PhD degree in Systems Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples “Federico II”

Reconstruction of the condition- and location-specific colon cancer microbiome from human RNA sequencing data

Settore disciplinare: MED/04

Gaia Sambruni

Tutor: Dr. Schaefer Martin

European Institute of Oncology

PhD Coordinator: Prof. Saverio Minucci

Anno accademico 2022-2023

Contents

Acronyms

16S ribosomal RNA 16S gene.

APC adenomatous polyposis coli.

BMI body mass index.

BRCA breast carcinoma.

CIMP CpG methylation phenotype.

CIN chromosomal instability.

CMS consensus molecular subtype.

COAD colon adenocarcinoma.

CpG cytosine-phosphate-guanine.

CPM copies per million.

CRC colorectal cancer.

DAPI 4',6-Diamidino-2-phenylindole.

DFS disease-free survival.

DNA deoxyribonucleic acid.

EGA European Genome-Phenome Archive.

FAP familial adenomatous polyposis.

FDR false discovery rate.

FFPE Formalin-Fixed Paraffin-Embedded.

FISH fluorescent in situ hybridization.

FPKM fragments per kilobase million.

GBM glioblastoma multiforme.

GDC Genomic Data Commons.

HNSC head and neck squamous cell carcinoma.

HPV Human papillomavirus.

IBD inflammatory bowel disease.

IEO European Institute of Oncology.

KRAS Kirsten rat sarcoma viral oncogene homolog.

LUAD lung adenocarcinoma.

LUSC lung squamous cell carcinoma.

MMR mismatch repair.

MSI microsatellite instability.

NCBI National Center for Biotechnology Information.

NGS next generation sequencing.

nMDS non-metric multidimensional scaling.

OS overall survival.

OV ovarian serous cystadenocarcinoma.

PBS phosphate buffered saline.

PC principal component.

PCA principal component analysis.

PCo principal coordinate.

PCoA principal coordinate analysis.

PCR polymerase chain reaction.

plate ID plate identifier.

pRB retinoblastoma protein.

RB retinoblastoma.

READ rectum adenocarcinoma.

RNA ribonucleic acid.

RNA-Seq RNA Sequencing.

rRNA ribosomal RNA.

SKCM skin cutaneous melanoma.

STAT3 signal transducer and activator of transcription 3.

TCA tricarboxylic acid cycle.

TCGA the cancer genome atlas.

TGF-β transforming growth factor β.

TPM transcript per million.

WGS whole genome sequencing.

WXS whole exome sequencing.

List of Figures

List of Tables

Acknowledgements

I would like to express my sincere gratitude to several individuals and institutions who have played a significant role in the completion of my thesis.

Firstly, I am grateful to my supervisor, Martin, for his guidance and feedback throughout this research project. His expertise and mentorship have been instrumental in shaping my work.

I would like to thank Luigi and his group, especially Angeli and Carlotta, for the supervision, help and suggestions they gave me during this academic journey.
Furthermore, I extend my appreciation to my internal advisor, Sara, and my external advisor, Georg, for their insights and suggestions. Their contributions have not only greatly enriched my research but had a profound impact on my personal growth.

I am grateful to the Schaefer group, Richard, Gokce, Danilo, Fabio, Tiziano, Marta and my office mates, Angeli and Gaurav, for creating a stimulating and collaborative environment and providing fruitful discussions.
I would like to acknowledge the SEMM School for providing me with an environment conducive to research and intellectual growth, with great opportunities for interdisciplinary collaboration, such as the ENABLE event.
I would also like to express my sincere appreciation to the IEO and the University of Milan for their support and access to resources.
Furthermore, I am grateful to the AIRC for providing the grant that supported my research.

Lastly, I extend my thanks to all the friends and family members who have offered their encouragement, understanding and support throughout this journey.

Abstract

The microbiota play a vital role in the organism's survival: the disruption of the delicate equilibrium between the host and its microbiota can lead to the development of diseases, prompting the exploration of the involvement of microbes in the growth and progression of cancer. Nonetheless, it remains uncertain whether specific patterns of microbial colonisation are associated with the molecular characteristics of tumours. This study introduces an approach for estimating microbial signals within human next generation sequencing (NGS) data, aiming to comprehend how much you can exploit this data to uncover insights about the link between microbes and the properties of tumours.

Human NGS data may be susceptible to contamination and technical issues, so we have implemented controls to identify the most appropriate approaches for analysing the data, including the selection of the optimal microbial reads extractor and dimensionality reduction method to understand the overall microbial composition trends.

We conducted the analyses on TCGA data and further evaluated it using a cohort of colon cancer patients. The majority of the identified microbes were bacteria, while only a small proportion of the signals could be attributed to viruses, eukaryota, or archaea. Nevertheless, we confirmed the presence of specific microbes, such as Human papillomavirus signals, in clinically positive samples but the sparsity of the data prevented us from conducting further analysis. Leveraging the higher detection of bacterial signals, we were able to establish associations between bacterial compositions and several tumour properties, including survival rates, tumour location, microsatellite instability, consensus molecular subtype and the infiltration of immune cells in colon tumours. Moreover, our findings suggest potential mechanisms through which colon tumours are linked to bacteria, such as interactions with immune cells and bacterial pathways. However, only a limited number of modest associations were identified in other cancer types.

Through the concurrent analysis of tumour properties along with the composition of the microbiome associated with them, we have enabled the exploration of the relationship between microbiota and tumours with a methodology that has the potential to improve patient stratification and investigations into the underlying mechanisms.

Introduction

2.1 Microbiota and human

The microbiota is defined as the group of microorganisms that colonise a habitat, e.g. soil, ocean and a host's body, where it usually shows a host-beneficial effect but that, in specific situations, can become detrimental (Kuziel and Rakoff-Nahoum, 2022). The microbiota is composed of microorganisms from different domains, i.e. Bacteria, Archaea, Eukaryota and viruses and can range from only one species to thousands (Dey and Ray Chaudhuri, 2022; Kuziel and Rakoff-Nahoum, 2022). In particular, in the human body there are estimated to be almost $3.8x10^{13}$ microbial cells, approximately as many as the number of the host cells ($3x10^{13}$) (Whisner and Athena Aktipis, 2019) and the number of microbial genes has been hypothesised to be around 6 million (Almeida et al., 2021) many more than the 20 thousand genes of the human genome (Nurk et al., 2022). In humans, the microbiota is present in all the tissues exposed to the external environment, i.e. the skin, respiratory, digestive and urogenital tract, but the gut holds the largest microbial biomass (trillions of living microbes for around one thousand species, Davenport et al., 2017). Gut bacteria are widely studied in the context of health and disease, but the findings from these studies can be applied to describe other organs and microorganisms.

The microbiota contributes to the healthy status of the host by maintaining the homeostasis of the colonised tissue: so, to better highlight the relevance of the microbiota in the host integrity, some authors suggested the term "holobiont", to define organisms in which the host is strictly defined together with the microbiota, sharing with it their evolution direction (Gordon et al., 2013; Davenport et al., 2017).

One of the most interesting aspects of the human microbiota, especially for the colon, is its heterogeneity: the microbiota composition of humans is so distinctive in each individual that it can be considered a fingerprint or a specific signature for every person and every body site (Mousa, Chehadeh, and Husband, 2022; Franzosa et al., 2015; Dekaboruah et al., 2020). This variability makes it very difficult to define a standard composition of healthy microbes for humans. Anyway, microbes are linked to their host in many ways that shape an interactive relationship in which the

microbiota and the host co-evolve, in a mutual influence. These interactions are affected by ecological, epigenetic and genetic factors and it is difficult to disentangle the involvement of these factors in the host characteristics, from body size to diseases and in the microbial composition. The study of the holobiont described three stochastic processes behind the transmission of the microbiota in the human: dispersal (organisms can move across space), drift (changes in the population linked to time due to random birth and death) and diversification (genome changes). In fact, the study of *Helicobacter pylori* strain (Moodley et al., 2009) *Prevotella copri* (Tett et al., 2019) and *Eubacterium rectale* (Karcher et al., 2020) distributions reflects human migrations, suggesting the importance of co-evolution of bacteria and human. There are other deterministic processes, such as the interaction with factors from the environment, that determine the fitness of the microbes and their filtering and so, possibly, their variability. Abiotic factors can be the gastrointestinal pH and the oxygen concentration, which select the organisms more tolerant to these conditions. The biotic factors are the interaction between microbes colonising the same host tissue and between the microbes and the host, its cells and especially with the immune system. These relationships can be defined as mutualism, commensalism, neutralism, amensalism, parasitism or competition. Moreover, a combination of abiotic and biotic factors can generate several landscapes that can alter the microbiota: it has been proven that the composition of bacteria can change in people with different lifestyles (e.g. stress, drug intake, smoke, exercise and environmental choices such as living in a polluted area), diets (fermented food is associated to higher microbial diversity and decrease of inflammation biomarkers (Wastyk et al., 2021; Menni et al., 2021), ethnicities, host genetics, etc. (Asnicar et al., 2021; Valles-Colomer et al., 2023; Yadav and Chauhan, 2021; Swann et al., 2020). Examples of studies that tried to focus on some of these factors demonstrated that Bifidobacteria abundance is related to lactase gene expression since Bifidobacteria consume undigested dietary lactose (Kato et al., 2018) and *Akkermansia muciniphila* is involved into health benefits of polyphenols (Jayachandran, S. S. M. Chung, and Xu, 2020). Moreover, host characteristics, from body size to diseases, are functions of genetic, environmental and microbial factors. To disentangle this relationship, large scale microbiome-wide association studies are required but not largely available at the moment.

Anyway, a clear example of abiotic factor involvement in bacterial composition is provided by Asian immigrants' change in microbiome after moving to the United States, which is linked to diet change and the tendency to develop obesity (Vangay et al., 2018). Moreover, these Asian people showed higher microbial diversity than USA ones before leaving Asia: the authors claim that this diversity can not only be linked to the diet but also to the host genetics, even if this factor is still poorly understood (Mousa, Chehadeh, and Husband, 2022). These studies suggested

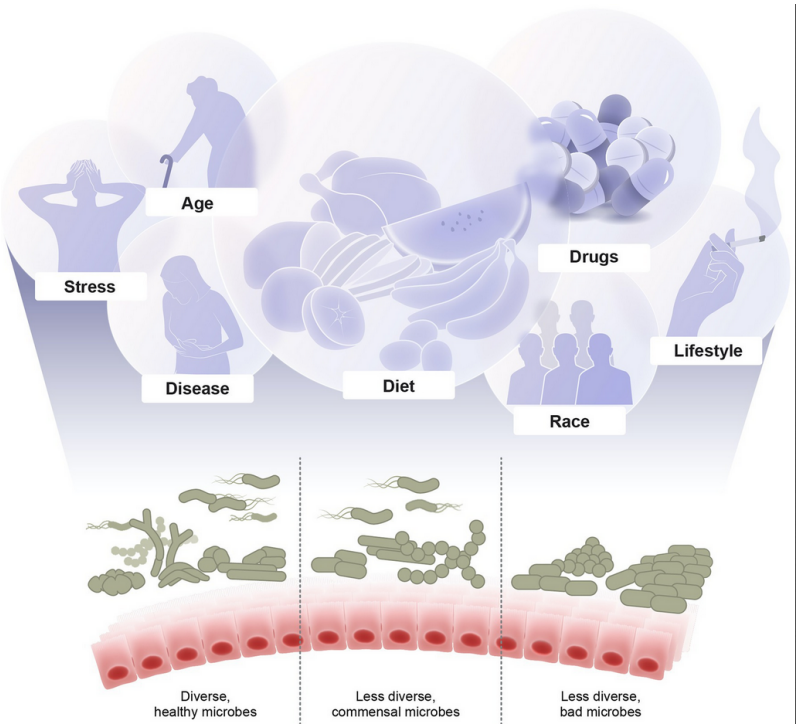


Figure 2.1: **Major factors shaping the microbiome diversity.** From Mousa, Chehadeh, and Husband, 2022

the theory of the disappearing microbiome: industrialisation leads to the loss of bacterial species, this is why Western microbiota have 15-30% fewer species than non-Western microbiota (Davenport et al., 2017). In general, all these factors contribute together to the microbiota composition, making it difficult to disentangle the impact of each one (Mousa, Chehadeh, and Husband, 2022). Other studies tried to investigate the impact of single factors, in fact it has been proven that diet can influence the microbial composition, giving selective advantages to the microbes that better metabolise the provided nutrients, as described by David et al., 2014, who showed that diet change can alter the microbial composition of the gut. Many studies focused on the relationship between microbes and humans and we are still far from understanding what is driving the differences, why there is such great diversity and how we could shape people's microbiota into the most healthy ones.

Many efforts are currently focused on elucidating the ways in which bacteria interact with their host. Furthermore, another goal in microbiota research is to go beyond the association of bacteria with functions or properties of the tissue or organism and establish causality and outcomes.

The microbial dysbiosis can drive diseases, as inflammatory bowel disease (IBD), obesity, metabolic and mental disorders (Mousa, Chehadeh, and Husband, 2022). Recently celiac disease and IBD symptoms are thought to be mediated by the microbiota and not directly by the diet itself (Krishnareddy, 2019; Glassner, Abraham, and Quigley, 2020). Gut dysbiosis is not only linked to gut diseases, but there are different communication axes to other organs, for example the gut-lung axis

or the gut-brain axis. The connection between these organs and the gut microbiota is due to the translocation of bacteria from the gut to the bloodstream that then reaches other tissues, where bacteria can increase the pro-inflammatory environment or activate immune responses (Lau et al., 2021; Irwin, 2019). Bacterial translocation to the bloodstream can also happen with diseases with the alteration of the barriers. A normal presence of bacteria in the blood is still under debate but a persistent and abundant presence of bacteria in the blood has been linked to various diseases, cancer included (H. S. Cheng et al., 2023).

Even if the mechanisms are still not understood, the gut microbiota seems to influence or be influenced by other tissue microbiotas too, for example by the lung microbiota (Mousa, Chehadeh, and Husband, 2022). In this way, an altered gut microbial composition is linked to lung disease like ashtma, pulmonary and cystic fibrosis, but it is also associated to psychiatric and neurological disorders, such as depression, sleep disorders, autism, Alzheimer disease and dementia (Renson et al., 2020; Mousa, Chehadeh, and Husband, 2022; Lau et al., 2021; Irwin, 2019; Khkheirouri, Kalejahi, and Noorazar, 2016).

2.2 Disease, cancer and microbes

As pointed out in the previous chapter, the human microbiota protects its host from illnesses but, when it is altered, this dysbiosis can cause or enhance the unhealthy status of the tissue. The relationship with the microbiota is also relevant to cancerous conditions: the first clinical study on the connection of bacteria and cancer is dated to 1868 when patients bearing a *Streptococcus pyogenes* infection showed spontaneous tumour regression (G. D. Sepich-Poore et al., 2021.) Later, the discovery of the Rous sarcoma virus (Rous, 1911) paved the path to the viral theory that indicates viruses as cancer aetiology. Many studies tried to find a viral origin for cancer growth but they failed, demolishing the viral theory. In the last decades, the link between cancer and microbes is gaining more and more interest and it is currently a hot topic in the cancer field, especially for bacteria. In fact, microbial dysbiosis has been recently added to the hallmarks of cancer (see Figure 2.2), the list of the common characteristics that neoplastic cells acquire when developing into tumour (Hanahan, 2022), as microbes from the tumour microenvironment are recognised to be involved in the growth and development of gut, lung, oral, vaginal and skin cancer, even if bacterial signals were found in other cancer tissues too, e.g. bone, brain and pancreas (Nejman et al., 2020). Even if the microbial role in cancer formation and development has been recognised as a critical point that deserves interest for the characterisation, classification and from the therapeutic point of view, there are few microbes that have been identified as causative of cancer. For the viral compartment, Human papillomavirus (HPV), linked to cervical, anal and some head and neck cancers (De Martel et al., 2017), or Hepatitis B and C viruses associated with liver cancer (Takano et al., 1995). Also bacteria can directly be involved in cancer formation and, for this reason, are proposed to be called oncobacteria, e.g. *H. pylori* and gastric cancer (Alipour, 2021) and *Fusobacterium nucleatum* in colon cancer (Abed et al., 2020).

Our knowledge about microbial involvement in cancer has highlighted a more complicated situation, in which the whole composition of the microbiota can be a contributing cause of cancer and not only the main etiological explanation. This is expected since, as discussed above, the microbiota composition takes part in the maintenance of a healthy status of the tissue and, in the same way, can take part in the disorder that drives the disease. For this reason and because of the generally high heterogeneity of human microbiota, the methods by which the microbiota enhance cancer growth and development are difficult to disentangle. Recent studies suggested some possible mechanistic ways by which bacteria interact with the host and, in this way, can affect or be affected by this relationship (Heintz-Buschart and Wilmes, 2018). One of these approaches involves the production of toxins that shapes a carcinogenic microenvironment and/or directly al-

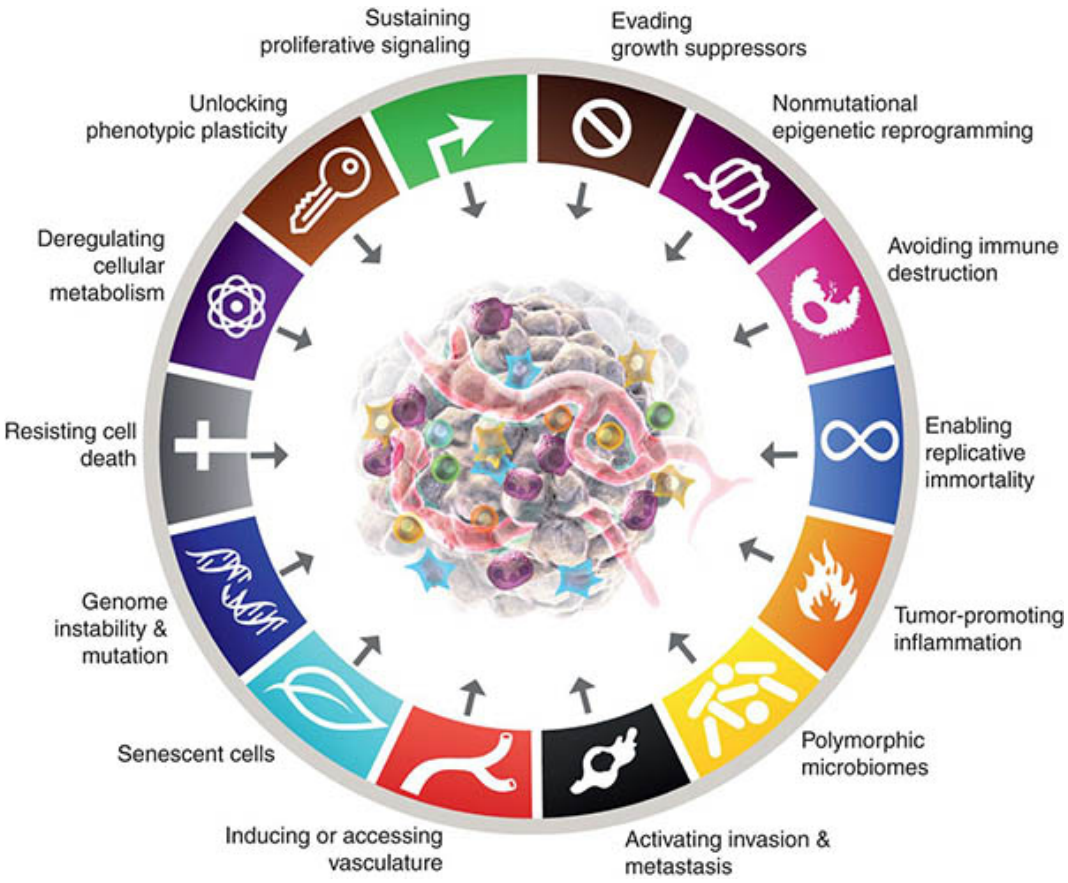


Figure 2.2: **Hallmarks of Cancer: New Dimensions.** From Hanahan, 2022

ters the host cells. For example, specific enterotoxigenic *Bacteroides fragilis* strains are known to bear the bft gene, whose protein, activated by the fragipain protease and fragilisn metalloprotease, causes barrier injuries and is associated with cancer (Valguarnera and Wardenburg, 2020; G.-T. Chung et al., 1999). *Escherichia coli* strains carry pks island, a locus that allows the production of the colibactin toxin. This toxin is known to increase intestinal cell proliferation and permeability, induce epigenetic alteration and activate protumourigenic immune responses, driving the transformation to colorectal cancer (Clay, Fonseca-Pereira, and Garrett, 2022; Nougayrède et al., 2006). Another example is given by *F. nucleatum*, an oral symbiont that can cause colorectal cancer activating cell proliferation, inducing immune cell apoptosis and, in certain situations, sustaining the chemoresistance to treatment (Clay, Fonseca-Pereira, and Garrett, 2022; N. Wang and Fang, 2022).

In a similar way, the production of specific molecules by bacteria can affect the microenvironment and directly affect the host cells. This procedure can also affect the tumour resident immune cell, for example by switching them off. Bacteria and immune cell can interact also through direct contact. The bacteria-immune cell interaction drives chronic inflammation, a situation that enhances tumour growth and development. Finally, in particular in colon, bacteria can modify the environment creating a matrix and also bacteria can alter the barrier of tissues in contact with

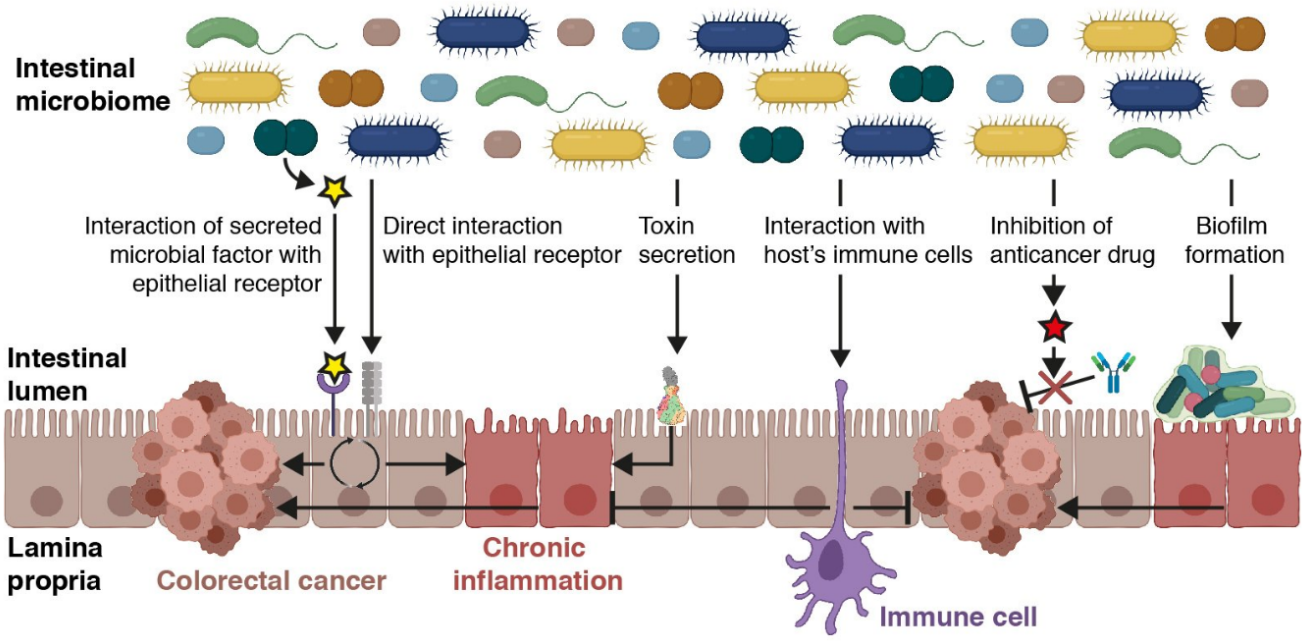


Figure 2.3: **Tumour microenvironment interactions.** Summary of some of the possible mechanisms of interaction in the tumour microenvironment. From https://leibniz-fmp.de/research/research-section/structural-biology/daniel-roderer/projects/standard-titel-2

the external body leading to an inflammation environment (which involves the immune system) that is associated with tumour development (Clay, Fonseca-Pereira, and Garrett, 2022).

These mechanisms describe a phenomenon that involves the tumour microenvironment: a tridimensional region of tissue that is characterised by many specific properties that directly initiate, enhance, sustain and/or maintain an environment fitted for the tumour. This idea to approach cancer resembles the ecological niche: as every ecological niche is populated by very different species that cooperate, compete and/or interact and are affected by chemical and physical characteristics of the environment, in the same way the tumour microenvironment presents different cells (the host healthy cells, the tumour cells, the infiltrating immune cells, the microbes and the mesenchymal cells) that cooperate (e.g. the tumour and the bacteria to switch off the immune cells), compete (e.g. immune cells and bacteria) and /or interact (e.g. the immune cells and bacteria) and are affected by the chemical (e.g. the molecules made by bacteria and food) or physical (e.g. pH and oxygen concentration) characteristics of the niche, see Figure 2.3.

Notably, all these relationships are not mutually exclusive but it is more likely that many of them are active in the same microenvironment from more than one bacterial species. It is also possible that the same mechanism is supported by different species in different hosts, due to the heterogeneity of microbial composition and the functional redundancy that can confer resilience and stabilise the ecosystem (Heintz-Buschart and Wilmes, 2018). Another aspect that must be considered is that these mechanisms can be present in different phases of tumour development,

e.g. some mechanisms can lead to cancer growth and other mechanisms sustain the development, hiding or switching off the previous ones. In this way, the bacteria involved in these mechanisms can be different: many authors suggested the driver-passenger theory (Tjalsma et al., 2012) in which some or one bacterial species are able to initiate the tumour growth, the driver, while others are only sustaining the tumour once it is part of the tumour niche. This theory depicts a first bacteria-to-tumour direction of interaction, in which the bacteria actively initiate or enhance the tumour growth and then a second tumour-to-bacteria direction in which the tumour shapes an environment that makes some bacteria fit better than others. As said before, this hypothesis can be true for some bacteria, while others cannot follow this theory.

As mentioned before, the characterisation of the tumour microenvironment is important also from a therapeutic point of view. The interaction of bacteria with the immune cell can alter immune therapies, triggering a good or bad prognosis. Also the generic chemical treatment can be affected by the altering of the absorption of the drug (Lehouritis et al., 2015). As the microbiota can alter cancer chemotherapy, even the microbiota, especially its bacteria composition, can be altered by drugs, e.g. antibiotics. It has been demonstrated that antibiotic treatments can cause colon dysbiosis, which drives severe diseases or neoplastic transformation (Morgun et al., 2015; Lange et al., 2016). In conclusion, the relationship between bacteria, disease and cancer is complex and not fully understood. While some bacteria have been associated with an increased risk of cancer, others may actually have protective effects. Certain bacteria in the gut are thought to be important for maintaining a healthy gut microbiome, which may help to prevent the development of colon cancer. On the other hand, some bacteria produce toxins or other virulence factors that can contribute to cancer development. More research is needed to better understand the complex interactions between bacteria and cancer.

2.3 Colon cancer

Colon cancer, the third most common cancer worldwide (H. Sung et al., 2021) is a complex and heterogeneous disease that can be caused by a combination of genetic and environmental factors, such as diet and lifestyle. While heredity plays an important role in colon cancer risk (30-35% of cases), with first-degree relatives of people with colon cancer having a higher risk of developing the disease themselves, other risk factors include age, diet (that comprehend processed and red meats but low in fruits and vegetables), obesity, physical inactivity, alcohol consumption and smoking (Keum and Giovannucci, 2019). Colon cancer is also the second leading cause of cancer death for men and women combined while colorectal cancer (CRC) is one of the most common cancers worldwide and a major cause of cancer-related death (H. Sung et al., 2021). However, despite its devastating potential, early detection and treatment can greatly improve outcomes (Keum and Giovannucci, 2019).

The complexity of colon cancer is further demonstrated by the various molecular alterations implicated in the pathogenesis of colon cancer, which have distinct clinical presentations, prognoses and treatment options. These multiple molecular alterations include mutations in oncogenes and tumour suppressor genes, deoxyribonucleic acid (DNA) methylation changes, alterations in microribonucleic acid (RNA) expression and in the epigenome. Understanding the complex nature of colon cancer is essential for improving screening, diagnosis and treatment for this deadly disease (Malki et al., 2020).

2.3.1 Mutated genes

One of the most commonly mutated tumour suppressor genes in colon cancer is the adenomatous polyposis coli (APC) gene, located on chromosome 5 (mutated in the 76% of patients), which plays a crucial role in several cellular processes (M. S. Lawrence et al., 2014). The most relevant role for cancer progression is its involvement in regulating cell growth and division by keeping cells from growing and dividing too fast or in an uncontrolled way. It helps control how often a cell divides, how it attaches to other cells within a tissue and whether a cell moves within or away from the tissue. Mutations in the APC gene can lead to the development of various cancers, such as those of the liver, endometrial cancer and gastric cancer but mutations in this gene are found in a majority of individuals with a genetic predisposition for hereditary colorectal cancer (mutations in the APC gene can cause a rare inherited disorder called familial adenomatous polyposis (FAP) that increase the risk of developing colorectal cancer, Aghabozorgi et al., 2019). In addition, somatic alterations in the APC gene have been identified in many sporadic cases of colorectal cancer.

The function of this gene is to regulate cell growth and division and mutations can disrupt this process leading to uncontrolled cell growth and the development of cancer (Dow et al., 2015).
Mutations in the Kirsten rat sarcoma viral oncogene homolog (KRAS), which regulates cell signalling pathways, are also frequently found in colon cancer. The KRAS gene is a member of the RAS family of genes that are involved in signal transduction pathways and it is a member of the small GTPase superfamily that encodes a protein involved in transmitting signals within cells' pathways. It is located on chromosome 12 and encodes a protein that is involved in the regulation of cell growth, differentiation and survival (Imperiale et al., 2014). Mutations in the KRAS gene are among the most common genetic alterations in human cancer and are associated with a variety of malignancies, including lung, CRC, pancreatic and ovarian cancer and is also involved in the development of certain types of leukaemia (Timar and Kashofer, 2020).
Other molecular features of colon cancer include alterations in the Wnt signalling and the PI3K/Akt/mTOR pathway. Wnt signalling is a complex and highly conserved pathway that plays a crucial role in cell proliferation, differentiation, development and homeostasis. The pathway is activated by Wnt proteins, which bind to the Frizzled family of receptors on the cell surface, leading to the activation of intracellular signalling cascades that ultimately regulate gene expression. For example, Wnt pathway can activate the transcription factors β-catenin, which then translocates to the nucleus where it binds to T-cell factor/lymphoid enhancer factor (Y. Zhang and X. Wang, 2020). Aberrant activation of the Wnt signalling pathway due to mutations in pathway components, such as APC, β-catenin, and Axin, have been linked to the development of colorectal cancer and other malignancies. Additionally, recent studies have identified multiple interactions between Wnt signalling and other pathways, such as the PI3K/Akt and ERK/MAPK pathways, which further highlight the complexity and significance of Wnt signalling in disease development (Flores-Hernández et al., 2020).
Another pathway that is often altered in colon cancer is the transforming growth factor β (TGF-β) signalling pathway. The TGF-β pathway is a signalling pathway involved in many cellular processes, including cell growth, differentiation, apoptosis and migration. It is activated by the binding of TGF-β to its receptor, which then activates a series of intracellular signalling molecules, including the smad pathway. These molecules then activate downstream pathways that lead to the desired cellular response (Itatani, Kawada, and Sakai, 2019). The TGF-β pathway is important for normal development and tissue homeostasis and its dysregulation has been linked to many diseases, including cancer, fibrosis and immune disorders . In cancer, the TGF-β pathway has a dual role, acting as both a tumour suppressor and a tumour promoter, depending on the stage of tumour development, microenvironment and cellular context. Recent studies have also identified

crosstalk between the TGF-β pathway and other signalling pathways, further highlighting the complexity of this pathway and its critical role in disease development and progression (Germann et al., 2020).

Furthermore, mutations in the TP53 and retinoblastoma (RB) pathways are common in colon cancer and are associated with a better prognosis. TP53 and RB genes are two of the most well-known tumour suppressor genes that play a crucial role in regulating cell growth, division, DNA repair and apoptosis. The TP53 gene encodes for the p53 protein, which plays a vital role in preventing cancer through cell cycle arrest and apoptosis induction in response to DNA damage or other cellular stresses, thereby preventing the accumulation of damaged cells. Mutations in the TP53 gene are among the most common genetic alterations in human cancer and loss of p53 function is associated with increased tumour development and progression (Bukholm and Nesland, 2000; M. Michel et al., 2021).

The RB gene encodes for the retinoblastoma protein (pRB), which acts as a tumour suppressor by inhibiting cell cycle progression and promoting differentiation and regulates the G1/S checkpoint of the cell cycle by inhibiting the activity of the E2F transcription factors, thereby preventing the inappropriate proliferation of cells. Mutations in TP53 or RB genes can cause aberrant cell proliferation and tumorigenesis, e.g. retinoblastoma and osteosarcoma. Studies have shown that the TP53 and RB pathways interact with each other and with other signalling pathways, such as the Wnt and TGF-β pathways, to regulate cell proliferation and tumour development (M. Michel et al., 2021).

2.3.2 Molecular properties

Beyond gene mutations and pathway alterations, understanding the molecular basis of colon cancer is critical for developing targeted therapies for colon cancer and improving patient outcomes. The molecular characterisation of CRC has revealed distinct properties, including chromosomal instability (CIN), CpG methylation phenotype (CIMP) and microsatellite instability (MSI). The complexity of studying these characteristics of colon cancer is amplified by the overlap and connection between mutations in the described genes and all the molecular phenotypes that altogether cooperate in cancer growth and progression (Singh et al., 2021; Keum and Giovannucci, 2019).

The most well-known characteristic of colon cancer is its MSI, which arises due to defects in the DNA mismatch repair (MMR) pathway due to defects in DNA MMR machinery, leading to the accumulation of errors at repetitive DNA sequences called microsatellites (repeating DNA sequences consisting of a small number of nucleotide pairs, usually 2-6 bp, that are dispersed throughout the genome). MSI is caused by mutations in one or more of the MMR genes, including MLH1, MSH2,

MSH6 and PMS2, which are inherited or acquired somatically during tumour development. It is characterised by a high rate of mutation and a strong association with the immune system. MSI occurs when there is a change in the number of repeats within these microsatellites, leading to frameshift mutations in the downstream genes (Haiwei Wang et al., 2019). MSI is frequently observed in a wide range of malignancies, including CRC cancer, gastric cancer, endometrial cancer and ovarian cancer and is often associated with a favourable response to immune checkpoint inhibitors. In CRC cancer, MSI is a hallmark of the Lynch syndrome and is associated with a better prognosis as compared to microsatellite-stable tumours (Laghi et al., 2020; Sawhney et al., 2006).

In addition, another most widely studied characteristic of colon cancer is CIN, which involves changes in the number and structure of chromosomes, e.g. chromosomal rearrangements and chromosomal numerical abnormalities, and is associated with a poor prognosis (Grady, 2004). The CIN subtype is characterised by frequent chromosomal abnormalities and mutations in *APC*, *TP53* and *KRAS* genes and it can arise due to defects in chromosomal segregation during cell division, DNA damage, or errors in DNA repair pathways. This instability has been linked to the development of many different types of cancer, as well as other genetic disorders mentioned above. It can be caused by a variety of factors, including genetic mutations, environmental exposures and certain medical treatments, so the mechanisms underlying chromosome instability are still debated. CIN can lead to a wide range of health problems, including birth defects, developmental delays and cancer (W. Zhao et al., 2022).

Another CRC cancer phenotype involves the epigenome, specifically the addition of a methyl group on cytosine-phosphate-guanine (CpG) dinucleotide islands. The CpG methylation plays a crucial role in regulating gene expression, as it can affect the binding of transcription factors and other DNA-binding proteins to the DNA, thereby controlling the accessibility of the gene for transcription (Arain et al., 2010). CIMP refers to the pattern of CpG hypermethylation across the genome and can have profound effects on an organism's development and health: aberrant CpG methylation patterns have been linked not only to cancer but also to a variety of diseases, e.g. neurological disorders (Grayson and Guidotti, 2013). In the context of cancer, methylation at CpG sites can lead to the silencing of tumour suppressor genes and activation of oncogenes. For example, hypermethylation of the promoter regions of tumour suppressor genes such as p16INK4a, BRCA1 and MLH1 has been observed in various types of cancer, including breast, ovarian and CRC cancer. This hypermethylation results in the inactivation of these genes, which can promote tumourigenesis by allowing for uncontrolled cell growth and proliferation. On the other hand, hypomethylation of certain CpG sites in the genome has also been linked to cancer development

and progression, as it can lead to genomic instability and altered gene expression (Malki et al., 2020).

Given the various molecular characteristics described for colon cancer, several classification systems have been developed to improve our understanding of its molecular characteristics and to guide treatment decisions. One widely used classification is the consensus molecular subtype (CMS), which divides colon cancer into four molecular subtypes (CMS1, CMS2, CMS3 and CMS4) based on various molecular features, including mutations, DNA methylation and gene expression patterns, see Figure 2.4. Each subtype is associated with distinct clinical and pathological characteristics, prognosis and response to therapy. For instance, CMS1 tumours are characterised by immune activation and are more likely to respond to immunotherapy, while CMS4 tumours are characterised by stromal invasion and have a poorer prognosis. CMS2 tumours are charac-

CMS1 MSI immune	CMS2 Canonical	CMS3 Metabolic	CMS4 Mesenchymal
14%	37%	13%	23%
MSI, CIMP high, hypermutation	SCNA high	Mixed MSI status, SCNA low, CIMP low	SCNA high
BRAF mutations		*KRAS* mutations	
Immune infiltration and activation	WNT and MYC activation	Metabolic deregulation	Stromal infiltration, TGF-β activation, angiogenesis
Worse survival after relapse			Worse relapse-free and overall survival

Figure 2.4: **CMS biological differences.** From Guinney et al., 2015

terised by upregulation of genes involved in epithelial differentiation and metabolism and they often show canonical WNT signalling activation, while CMS3 tumours have prominent metabolic dysregulation and often harbour mutations in KRAS or BRAF oncogenes, which are known to drive tumourigenesis in colon cancer, as described above (Guinney et al., 2015).

2.3.3 Colon location

These tumour properties can manifest within the same tumour, e.g. CIMP high tumours are enriched in MSI high tumours (Goel et al., 2007), and are not uniformly distributed along the colon. In fact, traditionally, colon cancer was studied based on its location of origin: either the left or right part of the colon. Tumours arising from different colon sides exhibit distinct characteristics, such as molecular properties, progression, and survival outcomes, which necessitate specific treatments (a short summary is depicted in Figure 2.5). This differentiation comes from various factors, including the embryonic origins of the two sides: the right part develops from the

midgut and the left one from the hindgut. Furthermore, the blood supply to the two sides of the colon varies, with the midgut being provided by the superior mesenteric arteries and the hindgut by the inferior ones (Lee, Menter, and Kopetz, 2017).

