Syrian Virtual University
Master in Bioinformatics

Investigating the correlation between Colorectal cancer mutational profile and the associated microbiota on Tumor and matched normal healthy tissue; A computational analysis

Authored By: Zena Jandali

Student ID: zeina_161419

Supervised by: Prof. Majd Al Jamali

Table of Contents

1. Abstract

Colorectal cancer is a prevalent and deadly malignancy with a significant global burden. It arises from the accumulation of genetic and epigenetic changes that transform normal colonic epithelial cells into adenocarcinomas. The microbiome plays a crucial role in CRC development. Bacterial biomarkers have prognostic value and hold potential for CRC detection and clinical outcome prediction.

The human gut microbiota is a vibrant ecosystem teeming with bacteria, viruses, fungi, and archaea, residing in a harmonious relationship with the host. It profoundly influences various aspects of human health, playing a crucial role in maintaining gut homeostasis, immune function, and metabolism.

In recent years, the association between colorectal cancer (CRC) and the microbiota has gained significant attention. Emerging evidence suggests that dysbiosis, a disruption in the gut microbiota's composition, may contribute to the initiation and progression of CRC. Studies have unveiled distinct alterations in the gut microbiota composition and diversity in individuals with CRC compared to healthy controls. These alterations encompass shifts in microbial taxa, decreased microbial diversity, and modifications in microbial metabolites. Specific bacterial species, such as Fusobacterium nucleatum, Bacteroides fragilis, and certain Enterococcus and Escherichia coli strains, have been implicated in CRC pathogenesis due to their capacity to promote inflammation, produce genotoxins, or modulate the tumor microenvironment. Thus, in this study we used 16S rRNA data from 60 samples belonging to 30 patients, from the tissue and the matched normal healthy tissue, the data went through characterization process using Linux shell command, bash programming language and R programming language with RStudio with various microbiome processing packages and tools, we implemented the DADA2 package, for Amplicon Sequencing Variants based approach. DADA2 (Denoising Amplicon Data with Adaptive Removal of Chimeras and Dereplication) is a widely used pipeline for analysing amplicon sequencing data. It employs a three-step approach to accurately identify and quantify microbial communities: error estimation, chimera detection, and denoising, the denoising algorithm employed by DADA2 is particularly effective in handling error-prone amplicon sequencing data and can significantly improve the accuracy of microbial community analysis.

The final product of the dada2 package is the corresponding taxonomy table of the data, next it is input to other packages for further manipulation, filtering and downstream analysis.

After further statistical analysis with various measure popular for microbiome studies, we compared the microbiome composition between tumor and matched healthy tissue in patients with colorectal cancer (CRC). Our findings align with previous studies highlighting the dominance of Firmicutes and Bacteroidetes phyla in the gut

microbiome. While overall diversity may not be affected, the presence of a tumor may influence the abundance of specific rare taxa. Differential abundance analysis identified the genus Ruminococcus within the Firmicutes phylum as significantly enriched in cancer tissues. This finding is intriguing, considering the potential role of Ruminococcus species in promoting tumor growth and pro-inflammatory responses.

2. Introduction

2.1 Colorectal cancer epidemiology and projections.

CRC ranks among the most prevalent with an incidence rate of 1.9 million in 2020 and deadly malignancies worldwide with 935,000 deaths in 2020 (Sung et al. 2021) According to the World Health Organization (WHO); by 2040, the burden of colorectal cancer will increase to 3.2 million new cases per year (an increase of 63%) and 1.6 million deaths per year (an increase of 73%).

The Colorectal cancer arises from the accumulation of genetic mutations and epigenetic alterations that drive the transformation of normal colonic epithelial cells into adenocarcinomas. The conventional model of CRC development involves the adenoma-carcinoma sequence of genetic changes and inflammatory-immunological factors to facilitate and shape a tumorigenic microenvironment, where benign adenomatous polyps gradually progress to invasive carcinomas over several years, allowing for potential intervention and early detection.(Farhana et al. 2018)

Colorectal cancer has several risk factors. These include male sex, metabolic syndrome, hypertension, diabetes, inflammatory bowel disease, obesity, sedentary behaviour, smoking, high alcohol consumption, high intake of sugar and red meat, and family history of colorectal cancer. Other risk factors include occupational exposure, and pollution. Genetic factors such as genetic mutations and inherited predisposition syndromes like Lynch syndrome and familial adenomatous polyposis also contribute to the risk of colorectal cancer (Ye, Chen, and Gu 2023).

Additionally, gut microbiota alterations, bacterial genotoxicity, biofilm formation, oxidative stress, bacterial metabolome, and dysbiosis are assessed as risk factors (Feizi et al). In the last few years, the role of the microbiome in the development of CRC has been increasingly emphasized. It is well known that the gut microbiome has an important role in the carcinogenesis of CRC, during the development of cancer, a complex interaction is established among the gut microbiome, tumour microbiome and immune system causing initial inflammation (Y.-Z. Zhang et al. 2018) and modulating different signalling pathways. Because bacterial biomarkers have the potential to detect CRC and predict clinical outcome, they have prognostic value (Rebersek 2021).

2.2 Microbiota and CRC: Emerging Paradigm

The human gut microbiota is a complex ecosystem comprising bacteria, viruses, fungi, and archaea, existing in a symbiotic relationship with the host. It influences various aspects of human health, and it plays a crucial role in maintaining gut homeostasis, immune function, and metabolism. Perturbations in the gut microbiota composition or function have been associated with numerous diseases, including inflammatory bowel diseases (IBD), metabolic disorders, and cancer.

The association between colorectal cancer (CRC) and the microbiota has garnered significant attention in recent years, Emerging evidence suggests that alterations in the gut microbiota composition, termed dysbiosis, may contribute to the initiation and progression of CRC.

Studies have revealed distinct alterations in the gut microbiota composition and diversity in individuals with CRC compared to healthy controls. These changes involve shifts in microbial taxa, decreased microbial diversity, and alterations in microbial metabolites. Specific bacterial species such as Fusobacterium Nucleatum, Bacteroides fragilis, and certain Enterococcus and Escherichia coli strains have been implicated in CRC pathogenesis due to their ability to promote inflammation, produce genotoxins, or modulate the tumor microenvironment (Gong et al. 2023).

2.3 Microbiota and CRC: Implications for understanding and Intervention

Differentiating the microbiota profiles between healthy colonic tissue and neoplastic cells within the colorectal environment is pivotal in understanding the dynamic interplay between the microbiota and tumorigenesis.(Masood et al. 2023) Characterizing the microbiota in both settings could offer insights into the early events that precede carcinogenesis, potential biomarkers for early detection (Burns et al. 2018). Furthermore, understanding these differences may aid in the development of innovative strategies, such as microbiota-based therapies or early diagnostic tools, to mitigate CRC incidence, progression, or improve treatment outcomes. (Chen et al. 2022)

some researchers have hypothesized that modulating the microbiota could be a lead for new targeted therapy. As a very prevailing therapy, microbiome-targeted therapy or cancer bacteriotherapy were designed based on how to modulate gut microbiota with a change of diet, probiotics, and faecal transplantation (Conti et al. 2023)

This field is evolving rapidly, and further exploration is necessary to comprehensively understand the complex interactions between the gut microbiota and CRC, ultimately aiming for improved clinical outcomes for patients.

2.4 Literature Review.

Colorectal cancer (CRC) is a multifactorial disease resulting from both genetic predisposition and environmental factors including the gut microbiota (GM) (Moskowitz et al. 2020)The relationship between CRC and the human GIT microbiome is a significant focus in current research. The microbiome serves as a crucial interface for environmental factors in the body and possesses a genomic makeup much larger than the unique human genes. This genetic richness contributes molecules that aid in maintaining the body's balance and health, supporting digestion, and educating the immune system. A healthy microbiome also prevents harmful bacteria from colonizing the gut by occupying spaces and competing for nutrients. However, in CRC patients, an imbalance (dysbiosis) in the gut microbiome exists, suggesting that bacteria might interfere with the molecular mechanisms behind CRC (Ternes et al. 2020)Most studies that have investigated the tumor microbiome composition have examined biopsies from the tumor sites and adjacent healthy tissues. Flemer et al. showed that a CRC-associated microbiota was found also in adjacent healthy tissues 2–30 cm away from the tumor and argued that a CRC distinctive microbiota was established prior to CRC development (Senthakumaran et al. 2023) (Dejea et al. 2014)

Though various effects of bacteria associated with CRC have been identified, the specific ways in which they promote cancer development remain unclear. To understand this link better, further exploration is needed to uncover how CRC-associated bacteria influence tumor initiation and progression.(Tjalsma et al. 2012) Bacteria potentially impact CRC by directly or indirectly affecting host cells or their surrounding environment through different mechanisms such as bacterial metabolism and secreted molecules (e.g., extracellular superoxide, genotoxins, short-chain fatty acids), attachment, invasion, translocation processes and modulation of host defences (Ternes et al. 2020).

Recent advancements in culture-based methods and qRT-PCR have allowed the identification of specific bacteria in colorectal tissue and patient stool samples. Next-generation sequencing techniques like 16S rRNA gene and metagenomic profiling provide essential data on CRC-associated microbiomes. (Wensel et al. 2022) Although 16S rRNA sequencing helps identify bacterial genera, it lacks resolution at the strain level. Nonetheless, it offers valuable insights into prevalent genera, serving as a starting point for more sensitive approaches like qRT-PCR and RNA/whole-genome sequencing studies (Lian et al. 2020). The choice of the hypervariable region for the 16S rRNA gene affects analysis depth and may introduce bias in microbial diversity. The 16S rRNA gene works as a rapid and effective marker for the identification of microorganisms in complex communities; hence, a huge number of microbiomes have been surveyed by 16S amplicon-based sequencing. The resolution of the 16S rRNA gene is always considered only at the genus level;

however, it has not been verified on a wide range of microbes yet (Zhang et al. 2023).

When comparing various studies, differences often exist between control and CRC patient groups in basic parameters like gender ratios, ethnic groups and age

(Fusobacterium nucleatum's role in CRC is significant. Its abundance in the gut is considered a possible biomarker for CRC, seen in both stool and tissue samples. FadA, a virulence factor in Fusobacterium, interacts with E-cadherin, fueling cancer cell growth through Wnt signaling (Smith et al. 2002). Additionally, higher FadA levels are detected in adenomas and adenocarcinomas compared to healthy tissues (Gong et al. 2023).

While certain bacteria like F. nucleatum, E. coli pks+, or B. fragilis directly engage with host receptors on tumor or immune cells, numerous bacterial effects might stem from secreted metabolites. The gut microbiome serves as a rich source of secretory proteins (secretome) and metabolites (metabolome), contributing to a shared reservoir of metabolites within the tumor microenvironment which can play a role in cancer progression [17]. Oncometabolites, such as l-2-hydroxyglutarate, succinate, and fumarate (upstream), or d-2-hydroxyglutarate and lactate (downstream), accumulate in cancer due to metabolic defects (Senthakumaran et al. 2023)

In the context of CRC, the microbiome associated with this cancer type serves as a potential source for such metabolites. For instance, B. fragilis, Prevotellaceae, and F. nucleatum have demonstrated the ability to produce succinate, which activates proinflammatory pathways through the succinate receptor 1 on immune cells].

The large intestine hosts a vast array of microbes that progressively increases from the small intestine (103–104 bacteria/mL) to the colon (1011 bacteria/mL), correlating with varying cancer risks along the GIT [24.]. Environmental selection and competition among microbes are key factors shaping the diversity of these microbial populations. As a result, distinct microenvironments emerge, potentially influencing tumor progression [25]. This becomes particularly pertinent when considering the contrasting biological, pathological, and epidemiological aspects of right-sided (RCC) versus left-sided (LCC) colorectal cancers.

Survival rates among CRC patients based on tumor location have sparked debate due to contradictory findings. Some studies suggest a poorer prognosis for tumors on the right side (caecum to ascending and transverse colon), while others indicate worse outcomes for tumors on the left side (splenic flexure and descending colon to rectosigmoid junction). These discrepancies in results may stem from differences in statistical methodologies and diverse characteristics within study cohorts, including variations in cohort size and age.

In CRC, certain characteristics like hypermutable microsatellite instability (MSI)-high, CpG island methylator phenotype (CIMP)-high phenotypes, and BRAF mutation rates are recognized to decrease from the ascending colon to the rectum [26]. This gradient might influence how intestinal microbiota potentially impacts the disease along the proximal–distal axis. Notably, research by (Dejea et al). highlights that the arrangement of bacteria into biofilm structures within the gut's mucus layer is consistently seen in RCC but not LCC [27]. Interestingly, these biofilms in RCC invade the colonic crypts and predominantly comprise bacteria associated with CRC [28]. Furthermore, Purcell et al. demonstrated that these bacteria are linked to the consensus molecular subtype (CMS) 1, known as the MSI immune subtype, characterized by MSI, CIMP-high, BRAF mutations, and immune cell infiltration [29]. Hence, evaluating the microbiomes of RCC and LCC separately is crucial to comprehend their unique traits.

However, bacterial distributions in the gut don't always directly interact with tumors. For instance, while B. fragilis is abundant in the proximal colon, its IL-17-dependent NF-κB activation triggers a gradient of chemokines from the proximal to distal mucosa, influencing immune cell infiltration and distal colon tumorigenesis [67]. When considered together, the correlation between tumor location and bacterial distribution becomes significant, particularly concerning patient prognosis and treatment strategies.

Among the various molecular events that contribute to CRC development and progression, mutations in key genes involved in cell proliferation, differentiation, apoptosis, and DNA repair are of particular importance [6]. The most common mutations in CRC include mutations in the KRAS gene, TP53 gene, and APC gene, which have been detected in 30-50%, 40-60%, and 60-80% of CRC cases, respectively.

The KRAS gene encodes a GTPase that acts as a molecular switch in the RAS/RAF/MEK/ERK signalling pathway, which regulates cell growth and survival [8]. Mutations in the KRAS gene result in a constitutively active protein that promotes oncogenic transformation and resistance to anti-EGFR therapy [9]. The most prevalent mutation in the KRAS gene is the codon 12-G12D mutation, where Glycine is changed to Aspartic acid, which accounts for 25-30% of all KRAS mutations in CRC. Other common mutations in the KRAS gene include codon 12-G12V, codon 12-G12A, and codon 12-G12S, which represent 15-20%, 10-15%, and 5-10% of KRAS mutations in CRC, respectively. These mutations have similar functional effects and clinical implications, as they impair the GTPase activity of the KRAS protein and confer a poor prognosis and reduced response to anti-EGFR therapy [10].

