Microbiome and Host Immune Responses in Colorectal Cancer Development and Radiotherapy Response

A thesis presented in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Genetics

Massey University, Auckland, New Zealand

Arielle Kae L. Sulit

2022

Abstract

Colorectal cancer (CRC) is a highly heterogeneous disease that manifests differently from patient to patient, making prognosis, management, and therapy more complex as no universal solution is available. With the advent of high throughput sequencing, descriptions of CRC tumors have moved from histopathological features and descriptions to molecular characterization, allowing for the subtyping of CRC into groups with similar characteristics. This inter-tumoral heterogeneity affects CRC development and patient response to treatments. Understanding the different mechanisms of CRC development and response to therapy is therefore crucial to personalized healthcare. The majority of CRC tumors do not have a familial background, suggesting the environment plays a large role in their development. Environmental factors include the microbiome, which has been shown previously to affect CRC development. However, the role of the microbiome has largely been overlooked in studies of the CRC subtyping and radiotherapy response in rectal cancer treatment.

In this thesis, I show that bacteria in CRC may affect immune responses that drive CRC development in different subtypes, and radiotherapy response. I used RNA sequencing to revisit the consensus molecular subtypes (CMS) of CRCs and identified microbes that have possible contributions to their different characteristics. As microbes have been associated with differing responses to therapy, I also looked at their putative roles in radiotherapy response in a rectal cancer cohort. I first developed a computational pipeline that takes raw sequencing reads as input and

yields host gene expression data, microbiome abundances and functional information. I then focused on two subtypes of CRC, CMS1 and CMS4. Analysis of host gene expression in these subtypes confirmed that their expression profiles are enriched in gene sets associated with immune responses. Analysis of the microbiome content found that lipopolysaccharides (LPS) from *Fusobacterium periodonticum* and *Bacteroides fragilis* in CMS1, and *Porphyromonas asaccharolytica* in CMS4 potentially affect the production of the immune infiltrates of their respective subtypes. *F. periodonticum* LPS enhanced cytokine production while LPS from the latter two bacteria suppressed cytokine production in peripheral blood mononuclear cells (PBMCs) *in vitro*. These data indicate possible roles of LPS from these microbes in CRC development via immune response. These also indicate possible roles of these molecules in CRC therapy. I also found that in complete responders of radiotherapy, there was an enrichment in host gene functions that are associated with complement activation, response to viruses, and B-cell activation, all of which indicated a link to immunotherapy responses triggered by radiotherapy. Furthermore, bacteria that had previous associations with immunotherapy responses were enriched in complete responders indicating a role in enhancement of these cytotoxic immune responses in radiosensitivity.

Immune infiltrates have always been a crucial element in cancers, and the type of infiltrates can have conflicting effects on cancer development and therapeutic responses. In this work, I show that the types of bacterial molecules and how they interact between species can affect specific immune responses in CRC development, that dampening of immune responses in CRC is as crucial as inducing

immunogenicity, and that specific bacteria also affect immune responses in a manner that may be similar to immunotherapy that will increase radiosensitivity. This work provides an initial look into mechanisms of how microbes interact with host immune responses affecting CRC development into two different subtypes, and radiosensitivity. It also provides initial experimental evidence for effects of LPS in CRC development and a possible mechanism of radiotherapy sensitivity to test in the laboratory.

Acknowledgements

Ephesians 3:20-21

First, I would like to thank my supervisory team: **Dr. Sebastian Schmeier**, **Dr.Olin Silander**, and **Dr. Rachel Purcell**. I am grateful for your expertise, guidance, and patience throughout the PhD. I am a better scientist and human being through your mentorship.

I am grateful to our collaborators, without whom this thesis would not be what it is: to **Dr. John Pearson** for your advice on my statistical analyses, to **Dr. Barry Hock** for advice on figures and editorial comments for Chapter 3 of this thesis, to **Dr. Judy McKenzie** for advice on the use of the flow cytometer, to **Michelle Daigneault** and **Professor Emma Allen-Vercoe** for the LPS I used in my experiments, to **Dr. Kasmira Wilson** and **Professor Alexander Heriot** for the samples I have used for Chapter 4 of this thesis, and **Professor Frank Frizelle**, for his clinical expertise. I would also like to express my gratitude to the students, scientists, and administrators at the University of Otago, Christchurch who have assisted me during my time as a visiting student there. Thank you, too, to **Dr. Xochitl Morgan**, who agreed to train me in some aspects of bioinformatics analyses. I would also like to thank **Tyler Kolisnik** for his inputs to the MetaFunc pipeline, and for being a great colleague and workmate.

To my family, especially **Mama Dang** and **Papa Gie**, thank you for your unwavering encouragement and support, as I completed a doctorate degree during a pandemic. This is for you!

Thank you too, to **Tita Lala**, for being my adoptive mother in New Zealand.

To my dgroup ladies, especially my leader **Pdee,** who are a constant reminder of God's presence, grace, and mercy in my life. Thank you for keeping me grounded and accountable.

To my feast light group, especially **Meann** and **Tito Butch**, thank you for the prayers. **Meann,** thank you for the roadtrips!

To **Richard, Owee,** and **Baby Ollie**, for being a source of inspiration, and cuteness that is very much needed.

To **Jerica, Carissa,** and **Angeli,** thank you for the laughter, encouragement, and expertise as I ask for advice from my Molecular Medicine gals.

To IPSO, and my patatas - **May, Kaye, Aidz -** I would be so lost without you. Thank you for keeping me sane, always.

Lastly, to **Brian**, who is all of the above and so much more. This is ours. I love you.