Tumours originating from different sides of the colon exhibit several notable distinctions. For example, the gastrointestinal structure itself exposes the right side of the colon to potential carcinogenic factors such as bile acids (Venturi et al., 1997). Moreover, tumours found in the right part of the colon tend to consist of poorly differentiated cells and are classified as sessile adenomas or mucinous adenocarcinomas. Mucinous adenomas are characterised by faster progression and are frequently diagnosed in patients with inflammatory bowel disease (Hugen et al., 2016). Interestingly, right-sided tumours have a higher incidence in females and older individuals (Nawa et al., 2008). Furthermore, hereditary cancer syndromes, excluding FAP, occur more frequently in the right part of the colon (Sarvepalli et al., 2018). On the other hand, left-sided tumours are tubular villous adenocarcinomas with a polypoid morphology, which makes them easier to detect and improves the prognosis for patients with this type of disease (Gualco et al., 2006). Unlike the right colon, these tumours more commonly arise in males at an earlier age (Nawa et al., 2008). From a molecular perspective, the two sides of the colon frequently exhibit distinct properties.

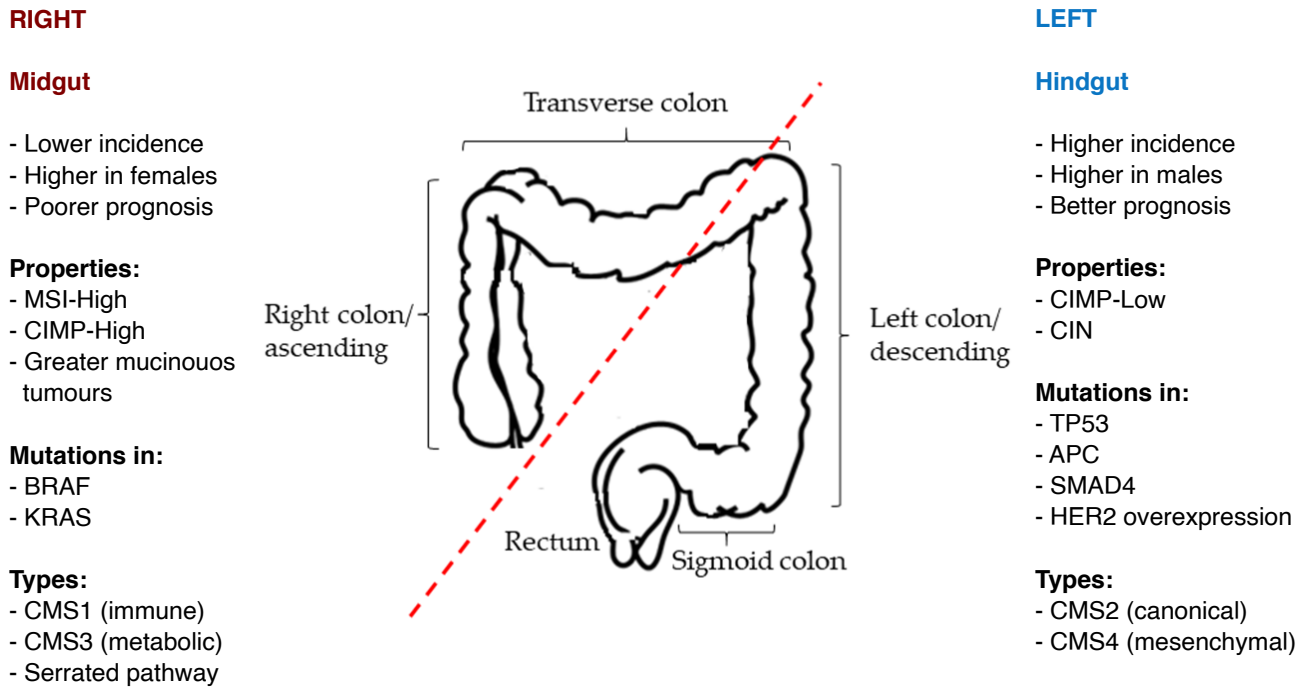


Figure 2.5: **Right-sided vs. left-sided CRC characteristics.** Adapted from Ahmad Zawawi and Musa, 2022

Tumours on the right side are more commonly characterised by high MSI, leading to increased immune infiltration due to the presence of immunogenic mutations. The infiltration of immune cells, particularly lymphocytes, is associated with improved prognosis and reduced metastasis

(Galon et al., 2006; Ogino et al., 2009. These characteristics make MSI-high tumours a promising target for immunotherapy (Le et al., 2015). Additionally, right-sided tumours often exhibit high CIMP and harbor BRAF mutations. Furthermore, in the right side of the colon, CMS1 and CMS3 tumours are more frequently observed. On the other hand, tumours on the left side of the colon tend to show CIN, following the traditional pathway.

Another notable distinction observed in colon tumours growing in the left or right colon is the site of metastasis. Right-sided tumours have a tendency to preferentially metastasise to the peritoneum, whereas left-sided tumours more commonly metastasise to the thorax or bones (Riihimäki et al., 2016).

2.3.4 Bacterial involvement

As mentioned in the previous chapter, there is increasing evidence suggesting that the bacteria in the gut may play a role in the development and progression of colon cancer. The gut microbiota can influence various biological processes, including inflammation, immune response and metabolism, processes that are known to be crucial in cancer growth and development. Studies have shown that alterations in the gut microbiota, such as dysbiosis or an imbalance in bacterial diversity and composition, can contribute to the development of CRC cancer, see Figure 2.6. For example, certain bacteria, such as *F. nucleatum* and *B. fragilis*, have been found to be more abundant in colon cancer tissue than in healthy tissue (Garrett, 2019). *F. nucleatum* is a gram-negative anaerobic bacterium that is commonly found in the oral cavity and gut microbiota. *F. nucleatum* can promote tumourigenesis by several mechanisms, including inducing inflammation, inhibiting immune response and modulating the gut microenvironment (Brennan and Garrett, 2019). It can also enhance the proliferation and invasion of cancer cells by activating oncogenic pathways, such as WNT signalling and epithelial-mesenchymal transition pathway (Rubinstein et al., 2019). Furthermore, *F. nucleatum* has been associated with a poorer prognosis and a higher risk of recurrence in patients with colon cancer (Dejea et al., 2018). Targeting *F. nucleatum* and its associated pathways may represent a promising strategy for the prevention and treatment of colon cancer (Bullman et al., 2017). *B. fragilis* is a gram-negative anaerobic bacterium that is a common constituent of the gut microbiota. This bacterium has been associated with several pathogenic mechanisms, including induction of inflammation, DNA damage and alteration of the gut microenvironment. *B. fragilis* produces a toxin called BFT (*B. fragilis* toxin), which can cause DNA damage and inhibit the function of T-cells, leading to impaired immune response. Additionally, *B. fragilis* can activate the signal transducer and activator of transcription 3 (STAT3)

pathway, which is involved in the regulation of cell proliferation, apoptosis and immune response (W. T. Cheng, Kantilal, and Davamani, 2020).

On the other hand, probiotics and prebiotics, which can modulate the gut microbiota, have been proposed as potential strategies to prevent or treat colon cancer (Holscher, 2017). Further research is needed to better understand the complex interactions between the gut microbiota and colon cancer and to identify novel therapeutic targets for this disease.

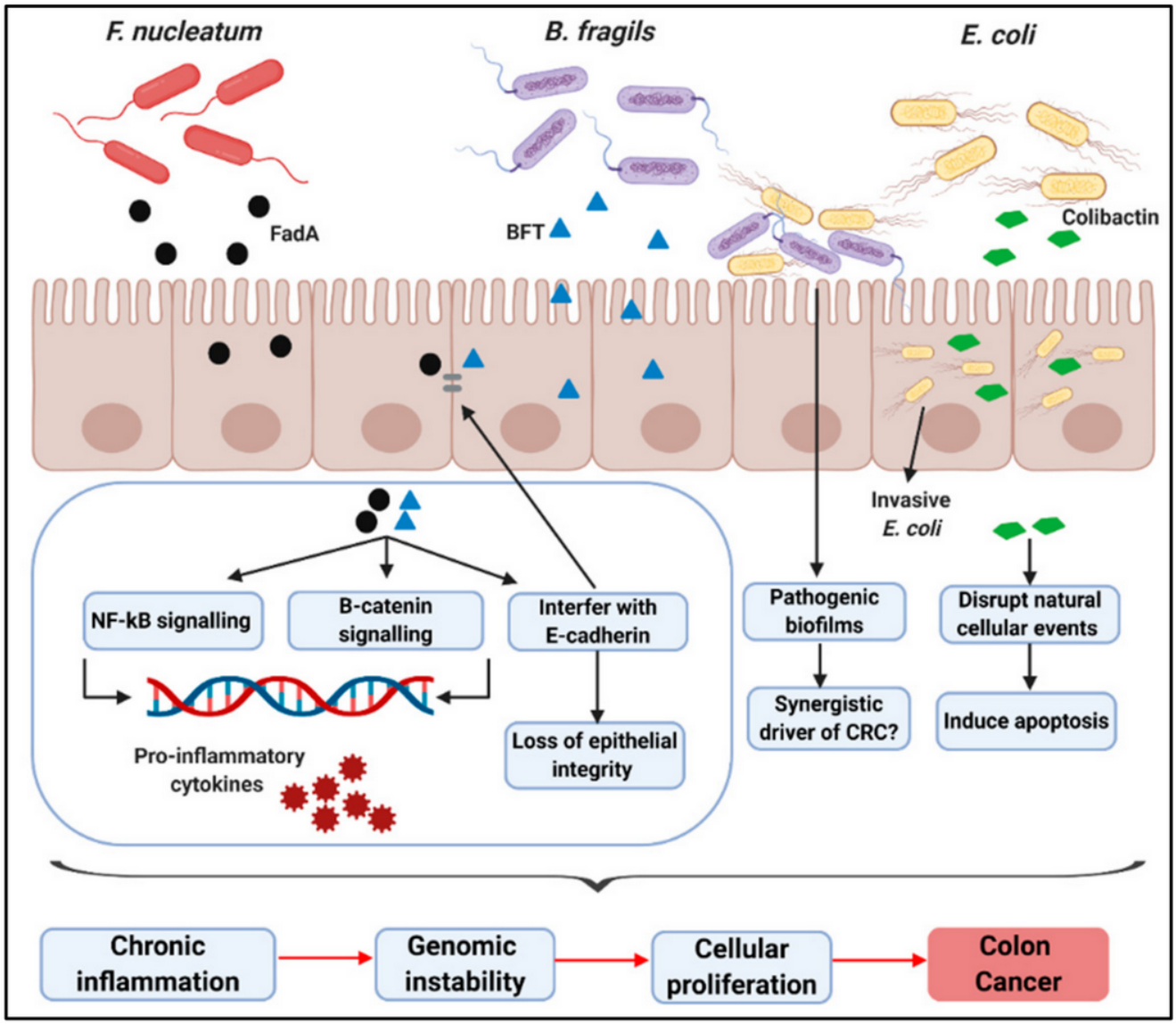


Figure 2.6: **Overview of bacterial mechanisms associated with CRC development.** From G. W. Lawrence et al., 2020

2.4 Bacterial detection methods

The first approaches to study bacteria date back to the 17th century, when Antonie van Leeuwenhoek observed and described microorganisms using a simple microscope he designed. However, it was not until the late 19th century that the field of bacteriology emerged as a scientific discipline, largely due to the pioneering work of Louis Pasteur, who developed methods to culture and isolate bacteria, as described in this 200th birthday celebration article (Høiby, 2022). Since then, many other approaches have been developed to study bacteria, the most common include microscopy, culture-based methods and molecular biology techniques. Microscopy allows for the direct observation of bacteria under a microscope and it is useful for studying the morphology of bacteria. Culture-based methods involve the growth of bacteria on nutrient-rich media to obtain pure cultures, which can be used for further analysis, such as biochemical and genetic tests (Lagier et al., 2015).

However, both of these methods have limitations, as microscopy can only provide information on the morphology of bacteria and usually cannot identify specific bacterial species or strains, while culture-based methods may fail to detect the bacterial species that cannot grow in laboratory, including those that require specific growth conditions or that are present in low abundance. Additionally, while some bacteria are viable, other difficult-to-culture or non-culturable bacteria may be missed by culture-based methods (Riley, 2017).

These limitations have led to the development of molecular biology techniques, such as polymerase chain reaction (PCR) and metagenomics, which allow for the direct detection and analysis of bacterial DNA from various samples, without the need for culture-based methods (Miller and Chiu, 2022).

Molecular biology techniques, such as fluorescent in situ hybridization (FISH), PCR and metagenomics have revolutionised the study of bacteria by allowing for the direct detection and analysis of bacterial DNA from various samples. These techniques are highly sensitive and specific and can be used to identify and quantify bacterial species and strains, as well as study the genetic diversity and evolution of bacterial populations. The combination of these approaches has led to significant advances in our understanding of bacterial diversity, ecology and pathogenesis, and has contributed to the development of new diagnostic and therapeutic strategies (Miller and Chiu, 2022).

FISH is a technique used to detect specific bacterial species or strains in samples that involves the use of fluorescently labelled DNA probes that are complementary to specific regions of bacterial genome. These probes bind to the bacterial DNA in the sample, allowing for the visualisation

and identification of the targeted bacteria under a fluorescent microscope. FISH is a sensitive and specific method for detecting bacteria, making it a valuable tool in microbiology research and clinical diagnosis (Prudent and Raoult, 2019).

Metagenomics studies the collective DNA of microbial communities present in a particular environment through NGS techniques. It allows the estimate of the genetic diversity and functional potential of microbial populations and can be applied in various fields, including environmental science, human health, agriculture, and biotechnology (Wooley, Godzik, and Friedberg, 2010).

ribosomal RNA 16S gene (16S) sequencing is a technique used to analyse the genetic material of bacteria and identify their taxonomic classification (Peterson et al., 2021). The 16S is a conserved region of bacterial DNA that codes for the 16S, an essential component of the bacterial ribosome. The 16S gene is present in all bacteria but its sequence varies between species, allowing for the identification and classification of different bacterial groups. 16S sequencing involves the amplification and sequencing of the 16S from a bacterial sample, which is then compared to a database of known 16S sequences to determine the bacterial taxonomy (J. S. Johnson et al., 2019). Even if not common due to the issues related to low biomass samples (Karstens et al., 2019; Eisenhofer et al., 2019; Kennedy et al., 2023), 16S sequencing can be applied to analyse bacteria from tissue (Nejman et al., 2020).

A novel method has emerged for identifying bacteria residing in tissue samples, which involves the extraction of bacterial reads from human NGS analyses. Human NGS techniques, such as RNA Sequencing (RNA-Seq), whole genome sequencing (WGS) and whole exome sequencing (WXS), have long been employed to investigate the transcriptome and genome of human samples for various research purposes (Q. Wang et al., 2023). The RNA-Seq approach was developed for studying gene expression and provides a comprehensive view of the transcriptome of a sample. It is particularly useful for identifying differentially expressed genes in various disease conditions and showing the alternative splicing of genes (Haas and Zody, 2010). WXS has been used to identify genetic variations in the protein-coding regions of the genome. This technique has been particularly useful for identifying mutations associated with various genetic diseases, including cancer. By sequencing only the exome, WXS enables researchers to focus on the most functionally relevant portions of the genome, allowing for a more efficient analysis of genetic variations (Lalonde et al., 2010). Finally, WGS sequences the entire genome, including non-coding regions, to provide a comprehensive view of an individual's genetic profile. This approach has been used to study genetic diversity and evolution, as well as to identify genetic variations associated with various diseases (Meienberg et al., 2016). Several authors have recently demonstrated that microbial signals can be detected in human NGS analyses, enabling the characterisation of the tissue-resident microbiota (Q. Wang

et al., 2023). This novel approach is particularly valuable because it allows for the simultaneous analysis of both human and microbial components from the same experiment. Moreover, NGS techniques such as RNA-Seq, WGS and WXS have become increasingly common in the study of human characteristics and have been applied to various sample types so the ability to extract microbial signals from these samples is of great importance because it enables comparisons between human characteristics and microbial populations, enabling the reconstruction of microbial communities from large human databases. This approach has the potential to significantly expand our understanding of the microbial world and its relationship to human health and disease by enabling the analysis of thousands of samples with different characteristics.

One of the earliest papers to take advantage of this approach was Poore et al., 2020, who extracted bacterial reads from the cancer genome atlas (TCGA) data to reconstruct the microbiome of various samples, enabling the tumour location prediction from the analysis of blood derived microbiota. Similarly, Dohlman et al., 2021, utilised WXS and WGS data from TCGA to create an atlas of tumour-resident bacteria. Other authors have used these data to explore other hypotheses, such as the association of bacteria with patient survival (Hermida, Gertz, and Ruppin, 2022; J. Wang et al., 2021). Additionally, smaller datasets have been analysed to investigate the association between the reconstructed microbiome and tumour stage (Bullman et al., 2017). Despite the increasing use of this approach, there has yet to be a comprehensive investigation into the association between the reconstructed microbiome and tumour properties.

Overall, this approach has the potential to significantly improve our understanding of the complex interactions between bacteria and human tissues, with important implications for fields such as medicine and microbiology. By utilising large tumour databases, researchers can gain a more comprehensive understanding of the role of the microbiome in cancer and potentially identify new targets for treatment and prevention.

Materials and methods

3.1 Samples

One of the goals of this work is to detect the microbial signal from human NGS cancer data. We took advantage of TCGA, the most important human NGS cancer database, and we internally collected a cohort of 30 colon cancer patients from the European Institute of Oncology (IEO) in Milan, Italy.

3.1.1 TCGA

TCGA collects data from 11'315 cancer patients of different ages, gender and origin; samples were taken from 54 different primary sites and 33 different cancer types from different hospitals and countries (Grossman et al., 2016). This large number of samples makes TCGA one of the biggest databases of NGS human cancer data. TCGA analysed samples with 10 experimental strategies, between them we find RNA-Seq, WXS and WGS, whose data can be downloaded either as processed data, e.g. gene counts, or as raw data, e.g. BAM files. The availability of raw data can be acquired after a proper request, given the sensibility of this type of human data.
We requested and obtained accession to the raw human files and downloaded 3'737 primary tumours and 318 solid tissue normal RNA-Seq BAM files from 8 different cancer types, see Table 3.1. The solid tissue normal samples were resected during the biopsy from a non-malignant portion of the tissue affected by the tumour, at 2 to 10cm far away from the tumour. For this reason, the sample is considered "normal", meaning that the tissue is composed of non-malignant cells, even if it was taken from a cancer patient.

For a few cancer types, samples have been analysed using different sample preparations and this could introduce intra-study bias, so we restricted the samples to Illumina Truseq method (in colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD) and breast carcinoma (BRCA)) and AllPrep RNA extraction (in ovarian serous cystadenocarcinoma (OV)), while for glioblastoma

Cancer type id	Cancer type	# of primary tumour samples	# of solid tissue normal samples
COAD	Colon adenocarcinoma	382	39
GBM	Glioblastoma multiforme	152	5
LUAD	Lung adenocarcinoma	512	59
LUSC	Lung squamous cell carcinoma	499	49
HNSC	Head and neck squamous cell carcinoma	499	44
OV	Ovarian serous cystadenocarcinoma	366	0
READ	Rectum adenocarcinoma	151	9
SKCM	Skin cutaneous melanoma	103	1
BRCA	Breast carcinoma	1073	112

Table 3.1: **TCGA RNA-Seq number of samples.** Number of RNA-Seq samples obtained from TCGA and used in this project by cancer and sample type.

multiforme (GBM) and lung squamous cell carcinoma (LUSC), we removed the cases with duplicated samples. The sample preparation information was downloaded from the Genomic Data Commons (GDC) Legacy Archive (Grossman et al., 2016). These RNA-Seq samples described have been outlined in Sambruni et al., 2023.

To compare RNA-Seq to other NGS methods data, we downloaded the BAM files of WGS and WXS from COAD patients whose samples have been analysed with these three approaches. We downloaded 33 WXS and 33 WGS tumour samples.

3.1.2 IEO

We enrolled 30 colon cancer patients from the IEO. All the patients signed informed consent and the IEO Ethical Committee approved the study with the number IEO 1149. From each patient, we retrieved a tissue sample from the tumour and two non-tumour adjacent regions (at 2 and 10 cm from the border of the pathologist-assessed neoplastic lesion) for a total of 90 samples. These samples underwent RNA-Seq, 16S sequencing and bacterial FISH: in RNA-Seq experiment, two samples with low amounts of bacterial reads were removed from the analyses, while subsets of samples were selected for the other two approaches. FISH was done to detect three bacteria: *Faecalibacterium prausnitzii*, *A. muciniphila* and *F. nucleatum*: these three bacteria are involved in colorectal microbiota in healthy status (*F. prausnitzii* and *A. muciniphila*, Verhoog et al., 2019) or in colorectal tumour (*F. nucleatum*, Abed et al., 2020).

The raw FASTQ files are available under controlled access from the European Genome-Phenome Archive (EGA), with the code EGAD00001009635.

These samples are described in Sambruni et al., 2023.

3.2 Bacterial detection methods

3.2.1 RNA extraction and sequencing analysis

30 IEO samples (tumour and non-tumour ones) underwent this method. The RNA was extracted with the AllPrep DNA/RNA kit (from Qiagen, manufacturer recommendations) from flash-frozen tissues. Illumina Truseq or Stranded Total RNA Prep Ligation with Ribo-Zero Plus kit (from Illumina) were used to prepare the RNA library with 100ng of extracted RNA. The process consists of the rRNA depletion, followed by the RNA fragmentation at 94° C for 2 minutes. After retrotranscription and anchors ligation, the library was amplified (13 cycles). We used Bioanalyser to measure the sample quality and Qubit to test sample quantity. We then sequenced the samples with Illumina NovaSeq 6000 with 50bp paired-end reads.
The following bioinformatic analyses were conducted following GDC/TCGA approaches to avoid in-silico technical biases when comparing IEO and TCGA samples. For this reason, we used the same tools and parameters defined by TCGA: we aligned the IEO FASTQ files using STAR (version v2.2.7a, GRCh38) on the IEO cluster with the parameters listed in TCGA documentation (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/).
This approach is described in Sambruni et al., 2023.

3.2.2 DNA extraction and 16S sequencing analysis

For each sample, a mucosal tissue scrap was taken and transferred to 1ml of phosphate buffered saline (PBS) solution. 500μl of this solution was combined with DNA Power Soil Pro Isolation kit (Qiagen) to extract the DNA. The extracted DNA was then quantified by Qubit and quality assessed by Nanodrop to then amplify and sequence the 16S V3-V4 regions by 16S Metagenomic Sequencing Library Preparation protocol (Illumina, 2013). In this protocol, we first performed 25 PCR cycles with 16S V3-V4 primers under the manufacturer's instructions with both forward and reverse primers (see Table 3.2) composed of Illumina overhang adapter sequences and the specific 16S sequences of primers.

forward	5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3'
reverse	5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'

Table 3.2: **16S sequencing primers.** Sequences of the 16S sequencing primers used for the first 25 PCR cycles in the 16S Metagenomic Sequencing Library Preparation protocol.

The second PCR of 8 cycles was conducted with Nextera XT Index kit to join the dual indices

and the Illumina sequencing adapters. After this, the DNA was pooled and its quality and quantity were controlled by Bioanalyser and Qubit. The DNA was then run on a MiSeq flowcell (Illumina), followed by 16S V3-V4 amplicon sequencing. Qiime2 package (versione v.2018.11, Bolyen et al., 2019) supported further analysis. Raw count tables were produced (q2-dada2 (version v.1.6.0, Callahan et al., 2016) truncation length parameters set to the primer length) and the taxonomic profiling was executed by trimming whole-gene 16S sequences from the SILVA 132 database (Quast et al., 2012) by V3-V4 flanking region. Q2-classifier skclassify plug-in (version v.2019.7.0, Pedregosa et al., 2011) was trained with these trimmed SILVIA sequences and then it was run on DADA2 representative sequences (Callahan et al., 2016): the majority of them were resolved to genus level. The taxonomy table was collapsed to genus level and transformed into relative frequencies.

This approach is described in Sambruni et al., 2023.

3.2.3 Comparison between RNA-Seq and 16S

To compare the bacteria signals detected in RNA-Seq and 16S datasets of the IEO cohort, we applied the Spearman correlation test on the bacteria genera quantification of these two approaches. We applied this analysis only to the bacteria detected by both approaches and, since the presence of many zeros can affect our correlations, we applied prevalence filters to both datasets to remove the rare and low-abundance taxa. Specifically, we fixed six thresholds based on bacteria prevalence in both methods, i.e. bacteria genera were selected if their relative abundance percentage was above 0%, 10%, 20%, 30%, 40%, 50% and 60% in the two datasets. The results of these analyses were shown in a density plot of Spearman coefficients across the thresholds of prevalence (ridge plot) and one-sample Wilcoxon test was applied on each threshold to evaluate if the median distribution of the Spearman coefficients was greater than zero. This approach was applied to compare PathSeq (Walker et al., 2018) and Kraken2 (Wood, Lu, and Langmead, 2019) results. This approach is described in Sambruni et al., 2023.

3.2.4 FISH

A subset of the IEO cohort patient tumour tissues underwent Carnoy's fixation, was paraffin-embedded and then submitted to a modified version of the protocol described by Greuter et al., 2016, to be analysed by FISH. Tumour tissue slides were deparaffinated, rehydrated and incubated for 3 hours with FISH probes with hybridisation buffer. We incubated the samples with bacteria universal probes (EUB probe) in combination with probes targeting one of the following species specifically: *A. muciniphila*, *F. prausnitzii* and *F. nucleatum*, see Table 3.3. The temperature of

incubation and the amount of formamide depends on each probe specifics.

GCTGCCTCCCGTAGGAGT	EUB338	CY3			Eubacteria
CCTTGCGGTTGGCTTCAGAT	MUC1437	CY5	48°C	formamide 30%	*A. muciniphila*
GTGCCCAGTAGGCCGCCTTC	FP698	CY5	50°C	formamide 0%	*F. prausnitzii*
CTTGTAGTTCCGCYTACCTC	FUS664	CY5	48°C	formamide 50%	*F. nucleatum*

Table 3.3: **FISH probes.** Sequences of the probes used to analyse IEO cohort samples with FISH.

After this, slides were washed, nuclei stained with 4',6-Diamidino-2-phenylindole (DAPI) and mounted to perform image acquisition with SP8 confocal microscope (Leica) at 63X magnification. This approach is described in Sambruni et al., 2023.

3.2.5 Comparison between RNA-Seq and FISH

To compare the bacteria signals detected in RNA-Seq and FISH datasets of the IEO cohort, we applied the Spearman correlation test on the bacteria species quantification of these two approaches. The bacterial species counts from FISH images were normalised by the number of EUB positive signals and total cells (DAPI) present in the images.
This approach is described in Sambruni et al., 2023.

3.3 Tumour properties

TCGA provides many clinical and technical information about the patients and their samples so we selected the ones that were the most relevant, available for most of the patients, lowly redundant and potentially associated with the microbiota composition. As clinical properties, we choose gender, body mass index (BMI), stage, history of other malignancy, location (side), age at initial pathological diagnosis, history of colon polyps and percentage of normal cells. We selected also the technical properties that we assumed could potentially influence our results. Moreover, TCGA has been deeply investigated and many relevant properties have been discovered and measured in previous studies. Thus, we expanded our analysis with the MSI level from Bonneville et al., 2017: we classified as MSI high those samples with a MANTIS score > 0.4 and with a low MSI the ones with MANTIS score $<$ or equal to 0.4, as suggested by the authors. We also included the CIMP status (Y. Liu et al., 2018), the CMS classification (measured from the tumour gene expression profile with the CMSclassifier R package (Guinney et al., 2015)), the stemness value (Malta et al., 2018) and the aneuploidy status (Taylor et al., 2018). For each TCGA sample, we added the mutation status of the most frequently mutated genes of each cancer type (M. S. Lawrence et al., 2014) from the GDC database collection (Ellrott et al., 2018): we considered a gene mutated if it carries any type of non-silent mutation (silent mutations: silent, 5'flank, RNA, intron, 3'flank). We also included the total number of mutations per sample.

Whenever we needed to compare continuous properties we used appropriate tests (i.e. Spearman or Pearson test) and, when needed (e.g. the independence test), we binned the continuous properties to convert them to discrete ones by automatically choosing the breakpoints. To this end, we decided that if the frequency of zero values is over 30%, we considered the presence or absence (anything above zero is considered as presence); if the distribution was normal (Shapiro test) we split the samples into two groups with lower or higher values from the mean; if the distribution was bimodal (is.bimodal function from LaplacesDemon R package), we defined low and high values taking the lowest value between the two peaks as breakpoint; and if none of the previous conditions was satisfied, we binned the values by quartiles (low, medium-low, medium-high and high levels). The choice of tumour properties has already been described in Sambruni et al., 2023.

3.4 Microbiome reconstruction workflow

3.4.1 Microbial signal quantification

Pathseq

To extract the microbial signals from the NGS data of the samples we applied PathSeq (version 4.0.10.0, Walker et al., 2018) from the Genome Analysis Toolkit (McKenna et al., 2010) with the provided reference genomes prepared on 12/04/2017. Briefly, PathSeq aligns the sequences of the input files to the host reference genome, the human GRCh38, and discards the aligned reads. The remaining reads are then aligned to a set of microbial genomes to reconstruct the microbial composition of the analysed sample, see Figure 3.1. We ran PathSeq with default parameters on

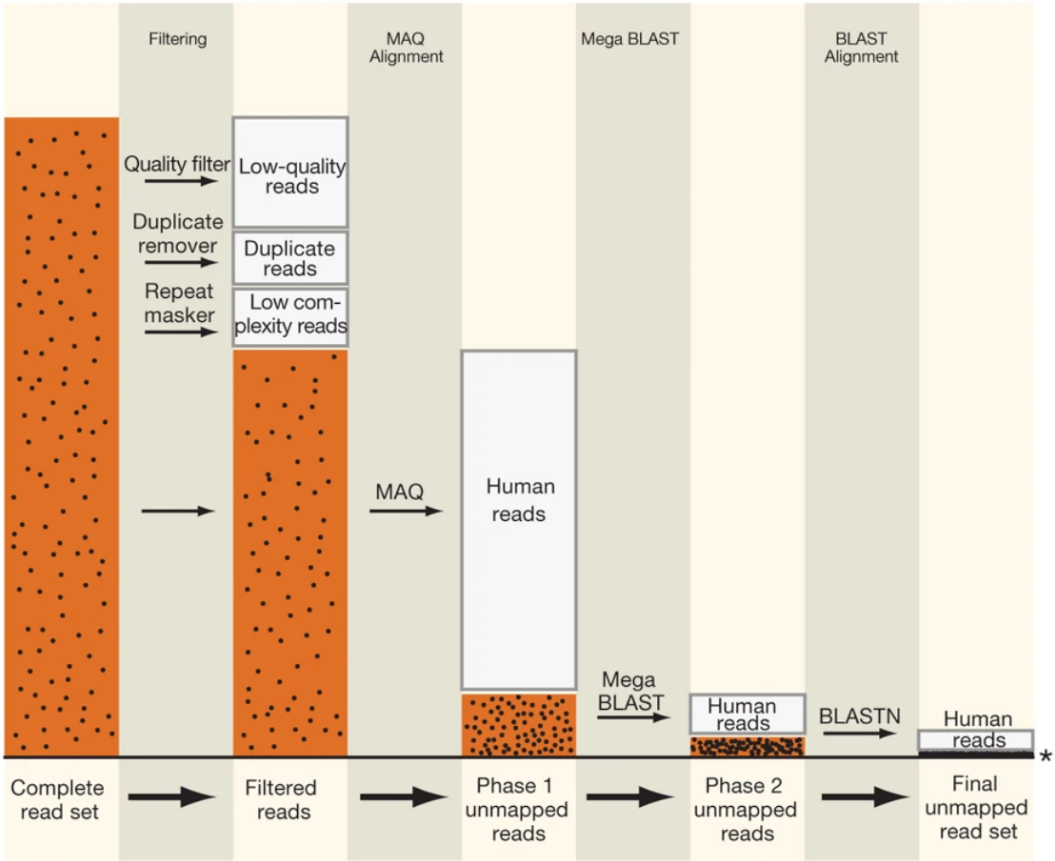


Figure 3.1: **The PathSeq subtractive phase workflow.** Orange bars represent the read set, the black dots inside them represent the microbial reads. The microbial reads become more and more concentrated at each step. From Kostic et al., 2011.

BAM files: these files were provided by the database (i.e. TCGA) or produced as described above, following TCGA protocol. Between the output files, PathSeq returns a table listing the detected taxa with different metrics to deal with taxa genome redundancy in different ways. Genomes of different species or strains can share homologous genomic regions and if a read maps to one of these common regions, it is impossible to assign it to only one taxon. PathSeq takes this into

account by reporting three metrics for each taxon: the score, the ambiguous and the unambiguous values. The unambiguous values include the reads assigned uniquely to one genome and represent the most stringent metrics. The ambiguous values enumerate all the reads mapping to the genome, no matter if the read maps to one or more genomes. Even if this approach is less stringent than the previous one, it inflates the total number of reads assigned to microbes (assigning the same read to more than one taxa). Lastly, the score represents a "weighted count" of the number of reads that map to the reference genome of the taxon considered. Considering taxon t, if a read maps only to the genome of the taxon t, it has a value of 1; if it does not map, it has a value of 0; and if the read maps to more than one genome (to a common region), it has a value of 1/(number of genomes to which the read maps). The bacterial score of the taxon t is the sum of the values from all the reads. These last two metrics, the score and the ambiguous values, can wrongly detect bacteria which are not present by taking into account the bacterial reads of non-detected species that share genomic regions with the real sample-derived ones. To avoid this issue, we decided to estimate the bacterial abundance with a modified version of the score values, the unambiguous scores: we only considered the bacterial scores of those species with at least one unambiguously mapping read. Finally, the bacterial unambiguous-score values were then intra-sample normalised so that all bacterial species scores sum up to 100 as a measure of bacterial relative abundance scaled to percentages.

We are aware that we are anyway missing or underestimating many microbes because the RNA-Seq hasn't been specifically developed to measure microbial signal but captures the majority of the human sequences available in the sample, given also the highly lower proportion of microbial cells than the human ones. Our approach suggests anyway an overview of the microbial composition of the known microbiota present in the sample and represents a great chance to increase our knowledge about human microbiota with big cohorts of samples.

This approach is described in Sambruni et al., 2023.

Kraken

Kraken is another commonly used tool for extracting microbial reads from NGS data (Wood and Salzberg, 2014). It employs a "classification tree" approach, as illustrated in Figure 3.2. The process involves breaking down input sequences into k-mers and mapping them to a set of reference genomes. Kraken then determines the lowest common ancestor for each k-mer: if a k-mer is assigned to multiple reference genomes, it is attributed to the taxonomic level that is closest to these reference genome taxa it maps to. This results in the construction of a classification tree,

where each node is assigned a weight based on the number of k-mers assigned to it. Following this, each path from the root to the leaf in the classification tree is assigned a score calculated by summing the weights of the nodes along the path. The path with the highest score determines the final classification: the sequence is assigned to the leaf associated with that path. In cases where multiple paths have the same score, the sequence is assigned to the lowest common ancestor of those paths. This explains why Kraken tends to provide classifications at higher taxonomic levels, such as genus, rather than assigning sequences to the lowest level, such as species or strains. We utilised

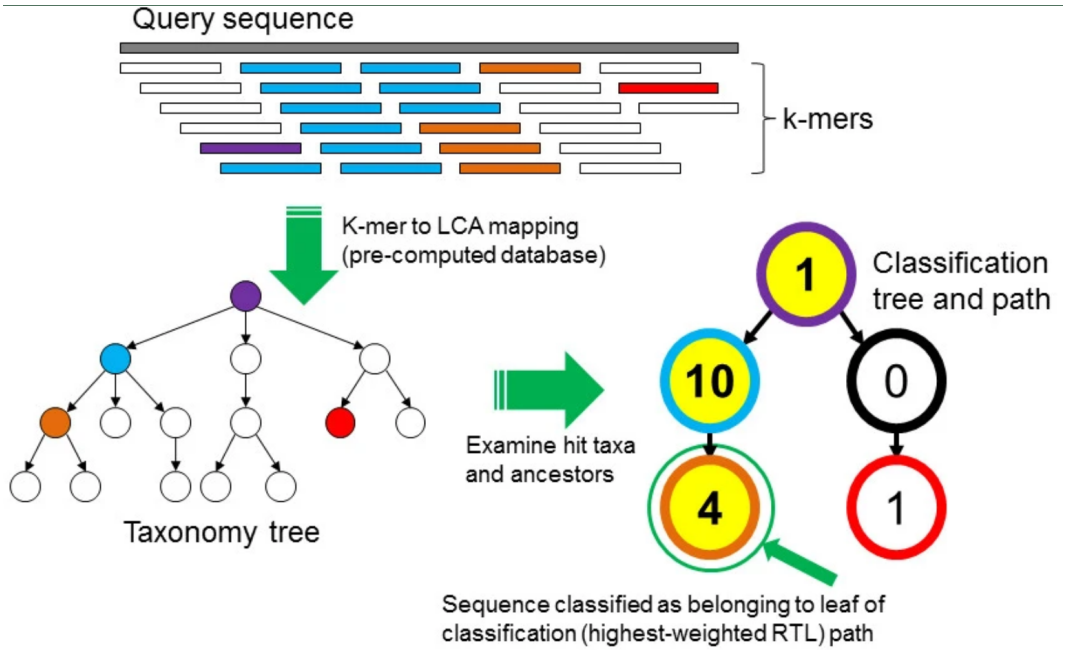


Figure 3.2: **The Kraken classification algorithm.** Summary of the classification algorithm used by Kraken to assign sequences to a taxon. From Wood and Salzberg, 2014.

Kraken2, described by Wood, Lu, and Langmead, 2019, using its Standard database containing RefSeq genomes of archaea, bacteria, viral, plasmid, human and UniVec_Cor. BAM files were converted to FASTQ files with Samtools (Danecek et al., 2021) and no human-read filtering steps were applied. Default parameters were used, with the addition of the "–report" and "–use-names" parameters. The "–report" parameter enabled the generation of a file providing information such as the percentage of assigned fragments, the number of fragments assigned to the root clade of each taxon and the number of fragments assigned to specific taxa at various taxonomic ranks. We estimated the bacterial signal present in each sample by considering the number of fragments directly assigned to each taxon and proceeded to select bacterial species or genera estimations. Next, we normalised the signal by dividing the number of sequences assigned to each taxon by the total number of bacterial sequences detected in each sample. This allowed us to calculate the percentage of sequences corresponding to each taxon in each sample. Additionally, by using the "–use-names" parameter, we included an extra column in the Kraken outputs containing the

names of the taxa.

3.4.2 Read annotation

PathSeq (Walker et al., 2018), the tool used to extract microbial reads from human NGS data, outputs the microbial abundance summary table described above and a BAM file. This BAM file contains the remaining reads after the first step of PathSeq, the human reads screening. We used this BAM file to further analyse the non-human reads.