The TP53 gene encodes a tumor suppressor protein that regulates cell cycle arrest, DNA repair, apoptosis, and senescence in response to cellular stress [11]. Mutations in the TP53 gene lead to the loss of its function and the accumulation of genomic instability and malignant cells [12]. In CRC, various substitution and insertion-deletion mutations have been observed in the TP53 gene, including nonsense and missense mutations that affect the DNA-binding domain and the oligomerization domain of the p53 protein. These mutations are associated with advanced tumor stage, lymph node metastasis, and poor survival in CRC patients [13]. Moreover, some TP53 mutations have been shown to confer a gain-of-function phenotype, which enhances the oncogenic potential and chemoresistance of CRC cells [14].

The APC gene encodes a multifunctional protein that regulates the WNT/β-catenin signaling pathway, which controls cell fate, polarity, and adhesion [15]. Mutations in the APC gene result in the activation of the WNT/β-catenin pathway and the transcription of target genes that promote cell proliferation and invasion [16]. The APC gene also shows pathogenic mutations, such as the c.4348C>T mutation, which causes a premature stop codon and a truncated protein that lacks the β-catenin binding domain. This mutation is found in 10-15% of CRC cases and is associated with familial adenomatous polyposis (FAP), a hereditary syndrome characterized by the development of hundreds of colorectal polyps and a high risk of CRC. Other APC mutations include frameshift and nonsense mutations that occur in the mutation cluster region (MCR) of the gene, which spans exons 14 and 15.

These mutations are detected in 60-70% of sporadic CRC cases and are considered as early events in the adenoma-carcinoma sequence.

These mutations in the KRAS, TP53, and APC genes play a significant role in CRC tumorigenesis and can affect treatment response and prognosis 5. Understanding the molecular diversity and frequency of these mutations is crucial for personalized therapy and better management of CRC patients. Therefore, molecular testing and profiling of these genes are recommended for CRC diagnosis, staging, and treatment selection.

To sum up, the unique spatial arrangement of bacterial species not only acts as a prognostic indicator but can also be a target for enhancing treatment effectiveness. It's crucial to analyse bacterial community structures and comprehend how they adapt within their environments to explore ways of modifying these communities, potentially leading to improved patient outcomes. (Louis, Hold, and Flint 2014))

2.5 Microbiota characterization: from raw sequences to Amplicon sequencing variants ASV (amplicon sequencing variant):

Recent explosion of research in the field of microbiome has led to the development of a wide range of tools, packages, and algorithms to analyse microbiome data. Advances in high-throughput sequencing (HTS), for example, Next Generation Sequencing (NGS) have fostered rapid developments in the area of microbiome research, with massive microbiome datasets are now being generated. The two most commonly used methodologies of Next Generation Sequencing for microbial identification and genotyping are based on gene amplicon/marker genes (e.g., 16S rRNA) and shotgun metagenomics.

2.5.I Gene Amplicon Sequencing:

(i) Overview on Gene Amplicon Sequencing.

Over the past decades, gene amplicon sequencing has been the primary technique utilized to examine the phylogeny and taxonomy of complex microbiomes that were previously considered difficult to characterize. Also, sequencing is useful for the discovery of rare somatic mutations in complex samples (such as tumors mixed with germline DNA).

For bacteria, archaea, fungi, and mycobacteria, there are several specific marker/target genes that have been identified and extensively used for amplicon sequencing. One of these genes is the 16S rRNA gene. The 16S rRNA gene is present in all bacteria and works as a rapid and effective marker for the identification of microorganisms in complex communities. It is sufficiently large for informatics and analysis purposes, at the same time, short enough for NGS, conserved in a way that allows for comparison across different organisms, and divergent enough to identify species unambiguously (Grimm 2019); hence, a huge number of microbiomes have been surveyed by 16S amplicon-based sequencing and it has been employed in numerous works such as the Human Microbiome Project (HMP), Earth Microbiome Project (EMP), and Metagenomics of the Human Intestinal Tract (MetaHIT) (W. Zhang et al. 2023).

More recently, 16S rRNA-based NGS analysis has helped to identify changes in microbial community structures along with its associated alterations in community functions. It helped obtaining a deeper understanding of several gut-associated diseases, including Crohn’s disease, ulcerative colitis, diabetes and gastrointestinal cancers.

Gene amplicon sequencing technique utilizes PCR to target and amplify specific portions of the hypervariable regions of the bacterial 16S ribosomal RNA subunit gene (Gao et al. 2021). There are nine hypervariable regions in the 16S rRNA gene (termed V1–V9), and most of these regions are used for metabarcoding in diverse ecosystems (Lee et al. 2023).

Most of the 16S rRNA-based genotyping protocols use V5–V6, V3–V4, or V4 hypervariable regions to identify and catalogue microbial profiles.

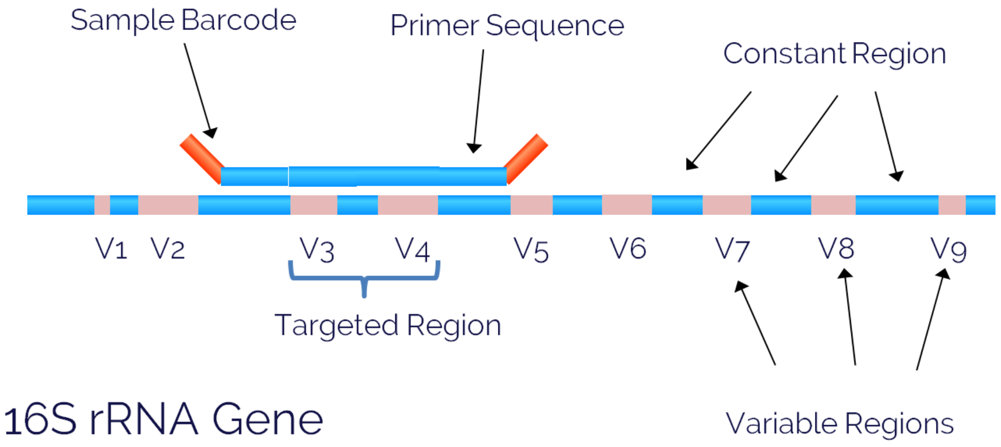


Figure 2.5.I.A 16S rRNA structure

For instance, dominated by Firmicutes and Bacteroidetes, the optimal amplification region for the gut microbiome is V4. Alternatively, the V3 region works better for specimens with abundant Proteobacteria and Actinobacteria (W. Zhang et al. 2023) Other variable regions, including V1–V2 and V3–V4, have been utilized for genotyping archaeal species in complex microbial communities. Therefore, different microbes have their preference for the 16S rRNA gene variable region in amplification sensitivity and nucleotide sequence recognizability. Thus, the choice of primer set design used to amplify the hypervariable regions of the 16S rRNA gene is critical, and can affect the specify and resolution of the sequencing results. Moreover, the length of sequencing reads and sequencing strategy can also impact the accuracy and coverage of the results. In summary, this highly conserved nature of the 16S rRNA gene plays a vital role in cellular function and survival, making it an essential tool for accurately classifying both known and unknown microbial taxa. Furthermore, the relatively short size of the 16S rRNA gene (~1542 bp) makes it easier to sequence even for very large sample sizes.(Gao et al. 2021).

(ii) NGS Platforms and Methodologies.

Currently, the most popular NGS method for single marker/target gene is based on the Illumina platform, Second-generation sequencers, e.g., Illumina’s MiSeq, enable sequencing of amplicons up to 600 bp with high accuracy. This length allows targeting about one to three adjacent variable regions of the 16S rRNA gene using “universal” primers for the conserved regions, while maintaining a significantly reduced sequencing cost in comparison to alternative high-throughput sequencers. The initial steps involve the amplification of target sequences using barcode primer pairs, followed by a subsequent PCR to add sequencing adapters to the amplicons (Abellan-Schneyder et al. 2021). Commonly utilized DNA isolation kits compatible with the Illumina platform encompass Nextera DNA Flex, Nextera XT, and TruSeq DNA PCR-Free, which accommodate diverse genome sizes and necessitate varying quantities of input DNA. After a clean-up step, the purified DNA libraries are prepared for sequencing on the Illumina MiSeq platform, which is predominantly employed for amplicon sequencing due to its ability to generate longer reads (2× 300 bp). The resulting reads are used to analyse similarities and differences between samples with different microbial compositions (e.g., alpha- and beta-diversity).

(iii) Bioinformatics Analysis of Amplicon Sequencing Data.

A variety of bioinformatics tools have been developed to analyse 16S rRNA amplicon sequencing data. These tools typically follow a three-step process: data pre-processing and quality control, taxonomic assignment, community characterization and downstream analysis. One of the key challenges of gene marker-based analysis is distinguishing between true biological signals and sequencing artifacts. To address this challenge, two main tool categories exist:

(i) operational taxonomic unit (OTU)-based (QIIME and Mothur).
(ii) amplicon sequence variant (ASV)-based (DADA2, Deblur, MED,and UNOISE).

Due to recent advances in high-throughput sequencing technologies, OTUs are increasingly being replaced by ASVs, which are un-clustered error-corrected reads (Callahan et al., 2017). After clustering (in case of OTUs) or denoising (in case of ASVs) and feature classification and annotation, the OTU/ASV table with the correspondent abundances is generated. OTUs are increasingly being replaced by amplicon sequence variants (ASVs), which are un-clustered error-corrected reads After clustering (in case of OTUs) or denoising (in case of ASVs) and feature classification and annotation, the OTU/ASV table with the correspondent abundances is generated (Marcos-Zambrano et al. 2021).

OUT Clustering:

In order to minimize the risks of sequencer error in targeted sequencing, clustering approaches were initially developed. Clustering approaches are based upon the idea that related/similar organisms will have similar target gene sequences and that any rare sequencing errors will have a minimal, if not none, impact to the consensus sequence for these clusters, known as operating taxonomic units (OTUs). These clusters are often being generated using a similarity threshold of 97% sequence identity.

This approach poses the risk as it may result with several similar species being clustered into a single OTU, with their unique identifications being lost within the abstraction of grouping. Alternatively, some have tried the approach of requiring extremely high levels of sequence identity to minimize the risk of losing diversity to clustering, with thresholds closer to 100% being used, but this created a significant risk of identifying sequencing errors as new species and false diversity.

While methods based on OTU clustering attempt to blur similar sequences into an abstracted consensus sequence, reducing the impact of potential sequencing errors within a set of reads, the Amplicon Sequence Variant (ASV) approach attempts to go the opposite direction (Chiarello et al. 2022).

The ASV approach

Amplicon sequencing variants (ASVs) are an alternative approach to operational taxonomic units. It will start by identifying which exact sequences were read and quantifying the frequency of each sequence.

These data will be integrated with an error model specific to the sequencing run, which enables the comparison of similar reads to determine the likelihood that a given read at a given frequency is not a result to sequencer error. (Jeske and Gallert 2022).

While there is also a risk of clustering ASVs from different species into the same cluster when using broad distance thresholds, the risk of splitting a genome into separate ASVs is a more significant concern, this risk was investigated by analysing the intragenomic variation of 16S rRNA genes from bacterial genomes.

The number of ASVs increased with the number of copies of the 16S rRNA gene in a genome (Schloss 2021).

2.5.II Analysis Pipelines.

As of January 23, 2020, the words “amplicon” and “metagenome” were mentioned more than 200,000 and 40,000 times in Google Scholar, respectively.

Analysis pipeline” refers to specialized software, programs, algorithms or scripts that are needed to convert raw sequencing data into biologically meaningful information by combining several or even dozens of software programs and tools organically in a certain order to complete the analysis task.

While different bioinformatic pipelines are available in a rapidly changing and improving field, users are often unaware of limitations and biases associated with individual pipelines and there is a lack of agreement regarding best practices (Prodan et al. 2020). So, it is important to select a workflow that is appropriate for the research question and the type of microbiome being analysed.

The choice of bioinformatics pipeline for analysing 16S rRNA gene sequencing data from the gut microbiome can significantly impact the downstream statistical analysis results. For instance, (Prodan et al. 2020; Szopinska-Tokov et al. 2023) compared 5 workflows for microbiome characterization, different pipelines resulted in variations in the number of ASVs/OTUs and genera obtained, as well as in the case versus control comparison results. Currently the best-practice for amplicon analysis-NGS based are performed with the Shell environment and R language (Liu et al. 2021).

(iv) Quality control

The first step in the analysis pipeline. It involves quality checking, adapter removal, filtering and trimming to remove artifacts and non-biological sequences, low-quality and contaminant sequencing reads resulting from sample impurities or inadequate samples preparation steps. Many quality control software packages use PHRED algorithm score to assess the base quality.(Reitmeier et al. 2021). A **Phred quality score** is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing. It was originally developed for the computer program Phred to help in the automation of DNA sequencing in the Human Genome Project.

(v) The taxonomic assignments.

A key step in the 16S rRNA sequencing data analysis pipeline, and the most important output file from amplicon analysis pipeline.

Taxonomic assignments refer to the process of classifying an organism or a gene sequence to a particular taxon based on their similarity to known sequences in a reference database.(Sharma et al. 2012)

As mentioned before, an OTU-based analysis, first clusters sequences into different OTUs and then performs taxonomic assignment. On the other hand, ASV-based methods utilize a denoising approach to infer the biological sequences in the sample before the introduction of amplification and sequencing errors. This allows to resolve sequences differing by as little as a single nucleotide. Therefore, an ASV-based analysis is able to provide a higher-resolution taxonomic result.

Currently, the 16S-based microbial taxonomy profiling and species recognition are still limited by the shortage of reference databases. Some of the most commonly used taxonomic reference databases are: the NCBI RefSeq database (National Centre for Biotechnology Information) which provides high-resolution annotations at the species level, while Greengenes; A database of the work of hundreds of scientists and Silva; for quality checked and aligned rRNA sequence data, have the advantage of comprehensiveness of taxonomy units (W. Zhang et al. 2023), and the RDP database (Ribosomal Database Project). The quality of the reference database can significantly impact the accuracy of taxonomic assignment. Thus, reference databases are constantly being updated, and it is important to use databases that are well-maintained, have a high level of coverage, and are appropriate for the type of microbiome being studied.

(vi) Within Sample Alpha Diversity.

Alpha diversity provides an idea of the diversity of species within a particular sample. This metric is often used as a biomarker in disease association studies (Prehn-Kristensenetal.,2018), and used as a check of sample quality (Schlossetal.,2009). Typically, alpha diversity metrics can be distinguished into two types: richness-and evenness-measures; Chao1 being the most used richness metric, and Shannon the most used evenness metric. The differences in alpha diversity among or between groups could be statistically evaluated using Analysis of Variance (ANOVA), Mann-Whitney U test, or Kruskal-Wallis test.

(vii) Between Sample (Beta) Diversity

Beta diversity represents the diversity of species across samples, commonly used to find clusters of similar samples.

Typically, this feature is calculated in the exploratory analysis, as it provides a first impression on which taxa are important to distinguish samples, also on how microbial compositions are related to environmental and personal meta data. Beta diversity analysis is expressed as a distance matrix calculation on relative ASV/ OUT abundance, which serves as an input for visual exploration of sample divergence and similarity.