To everyone, thank you from my heart. To God be all the glory.

Table of Contents

List of Abbreviations

5FU	fluorouracil
AJCC	American Joint Committee on Cancer
BC	British Columbia
bft	*B. fragilis* toxin
BH	benjamini-hochberg
BP	biological process
CAC	colitis-associated cancer
CAF	cancer associated fibroblasts
cCR	clinical complete response
CIMP	CpG island methylator phenotype
CIN	chromosomal instability
CMS	consensus molecular subtype
CRC	colorectal cancer
cRT	chemoradiotherapy
DA	differentially abundant
DAG	directed acyclic graph
DEGs	differentially expressed genes
DGEA	differential gene expression analysis
dMMR	deficient mismatch repair
ds	double stranded
EMT	epithelial-to-mesenchymal transition
ETBF	enterotoxigenic *B. fragilis*
FAP	familial adenomatous polyposis
FDR	false discovery rate
FMT	fecal microbial transplantation
FOLFIRI	folinic acid, fluorouracil, irinotecan
FOLFOX	folinic acid, fluorouracil, oxaliplatin
GO	gene ontology
GOA	gene ontology annotation
GSEA	gene set enrichment analysis
HNPCC	hereditary non-polyposis colon cancer
ICIs	immune checkpoint inhibitors
IFN	interferon
IL-	interleukin
LCCRT	long-course chemoradiotherapy
lncRNAs	long non-coding RNAs
LPS	lipopolysaccharide
MF	molecular function

miRNA	microRNA
MMR	mismatch repair
MSS	microsatellite stable
MSI	microsatellite instability
MSigDB	molecular signatures database
nCRT	neoadjuvant chemoradiotherapy
NES	normalized enrichment score
NK	natural killer
NTBF	non-toxigenic *B. fragilis*
ORF	open reading frame
PAMPs	pathogen-associated molecular patterns
PAs	polyamines
PBMCs	peripheral blood mononuclear cells
PCNA	proliferating cell nuclear antigen
pCR	pathological complete response
pMMR	proficient mismatch repair
ROS	reactive oxygen species
RPK	reads-per-kilobase
SCFA	short chain fatty acid
SCNA	somatic copy number alterations
SCRT	short course radiotherapy
SMO	spermine oxidase
STAT-3	signal transducer and activator of transcription-3
STING	stimulator of interferon genes
TA	transit-amplifying
TAMs	tumor-associated macrophages
TILs	tumor infiltrating lymphocytes
TME	tumor microenvironment

List of Figures

Figures **Page No.**

Figures **Page No.**

List of Tables

Chapter 1:

Background

1.1 The Heterogeneity of Colorectal Cancer

Colorectal Cancer (CRC) is among the top causes of cancer mortality worldwide (Arnold et al., 2017), and the second highest cause of cancer deaths in New Zealand (*Bowel Cancer*, 2021). Several studies had been carried out to study the mechanisms of CRC progression. A widely accepted fact is that CRC is a highly heterogeneous disease that manifests differently from patient to patient, arising from different molecular backgrounds, and is affected by the tumor microenvironment, human genetics, and the interplay of both (Burns et al., 2018; Lawler et al., 2018). In this section I discuss the etiology of CRC and the pathways colonic epithelial cells undergo to become tumors, the subtypes of CRC and how they differ, and rectal cancers as opposed to colon cancers.

1.1.1 Etiology of Colorectal Cancer

1.1.1.1 Genetic Pathways of CRC Development

Fearon & Vogelstein (1990) first theorized that an epithelial cell undergoes a series of mutations as it follows the steps of tumor initiation, promotion, and progression to a malignancy. Initiation involves mutations that transform normal cells into hyperproliferative cells. Promotion and progression involves the clonal expansion of

these cells which is thought to give rise to adenomas, and finally, through further mutations, leads to carcinoma development. This, in turn, eventually leads to metastatic tumors (Carethers & Jung, 2015; Fearon & Vogelstein, 1990). This model proposes that an accumulation of mutations is necessary for an epithelial cell to undergo an adenoma-carcinoma sequence (Carethers & Jung, 2015; Fearon & Vogelstein, 1990; W. Wang et al., 2019). Here I describe common mutations in CRC development, and the three accepted pathways by which CRC develops.

1.1.1.1.1 Microsatellite Instability

Microsatellites are regions of repetitive DNA elements, mostly associated with polyadenine tracts in *Alu* sequences (Boland et al., 1998; Boland & Goel, 2010). Defects in DNA mismatch repair (MMR) genes result in accumulation of single nucleotide mutations in microsatellites which alter their lengths and lead to hypermutations in the tumors (Boland et al., 1998). Microsatellite instability (MSI) is assessed via a panel of biomarkers, comparing whether a tumor sample's microsatellite biomarkers have the same lengths (i.e. same number of repeats) as its respective normal tissue. A tumor is MSI-high if more than 30% of the panel shows instability; it is MSI-low if greater than 0% but less than 30% of the panel is unstable; and it is MSI-stable if none of the biomarkers show instability (Kohlmann & Gruber, 2018). Two examples of biomarker panels are the Bethesda panel, consisting of screens for MSI in the *BAT25, BAT26, D2S123, D5S346,* and *D17S250* loci; and a pentaplex panel consisting of screens for MSI at *BAT25, BAT26, NR21, NR24,* and *NR27* (Evrard et al., 2019).