Gene biotype identification

We used PathSeq output BAM file to map the non-human reads to the bacterial genome of a specific taxa of interest, annotate the reads and analyse the biotypes of the loci the reads map to. Firstly, for each bacteria of interest, we obtained the FASTA file of its genome and downloaded the annotation information in a GTF file from National Center for Biotechnology Information (NCBI). These files were used to create the index file needed by the Bowtie2 aligner (version 2.4.2, Langmead and Salzberg, 2012), using the bowtie2-build command. Then, after extracting the reads flagged as assigned to the bacteria of interest by PathSeq, we realigned them on the bacteria of interest annotated genome with Bowtie2 using default parameters. For each sample, we collected the information of genes or loci each read maps to and the biotype of that gene/locus. The gene biotypes available were further grouped into 7 categories:

- not aligned: alignment not unique, ambiguous, not aligned, too low aQual
- no classification: no feature, CDS, empty
- protein coding: protein coding
- pseudogene: pseudogene
- ncRNA: ncRNA
- rRNA: RNase P RNA, rRNA, SRP RNA
- tRNA: tmRNA, tRNA

The genomes used for this analysis are listed in Table 3.4.

A. muciniphila	GCF_000020225.1_ASM2022v1_genomic
F. prausnitzii	GCF_000162015.1_ASM16201v1_genomic
F. nucleatum	GCA_000007325.1_ASM732v1_genomic
B. fragilis	GCF_000009925.1_ASM992v1_genomic

Table 3.4: **Reference genomes for bacterial read annotation.** Reference genomes of *A. muciniphila*, *F. prausnitzii*, *F. nucleatum* and *B. fragilis* for their bacterial read annotation.

Toxin research

We used PathSeq output BAM file to identify the bacterial toxin signals or the expression of specific bacterial genes in the samples. We downloaded the sequences of the genes of interest from NCBI, see Table 3.5, and we created an index with Bowtie2 command bowtie2-build for each of them. We converted the BAM files obtained from PathSeq to FASTQ files with Samtools (Danecek et al., 2021) and then we aligned the bacterial reads with Bowtie2 to each reference sequence.

Gene	Locus	Name	Notes
Colibactin	AM229678	*E. coli* colibactin polyketide biosynthesis gene cluster	selected genes: clbA-R
Fragipain	NZ_CP069563	*B. fragilis* strain FDAARGOS_1225	region: 51311-52492
Fragilisin	AF081785	*B. fragilis* metalloprotease enterotoxin gene	
	AB026624	*B. fragilis* bft-3 (metalloprotease)	
	AB026625	*B. fragilis* bft-1 (metalloprotease)	
	AB026626	*B. fragilis* bft-2 (metalloprotease)	
Spermidine/ Putrescine	CR626927	*B. fragilis* NCTC 9343	selected genes: speA, speG, PotF, PotA, PotH nspC, PotC

Table 3.5: **Genes of interest.** Bacterial gene of interest loci and descriptions from NCBI database.

3.4.3 Microbiome composition

Given the overall high number of detected bacteria, ∼11,000, and the high number of missing species per patient, we decided to perform a dimensionality reduction approach to reduce the space and to capture the general tendency of the data (Murphy, 2012). Various techniques have been proposed for this purpose, with some commonly used across multiple fields, such as principal component analysis (PCA), while others are more specific to the microbiome field, such as principal coordinate analysis (PCoA) and non-metric multidimensional scaling (nMDS). The selection of the appropriate approach typically depends on the data characteristics and the specific question

being addressed.

PCA

PCA is a simple and established linear method that measures sample similarity. It represents the rotation of a data matrix and is highly versatile, allowing for the analysis of features measured with different units or scales. PCA can be described as a process of projecting samples onto a new set of orthogonal axes, referred to as principal components (PCs). The first PC captures the maximum variance in the data, while the second PC represents the maximum variation uncorrelated with the first PC and so on. So these projections resemble a linear combination of the original variables. PCA generates a significant number of PCs, typically equal to the number of samples or the number of species minus one, whichever is smaller. However, only a few PCs are typically meaningful for interpretation, as they explain a significant percentage of the overall variation, as observed in previous studies (Ramette, 2007).

PCoA

Another commonly used linear method is PCoA, also known as metric multidimensional scaling, commonly used in the field of microbiology. PCoA aims to represent the dissimilarity between objects in a low-dimensional space. Similar to PCA, it fits linear axes in a multivariate space to capture the maximum variance but in PCoA these axes are referred to as principal coordinate (PCo). However, the main difference arises because PCoA accepts dissimilarity matrices as input instead of raw data: PCoA aims to maximise the linear correlation between distances in a distance matrix and distances in the low-dimensional space obtained. There are anyway many similarities between PCA and PCoA, to the extent that when the distance metric is Euclidean, PCoA is equivalent to PCA. In order to prevent the generation of imaginary numbers during the analysis, PCoA only accepts only positive values and therefore requires pre-analysis data transformation. However, PCoA is prone to the horse-shoe or arch effect, which is partially attributed to its emphasis on maximising linear correlation. Additionally, PCoA is less affected by zeros and, unlike PCA where components are linear combinations of the original variables, PCoA coordinates are complex functions of the original variables based on the selected dissimilarity measure. Consequently, interpreting variable contributions in PCoA may be more challenging compared to PCA and the exact calculation of the percentage of total variance explained is not feasible (Ramette, 2007).

nMDS

nMDS is an alternative technique for dimensionality reduction that is particularly effective in identifying underlying gradients and representing relationships based on different types of distance measures. Unlike PCoA, nMDS does not utilise the raw dissimilarity values directly. Instead, it transforms them into ranks and incorporates these ranks in the calculation. Additionally, nMDS is an iterative algorithm that begins with the initial distribution of samples in the ordination space and iteratively rearranges the samples to search for the optimal final distribution. During these iterations, a parameter called "stress" is employed to assess how accurately the distances between points in the plot align with the true underlying distances. The objective is to minimise this stress value. Like the other methods discussed, nMDS positions samples in a low-dimensional ordination space in such a way that the Euclidean distances between these samples correspond to the dissimilarities represented by the original dissimilarity index. nMDS is particularly suitable for species abundance matrices that contain numerous zero values since it makes fewer assumptions in its analysis (Ramette, 2007).

Both PCoA and nMDS are dimensionality reduction techniques that can be employed on dissimilarity matrices, such as the Bray-Curtis distance matrix. The Bray-Curtis similarity index serves as a normalisation method used to compare the relative abundances of a community between two samples. It quantifies the dissimilarity or similarity between samples based on the relative proportions of shared and non-shared taxa. A value of 0 in the Bray-Curtis index indicates complete dissimilarity between samples, while a value of 1 indicates complete similarity, suggesting identical community compositions.

Our approach

We represented the whole reconstructed bacterial microbiome of samples applying the PCA on the intra-sample normalised unambiguous-score relative abundance, see section 3.4.1. To apply the PCA, we removed the bacterial species with zero relative abundances in all the samples analysed and then we selected the 1,000 species with the highest standard deviation relative abundances to focus on the most variable and more abundant bacteria. After this, we used the prcomp function from the stats R package to apply the PCA to the bacterial relative abundance. Since outliers can affect the results, we measured the Euclidean distances between samples and, if one sample was the most distant to 95% of the other samples (or more), it was considered an outlier and

removed. After removing an outlier, we reran the outlier identification method to identify and remove further outliers until no further ones could be detected. This approach has been previously described in Sambruni et al., 2023.

In a similar manner to the PCA approach, we employed PCoA and nMDS to represent the reconstructed bacterial microbiome of the samples as an additional control. To ensure compatibility with PCoA and nMDS, we eliminated bacteria with zero relative abundance in all the samples and selected the most variable bacteria. Subsequently, we converted the bacterial data into positive values by adding the smallest detectable value. This conversion was necessary because both PCoA and nMDS require positive values, while batch-corrected values can be negative. For the PCoA analysis, we utilised the vegdist function from the vegan package in R to measure the Bray-Curtis distance. Subsequently, we employed the cmdscale function in R to obtain the PCoA representation. For nMDS, we employed the metaMDS function from the vegan package in R, with the distance parameter set to "bray" to select the Bray-Curtis distance. Similar to PCA, we also performed iterative analyses for PCoA and nMDS to remove outliers, as previously described.

3.4.4 Batch effect

Our approach not only can underestimate bacterial detection, as explained above in section 3.4.1, but can be affected by biases that have been already highlighted in RNA-Seq analyses (Sprang, Andrade-Navarro, and Fontaine, 2022). Our detection of bacterial species from human RNA-Seq can also be affected by contamination at different stages of processing (from surgery to sequencing, Robinson et al., 2017). Moreover, the wrong identification of species can occur due to technical reasons, such as technical sequencing errors (usually discarded in human analyses) randomly mapping to bacterial genomes. To minimise this issue, we developed two steps to detect the strongest batch effect and in-silico correct for it.

Batch detection approach

After reconstructing the microbiome composition of tumour samples from human NGS data, we need to evaluate if our measures are affected by a technical bias. To detect the major technical batch affecting the bacterial estimation of each cancer type, we measured the Euclidean distances of the bacterial composition of samples in the first six PC of the PCA. The first six PC collectively explain more than 10% of the variability of each cancer type. We then compared the distributions of these distances of samples belonging to the same level of the technical property to the distances of samples belonging to different levels of that property applying the Wilcoxon test. For example, we compared the distribution of the distances between the samples belonging to the same 96-well

plate identifier (plate ID) to the distances between samples from different plate IDs. The technical property showing the lowest p value was considered the major batch effect in the analysed cancer type. This approach is described in Sambruni et al., 2023.

For the majority of the TCGA cancer types analysed, we detected the plate IDs as the strongest batch effect. To assure a minimum number of samples per plate ID, we pooled the plates with a low number of samples (less than 10% of the total number of samples (frac) and if frac $>$ 5, frac was set to 5). In a similar way, we detected the sequencing step as major batch effect in the IEO cohort.

In a similar fashion to the PCA method, we evaluated the significant batch effect found in alternative dimensionality reduction approaches by computing the Euclidean distances of the bacterial composition for samples in the first six dimensions of these methods, specifically PCoA and nMDS. Subsequently, we compared the resulting distributions using the same methodology mentioned earlier.

Batch correction

After identifying the strongest batch effect, we want to correct for it to reduce as much as possible its effect on the reconstructed bacterial composition. Firstly, the reconstructed bacterial microbiome relative abundances were scaled and log-transformed. After that, we applied the ComBat function from the sva package in R (W. E. Johnson, C. Li, and Rabinovic, 2007), controlling for the known batch covariate. This function is one of the most popular approaches to deal with batch effects and has been widely applied to many methods (Leek, W. E. Johnson, et al., 2012). ComBat works with the empirical Bayesian framework to adjust a known batch effect and it represents a robust approach towards outliers with small sample sizes, avoiding overcorrection.

This procedure is explained in Sambruni et al., 2023.

3.4.5 Microbiome composition association with tumour properties

One of the main goals of this study is to find links between the bacteria composition of tumour samples with their properties. So we used the Wilcoxon or Kruskal-Wallis test to compare the distributions of the PC coordinates of the batch corrected data PCA on the reconstructed microbiome of the samples, with the property subgroups we were analysing. We also tested the correlation between PC coordinates and the tumour properties values with the Spearman correlation test when feasible. We considered the first six PCs since they can explain more than 10%

of the total variability of the reconstructed microbiome for all the cancer types tested. The same procedure was applied with the other dimensionality reduction approaches.
This approach is described in Sambruni et al., 2023.

3.4.6 Bacteria filter criteria

As explained above, the bacterial species detected include the real sample-derived ones but also other signals from different sources of contamination. After applying the corrections explained above, we defined two approaches to filter the species:

1. High confidence set of species: we defined three filters to remove the batch-affected, low-present bacteria and select the cancer type-specific ones. To remove the bacteria whose distribution is affected by the dominant batch effect of the cancer type (see section 3.4.4), we applied the Wilcoxon test to their relative abundances and removed the bacteria with false discovery rate (FDR) multiple-test corrected q value < 0.1. To filter out the low prevalent bacteria, we selected those bacteria detected in at least 10% of the samples of the cancer type of interest. To select cancer-type specific bacteria, we finally selected those bacteria showing a higher mean in the cancer type analysed than in the other types.

2. Colon specific set of species: we screened the bacteria by applying the presence and cancer type-specificity filters described above. After that, in order to test for differentially abundant species between different levels of properties, we applied a non-parametric Mann–Whitney test from the independence_test function as implemented in the R package coin (Hothorn et al., 2008). This approach tests the independence of two groups of multivariate variables and refers to the general framework for conditional inference (Strasser and Weber, 1999). In our analyses, we applied the independence_test blocking for the property we considered the dominant technical batch. Finally, we considered bacteria statistically significantly associated with a property, if their multiple-testing corrected q value (FDR method) was below 0.1.

These methods are reported in Sambruni et al., 2023.

3.5 Immune cell quantification

Recently many authors demonstrated that it is possible to infer the immune cell infiltration abundance from bulk human RNA-Seq data (Finotello and Trajanoski, 2018). Firstly, we downloaded the fragments per kilobase million (FPKM) values from the GDC. Then we calculated the transcript per million (TPM) from FPKM tables by dividing each FPKM value with the sum of the FPKM values of that sample and then multiplied by 1 million. Since the GDC provides the expression of Ensembl IDs, we converted them to HUGO gene names using the annotation version v22. After that, we inferred the immune cell infiltration by running CIBERSORTx (Newman et al., 2019) from their web page on TPM gene expression quantification. We used their default signature matrix LM22, activated the B-mode batch correction (as suggested by the authors) and run with 1000 permutations, the highest number available. We run CIBERSORTx in both absolute and relative mode for COAD, but we mainly rely on the relative values. We considered only the significant ($p < 0.05$) immune estimates.

This approach is described in Sambruni et al., 2023.

3.6 Pathway analysis

We used the human RNA-Seq data to disentangle the microbial functions activated in the tumour samples. The microbial signals extracted from the human RNA-Seq data are too low to reproduce the expression of a complete pathway so, to approximate the active microbial pathways expressed in samples with the same tumour property, we pooled the PathSeq output BAM files of the primary tumour samples from the same sublevel (e.g. from the left and the right colon sections). We then analysed these pooled reads with HUMAnN 3.0 (Beghini et al., 2021), a tool to profile microbial pathways from NGS data. As suggested by the authors, we normalised the pathway abundances to copies per million (CPM). After this, we filtered out the low abundance pathways and, to select the differentially active pathways, we considered the pathways showing at least one-third higher abundance than the other sublevel.

To further confirm our results, we applied bootstrapping to estimate the significance of our observations. Firstly, we randomly picked and pooled one-third of the samples from each sublevel in 50 independent permutations. Then we applied HUMAnN 3.0 as described above to obtain pathway distributions of the sublevels. The distributions of the previously identified pathways were compared with the Wilcoxon test and a FDR multiple-test corrected q value < 0.2.

This method is reported in Sambruni et al., 2023.

3.7 Survival analysis

TCGA provides follow-up information about the patients involved in the study. We downloaded the disease-free survival (DFS) and overall survival (OS) data from cBioPortal for Cancer Genomics website (Cerami et al., 2012; J. Gao et al., 2013) to measure the association between these survival details and the bacterial reconstructed composition of the tumour analysed. To do this, we applied Cox proportional-hazard models with the coxph function of the survival R package: we ran univariate Cox models on the top six PC coordinates separately and selected the significant ones. To take into account values with different scales, we scaled continuous properties to be in the range 0-1. The tumour properties associated with the PCs could drive the association of the PCs with the survival of the patients and bias the results. To exclude this possibility, we tested a multivariate model with the selected PC coordinates together with their associated properties and checked whether PCs remained significant. The significant associations were further validated by running the Kaplan-Meier analysis on the original PC coordinates: patients were stratified into "high" and "low" groups by maximally selected rank statistics with surv_cutpoint function from survminer R package.

To detect which bacterium is associated with the relapse probability, we applied univariate Cox analysis to the batch corrected values of the first 100 bacteria with the highest loadings of the PCs associated with relapse probability. We then multiple-test correct the p value of the Wald test, selecting for $q < 0.2$ species.

This approach is described in Sambruni et al., 2023.

Results

4.1 Microbiome reconstruction

After collecting the data from TCGA and the IEO cohort, the first step in our project was to reconstruct the microbial composition of the samples. By using human NGS data, we were able to obtain a microbial signal that provided insight into the types of microorganisms present in each sample. We developed and applied a workflow to extract and evaluate microbial signals, as explained in section 3.4 of the Methods chapter. This workflow involved utilising a tool called PathSeq (Walker et al., 2018) to identify microbial reads and associate them with specific microbial species. Then we took several technical considerations into account, such as genome redundancy, to reduce noise and ensure accuracy in our analysis.

4.1.1 Microbial signal detection

PathSeq detected microbial signals in all the tissues that were analysed, for a total of 71,997,702 microbial reads in TCGA samples, including both tumour and solid normal tissues. These reads accounted for 0.025% of the total reads in the samples. Among them, the majority, 67,501,963 reads, were identified in tumour samples, while only 4,495,739 were found in solid normal tissue samples. The proportion of microbial reads relative to the total reads in the samples was similar for both tumour and solid normal tissues, at 0.026% and 0.020%, respectively. IEO samples showed a total of 7,933,012 microbial reads (0.15% of the total), a higher proportion compared to TCGA.

Even if we detected a higher percentage of archeal reads in IEO samples than in TCGA, in both cohorts, the bacteria superkingdom accounted for the majority of microbial reads in both tumour and solid normal tissue samples (see 4.1). As a result of the high proportion of bacterial reads, further analyses primarily focused on this superkingdom.

Some of the results presented here have already been published in Sambruni et al., 2023.

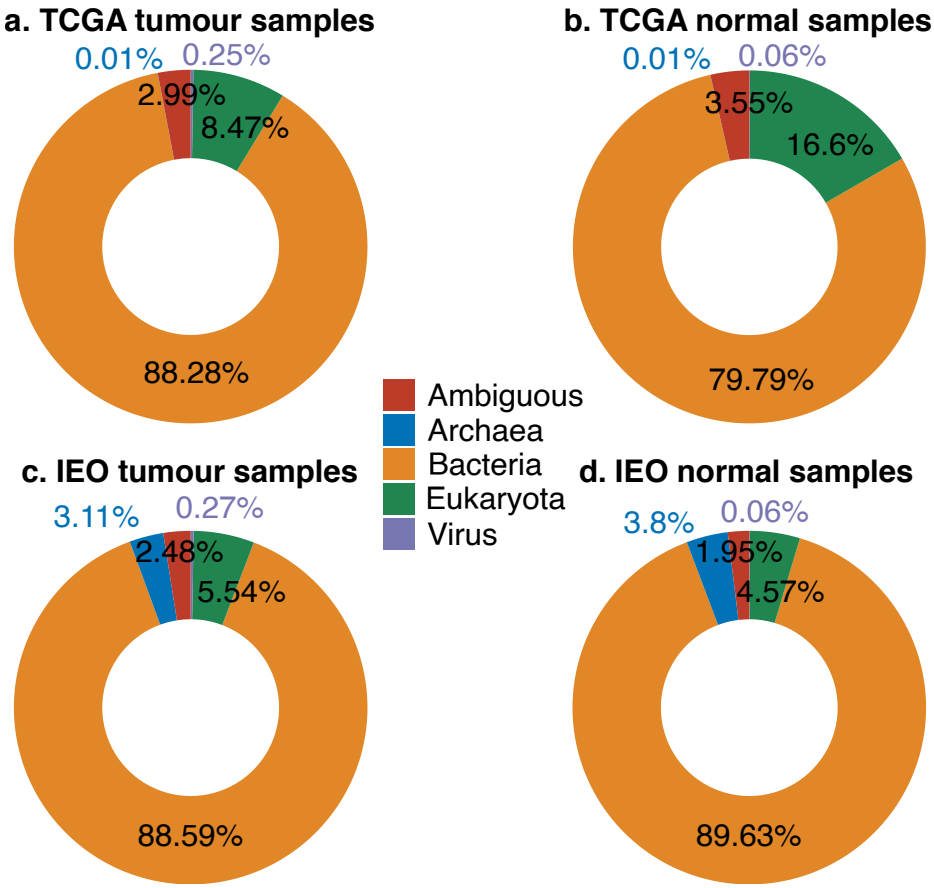


Figure 4.1: **Percentage of microbial reads in TCGA and IEO samples.** Percentage of microbial reads (after human read removal) per superkingdom in (a) tumour and (b) solid normal tissue samples of TCGA and in (c) tumour and (d) normal samples (at 2 and 10 cm from the tumour) of the IEO cohort.

4.1.2 Bacterial signal detection

In total, we identified 63,179,002 bacterial reads in TCGA samples across all analysed cancer types. BRCA exhibited the highest number of bacterial reads, while skin cutaneous melanoma (SKCM) had the lowest, as shown in Figure 4.2. Notably, GBM showed a substantial number of bacterial reads, even though it is considered a sterile organ, colonised by bacteria only after the blood-brain barrier disruption (Le Guennec et al., 2020) or in particular brain diseases, such as Alzheimer's one (Parra-Torres et al., 2023).

Due to the higher number of samples, most of the collected reads were from tumour samples

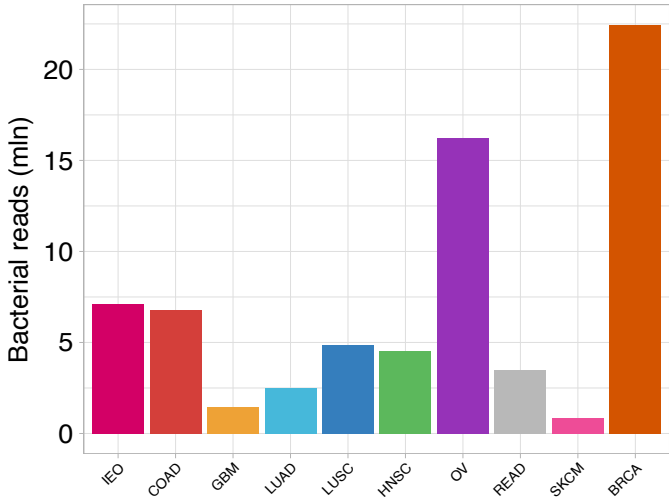


Figure 4.2: **Bacterial reads detected in TCGA cancer type.** Number of bacterial reads found in TCGA and IEO by cancer type. Adapted from Sambruni et al., 2023

(59,592,060 reads in total). However, solid normal tissue and primary tumour samples produced a similar number of reads in each cancer type, as illustrated in Figure 4.3a. Some exceptions include GBM, LUAD and rectum adenocarcinoma (READ), where the number of bacterial reads per sample was higher in tumour samples. In contrast, BRCA exhibited the opposite trend. The limited number of normal samples in GBM (only 5 samples, as shown in Table 3.1) precludes any definitive conclusions. Nevertheless, the observed difference in the number of reads between the different LUAD sample types was unexpected, given previous studies reporting a similar bacterial signal load between tumour and normal adjacent tissue (Nejman et al., 2020).

These reads corresponded to a total of 11,961 bacterial species detected, with OV samples exhibiting the highest number of bacterial species detected per sample, as depicted in Figure 4.3b. Some of the results presented here have already been published in Sambruni et al., 2023.

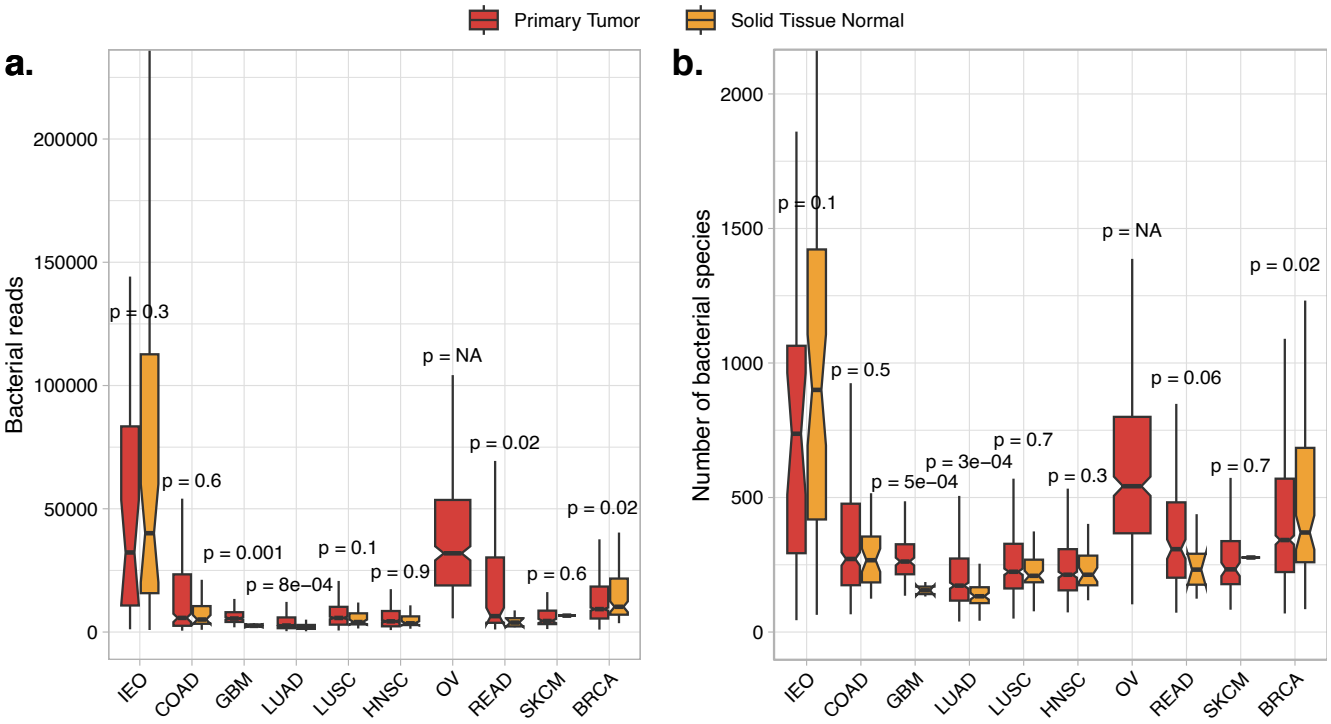


Figure 4.3: **Bacterial reads and species in TCGA samples.** Number of (a) bacterial reads and (b) bacteria species per sample in TCGA by sample type.

4.1.3 Signals from other superkindoms

As shown in Figure 4.1a-b, bacterial reads accounted for a larger proportion than eukaryotic, archaeal and viral reads in TCGA samples. In particular, we detected 168,059 viral, 5,714,956 eukaryotic and 9,469 archaeal reads present in the tumour samples of TCGA cancer types. OV and head and neck squamous cell carcinoma (HNSC) showed the highest number of viral reads (79,497 and 62,549, respectively) and tumour samples had more viral reads than normal samples only in COAD, as represented in Figure 4.4a.

Eukaryotic reads showed different behaviour, with BRCA, HNSC and LUAD having the highest number of reads (1,7639,599, 1,456,579 and 1,196,285 reads, respectively). Interestingly, BRCA, LUSC and HNSC also showed a higher number of reads in normal samples, as shown in Figure 4.4b.

Finally, in OV and LUAD we detected most archaeal reads (2,003 and 1,183 reads, respectively) but no cancer types showed statistically significant differences between tumour and normal samples, as presented in Figure 4.4c.

The smaller number of microbial reads assigned to eukaryota, viruses and archaea resulted in detecting fewer species, as shown in Table 4.1.

Due to the low number of reads and species detected for non-bacterial superkingdoms, we were unable to apply any correction to remove technical biases or capture global trends. Nevertheless,

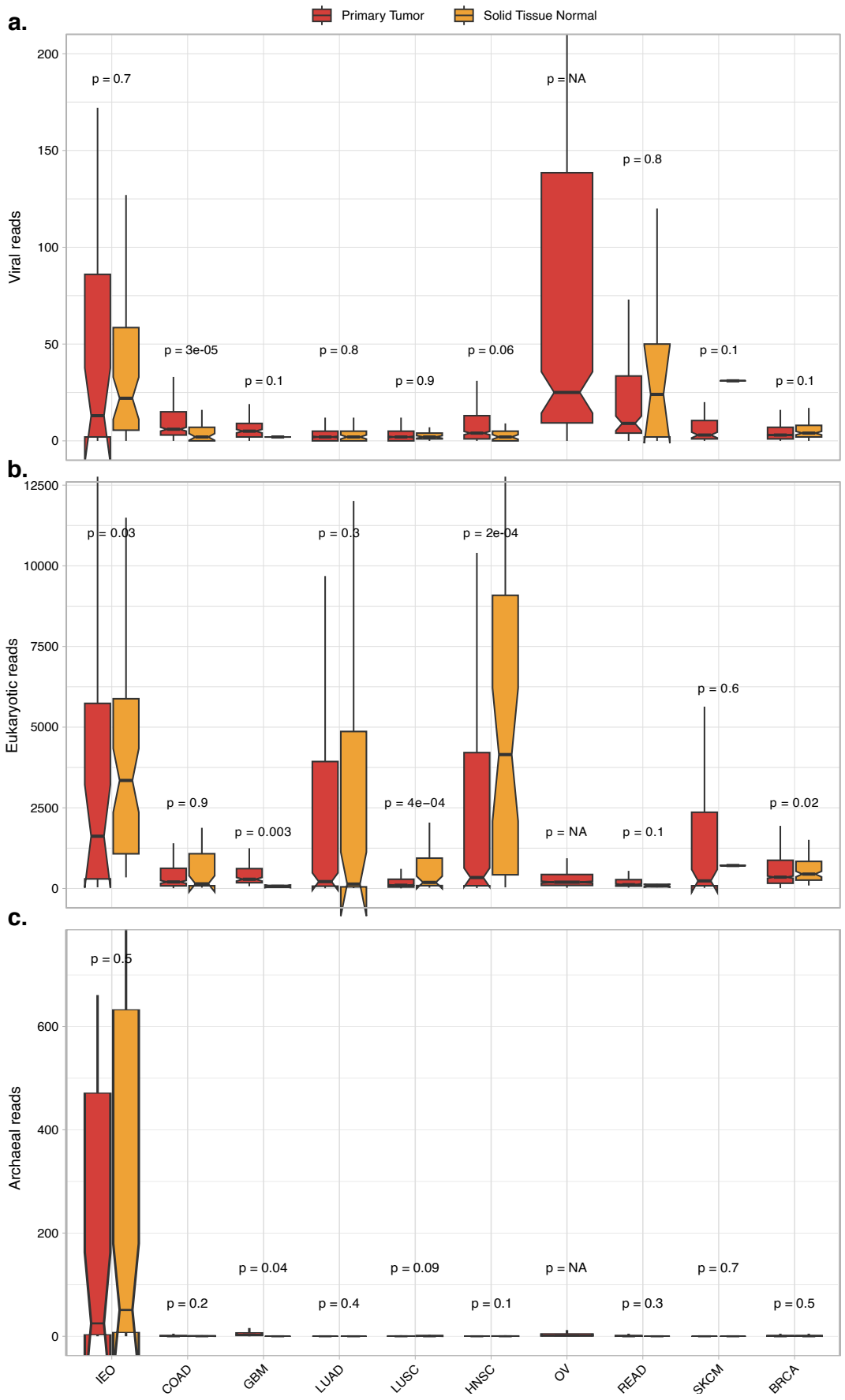


Figure 4.4: **Superkingdoms reads in TCGA samples.** Number of (a) viral, (b) eukaryotic and (c) archaeal reads per sample in TCGA by sample type.

when examining the detected taxa, we found some interesting species. HPV, one of the most common cancer-causing pathogens, affects cervical and oral cavity cancer (Martel et al., 2020). TCGA provides details of HPV16, HPV18, HPV33 and HPV35 calls for HNSC samples. These detections were performed using Multiplex PCR and Sequenom-based Mass Spectrometry, which rely on the identification of DNA sequence variations within the viral E6 regions. Hence, we tested if we could detect HPV viral reads in HPV-positive samples from HNSC. We analysed Alphapapillomavirus 9 (the viral species of the serotypes HPV16, 33 and 35) and Alphapapillomavirus 7 (the species of the serotype HPV18). Since TCGA reported only 3 HPV18 positive samples, we focused on the Alphapapillomavirus 9. Our method detected Alphapapillomavirus 9 in 78 HNSC samples, 57 of which were detected as HPV-positive by TCGA (p = 2.94e-45, Chi-squared test). Among the most common eukaryotic signals found in COAD samples, we detected *Blastocystis hominis* (present in 38 samples) known to be associated with colorectal cancer and other human diseases (Kumarasamy et al., 2022). *Blastocystis sp. Subtype 4*, another blastocystis species previously associated with acute diarrhea (Kosik-Bogacka et al., 2021), was detected in 36 samples from TCGA colon cancer patients. Finally, *Toxoplasma gondii* was detected in both LUAD and LUSC (in 53 samples in both cancer types) and it is known to infect patients with lung disease (Y.-X. Li et al., 2020) and lung cancer (Bajnok et al., 2019). *T. gondii* was also found in OV (31 samples), together with the closer related parasite *Hammondia hammondi* (38 samples) (Sokol-Borrelli, Coombs, and Boyle, 2020).

Cancer type	Virus	Eukaryota	Archaea
COAD	304	182	68
GBM	126	97	34
LUAD	180	185	38
LUSC	170	173	50
HNSC	197	181	54
OV	331	148	95
READ	166	128	39
SKCM	106	103	19
BRCA	327	233	117

Table 4.1: **Number of species detected in TCGA cancer type.** Number of species detected in TCGA cancer type in tumour and normal samples per each superkingodom.

4.1.4 Bacterial read annotation

The PathSeq tool (Walker et al., 2018) selected the non-human reads from COAD samples and we then aligned and annotated them to the genomes of *A. muciniphila*, *F. prausnitzii*, *F. nucleatum* and *B. fragilis*, which are known to be part of the colon microbiota (Verhoog et al., 2019) or to be associated with colon cancer (Clay, Fonseca-Pereira, and Garrett, 2022). Most of the reads were successfully classified into a specific gene biotype, with the majority of them being associated with ribosomal RNA (rRNA) genes. Figure 4.5a illustrates that only a minor fraction of the reads were mapped to other genes. Although *F. nucleatum* had a poorly annotated genome, most of the non-assigned reads also mapped to rRNA loci when aligned to the latest genome version.

The read load reflected the healthy or diseased state, see Figure 4.5b, with *F. prausnitzii* being more prevalent in the normal tissue and *F. nucleatum* being more common in tumour samples (relative to their matching normal counterparts). The presence of *A. muciniphila* and *B. fragilis* was similar in both normal and tumour samples, even if they are known to be normally present in the colon (only the enterotoxic strains of *B. fragilis* are associated with cancer (Clay, Fonseca-Pereira, and Garrett, 2022). However, additional analyses were not possible due to the low number of reads assigned to each bacteria per sample and the small proportion of reads that aligned with genes.

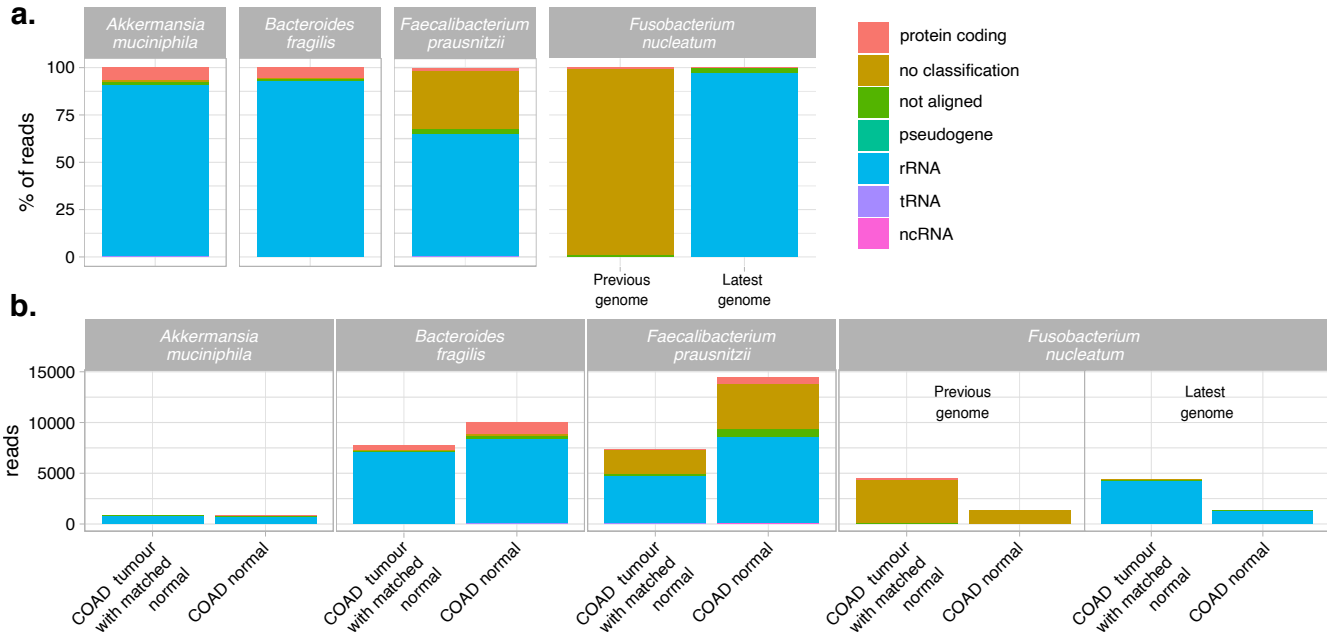


Figure 4.5: **Species reads in COAD samples.** (a) Percentage of reads coloured by biotype mapping to *A. muciniphila*, *B. fragilis*, *F. prausnitzii* and *F. nucleatum* in tumour COAD samples. (b) Number of reads assigned to the same species coloured by biotype in normal and tumour with matching normal COAD samples. Read annotation of *F. nucleatum* is shown for both the genome in PathSeq database (previous genome) and the latest, better annotated one.

4.2 Bacterial signal validation

In recent years, the retrieval of bacterial signals from human NGS data has emerged as a prime example of how valuable information can be extracted from existing data with innovative techniques (Q. Wang et al., 2023). Previous studies have confirmed these signals by quantifying the abundance of specific bacteria using qPCR, among other methods (Bullman et al., 2017). In this study, we propose two different approaches for validation. Specifically, we compared the reconstructed bacterial composition from RNA-Seq with those from 16S and for *F. prausnitzii*, *A. muciniphila* and *F. nucleatum* with those from FISH.

4.2.1 RNA-Seq vs 16S

Our method was compared to one of the most common approaches for detecting bacteria in NGS, which is the 16S sequencing. This approach specifically sequences the 16S to classify bacterial genera (Lane et al., 1985). Due to the lower specificity of the 16S approach, we compared microbial abundances on genes level in a subset of 61 samples from the IEO cohort for which both RNA-Seq and 16S data were available. We applied different prevalence cutoffs, as shown in Figure 4.6. We

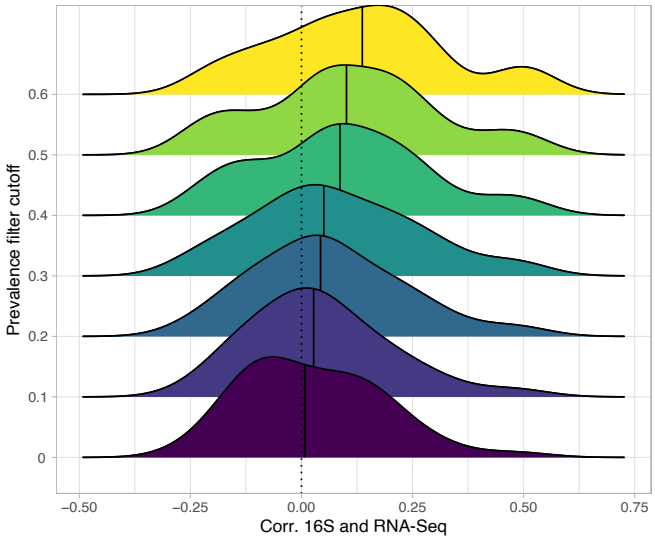


Figure 4.6: **RNA-Seq vs. 16S ridge plot correlation.** Ridge plot of the Spearman's coefficients of RNA-Seq and 16S bacteria genera quantification correlations. Different filter cutoffs for bacterial presence are shown. Black vertical lines represent the median of the distribution. Adapted from Sambruni et al., 2023

observed a reasonable correlation between the estimates of the bacterial genera detected by the two approaches and the distributions of the Spearman's coefficient values were significantly above zero ($p < 0.05$; One-sample Wilcoxon test). The median of the Spearman's coefficients increased with the increase of the cutoff, suggesting that highly prevalent bacteria are more reliably quantified

by our method.