Often occurring distance metrics are: (weighted) UniFrac, Jaccard, Bray-Curtis and Jenson-Shannon (Oliveira et al., 2018; Chong et al., 2020; Shamsaddini et al., 2020). However it is important to note that none of these measures account for the compositional nature of the data, as there are newer methods that have been developed and designed to be “compositionally aware” and can better resolve microbiomes associated with phenotype (Martino et al. 2022)

(viii) Differential Abundance

With differential abundance analysis, OTUs/ASVs that differ significantly between samples, cohorts or populations are identified using statistical hypothesis testing. In doing so, taxa can be related to a certain response (e.g., disease state, growth process). (Peeters et al. 2021).

3. Purpose of the study

In this study, an overview of the aforementioned purviews is provided by covering the three main areas of:

(i) Identification of CRC-associated bacteria.

(ii) Comparison between cancer tissue and matched healthy tissue-associated microbiota.
(iii) Investigate the correlation between CRC mutational profile and the significantly differentiated microbiome.

4. Materials and Methods

4.1 Collection of biological samples and tissue processing (DNA extraction and amplification of 16SrRNA gene):

In order to detect differences specific to the cancer-associated microbiome, samples taken directly from the tumor microenvironment are preferable to bulk stool samples, at least at the initial characterization phase. Minimising unwanted variation in an experiment is dependent on good experimental design.

Moreover, the use of traditional case-control studies of the colon cancer microbiome makes it difficult to control for all of the external effects on the microbiome. For example, Diet and the variations of host genetics are two of the strongest influences on the composition of gut microbiota.(Burns et al. 2014).

The microbiome data used in this study was generated previously and is described exhaustively in (Roelands et al. 2023)***.***

Briefly, snap-frozen tumor and healthy colon tissue were extracted from patient tumor tissue and matched normal samples were collected from colon cancer patients who underwent surgical resection of the primary tumor between 2001 and 2015 at Leiden University Medical Centre, resulting in a final cohort of 348 patients with available clinicopathological and survival data. Tissue processing involved sectioning tumor and healthy tissue samples, removing non-target tissue, and collecting frozen tissue for DNA and RNA extraction. DNA and RNA extraction was performed using the QIAGEN AllPrep DNA/RNA Mini kit, and the samples were stored at -80 °C. The Hypervariable regions V3–V4 of 16S rRNA gene were amplified with PCR using the amplicon primers with Illumina adaptors:
Forward:
5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNG-GCWGCAG'3
Reverse:
5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACH-VGGGTATCTAATCC'3.

PCR was performed in a 25-μl reaction mixture containing primers, template DNA, and Hot Master Mix, followed by amplification cycles and confirmation of PCR products by electrophoresis. The amplicons were purified and multiplexed using a dual-index approach, and the concentration was determined before sequencing on the Illumina MiSeq platform. Samples were multiplexed using a dual-index approach with the Nextera XT Index kit, determining amplicon concentration with the Qubit HS dsDNA assay kit, pooling to achieve an equimolar library concentration, and performing paired-end sequencing on the Illumina MiSeq platform.

The microbiome cohort in this study consisted of 246 patients whose matched tumor and healthy colon tissues were sequenced for the 16S rRNA gene. This cohort is referred to in the study as the AC-ICAM246 cohort.

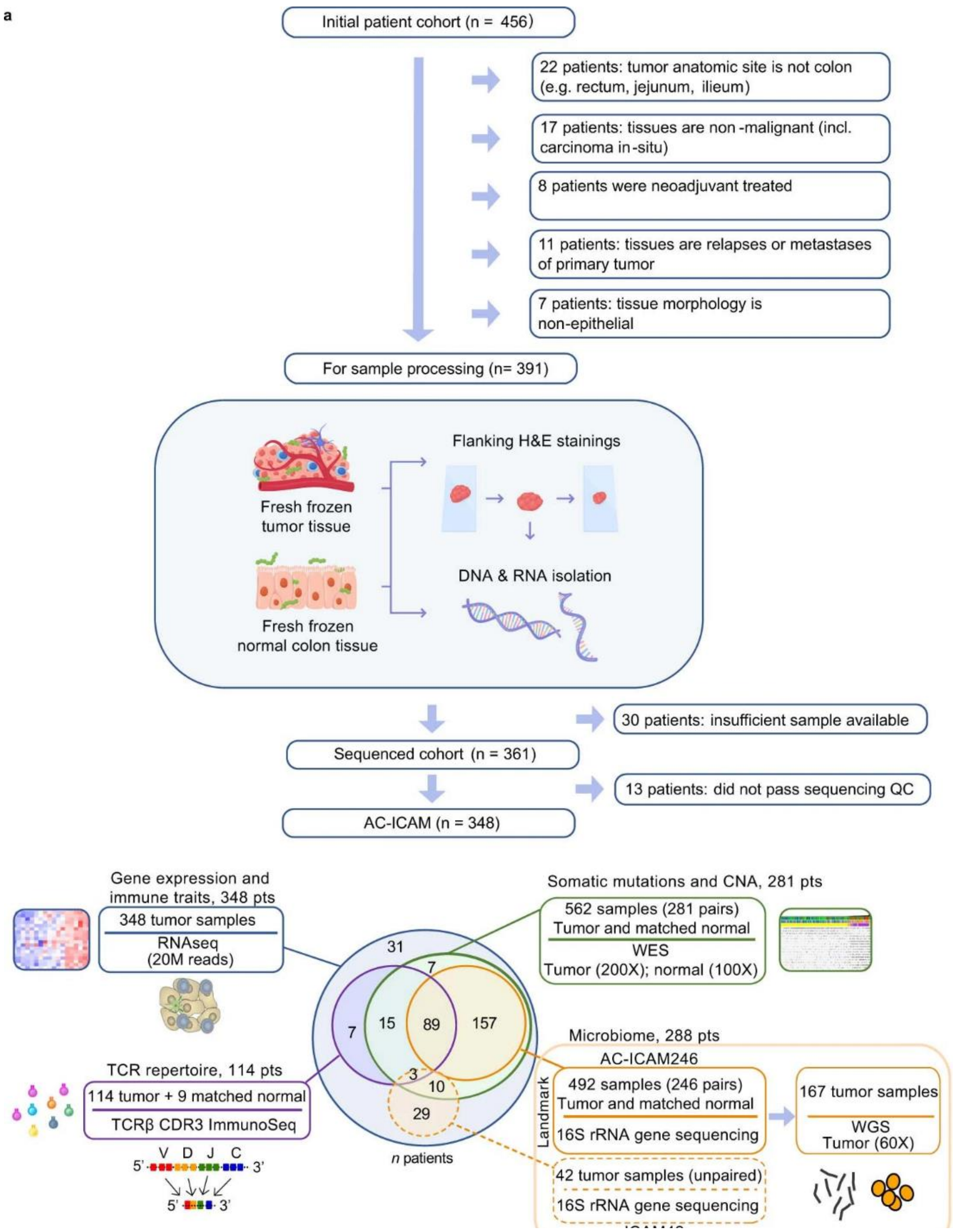

a
Initial patient cohort (n = 456)
22 patients: tumor anatomic site is not colon (e.g. rectum, jejunum, ileum)
17 patients: tissues are non -malignant (incl. carcinoma in-situ)
8 patients were neoadjuvant treated
11 patients: tissues are relapses or metastases of primary tumor
7 patients: tissue morphology is non-epithelial
For sample processing (n= 391)
Fresh frozen tumor tissue
Fresh frozen normal colon tissue
Flanking H&E stainings
DNA & RNA isolation
30 patients: insufficient sample available
Sequenced cohort (n = 361)
13 patients: did not pass sequencing QC
AC-ICAM (n = 348)
Gene expression and immune traits, 348 pts
348 tumor samples
RNAseq (20M reads)
TCR repertoire, 114 pts
114 tumor + 9 matched normal
TCRβ CDR3 ImmunoSeq
V D J C
5' 3'
5' 3'
31
7
7
15
89
157
3
10
29
n patients
Somatic mutations and CNA, 281 pts
562 samples (281 pairs)
Tumor and matched normal
WES
Tumor (200X); normal (100X)
Microbiome, 288 pts
AC-ICAM246
Landmark
492 samples (246 pairs)
Tumor and matched normal
16S rRNA gene sequencing
167 tumor samples
WGS
Tumor (60X)
42 tumor samples (unpaired)
16S rRNA gene sequencing

Figure 2.5.II.A AC-ICAM study design. **a**, Visual representation of exclusion criteria and number of excluded samples from the 456 available samples in the LUMC biobank, followed by overview of tissue processing and genomic profiling of fresh-frozen tumor and matched normal colon tissue samples. Samples of a total of 348 colon cancer patients were included in AC-ICAM. Number of profiled samples and technical specifications are indicated for each platform, including RNA Sequencing (RNA-Seq), Whole-Exome Sequencing (WES), TCR sequencing (immunoSEQ TCRβ assay) and 16 S rRNA gene sequencing to profile the microbiome. AC-ICAM246 is a subset of AC -ICAM with tumor–normal matched rRNA 16 S microbiome data, while AC-ICAM42 only has tumor samples with 16 S rRNA gene sequencing. Venn diagram reflects overlap in number of patients between the different platforms applied. **b**, Summary of patient characteristics of colon cancer cohort (*n* = 348). Number in pie chart indicates number of patients in each category.

(Roelands et al. 2023)

4.2 Microbiome data processing and computational analysis:

4.2.I Data retrieving and pre-processing:

the microbiota data with number (SRP426032) were retrieved from the NCBI Sequence Read Archive (SRA) database: project accession number (PRJNA941834; 16S) using "efetch 14.6" and "prefetch 2.11.3" command from the SRA-toolkit version 3.0.6.

This dataset contained 16S rRNA gene amplicon sequencing data of 60 samples from 30 patients, each comprising one tumor and one normal sample.

The study cohort for this project were chosen randomly from the cBioportal for interactive data exploration (Sidra-LUMC AC-ICAM dataset; https://www.cbioportal.org/)".

The raw sequenced data were converted to fastq format (or "demultiplexed") using "fasterq_dump" software (parameter: split-3). The overall Sequencing quality was evaluated using "FastQC v0.12.1" reports.

FastQC gives general quality metrics about sequenced reads. providing information about the quality score distribution across reads, per base sequence content (%A/T/G/C), adapter contamination and overrepresented sequences. The output from FastQC, after analyzing a FASTQ file of sequence reads, is an HTML file that may be viewed in browser. The interpretation of these plots can vary depending on the nature and context of your sequencing data. Comprehensive quality assessment report generated using "MultiQC version 0.4". to summarize all fastqc report for all samples.

All reads from this type of library are expected to be nearly identical, expected results are:

- Extremely biased per base sequence content.
- Extremely narrow distribution of GC content.
- Very high sequence duplication levels.
- Abundance of over-represented sequences.
- in cases where the PCR target is shorter than the read length, the sequence will read through into adapters.

"Cutadapt version 1.18" was employed to remove adapters, and the quality assessment was repeated using FastQC and MultiQC and further summarized using 'fastqcr' R-package, to ensure the effectiveness of the adapter removal process.

The data consisted of paired-end FASTQ reads with a median sequencing depth of 139,428,563 reads per sample."

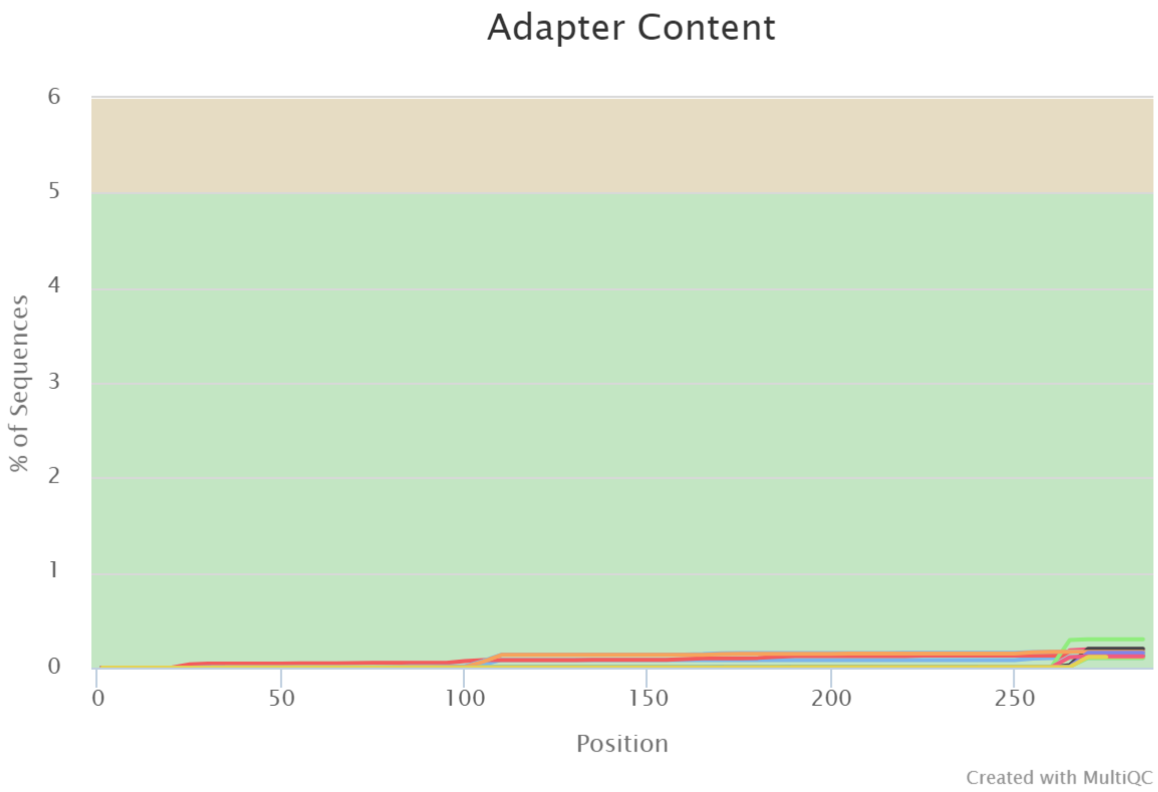


Figure 4.2.I.A Adapter content plot exported from MultiQC report showing the process of adapter removal from the reads.

4.2.II ASV-based pipeline.

In this project, we used DADA2 R-package (Denoising Amplicon Data with Adaptive Removal of Chimeras and Dereplication) v3.16 (Callahan et al., 2016) on R v3.6.3 Sequences were processed according to following the general tutorial available on the GitHub of the software (https://benjjneb.github.io/dada2/tutorial.html, November 2020), the tutorial instructions for the DADA2 workflow for paired end Illumina Miseq data and the Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses (https://doi.org/10.12688/f1000research.8986.2) (Callahan BJ, Sankaran K, Fukuyama JA et al)

the starting point to DADA2 pipeline is sequences that meet the following criteria:

- Samples have been demultiplexed, i.e. split into individual per-sample fastq files.