One well known cause of MSI is the hypermethylation of the MMR gene *MLH1*, which prevents its expression. MMR genes serve to repair replication errors and their inactivity can lead to more than a 100-fold increase in mutation rate (Colussi et al., 2013). Aside from affecting CRC driver genes that contain microsatellites, frameshift mutations occurring within the genes result in truncated proteins that are neo-antigenic, and may trigger the immune system (Carethers & Jung, 2015). MSI tumors are frequently characterized by the BRAF V600E mutation, the activating oncogenic mutation of *BRAF*, a gene involved in responses to growth signals (Carethers & Jung, 2015; Colussi et al., 2013).

The clinical relevance of MSI tumors are also well established. They are mostly considered to have better prognosis (Colussi et al., 2013; Q. Wang et al., 2019), and are thought to be more responsive to immunotherapy (W. Wang et al., 2019).

1.1.1.1.2 Chromosomal Instability

The majority of CRC tumors exhibit chromosomal instability (CIN), and are characterized as “classical” types of cancer and non-hypermutated (Carethers & Jung, 2015; W. Wang et al., 2019). CIN is characterized by chromosomal aberrations, and high somatic copy number alterations (SCNA). The underlying molecular mechanisms driving CIN are still not fully understood. Most of the aberrant genes affected by CIN, including *APC, KRAS,* and *TP53*, are involved in cell cycle checkpoints, chromosome segregation, and sister chromatid cohesion. Mutations in the tumor suppressor *APC*

gene results in the failure to degrade β-catenin, which results in increased signals inducing cell proliferation; mutations in the *KRAS* gene converts it to a constitutively active state and the RAS growth factor activates a cascade that triggers the G1-S phase cell cycle transition; and *p53* loss of function allows for cell survival during stress caused by DNA damage, hypoxia, reduced nutrient access, and aneuploidy (Carethers & Jung, 2015; Colussi et al., 2013; Hagland et al., 2013; Inamura, 2018; Rodriguez-Salas et al., 2017).

1.1.1.1.3 CpG Island Methylator Phenotype

CpG Island Methylator Phenotype (CIMP) is characterized by hypermethylation of CpG islands typically located in promoters, leading to loss of gene expression (Colussi et al., 2013; W. Wang et al., 2019). Tumors with CIMP can be stratified as CIMP-high or CIMP-low depending on the frequency of CpG loci methylated (Carethers & Jung, 2015; Colussi et al., 2013; Inamura, 2018). CIMP-high tumors typically have the BRAF V600E mutation, and have hypermethylated *hMLHI* showing considerable overlap with MSI tumors (Carethers & Jung, 2015). *KRAS* mutations, meanwhile, are mutually exclusive with *BRAF* mutations and occur in CIMP-low tumors which are described as microsatellite stable (MSS) and non-hypermutable (Carethers & Jung, 2015; Colussi et al., 2013).

1.1.1.2 Environmental Risk Factors of CRC

Only a small percentage of CRCs have a familial background. Among the most common of these include hereditary non-polyposis colon cancer (HNPCC) and

familial adenomatous polyposis (FAP). Both syndromes signify a family history of early-onset CRC, with HNPCC exhibiting germline mutations in DNA mismatch repair enzymes, and FAP showing germline mutations in the tumor-suppressor *APC* gene (Hardy, 2000). Another common hereditary syndrome is Lynch syndrome, caused by mutations in DNA repair genes. These tumors are therefore microsatellite unstable (Kuipers et al., 2015).

The rest of CRC occurrences have a sporadic origin, affected by environmental factors. Smoking, alcohol intake, and obesity all increase the risk for CRC development. Meanwhile, physical activity and the use of NSAIDs that inhibit cyclooxygenase-2 both decrease CRC risk (Potter, 1999). Diet is also considered one of the most important exogenous factors that affect CRC development (De Almeida et al., 2019). For instance, red meat consumption is associated with blood-borne carcinogens and therefore CRC risk, while vegetables are considered good sources of folate, antioxidants, and detoxifying enzymes (Potter, 1999), and associated with decreased cancer risk. Moreover, obese conditions could be affected by diets, and obesity is associated with chronic inflammation increasing propensity for cancer development. On the other hand, short-term fasting is associated with decrease of regulatory T-cells and isocaloric diets are associated with increased cytotoxic CD8+ T-cells, which are associated with positive responses to immunotherapy (De Almeida et al., 2019). Lastly, diet is also associated with modulation of the gut microbiome (De Almeida et al., 2019) which, along with immunomodulation, will be discussed in this thesis as an important player in CRC development. These lifestyle factors, to an

extent, explain the geographical and socio-economical differences seen in CRC prevalence and incidence (Kuipers et al., 2015).

1.1.2 Subtypes of Colorectal Cancer

In the clinic, clinicopathological descriptions and mutation status may be used to determine the prognosis and diagnosis of patients. However, patients defined under these characteristics still differ in outcome (Rodriguez-Salas et al., 2017), and it is now thought that the heterogeneity between CRC tumors is better captured in transcriptomic data rather than mutations and clinical characteristics (W. Wang et al., 2019)**.** Several studies have resulted in different categories which CRC tumors can be classified into, based on transcriptomic profiles (Rodriguez-Salas et al., 2017; Singh et al., 2019; W. Wang et al., 2019).

These classifications, summarized in **Table 1.1** (De Sousa E Melo et al., 2013; Marisa et al., 2013; Roepman et al., 2014; Sadanandam et al., 2013), differ in stratifying tumors, where categories range from three to six groups among different studies. Criteria used to separate groups include mutation status of *KRAS, TP53*, and *BRAF*; MSI status; CpG island status; Wnt pathway activation; epithelial-to-mesenchymal transition (EMT) pathway activation; and immune system activation. Such studies have also explored the prognosis and treatment response of the subtypes highlighting the importance of subtyping to precision medicine.