These results were previously shown in Sambruni et al., 2023.

4.2.2 RNA-Seq vs. FISH

We evaluated the quantification of three bacteria, namely *F. prausnitzii*, *A. muciniphila* and *F. nucleatum* in a subset of ten samples from the IEO cohort, by comparing their estimation from their RNA-Seq reconstructed microbiome with their quantification using FISH probes specific to each bacterium. We tested *F. prausnitzii* and *A. muciniphila* as controls since they are associated

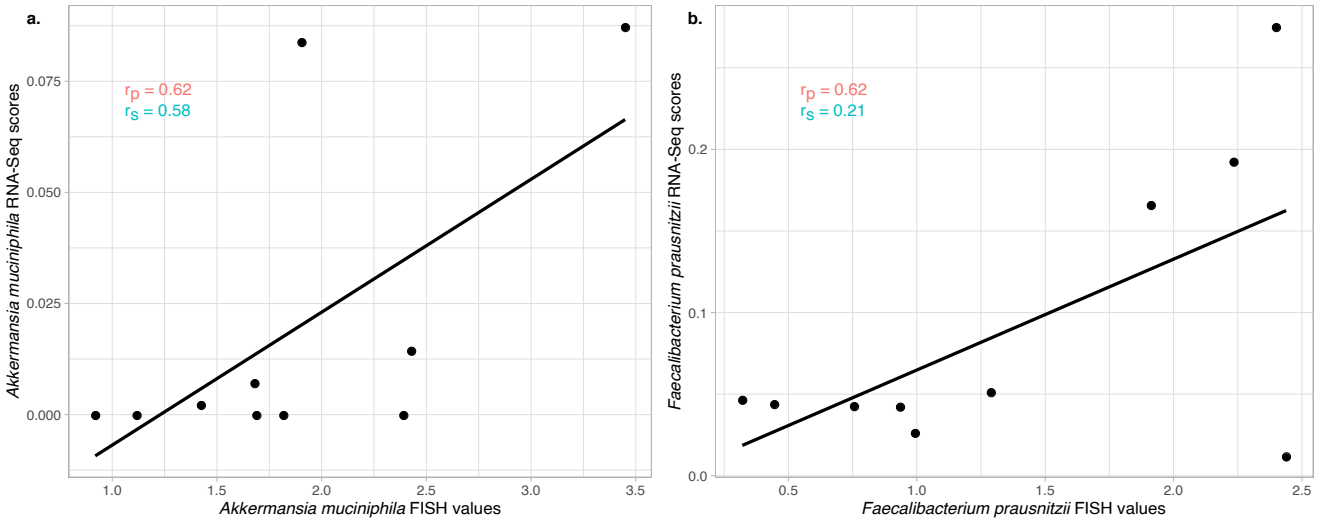


Figure 4.7: ***A. muciniphila* and *F. prausnitzii* RNA-Seq vs. FISH correlation.** Correlation of the RNA-Seq bacterial relative abundances and FISH quantification of (a) *A. muciniphila* and (b) *F. prausnitzii*. Pearson (r_p) and Spearman (r_s) coefficients are indicated.

with healthy colon, while *F. nucleatum* is linked to colon cancer (Clay, Fonseca-Pereira, and Garrett, 2022). We expected to find signals of *A. muciniphila* and *F. prausnitzii* in the distal and peripheral samples and potentially in tumour samples. We observe a good agreement in the quantification of both methods for *A. muciniphila* and *F. prausnitzii*: we observe a favourable trend, particularly for *F. prausnitzii*, as depicted in figures 4.7. In contrast, *F. nucleatum* displays a negative trend, primarily attributed to the numerous samples with zero detection of *F. nucleatum* signals, as illustrated in figure 4.8. This could be due to an inefficient FISH probe or the RNA-Seq microbiome reconstruction approach's inability to detect the bacteria. Experimental reasons such as an inability to detect *F. nucleatum* signals or technical issues such as genome similarity with other bacteria could also contribute to this outcome. Since there were numerous instances in which no signal of *F. nucleatum* was detected in the samples, it is probable that the microbiome reconstruction from human RNA-Seq data approach is not accurately reflecting the presence of *F. nucleatum* in the samples. Anyway, experimental reasons and technical issues are not mutually

exclusive and both of them could have affected our results.

These findings suggest that our microbiome reconstruction workflow can detect some bacteria better than others and the workflow can be improved to enhance bacterial detection accuracy.

Part of these results has been presented in Sambruni et al., 2023.

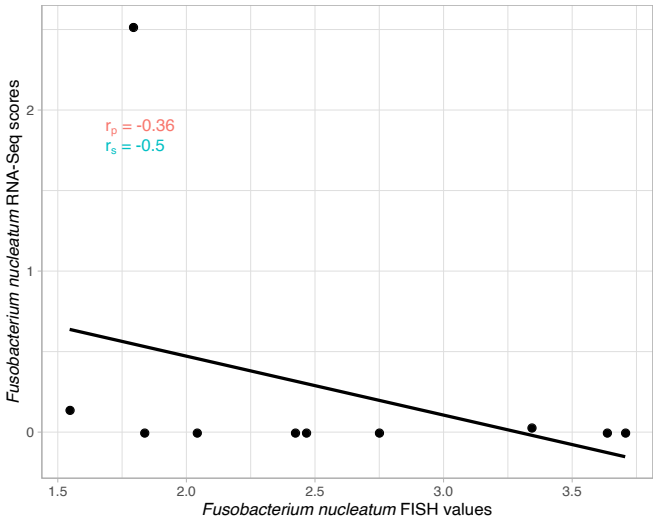


Figure 4.8: ***F. nucleatum* RNA-Seq vs. FISH correlation.** Correlation of the RNA-Seq bacterial relative abundances and FISH quantification of *F. nucleatum*. Pearson (r_p) and Spearman (r_s) coefficients are indicated.

4.2.3 NGS method comparison

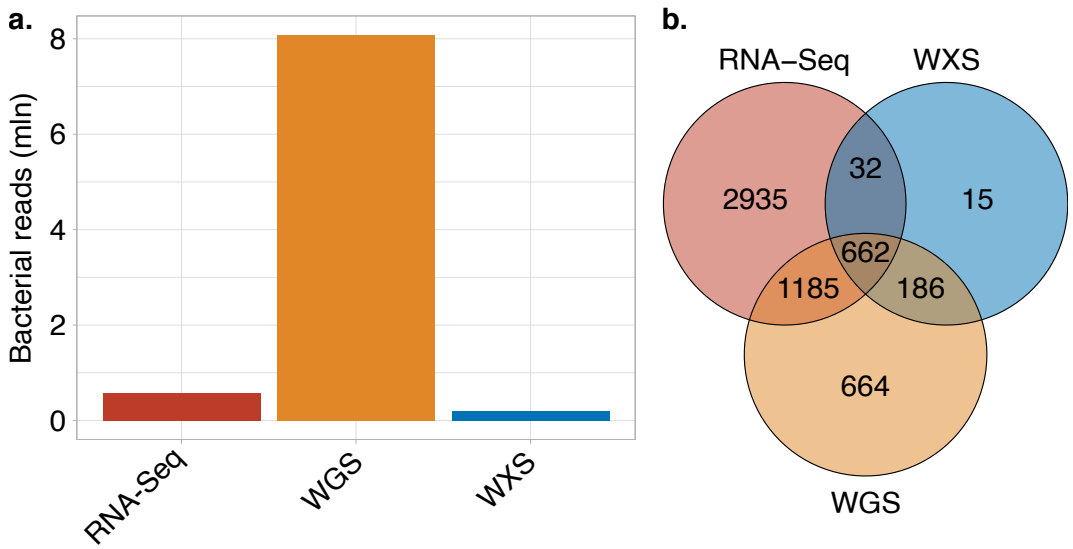


Figure 4.9: **NGS methods comparison in COAD samples.** (a) Number of bacterial reads detected in RNA-Seq, WGS and WXS methods in a subset of 33 COAD samples analysed by the three approaches. (b) Number of bacteria species detected in the same subset of COAD samples.

The TCGA dataset includes various types of NGS methods that can undergo our microbiome reconstruction workflow. We compared the bacterial microbiome reconstruction from different NGS

approaches applied on the same sample: we specifically examined RNA-Seq, WGS and WXS. We selected 33 COAD patients whose tumour sample was analysed using all three approaches. The results of our analysis, shown in Figure 4.9a, revealed that WXS detected the lowest number of reads and species, while WGS samples detected the highest number of bacterial reads compared to the other approaches. Figure 4.10 demonstrates that the estimation of bacteria detected by

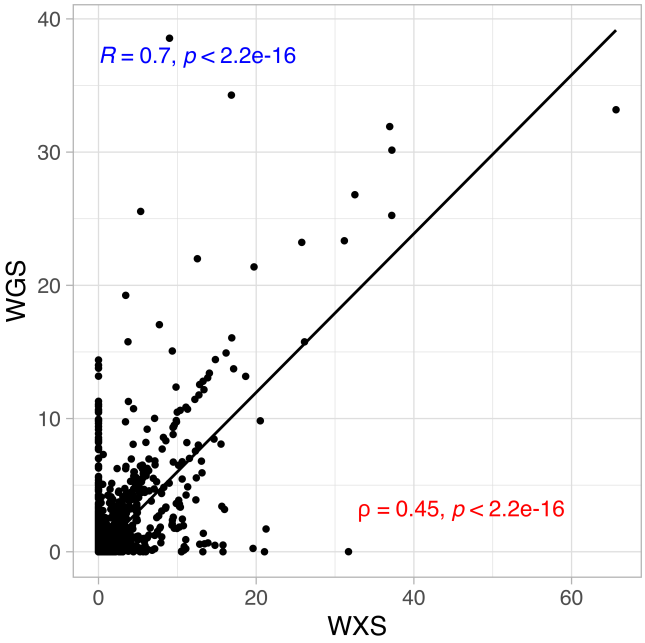


Figure 4.10: **Correlation of COAD WGS and WXS bacteria estimation.** Correlation of the WGS and WXS bacterial estimations of common species in 33 common samples.

both WXS and WGS approaches are highly comparable (Spearman correlation $\rho = 0.7$), with very few bacteria detected by WXS and not by WGS (bottom right area of Figure 4.10) and a few bacteria detected by WGS but not by WXS (top left area). As expected, given that WXS protocol uses targeted primers to amplify human genes, WGS seems to detect a higher bacterial signal and more species than WXS, even though the bacteria detected by both approaches are similar.

The comparison between WGS and RNA-Seq is less straightforward, as WGS detected the highest number of bacterial reads but RNA-Seq reads mapped to the highest number of species, see Figure 4.9b. Interestingly, the majority (68%) of the species detected by WGS were also detected by RNA-Seq. The comparison of the species detected by both approaches reveals a similar tendency, although not as strong as that seen between the other two approaches, as shown in Figure 4.11. Finally, we annotated and compared the bacterial reads from the three approaches to the genomes of four species due to their association with the normal colon microbiome or colon cancer, namely *A. muciniphila*, *F. prausnitzii*, *B. fragilis* and *F. nucleatum*. The majority of both WGS and WXS reads mapped to genes, while, as expected, the majority of RNA-Seq reads were assigned

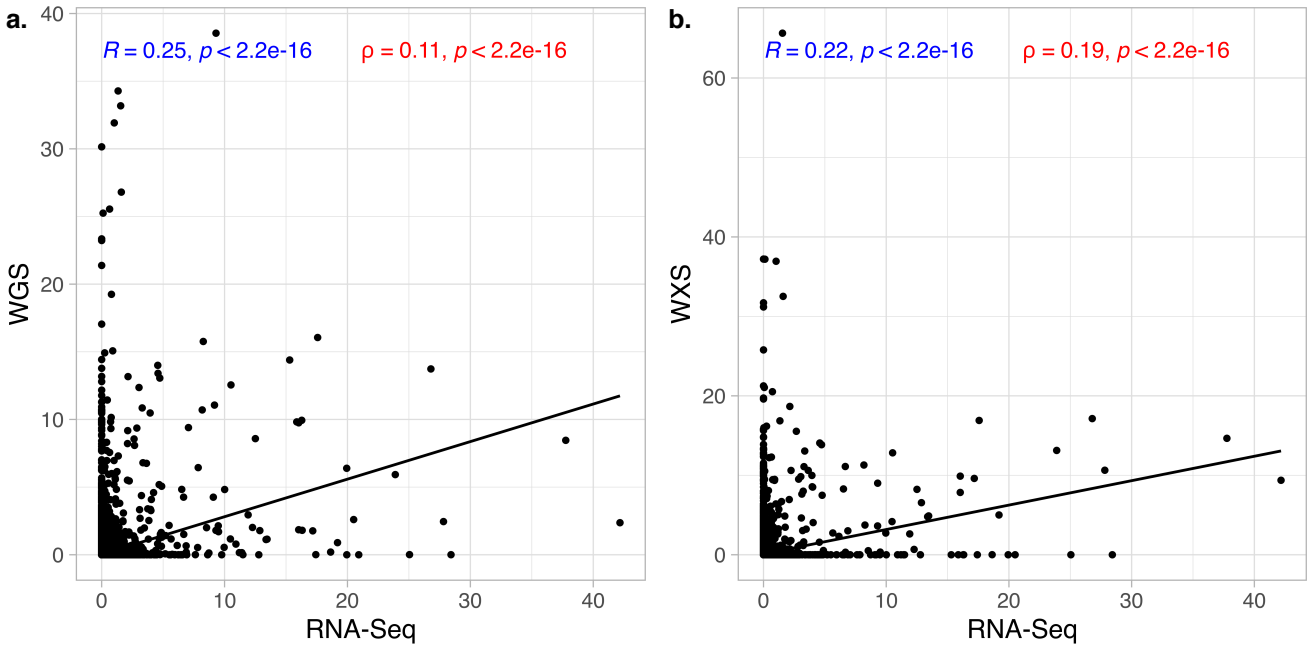


Figure 4.11: **Correlation of COAD RNA-Seq with other methods.** Correlation of the RNA-Seq bacterial estimation with (a) WGS and (b) WXS bacterial estimations of common species in 33 common samples.

to rRNA, as shown in Figure 4.12. In summary, while there were differences in the number of bacterial reads and species detected among the three approaches tested, a significant portion of the species were identified by multiple methods. Additionally, the common species exhibited a consistent level of concordance across the different approaches. Notably, the WGS and WXS methods demonstrated the highest agreement, which is expected given their shared detection of DNA. Despite the inherent differences between RNA-Seq and the two DNA sequencing techniques, the agreement with RNA-Seq was reasonable.

For this project, RNA-Seq was selected due to its widespread availability in TCGA and the ability to validate it with the IEO cohort.

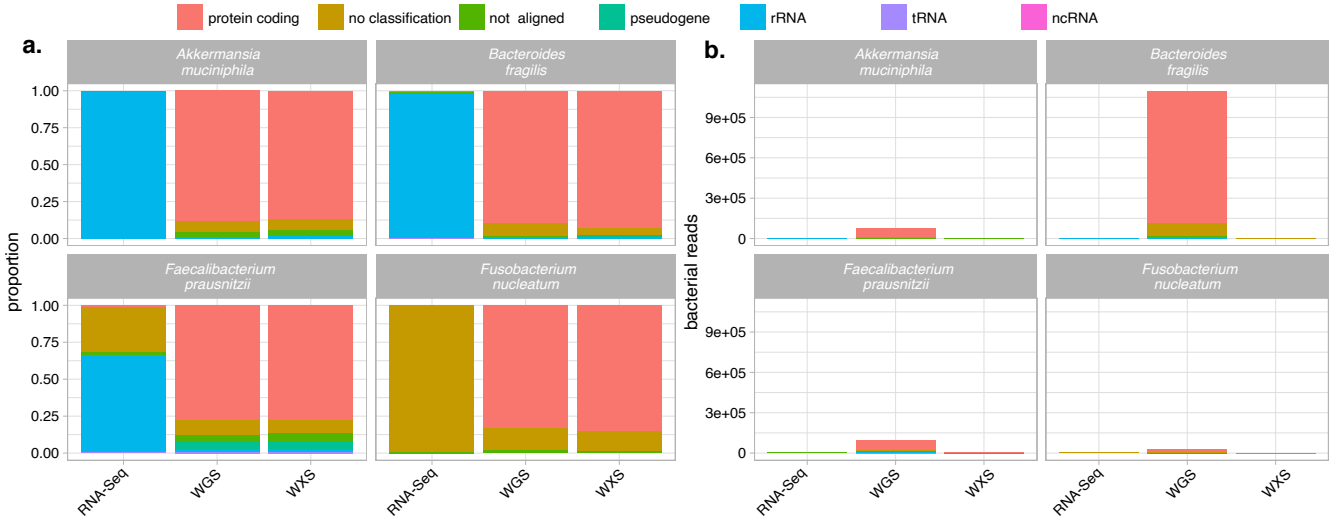


Figure 4.12: **Species read annotation in a COAD subset of samples in RNA-Seq, WGS and WXS.** (a) Percentage of reads mapping to *A. muciniphila*, *B. fragilis*, *F. prausnitzii* and *F. nucleatum* and (b) number of reads assigned to the same species coloured by biotype in the subset of 33 common samples from COAD project analysed by RNA-Seq, WGS and WXS.

4.3 Identification of toxin reads

Bacteria can potentially be oncogenic by expressing toxins that can damage or activate host cells, leading to the development of malignant cells by several mechanisms (Clay, Fonseca-Pereira, and Garrett, 2022). In the context of investigating the relationship between colon cancer and bacteria, some toxins have been suggested as possible promoters of tumour development when expressed and released in the tumour microenvironment. In this study, we aimed to detect the presence of these toxins in TCGA samples by searching for reads of genes that express these toxins in RNA-Seq data. We specifically tested for the presence of spermidine/putrescine, fragilisin, fragipain and colibactin, which have been previously reported to be toxic metabolites affecting tumour growth and development in colon cancer (Arthur, 2020; Farriol et al., 2001; Goodwin et al., 2011; Bao et al., 2021). Toxic reads, particularly those related to colibactin and spermidine, were not detected in most of the samples that we analysed, see Table 4.2, first column. Due to this lack of signals, we expanded our search for genes expressing these toxins in other NGS approaches, including WXS and WGS experiments, as each experiment can describe and detect different aspects of bacterial life. Out of the 33 patients whose samples were analysed by RNA-Seq, WXS and WGS, we found only a few samples that tested positive for toxins, with a minimal number of reads assigned to the genes of interest (Table 4.2).

	all RNA-Seq samples	common RNA-Seq samples	common WXS samples	common WGS samples
Colibactin	17 (7)	0	50 (3)	1229 (9)
Fragilisin	0	0	0	14 (1)
Fragipain	0	0	0	68 (7)
Spermidine/putrescine	5 (4)	0	13 (7)	640 (16)

Table 4.2: **Number of reads belonging to toxins in COAD samples.** Number of reads and samples (in brackets) assigned to each toxin gene in all the RNA-Seq samples available (first column) and common samples (last three columns) analysed both with RNA-Seq, WXS and WGS.

Moreover, the results regarding toxin positivity were inconsistent across samples analysed by different experimental methods. For instance, while the three colibactin-positive samples from WXS also tested positive in WGS analysis, only three out of the seven spermidine-positive samples in WXS resulted also positive in WGS detection, even if 16 resulted positive with the latter approach. Given the low number of reads, positive samples and low overlap between NGS approaches, these results cannot be trusted as they may simply be due to chance.

4.4 Bacterial composition

As described in detail in section 3.4.1, we employed a method for quantifying the bacteria present in the samples that involved adjusting the number of reads assigned by accounting for the number of genomes to which each read mapped. Additionally, we limited our analysis to only those bacteria that were detected without any ambiguity. By using this approach, we were able to obtain a more accurate and precise estimation of the bacterial abundance in the samples. Anyway, due to the large number of species present in the TCGA samples, the potential for contamination and the sparsity of the signals, we selected the 1000 species with the highest standard deviation. By doing so, we filtered the most frequently occurring species with a high number of reads and greater variability across the analysed samples. We then utilised PCA to describe the bacterial composition of the samples, with the PCs summarising the variability of the bacterial composition.

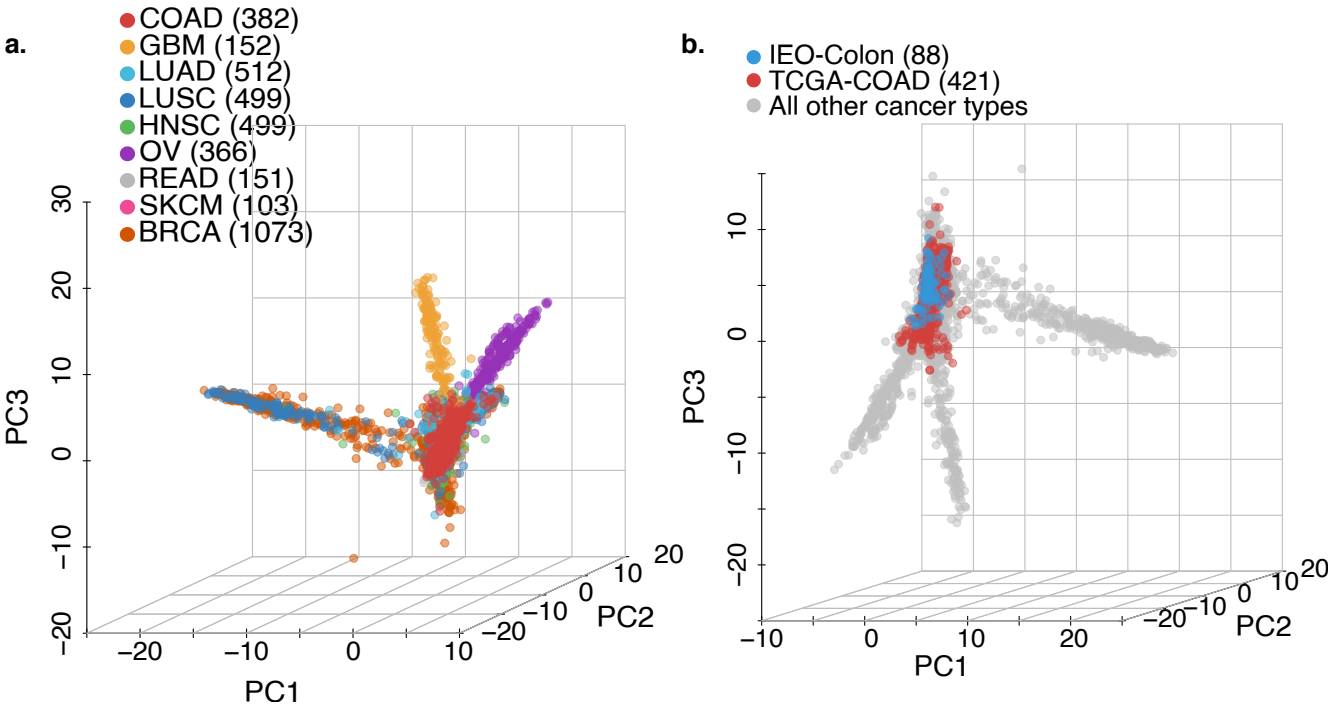


Figure 4.13: **PCA of all the reconstructed microbiome of samples. a.** PCA of all the reconstructed microbiome of TCGA samples analysed. **b.** PCA of all the reconstructed microbiome of TCGA samples analysed and the IEO cohort. Adapted from Sambruni et al., 2023

Initially, we applied this method to all TCGA cancer types and as shown in Figure 4.13a, the samples clustered according to cancer type, indicating that the bacterial compositions of samples from the same cancer type were more similar to each other than those from different types. Interestingly, when we applied the same workflow to the IEO cohort and plotted the results alongside the TCGA data, we observed that COAD and IEO samples clustered together, Figure 4.13b. This finding suggests that despite the potential for technical noise, the bacterial compositions of

samples from colon tumours are similar.
These results are reported in Sambruni et al., 2023.

4.4.1 Technical biases and batch effect

Despite the approaches applied, it is plausible that contamination, noise and technical variation may have influenced the bacterial microbiome reconstruction of the samples. To address this potential issue, we included a computational step in the workflow to detect and correct for any technical variation that could have affected the reconstructed microbiome. This correction step is crucial for ensuring the accuracy and reliability of the results obtained. Further information regarding this step can be found in the Methods chapter, section 3.4.4.

Batch effect detection

In the analysis of each type of cancer, technical factors influenced the clustering. To identify the strongest technical batch affecting the reconstructed microbiome, we utilised the approach outlined in section 3.4.4 of the Methods chapter. The technical features provided by TCGA were evaluated and it was determined that plate ID was the strongest contributor across all tested cancer types except GBM. As a result, no further correction was performed on the reconstructed microbiome of GBM. In the case of COAD and READ, plotting the bacterial reconstructed microbiome in a 2D space revealed the plate ID batch effect, as well as another technical factor that clustered the reconstructed microbiome by the length of the read, used to sequence the samples (50 or 76 bp), as shown in Figure 4.14.
Notably, in the IEO cohort, the sequencing run was identified as the dominant technical batch, which corresponds to what TCGA provides as plate ID.
These results are shown in Sambruni et al., 2023.

Batch effect correction

In the Methods chapter, section 3.4.4 describes how we addressed the main batch effect on the reconstructed microbiome, the plate ID, using the ComBat function (Leek, W. E. Johnson, et al., 2012). After correcting for this batch effect, we observed only minor effects from technical factors on the reconstructed microbiomes (see Figure 4.15 and Figure 4.16a). This indicates that our method can accurately measure bacterial reads across a broad range of samples while controlling for unwanted technical variation and noise in the data. As discussed in the previous section, we

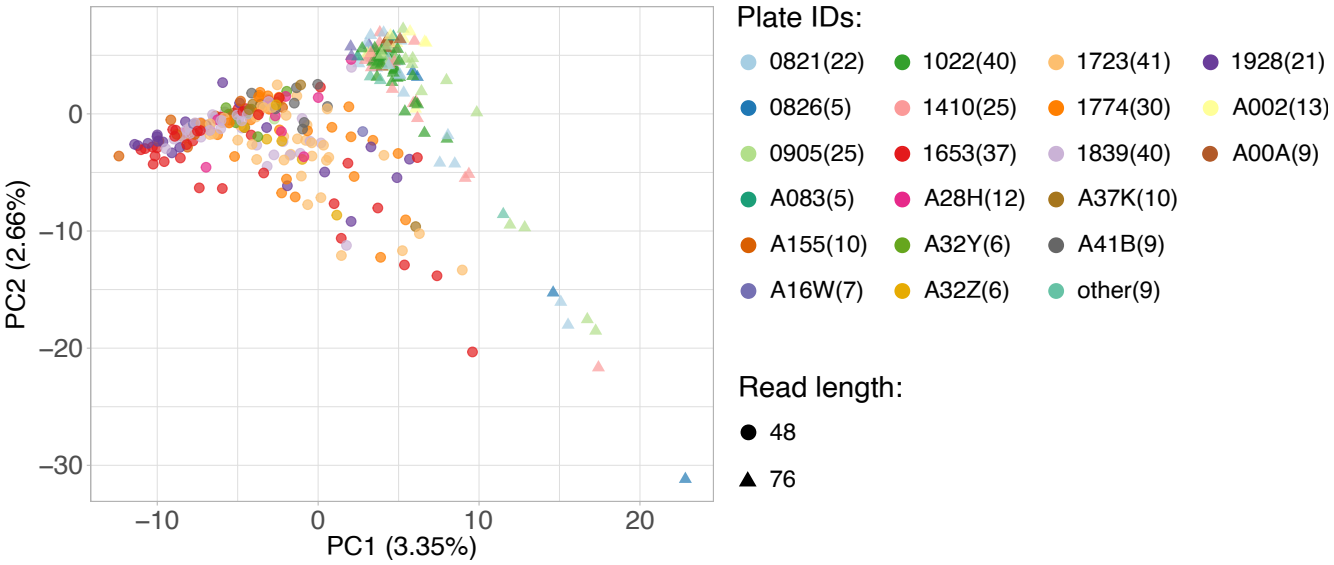


Figure 4.14: **PCA of reconstructed microbiome of COAD samples.** PCA of the reconstructed microbiome of TCGA COAD samples analysed. Colours highlight the plate ID, shape corresponds to the read length. Adapted from Sambruni et al., 2023

also found a secondary batch effect associated with the sequencing read length for COAD and READ. Since the read length is strongly correlated with the plate ID, correcting for the latter also resulted in a decrease in the read length effect (as shown in Figure 4.16b).

IEO samples underwent the same approach to correct for the sequencing run.

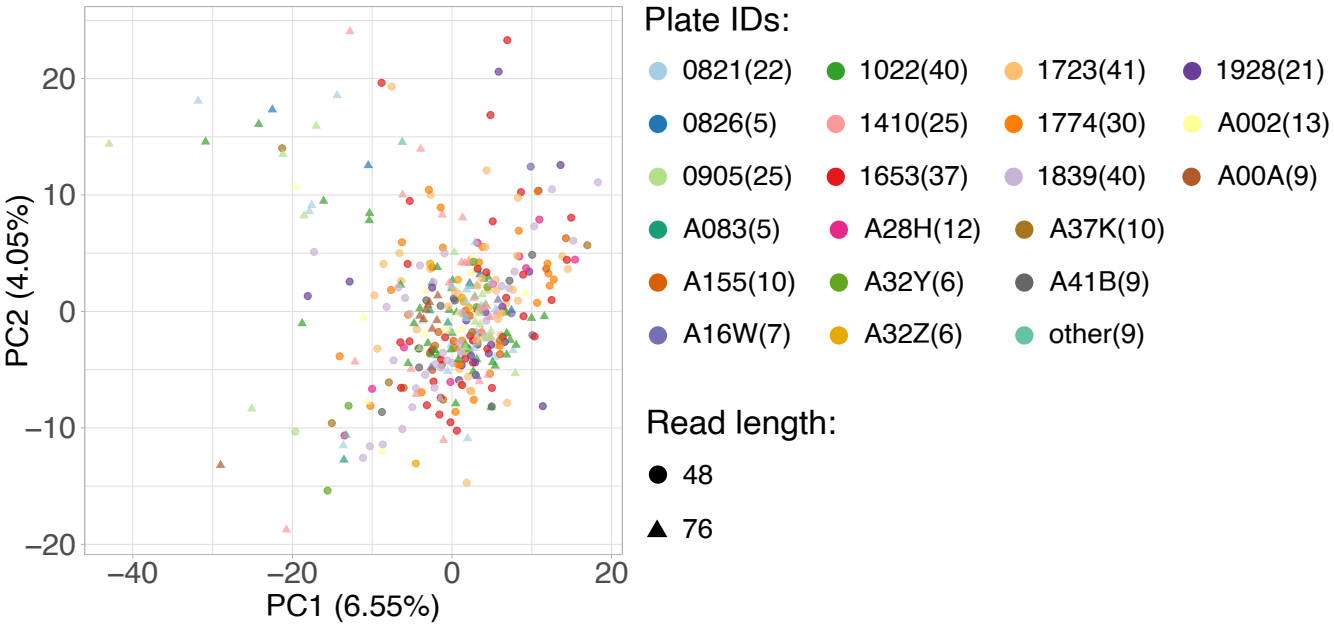


Figure 4.15: **PCA of reconstructed microbiome of COAD samples.** PCA of the reconstructed microbiome of TCGA COAD samples analysed after the batch correction for plate ID. Colours highlight the plate ID, shape corresponds to the read length. Adapted from Sambruni et al., 2023

Our method effectively reduced technical biases in the reconstructed microbiomes of the samples analysed and we used the PCs of the batch-effect corrected data PCA as a proxy for the bacterial reconstructed microbiome composition of the samples.

These results are presented in Sambruni et al., 2023.

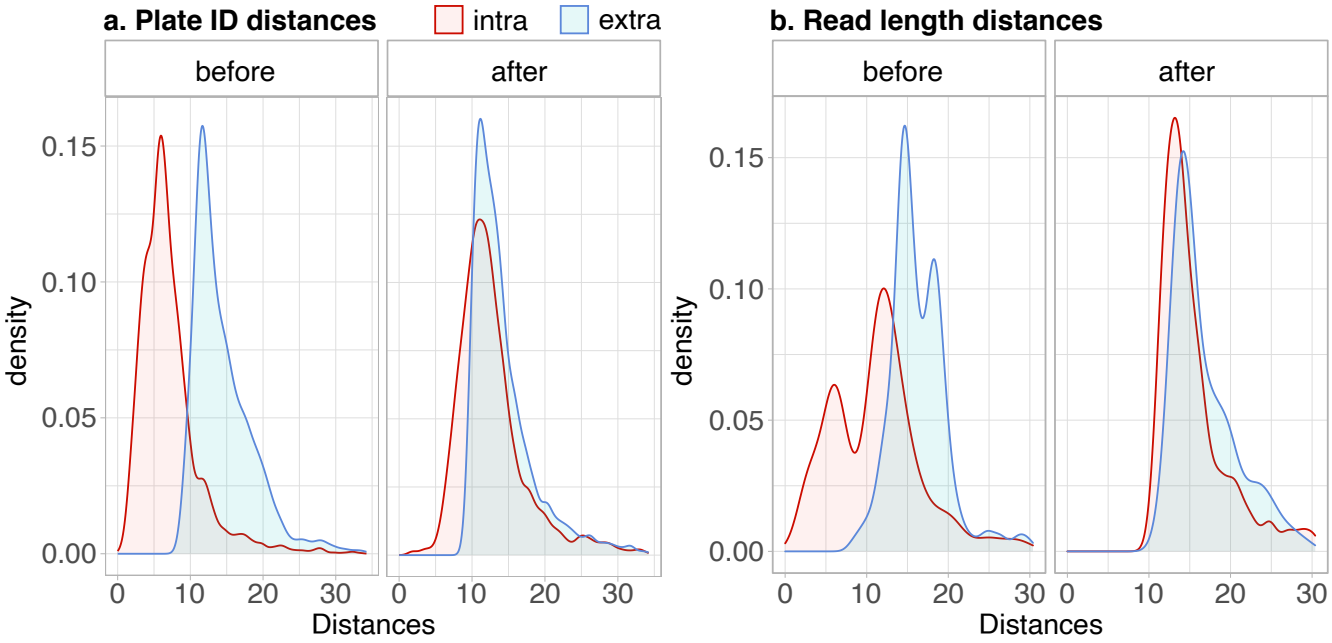


Figure 4.16: **Distances of samples between different technical biases.** (a) Comparison of the distances between COAD, LUAD, LUSC, HNSC, OV, READ, SKCM and BRCA samples analysed in the same plate (intra, red) and the distances between samples analysed in different plates (extra, blue), shown before and after the batch correction by plate ID.(b) Comparison of the distances between COAD and READ samples analysed with the same read length (intra, red) and the distances between samples analysed with different read length (extra, blue), shown before and after the batch correction by plate ID. Even if we corrected by another technical bias (the plate ID), the read length effect bias was reduced. Adapted from Sambruni et al., 2023

4.5 Association between bacteria and cancer properties

We explored whether the bacterial microbiome reconstruction, which was corrected for batch effects, could be utilised to establish associations between the bacterial composition and clinical characteristics of tumours. As the accuracy of extracting bacteria from RNA-Seq data varies for different species, we initially examined the connection between the coordinates of TCGA samples in the microbial abundance space PCA and specific tumour properties. We studied the first six PCs, which accounted for more than 10% of the variance in microbiome composition (refer to section 3.4.3 in the Methods chapter). The tumour properties showing a significant association with the reconstructed bacterial composition of the samples were further analysed with a different approach to detect specific bacterial species associated with the properties to highlight the most relevant bacteria of each property.

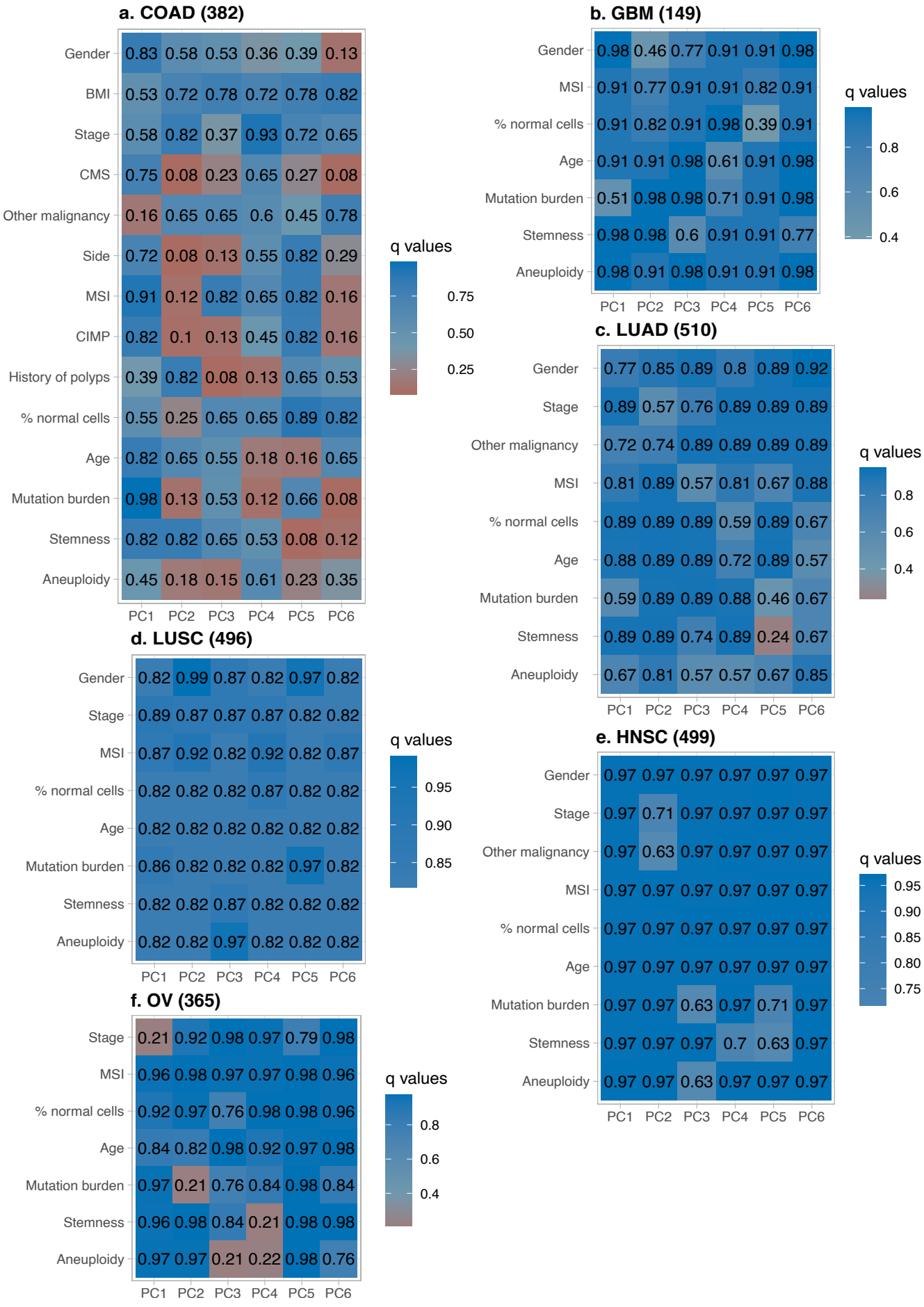
The findings are presented in Sambruni et al., 2023.

