEA2466938	38
	SAMEA2466911	2	SAMEA2466952	3			SAMEA2466952	3	SAMEA2466986	6
	SAMEA2467030	3	SAMEA2466960	1			SAMEA2466960	1	SAMEA2466989	41
	SAMEA2467033	4	SAMEA2466971	2			SAMEA2466971	2	SAMEA2466993	3
	SAMEA2467035	3	SAMEA2466999	6			SAMEA2466999	6	SAMEA2466998	3
									SAMEA2467002	3
HCP-3	SAMEA2466891	3	SAMEA2466923	4	SAMEA2466949	2	SAMEA2466923	4	SAMEA2466915	5
	SAMEA2466911	3	SAMEA2466966	3			SAMEA2466966	3	SAMEA2466927	2
	SAMEA2466913	1					SAMEA2466949	2	SAMEA2466941	2
	SAMEA2466979	2							SAMEA2466946	9
	SAMEA2467012	3							SAMEA2466948	8
	SAMEA2467013	6							SAMEA2466957	5
	SAMEA2467016	17							SAMEA2466984	1
	SAMEA2467018	14							SAMEA2467001	2
	SAMEA2467019	2							SAMEA2467002	4
	SAMEA2467021	3								
	SAMEA2467032	3								
	SAMEA2467033	5								
	SAMEA2467035	15								
	SAMEA2467042	5								
tcpC	SAMEA2466979	2							SAMEA2466927	1
	SAMEA2467013	7							SAMEA2466932	2
	SAMEA2467018	14							SAMEA2466946	3
									SAMEA2466948	3
									SAMEA2467004	1
tssK	SAMEA2466891	4	SAMEA2466975	1	SAMEA2466949	7	SAMEA2466975	1	SAMEA2466922	86
	SAMEA2466897	1			SAMEA2466962	9	SAMEA2466949	7	SAMEA2466927	2
	SAMEA2466911	2					SAMEA2466962	9	SAMEA2466932	2
	SAMEA2467012	5							SAMEA2466946	14

	SAMEA2467017	3							SAMEA2466948	14
	SAMEA2467018	2							SAMEA2466957	4
	SAMEA2467019	5							SAMEA2466984	3
	SAMEA2467022	2							SAMEA2466998	8
	SAMEA2467032	4							SAMEA2467001	3
	SAMEA2467035	2							SAMEA2467004	6
	SAMEA2467041	3								
USP	SAMEA2466891	4	SAMEA2466960	32	SAMEA2466962	4	SAMEA2466960	32	SAMEA2466904	4
	SAMEA2466911	3					SAMEA2466962	4	SAMEA2466915	18
	SAMEA2466979	3							SAMEA2466922	13
	SAMEA2467012	3							SAMEA2466941	3
	SAMEA2467013	13							SAMEA2466946	6
	SAMEA2467018	24							SAMEA2466948	10
	SAMEA2467019	5							SAMEA2466956	2
	SAMEA2467032	5							SAMEA2466957	4
	SAMEA2467033	7							SAMEA2467001	4
	SAMEA2467035	22							SAMEA2467002	6
	SAMEA2467041	2							SAMEA2467004	2
vgrG1	SAMEA2466891	152	SAMEA2466899	1	SAMEA2466893	2	SAMEA2466899	1	SAMEA2466926	2
	SAMEA2466897	1	SAMEA2466923	30	SAMEA2466906	3	SAMEA2466923	30	SAMEA2466927	1
	SAMEA2466911	15	SAMEA2466929	1	SAMEA2466921	3	SAMEA2466929	1	SAMEA2466937	3
	SAMEA2466940	13	SAMEA2466943	3	SAMEA2466935	2	SAMEA2466943	3	SAMEA2466938	124
	SAMEA2467017	9	SAMEA2466951	8	SAMEA2466945	6	SAMEA2466951	8	SAMEA2466944	61
	SAMEA2467018	2	SAMEA2466952	25	SAMEA2466949	16	SAMEA2466952	25	SAMEA2466947	1
	SAMEA2467024	5	SAMEA2466959	3	SAMEA2466962	2	SAMEA2466959	3	SAMEA2466963	1
	SAMEA2467025	5	SAMEA2466960	16			SAMEA2466960	16	SAMEA2466983	2
	SAMEA2467026	2	SAMEA2466971	8			SAMEA2466971	8	SAMEA2466986	21
	SAMEA2467030	14	SAMEA2466999	26			SAMEA2466999	26	SAMEA2466989	224
	SAMEA2467033	44	SAMEA2467008	3			SAMEA2467008	3	SAMEA2466991	7
	SAMEA2467035	12					SAMEA2466893	2	SAMEA2466993	25
	SAMEA2467037	1					SAMEA2466906	3	SAMEA2466998	11
	SAMEA2467040	1					SAMEA2466921	3	SAMEA2467002	6
							SAMEA2466935	2	SAMEA2467004	1