Table 1.1 Comparisons of some CRC subtyping systems by publication, and their descriptions. Shaded rows are categories of each subtype that share similar characteristics.

Roepman et al 2014	*de Sousa et al 2013*	*Sadanandam 2013*	*Marisa et al 2013*	*Guinney et al 2015*
Type A: • MMR deficient • MSI • High mutation rate	**CCS2:** • MSI, CIMP phenotype • Inflammatory cell infiltration • Located in right colon	**Inflammatory:** • ↑ Interferon gene expression • ↑ cytokine gene expression • ↑ chemokine gene expression	**C2:** • MSI, CIMP • *BRAF* mutations • ↑ Immune, proliferative pathways • ↓ Wnt pathway activity	**CMS1 (Immune):** • MSI, CIMP • *BRAF* mutations • Immune infiltration
Type B: • Epithelial proliferative • MSS	**CCS1:** • *KRAS, TP53* mutations • Chromosomal Instability • ↑ Wnt Pathway activity	**Goblet like:** • ↑ *MUC2, TFF3* expression • ↓ Wnt and stem cell markers	**C1:** • Chromosomal Instability • *KRAS, TP53* mutations • ↓ Immune, EMT pathways activation	**CMS2 (Canonical):** • MSS • Chromosomal Instability • ↑ Wnt, MYC Pathway activation • Elevated *EGFR* with *TP53* mutations
		Transit-Amplifying (TA): • Heterogeneous group • Altered Wnt targets • Variable expression of stemness genes	**C3:** • MSS • *KRAS* mutations • ↓ Immune, EMT pathways activation	
		Enterocyte: • ↑ enterocyte cell gene expression	**C5:** • Chromosomal Instability • *KRAS, TP53* mutations • ↑ Wnt Pathway genes	**CMS3 (Metabolic):** • Mixed MSI • *KRAS* mutations • metabolic deregulation
Type C: • Mesenchymal • MSI • EMT expression	**CCS3:** • ↑ EMT genes • Matrix remodeling • Cell migration • *BRAF, P13CA* mutations	**Stem-like:** • ↑ Wnt Pathway genes • Mesenchymal stem cell features • ↓ differentiation markers expression	**C4:** • CIN, CIMP • *KRAS, BRAF, TP53* mutations • ↑ EMT pathways	**CMS4 (EMT-Mesenchymal):** • Mesenchymal • Angiogenic characteristics • EMT gene activation
			C6: • Chromosomal Instability • *KRAS, TP53* mutations • ↑ EMT genes • Serrated neoplasia pathway activation	

Legend:

↑ - upregulated/increased
↓ - downregulated/decreased
MSI – Microsatellite Instability
MMR – Mismatch Repair
MSS – Microsatellite Stable
EMT – Epithelial-Mesenchymal Transition
MUC2 – mucin 2
TFF3 – Trefoil Factor 3

As there are intrinsic differences among these classification efforts, the Colorectal Cancer Subtyping Consortium performed a meta-analysis of six subtyping systems to come up with four consensus molecular subtypes (CMS) (Guinney et al., 2015; W. Wang et al., 2019). **Figure 1.1** (Inamura, 2018), and **Table 1.1, fifth column**, summarizes the characteristics of these four subtypes. The four subtypes based on this meta-analysis are described as follows (Guinney et al., 2015): CMS1 tumors are usually microsatellite instability positive (MSI+), are hypermutated and hypermethylated, with frequent *BRAF* mutations, low prevalences of SCNA, and characterized by increased expression of immune infiltrates. CMS2, CMS3, and CMS4 are all chromosomally unstable (CIN). CMS2, known as the canonical subtype of CRC, has the highest copy number gains in oncogenes, as well as losses in tumor suppressor genes. CMS2 has upregulated WNT and MYC pathways, displays epithelial differentiation, and is characterized by upregulation of the microRNA (miRNA) cluster miR-17-92, which is a target of MYC. CMS3 has fewer SCNAs, characterized as CIMP-low, has *KRAS* mutations, some degree of hypermutation, and is enriched in multiple metabolism signatures. CMS4 is distinguished by upregulation of EMT pathway genes, activation of TGF-β signaling, angiogenesis, matrix remodeling, and complement-mediated inflammation.

Of the different subtyping systems that have been published by different groups, including the CMSs, two distinct subtype characteristics are commonly identified (Rodriguez-Salas et al., 2017). CMS1 (immune subtype) has common characteristics with CCS2 (De Sousa E Melo et al., 2013), Type A (Roepman et al., 2014), C2 (Marisa

et al., 2013), and Inflammatory subtypes; while CMS4 (EMT-mesenchymal) has common characteristics with CCS3 (De Sousa E Melo et al., 2013), Type C (Roepman et al., 2014), C4 and C6 (Marisa et al., 2013), and Stem-like (Sadanandam et al., 2013) subtypes. Shaded rows in **Table 1.1** summarize these similarities.