4.5.1 Clinical and molecular property association

In our study, we investigated 14 clinical properties in each cancer type, such as age, gender, tumour location (i.e. side), BMI, presence of previous malignancy, history of polyps, stemness of the sample, percentage of normal cells, stage of the tumour, MSI status, CIMP status, CMS (the gene expression–based classification of colon cancer subtypes, Guinney et al., 2015), aneuploidy status and mutation burden, when available (see Methods, section 3.3). Although we tested the microbiome compositions in TCGA cancer types available, measured by bacterial PCs, we did not observe significant associations in most of these cancer types, as shown in Figure 4.17b-i. However, in COAD samples, we did find significant associations between the reconstructed microbiome and several properties, including side, MSI status, CIMP status, CMS, mutation load, age, aneuploidy status, gender, stemness and history of polyps and other malignancies, see Figure 4.17a.

In order to identify the specific bacteria involved in the significant associations in COAD, we conducted further analysis, as outlined in section 3.4.6 of the Methods chapter. Specifically, we tested a subset of species that were colon-specific and prevalent in colon cancer samples, see section 3.4.6, Methods chapter) and controlled for technical variation using the independence test blocking for plate ID (section 3.4.6 of the Methods chapter). This analysis revealed significant bacteria for only three properties: side, MSI and CIMP.

These results are shown in this section arepresented in Sambruni et al., 2023.


a. COAD (382)
Gender 0.83 0.58 0.53 0.36 0.39 0.13
BMI 0.53 0.72 0.78 0.72 0.78 0.82
Stage 0.58 0.82 0.37 0.93 0.72 0.65
CMS 0.75 0.08 0.23 0.65 0.27 0.08
Other malignancy 0.16 0.65 0.65 0.6 0.45 0.78
Side 0.72 0.08 0.13 0.55 0.82 0.29
MSI 0.91 0.12 0.82 0.65 0.82 0.16
CIMP 0.82 0.1 0.13 0.45 0.82 0.16
History of polyps 0.39 0.82 0.08 0.13 0.65 0.53
% normal cells 0.55 0.25 0.65 0.65 0.89 0.82
Age 0.82 0.65 0.55 0.18 0.16 0.65
Mutation burden 0.98 0.13 0.53 0.12 0.66 0.08
Stemness 0.82 0.82 0.65 0.53 0.08 0.12
Aneuploidy 0.45 0.18 0.15 0.61 0.23 0.35
PC1 PC2 PC3 PC4 PC5 PC6
q values
0.75
0.50
0.25
b. GBM (149)
Gender 0.98 0.46 0.77 0.91 0.91 0.98
MSI 0.91 0.77 0.91 0.91 0.82 0.91
% normal cells 0.91 0.82 0.91 0.98 0.39 0.91
Age 0.91 0.91 0.98 0.61 0.91 0.98
Mutation burden 0.51 0.98 0.98 0.71 0.91 0.98
Stemness 0.98 0.98 0.6 0.91 0.91 0.77
Aneuploidy 0.98 0.91 0.98 0.91 0.91 0.98
PC1 PC2 PC3 PC4 PC5 PC6
q values
0.8
0.6
0.4
c. LUAD (510)
Gender 0.77 0.85 0.89 0.8 0.89 0.92
Stage 0.89 0.57 0.76 0.89 0.89 0.89
Other malignancy 0.72 0.74 0.89 0.89 0.89 0.89
MSI 0.81 0.89 0.57 0.81 0.67 0.88
% normal cells 0.89 0.89 0.89 0.59 0.89 0.67
Age 0.88 0.89 0.89 0.72 0.89 0.57
Mutation burden 0.59 0.89 0.89 0.88 0.46 0.67
Stemness 0.89 0.89 0.74 0.89 0.24 0.67
Aneuploidy 0.67 0.81 0.57 0.57 0.67 0.85
PC1 PC2 PC3 PC4 PC5 PC6
q values
0.8
0.6
0.4
d. LUSC (496)
Gender 0.82 0.99 0.87 0.82 0.97 0.82
Stage 0.89 0.87 0.87 0.87 0.82 0.82
MSI 0.87 0.92 0.82 0.92 0.82 0.87
% normal cells 0.82 0.82 0.82 0.87 0.82 0.82
Age 0.82 0.82 0.82 0.82 0.82 0.82
Mutation burden 0.86 0.82 0.82 0.82 0.97 0.82
Stemness 0.82 0.82 0.87 0.82 0.82 0.82
Aneuploidy 0.82 0.82 0.97 0.82 0.82 0.82
PC1 PC2 PC3 PC4 PC5 PC6
q values
0.95
0.90
0.85
e. HNSC (499)
Gender 0.97 0.97 0.97 0.97 0.97 0.97
Stage 0.97 0.71 0.97 0.97 0.97 0.97
Other malignancy 0.97 0.63 0.97 0.97 0.97 0.97
MSI 0.97 0.97 0.97 0.97 0.97 0.97
% normal cells 0.97 0.97 0.97 0.97 0.97 0.97
Age 0.97 0.97 0.97 0.97 0.97 0.97
Mutation burden 0.97 0.97 0.63 0.97 0.71 0.97
Stemness 0.97 0.97 0.97 0.7 0.63 0.97
Aneuploidy 0.97 0.97 0.63 0.97 0.97 0.97
PC1 PC2 PC3 PC4 PC5 PC6
q values
0.95
0.90
0.85
0.80
0.75
f. OV (365)
Stage 0.21 0.92 0.98 0.97 0.79 0.98
MSI 0.96 0.98 0.97 0.97 0.98 0.96
% normal cells 0.92 0.97 0.76 0.98 0.98 0.96
Age 0.84 0.82 0.98 0.92 0.97 0.98
Mutation burden 0.97 0.21 0.76 0.84 0.98 0.84
Stemness 0.96 0.98 0.84 0.21 0.98 0.98
Aneuploidy 0.97 0.97 0.21 0.22 0.98 0.76
PC1 PC2 PC3 PC4 PC5 PC6
q values
0.8
0.6
0.4

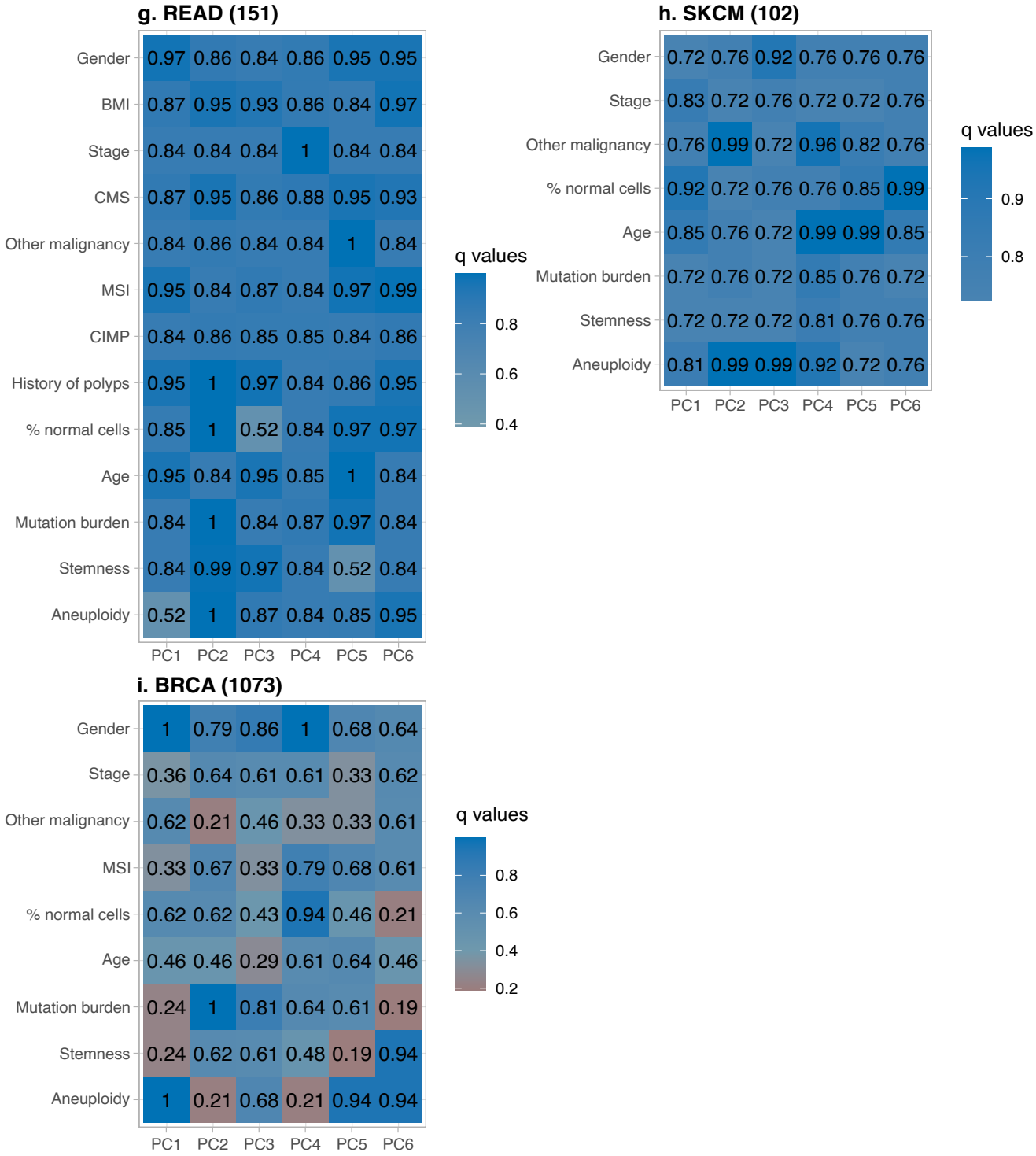


Figure 4.17: **Heatmaps of the associations between the reconstructed microbiome of TCGA cancer types and clinical and molecular properties.** Heatmaps of the q values of the associations and correlation between the first six PCs of the PCA on the reconstructed microbiome and the clinical properties from the metadata of (a) COAD, (b) GBM, (c) LUAD, (d) LUSC, (e) HNSC, (f) OV, (g) READ, (h) SKCM and (i) BRCA. No batch correction has been applied on GBM bacteria quantification, while the other tissues underwent plate ID correction. Number of samples analysed in brackets. Adapted from Sambruni et al., 2023

Side: Our analysis identified nine species with differential abundances between the left and right sides of the colon. Specifically, we observed that *F. prausnitzii*, *Coprococcus comes* and two *Bacteroides spp.* (*Bacteroides vulgatus* and *Bacteroides thetaiotaomicron*) were more abundant in samples from the right side of the colon (Figure 4.18). Interestingly, these four bacterial species were among the top 20% of species contributing to PC2, the most robust side-associated PC (Figure 4.18).

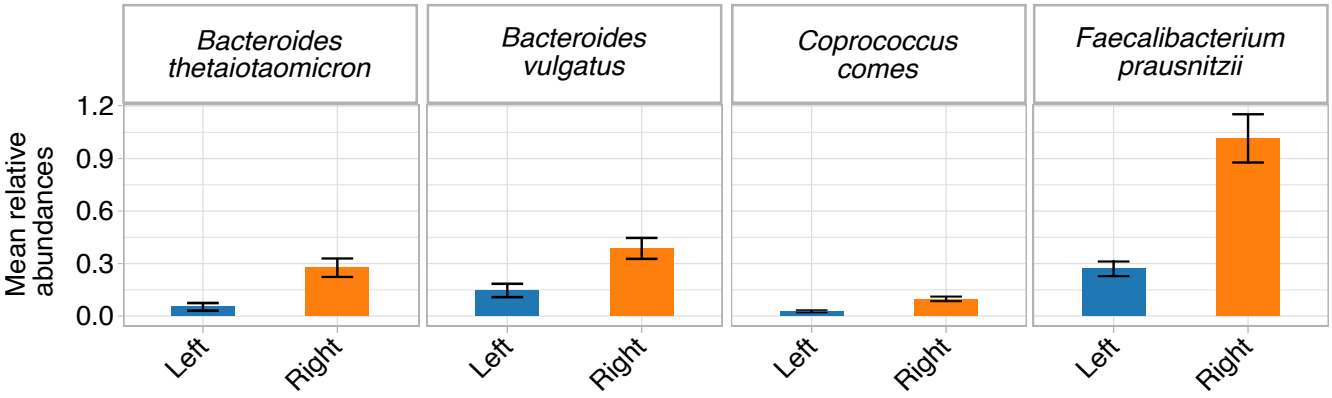


Figure 4.18: **Species associated with the side of COAD samples**. Barplots of the mean scores of a subset of the bacterial species with differential distribution in the side. In total, we found nine bacteria in side. Adapted from Sambruni et al., 2023

MSI: In our analysis, we observed that five bacterial species exhibited significantly higher abundances in samples with high MSI compared to those with low MSI (Figure 4.19). Among these species were *B. fragilis*, *Clostridium asparagiforme*, *Fusobacterium sp. OBRC1* and *Bacteroides sp. 3_2_5*, which were strongly associated with two principal components (PC2 and PC6) that were linked to MSI level (Figure 4.19).

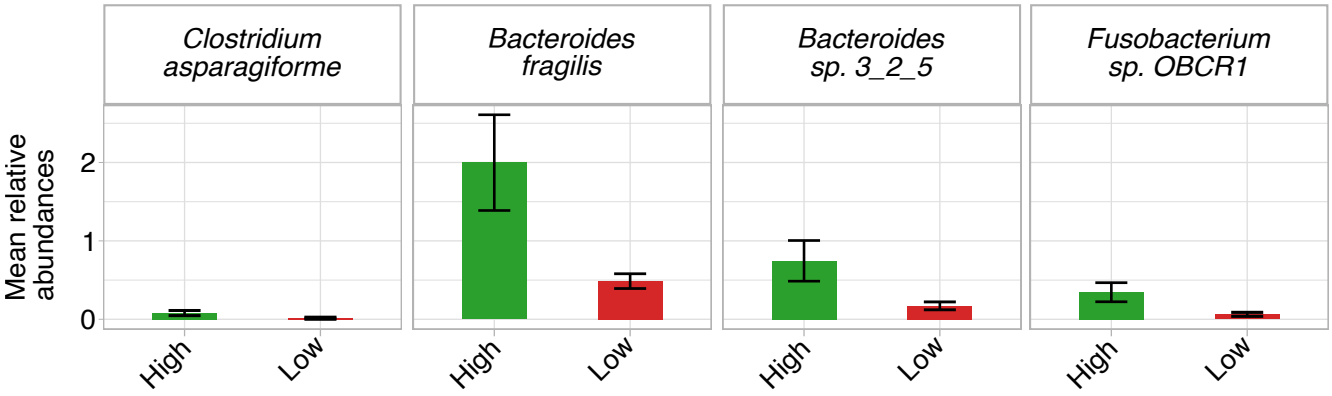


Figure 4.19: **Species associated with the MSI of COAD samples.** Barplots of the mean scores of a subset of the bacterial species with differential distribution in the MSI. In total, we found five bacteria in MSI. Adapted from Sambruni et al., 2023

CMS: Given the specific molecular and clinical properties characterising colon cancer CMS, we investigated whether the microbial composition of each CMS was distinct too. As expected, we

observed significant variations in microbiome composition across different CMS types, with CMS1 displaying a unique microbiome profile (Figure 4.20). CMS1 is characterised by robust immune

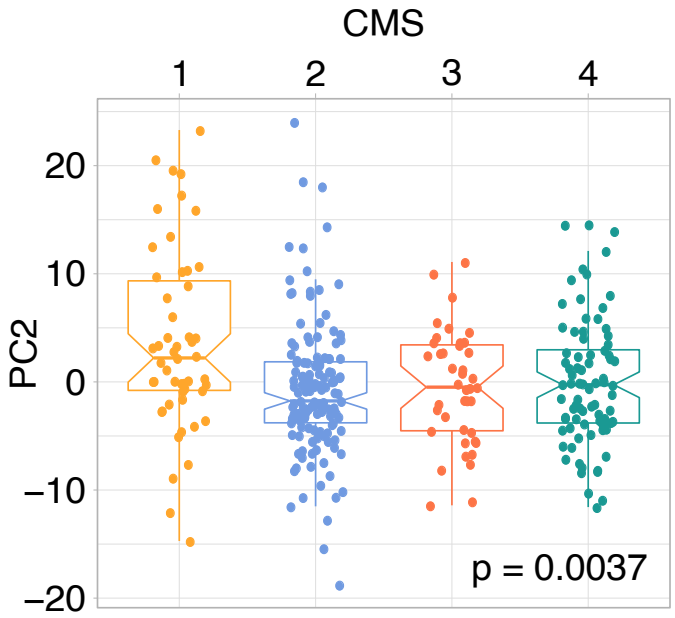


Figure 4.20: **Species associated with the CMS of COAD samples.** Boxplot of PC2 coordinates by CMS, highlighting the particular behaviour of CMS1 reconstructed microbiomes. Adapted from Sambruni et al., 2023

cell infiltration and activation of immune evasion pathways (Guinney et al., 2015), suggesting a possible link between the microbiota and the immune landscape of colon cancer. Therefore, we estimated the immune cell proportion of the samples and tested the correlation of these values with the microbial composition of COAD samples (see the next section 4.5.2). Additionally, we investigated the bacterial species associated with the highly immune-infiltrated CMS1 (Figure 4.21) and identified 16 bacteria from *Clostridium*, *Bacteroides*, *Fusobacterium*, *Peptostreptococcus* and *Selenomonas* genera and *Firmicutes* phylum. Among these, we found five *Fusobacterium* species with higher levels in the CMS1 subgroup, although not *F. nucleatum* itself, which has previously been linked to colorectal cancer growth and progression (Clay, Fonseca-Pereira, and Garrett, 2022). These bacteria strongly contributed to PC2 or PC6, the two PCs associated with CMS. Moreover, we found that five species of *Clostridium* were associated with CMS1, while *Clostridium perfringens* did not contribute to PC2.

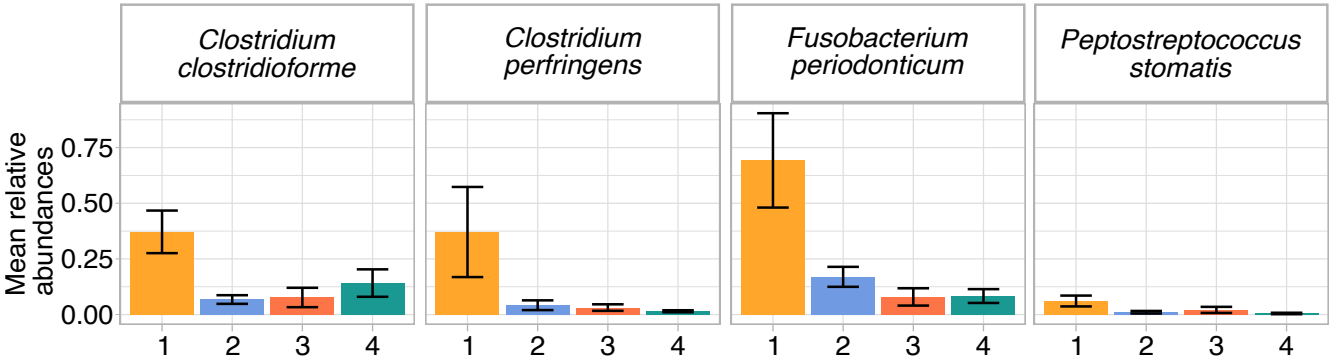


Figure 4.21: **Species associated with the CMS of COAD samples**. Barplots of the mean scores of a subset of the bacterial species with differential distribution in the CMS. In total, we found 18 bacteria in CMS. Adapted from Sambruni et al., 2023

4.5.2 Immune infiltration association

As CMS1 is known for its strong immune cell infiltration and activation of immune evasion pathways (Guinney et al., 2015) and the microbial composition of COAD samples was significantly associated with CMS1 samples (see the previous section 4.5.1), we used the gene expression data of these samples to estimate their immune landscape using the tool CIBERSORTx (Newman et al., 2019). This tool is able to quantify the proportion of 22 immune cell types in each sample, see section 3.5 in Methods chapter. We tested this immune landscape against the reconstructed microbiome of COAD samples and we observed a significant association between the PCs and the estimates of dendritic cells and mast cells in COAD samples ($q < 0.2$; determined using Wilcoxon, Kruskal-Wallis or Spearman correlation tests, as described in the Methods chapter, section 3.4.5). This association was evident not only in terms of immune cell proportion but also in terms of the abundance of immune cells present in each sample, measured again by CIBERSORTx (Figure 4.22).

In a similar way described above, we were interested in identifying the specific bacteria involved in the significant associations between the microbial composition and the immune cells in COAD samples. We tested a subset of species that were colon-specific and prevalent in colon cancer samples and controlled for technical variation using the independence test blocking for plate ID (section 3.4.6 of the Methods chapter). We found that the absence of resting mast cells was linked to five bacterial species, while activated mast cells displayed an opposite tendency. These bacterial species included *B. fragilis*, *Clostridium clostridioforme*, *Clostridiales bacterium 1_7_47FAA* and *Clostridium sp. FS41* (Figure 4.23). Interestingly, all of these bacteria were found to contribute to PC2, which was associated with mast cell infiltration. Additionally, *C. clostridioforme* was among the bacterial species linked to CMS1.

These results are presented in Sambruni et al., 2023.

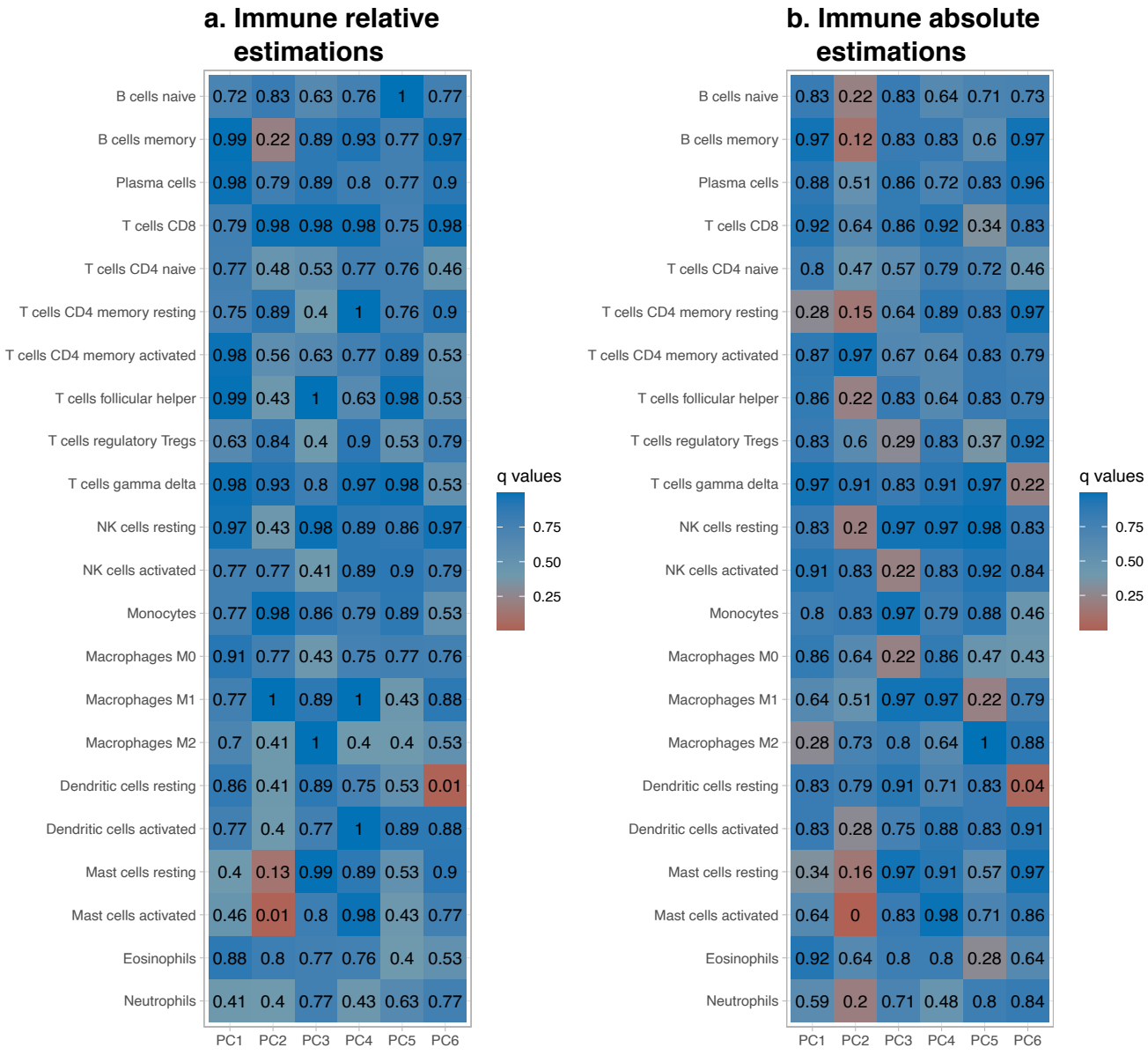


Figure 4.22: **Heatmaps of the associations between the reconstructed microbiome of COAD and immune cell estimations**. Heatmaps of the q values of the association and correlation between the first six PCs of COAD microbiome profiles (PCs in rows, clinical properties in columns). Immune cell estimation in (a) relative and (b) absolute values. Adapted from Sambruni et al., 2023

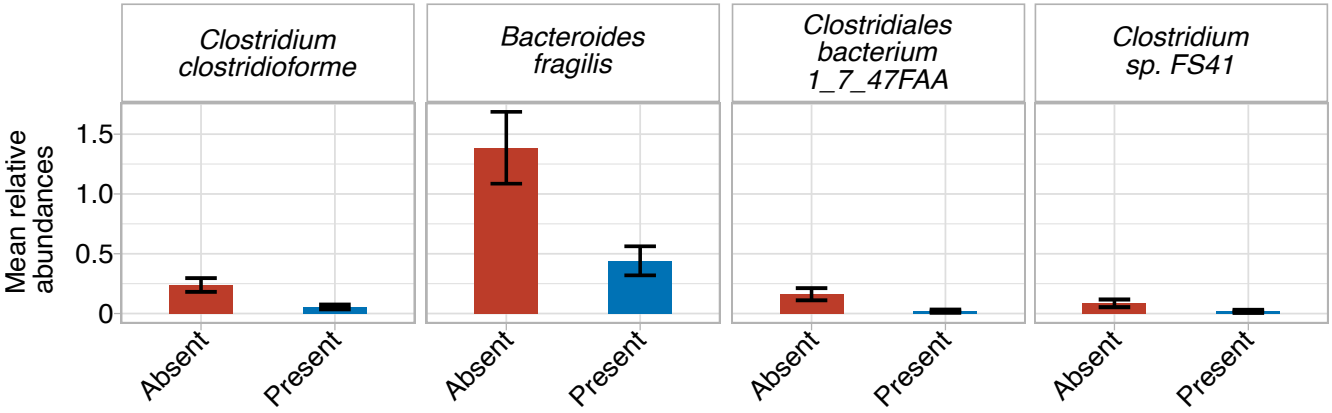


Figure 4.23: **Species associated with the mast cell level of COAD samples**. Barplots of the means of the scores of a subset of the bacterial species with differential distribution in high or low level of mast cell estimations. In total, we found 12 bacteria associated with mast cell level. Adapted from Sambruni et al., 2023

4.5.3 Gene mutation status associations

Host genetics plays a crucial role in determining the growth and progression of tumours and it has been proposed that it may also affect the microbial ecosystem associated with the tumour (Mousa, Chehadeh, and Husband, 2022; Yadav and Chauhan, 2021). Therefore, we investigated the relationship between the bacterial composition found in tumour samples and the mutation status of commonly mutated genes, see Methods chapter, section 3.3. However, we did not find any significant associations between them (Figure 4.24). This indicates that there may not be a direct correlation between the bacterial composition of the samples and the commonly mutated genes or our analysis may not be sensitive enough to detect such a link.

These findings are represented in Sambruni et al., 2023.

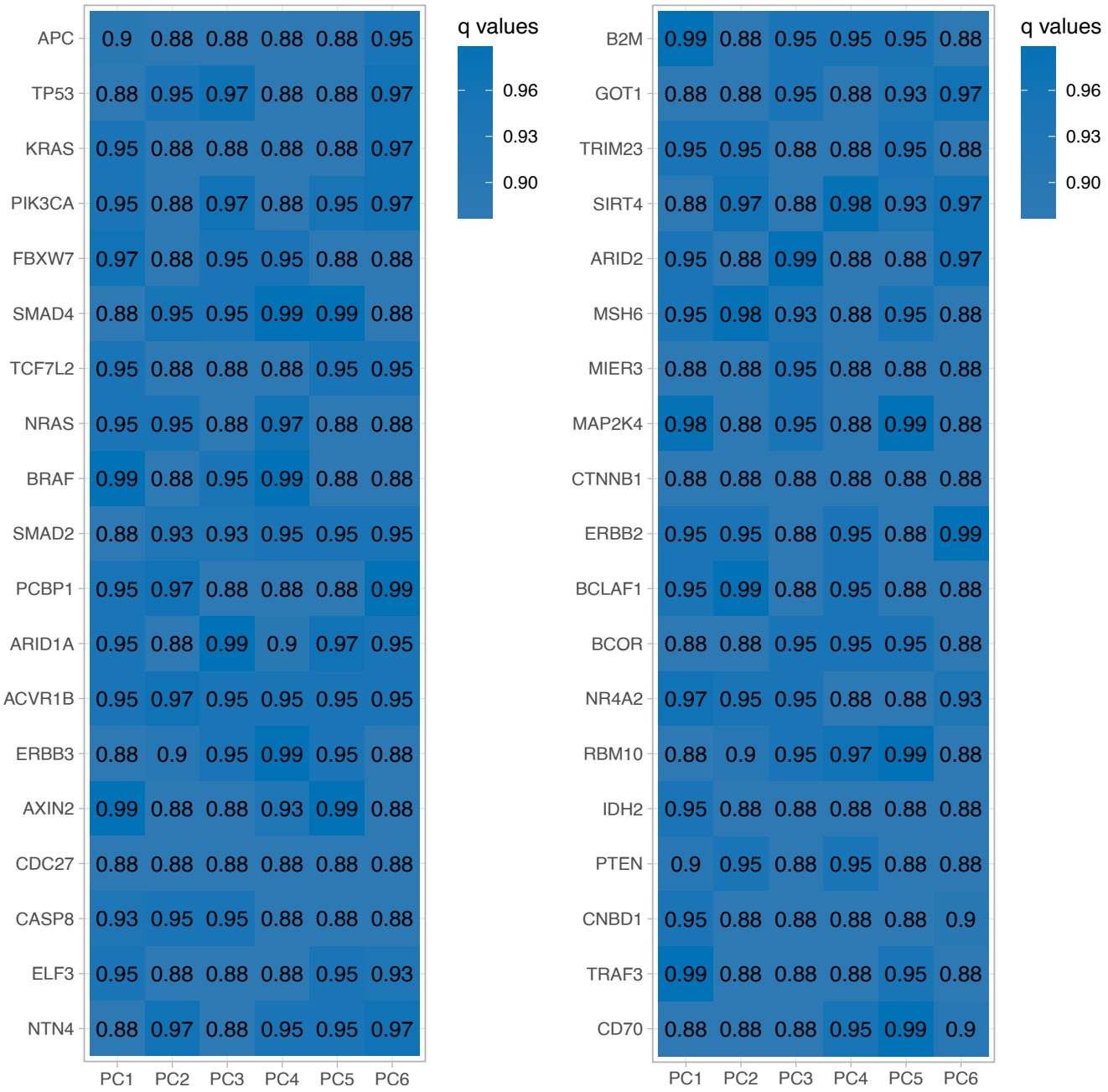
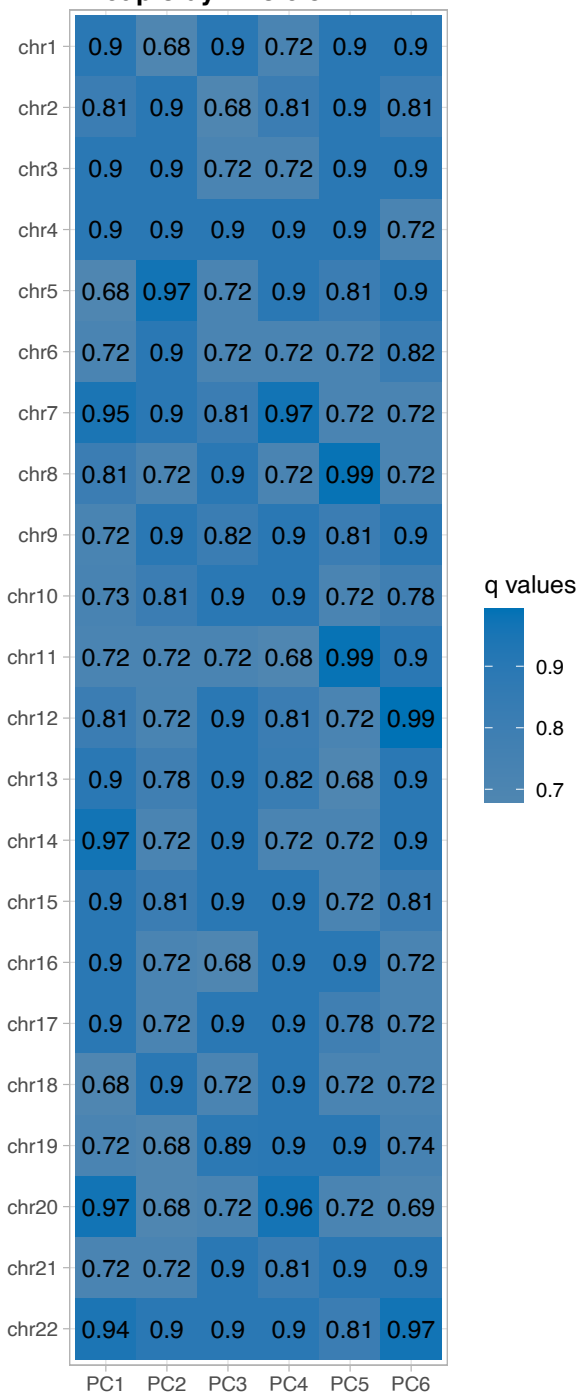


Figure 4.24: **Heatmaps of the associations between the reconstructed microbiome of COAD and mutation status of commonly mutated genes in colorectal cancer.** Heatmaps of the q values of the associations and correlation between the first six PCs of the PCA on the reconstructed microbiome and the mutation status of commonly mutated genes of colorectal cancer. COAD bacteria quantification underwent plate ID correction. Adapted from Sambruni et al., 2023

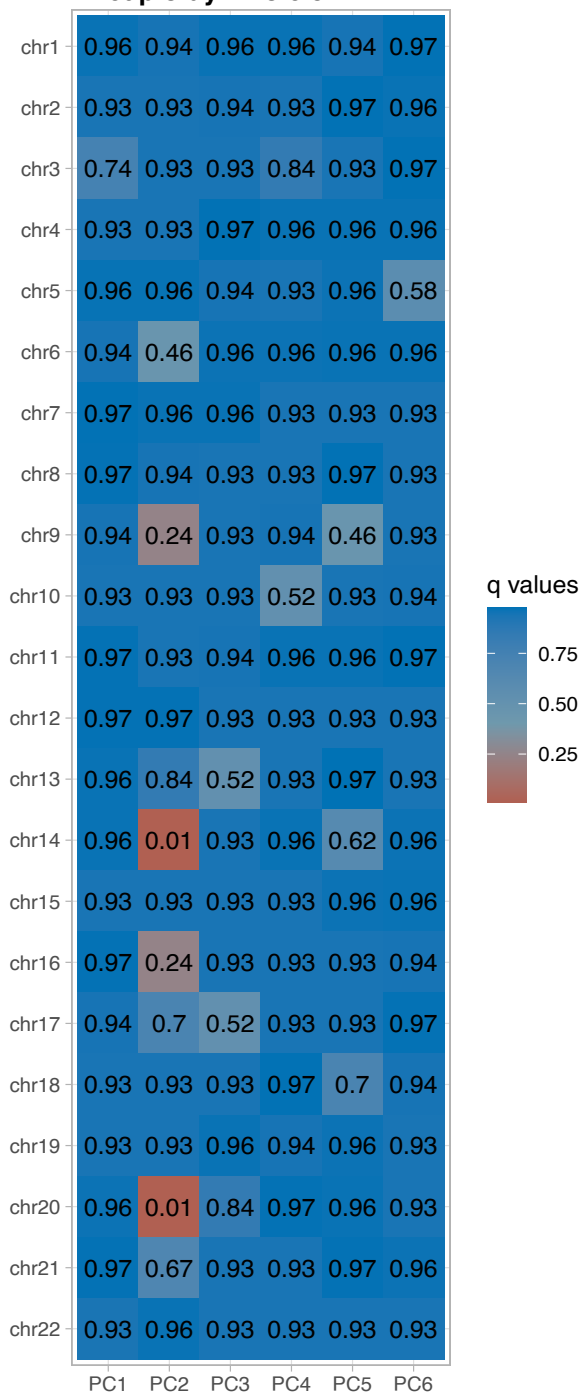
4.5.4 Aneuploidy associations

As mentioned in the previous paragraph, we investigated the relationship between the host somatic genetic background and the bacterial composition of the tumour. Next, we explored whether there was a correlation between the microbiome composition and chromosomal gain or loss in COAD. Our analysis showed no significant association between the microbiome composition and chromosomal gain or loss in COAD, as demonstrated by the results depicted in Figure 4.25a. It is intriguing to note that while there is a correlation between the overall chromosomal aberration in COAD (see Figure 4.17a), there is no specific association with the change of any particular chromosome. However, we did observe a correlation between the microbes and the aneuploidy status in other cancer types such as HNSC (chromosome loss at 14 and 20), OV (chromosome 14 alteration) and READ (chromosome 2 deletion), see Figure 4.25b-d.
These results are shown in Sambruni et al., 2023.

a. COAD (382) - Aneuploidy whole chr



b. HNSC (499) - Aneuploidy whole chr



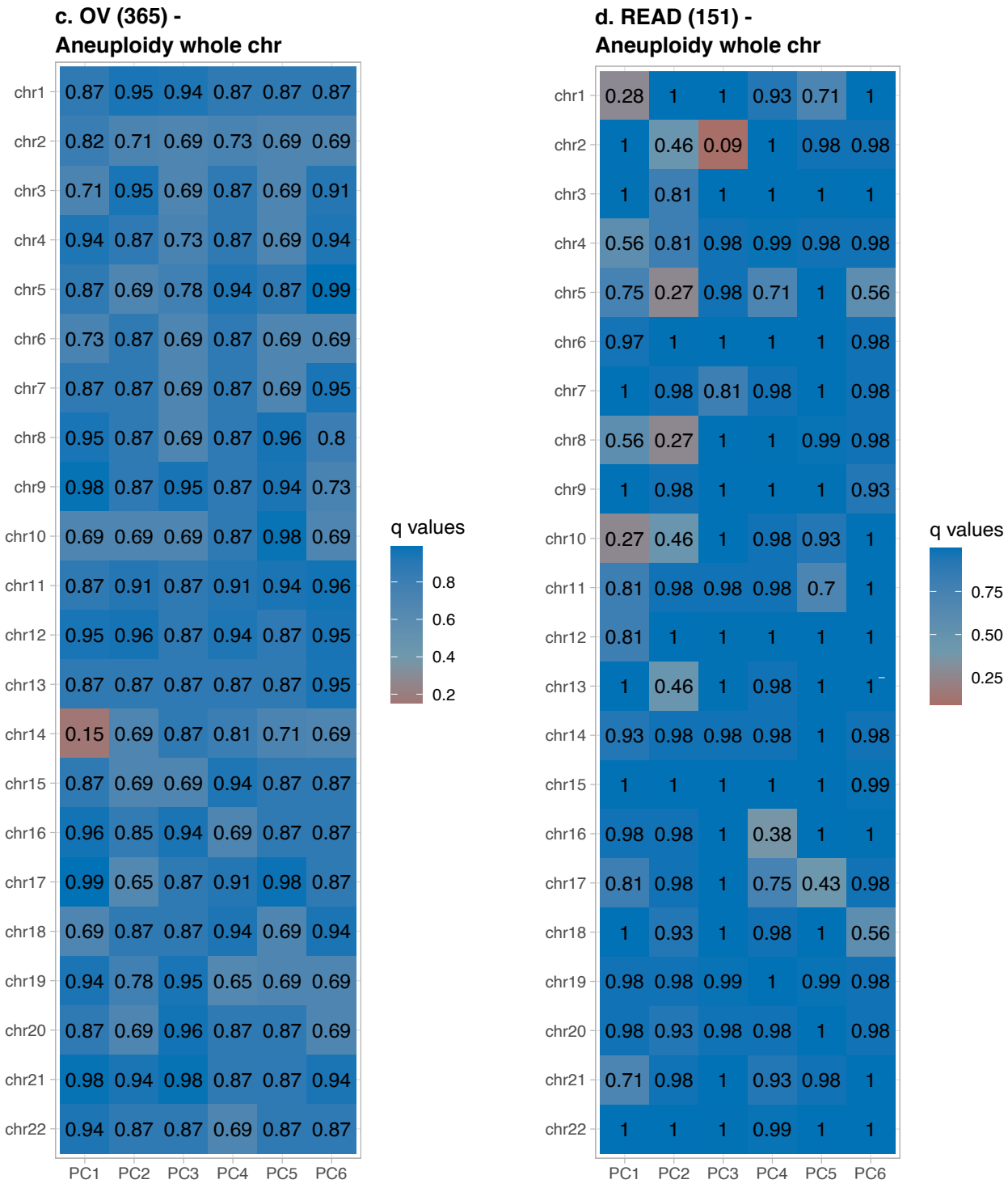


Figure 4.25: **Heatmaps of the associations between the reconstructed microbiome of COAD, HNSC, OV and READ and aneuploidy status.** Heatmaps of the q values of the associations between the first six PCs of the PCA on the reconstructed microbiome and the aneuploidy status of whole chromosomes of (a) COAD, (b) HNSC, (c) OV and (d) READ. Bacteria quantification underwent plate ID correction. Number of samples analysed in brackets. Adapted from Sambruni et al., 2023

4.6 Bacterial association in normal samples

In order to determine if the observed associations between microbiome composition and clinical properties described in the section 4.5.1 were specific to COAD tumour microenvironment or reflected a more general dysbiosis of the colon, we tested if the associations held in non-malignant tissues available from TCGA. The non-malignant samples were collected from areas of the colon that were not affected by the tumour and therefore considered "normal". However, the microbiota in the colon of a cancer patient may have been influenced by the disease. Therefore, we compared the microbiome composition of tumour vs. non-pathologic microenvironments rather than a real tumour vs. healthy ones.