- Non-biological nucleotides have been removed, e.g. primers, adapters, linkers, etc.
- If paired-end sequencing data, the forward and reverse fastq files contain reads in matched order.

The data input to DADA2 as paired-end trimmed for adapters fastq files and the end product of this package is an amplicon sequence variant (ASV) table.

After the package successfully loaded and performed some string manipulation to get matched lists of the forward and reverse fastq files, we start with visualizing the quality profile of the forward and reverse reads.

plotQualityProfile(fnFs[1:2]) for forward reads and plotQualityProfile(fnRs[1:2]) for reverse reads

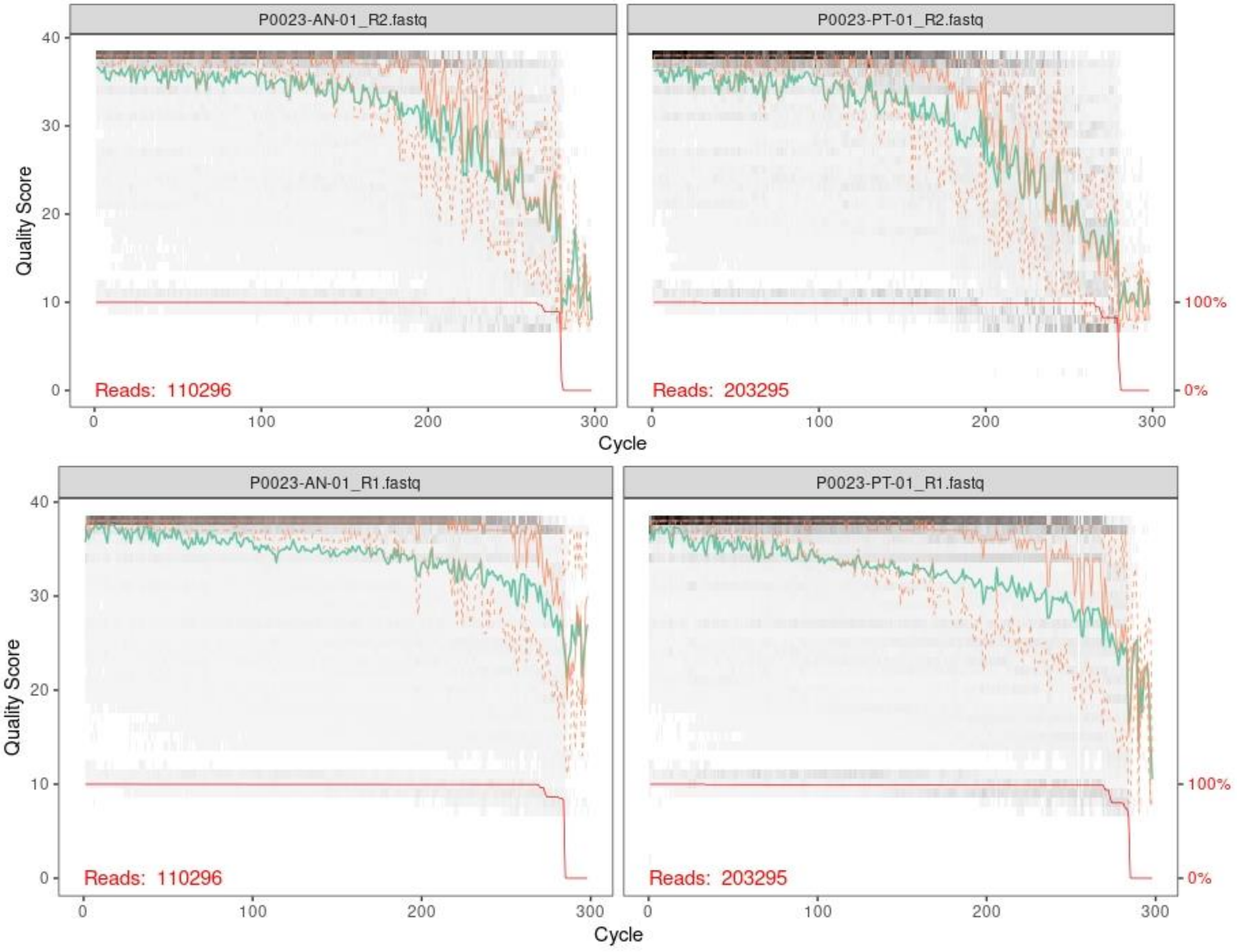
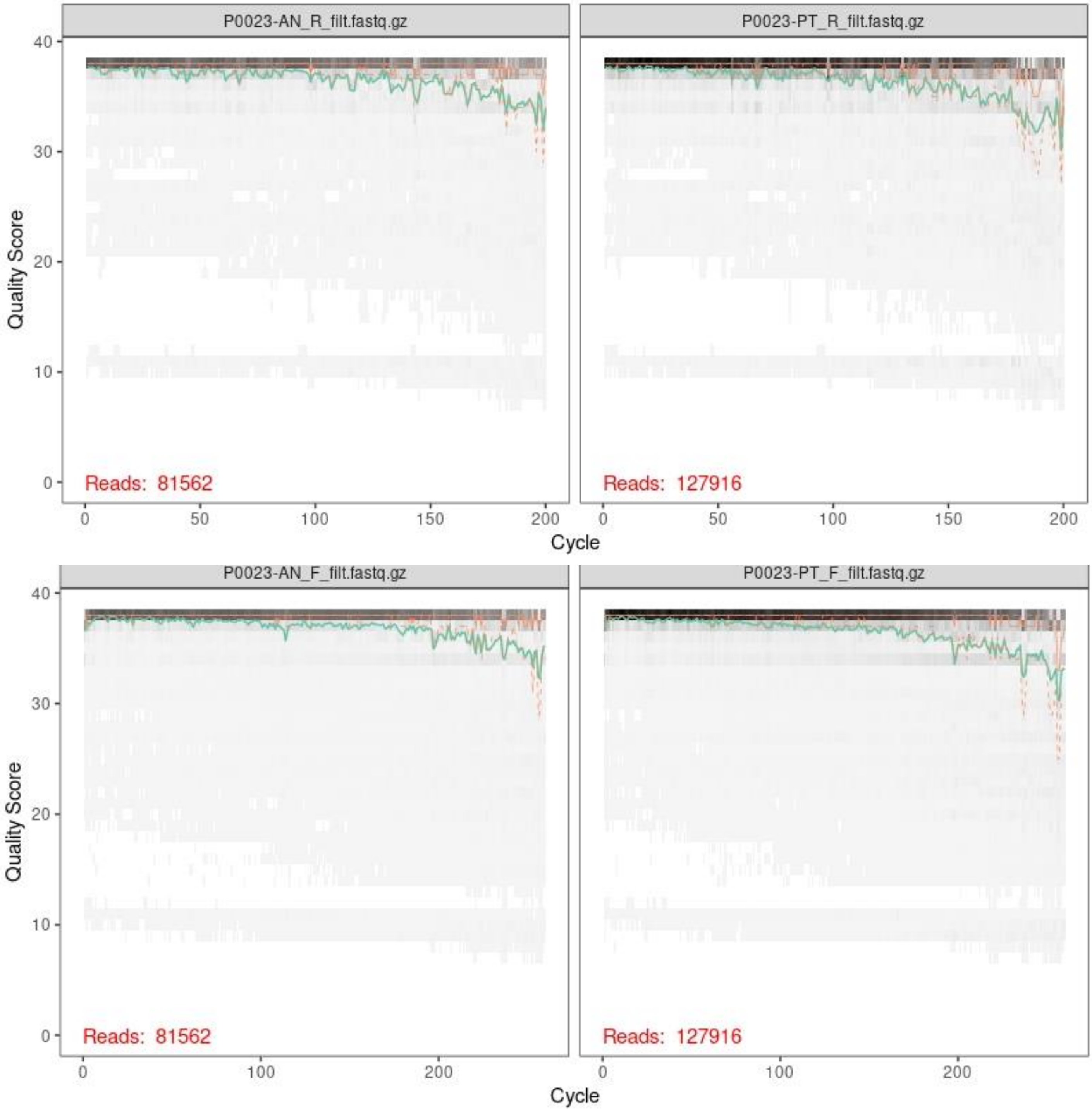


Figure 4.2.II.A a heatmap depicting the frequency of each quality score at each base position is presented in grayscale. The X axis represents the read length. The y axis represents the PHRED quality score. The green line represents the average quality score at each position. while the orange lines indicate the quartiles of the quality score distribution. The red line shows the scaled proportion of reads that extend to at least that position and the quartiles of the quality score distribution by the orange lines. The forward reads maintain high quality throughout, while the quality of the reverse reads drops significantly at about position 200, which is common in Illumina sequencing.

Filter and trim.

Specifically, paired 300-bp reads were trimmed of the initial five low-quality bases. The Forward and reverse reads with more than 2 estimated errors, filtered and truncated at the 3’ end, where read quality dropped below a quality score of 2 (TrunQ = 2). Based on the quality plots we choose to truncate the forward reads at position (260) and the reverse reads at position (200), This stringent filtering step ensured the quality and integrity of the data for subsequent ASV inference, yielding a high-quality dataset for downstream microbiota analyses.



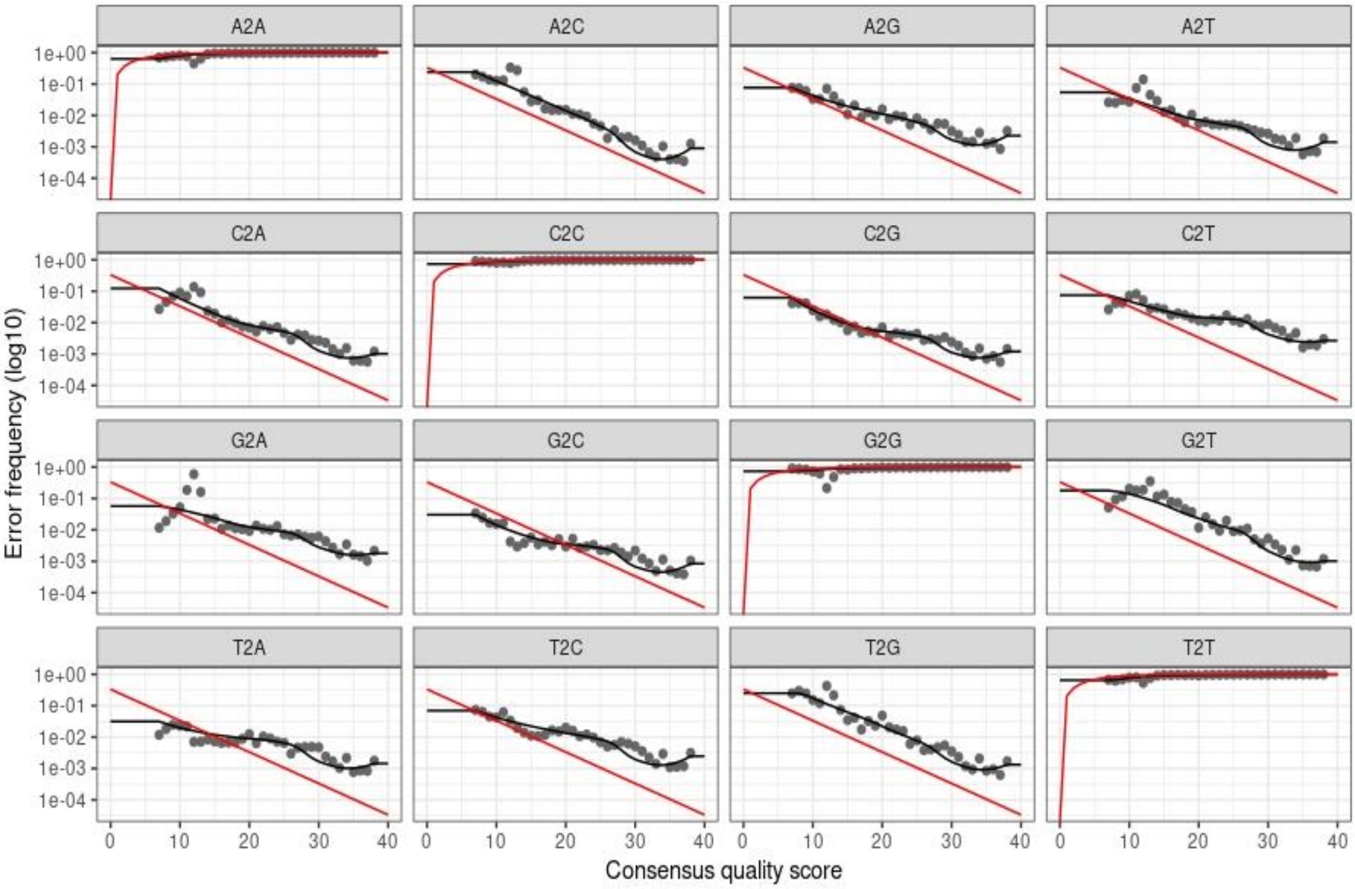
ASV Inference and error rate estimation

on the filtered fastq sequence data files, a dereplication process was performed, dereplication combines all identical sequencing reads into “unique sequences” with a corresponding “abundance”: the number of reads with that unique sequence to eliminate redundant information and reduces the memory requirements for downstream analysis.

Subsequently, separate error rate estimations were conducted for forward and reverse reads, using 264,837,560 total bases in 1018,606 reads from 6 samples and 203,721,200 total bases in 1018,606 reads from 6 samples, respectively.