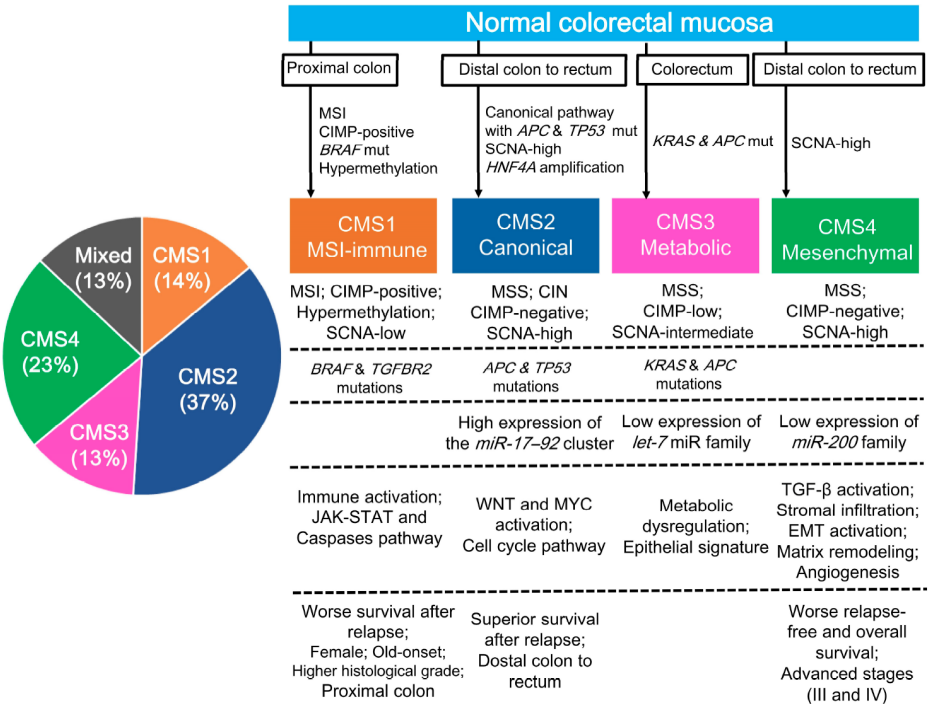


Figure 1.1. Summary characteristics of the four Consensus Molecular Subtypes (Inamura, 2018, under Creative Commons Attribution License). Illustration of the defining molecular characteristics, prevalence, location, and definitive prognosis of tumors under each consensus molecular subtype.

The heterogeneity between CRC tumors necessitates the use of personalized medicine for individual patient management. Prognostic and predictive biomarkers aid in treatment decisions, and therapeutic outcome predictions, respectively (Singh et al., 2019). Subtyping of CRCs could potentially aid in this. For instance, some of the older classification systems have been shown to predict survival and response to

drugs. However, the inconsistency between the different classifications hindered their clinical utility (W. Wang et al., 2019). The CMS consortium circumvents this, but there are still some gaps to be filled for CMS to be translated for prognostic or therapeutic decisions in the clinic. For instance, the context of the stage and sample source is important. The majority of samples used for the subtyping efforts were stage II and III tumors, and were primarily colon cancers as opposed to rectal cancers. Intra-tumoral heterogeneity also exists, and different subtyping results could arise depending on the area of the tumor sampled for subtyping; indeed a fraction of samples could be unclassified and this could be attributed to heterogeneity within the tumor itself (Fontana et al., 2019). It should also be taken into consideration that the classifications may not be static, and different treatments could change molecular phenotypes (Rodriguez-Salas et al., 2017). Other factors that could affect definitive subtyping are the method used for CMS classification, as two different algorithms are available, the platform for transcriptomic data collection such as RNA sequencing versus microarrays, or assays to quantify gene expression and the gene sets used to classify the tumors (Fontana et al., 2019).

1.1.3 Rectal Cancers

Rectal cancers account for 30% or about a third of CRCs, and are associated with worse clinical outcomes (Dayde et al., 2017; Tamas et al., 2015). These tumors are histologically similar to cancers in other parts of the colon (Feeney et al., 2019), but their location in the narrow pelvis and being surrounded by vital structures require a more aggressive local treatment (Tamas et al., 2015). Molecular characteristics

among the sites of CRC also differ with rectal cancers being more similar to distal colon carcinomas. Proximal tumors are more hypermutated, with incidence of MSI-high, *BRAF* mutations, and CIMP-high tumors gradually decreasing from proximal to rectal sites (Tamas et al., 2015)**.** Furthermore, it has been shown that mutational frequencies of *PIK3CA, CTNNB1, ATM*, and *PTEN* also decrease from proximal to rectal locations, while *TP53* and *APC* mutational frequencies increase (Salem et al., 2017). These molecular differences among CRC sites may have to be taken into account as they can affect therapeutic strategies (Salem et al., 2017; Tamas et al., 2015). Rectal cancers usually require total mesorectal excision, preceded by radiotherapy or chemoradiotherapy.

1.2 Therapy in Colorectal Cancer

1.2.1 Surgery and Chemotherapy

Current CRC treatment involves complete removal of the tumor from primary and metastatic sites, usually involving surgical resection (Xie et al., 2020). Early stage cancers may be treated with different surgical approaches such as polypectomy and laparoscopic surgery (*Colon Cancer - Diagnosis and Treatment - Mayo Clinic*, n.d.; Granados-Romero et al., 2017). Cancers that have spread may need chemotherapy - the use of drugs to destroy cancer cells - to shrink tumors and suppress tumor growth. Common drugs used include the fluoropyrimidines 5-FU and capecitabine, oxaliplatin, and irinotecan, either alone or in combination (Xie et al., 2020). Chemotherapy may also be used as an adjuvant to surgery to help shrink tumors

prior to the procedure (*Colon Cancer - Diagnosis and Treatment - Mayo Clinic*, n.d.; Xie et al., 2020).