				SAMEA2466945	6	
				SAMEA2466949	16	
				SAMEA2466962	2	

Table S3. The preceding table groups all the patients considered in the study (all those who displayed at least one hit in each replica group for each individual genes). The unique Patient IDs, which identify each patient of the study, as well as the number of hits each patient had for each individual gene are presented. Following the first column, which corresponds to all the genes studied, the five patient cohorts are presented (from left to right: healthy, small adenoma, large adenoma, grouped adenomas, and cancer).

Table S4. Presence and frequency of genes of interest in the selected metagenomic databases cohorts without outlier values

Healthy Cohort (N=61)										
Gene	**AMmurB**	**ECmurB**	**EFmurB**	**Hcp**	**Hcp-1**	**Hcp-2**	**Hcp-3**	**tssK**	**Usp**	**vgrG1**
Q1	3	3	1	2	2	3	3	2	3	2
Q3	10	16	4	14	14	16	6	4	13	14
IQR	8	13	3	12	12	13	3	2	10	12
Upper Limit	21	35	9	31	32	35	11	7	28	32
Lower Limit	-9	-17	-4	-15	-16	-17	-2	-1	-12	-16
Total Hits	98	174	13	66	24	38	34	29	88	78
Mean Hits ± σ	5 ± 4	9 ± 9	2 ± 1	7 ± 8	4 ± 5	8 ± 11	3 ± 1	3 ± 1	8 ± 8	7 ± 5
Positives	19	20	6	9	6	5	11	11	11	12
Frequency % of Positives	31	33	10	15	10	8	18	18	18	20
Adenomas Cohort (N=42)										
Gene	**AMmurB**	**ECmurB**	**EFmurB**	**Hcp**	**Hcp-1**	**Hcp-2**	**Hcp-3**	**tssK**	**Usp**	**vgrG1**
Q1	3	2	3	3	-	2	-	-	-	2
Q3	13	7	9	15	-	6	-	-	-	16
IQR	10	5	6	12	-	4	-	-	-	14
Upper Limit	28	15	18	33	-	12	-	-	-	37
Lower Limit	-12	-6	-7	-15	-	-5	-	-	-	-19
Total Hits	132	58	21	66	8	16	9	17	36	154
Mean Hits ± σ	8 ± 7	4 ± 3	5 ± 4	7 ± 11	4 ± 3	3 ± 2	3 ± 1	6 ± 4	18 ± 20	9 ± 10

Positives	16	13	4	9	2	5	3	3	2	18
Frequency of positives (P-value)	38 (0.53)	31 (1.00)	10 (1.00)	21 (0.43)	5 (0.47)	12 (0.74)	7 (0.15)	7 (0.15)	5 (0.07)	43 (0.02)
Cancer Cohort (N=53)										
Gene	**AMmurB**	**ECmurB**	**EFmurB**	**Hcp**	**Hcp-1**	**Hcp-2**	**Hcp-3**	**tssK**	**Usp**	**vgrG1**
Q1	4	4	2	3	3	3	2	3	3	1
Q3	18	17	7	12	17	38	7	14	10	25
IQR	15	13	6	9	14	35	5	11	7	24
Upper Limit	40	36	15	26	37	91	13	31	21	61
Lower Limit	-18	-16	-7	-11	-17	-50	-5	-14	-8	-35
Total Hits	196	134	17	46	51	93	36	53	68	141
Mean Hits ± σ	9 ± 9	7 ± 6	4 ± 3	7 ± 5	7 ± 7	16 ± 18	4 ± 3	6 ± 5	6 ± 5	11 ± 17
Positives	22	18	4	7	7	6	9	9	11	13
Frequency of positives (P-value)	42 (0.33)	34 (1.00)	8 (0.75)	13 (1.00)	13 (0.77)	11 (0.75)	17 (1.00)	17 (1.00)	21 (0.81)	25 (0.65)

Table S4. In order to tend to one of the study's limitations, the statistical outliers were removed by finding the first (Q1) and third (Q3) quartile values, as well as the interquartile range (IQR) of the hit values obtained for each patient in each cohort (Table S3). From these values, the upper and lower limits were calculated by multiplying the IQR value by 1.5 and adding or subtracting that value to Q1 and Q3, respectively. After eliminating these, the total amount of hits obtained, the average amount of these hits, as well as the corresponding positive patients and their frequency percentages was found. Genes with three or less sample hits were not included in this statistical analysis (-).