These error rate estimations employed a novel unsupervised learning approach that involved alternating sample inference with parameter estimation until both were mutually consistent. As in many machine-learning problems, the algorithm must begin with an initial guess, for which the maximum possible error rates in this data are used;

```
264837560 total bases in 1018606 reads from 6 samples will be used for
learning the error rates.
203721200 total bases in 1018606 reads from 6 samples will be used for
learning the error rates.
```



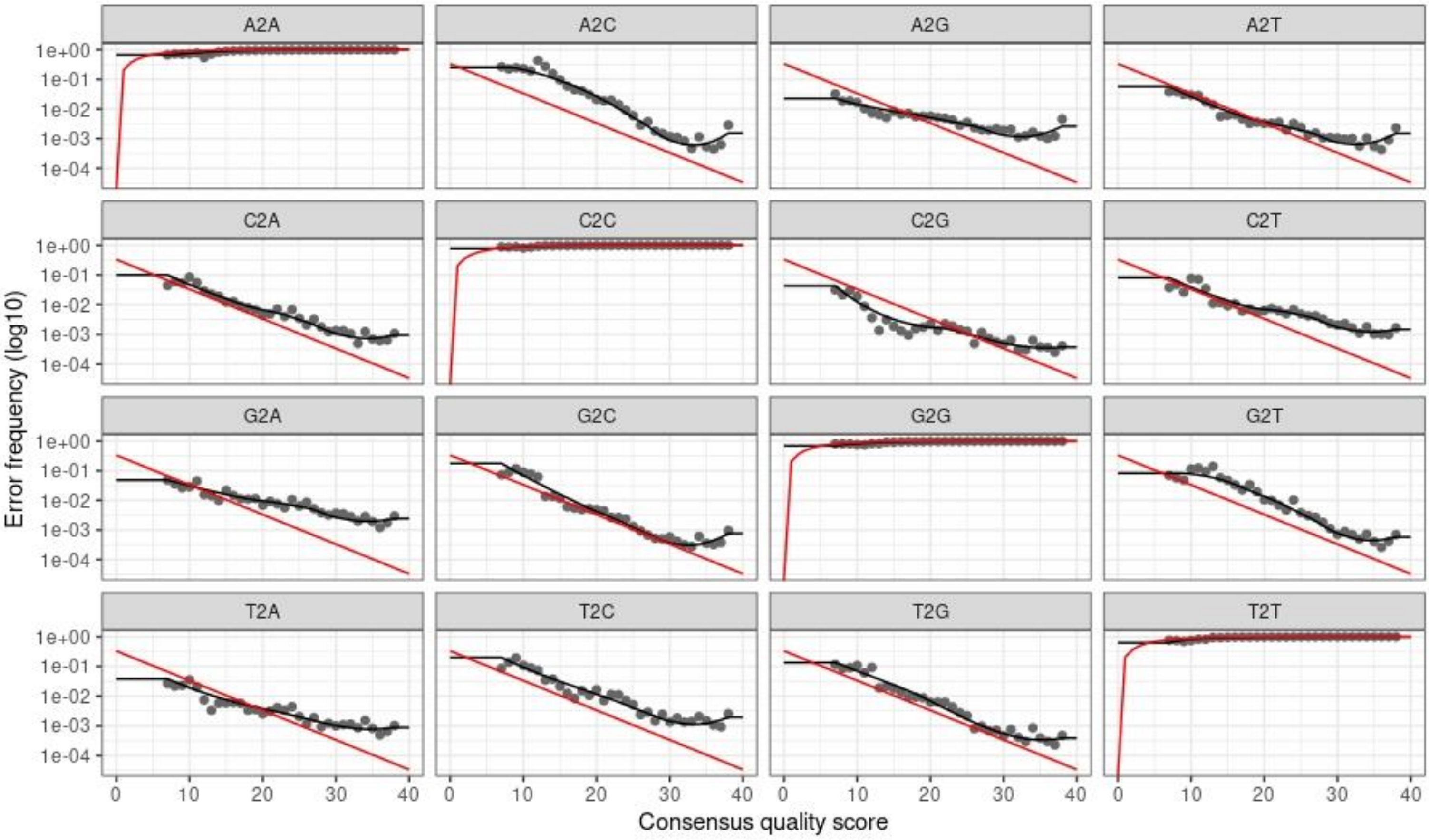


Figure 4.2.II.B illustration of The error rates for each possible transition (A to C, A to G, …). The points represent the observed error rates for each consensus quality score. The black line depicts the estimated error rates after the machine-learning algorithm converges. The red line indicates the error rates projected under the standard definition of the Q-score. Here, the estimated error rates (black line) closely match the observed rates (points), and the error rates decline with increasing quality as anticipated.

Everything appears to be in good order, and we proceed with certainty.

After learning error rates, we applied the code sample inference algorithm to the filtered and trimmed sequence data (Callahan et al. 2016). When inspected the dada_class object which describes DADA2 denoising results for the first sample:

```
dada-class: object describing DADA2 denoising results
344 sequence variants were inferred from 13626 input unique sequences.
Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16
dada-class: object describing DADA2 denoising results
230 sequence variants were inferred from 14199 input unique sequences.
Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16
```

Construct sequence table and Chimera Removal.

ASVs were predicted and merged to obtain full denoised sequences. using a minimal overlap of 60 bases, which was necessary to maintain sufficient overlap between the forward and reverse reads, considering the biological length variation expected for the primer set used.

By default, merged sequences are only output if the forward and reverse reads overlap by at least 12 bases and are identical to each other in the overlap region and is performed by aligning the denoised forward reads with the reversed-complement of the corresponding denoised reverse reads and then constructing the merged “contig” sequences. To inspect the merging process:

```
  abundance forward reverse nmatch nmismatch nindel prefer accept
1       56523       1       1     33         0      0      1   TRUE
2       16905       2       3     58         0      0      1   TRUE
3       15120       3       2     58         0      0      1   TRUE
4       14234       4       4     38         0      0      2   TRUE
5       11582       5       7     58         0      0      2   TRUE
6       10898       6       6     38         0      0      2   TRUE
7       10535       7       8     58         0      0      2   TRUE
```

the next step in the DADA2 pipeline is to remove chimeric sequences.

Chimeras are sequences that are composed of two or more distinct microbial genomes. They can arise from a variety of sources, including PCR artifacts, contamination, and horizontal gene transfer. The presence of chimeras can distort the composition of an amplicon sequence variant (ASV) table, leading to inaccurate estimates of microbial diversity and abundance. Chimeras were removed using the consensus method with the *removeBimeraDenovo()* function.

In this dataset we identified and removed a significant number of chimeric sequences (146,943 out of 155,368 input sequences), representing approximately 93% of the original chimeras. This is a relatively high level of chimera removal, but it is not uncommon for datasets that have been sequenced from environmental samples. The number of ASVs remaining after chimera removal was also relatively low (8,425), representing a reduction of approximately 44% of the original ASVs. This reduction is again not unusual, as chimera removal can often result in a significant reduction in the number of sequences identified.

Before moving forward in the downstream analysis we’ll look at the number of read that made it through each step in the pipeline, and assess sample-level variations in sequencing depth or read count distribution. No samples were excluded from the analysis.

```
input filtered denoisedF denoisedR merged nonchim
P0023-AN 110296    81562     80868     81136  78482   73047
P0023-PT 203295   127916    125417    126819 119467   97731
P0030-AN 257782   172904    167126    169906 146405   92989
P0030-PT 263334   183971    176739    180186 159679   98249
P0045-AN 140118   100493     99209     99864  94860   86374
P0045-PT 491350   351760    348157    349986 333858  285694
```

Taxonomy assignation.

Following constructing sequence table, ASVs were classified using the RDP naïve Bayesian classifier method (Wang et al. 2007) implemented in the DADA2 using the assignTaxonomy() function and the SILVA 16S rRNA gene data base (“Release 138.1,” n.d.), For the RDP training set.
This database enabled classification of ASVs across six taxonomic levels: Kingdom, Phylum, Class, Order, Family, and Genus. Remarkably, all ASVs were successfully classified to the genus level, providing a comprehensive overview of the microbial composition within the samples. the dada2 package also implements a method to make species level assignments based on exact matching between ASVs and sequenced reference strains (Edgar 2018). To achieve species-level resolution, we employed the *addSpecies()* function, utilizing the SILVA species assignment database (McLaren and Callahan 2021) (Release 138.1) as a reference.
Only in instances, where a query sequence exhibits a perfect match (100% identity) with a reference sequence is a species-level assignment confidently made. This rigorous approach ensures the elimination of ambiguous matches, mitigating the risk of miss identification. Emerging scientific evidence strongly supports the notion that exact matching with amplicon sequence variants (ASVs) represents the most reliable and appropriate method for species-level assignment within high-throughput 16S amplicon datasets.
This additional step enabled more precise identification of the microbial species present in the samples. using DADA2’s default parameter.

Head of ASV taxonomy table:

```
Kingdom    Phylum          Class               Order            Family            Genus               Species
[1,] "Bacteria" "Proteobacteria" "Gammaproteobacteria" "Enterobacterales" "Enterobacteriaceae" "Escherichia-
Shigella" NA
[2,] "Bacteria" "Bacteroidota"   "Bacteroidia"        "Bacteroidales"    "Bacteroidaceae"     "Bacteroides"        "fragilis"
[3,] "Bacteria" "Proteobacteria" "Alphaproteobacteria" "Rickettsiales"    "Mitochondria"      NA                 NA
[4,] "Bacteria" "Proteobacteria" "Alphaproteobacteria" "Rickettsiales"    "Mitochondria"      NA                 NA
[5,] "Bacteria" "Firmicutes"     "Clostridia"         "Oscillospirales"  "Ruminococcaceae"    "Faecalibacterium"
"prausnitzii"
[6,] "Bacteria" "Proteobacteria" "Alphaproteobacteria" "Rickettsiales"    "Mitochondria"      NA                 NA
```

Combine data into a phyloseq object.

Phyloseq (McMurdie and Holmes 2013) is an R package used to import, store, analyse and graphically display complex phylogenetic sequencing data that has already been clustered into amplicon sequencing variants and assigned to a taxonomy reference.

This package leverages many of the tools available in R for ecology and phylogenetic analysis (vegan, ade4, ape, picante), while also using advanced/flexible graphic systems (ggplot2) to easily produce publication-quality graphics of complex phylogenetic data.

The “phloseq” package uses a specialized system of S4 data classes to store all related phylogenetic sequencing data as a single, self-consistent, self-describing experiment-level object making it easy to use R for efficient interactive and reproducible analysis of amplicon count data jointly with important sample covariate.

The full suite of data for this study; the sample-by-sequence feature table, the samples metadata, the taxonomy table are combined into a single object for storing and further analysing the microbiome data.