1.2.2 Immunotherapy

The immune response in CRC can both constrain and promote carcinogenesis. Immunosurveillance is a constant occurrence where the immune system continues to patrol the body for invading microbes and aberrant cells (Burkholder et al., 2014). Activation of T-cells and cytotoxic lymphocytes by antigen presentation is important in tumor elimination (Burkholder et al., 2014) and may be the reason why T-cell infiltration is associated with favorable outcomes in cancer (Ganesh et al., 2019). Overtime, this elimination phase gets replaced by the equilibrium phase where tumor cells coexist with immune responses in a balanced state. This further leads to the escape phase where tumors are able to escape anti-tumor immune responses, facilitate immunosuppression, and lead to tumor progression (Burkholder et al., 2014; Zaborowski et al., 2021). Prolonged antigen exposure and T-cell receptor signaling results in expression of checkpoints that regulate immune responses, resulting in an exhausted T-cell state (Zaborowski et al., 2021) .

Researchers responsible for the development of immunotherapy in cancer studies won the Nobel prize in medicine in 2018. They discovered that the blocking of CTLA-4 and PD-1, two immune checkpoints on the surfaces of T-cells, releases a brake that allows the immune system to attack cancer cells (Z. S. Guo, 2018). This further

recognizes the importance of immune responses on cancer development, progression, and treatment.

However, immunotherapy only works for a certain subset of CRCs, namely those with a high mutational burden that have mismatch repair deficiencies. This high mutational burden is thought to result in more peptide neoantigens that are recognized as foreign by immune cells (Ganesh et al., 2019).

1.2.3 Radiotherapy

As mentioned above, rectal cancers are a subset of colorectal cancers that are unique in their location. As such, the standard therapy for rectal cancers is surgery, pre-treated with chemoradiotherapy (cRT). cRT serves to downstage or downgrade tumors and is beneficial to surgery. Neoadjuvant cRT is usually followed by total mesorectal excision, which improves resectability, anal sphincter preservation, and local control (Dayde et al., 2017). Around 8-20% of patients exhibit a pathological complete response and for this subset, further surgery confers no other advantage other than confirmation of a complete response. As surgery brings with it side effects and potential morbidity, it has been found that a watch-and-wait strategy consisting only of clinical, endoscopic, and radiologic follow-up of patients with complete response to CRT is enough for excellent long term results (Habr-Gama et al., 2004). Biomarkers that could stratify patients that would respond well to radiotherapy would therefore be beneficial for a more confident approach to the watch-and-wait strategy. However, to date, no biomarker has been identified for use in the clinic. One

reason contributing to this difficulty may include the heterogeneity of CRCs. Therefore, aside from identifying biomarkers, the mechanistic link of these markers to tumor biology is needed (Dayde et al., 2017)

1.3 The Microbiome in Colorectal Cancer

The majority (more than two-thirds) of CRCs are sporadic (Carethers & Jung, 2015), suggesting a large role for environmental factors in their development, not the least of which, is the microbiome. The role of the microbiome in the progression of various diseases, including CRC, has been gaining traction. There is a growing consensus that there is no singular definition of what constitutes a healthy microbiome (Eisenstein, 2020; Karamalegos et al., 2020; Lloyd-Price et al., 2019). This contributes further to the heterogeneity found in CRC. Factors such as birthing method, early environmental exposure - which may be affected by geography or even rural versus urban conditions, lifestyle, and diet all affect the development of a person's microbiome growing up, until they reach a certain age where microbial composition is relatively stable. As such, the healthy microbiome of one person may not be healthy for somebody else (Eisenstein, 2020). An alternate view of this is the presence of a healthy functional core instead of a taxonomic core, where core functions are necessary for a healthy microbiome, but which is not necessarily provided for by the same organisms in different persons (Lloyd-Price et al., 2016). This suggests the importance of function in microbiota and not just taxonomic identity. To date, there is no single species that has been universally attributed to CRC development (Loftus et al., 2021), and even *Fusobacterium nucleatum*, the most

characterized microbe in CRC, is not found in every instance of CRC (Burns et al., 2018).

The microbiome's role in CRC progression has long been hypothesized. As far back as 1974, it has been shown that germ-free rats had lesser tumor susceptibility than conventional rats upon introduction of 1,2-dimethylhydrazine (Reddy et al., 1974). More recently, it has been seen that, with carcinogen introduction, mice gavaged with fecal microbiota from patients with CRC develop more polyps, dysplasia, and inflammation markers compared to mice fed with microbiota from normal patients (Wong et al., 2017). In 2020, a study showed that mice with different microbiota populations also showed different tumor susceptibilities (A. I. Yu et al., 2020). These studies, among others, indicate that microbes influence the course of carcinogenesis in the colon.

1.3.1 Abundant Microbes in Colorectal Cancer Disease

Several studies have compared microbial abundance in CRC compared to normal controls. The *Firmicutes* and *Fusobacteria* phyla were found to be enriched in CRC, while *Proteobacteria* was underrepresented. It was further identified that the *Fusobacterium, Prevotella, and Peptostreptococcus* genera were among those that contribute to the dysbiosis of the microbiota of CRC patients (Gao et al., 2015). A multi-cohort analysis from geographically distinct populations identified *Fusobacterium nucleatum*, *Bacteroides fragilis, Porphyromonas asaccharolytica, Parvimonas micra, Prevotella intermedia, Alistipes finegoldii,* and *Themanaerovibrio*

acidaminivorans as enriched in CRC fecal metagenomic samples (Dai et al., 2018) . They were able to use these microbes to successfully differentiate between cases and controls in the different cohorts, as well as to identify early stage CRCs.