The reconstructed microbiome of these non-malignant samples of the colon did not show any

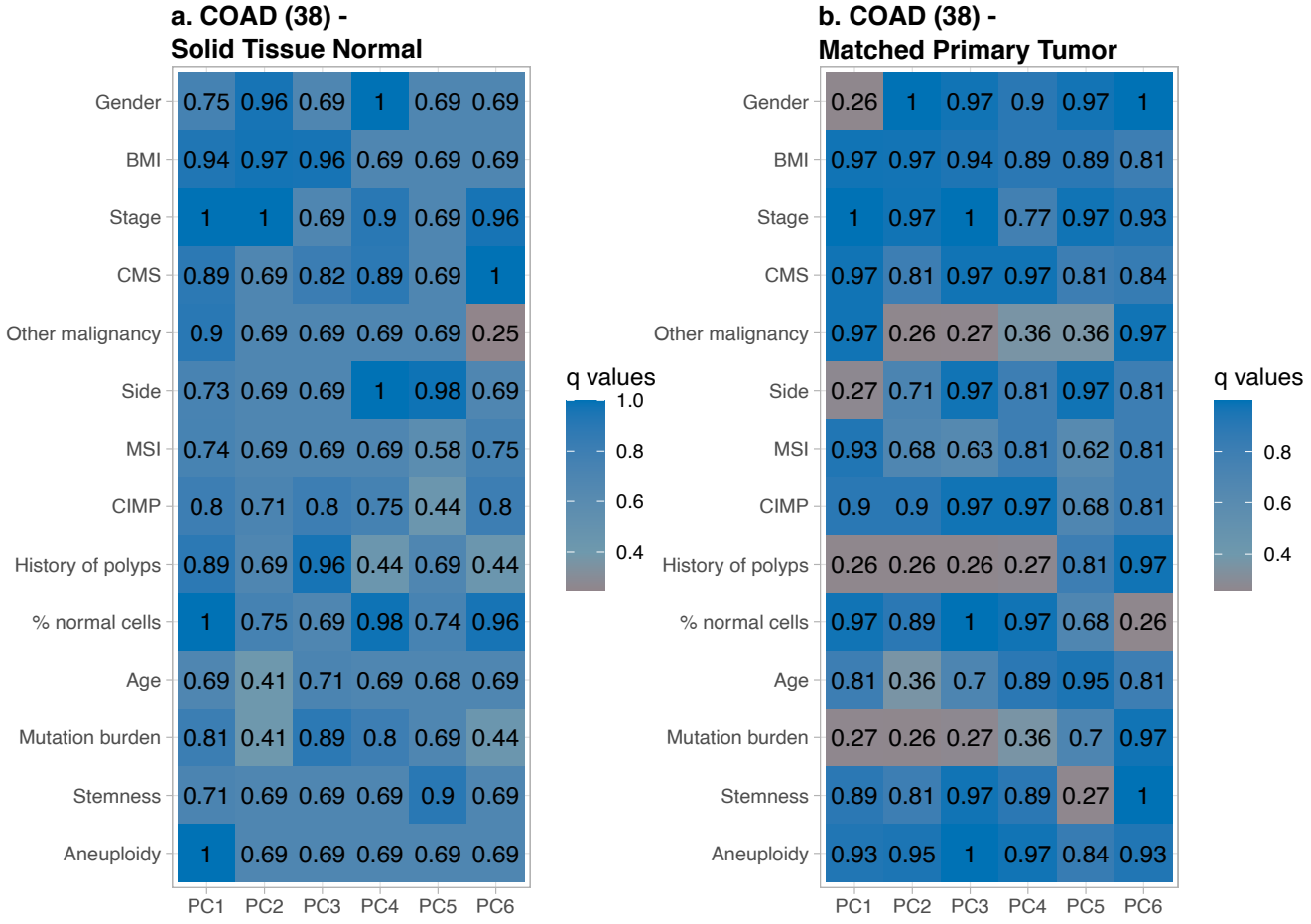


Figure 4.26: **Heatmaps of the associations between the reconstructed microbiome of the subset of paired normal-tumour COAD samples and clinical and molecular properties.** Heatmaps of the q values of the associations and correlation between the first six PCs of the PCA on the reconstructed microbiome of (a) normal and (b) subset of tumour samples derived from the same patient of the normal ones (matching) and the clinical and molecular properties of COAD. No corrections applied. Number of samples analysed in brackets. Adapted from Sambruni et al., 2023

significant association with the available clinical properties (Figure 4.26a). Nonetheless, the lack

of significance may be due to the lower statistical power of the TCGA dataset, which only contains 39 non-malignant colon samples. To exclude this possibility, we tested the same associations in the subset of COAD tumour samples matching with the non-malignant ones (extracted from the same patient). Among the significant associations, we found lower significance levels in seven out of ten cases in the paired, reduced tumour cohort compared to non-malignant samples, implying that these associations are specific to the tumour microenvironment rather than the non-malignant tissue (Figure 4.26b). Although we cannot rule out the possibility that significant associations may be found with a larger number of non-malignant samples, the complete absence of associations in our pool of samples suggests that the detected associations are tumour-microenvironment specific.
In section 4.5.1, we detected bacterial species associated with tumour properties and we wondered if these bacteria exhibited similar behaviour in non-malignant samples. To explore this possibility, we measured the abundance ratio of the bacterial species across the property tested, see Methods chapter, section 4.5.1. This allowed us to investigate whether the bacteria detected in tumour samples were also present in non-cancerous tissues and whether their presence correlated with any distinct properties. We tested if the bacteria detected as differentially distributed between the levels of side, MSI and CMS in tumour samples show the same trend in normal samples. We were able to detect a higher difference between the levels in tumour samples compared to the normal ones, even if we subset the tumour samples to comprehend only the ones with matched normal samples (derived from the same patients of the normal samples). Only *C. asparagiforme* and *Fusobacterium periodonticum* showed ambiguous behaviour (Figure 4.27).
Some results presented in this section have been previously reported in Sambruni et al., 2023.

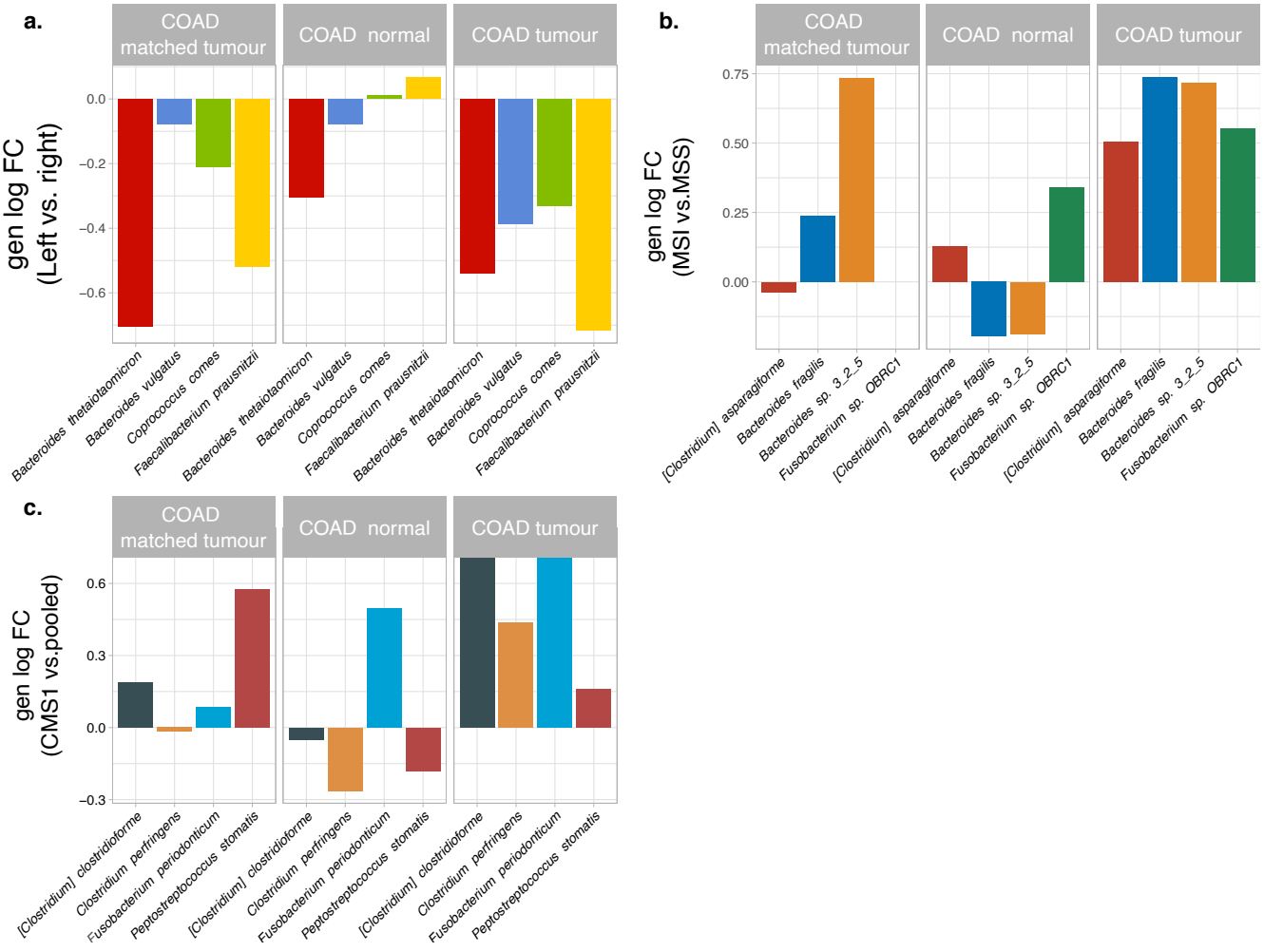


Figure 4.27: **Tendency of bacteria associated with tumour properties in normal and tumour samples with matched normal ones of COAD.** Generalised log fold change of the scores of the bacteria detected as differentially distributed in tumour COAD samples between the levels of (a) side, (b) MSI and (c) CMS, shown in tumour, normal and normal-matching tumour samples of COAD.

4.7 Association between bacteria and clinical outcome

Since the molecular and immunological characteristics of the tumour have been extensively studied and shown to play a critical role in determining clinical outcomes in colon cancer (Buikhuisen, Torang, and Medema, 2020), understanding the potential relationship between the microbial composition of cancer microenvironment and clinical prognosis could provide other important insights into the progression of this disease and highlight the involvement of bacterial to tumour microenvironment. Hence, we investigated the possible link between the bacterial composition derived from RNA-Seq data and clinical prognosis of TCGA samples. We used Cox proportional-hazard models to analyse the top six PCs coordinates and conducted univariate analyses to assess the impact of each PC coordinate on OS and DFS, see 4.3.

Interestingly, among all the cancer types, we found that only COAD reconstructed microbiome

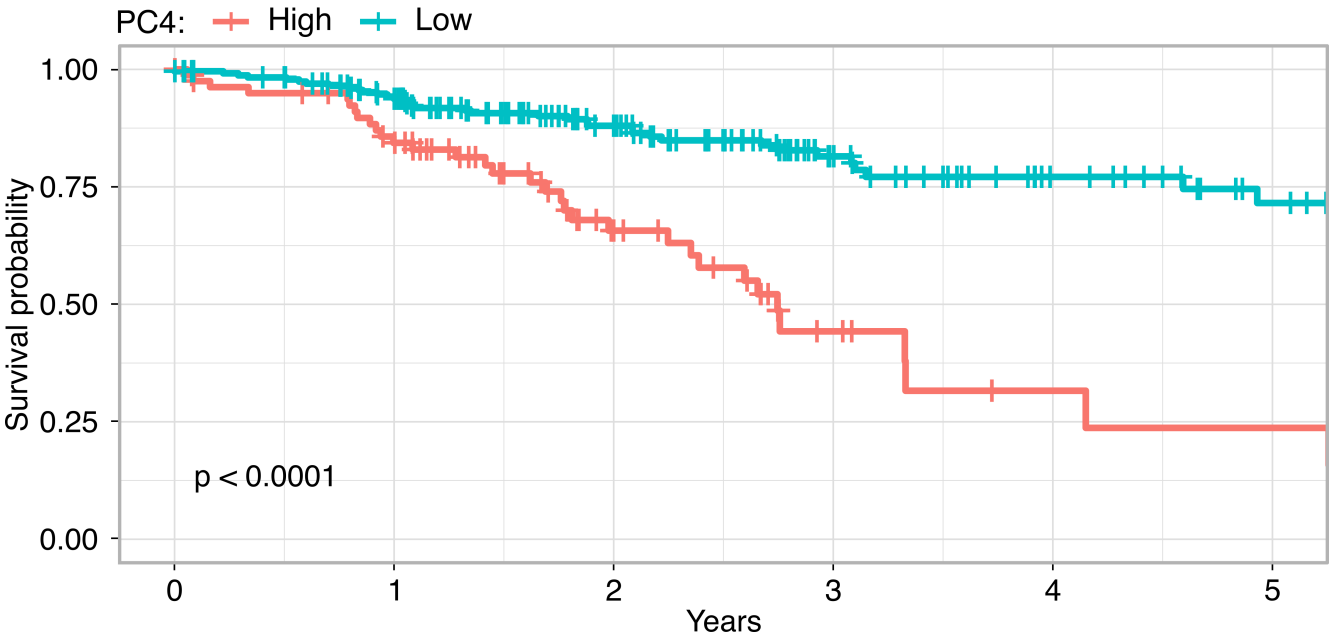


Figure 4.28: **Disease-free survival results for COAD samples.** Kaplan-Meier analysis of disease-free survival on COAD patients stratified by high or low values of PC4. Adapted from Sambruni et al., 2023

was associated with survival data. In particular, PC4 was significantly associated with DFS. The top 20 bacteria contributing to PC4 included *Cutibacterium granulosum*, *Corynebacterium tuberculostearicum*, *Moraxella osloensis*, *Gemella haemolysans*, *Staphylococcus epidermidis*, *Finegoldia magna*, *Lawsonella clevelandensis* and *Acinetobacter baumannii*. To further examine this connection between bacteria and survival data, we then divided patients into "high" and "low" groups based on their PC4 coordinates and applied Kaplan-Meier analysis, revealing that patients with higher PC4 coordinates had a higher probability of relapsing, see Figure 4.28. As mentioned before, patient outcome is influenced by various clinical and molecular tumour properties, we tested if the

association of PC4 with DFS was independent of other properties associated with PC4, namely age, polyps history and mutation load (see Figure 4.17a). We applied the Cox proportional-hazard model and demonstrated that the bacterial association with DFS was independent of molecular properties suggesting that the microbiome composition may be directly connected to the risk of relapse in patients with COAD.

Following the approach mentioned above, we aimed to identify the specific bacteria that were responsible the most for the significant associations between microbial composition and patient relapse. Based on the significant association of PC4 with DFS in COAD patients, we selected the top 100 bacteria with the highest PC4 loading values and performed a univariate Cox analysis. As a result, 17 bacteria showed significant positive associations with the relapse probability (Hazard ratio > 3; $q < 0.2$; Wald test), including *Corynebacterium matruchotii*, *A. baumannii*, *Pseudomonas stutzeri* and *Propionibacterium namnetense*.

These results were reported in Sambruni et al., 2023.

Cancer Type	PCs	beta	HR	lower 95 HR CI	upper 95 HR CI	wald.test	p.value	q.value
COAD	PC1	-1.3	0.28	0.051	1.6	2.1	0.15	1
COAD	PC2	-0.18	0.84	0.16	4.4	0.04	0.83	1
COAD	PC3	2.1	8.3	0.71	97	2.8	0.091	1
COAD	PC4	2.7	15	3.3	72	12	0.00052	0.02808
COAD	PC5	0.36	1.4	0.12	18	0.08	0.78	1
COAD	PC6	-1.4	0.24	0.035	1.6	2.1	0.15	1
GBM	PC1	-0.018	0.98	0.39	2.5	0	0.97	1
GBM	PC2	0.29	1.3	0.41	4.4	0.23	0.63	1
GBM	PC3	1.1	3	0.42	22	1.2	0.27	1
GBM	PC4	-0.48	0.62	0.014	27	0.06	0.81	1
GBM	PC5	0.73	2.1	0.29	15	0.52	0.47	1
GBM	PC6	0.092	1.1	0.13	9.2	0.01	0.93	1
HNSC	PC1	-1.4	0.25	0.047	1.3	2.8	0.096	1
HNSC	PC2	-0.51	0.6	0.15	2.4	0.52	0.47	1
HNSC	PC3	-0.88	0.41	0.044	3.9	0.6	0.44	1
HNSC	PC4	-0.38	0.68	0.051	9.1	0.08	0.77	1
HNSC	PC5	2	7.4	0.72	77	2.8	0.093	1
HNSC	PC6	-1.4	0.24	0.0051	11	0.53	0.47	1
LUAD	PC1	-0.31	0.74	0.3	1.8	0.45	0.5	1
LUAD	PC2	-1.2	0.29	0.067	1.3	2.7	0.099	1
LUAD	PC3	1.1	3	0.3	30	0.86	0.35	1
LUAD	PC4	0.37	1.4	0.27	7.9	0.18	0.67	1
LUAD	PC5	1.2	3.2	0.55	19	1.7	0.19	1
LUAD	PC6	0.17	1.2	0.31	4.6	0.06	0.81	1
LUSC	PC1	1.5	4.3	1.3	14	5.7	0.017	0.901
LUSC	PC2	-0.77	0.46	0.078	2.8	0.71	0.4	1
LUSC	PC3	-0.52	0.59	0.096	3.7	0.31	0.58	1
LUSC	PC4	-1.1	0.32	0.06	1.7	1.8	0.18	1
LUSC	PC5	-0.56	0.57	0.063	5.1	0.25	0.61	1
LUSC	PC6	-1.7	0.18	0.018	1.9	2	0.15	1
OV	PC1	0.39	1.5	0.45	4.8	0.41	0.52	1
OV	PC2	-0.18	0.84	0.28	2.5	0.1	0.76	1
OV	PC3	0.088	1.1	0.24	4.9	0.01	0.91	1
OV	PC4	1	2.8	0.85	9.4	2.9	0.089	1
OV	PC5	0.7	2	0.35	12	0.62	0.43	1
OV	PC6	0.1	1.1	0.21	5.9	0.01	0.9	1

Cancer Type	PCs	beta	HR	lower 95 HR CI	upper 95 HR CI	wald.test	p.value	q.value
READ	PC1	1.2	3.5	0.5	24	1.6	0.21	1
READ	PC2	1.2	3.4	0.14	80	0.57	0.45	1
READ	PC3	-0.18	0.84	0.1	6.8	0.03	0.87	1
READ	PC4	1.3	3.7	0.29	47	1	0.31	1
READ	PC5	0.089	1.1	0.016	73	0	0.97	1
READ	PC6	-2.9	0.055	0.00025	12	1.1	0.29	1
SKCM	PC1	0.51	1.7	0.25	11	0.28	0.6	1
SKCM	PC2	2.5	12	0.057	2400	0.82	0.37	1
SKCM	PC3	0.24	1.3	0.068	24	0.03	0.87	1
SKCM	PC4	0.086	1.1	0.024	50	0	0.97	1
SKCM	PC5	0.4	1.5	0.051	44	0.05	0.82	1
SKCM	PC6	-0.89	0.41	0.057	3	0.78	0.38	1
BRCA	PC1	-1.5	0.22	0.028	1.7	2.2	0.14	1
BRCA	PC2	0.21	1.2	0.11	13	0.03	0.86	1
BRCA	PC3	-1.9	0.15	0.019	1.2	3.2	0.074	1
BRCA	PC4	-0.81	0.45	0.07	2.8	0.73	0.39	1
BRCA	PC5	-0.27	0.76	0.056	10	0.04	0.84	1
BRCA	PC6	-0.66	0.52	0.064	4.2	0.38	0.54	1

Table 4.3: **Survival analyses for reconstructed microbiome of all the TCGA samples.** Table of the DFS univariate Cox model analysis results of all the cancer types analysed.

4.8 Bacterial pathway associations with molecular and clinical properties

The communication between bacteria and host relies on mutual metabolic exchanges in both physiological and pathological conditions. These metabolic interactions have been shown to play a crucial role in modulating host physiology, immune response and even influencing disease progression (Mousa, Chehadeh, and Husband, 2022). Through these exchanges, the host provides a hospitable environment for the bacteria to colonise and thrive, while bacteria can produce or consume metabolites. This intricate interplay between bacterial and host metabolism has been demonstrated in various diseases, including cancer, where bacterial metabolic pathways have been shown to be dysregulated and associated with tumour growth and progression, as described in the Introduction chapter, see 2.2. Understanding the metabolic exchanges between bacteria and host can provide valuable insights into disease pathogenesis and help identify potential therapeutic targets. Thus, to investigate bacterial pathway activity in association with the tumour properties, we used HUMAnN, a tool to profile the abundance of bacterial metabolic pathways from metagenomics or transcriptomics data (Beghini et al., 2021), see Methods chapter, section 3.6. Since our approach detected substantially fewer bacterial reads than direct bacterial sequencing methods, we decided to pool the reads of COAD samples belonging to the same tumour property.

The side of the tumour showed a strong association with the reconstructed microbiome, so we quantified the differential signals of bacterial metabolic pathways by comparing the pooled left and right COAD tumours. After applying a filter to exclude low-abundance pathways, we selected only those pathways that showed a differential abundance of at least 30% and validated their differential abundance by bootstrapping subsets of samples (refer to Methods chapter, section 3.6). This revealed stronger signals for pathways of fatty acid biosynthesis on the left side of the colon, particularly in the palmitate to cis-vaccenate synthesis pathway (we found a higher abundance of (5Z)-dodecenoate biosynthesis I, palmitoleate biosynthesis I (from (5Z)-dodec-5-enoate) and cis-vaccenate biosynthesis) and stearate biosynthesis, which is consistent with the literature (Choi et al., 2019; Pickens et al., 2016; Schirmer et al., 2016; Butler et al., 2017; Akazawa et al., 2021). The tricarboxylic acid cycle (TCA) cycle was also associated with the left side of the colon, which has been previously found to be enriched in colorectal cancer-associated bacteria (Dai et al., 2018). We employed the same methodology to analyse the association of CMS and mutation burden with the reconstructed microbiomes, as they exhibited the strongest correlation with the microbiomes along with the colon side in the PCA analyses, see section 4.5.1. The CMS2, 3 and 4 samples were grouped and compared with the metabolic pathways of CMS1 samples, however, no pathways showed significant differences in abundance between the two groups. On the other hand, when comparing the metabolic pathways of high and low mutation burden samples, we observed that the high mutation burden samples displayed two subgroups with higher abundance. One subgroup was associated with DNA degradation (inosine 5'-phosphate, purine ribonucleosides, adenosine and guanosine nucleotides degradation), while the other was related to sugar metabolism (starch, D-glucarate and D-galactarate, GDP-mannose, glucose, glucose-1-phosphate and xylose degradation).

These findings were previously shown in Sambruni et al., 2023

4.9 Analyses of the IEO cohort

In section 4.4, we utilised the IEO cohort to validate our bacterial microbiome reconstruction approach by comparing its measurements to other commonly used bacterial detection approaches, as described in the Methods chapter, section 3.2. Furthermore, we demonstrated that the microbiome composition of the IEO cohort was similar to that of the COAD samples from the TCGA dataset, as evidenced by their clustering in the PCA space of all TCGA tissues. With a smaller

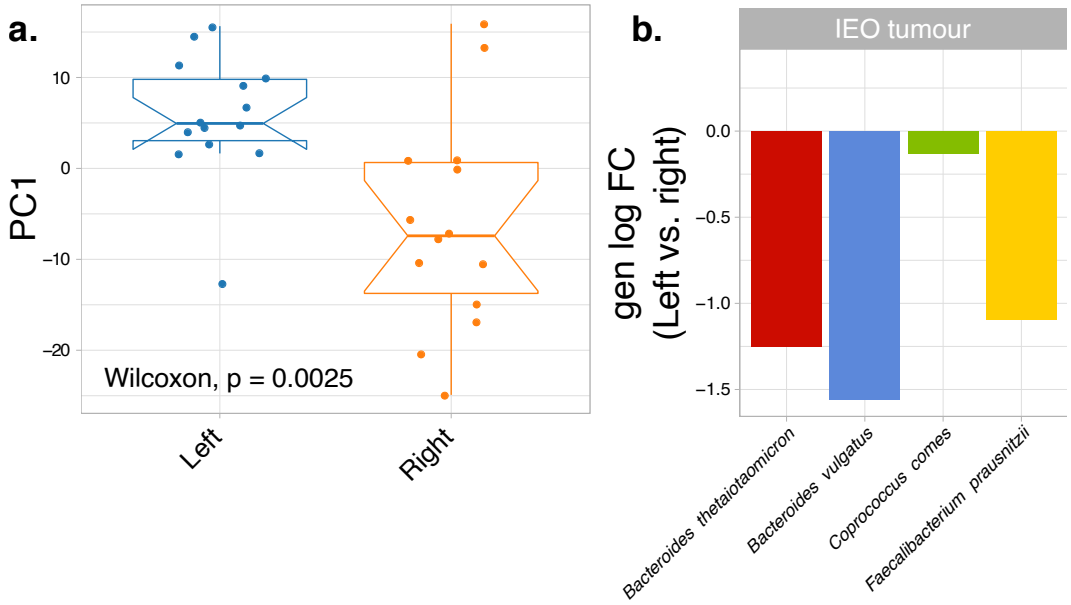


Figure 4.29: **Association of the reconstructed microbiome of IEO samples and the side of colon.** (a) Boxplot of PC1 values of PCA on IEO cohort reconstructed microbiomes by the side of tumour samples. (b) Tumour IEO cohort generalised log fold change of the scores of the bacteria detected as differentially distributed in tumour COAD samples between left and right side.

sample size and less clinical and molecular information about the tumours analysed, we sought to test the association between the reconstructed microbiome of the IEO samples and tumour properties. While we collected clinical and molecular information for the IEO patients, we limited our analysis to the side of the colon and the age of the patients due to the small number of tumour samples (29 samples). We applied the described batch effect detection approach to the microbiomes of the IEO tumour samples and corrected for the main batch effect, which was associated with the sequencing runs, as already described in section 4.4.1. We then employed PCA and investigated the associations between the first six PCs and the side of the colon and age of patients. Our analysis revealed a statistically significant association between the side of the colon and the microbiome composition of PC1 ($p = 0.0025$; Wilcoxon test; Figure 4.29a), while a weaker association was observed for the age and PC2 ($p = 0.067$; Spearman correlation test). Interestingly, we found a significant overlap between the species contributing the most to PC2 in

the TCGA COAD PCA and PC1 in the IEO cohort PCA (46 species; $p = 0.00002$; Fisher exact test). Both of these PCs were associated with the side of the colon. Although we tested for specific species associated with the left or right side of the colon, we were unable to find any significant associations, possibly due to the limited number of samples analysed. However, the species that were detected as significantly associated with the right side of the colon in COAD samples showed a similar trend in the IEO cohort, as shown in Figure 4.29b.

Taken together, these results indicate a strong agreement between the findings of the TCGA and IEO cohorts.

4.10 Controls

4.10.1 Microbial read extraction tool (Kraken)

In previous studies, various methods were employed to extract microbial reads from NGS data. For instance, in their paper analysing the entire TCGA for microbial reads extraction, Poore et al., 2020, utilised Kraken (Wood and Salzberg, 2014), an ultra-fast taxonomic sequence classifier that assigns taxonomic labels to DNA sequences. This tool differs from PathSeq in two key aspects: it utilises a k-mer approach to align reads to the microbial reference genomes and assigns them to the lowest common ancestor, typically at the genus level. To benchmark the classification obtained using PathSeq, we applied Kraken2 (Wood, Lu, and Langmead, 2019) with default parameters with its Standard reference genome, which comprehends Refeq genomes of archaea, bacteria, viral, plasmid, human and UniVec_Cor, to the IEO cohort.

We utilised the Kraken report files, which provide a summary of the percentage and number

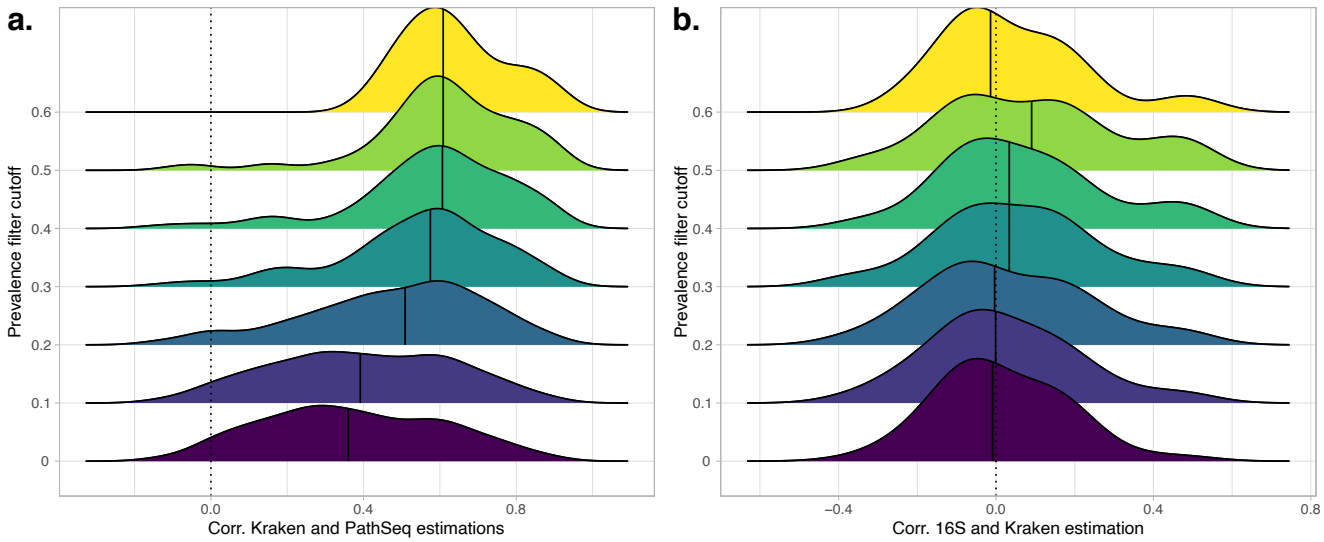


Figure 4.30: **Comparison of Kraken estimation.** Ridge plot of the Spearman's coefficients of Kraken bacterial estimations and (a) PathSeq reconstructed microbiome (bacterial species level) and (b) 16S bacteria genera quantification correlations. Different filter cutoffs for bacterial presence are shown. Black vertical lines represent the median of the distribution.

of fragments assigned to each taxon, to calculate the percentage of bacterial species identified in each sample based on their directly assigned fragments. Initially, we compared the results obtained from PathSeq and Kraken. Interestingly, Kraken detected a greater number of bacterial species compared to PathSeq (4494 species detected by Kraken vs. 2715 species detected by PathSeq in IEO cohort). Among these, we selected 668 species that were common to both methods and observed a high level of agreement. The agreement improved further when we specifically selected species that were prevalent in both Kraken and PathSeq estimations, as depicted in Figure 4.30a.

After that, we conducted a comparison between Kraken's reconstructed microbiomes and the

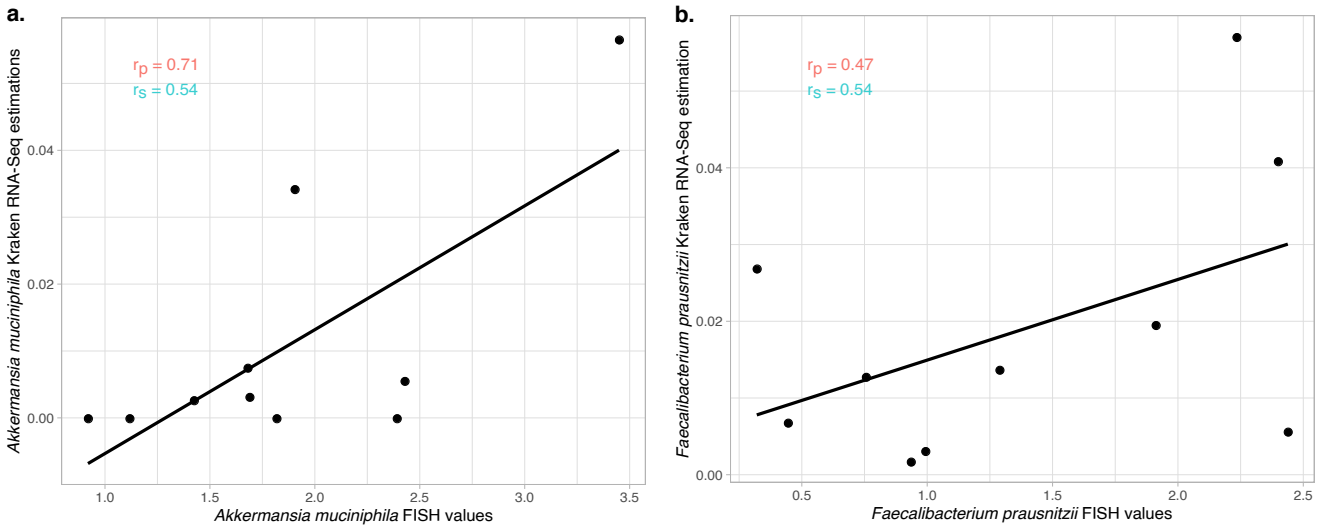


Figure 4.31: **Comparison of FISH and Kraken RNA-Seq estimations.** Correlation of the Kraken RNA-Seq bacterial abundances and FISH quantification of (a) *A. muciniphila* and (b) *F. prausnitzii*. Pearson (r_p) and Spearman (r_p) coefficients are indicated

results obtained from 16S and FISH techniques. Similar to our previous analysis with PathSeq described in section 4.2.1, we evaluated the performance of Kraken in estimating bacterial genera employing increasing prevalence cutoffs, comparing Kraken estimations with 16S. However, as illustrated in Figure 4.30b, unlike for the comparison between PathSeq- and 16S-detected genera, we did not observe a clear trend with respect to higher cutoffs on the prevalence. Notably, at the cutoff of 0.6, the median of the correlation values shifted towards the negative direction.

Unlike the 16S comparison, we observed a strong correlation between Kraken's quantification of *A. muciniphila* and *F. prausnitzii* and the corresponding values obtained from FISH analysis (both Spearman's $r_s = 0.54$), consistent with the results obtained using PathSeq (Spearman's $r_s = 0.52$ and $r_s = 0.21$, respectively), see Figure 4.31. However, similar to PathSeq, Kraken also failed to detect *F. nucleatum*, as shown in Figure 4.32.

In general, Kraken identifies a greater number of bacterial species but when, comparing to PathSeq and 16S, PathSeq demonstrates slightly better performance. Therefore, we consider PathSeq to be the preferable option.