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 8227 taxa and 60 samples ]
sample_data() Sample Data:       [ 60 samples by 2 sample variables ]
tax_table()   Taxonomy Table:    [ 8227 taxa by 7 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 8227 tips and 8225 internal nodes ]
refseq()      DNAStringSet:      [ 8227 reference sequences ]
```

Phyloseq filtering.

To further refine the taxonomic classification of the microbiome data, the *subset_taxa()* function from the phyloseq package was employed. This function was used to filter the taxonomic classification to include only bacteria, excluding other domains at the Kingdom taxonomic level. Additionally, sequences from chloroplasts and mitochondria were removed from the analysis by filtering at the Order and Family taxonomic levels, respectively. This step helped to eliminate potentially contaminating sequences from the analysis, ensuring that the results were more representative of the true microbial community present in the samples.

Contamination assessment.

After refining the taxonomic classification, a list of unique genera was extracted using the *get_taxa_unique()* function. This list was then compared to a list of known contaminant genera, denoted as “likely_contaminant”, to identify potentially problematic sequences that may require further investigation or removal.

The “genera_in_contaminant_list” variable was created to flag genera that were present in both the refined dataset and the contaminant list, highlighting potential contaminants that could bias the results of downstream analyses. we used a list of microbial taxa that are typically found in negative blank reagents,as described by (Salter et al). This list has previously been curated and annotated by (Poore et al).

Following the refinement of taxonomic classification, the *tax_table()* function from the phyloseq package was used to assess the completeness of taxonomic annotation and identify ASVs that lacked phylum-level annotation, the presence of ASVs without phylum-level annotation suggests that the taxonomic classification may not be comprehensive, potentially impacting the accuracy of downstream analyses;

Number of ASVs without phylum-level annotation: 0

4.2.III Statistical Analysis

First, we assessed sample-level variations in sequencing depth (read count distribution).

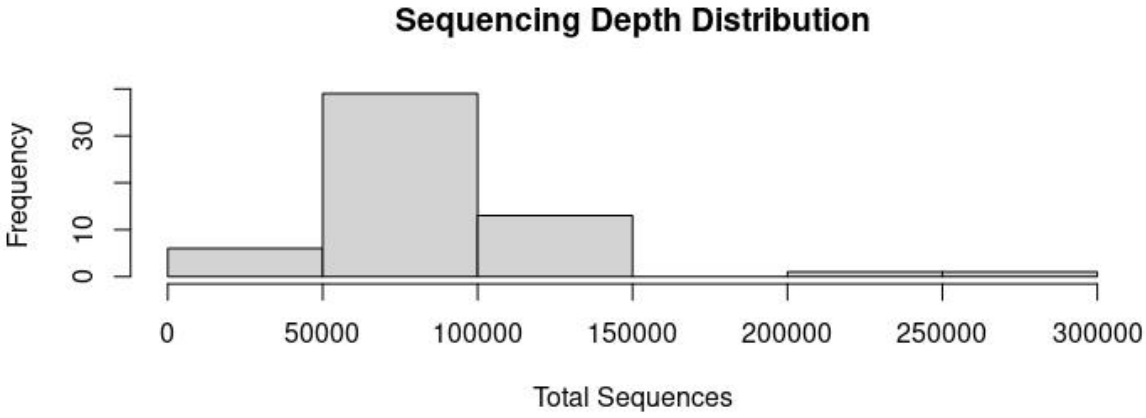


Figure 4.2.III.A X-axis (Total Sequences): This represents the total number of sequences observed in the study. It ranges from 0 to 300,000, Y-axis (Frequency): This represents the frequency of the total sequences observed. It ranges from 0 to over 30, There are three bars in the graph, each representing the frequency of sequences at different total sequences. The first bar shows a small frequency of sequences at around 50,000 total sequences. The second bar, which is significantly taller, indicates that the highest frequency of over 30 occurs at around 100,000 total sequences. This suggests that most sequences in the study were observed at this depth The third bar shows another small frequency of sequences occurring at approximately between 150,000 and 200,000 total sequences.

Sequencing depth is the number of sequences that we obtained from each sample, and it can affect the quality and accuracy of the analysis. We wanted to make sure that our samples had enough sequencing depth to capture the microbial diversity and abundance. To do this, we plotted the distribution of the total number of sequences per sample, as shown in Figure 4.2.III.A, most of the samples had around 100,000 sequences, which is considered sufficient for 16S rRNA gene sequencing. Some samples had lower or higher sequencing depth, but they were not very different from the majority. This means that our samples had relatively consistent and adequate sequencing depth.

To further explore the characteristics of the microbial communities, we conducted analyses of ASV prevalence and abundance across different taxonomic levels.

The **dplyr** package was employed to calculate prevalence, defined as the number of samples in which a given ASV was present, and total abundance, representing the sum of its abundances across all samples. This information was integrated with taxonomic metadata using *cbind().* Phylum-level summaries were generated using *plyr::ddply().*

Alpha diversity within samples was assessed by observed richness index on ASVs (sum of unique AVSs per sample), Shannon's index was also calculated on the raw ASV counts table for the dataset of both tumor and healthy samples, Simpson which is more dependent on highly abundant ASVs and less sensitive to rare ASVs, The

Chao2 (Chao 1987) an abundance-based richness estimator that is sensitive to rare ASVs was also included as complimentary measures, Incendies were read using R package “Vegan (v.2.5-6)

Multiple ordination techniques were employed to comprehensively assess relationships between samples and potential clustering, normalized abundance data was used for all analyses to minimize compositional biases, the *phyloseq_to_deseq2()* function converted a phyloseq object containing the microbiome count data into *DESeqDataSet* object.

DESeq2 was employed to estimate size factors for each sample to account for sequencing depth differences, using the calculated geometric means for each ASV using a custom function (*gm_mean*) and supplied as prior information.

Geometric mean is often preferred for microbiome data as it’s less sensitive to outliers and more robust to zero-inflation compared to other methods like total sum scaling. this step addresses compositional bias, ensuring comparisons are based on relative proportions rather than absolute counts.

Beta diversity was evaluated using weighted, unweighted UniFrac distances and Bray-Curtis dissimilarity which focuses on compositional differences in abundance, disregarding phylogenetic information.

Non-metric multidimensional scaling (NMDS) ordinations were generated for each distance metric to visualize community dissimilarities in two-dimensional space. Additionally, PCA was performed to provide a complementary perspective on sample relationships and variance structure.

Differential abundance analysis using *DESeq()* function from DESeq package was performed to identify ASVs that significantly differ in abundance analysis, the significance threshold was set at a p value of 0.01 and the log fold change <1. The obtained results were further visualized using ggplot2. The results were visualized to identify taxa exhibiting significant changes in abundance.

5. Results

There was a total of 8425 ASVs in the unrarefied dataset of 8425 samples, based on 39870–285694 valid sequence reads per sample.

The average and median read counts were 85456.7 and 73906.5, respectively.

The most abundant phylum in the dataset is Firmicutes with a prevalence of 0.76, which means that Firmicutes was detected in an average of 76% of the samples, followed by Bacteroidetes the average prevalence of Bacteroidetes is 0.69. Proteobacteria, the average prevalence of Proteobacteria is 0.32, which means that Proteobacteria was detected in an average of 32% of the samples and Actinobacteria was detected in an average of 2% of the samples. These phyla are typically the most dominant in the gut microbiome of. Examining the prevalence of genera revealed a diverse landscape. While some, like Bacteroides and Faecalibacterium, exhibit moderate prevalence across samples, others like Lachnoclostridium and Blautia stand out with consistently high presence. This suggests the presence of key players alongside potentially specialized taxa within the microbial community.

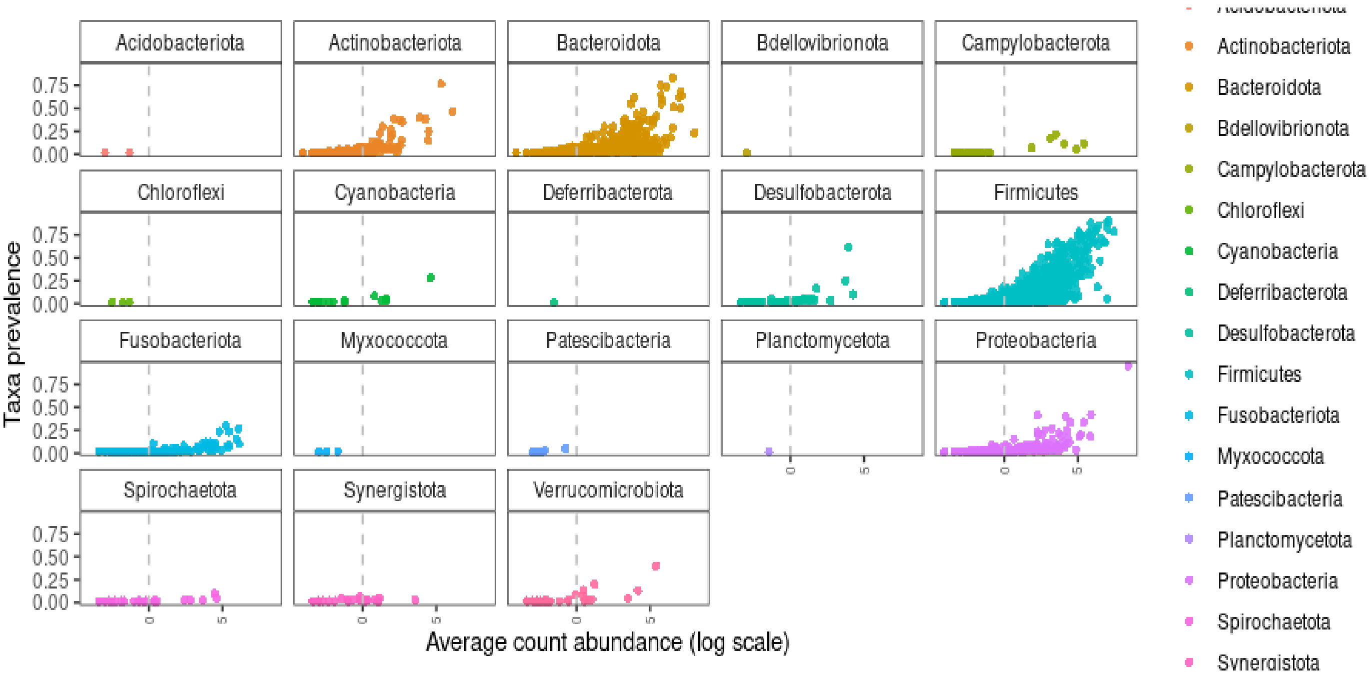


Figure 4.2.III.A Prevalence plot on the Phylum level

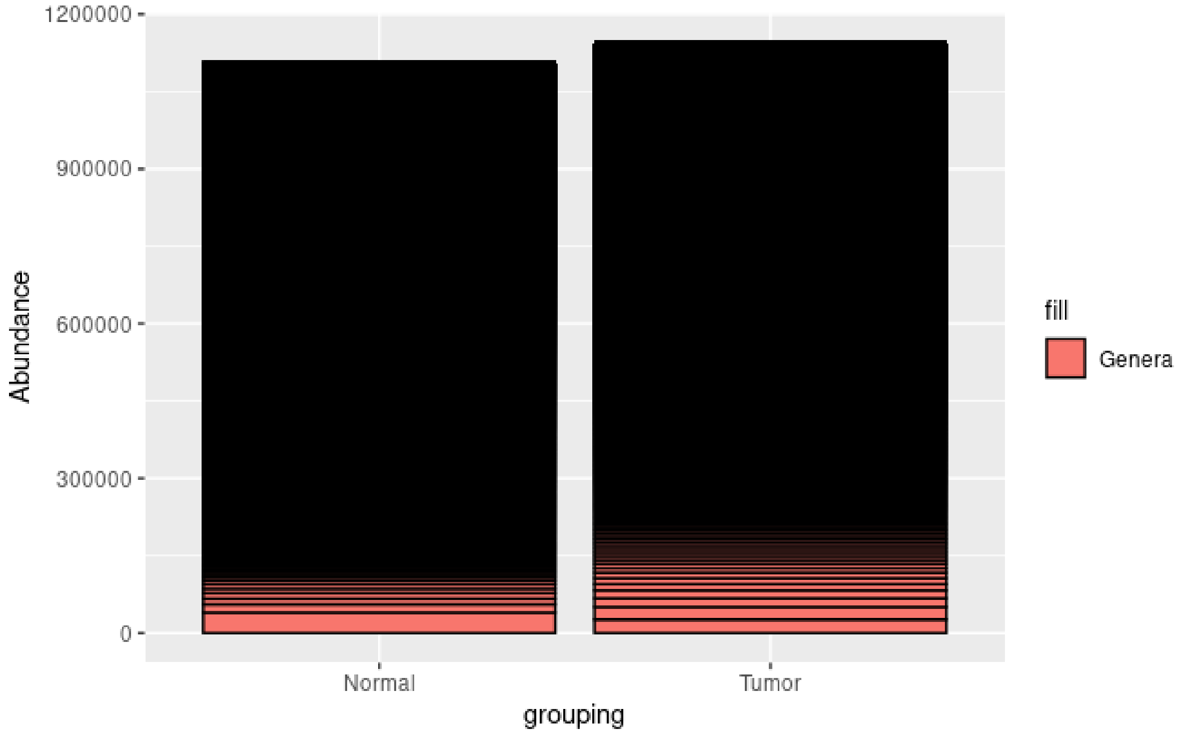
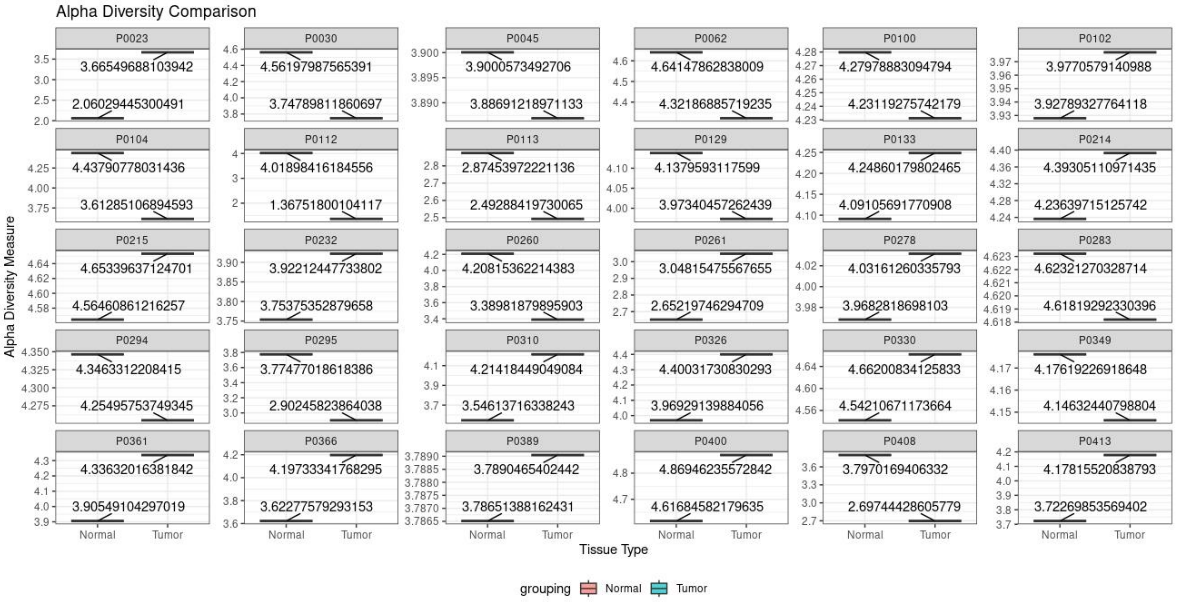


Figure 4.2.III.B difference in abundance on the Genus level

Alpha diversity, when comparing the microbiota composition in tumor tissue versus matched normal healthy tissue within the same patients, the results suggest that there is no significant difference in the microbial diversity (as measured by the Shannon index) between the tumor and normal tissues. This could mean that the presence of the tumor does not significantly alter the overall diversity of the microbiota in the tissue samples analysed.

The combined analysis of the Shannon and observed richness measures suggests that there is a real difference in the alpha diversity of tumor and normal tissues. However, this difference is more pronounced when using the observed richness measure, which is more sensitive to rare taxa.



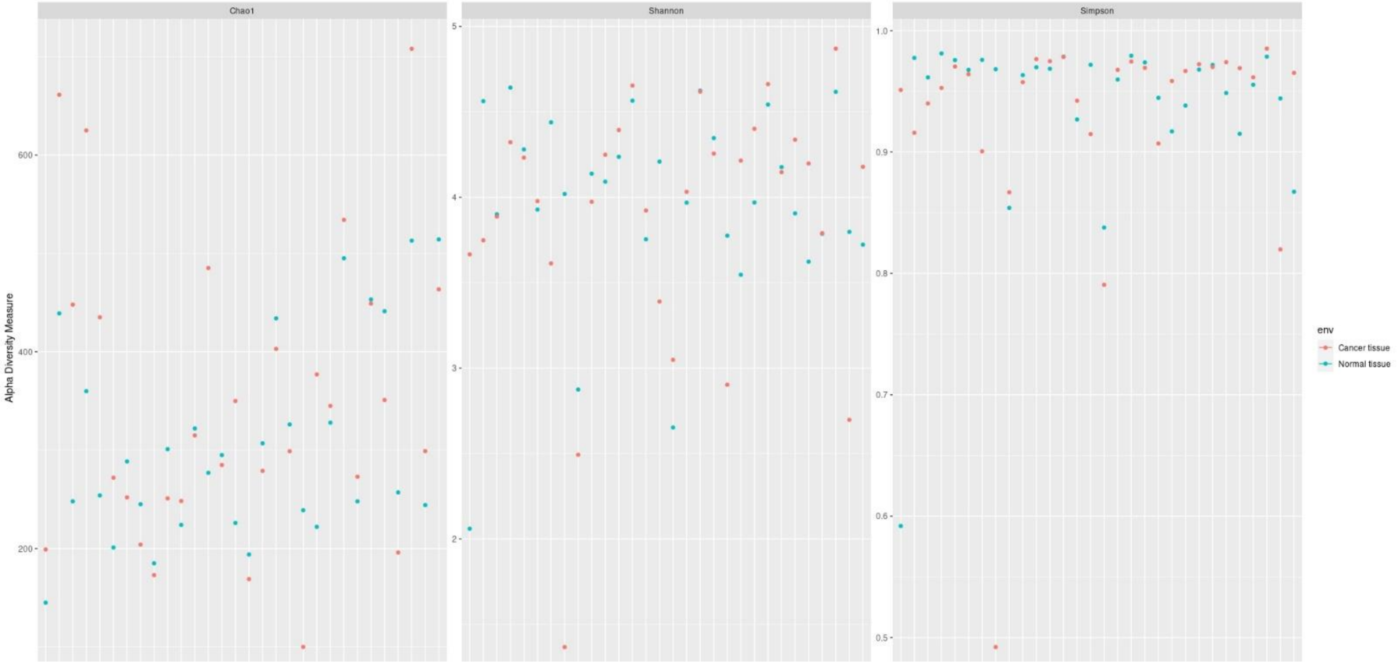


Figure 4.2.III.C Alpha Diversity within samples plot

Beta diversity analysis, the weighted UniFrac NMDS analysis has a stress of 0.1906, which is relatively low. This suggests that the ordination captures the underlying structure of the data well. The samples are clustered into two distinct groups, with tumor samples being more concentrated in one group, and normal samples being more concentrated in the other group.

Bray-Curtis NMDS: The Bray-Curtis NMDS analysis has a stress of 0.2053, which is slightly higher than the weighted UniFrac NMDS analysis. However, it is still low enough to suggest that the ordination captures the underlying structure of the data well. The samples are also clustered into two distinct groups in this analysis.

Unweighted UniFrac NMDS: The unweighted UniFrac NMDS analysis has a stress of 0.1749, which is the lowest of the three analyses. This suggests that it is the most accurate at capturing the underlying structure of the data. After filtering and applying significance criteria (adjusted p-value < 0.01 and log2 fold change > 1), differential abundance analysis revealed a single ASV (ASV149) exhibiting significant differential abundance between normal and cancer tissues (adjusted p-value = 0.999886). This ASV, taxonomically classified as a member of the genus Ruminococcus within the phylum Firmicutes, demonstrated a remarkable 5.95-fold higher abundance in cancer tissues compared to normal tissues, corresponding to a 384-fold increase in absolute abundance.

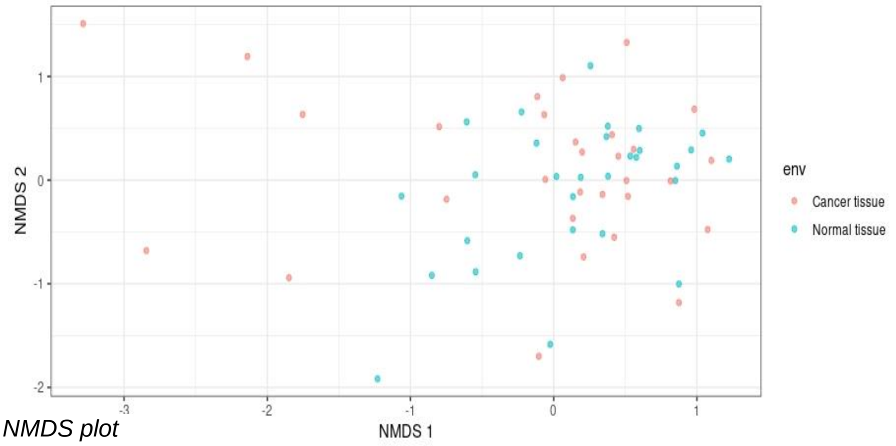


Figure 4.2.III.D Bray-Curtis NMDS plot

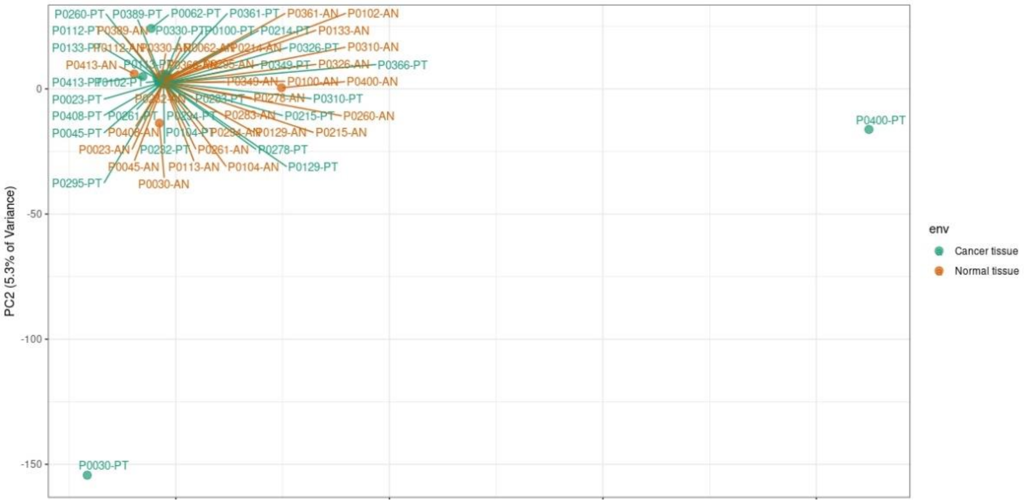


Figure 4.2.III.E PCA Analysis plot

6. Disscussion

6.1 Microbiome Composition and Diversity

Our exploration of the microbiome composition in colorectal cancer (CRC) between matched tumor and healthy colon tissues unveiled a dominance of Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria. At the phylum level This finding aligns with their known prevalence in the gut microbiome of healthy individuals. Interestingly, we observed a moderate prevalence of genera such as Bacteroides and Faecalibacterium, while Lachnoclostridium and Blautia demonstrated a consistently high presence. This suggests a complex microbial community where key players coexist with potentially specialized taxa. In assessing alpha diversity, we initially found no significant differences between the microbiota in tumor and matched normal healthy tissues when using metrics like the Shannon index. While the Shannon index is a widely used measure of alpha diversity, it has some limitations in terms of its sensitivity to subtle differences between communities. This is because the Shannon index is based on the assumption that all rare taxa contribute equally to the overall diversity, which may not always be the case.