Two recent studies (Thomas et al., 2019; Wirbel et al., 2019) performed meta-analyses on several different heterogeneous cohorts of CRC. One showed higher species richness in CRC samples compared to controls, which may be explained by an influx of species from the oral cavity (Thomas et al., 2019). Oral pathogens have also been found among microbes associated with CRC tumors (Loftus et al., 2021; Nakatsu et al., 2015; Purcell, Visnovska, et al., 2017). Furthermore, *F. nucleatum, S. moorei, P. asaccharolytica, P. micra,* and *P. stomatis* have been identified as biomarkers of CRC (Thomas et al., 2019). Meanwhile, the genera *Fusobacterium, Porphyromonas, Parvimonas, Peptostreptococcus, Gemella, Prevotella, and Solobacterium* have also been identified as associated with CRC (Wirbel et al., 2019).

We can see from each of these studies that common microorganisms that are enriched with CRC compared to normal controls include *Fusobacterium, Porphyromonas, Prevotella,* and *Peptostreptococcus*.

1.3.2 Microbial Mechanisms in Colorectal Carcinogenesis

Microbes enriched in CRC may be drivers (those that directly affect carcinogenesis) or passengers (those that thrive in the microenvironment brought about by

carcinogenesis) (Coleman & Nunes, 2016; Saus et al., 2019; Tjalsma et al., 2012; Wong & Yu, 2019). Some drivers promote carcinogenesis by changing the microenvironment of the cells. For instance, the presence of microbes can influence immune processes and promote chronic inflammation, which may lead to epithelial damage, high levels of reactive oxygen species (ROS), induction of stress response pathways and oncogene expression, as well as down-regulation of tumor suppressor genes (Coleman & Nunes, 2016; Goodwin et al., 2011; X. Ye et al., 2017). *F. nucleatum* subspecies *animalis* induces production of the chemokine CCL20 in CRC cells, as well as THP-1 monocyte activation and migration. CCL20, and its receptor CCR6 play roles in recruitment of Th17 cells, regulatory T-cells, and dendritic cells; it is stimulated by lipopolysaccharides and tumor necrosis factors (X. Ye et al., 2017). Enterotoxigenic strains of another CRC-associated bacterium, *B. fragilis* is found to be associated with colon tumor formation in mice through the induction of STAT-3 (signal transducer and activator of transcription-3) activation with a Th17 response involving IL-17 and IL-23 (Wu et al., 2009).

Some microbial functional pathways have also been identified as being enriched in CRC. This includes the TCA cycle, LPS biosynthesis, ubiquinone and other terpenoid-quinone biosynthesis, lipoic acid metabolism, valine, leucine and isoleucine degradation, phosphate and phosphonate metabolism, and other glycan degradation (Dai et al., 2018); pathways involved in gluconeogenesis, and uptake and metabolism of amino acids via putrefaction and fermentation pathways including those that can convert amino acids to tumor-promoting compounds (Thomas et al., 2019); and pathways for degradation of amino acids, mucins, and organic acids

(Wirbel et al., 2019). Conversely, a depletion of bacterial genes for carbohydrate degradation is found in CRC samples, indicating a shift from dietary carbohydrate utilization to amino acid degradation (Wirbel et al., 2019).

Other bacteria, meanwhile, directly affect carcinogenesis through the products they secrete. There are strains of *E. coli* known to have genomic islands called *pks* islands that produce colibactin, a peptide-polyketide genotoxin known to induce DNA breakages (Cuevas-Ramos et al., 2010). *Pks+ E. coli* infection of epithelial cell lines induced impaired cell division with incomplete DNA repair, anaphase bridges, and chromosome aberrations, including aneuploidy and polyploidy (Coleman & Nunes, 2016; Cuevas-Ramos et al., 2010; Fulbright et al., 2017). Aside from the immune processes *B. fragilis* affects, its enterotoxigenic strain*, ETBF,* produces a metalloprotease toxin that works to induce an upregulation of spermine oxidase (SMO). A by-product of the spermine to spermidine conversion is H_2O_2, a ROS that leads to DNA damage (Goodwin et al., 2011). Furthermore, the *ETBF* toxin also disrupts e-cadherin, resulting in β-catenin pathway activation, increased *c-Myc* expression, and cell proliferation (Goodwin et al., 2011; Wu et al., 2009). *Streptococcus gallolyticus* subspecies *gallolyticus* has been shown to increase cell proliferation, β-catenin, *c-Myc*, and proliferating cell nuclear antigen (PCNA) levels (Kumar et al., 2017). *F. nucleatum,* one of the bacteria most commonly associated with CRC, expresses FadA adhesins that bind to E-cadherin, resulting in the blockage of its tumor suppressor activity and thereby activating the β-catenin pathway that leads to an increase of oncogenic factors in human CRC cell lines. This leads to the

stimulation of the expression of inflammatory elements such as NF-κB and cytokines (IL-6, IL-8, IL-18), and promotes cell proliferation (Rubinstein et al., 2013).

1.3.3 The Contribution of Microbes to Cancer Therapy

Microbes and their by-products, as discussed above, affect CRC progression. They may also influence the therapies used to treat CRC, as they do for other types of cancers. For example, it has been found that *Bifidobacterium longum, Collinsella aerofaciencs,* and *Enterococcus faecium* were more abundant in responders to anti-PD-1 immunotherapy in metastatic melanoma (Matson et al., 2018). Meanwhile, an increased relative abundance of the *Ruminococcaceae* family has been found in patients responding to immune checkpoint inhibitors (ICIs), also in melanoma patients (Gopalakrishnan et al., 2018). In epithelial tumors, it has been shown that *Akkermansia municiphila* correlates with clinical response to ICIs, and that antibiotics can induce resistance to ICIs (Routy et al., 2018). *Bacteroides* species have immunostimulatory effects in CTLA-4 blockade, evidenced in fecal microbial transplantation (FMT) from human to mice in melanoma studies (Vétizou et al., 2015). Germ-free mice, treated with FMT from responding patients, also showed improved response to PD-1-based immunotherapy for melanoma (Gopalakrishnan et al., 2018) and epithelial tumors (Routy et al., 2018). FMT, which involves the transfer of lyophilized and encapsulated feces from donor to recipient, orally or rectally, is a way of re-shaping the microbiome, and has supporting clinical evidence for its use, with human clinical trials currently underway (Daillère et al., 2020; Park et al., 2020).