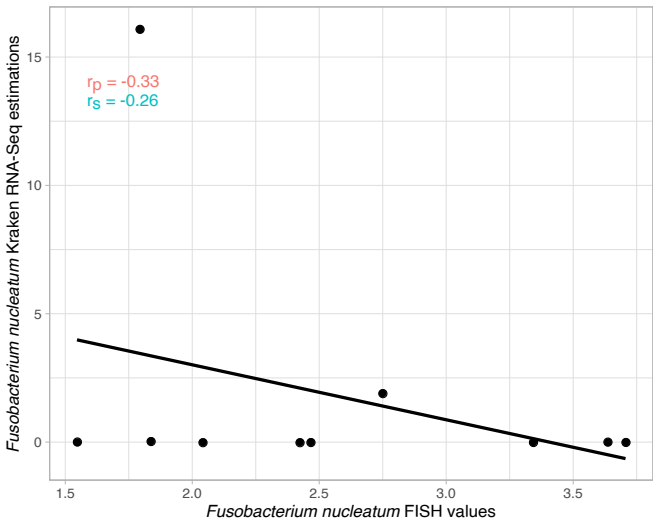


Figure 4.32: **Comparison of FISH and Kraken RNA-Seq estimations.** Correlation of the Kraken RNA-Seq bacterial abundances and FISH quantification of *F. nucleatum*. Pearson (r_p) and Spearman (r_p) coefficients are indicated

4.10.2 Stringent filters: subset of species

The reconstructed microbiome from human RNA-Seq data is affected by several technical biases that can occur at various steps of the analysis, as described in 3.4. On the other hand, this technique allows for the measurement of the entire bacterial composition in each sample. To ensure the reliability of our findings and eliminate the possibility of technical biases affecting our observations, we conducted a rigorous analysis of the associations between clinical properties and a limited set of highly confident bacterial species in COAD samples, the cancer type showing the most associations between bacteria and tumour properties. To achieve this, we applied stringent filtering of the bacteria detected in COAD samples based on prevalence, cancer type specificity and species that co-vary with the strongest technical property of the samples, i.e. the plate ID, as described in the Methods chapter, section 3.4.6. We quantified only 44 species with high confidence, the majority of which have been isolated from the human gut or body, as shown in Table 4.4.

We then applied the final steps of our approach and obtained the PCA on this subset of reconstructed bacterial microbiome. Even with this small cohort of species, we were able to confirm the strongest associations of microbial composition with clinical properties detected in section 4.5.1, including side, MSI and aneuploidy, as shown in Figure 4.33. This control suggests that while some associations are driven by a subset of high-confidence bacteria, we were able to detect other associations only by considering the entire bacterial composition of tumour samples.

This results were shown in Sambruni et al., 2023.

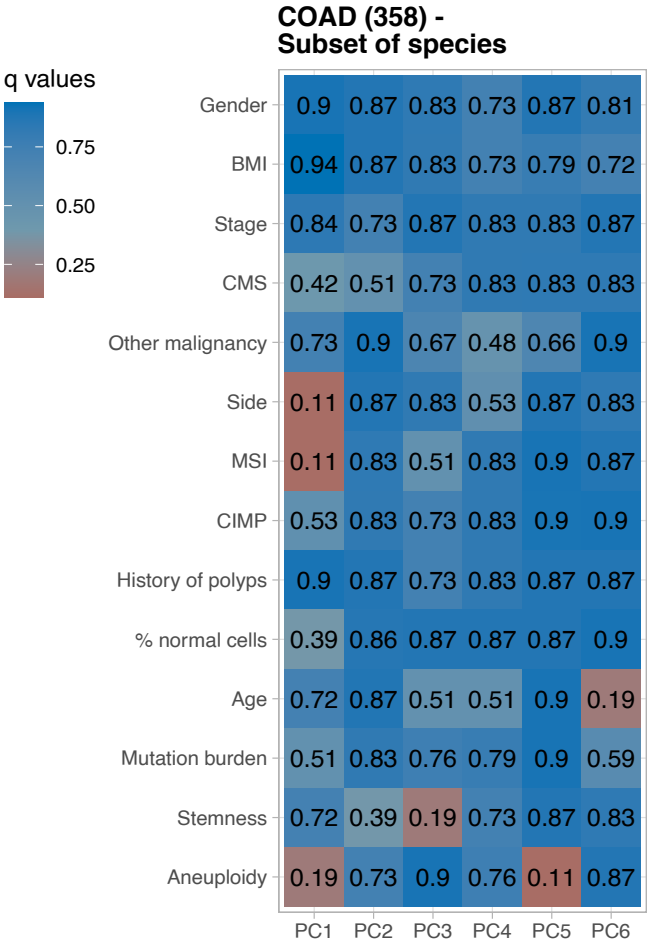


Figure 4.33: **Heatmaps of the associations between the high confident subset of bacteria reconstructed microbiome of COAD and clinical and molecular properties.** Heatmaps of the q values of the associations and correlation between the first six PCs of the PCA on a strict subset of abundant, tissue-specific and reduced in technical variability species from the reconstructed microbiome of the associations between clinical tumoural properties of COAD primary tumours. Number of samples analysed in brackets. Adapted from Sambruni et al., 2023

tax id	bacterial name
40545	*Sutterella wadsworthensis*
214856	*Alistipes finegoldii*
328814	*Alistipes shahii*
674529	*Bacteroides faecis*
333367	*Clostridium asparagiforme*
437898	*Sutterella parvirubra*
74426	*Collinsella aerofaciens*
1531	*Clostridium clostridioforme*
239935	*Akkermansia muciniphila*
901	*Desulfovibrio piger*
1892897	*Shigella sp. FC569*
68259	*Streptomyces purpurogeneiscleroticus*
1450439	*Bacteroides sp. UW*
585543	*Bacteroides sp. D20*
1581131	*Actinomyces sp. HMSC08A01*
469610	*Burkholderiales bacterium 1_1_ 47*
1207055	*Sphingobium sp. C100*
712117	*Actinomyces sp. oral taxon 170*
1807	*Mycobacterium obuense*
1768781	*Sphingobium sp. CCH11-B1*
13690	*Sphingobium yanoikuyae*
1813946	*Acidovorax sp. GW101-3H11*
1739435	*Fusobacterium sp. HMSC065F01*
1454008	*Aeromonas sp. HZM*
818	*Bacteroides thetaiotaomicron*
1030157	*Sphingomonas sp. KC8*
1310601	*Acinetobacter sp. 479375*
817	*Bacteroides fragilis*
204516	*Bacteroides massiliensis*
291644	*Bacteroides salyersiae*
100886	*Catenibacterium mitsuokai*
1852365	*Fusobacterium massiliense*
293	*Brevundimonas diminuta*
457392	*Bacteroides sp. 3_2_5*
69823	*Selenomonas sputigena*
39488	*Eubacterium hallii*
1032505	*Fusobacterium sp. OBRC1*
1217710	*Acinetobacter sp. NIPH 899*
407152	*Ochrobactrum cytisi*
341694	*Peptostreptococcus stomatis*
823	*Parabacteroides distasonis*
204	*Campylobacter showae*
936563	*Fusobacterium sp. CM22*
144185	*Leifsonia aquatica*

Table 4.4: **Highly confident bacterial species in COAD samples.** Selected bacteria species after the application of three filters to remove the low-present, batch-affected bacteria and select the cancer type-specific ones.

4.10.3 Bacterial quantification comparison (unambiguous reads vs ambiguous reads vs scores)

Given that the estimation of bacterial signals can be influenced by the handling of reads mapping to sequence-redundant regions of bacterial genomes, we sought to investigate whether alternative approaches to estimating bacterial abundance could have an impact on our findings. Our aim was to assess if these alternative methods would reveal associations that were not previously identified using the approach described in steps 1 and 2, section 3.4.1. Therefore, we conducted a comparative analysis of COAD bacterial abundance estimation techniques to evaluate their ability to identify novel associations between bacterial composition and tumour properties. We applied the workflow described in section 3.4.1 to the uniquely mapping reads per species (i.e. unambiguous reads) or the sum of all mapping reads, including also those that mapped to redundant regions (i.e. ambiguous reads). We corrected for the strongest batch effect (i.e. the sequencing plate, as the other bacterial abundance estimation approach reported before) and tested for significant associations between tumour properties and the first six PCs of the reconstructed microbiome PCA. All approaches identified side, MSI, CIMP and aneuploidy status as associated with bacterial compositions (see Figure 4.34). However, the relaxed approach (including ambiguous reads) also detected an association with the percentage of normal cells that was not identified before. These results are reported in Sambruni et al., 2023.

4.10.4 Comparison with other dimensionality reduction approaches

Considering the number of bacterial species we detected in the samples and our objective of examining the overall bacterial composition rather than individual species, we employed a dimensional reduction technique to reduce the dimensionality of the data, capture the general patterns and identify similarities or differences between sample compositions. To achieve this, we initially applied PCA, which is one of the earliest and most well-known dimensionality reduction approaches. We selected PCA due to its ease of application and interpretation. However, we also explored other approaches, namely PCoA and nMDS. These methods are widely used in the field of metagenomics, as PCoA and nMDS are commonly utilised to represent bacterial signals.

We applied these approaches to the reconstructed microbiome of COAD bacteria and focused on the first six dimensions. In each approach, we addressed the most prominent batch effect and once again identified the plate ID as the primary source of technical bias across all approaches. Similar to the PCA, the second strongest technical effect was attributed to the read length used

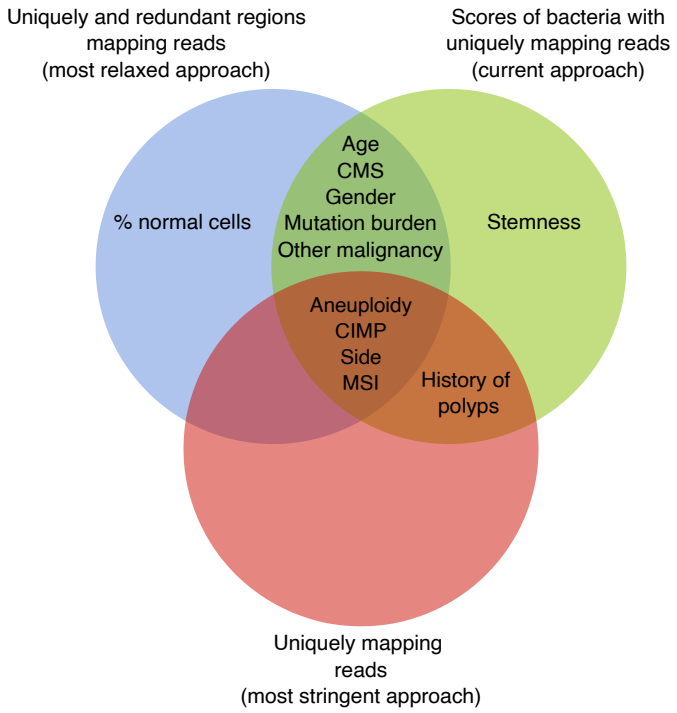


Figure 4.34: **Tumour properties associated with bacterial microbiome reconstructed with different approaches in COAD samples.** Venn diagram of the tumour properties associated with any of the first six PCs of the PCA on different bacterial quantification approaches, namely uniquely mapping reads (unambiguous reads), uniquely and redundant region mapping reads (ambiguous reads) and scores of uniquely mapping reads (unambiguous score, current approach) in COAD samples. Adapted from Sambruni et al., 2023

for sequencing the samples. After applying batch correction to the COAD samples' bacterial composition based on plate ID, we examined the associations with tumour properties. Notably, nMDS exhibited effective batch correction for both plate ID and read length, which were the two most significant batch effects observed. However, in the first two PCoA dimensions, clusters related to read length were still visible, as depicted in Figure 4.35a-b. When exploring the associations between these dimensions and the tumour properties of COAD samples, we observed that both techniques detected associations with several key properties that were identified before using PCA, such as CMS, side, MSI and CIMP. It is relevant to note that the first two PCoA dimensions were not significantly associated with properties, except for a few ones like history of polyps, % of normal cells and age. Figure 4.35a, which displays these two dimensions, shows samples clustering by plate ID and read length, suggesting that PCoA was still able to identify residual batch effects even after the correction. However, these dimensions, which were sensitive to batch effects, did not show any significant associations with relevant tumour properties.

The outcomes obtained from nMDS were similar to those obtained from PCA. The second dimension of both methods captured the most significant biological association between bacteria and tumour properties. As outlined in section3.4.3 from Methods chapter, there are several technical distinctions between PCA and nMDS. However, due to the advantage of the ability to identify the

bacteria with the greatest impact on each component associated with tumour properties, PCA appears to be a more favourable approach than nMDS.

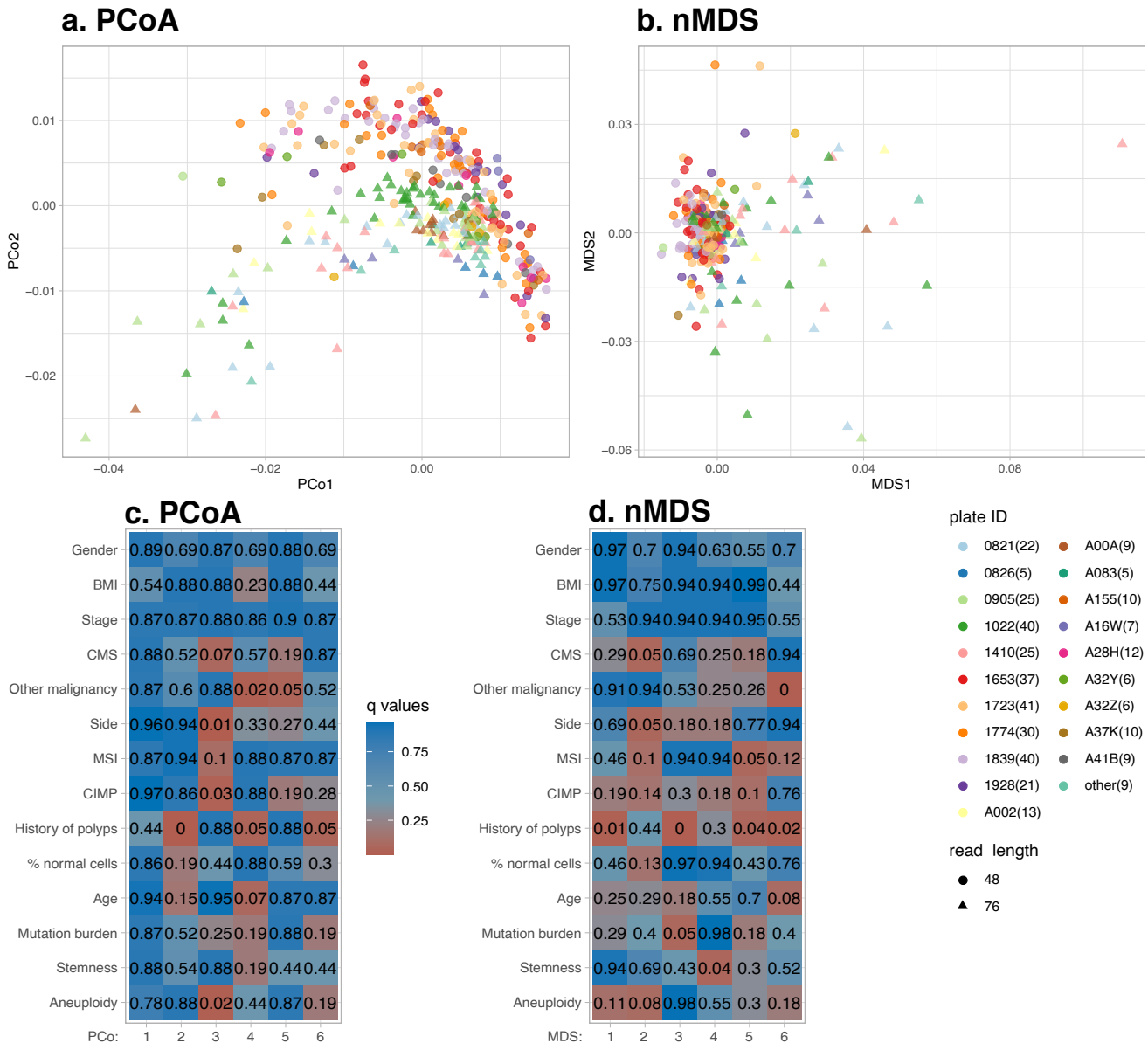


Figure 4.35: **Bacterial composition reconstruction with other dimensionality reduction approaches.** Bacterial composition reconstruction of COAD samples using (a) PCoA or (b) nMDS scatterplots after plate ID batch effect correction, showing plate ID in colour and the read length in symbol. Tumour property association of the dimensions of (c) PCoA and (d) nMDS of the same samples.

Discussion

The focus of cancer research has been to differentiate between abnormal and normal cells, with the aim of understanding tumour-specific characteristics to identify potential biomarkers or therapeutic targets. Although cancer is a term that encompasses a large group of diseases, which share the common feature of uncontrolled cell growth, its causes, progression and properties vary widely depending on the tissue where it originates and other factors.

The understanding of the immune cells' role in cancer has led to a new perspective on studying the disease, which involves examining the tumour microenvironment - a distinct space that promotes the growth and development of cancer. In this microenvironment, various players such as the host cells, microbes and other factors contribute to the disease in different ways that are dependent on the tissue and cancer type. This approach goes beyond the study of tumour cells alone and extends to other players that either support or are supported by the disease. Therefore, targeting or detecting the presence or type of the disease can involve considering these additional contributors.

Our study adopts the aforementioned approach and seeks to characterise the tumour microenvironment as an ecological niche, where cellular species are interdependently connected. More specifically, we aim to examine the unique peculiarities of the microbes residing in the tumour microenvironment and explore the links between the composition of these microbes and the properties of the tumour, as well as other factors that contribute to the formation of the tumour microenvironment. In our project, whose main results are published in Sambruni et al., 2023 and other additional analysis and methodological choices are reported here, we used PathSeq to reconstruct the microbial composition of nine cancer types from the TCGA database. We also corrected the major batch effect affecting bacterial composition and tested the association between tumour properties and bacterial composition in a methodological manner, identifying the species associated with these properties. Furthermore, we identified the bacterial composition and specific species linked to bad prognosis in colon cancer patients. We developed several approaches to validate our approach, such as the comparison with other experimental approaches in our internal cohort.

5.1 Technical issues

5.1.1 Bacterial heterogeneity

This intriguing question poses several challenges, with one of the most difficult being microbial heterogeneity. As outlined in the Introduction chapter, various human body surfaces, from the skin to the gastrointestinal and respiratory tracts, are inhabited by microbes that vary in composition from person to person to such an extent that they can serve as a unique identifier or fingerprint for each person (Mousa, Chehadeh, and Husband, 2022). Furthermore, the microbial composition appears to be linked to the host and is influenced by factors such as lifestyle, food consumption and age (Wastyk et al., 2021; Mousa, Chehadeh, and Husband, 2022; Yadav and Chauhan, 2021).

The significant variability in microbial composition presents a technical challenge in that numerous samples are needed to identify statistically significant similarities between the tumour microbes of different hosts. Unfortunately, microbial studies typically involve fewer patients than required and study biases can heavily influence microbial detection. This can make it difficult to differentiate between technical biases and biologically relevant variation. Additionally, few studies report both the characterisation of the microbiome and the analysis of tumour properties. In the case of colon cancer, the microbial composition is often detected using faecal samples, which may serve as a proxy for colon dysbiosis but do not accurately represent the composition of the tumour microenvironment, as it contains bacteria from the entire digestive tract and is not localised.

A recent approach has effectively addressed these challenges by utilising human NGS data to describe tumour properties and serve as a source of microbial signal. This involves using large databases of NGS data that have already been extensively studied to characterise molecular properties of samples and collect clinical information of patients. By doing so, it becomes possible to reconstruct the microbial composition of the samples. One of the pioneering studies to apply this approach to a large database was Poore et al., 2020, where the authors extracted bacterial signals from TCGA, one of the largest databases comprising over 30 human cancer types and more than 10,000 patients. They employed the Kraken tool (Wood and Salzberg, 2014) to assign reads to microbes at the genus level, normalised the estimations to reduce technical biases and used the bacterial composition of blood samples to predict the tissue of origin of cancer. Another influential study, Dohlman et al., 2021, created a bacterial species atlas of oropharyngeal, esophageal, gastrointestinal and colorectal cancer types using the PathSeq tool as well (Dohlman et al., 2021), which underwent multiple decontamination steps, including comparing it with the

reconstructed microbiomes of blood samples to remove highly prevalent contaminants from tissue-specific bacteria. This atlas was then used to identify prognostic bacterial species and bacterial signatures in blood for mucosal barrier injuries. These breakthrough studies paved the way for further investigations into the association between bacterial composition and tumour properties, but none of these studies systematically analysed the link between bacteria and a comprehensive set of clinical and molecular properties.

5.1.2 Previous literature comparison

Poore et al., 2020 employed Kraken, an ultra-fast aligner, along with the RepoPhlan microbiome collection (Q. Zhu et al., 2019) to analyse the unaligned human reads from both the RNA-Seq and WGS. They applied the Kraken algorithm to provide a taxonomy assignment at the genus level. In our view, a lower taxonomic level gives a more accurate and relevant depiction of the tumour microbiome environment. This becomes particularly apparent when considering research papers that argue even species-level resolution may be sometimes insufficient: for instance, *B. fragilis*, where the presence of a specific enterotoxin distinguishes different strains and influences their carcinogenic potential (W. T. Cheng, Kantilal, and Davamani, 2020). To partially address this limitation, we opted for alternative approaches that allowed us to quantify the abundance of bacterial species. Although we recognise that this choice still has its limitations, as exemplified by the case of *B. fragilis*, we believe that opting for a lower taxonomic specificity, such as the species level, is a more favourable choice for studying the tumour microenvironment. This is particularly relevant considering the nature of the data, which allows for greater precision in quantifying bacteria.

Moreover, Poore et al., 2020 extensively discuss contamination and decontamination strategies, acknowledging the absence of a gold standard for decontamination. They propose three approaches with varying levels of stringency and the choice among them depends on the specific hypothesis being investigated. Furthermore, they employed the Voom algorithm (Law et al., 2014) and supervised normalisation methods on log-count data, assuming normal distribution. To apply this approach, the authors incorporated sample type as a biological variable that needs to be preserved and considered several technical covariates, such as sequencing center, sequencing platform, experimental strategy, tissue source site and Formalin-Fixed Paraffin-Embedded (FFPE) status. Both Voom and ComBat, the chosen tool for our analysis, have extensive support in the RNA-Seq literature, making it challenging to favour one over the other. However, we consider our approach more conservative: we applied batch correction exclusively to the RNA-Seq samples within the same cancer type and sample type, excluding FFPE samples (sample selection criteria

are detailed in section 3.1). This strategy enabled us to analyse a subset of samples with fewer technical confounders involved.

The second important study that examined TCGA data to extract microbiome information was published by Dohlman et al., 2021. This study specifically focused on oropharyngeal, oesophagal, gastrointestinal and colorectal samples analysed using WGS and WXS. The authors of this paper opted for Pathseq (Walker et al., 2018) as the microbial-extraction tool, enabling them to delve into bacterial species level, although the lowest available data from the online atlas was provided at the genus one. In their methodology, the authors selected uniquely mapping reads, the unambiguous ones, for estimating bacterial abundance. By adopting this approach, they significantly reduced the bacterial signal identified within the samples.

Similar to Poore et al., 2020, the authors extensively discussed contamination-related concerns and the associated procedures for its removal, although they ultimately adopted a distinct approach. Their approach involves comparing the prevalence of bacteria in tumour samples to that in blood samples, employing a minimum read assignment filter. They identified bacteria enriched in tumour samples, the tumour-resident species, created a blacklist of contaminants that should be eliminated from the data and supported their findings with several tests. While the authors developed their method based on WGS reconstructed microbiome and extended it to WXS for microbiome reconstruction decontamination, we believe that the differences between these experiments (and so even with RNA-Seq one) could cover other experiment-specific contaminants. The strong correlation observed between WGS and WXS, as depicted in Figure 4.10, validates the applicability of the selection method developed by Dohlman et al., 2021 directly to WXS data. However, this raises concerns about the suitability of applying the same approach to microbiome reconstruction from RNA-Seq data, as the correlation between RNA-Seq and WGS was good but not as strong. Moreover, the selection of WGS tumour-resident bacteria could result in an underestimation of biologically interesting bacteria that are not detected in WGS, potentially leading to their removal.

An additional noteworthy aspect of this study is the exclusive reliance on contaminant removal as the sole strategy for addressing technical biases, without incorporating any batch correction approaches. In our analysis, we observed the influence of plate ID on the bacterial reconstructed microbiome in both WXS and WGS samples (section 4.2.3) but we also illustrated that by stringently selecting bacteria by prevalence, batch-affection and tissue-specificity, we were able to build a bacterial composition that still captured certain associations identified in the complete microbiome (section 4.10.2). Hence, although we recognise the potential advantages of addressing plate ID contamination, we cannot disregard the possibility that the approach suggested by Dohlman

et al., 2021 represents the most optimal method for contaminant removal.

5.1.3 Microbiome reconstruction

When applying PathSeq to extract bacterial reads from human RNA-Seq data, we encountered an initial challenge concerning the choice of bacterial quantification that would best represent the microbial data. As mentioned earlier in section 3.4.1, PathSeq addresses the issue of bacterial genome redundancy and provides three quantification options: uniquely mapping reads (also called unambiguous reads, reads mapping to a single genome), ambiguous reads (comprising both uniquely mapping reads and those mapping to multiple genomes) and a middle-ground approach known as "score", which counts the uniquely mapping reads and assigns a weighted quantification to the multi-mapping reads based on the number of genomes they map to, distributing their weight accordingly. While utilising uniquely mapping reads resulted in the exclusion of a significant number of reads (as evident from the previous results, see section 4.1.4, many reads mapped to rRNA genes, which are highly conserved in bacteria), which is already scarce in human RNA-Seq experiments, the use of multi-mapping reads introduced excessive noise, leading to an inflation in the total count of bacterial reads. The score quantification approach, on the other hand, could detect numerous bacteria that share common genomic regions with the in-sample bacteria, thereby exacerbating the challenge of identifying trustworthy bacteria. After careful consideration, we opted to use the "unambiguous score" quantification as it represents a more balanced approach compared to the other options. This approach successfully retrieved a considerable number of reads that were previously disregarded by Dohlman et al., 2021, which used uniquely mapping reads.

To validate the efficacy of our quantification method, we performed a control experiment to examine the differences in the association between the bacterial composition of COAD samples and tumour properties when considering either the ambiguous reads or the unambiguous reads, mentioned in section 4.10.3. Notably, all approaches, including our metric, detected strong associations between bacterial composition and tumour properties. However, our metric stood out by detecting the highest number of associations among the three approaches. This finding further supports the validity and robustness of our chosen quantification method. The ability to uncover a greater number of associations suggests that our approach provides a more comprehensive and nuanced understanding of the relationship between bacterial composition and tumour properties. Furthermore, when examining the bacterial signal extracted from human NGS experiments, another commonly used tool, Kraken2 (Wood, Lu, and Langmead, 2019), introduced a different approach. Unlike assigning a weight to a sequence with multiple mappings, Kraken2 simply assigns such reads to a higher taxonomic level on the classification tree. Consequently, ambiguous

reads are not assigned to lower taxonomic levels, such as species and strains but rather to the genus or higher levels. As discussed earlier, this limitation is undesirable considering the opportunity to achieve greater specificity. Nonetheless, when bacterial species were detected in both PathSeq and Kraken2 for comparison, they exhibited similar quantification. Indeed, the estimations of *A. muciniphila* and *F. prausnitzii* by Kraken2 showed a correlation with FISH estimations similar to that of PathSeq. Strangely, Kraken2 appeared to be less precise when assessing the genus level. In fact, Kraken2's estimations displayed a weaker correlation with 16S bacterial quantification compared to PathSeq, see section 4.10.1. Although unexpected, this final outcome convinced us to proceed with PathSeq for our analysis. Furthermore, although Kraken2 exhibited a runtime advantage of one-third compared to PathSeq, the computational capabilities and requirements of both tools did not influence our decision. Specifically, we ran PathSeq with 8-16 CPUs, depending on the cluster availability, while Kraken2 used 8 CPUs. However, it should be noted that Kraken2 generally demonstrated a higher demand for memory compared to PathSeq.

As explained above, our research primarily focused on analysing the comprehensive bacterial composition within the tumour microenvironment and utilising specific bacteria as a means to validate the presence of relevant gut-resident and colon cancer-related bacteria. It is important to acknowledge that our data is susceptible to various technical issues such as contamination, the detection of bacteria originating from external sources such as the human breath and skin, air and reagents (Salter et al., 2014) or read misclassification, low-quality human reads or technical errors resulted in the misassignment of nonsensical bacteria. To address these issues, we employed a bacterial composition-based approach to select the bacteria included in our analysis, summarised in section 3.4. This data-driven approach aimed to identify the most variable bacteria while excluding those with sparse abundance to mitigate the impact of contamination and read misclassification on our results. This simple approach resulted in a very effective one since, despite implementing the previously described procedure for bacterial quantification, our bacterial microbiome reconstruction revealed a large number of bacterial species, which resulted in noise that hindered the effective analysis of our data.

This issue of misclassification was also stressed in a recent preprint paper by Gihawi, Cooper, and S, 2023. The authors highlighted various problems with the work of Poore et al., 2020, including the misclassification issue. In their manuscript, Gihawi, Cooper, and S, 2023 proposed additional steps for filtering human reads before data modelling, as implemented in the study by Dohlman et al., 2021. These suggestions aimed to address the challenges posed by misclassification and enhance the accuracy of the analysis. It is evident that the issue of misclassification is a significant concern within the field, as acknowledged by the aforementioned studies. In our

opinion, additional filtering steps could be implemented prior to bacterial microbiome reconstruction to enhance the identification of in-sample bacteria. These steps could be based on both data-driven considerations and biological knowledge, such as screening the results for well-known tissue-resident bacteria. However, as highlighted in the response provided in the preprint by G. Sepich-Poore et al., 2023, incorporating such filtering steps is unlikely to significantly alter the results obtained through our less stringent filtering approach. In fact, to address our concerns about misclassification, we specifically selected a small subset of bacteria that displayed minimal sensitivity to the batch effect associated with plate ID, exhibited high prevalence across samples and exhibited tissue-specific patterns distinct from other examined tissues, see section 4.10.2. Through this approach, we employed a data-driven strategy that eliminated batch-affected bacteria (even if their potential biological significance), prioritised tissue-specific species and selected the most prevalent ones. Despite the relatively small number of bacterial species considered, we successfully identified associations with key tumour properties, previously associated with the wider bacterial reconstruction composition. This finding not only suggests alternative options for bacterial selection and investigation but also validates the results obtained using the wider bacterial composition analysis to detect novel associations. The majority of the detected species were well-known gut-related species, although some of them lacked a clear explanation. While it is plausible that some of these species may indeed have biological relevance as colon-resident or cancer-related bacteria, it is important to acknowledge the possibility of residual noise, contamination, or technical issues that may still influence the results and cannot be entirely eliminated.

As already explained, the key focus of our approach is to investigate the overall trend in bacterial composition within the tumour microenvironment, rather than solely examining specific bacteria of interest. To achieve this, we employed various dimensionality reduction techniques, namely PCA, PCoA and nMDS to reconstruct the bacterial microbiome of COAD. These approaches differ in their ability to highlight different aspects of the data: PCA emphasises the similarities between bacterial compositions across samples, while PCoA and nMDS utilise dissimilarity matrices to underscore their differences. These methods are commonly utilised across various research domains, with PCoA and nMDS being particularly prevalent in bacterial studies. Despite the inherent differences among these approaches, when applied to the raw reconstructed microbiomes, all three methods identified the plate ID of TCGA as the primary source of contamination, significantly impacting the relative abundances observed in the samples. One key difference between the dimensionality reduction approaches we used is their sensitivity to technical batches, particularly in the case of PCoA when applied to corrected bacterial abundances. This sensitivity affected the association of the reconstructed bacterial composition with tumour properties, which was not

detected in the corresponding PCos. This finding prompted us to explore a different approach to reconstruct the bacterial microbiome but it also confirmed that the associations between tumour properties and the reconstructed microbiome were not driven by residual technical bias that could not be removed by our correction. In fact, the components that did not show a clear association with batches were still able to detect associations with tumour properties, which were also detected by the other approaches. This highlights the availability of multiple approaches, yet in our perspective, PCA emerged as the optimal method. This choice is influenced by factors such as the sensitivity to the batch effect and the ability to assess the bacteria that contribute the most to each PC. Indeed, the bacteria that exhibited the most substantial impact on the PC linked to tumour properties were once again identified when examining the distribution of species across various tumour properties. These results further confirm the reliability of the association of these bacterial with tumour properties.

5.1.4 Batch effect correction

As explained in the previous section, we reconstructed the bacterial composition of nine different cancer types RNA-Seq data that were available from TCGA. In order to evaluate the overall trend of bacterial compositions, we performed a PCA on the estimated abundance of the most variable bacterial species. This allowed us to illustrate the general behaviour of the bacteria and identify similarities between samples that contained the same bacterial species in their microenvironment. By considering the bacterial composition as a whole, we can overcome the limitations of relying on the estimation of a single species that may be influenced by technical issues and noise. Moreover, analysing the bacterial composition allows us to detect and correct other technical biases, as suggested by Gihawi, Cooper, and S, 2023 when considering the entire bacterial composition of the samples. The authors express doubts about the feasibility of using single-species batch-corrected estimations and we generally agree with their perspective. In order to highlight some of the most strongly associated bacteria with tumour properties, we applied statistical tests that are capable of controlling for the batch effect on raw values. However, it is worth noting that the Kaplan-Meier test does not correct for technical biases. Consequently, we opted to divide the samples into high- or low-level abundance of the bacterium based on batch-corrected data. Although we acknowledge the concern raised by Gihawi, Cooper, and S, 2023 regarding the potential conversion of zeros to non-zero values after the correction, which can lead to misinterpretation and overestimation of single bacteria values, we still believe that our approach is the most suitable for the analysis we intended to perform. In their paper, Gihawi, Cooper, and S, 2023, highlights an example from

Poore et al., 2020, where a false taxonomic assignment was made for a very low-abundance microbe. However, in our analysis, we specifically selected the bacteria that have the greatest impact on the principal component associated with DFS, which were already chosen from the pool of the most variable bacteria. Hence, the concerns raised by Gihawi, Cooper, and S, 2023 regarding false taxonomic assignments do not apply to our analysis.

In response to the concerns raised by Gihawi, Cooper, and S, 2023, G. Sepich-Poore et al., 2023 argue that they were able to demonstrate their results even without applying their batch correction step. In our study, however, the batch correction approach played a crucial role in uncovering the biologically meaningful associations between bacteria and tumour samples. Nevertheless, as discussed in the previous section, we can employ several bacteria selection approaches to avoid batch correction, e.g. the more rigorous one previously explained in section 4.10.2 that also revealed significant associations between the reconstructed microbiome of COAD and tumour properties. These findings support the conclusion put forth by G. Sepich-Poore et al., 2023, that different approaches can be employed depending on the specific hypotheses being tested.

Our approach also allowed us to pinpoint the plate ID as the strongest source of contamination in the TCGA database, as explained in section 4.4.1. Likewise, in the IEO cohort, the sequencing run was found to be the strongest factor causing batch effects (see section 4.4.1). The similarity between the TCGA plate ID and the IEO sequencing run, both identified as the main sources of technical biases, confirms the effectiveness of accounting for the sequencing as a significant source of bias in this type of analysis. However, the specific cause of this effect remains unclear: it could be attributed to various factors such as laboratory staff handling (as human data does not directly involve manipulation of bacterial source material and may be handled without strict precautions), reagent contamination (which has been previously documented in the literature, see Salter et al., 2014), or other sources. In order to gain a more comprehensive understanding of the issue, it is necessary to analyse additional cohorts using the same workflow and conduct further investigations to identify the potential vulnerabilities within the sequencing process. For instance, it would be intriguing to compare the reconstructed bacterial microbiome of the samples with the microbes found in the experimental technical negative controls used during the sequencing step. However, this comparison was not feasible as neither the TCGA dataset nor our IEO cohort provided the negative blank reagent tube analysis. However, while it is expected to observe sequencing batch-related variability in the sequencing process for human RNA-Seq experiments (Leek, Scharpf, et al., 2010), it was unexpected to encounter it when dealing with bacteria microbiome reconstruction. Other prominent technical biases, such as the hospital of origin of the patients, would intuitively have a more pronounced impact on the results. As described in the results section 4.4.1

regarding plate ID and read length, it is likely that several of these technical bias sources overlap. Notably, not only are read length and plate ID associated with each other but we detected other technical features associated with them. In fact, the majority of the COAD samples sequenced with a read length of 76bp were collected from the same hospital, which we initially identified as the strongest batch effect. By applying our data correction technique for the sequencing step, we effectively mitigated the impact of this bias as well as other minor biases that may be associated with it. This suggests that other technical biases could potentially be corrected along with the major one, as demonstrated for read length. However, we cannot role out that all of the approaches employed eliminated or reduced all of the technical issues and contamination present in the samples. Nonetheless, we were able to eliminate the most prominent biases, allowing us to capture at least some of the biological differences in the reconstructed microbiomes of the samples.

5.1.5 Technical validation

The approach of extracting bacterial reads from human NGS data is relatively new and although other researchers have used this method to detect bacteria, we opted to validate it using the IEO internal cohort. In our perspective, this represents a crucial aspect as our study offers a direct comparison the estimations from diverse methodologies of the bacterial population within the tumour microenvironment. This is similar to the control conducted by Dohlman et al., 2021, although our research encompasses a larger sample size, which gave us the opportunity to apply tests to measure the correlations between the different approaches employed.

We applied two of the most commonly used methods to detect bacteria, such as FISH and 16S, to the IEO cohort and then compare the bacterial estimation with our RNA-Seq bacterial reconstructed microbiome, see section 4.2. The comparison of both methods showed that our approach was detecting certain bacteria better than others. For example, when comparing FISH quantifications, we found that *F. prausnitzii* and *A. muciniphila* showed good correlations, while *F. nucleatum* results were poor. While it is possible that the FISH probe for *F. nucleatum* may not have functioned well, the low frequency of *F. nucleatum* detected in the RNA-Seq reconstructed microbiome suggests that our approach may not accurately detect this bacteria. Several factors could potentially account for our inability to detect *F. nucleatum* in our analysis. For instance, it is possible that experimental challenges, such as physical barriers or low signal strength, may have impeded the detection of *F. nucleatum* genetic material. Additionally, the genetic similarity of *F. nucleatum* to other bacteria, particularly those that are closely related to it, may have hindered our ability to accurately detect and map reads specifically to the *F. nucleatum* genome, thereby

leading to its underrepresentation in our analysis.

Similar factors could be at play for the comparison results between the 16S method and the microbiome reconstruction from RNA-Seq. We found that some bacterial genera are detected similarly by both methods, while others show differences. The results highlight the importance of signal frequency in determining the agreement: the more frequently a bacterium is detected, the greater the agreement between the two approaches. Both comparisons indicate a consistent pattern, supporting the use of the RNA-Seq reconstruction approach. However, it is possible that the presence of some bacteria is underestimated when reconstructing the microbiome from human RNA-Seq data.