Additionally, the lack of significant difference could be due to a variety of factors, including small sample size, high variability within sample groups, or the possibility that the Shannon index is not sensitive enough to detect the differences present. However, a more nuanced examination using observed richness revealed subtle but significant differences. This underscores the importance of employing multiple diversity metrics for a comprehensive understanding. The observed richness, being more sensitive to rare taxa, highlighted distinctions not apparent with the Shannon index However, it's important to note that while the overall diversity may not differ, the composition of the microbiota (which species are present and their relative abundances) could still vary significantly.

6.2 Beta Diversity and Community Structure

In this study, beta diversity analysis was conducted between the tumor and healthy tissue for the same patient, employing weighted UniFrac, Bray-Curtis, and unweighted UniFrac NMDS, uncovered distinct clustering of tumor and normal tissues. The weighted UniFrac NMDS, with its low stress value, captured the underlying structure well, revealing two distinct groups with tumor tissues more concentrated in one group and normal tissues in another. The Bray-Curtis analysis exhibited slightly higher stress but still effectively captured the separation, while the unweighted UniFrac analysis proved the most accurate, emphasizing the importance of considering different metrics for a comprehensive assessment of community structure.

6.3 Differential Abundance Analysis

a single ASV emerged as significantly differentially abundant between the two tissue types. ASV149 which represents the Ruminococcus, a genus within the Phylum Firmicutes. This finding implicates Ruminococcus as a potential key player in the CRC-associated microbiota which aligns recent research findings, A study by Cai at al. found that the abundance of Rumnicoccus was significantly increased in colorectal neoplasms, this increase in Rumnicoccus, along with other genera such as Blautia, may lead to gut microbiota imbalance which could potentially increase the severity of diseases like COVID-19

Meanwhile, the exact species was not classified, Ruminococcus bromii is a dominant member of the human colonic microbiota that plays a 'keystone' role in degrading dietary resistant starch.

7. Conclusion

It is clear that there are numerous taxa in the colorectal cancer microbiota correlated with the disease. Here, in this study we investigated the potential role of microbiota in CRC by comparing microbial composition between the tumor tissue and matched normal tissue in patients' samples. Our findings align with previous studies highlighting the dominance of Firmictes and bacteroides phyla in the gut microbiome.

Further more we observed distinct community profiles within specific genus indicating the role of microbiome as key players and potentially specialized taxa.

Additionally, while alpha diversity did not reveal significant a differences between tumor and normal tissue, within the patient, a combined analysis including observed richness suggested a subtle shift in community composition. This suggests that, while overall diversity may not be affected, the presence of a tumor may influence the abundance of specific rare taxa.

Differential abundance analysis identified the genus Ruminococcus within the Firmicutes phylum as significantly enriched in cancer tissues. This finding is intriguing, considering the potential role of Ruminococcus species in promoting tumor growth and pro-inflammatory responses.

While not revealing drastic changes in overall diversity, our findings suggest subtle compositional shifts and highlight the potential significance of specific taxa, particularly Ruminococcin species. Further research exploring these specific taxa and their functional roles is crucial to deepen our understanding of the microbiome's contribution to colorectal cancer progression and potentially pave the way for novel therapeutic strategies.

8. Limitations

Despite its power, such projects on microbiota characterization faces several limitations that can influence the workflow and hinder reliable interpretation. Some key limitations include:

Data size, big data challenges and Algorithm complexity.

Microbiome sequencing generates massive datasets requiring specialized hardware and software for efficient storage, processing and analysis. Limited computational resources can constrain research, particularly for high resolution techniques like 16S rRNA sequencing. Moreover, the analysis algorithms are often complex and involve numerous processing steps, each has its own computational demands. For instance, the DADA2 package which was integral in our project typically requires an average of 32GB of RAM to execute certain commands and codes.

Therefore, adjusting pipelines for efficiency, accuracy and automation while managing resources limitations can be challenging. For instance, it is important to be aware of the limitations of different taxonomic assignment methods. OTU-based methods, which cluster sequences into OTUs based on their similarity, can be inaccurate if the reference database is of poor quality or if the sequences are highly divergent. ASV-based methods, which denoise and identify individual sequences, can provide more accurate taxonomic assignment but are computationally more demanding. Researchers should carefully consider the trads-offs between these two methods when choosing an approach for their study.

Inconsistency in analysis pipelines and the lack of standardization.

the analysis of the microbiome is affected by experimental conditions (e.g. sequencing errors and genomic repeats, wet library preparations, errors from the sequencing machines) and is computationally intensive and cumbersome downstream analysis (e.g. quality control, assembly, binning and statistical analyses). The lack of consistency between studies results even when analyzing the same data, is a significant limiting factor in this field. One of the reasons for this inconsistency is the variation in applied bioinformatics analysis pipelines. The field currently lacks universal standardization for analysis pipelines, variations in algorithms, databases, and processing parameters can lead to discrepancies between studies. Thus, the choice of bioinformatics analysis pipeline significantly impacts results. Development of robust algorithms, and standardization of analysis pipelines will pave the way for more reliable and reproducible studies, ultimately leading to a deeper understanding of the microbiome's role in health and disease.

9. Acknowledgments

With hearts full of gratitude, we acknowledge the invaluable contributions for everyone who made the establishment of the Bioinformatics Master’s program possible. A special thank you to the visionary leaders, professors who recognized the demand for bioinformatics expertise and championed the creation of this program at the SVU, to the dedicated faculty and staff who have poured their passion and expertise into crafting a world-class curriculum.

Moreover, I would like to extend my profound gratitude to master’s colleague

Dr. Haider Abd Aldaim (Germany-Wuppertal) for his invaluable assistance in setting up the working environment and downloading the list of required packages, after numerous personal setbacks and seemingly inexplicable errors.

The project was executed on a Virtual Box, utilizing an Ubuntu Virtual Machine. This setup streamlined the data retrieval and pre-processing step through Linux shell commands using Bash language, Data were then directed to the working directory and subsequently accessed using R.

All R packages essential for on microbiota characterization were sourced by Dr. Haidar either directly or through GitHub's development platform. However, it's important to note that access to many of these resources is IP-restricted and hence, not universally accessible. Furthermore, this project was notably demanding in terms of computational resources.

Despite These challenges, the setup has been instrumental in driving the project forward and we remain committed to optimizing it further to align with our objectives.

10. References

1. Burns, Michael B., Joshua Lynch, Timothy K. Starr, Dan Knights, and Ran Blekhman. 2014. "Virulence Genes Are a Signature of the Microbiome in the Colorectal Tumor Microenvironment." bioRxiv. https://doi.org/10.1101/009431.
2. Burns, Michael B., Emmanuel Montassier, Juan Abrahante, Sambhawa Priya, David E. Niccum, Alexander Khoruts, Timothy K. Starr, Dan Knights, and Ran Blekhman. 2018. "Colorectal Cancer Mutational Profiles Correlate with Defined Microbial Communities in the Tumor Microenvironment." Edited by Eric R. Fearon. *PLOS Genetics* 14 (6): e1007376. https://doi.org/10.1371/journal.pgen.1007376.
3. Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.
4. Chen, Maohua, Wei Lin, Nan Li, Qian Wang, Shaomi Zhu, Anqi Zeng, and Linjiang Song. 2022. "Therapeutic Approaches to Colorectal Cancer via Strategies Based on Modulation of Gut Microbiota." *Frontiers in Microbiology* 13. https://www.frontiersin.org/articles/10.3389/fmicb.2022.945533.
5. Chiarello, Marlène, Mark McCauley, Sébastien Villéger, and Colin R. Jackson. 2022. "Ranking the Biases: The Choice of OTUs vs. ASVs in 16S rRNA Amplicon Data Analysis Has Stronger Effects on Diversity Measures than Rarefaction and OTU Identity Threshold." *PLOS ONE* 17 (2): e0264443. https://doi.org/10.1371/journal.pone.0264443.

6. Conti, Gabriele, Federica D’Amico, Marco Fabbrini, Patrizia Brigidi, Monica Barone, and Silvia Turroni. 2023. “Pharmacomicrobiomics in Anticancer Therapies: Why the Gut Microbiota Should Be Pointed Out.” *Genes* 14 (1). https://doi.org/10.3390/genes14010055.
7. Edgar, Robert C. 2018. “Updating the 97% Identity Threshold for 16S Ribosomal RNA OTUs.” *Bioinformatics* 34 (14): 2371–75. https://doi.org/10.1093/bioinformatics/bty113.
8. Farhana, Lulu, Hirendra Nath Banerjee, Mukesh Verma, and Adhip P. N. Majumdar. 2018. “Role of Microbiome in Carcinogenesis Process and Epigenetic Regulation of Colorectal Cancer.” In *Cancer Epigenetics for Precision Medicine : Methods and Protocols*, edited by Ramona G. Dumitrescu and Mukesh Verma, 35–55. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-8751-1_3.
9. Gao, Bei, Liang Chi, Yixin Zhu, Xiaochun Shi, Pengcheng Tu, Bing Li, Jun Yin, Nan Gao, Weishou Shen, and Bernd Schnabl. 2021. “An Introduction to Next Generation Sequencing Bioinformatic Analysis in Gut Microbiome Studies.” *Biomolecules* 11 (4): 530. https://doi.org/10.3390/biom11040530.
10. Gong, Dengmei, Amma G. Adomako-Bonsu, Maijian Wang, and Jida Li. 2023. “Three Specific Gut Bacteria in the Occurrence and Development of Colorectal Cancer: A Concerted Effort.” *PeerJ* 11 (August): e15777. https://doi.org/10.7717/peerj.15777.
11. Grimm, Dominik G. 2019. “Current Challenges and Best-Practice Protocols for Microbiome Analysis.” *Briefings in Bioinformatics* 22 (1): 178–93. https://doi.org/10.1093/bib/bbz155.
12. Jeske, Jan Torsten, and Claudia Gallert. 2022. “Microbiome Analysis via OTU and ASV-Based Pipelines-A Comparative Interpretation of Ecological Data in WWTP Systems.” *Bioengineering (Basel, Switzerland)* 9 (4): 146. https://doi.org/10.3390/bioengineering9040146.
13. Lee, Hyeon Been, Dong Hyuk Jeong, Byung Cheol Cho, and Jong Soo Park. 2023. “Comparative Analyses of Eight Primer Sets Commonly Used to Target the Bacterial 16S rRNA Gene for Marine Metabarcoding-Based Studies.” *Frontiers in Marine Science* 10. https://www.frontiersin.org/articles/10.3389/fmars.2023.1199116.
14. Lian, Wenping, Huifang Jin, Jingjing Cao, Xinyu Zhang, Tao Zhu, Shuai Zhao, Sujun Wu, et al. 2020. “Identification of Novel Biomarkers Affecting the Metastasis of Colorectal Cancer through Bioinformatics Analysis and Validation through qRT-PCR.” *Cancer Cell International* 20 (1): 105. https://doi.org/10.1186/s12935-020-01180-4.
15. Liu, Yong-Xin, Yuan Qin, Tong Chen, Meiping Lu, Xubo Qian, Xiaoxuan Guo, and Yang Bai. 2021. “A Practical Guide to Amplicon and Metagenomic Analysis of Microbiome Data.” *Protein & Cell* 12 (5): 315–30. https://doi.org/10.1007/s13238-020-00724-8.

16. Louis, Petra, Georgina L. Hold, and Harry J. Flint. 2014. “The Gut Microbiota, Bacterial Metabolites and Colorectal Cancer.” *Nature Reviews Microbiology*. https://doi.org/10.1038/NRMICRO3344.
17. Marcos-Zambrano, Laura Judith, Kanita Karaduzovic-Hadziabdic, Tatjana Loncar Turukalo, Piotr Przymus, Vladimir Trajkovik, Oliver Aasmets, Magali Berland, et al. 2021. “Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment.” *Frontiers in Microbiology* 12. https://doi.org/10.3389/fmicb.2021.634511.
18. Martino, Cameron, Daniel McDonald, Kalen Cantrell, Amanda Hazel Dilmore, Yoshiki Vázquez-Baeza, Liat Shenhav, Justin P. Shaffer, et al. 2022. “Compositionally Aware Phylogenetic Beta-Diversity Measures Better Resolve Microbiomes Associated with Phenotype.” *mSystems* 7 (3): e00050-22. https://doi.org/10.1128/msystems.00050-22.
19. Masood, Muqaddas, Moussa Ide Nasser, Muhammad Bilal, Muqaddas Masood, Moussa Ide Nasser, and Muhammad Bilal. 2023. “Gut Microbial Metabolites and Colorectal Cancer.” https://doi.org/10.1016/B978-0-323-99476-7.00011-9.
20. McLaren, Michael R., and Benjamin J. Callahan. 2021. “Silva 138.1 Prokaryotic SSU Taxonomic Training Data Formatted for DADA2.” Zenodo. https://doi.org/10.5281/zenodo.4587955.
21. McMurdie, Paul J., and Susan Holmes. 2013. “Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data.” *PLOS ONE* 8 (4): 1–11. https://doi.org/10.1371/journal.pone.0061217.
22. Moskowitz, Jacob E., Anthony G. Doran, Zhentian Lei, Susheel B. Busi, Marcia L. Hart, Craig L. Franklin, Lloyd W. Sumner, Thomas M. Keane, and James M. Amos-Landgraf. 2020. “Integration of Genomics, Metagenomics, and Metabolomics to Identify Interplay between Susceptibility Alleles and Microbiota in Adenoma Initiation.” *BMC Cancer* 20 (1): 600. https://doi.org/10.1186/s12885-020-07007-9.
23. Peeters, Jannes, Olivier Thas, Ziv Shkedy, Leyla Kodalci, Connie Musisi, Olajumoke Evangelina Owokotomo, Aleksandra Dyczko, et al. 2021. “Exploring the Microbiome Analysis and Visualization Landscape.” *Frontiers in Bioinformatics* 1. https://www.frontiersin.org/articles/10.3389/fbinf.2021.774631.
24. Prodan, Andrei, Valentina Tremaroli, Harald Brolin, Aeilko H. Zwinderman, Max Nieuwdorp, and Evgeni Levin. 2020. “Comparing Bioinformatic Pipelines for Microbial 16S rRNA Amplicon Sequencing.” *PLOS ONE* 15 (1): e0227434. https://doi.org/10.1371/journal.pone.0227434.
25. Rebersek, Martina. 2021. “Gut Microbiome and Its Role in Colorectal Cancer.” *BMC Cancer* 21 (1): 1325. https://doi.org/10.1186/s12885-021-09054-2.
26. Reitmeier, Sandra, Thomas C. A. Hitch, Nicole Treichel, Nikolaos Fikas, Bela Hausmann, Amanda E. Ramer-Tait, Klaus Neuhaus, et al. 2021. “Handling of

Spurious Sequences Affects the Outcome of High-Throughput 16S rRNA Gene Amplicon Profiling." *ISME Communications* 1 (1): 31. https://doi.org/10.1038/s43705-021-00033-z.

27. "Release 138.1." n.d. Accessed January 3, 2024. https://www.arb-silva.de/documentation/release-138.1/.
28. Roelands, Jessica, Peter J. K. Kuppen, Eiman I. Ahmed, Raghvendra Mall, Tariq Masoodi, Parul Singh, Gianni Monaco, et al. 2023. "An Integrated Tumor, Immune and Microbiome Atlas of Colon Cancer." *Nature Medicine* 29 (5): 1273–86. https://doi.org/10.1038/s41591-023-02324-5.
29. Schloss, Patrick D. 2021. "Amplicon Sequence Variants Artificially Split Bacterial Genomes into Separate Clusters." *mSphere* 6 (4): 10.1128/msphere.00191-21. https://doi.org/10.1128/msphere.00191-21.
30. Senthakumaran, Thulasika, Aina E. F. Moen, Tone M. Tannæs, Alexander Endres, Stephan A. Brackmann, Trine B. Rounge, Vahid Bemanian, and Hege S. Tunsjø. 2023. "Microbial Dynamics with CRC Progression: A Study of the Mucosal Microbiota at Multiple Sites in Cancers, Adenomatous Polyps, and Healthy Controls." *European Journal of Clinical Microbiology & Infectious Diseases* 42 (3): 305–22. https://doi.org/10.1007/s10096-023-04551-7.
31. Sharma, Vineet K., Naveen Kumar, Tulika Prakash, and Todd D. Taylor. 2012. "Fast and Accurate Taxonomic Assignments of Metagenomic Sequences Using MetaBin." *PLOS ONE* 7 (4): e34030. https://doi.org/10.1371/journal.pone.0034030.
32. Smith, Gillian, Francis A. Carey, Julie Beattie, Murray J. V. Wilkie, Tracy J. Lightfoot, Jonathan Coxhead, R. Colin Garner, Robert J. C. Steele, and C. Roland Wolf. 2002. "Mutations in APC, Kirsten-Ras, and P53—Alternative Genetic Pathways to Colorectal Cancer." *Proceedings of the National Academy of Sciences* 99 (14): 9433–38. https://doi.org/10.1073/pnas.122612899.
33. Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." *CA: A Cancer Journal for Clinicians* 71 (3): 209–49. https://doi.org/10.3322/caac.21660.
34. Szopinska-Tokov, Joanna, Mirjam Bloemendaal, Jos Boekhorst, Gerben DA Hermes, Thomas HA Ederveen, Priscilla Vlaming, Jan K. Buitelaar, Barbara Franke, and Alejandro Arias-Vasquez. 2023. "A Comparison of Bioinformatics Pipelines for Compositional Analysis of the Human Gut Microbiome." bioRxiv. https://doi.org/10.1101/2023.02.13.528280.
35. Ternes, Dominik, Jessica Karta, Mina Tsenkova, Paul Wilmes, Serge Haan, and Elisabeth Letellier. 2020. "Microbiome in Colorectal Cancer: How to Get from Meta-Omics to Mechanism?" *Trends in Microbiology* 28 (5): 401–23. https://doi.org/10.1016/j.tim.2020.01.001.

36. Tjalsma, Harold, Annemarie Boleij, Julian R. Marchesi, and Bas E. Dutilh. 2012. “A Bacterial Driver–Passenger Model for Colorectal Cancer: Beyond the Usual Suspects.” *Nature Reviews Microbiology* 10 (8): 575–82. https://doi.org/10.1038/nrmicro2819.
37. Wang, Qiong, George M. Garrity, James M. Tiedje, and James R. Cole. 2007. “Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy.” *Applied and Environmental Microbiology* 73 (16): 5261–67. https://doi.org/10.1128/AEM.00062-07.
38. Wensel, Caroline R., Jennifer L. Pluznick, Steven L. Salzberg, and Cynthia L. Sears. 2022. “Next-Generation Sequencing: Insights to Advance Clinical Investigations of the Microbiome.” American Society for Clinical Investigation. April 1, 2022. https://doi.org/10.1172/JCI154944.
39. Ye, Xunwen, Yonglin Chen, and Jialin Gu. 2023. “Risk Factors for Advanced Colorectal Neoplasm in Young Adults: A Meta-Analysis.” *Future Oncology* 19 (18): 1293–1302. https://doi.org/10.2217/fon-2023-0165.
40. Zhang, Wenke, Xiaoqian Fan, Haobo Shi, Jian Li, Mingqian Zhang, Jin Zhao, and Xiaoquan Su. 2023. “Comprehensive Assessment of 16S rRNA Gene Amplicon Sequencing for Microbiome Profiling across Multiple Habitats.” *Microbiology Spectrum* 11 (3): e00563-23. https://doi.org/10.1128/spectrum.00563-23.
41. Zhang, Yong-Zhen, Xin Yu, Enda Yu, Na Wang, Quancai Cai, Qun Shuai, Feihu Yan, et al. 2018. “Changes in Gut Microbiota and Plasma Inflammatory Factors across the Stages of Colorectal Tumorigenesis: A Case-Control Study.” *BMC Microbiology*. https://doi.org/10.1186/S12866-018-1232-6.
42. Zhao, Liuyang, et al. “The composition of colonic commensal bacteria according to anatomical localization in colorectal cancer.” Engineering, vol. 3, no. 1, 2017, pp. 90–97, https://doi.org/10.1016/j.eng.2017.01.012.
43. 2. Qin, Junjie, et al. “A human gut microbial gene catalogue established by metagenomic sequencing.” Nature, vol. 464, no. 7285, 2010, pp. 59–65, https://doi.org/10.1038/nature08821.
44. 3. Sung, Hyuna, et al. “Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries.” CA: A Cancer Journal for Clinicians, vol. 71, no. 3, 2021, pp. 209–249, https://doi.org/10.3322/caac.21660.
45. 4. Sepulveda, Antonia R., et al. “Molecular biomarkers for the evaluation of colorectal cancer: Guideline from the American Society for Clinical Pathology, College of American Pathologists, Association for Molecular Pathology, and the American Society of Clinical Oncology.” Journal of Clinical Oncology, vol. 35, no. 13, 2017, pp. 1453–1486, https://doi.org/10.1200/jco.2016.71.9807.
46. 5. Petersen, C., and Round, J. L. (2014). Defining dysbiosis and its influence on host immunity and disease. Cell. Microbiol. 16, 1024–1033. doi: 10.1111/cmi.12308

47. 6. Kostic AD, Chun EY, Robertson L, Glickman JN, Gallini CA, Michaud M, Clancy TE, Chung DC, Lochhead P, Hold GL, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe. 2013;14(2):207–15.
48. 7. Goodwin AC, Shields CED, Wu SG, Huso DL, Wu XQ, Murray-Stewart TR, Hacker-Prietz A, Rabizadeh S, Woster PM, Sears CL, et al. Polyamine catabolism contributes to enterotoxigenic Bacteroides fragilis-induced colon tumorigenesis. P Natl Acad Sci USA. 2011;108(37):15354–9.
49. 8. Ternes, D., Karta, J., Tsenkova, M., Wilmes, P., Haan, S., & Letellier, E. (2020b). Microbiome in colorectal cancer: How to get from Meta-omics to mechanism? Trends in Microbiology, 28(5), 401–423. https://doi.org/10.1016/j.tim.2020.01.001
50. 9.Duvallet C. et al. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat. Commun. 2017; 8: 1784
51. Whelan F.J. Surette M.G. A comprehensive evaluation of the sl1p pipeline for 16S rRNA gene sequencing analysis. Microbiome. 2017; 5: 100
52. Rubinstein M.R et al. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/beta-catenin signaling via its FadA adhesin. Cell Host Microbe. 2013; 14: 195-206
53. Nosho K. et al.Association of Fusobacterium nucleatum with immunity and molecular alterations in colorectal cancer. World J. Gastroenterol. 2016; 22: 557-566
54. Gur C. et al. Binding of the Fap2 protein of Fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. Immunity. 2015; 42: 344-355
55. Abed J.et al.Fap2 mediates Fusobacterium nucleatum colorectal adenocarcinoma enrichment by binding to tumor-expressed Gal-GalNAc.Cell Host Microbe. 2016; 20: 215-225
56. Yu T.et al.Fusobacterium nucleatum promotes chemoresistance to colorectal cancer by modulating autophagy.Cell. 2017; 170: 548-563.e16
57. Zhang S.et al.Fusobacterium nucleatum promotes chemoresistance to 5-fluorouracil by upregulation of BIRC3 expression in colorectal cancer. J. Exp. Clin. Cancer Res. 2019; 38: 14
58. Gagic D.et al.Exploring the secretomes of microbes and microbial communities using filamentous phage display.Front. Microbiol. 2016; 7: 429
59. Collins R.R.J.et al.Oncometabolites. A new paradigm for oncology, metabolism, and the clinical laboratory.Clin. Chem. 2017; 63: 1812-1820
60. Connors J.et al.The role of succinate in the regulation of intestinal inflammation. Nutrients. 2019;11: 25
61. Chowdhury R.et al.The oncometabolite 2-hydroxyglutarate inhibits histone lysine demethylases. EMBO Rep. 2011; 12: 463-469

62. Chriett S.et al.Prominent action of butyrate over β-hydroxybutyrate as histone deacetylase inhibitor, transcriptional modulator and anti-inflammatory molecule. Sci. Rep. 2019; 9: 742
63. Wu X.et al. Effects of the intestinal microbial metabolite butyrate on the development of colorectal cancer. J. Cancer. 2018; 9: 2510-2517
64. Kaiko G.E. et al. The colonic crypt protects stem cells from microbiota-derived metabolites. Cell. 2016; 165: 1708-1720
65. Sender R.et al. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol. 2016; 14e1002533
66. Flynn K.J.et al. Spatial variation of the native colon microbiota in healthy adults. Cancer Prev. Res. (Phila.). 2018; 11: 393-402
67. Yamauchi M.et al. Assessment of colorectal cancer molecular features along bowel subsites challenges the conception of distinct dichotomy of proximal versus distal colorectum. Gut. 2012; 61: 847-854
68. Dejea C.M.et al. Microbiota organization is a distinct feature of proximal colorectal cancers. Proc. Natl. Acad. Sci. U. S. A. 2014; 111: 18321-18326
69. Drewes J.L.et al.High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. NPJ Biofilms Microbiomes. 2017; 3: 34
70. Purcell R.V.et al. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. Sci. Rep. 2017; 7: 11590
71. Chung L.et al. Bacteroides fragilis toxin coordinates a pro-carcinogenic inflammatory cascade via targeting of colonic epithelial cells. Cell Host Microbe. 2018; 23: 203-214.e5.