Because we had access to this type of experiment in the IEO cohort, we opted to employ human RNA-Seq data to reconstruct the microbiome composition of the samples in our study. However, it is feasible to extract bacterial reads from other types of NGS approaches, such as WXS and WGS, without requiring any modifications of the workflow. While the advantages and limitations of these approaches when analysing human tissues are well-understood (Leek, Scharpf, et al., 2010), less is known about the accuracy of the detected reconstructed microbiome. Despite some obvious differences, such as the high proportion of bacterial RNA sequences detected in the RNA-Seq reconstructed microbiome (mainly rRNA), all three methods were able to detect bacteria in a similar manner, although with a different number of total bacterial reads (WGS showed a higher number compared to the other methods), see section 4.2.3. It is expected that WGS and WXS show concordance since both methods measure DNA, although some bacteria were identified better by one approach than the other. The comparison between RNA-Seq and the other approaches showed more variability, but both comparisons were statistically significantly correlated. We opted to use RNA-Seq data as the main method to reconstruct the microbiome composition not only because already available from IEO cohort, but also due to its easier manipulation (shorter download and PathSeq run-time) compared to other methods, particularly WGS which can be slow to download and analyse due to the large size of its BAM files. Additionally, RNA-Seq had the highest number of primary tumour samples analysed for COAD, making it a practical choice for our study.

5.2 Microbiome characterisation and association

5.2.1 Non-bacterial superkingdom detection

Our approach and the literature share a common focus on bacterial reads, as other superkingdom signals were too low to explore in depth, even though they showed promise. For instance, we de-

tected the signals of known microbes (such as Alphapapillomavirus 9 reads in patients positive for HPV16, 33 and 35) and some eukaryotic organisms associated with colon diseases in COAD samples, in section 4.1.3. The paucity of viral reads may be attributed to an overly stringent filtering of human reads. Since the human genome encompasses integrated viral reads, the proviruses, the similarity between these proviral sequences and other viral reads might have led to the premature exclusion of viral content during the filtering stage. Concerning fungi, a recent investigation by Narunsky-Haziza et al., 2022 explored the pan-cancer mycobiome of TCGA samples using both WGS and RNA-Seq approaches. This study revealed variations in fungal compositions across various cancer types, which exhibited associations with bacterial and immune infiltration compositions. Additionally, the authors highlighted the lower prevalence of fungal signals compared to bacterial ones, posing significant challenges in this analysis. The analysis of fungal data is influenced by issues arising from low biomass organisms, as explained by Eisenhofer et al., 2019, which amplifies the vulnerability to technical noise. Furthermore, the limited availability of published fungal genomes further complicates the analysis of fungal data. Furthermore, Narunsky-Haziza et al., 2022, highlighted the possible impact of mobile genetic elements as a potential cause of biases in fungal classification. In general, exploring these superkingdoms in greater detail may necessitate deeper sequencing depth.

5.2.2 Toxin search

Similarly to what has been explained for the non-bacterial superkingdoms, our capability to identify bacterial toxin signals in the RNA-Seq data was restricted due to the low number of reads, see section 4.3. We were only able to detect a few toxin sequences, which made it difficult to conduct further analysis. Moreover, it is challenging to rely on these results as there was no overlap between the samples identified as toxin-positive using the three distinct NGS approaches we examined. However, out of the three methods, WGS demonstrated the highest abundance of toxin reads and the largest number of positive samples. Among the 33 samples that were analysed, colibactin signals were detected in 9 samples, while spermidine signals were found in 16 samples. Although these results were not corroborated by the other NGS techniques, they may indicate the presence of true toxin signals and so the analysis of the complete WGS dataset could be a promising approach. Nonetheless, due to the significant computational demands and time limitations, we did not perform the analysis on the entire WGS dataset. However, this presents a potential area for future investigations.

5.2.3 Bacterial association with tumour properties

Although the bacterial reconstructed microbiome of samples from different cancer types showed varying bacterial compositions, as illustrated in Figure 4.13, we were mostly unable to identify associations between the microbial composition and the tumour properties of the cancer types, with the exception of colon cancer, see Figure 4.17. The absence of associations between the microbial composition and the tumour properties of the majority of cancer types does not necessarily mean that there are no connections between the bacteria and various properties of the tumour in some or all of the cancer types we tested. In fact in colon cancer, which has the highest microbial biomass in the human body (Davenport et al., 2017), we were able to detect associations between bacteria and tumour properties. However, it is possible that certain cancer types exhibit interactions with specific bacterial species or pathways that are too subtle for our approach to detect. Furthermore, considering the lower abundance of bacteria in these tissues, adopting a more stringent approach to identify tissue-resident bacteria may lead to a more precise characterisation of the tumour bacterial composition in these other cancer types. Interestingly, in addition to COAD, we observed associations between the bacterial composition of HNSC, OV and READ with the aneuploidy of specific chromosomes, as well as between BRCA and stemness and mutation burden, even if mild. These results suggest the presence of a potential tumor-bacteria interaction even in these cancer types but, as we did not find any previous reports on these associations in the literature, we chose to narrow our focus to COAD.

When testing colon cancer properties association with bacterial reconstructed microbiome, we discovered several intriguing associations and for the identified properties, we detected specific bacterial species linked to the properties of the cancer, see section 4.5.1. Notably, we observed correlations with side, MSI and CIMP, which are well-known colon cancer properties that are often associated with one another (e.g. right-sided colon tumours frequently exhibit MSI and CIMP). The observed associations do not provide a clear cause-and-effect relationship, making it difficult to determine whether the bacterial composition influences the development of specific molecular characteristics in tumour cells or if the tumour characteristics shape the microenvironment, leading to the selection of specific bacteria. Nonetheless, these findings indicate that, within the tumour microenvironment, specific bacterial compositions coexist alongside tumour cells with distinct tumour properties. Understanding how this interaction evolves during tumour development requires further investigation, shedding light on potential mechanisms through which the components of the tumour microenvironment interact with one another. Additionally, we discovered a correlation with CMS, which is a classification of colon cancer tumours based on their gene expression. We found that CMS1, which is considered the ”immune cell subtype” due to

the high levels of immune cell infiltration, had a distinct bacterial composition compared to the other subtypes. For this reason, we explored the potential relationship between bacteria and various types of immune cells, see section 4.5.2. We found a notable correlation between bacterial compositions and the presence of mast cells, a type of immune cell that is known to interact with bacteria (Johnzon, Rönnberg, and Pejler, 2016). This finding is particularly intriguing because it suggests one of the earlier mentioned possible mechanisms through which different players shape the tumour microenvironment: the bacterial composition may be linked to immune cell activity and could contribute to a specific tumour microenvironment. This finding is particularly intriguing considering the influence of gut bacteria on the effectiveness of immunotherapy, as mentioned in the study by Temraz et al., 2019. Consequently, the composition of bacteria residing in the tumour microenvironment could emerge as a crucial factor to consider when determining the most suitable treatment for patients.

We investigated another potential mechanism linking bacteria to the microenvironment: we tested the link of bacteria metabolism with tumour properties in section 4.8. Our findings revealed the enrichment of fatty acid pathways in left-sided tumours, while tumours with high mutation burden were associated with DNA degradation and sugar metabolism. It is intriguing that we were able to link bacteria to these pathways, as they have been previously implicated in tumorigenesis (Choi et al., 2019; Pickens et al., 2016). Due to the low read count and the predominant alignment to rRNA sequences, explained in Figure 4.3a and section4.1.4, the research of bacterial pathway signals in each RNA-Seq sample yielded unsatisfactory results. However, our strategy of merging microbial reads from samples exhibiting similar characteristics (e.g. samples from the left side of the colon) proved effective in increasing the read count and identifying specific bacterial pathways. Among the detected pathways, some have previously been linked to colon cancer, such as the fatty acid pathways, while others are reasonably associated with tumour characteristics, such as DNA degradation pathways and a high mutation burden. Nonetheless, since this approach lacks high resolution, we were unable to uncover more nuanced associations.

Finally, we found that the bacterial composition of the tumour microenvironment is strongly correlated with the prognosis of colon cancer patients, described in section 4.7. Our analysis showed that patients with different bacterial compositions exhibited different relapse rates and that specific bacterial species were associated with varying disease-free survival. This is particularly intriguing because it indicates that not only individual bacteria but also the overall bacterial composition can serve as outcome indicators for patients, as previously demonstrated by Dohlman et al., 2021. This finding further underscores the significant role of the bacterial microenvironment in shaping cancer, particularly from a therapeutic perspective.

5.3 Conclusion

In conclusion, in our project, we develop a workflow to use NGS data to extract microbial reads and estimate the microbiome composition of the samples analysed. We focused on RNA-Seq data and we validated our findings in various ways, including the comparison with other experimental approaches, other NGS microbiome reconstruction and the detection of expected microbes such as Alphapapillomavirus reads in HPV-positive samples. Further analyses, such as the research of specific toxin signals, were not feasible with RNA-Seq data. We used this bacterial estimation to develop an integrative approach that associates the entire bacterial composition of the samples with the properties of the tumours from which they were extracted.

Some of the associations we found depict a complex situation in which the bacterial composition is associated with specific tumour properties. While we cannot determine the direction of this association, whether the bacteria affect the tumour during growth and development or if the tumour shapes the environment to select particular bacteria, our findings support the link between bacteria and colon tumours. Our findings highlight the significance of the relationship between the microbiota and cancer, not only in terms of understanding the disease but also in its potential implications for patient outcomes. Additionally, studies such as those conducted by Salvucci et al., 2022 and Duggan et al., 2023, on *F. nucleatum* suggest that the impact of microbiota depends on underlying tumour properties, further emphasising the relevance of our systematic study.

Despite our approach focusing on associations, we believe that our work reinforces the notion of a connection between tumours and their microenvironment, resulting in an ecosystem with distinct interaction roles between tumour cells, microbes and other host cells like immune infiltrates. This perspective is intriguing because it presents numerous possibilities, including the discovery of novel biomarkers (such as bacteria or other microbes) and the identification of new therapeutic targets, with microbes as a possible target alongside radio- or chemotherapy.

References

Abed, Jawad et al. (2020). "Colon cancer-associated Fusobacterium nucleatum may originate from the oral cavity and reach colon tumors via the circulatory system". In: *Frontiers in cellular and infection microbiology* 10, p. 400.

Aghabozorgi, Amirsaeed Sabeti et al. (2019). "Role of adenomatous polyposis coli (APC) gene mutations in the pathogenesis of colorectal cancer; current status and perspectives". In: *Biochimie* 157, pp. 64–71.

Ahmad Zawawi, Sahira Syamimi and Marahaini Musa (2022). "Dynamic Co-evolution of cancer cells and cancer-associated fibroblasts: Role in right-and left-sided colon cancer progression and its clinical relevance". In: *Biology* 11.7, p. 1014.

Akazawa, Yuko et al. (2021). "Significance of serum palmitoleic acid levels in inflammatory bowel disease". In: *Scientific Reports* 11.1, p. 16260.

Alipour, Majid (2021). "Molecular mechanism of Helicobacter pylori-induced gastric cancer". In: *Journal of gastrointestinal cancer* 52, pp. 23–30.

Almeida, Alexandre et al. (2021). "A unified catalog of 204,938 reference genomes from the human gut microbiome". In: *Nature biotechnology* 39.1, pp. 105–114.

Arain, Mustafa A et al. (2010). "CIMP status of interval colon cancers: another piece to the puzzle". In: *Official journal of the American College of Gastroenterology— ACG* 105.5, pp. 1189–1195.

Arthur, Janelle C (2020). "Microbiota and colorectal cancer: colibactin makes its mark". In: *Nature Reviews Gastroenterology & Hepatology* 17.6, pp. 317–318.

Asnicar, Francesco et al. (2021). "Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals". In: *Nature Medicine* 27.2, pp. 321–332.

Bajnok, Jaroslav et al. (2019). "High frequency of infection of lung cancer patients with the parasite Toxoplasma gondii". In: *ERJ Open Research* 5.2.

Bao, Yiqiao et al. (2021). "A common pathway for activation of host-targeting and bacteria-targeting toxins in human intestinal bacteria". In: *Mbio* 12.4, e00656–21.

Beghini, Francesco et al. (2021). "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3". In: *elife* 10, e65088.

Bolyen, Evan et al. (2019). "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2". In: *Nature biotechnology* 37.8, pp. 852–857.

Bonneville, Russell et al. (2017). "Landscape of microsatellite instability across 39 cancer types". In: *JCO precision oncology* 1, pp. 1–15.

Brennan, Caitlin A and Wendy S Garrett (2019). "Fusobacterium nucleatum—symbiont, opportunist and oncobacterium". In: *Nature Reviews Microbiology* 17.3, pp. 156–166.

Buikhuisen, Joyce Y, Arezo Torang, and Jan Paul Medema (2020). "Exploring and modelling colon cancer inter-tumour heterogeneity: opportunities and challenges". In: *Oncogenesis* 9.7, p. 66.

Bukholm, Ida K and Jahn M Nesland (2000). "Protein expression of p53, p21 (WAF1/CIP1), bcl-2, Bax, cyclin D1 and pRb in human colon carcinomas". In: *Virchows Archiv* 436, pp. 224–228.

Bullman, Susan et al. (2017). "Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer". In: *Science* 358.6369, pp. 1443–1448.

Butler, Lesley M et al. (2017). "Plasma fatty acids and risk of colon and rectal cancers in the Singapore Chinese Health Study". In: *NPJ precision oncology* 1.1, p. 38.

Callahan, Benjamin J et al. (2016). "DADA2: High-resolution sample inference from Illumina amplicon data". In: *Nature methods* 13.7, pp. 581–583.

Cerami, Ethan et al. (2012). "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data". In: *Cancer discovery* 2.5, pp. 401–404.

Cheng, Hong Sheng et al. (2023). "The Blood Microbiome and Health: Current Evidence, Controversies, and Challenges". In: *International Journal of Molecular Sciences* 24.6, p. 5633.

Cheng, Wai Teng, Haresh Kumar Kantilal, and Fabian Davamani (2020). "The mechanism of Bacteroides fragilis toxin contributes to colon cancer formation". In: *The Malaysian journal of medical sciences: MJMS* 27.4, p. 9.

Choi, SeokGyeong et al. (2019). "Clinical and biochemical relevance of monounsaturated fatty acid metabolism targeting strategy for cancer stem cell elimination in colon cancer". In: *Biochemical and Biophysical Research Communications* 519.1, pp. 100–105.

Chung, Gyung-Tae et al. (1999). "Identification of a third metalloprotease toxin gene in extraintestinal isolates of Bacteroides fragilis". In: *Infection and immunity* 67.9, pp. 4945–4949.

Clay, Slater L, Diogo Fonseca-Pereira, and Wendy S Garrett (2022). "Colorectal cancer: the facts in the case of the microbiota". In: *Journal of Clinical Investigation* 132.4, e155101.

Dai, Zhenwei et al. (2018). "Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers". In: *Microbiome* 6.1, pp. 1–12.

Danecek, Petr et al. (2021). "Twelve years of SAMtools and BCFtools". In: *Gigascience* 10.2, giab008.

Davenport, Emily R et al. (2017). "The human microbiome in evolution". In: *BMC biology* 15.1, pp. 1–12.

David, Lawrence A et al. (2014). "Diet rapidly and reproducibly alters the human gut microbiome". In: *Nature* 505.7484, pp. 559–563.

De Martel, Catherine et al. (2017). "Worldwide burden of cancer attributable to HPV by site, country and HPV type". In: *International journal of cancer* 141.4, pp. 664–670.

Dejea, Christine M et al. (2018). "Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria". In: *Science* 359.6375, pp. 592–597.

Dekaboruah, Elakshi et al. (2020). "Human microbiome: an academic update on human body site specific surveillance and its possible role". In: *Archives of microbiology* 202, pp. 2147–2167.

Dey, Priyankar and Saumya Ray Chaudhuri (2022). "The opportunistic nature of gut commensal microbiota". In: *Critical Reviews in Microbiology*, pp. 1–25.

Dohlman, Anders B et al. (2021). "The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants". In: *Cell host & microbe* 29.2, pp. 281–298.

Dow, Lukas E et al. (2015). "Apc restoration promotes cellular differentiation and reestablishes crypt homeostasis in colorectal cancer". In: *Cell* 161.7, pp. 1539–1552.

Duggan, William P et al. (2023). "Increased Fusobacterium tumoural abundance affects immunogenicity in mucinous colorectal cancer and may be associated with improved clinical outcome". In: *Journal of Molecular Medicine*, pp. 1–13.

Eisenhofer, Raphael et al. (2019). "Contamination in low microbial biomass microbiome studies: issues and recommendations". In: *Trends in microbiology* 27.2, pp. 105–117.

Ellrott, Kyle et al. (2018). "Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines". In: *Cell systems* 6.3, pp. 271–281.

Farriol, Mireia et al. (2001). "Role of putrescine in cell proliferation in a colon carcinoma cell line". In: *Nutrition* 17.11-12, pp. 934–938.

Finotello, Francesca and Zlatko Trajanoski (2018). "Quantifying tumor-infiltrating immune cells from transcriptomics data". In: *Cancer Immunology, Immunotherapy* 67.7, pp. 1031–1040.

Flores-Hernández, Eric et al. (2020). "Canonical and non-canonical Wnt signaling are simultaneously activated by Wnts in colon cancer cells". In: *Cellular Signalling* 72, p. 109636.

Franzosa, Eric A et al. (2015). "Identifying personal microbiomes using metagenomic codes". In: *Proceedings of the National Academy of Sciences* 112.22, E2930–E2938.

Galon, Jérôme et al. (2006). "Type, density, and location of immune cells within human colorectal tumors predict clinical outcome". In: *Science* 313.5795, pp. 1960–1964.

Gao, Jianjiong et al. (2013). "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal". In: *Science signaling* 6.269, pl1–pl1.

Garrett, Wendy S (2019). "The gut microbiota and colon cancer". In: *Science* 364.6446, pp. 1133–1135.

Germann, Markus et al. (2020). "Neutrophils suppress tumor-infiltrating T cells in colon cancer via matrix metalloproteinase-mediated activation of TGF β". In: *EMBO molecular medicine* 12.1, e10681.

Gihawi, Abraham, Colin S Cooper, and Brewer Daniel S (2023). "Caution Regarding the Specificities of Pan-Cancer Microbial Structure". In: *bioRxiv*. eprint: https://www.biorxiv.org/content/early/2023/01/18/2023.01.16.523562.full.pdf.

Glassner, Kerri L, Bincy P Abraham, and Eamonn MM Quigley (2020). "The microbiome and inflammatory bowel disease". In: *Journal of Allergy and Clinical Immunology* 145.1, pp. 16–27.

Goel, Ajay et al. (2007). "The CpG island methylator phenotype and chromosomal instability are inversely correlated in sporadic colorectal cancer". In: *Gastroenterology* 132.1, pp. 127–138.

Goodwin, Andrew C et al. (2011). "Polyamine catabolism contributes to enterotoxigenic Bacteroides fragilis-induced colon tumorigenesis". In: *Proceedings of the National Academy of Sciences* 108.37, pp. 15354–15359.

Gordon, Jeffrey et al. (2013). "Superorganisms and holobionts". In: *Microbe* 8.4, pp. 152–153.

Grady, William M (2004). "Genomic instability and colon cancer". In: *Cancer and metastasis reviews* 23, pp. 11–27.

Grayson, Dennis R and Alessandro Guidotti (2013). "The dynamics of DNA methylation in schizophrenia and related psychiatric disorders". In: *Neuropsychopharmacology* 38.1, pp. 138–166.

Greuter, Daniel et al. (2016). "probeBase—an online resource for rRNA-targeted oligonucleotide probes and primers: new features 2016". In: *Nucleic Acids Research* 44.D1, pp. D586–D589.

Grossman, Robert L et al. (2016). "Toward a shared vision for cancer genomic data". In: *New England Journal of Medicine* 375.12, pp. 1109–1112.

Gualco, Gabriela et al. (2006). "Flat elevated lesions of the colon and rectum: a spectrum of neoplastic and nonneoplastic entities". In: *Annals of Diagnostic Pathology* 10.6, pp. 333–338.

Guinney, Justin et al. (2015). "The consensus molecular subtypes of colorectal cancer". In: *Nature medicine* 21.11, pp. 1350–1356.

Haas, Brian J and Michael C Zody (2010). "Advancing RNA-seq analysis". In: *Nature biotechnology* 28.5, pp. 421–423.

Hanahan, Douglas (2022). "Hallmarks of cancer: new dimensions". In: *Cancer discovery* 12.1, pp. 31–46.

Heintz-Buschart, Anna and Paul Wilmes (2018). "Human gut microbiome: function matters". In: *Trends in microbiology* 26.7, pp. 563–574.

Hermida, Leandro C, E Michael Gertz, and Eytan Ruppin (2022). "Predicting cancer prognosis and drug response from the tumor microbiome". In: *Nature communications* 13.1, p. 2896.

Høiby, Niels (2022). "Louis Pasteur and the birth of microbiology in Denmark". In: *APMIS*.

Holscher, Hannah D (2017). "Dietary fiber and prebiotics and the gastrointestinal microbiota". In: *Gut microbes* 8.2, pp. 172–184.

Hothorn, Torsten et al. (2008). "Implementing a class of permutation tests: the coin package". In: *Journal of statistical software* 28, pp. 1–23.

Hugen, Niek et al. (2016). "Advances in the care of patients with mucinous colorectal cancer". In: *Nature reviews Clinical oncology* 13.6, pp. 361–369.

Illumina (2013). "16s metagenomic sequencing library preparation". In: URL: https://support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html.

Imperiale, Thomas F et al. (2014). "Multitarget stool DNA testing for colorectal-cancer screening". In: *New England Journal of Medicine* 370.14, pp. 1287–1297.

Irwin, Michael R (2019). "Sleep and inflammation: partners in sickness and in health". In: *Nature Reviews Immunology* 19.11, pp. 702–715.

Itatani, Yoshiro, Kenji Kawada, and Yoshiharu Sakai (2019). "Transforming growth factor-β signaling pathway in colorectal cancer and its tumor microenvironment". In: *International journal of molecular sciences* 20.23, p. 5822.

Jayachandran, Muthukumaran, Stephen Sum Man Chung, and Baojun Xu (2020). "A critical review of the relationship between dietary components, the gut microbe Akkermansia muciniphila, and human health". In: *Critical reviews in food science and nutrition* 60.13, pp. 2265–2276.

Johnson, Jethro S et al. (2019). "Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis". In: *Nature communications* 10.1, p. 5029.

Johnson, W Evan, Cheng Li, and Ariel Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1, pp. 118–127.

Johnzon, Carl-Fredrik, Elin Rönnberg, and Gunnar Pejler (2016). "The role of mast cells in bacterial infection". In: *The American journal of pathology* 186.1, pp. 4–14.

Karcher, Nicolai et al. (2020). "Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations". In: *Genome biology* 21, pp. 1–27.

Karstens, Lisa et al. (2019). "Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments". In: *MSystems* 4.4, e00290–19.

Kato, Kumiko et al. (2018). "Association between functional lactase variants and a high abundance of Bifidobacterium in the gut of healthy Japanese people". In: *PLoS One* 13.10, e0206189.

Kennedy, Katherine M et al. (2023). "Questioning the fetal microbiome illustrates pitfalls of low-biomass microbial studies". In: *Nature* 613.7945, pp. 639–649.

Keum, NaNa and Edward Giovannucci (2019). "Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies". In: *Nature reviews Gastroenterology & hepatology* 16.12, pp. 713–732.

Khkheirouri, Sorayya, Parinaz Kalejahi, and Seyyed Gholamreza Noorazar (2016). "Plasma levels of serotonin, gastrointestinal symptoms, and sleep problems in children with autism". In: *Turkish Journal of Medical Sciences* 46.6, pp. 1765–1772.

Kosik-Bogacka, Danuta et al. (2021). "Prevalence, subtypes and risk factors of Blastocystis spp. infection among pre-and perimenopausal women". In: *BMC Infectious Diseases* 21.1, pp. 1–14.

Kostic, Aleksandar D et al. (2011). "PathSeq: software to identify or discover microbes by deep sequencing of human tissue". In: *Nature biotechnology* 29.5, pp. 393–396.

Krishnareddy, Suneeta (2019). "The microbiome in celiac disease". In: *Gastroenterology Clinics* 48.1, pp. 115–126.

Kumarasamy, Vinoth et al. (2022). "Association of Blastocystis hominis with colorectal cancer: a systematic review of in vitro and in vivo evidences". In: *World Journal of Gastrointestinal Oncology* 14.3, p. 734.

Kuziel, Gavin A and Seth Rakoff-Nahoum (2022). "The gut microbiome". In: *Current Biology* 32.6, R257–R264.

Laghi, Luigi et al. (2020). "Prognostic and predictive cross-roads of microsatellite instability and immune response to colon cancer". In: *International Journal of Molecular Sciences* 21.24, p. 9680.

Lagier, Jean-Christophe et al. (2015). "The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota". In: *Clinical microbiology reviews* 28.1, pp. 237–264.

Lalonde, Emilie et al. (2010). "Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing". In: *Human mutation* 31.8, pp. 918–923.

Lane, David J et al. (1985). "Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses." In: *Proceedings of the National Academy of Sciences* 82.20, pp. 6955–6959.

Lange, Kathleen et al. (2016). "Effects of antibiotics on gut microbiota". In: *Digestive Diseases* 34.3, pp. 260–268.

Langmead, Ben and Steven L Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4, pp. 357–359.

Lau, Agnes Wei Yin et al. (2021). "The chemistry of gut microbiome in health and diseases". In: *Progress In Microbes & Molecular Biology* 4.1.

Law, Charity W et al. (2014). "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". In: *Genome biology* 15.2, pp. 1–17.

Lawrence, Garreth W et al. (2020). "Potential use of biotherapeutic bacteria to target colorectal cancer-associated taxa". In: *International Journal of Molecular Sciences* 21.3, p. 924.

Lawrence, Michael S et al. (2014). "Discovery and saturation analysis of cancer genes across 21 tumour types". In: *Nature* 505.7484, pp. 495–501.

Le, Dung T et al. (2015). "PD-1 blockade in tumors with mismatch-repair deficiency". In: *New England Journal of Medicine* 372.26, pp. 2509–2520.

Le Guennec, Loic et al. (2020). "Strategies used by bacterial pathogens to cross the blood–brain barrier". In: *Cellular microbiology* 22.1, e13132.

Lee, Michael S, David G Menter, and Scott Kopetz (2017). "Right versus left colon cancer biology: integrating the consensus molecular subtypes". In: *Journal of the National Comprehensive Cancer Network* 15.3, pp. 411–419.

Leek, Jeffrey T, W Evan Johnson, et al. (2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". In: *Bioinformatics* 28.6, pp. 882–883.

Leek, Jeffrey T, Robert B Scharpf, et al. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data". In: *Nature Reviews Genetics* 11.10, pp. 733–739.

Lehouritis, Panos et al. (2015). "Local bacteria affect the efficacy of chemotherapeutic drugs". In: *Scientific reports* 5.1, pp. 1–12.

Li, Yong-Xin et al. (2020). "Toxoplasma gondii infection in patients with lung diseases in Shandong province, eastern China". In: *Acta Tropica* 211, p. 105554.

Liu, Yang et al. (2018). "Comparative molecular analysis of gastrointestinal adenocarcinomas". In: *Cancer cell* 33.4, pp. 721–735.

Malki, Ahmed et al. (2020). "Molecular mechanisms of colon cancer progression and metastasis: recent insights and advancements". In: *International journal of molecular sciences* 22.1, p. 130.

Malta, Tathiane M et al. (2018). "Machine learning identifies stemness features associated with oncogenic dedifferentiation". In: *Cell* 173.2, pp. 338–354.

Martel, Catherine de et al. (2020). "Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis". In: *The Lancet Global Health* 8.2, e180–e190.

McKenna, Aaron et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data". In: *Genome research* 20.9, pp. 1297–1303.

Meienberg, Janine et al. (2016). "Clinical sequencing: is WGS the better WES?" In: *Human genetics* 135, pp. 359–362.

Menni, Cristina et al. (2021). "High intake of vegetables is linked to lower white blood cell profile and the effect is mediated by the gut microbiome". In: *BMC medicine* 19, pp. 1–10.

Michel, Maurice et al. (2021). "The role of p53 dysfunction in colorectal cancer and its implication for therapy". In: *Cancers* 13.10, p. 2296.

Miller, Steve and Charles Chiu (2022). "The role of metagenomics and next-generation sequencing in infectious disease diagnosis". In: *Clinical chemistry* 68.1, pp. 115–124.

Moodley, Yoshan et al. (2009). "The peopling of the Pacific from a bacterial perspective". In: *Science* 323.5913, pp. 527–530.

Morgun, Andrey et al. (2015). "Uncovering effects of antibiotics on the host and microbiota using transkingdom gene networks". In: *Gut* 64.11, pp. 1732–1743.

Mousa, Walaa K, Fadia Chehadeh, and Shannon Husband (2022). "Recent advances in understanding the structure and function of the human microbiome". In: *Frontiers in Microbiology* 13, p. 825338.

Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press, p. 11.

Narunsky-Haziza, Lian et al. (2022). "Pan-cancer analyses reveal cancer-type-specific fungal ecologies and bacteriome interactions". In: *Cell* 185.20, pp. 3789–3806.

Nawa, Toru et al. (2008). "Differences between right-and left-sided colon cancer in patient characteristics, cancer morphology and histology". In: *Journal of gastroenterology and hepatology* 23.3, pp. 418–423.

Nejman, Deborah et al. (2020). "The human tumor microbiome is composed of tumor type–specific intracellular bacteria". In: *Science* 368.6494, pp. 973–980.

Newman, Aaron M et al. (2019). "Determining cell type abundance and expression from bulk tissues with digital cytometry". In: *Nature biotechnology* 37.7, pp. 773–782.

Nougayrède, Jean-Philippe et al. (2006). "Escherichia coli induces DNA double-strand breaks in eukaryotic cells". In: *Science* 313.5788, pp. 848–851.

Nurk, Sergey et al. (2022). "The complete sequence of a human genome". In: *Science* 376.6588, pp. 44–53.

Ogino, Shuji et al. (2009). "Lymphocytic reaction to colorectal cancer is associated with longer survival, independent of lymph node count, microsatellite instability, and CpG island methylator phenotype". In: *Clinical Cancer Research* 15.20, pp. 6412–6420.

Parra-Torres, Valeria et al. (2023). "Periodontal bacteria in the brain—Implication for Alzheimer's disease: a systematic review". In: *Oral Diseases* 29.1, pp. 21–28.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12, pp. 2825–2830.

Peterson, Danielle et al. (2021). "Comparative analysis of 16S rRNA gene and metagenome sequencing in pediatric gut microbiomes". In: *Frontiers in microbiology* 12, p. 670336.

Pickens, C Austin et al. (2016). "Altered Saturated and Monounsaturated Plasma Phospholipid Fatty Acid Profiles in Adult Males with Colon AdenomasAssociations between Fatty Acids and Adenomas". In: *Cancer Epidemiology, Biomarkers & Prevention* 25.3, pp. 498–506.

Poore, Gregory D et al. (2020). "Microbiome analyses of blood and tissues suggest cancer diagnostic approach". In: *Nature* 579.7800, pp. 567–574.

Prudent, Elsa and Didier Raoult (2019). "Fluorescence in situ hybridization, a complementary molecular tool for the clinical diagnosis of infectious diseases by intracellular and fastidious bacteria". In: *FEMS Microbiology Reviews* 43.1, pp. 88–107.

Quast, Christian et al. (2012). "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools". In: *Nucleic acids research* 41.D1, pp. D590–D596.

Ramette, Alban (2007). "Multivariate analyses in microbial ecology". In: *FEMS microbiology ecology* 62.2, pp. 142–160.

Renson, Audrey et al. (2020). "Gut bacterial taxonomic abundances vary with cognition, personality, and mood in the Wisconsin Longitudinal Study". In: *Brain, Behavior, & Immunity-Health* 9, p. 100155.

Riihimäki, Matias et al. (2016). "Patterns of metastasis in colon and rectal cancer". In: *Scientific reports* 6.1, pp. 1–9.

Riley, Peter A (2017). "Principles of microscopy, culture and serology-based diagnostics". In: *Medicine* 45.10, pp. 639–644.

Robinson, Kelly M et al. (2017). "Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data". In: *Microbiome* 5.1, pp. 1–17.

Rous, Peyton (1911). "A sarcoma of the fowl transmissible by an agent separable from the tumor cells". In: *The Journal of experimental medicine* 13.4, p. 397.

Rubinstein, Mara Roxana et al. (2019). "Fusobacterium nucleatum promotes colorectal cancer by inducing Wnt/β-catenin modulator Annexin A1". In: *EMBO reports* 20.4, e47638.

Salter, Susannah J et al. (2014). "Reagent and laboratory contamination can critically impact sequence-based microbiome analyses". In: *BMC biology* 12, pp. 1–12.

Salvucci, Manuela et al. (2022). "Patients with mesenchymal tumours and high Fusobacteriales prevalence have worse prognosis in colorectal cancer (CRC)". In: *Gut* 71.8, pp. 1600–1612.

Sambruni, Gaia et al. (2023). "Location and condition based reconstruction of colon cancer microbiome from human RNA sequencing data". In: *Genome Medicine*, submitted.

Sarvepalli, Shashank et al. (2018). "Natural history of colonic polyposis in young patients with familial adenomatous polyposis". In: *Gastrointestinal Endoscopy* 88.4, pp. 726–733.

Sawhney, Mandeep S et al. (2006). "Microsatellite instability in interval colon cancers". In: *Gastroenterology* 131.6, pp. 1700–1705.

Schirmer, Melanie et al. (2016). "Linking the human gut microbiome to inflammatory cytokine production capacity". In: *Cell* 167.4, pp. 1125–1136.

Sepich-Poore, Gregory et al. (2023). "Reply to: Caution Regarding the Specificities of Pan-Cancer Microbial Structure". In: eprint: https://www.biorxiv.org/content/biorxiv/early/2023/02/13/2023.02.10.528049.full.pdf.

Sepich-Poore, Gregory D et al. (2021). "The microbiome and human cancer". In: *Science* 371.6536, eabc4552.

Singh, Manish Pratap et al. (2021). "Molecular subtypes of colorectal cancer: An emerging therapeutic opportunity for personalized medicine". In: *Genes & diseases* 8.2, pp. 133–145.

Sokol-Borrelli, Sarah L, Rachel S Coombs, and Jon P Boyle (2020). "A comparison of stage conversion in the coccidian apicomplexans Toxoplasma gondii, Hammondia hammondi, and Neospora caninum". In: *Frontiers in Cellular and Infection Microbiology* 10, p. 608283.

Sprang, Maximilian, Miguel A Andrade-Navarro, and Jean-Fred Fontaine (2022). "Batch effect detection and correction in RNA-seq data using machine-learning-based automated assessment of quality". In: *BMC bioinformatics* 23.6, pp. 1–15.

Strasser, Helmut and Christian Weber (1999). "On the asymptotic theory of permutation statistics". In.

Sung, Hyuna et al. (2021). "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3, pp. 209–249.

Swann, JR et al. (2020). "Considerations for the design and conduct of human gut microbiota intervention studies relating to foods". In: *European journal of nutrition* 59, pp. 3347–3368.

Takano, Susumu et al. (1995). "Incidence of hepatocellular carcinoma in chronic hepatitis B and C: a prospective study of 251 patients". In: *Hepatology* 21.3, pp. 650–655.

Taylor, Alison M et al. (2018). "Genomic and functional approaches to understanding cancer aneuploidy". In: *Cancer cell* 33.4, pp. 676–689.

Temraz, Sally et al. (2019). "Gut microbiome: a promising biomarker for immunotherapy in colorectal cancer". In: *International journal of molecular sciences* 20.17, p. 4155.

Tett, Adrian et al. (2019). "The Prevotella copri complex comprises four distinct clades underrepresented in westernized populations". In: *Cell host & microbe* 26.5, pp. 666–679.

Timar, Jozsef and Karl Kashofer (2020). "Molecular epidemiology and diagnostics of KRAS mutations in human cancer". In: *Cancer and Metastasis Reviews* 39, pp. 1029–1038.

Tjalsma, Harold et al. (2012). "A bacterial driver–passenger model for colorectal cancer: beyond the usual suspects". In: *Nature Reviews Microbiology* 10.8, pp. 575–582.

Valguarnera, Ezequiel and Juliane Bubeck Wardenburg (2020). "Good gone bad: one toxin away from disease for Bacteroides fragilis". In: *Journal of molecular biology* 432.4, pp. 765–785.

Valles-Colomer, Mireia et al. (2023). "The person-to-person transmission landscape of the gut and oral microbiomes". In: *Nature*, pp. 1–11.

Vangay, Pajau et al. (2018). "US immigration westernizes the human gut microbiome". In: *Cell* 175.4, pp. 962–972.

Venturi, Miro et al. (1997). "Genotoxic activity in human faecal water and the role of bile acids: a study using the alkaline comet assay." In: *Carcinogenesis* 18.12, pp. 2353–2359.

Verhoog, Sanne et al. (2019). "Dietary factors and modulation of bacteria strains of Akkermansia muciniphila and Faecalibacterium prausnitzii: a systematic review". In: *Nutrients* 11.7, p. 1565.

Walker, Mark A et al. (2018). "GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts". In: *Bioinformatics* 34.24, pp. 4287–4289.

Wang, Haiwei et al. (2019). "Analysis of the transcriptomic features of microsatellite instability subtype colon cancer". In: *BMC cancer* 19.1, pp. 1–16.

Wang, Jihan et al. (2021). "Global analysis of microbiota signatures in four major types of gastrointestinal cancer". In: *Frontiers in Oncology* 11, p. 685641.

Wang, Ni and Jing-Yuan Fang (2022). "Fusobacterium nucleatum, a key pathogenic factor and microbial biomarker for colorectal cancer". In: *Trends in Microbiology*.

Wang, Qi et al. (2023). "Computational methods and challenges in analyzing intratumoral microbiome data". In: *Trends in Microbiology*.

Wastyk, Hannah C et al. (2021). "Gut-microbiota-targeted diets modulate human immune status". In: *Cell* 184.16, pp. 4137–4153.

Whisner, Corrie M and C Athena Aktipis (2019). "The role of the microbiome in cancer initiation and progression: how microbes and cancer cells utilize excess energy and promote one another's growth". In: *Current nutrition reports* 8, pp. 42–51.

Wood, Derrick E, Jennifer Lu, and Ben Langmead (2019). "Improved metagenomic analysis with Kraken 2". In: *Genome biology* 20, pp. 1–13.

Wood, Derrick E and Steven L Salzberg (2014). "Kraken: ultrafast metagenomic sequence classification using exact alignments". In: *Genome biology* 15.3, pp. 1–12.

Wooley, John C, Adam Godzik, and Iddo Friedberg (2010). "A primer on metagenomics". In: *PLoS computational biology* 6.2, e1000667.

Yadav, M and NS Chauhan (2021). "Overview of the rules of the microbial engagement in the gut microbiome: a step towards microbiome therapeutics". In: *Journal of Applied Microbiology* 130.5, pp. 1425–1441.

Zhang, Ya and Xin Wang (2020). "Targeting the Wnt/β-catenin signaling pathway in cancer". In: *Journal of hematology & oncology* 13, pp. 1–16.

Zhao, Wenchao et al. (2022). "The Association of R-Loop Binding Proteins Subtypes with CIN Implicates Therapeutic Strategies in Colorectal Cancer". In: *Cancers* 14.22, p. 5607.

Zhu, Qiyun et al. (2019). "Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea". In: *Nature communications* 10.1, p. 5477.