In radiotherapy of melanoma and lung cancer mouse models, vancomycin reduction of gram-positive bacteria, and their corresponding short chain fatty acid (SCFA) products, has been shown to enhance response to the aforementioned treatment (Uribe-Herranz et al., 2020).

1.4 Computational Tools in Microbiome Studies

Microbiome studies usually involve determining which microbes are present (taxonomic identification) in an environment or specific population, and their effects on the respective environments (functional analysis). There are several tools available for microbial taxonomic identification from sequencing reads. These can be based on nucleotide sequences (e.g. Kraken, Kraken2, Clarke), translated nucleotide sequences (e.g. Kaiju, Diamond), or marker-gene nucleotide sequences (e.g. MetaPhlan2). Benchmarking efforts by S. H. Ye et al., 2019, concluded that these classifiers performed similarly, but nucleotide-based classifiers had better estimates probably due to absence of non-coding sequences in databases of translated nucleotide-based classifiers, while marker-gene based classifiers had lower performances based on precision and recall because of database limitations. Database composition, therefore, has a large effect on performance differences between classifiers. This should then be taken into consideration alongside other factors such as computational resources, and ease of use.

A more complete picture of interactions driving community dynamics can be gleaned if functional contributions of the microbiome can be determined (Langille, 2018).

Some of the more widely used packages used to identify microbial function include PICRUSt and PICRUSt2 (Douglas et al., 2019; Langille et al., 2013), which uses 16S gene marker sequences to predict function based on an OTU's phylogenetic similarity to other OTUs with known gene content, and HUMAnN2 (Franzosa et al., 2018) whose taxonomic profiling relies on MetaPhlan2 (Segata et al., 2012; Truong et al., 2015), which is limited due to its marker-gene based and non-customizable databases. More recently, SAMSA2 (Westreich et al., 2018), was developed. SAMSA2 uses NCBI reference databases, and SEED subsystems for taxonomic and functional annotations.

Comparing microbiome contents between populations is another common aspect of microbiome studies. Microbiome datasets are described as sparse (having many zeroes) and compositional (Fernandes et al., 2014; Gloor et al., 2017; Gloor, Wu, et al., 2016; Gloor & Reid, 2016; Weiss et al., 2017). Tools such as ALDEx2 (Fernandes et al., 2013, 2014; Gloor, Macklaim, et al., 2016) and ANCOM (Mandal et al., 2015) have been developed specifically for microbiome datasets. However, these tools do not consistently outperform other methods (Calgaro et al., 2020) such as those developed for RNA-seq data (e.g. edgeR (Robinson et al., 2010) and DESeq2 (Love et al., 2014)), especially as the normalization factors of these methods are mathematically similar to centered log ratios proposed in compositional methodologies (Calgaro et al., 2020). Indeed, edgeR has been cited as among tools able to distinguish true differentially abundant OTUs in 16S datasets although this is offset by higher false positive rates (Thorsen et al., 2016). DESEq2, meanwhile, has

been found to perform consistently in terms of false positive rates, concordance, power, and computational time (Calgaro et al., 2020).

1.5 Summary and Insights

Colorectal cancer is a highly heterogeneous disease that manifests differently from patient to patient. There are different pathways to colorectal carcinogenesis, the most widely accepted ones being microsatellite instability (MSI), Chromosomal Instability (CIN), and CpG Island Methylator phenotype (CIMP). Different locations - proximal, distal, or rectal - of colorectal cancer tumors also affect their characteristics. CRCs have also been categorized into subtypes, having similar molecular characteristics within a subtype. This will hopefully aid in the categorization of patients in terms of diagnosis, prognosis, and treatments. It has also been shown that the microbiome plays an important role in colorectal carcinogenesis, affecting not only its development, but also therapeutics involved in disease management.

Previous studies have looked at correlations between microbes and tumor staging (Nakatsu et al., 2015), loss of function mutations (Burns et al., 2018), and even immune cell markers (Cremonesi et al., 2018). However, I found the need to augment studies of CRC microbiome with corresponding host gene expression. This is important as treatment strategies depend on the genetic background of CRC tumors, and the heterogeneity of this background may be affected by the microbiome associated with it. In line with this, I focused on studying the microbiome of CRC in

the context of the consensus molecular subtypes, as well as the microbiome and host gene expression in radiotherapy of rectal cancers, as an important subset of CRC.

In this thesis, I aimed to determine host gene and microbiome taxonomic and functional contributions to different cases of colorectal cancer. Below, I discuss the specific rationales behind the methods and goals of this thesis, and in the next section, I summarize the objectives of this study based on these rationales.

1.5.1 Computational Pipeline for Host - Microbiome Studies

Various computational tools have been developed to study the microbiome, but a consensus on a gold standard tool does not exist, at the time of writing. Many publications (Calgaro et al., 2020; Russel et al., 2018; S. H. Ye et al., 2019) recommend a case-by-case inspection of the data to determine the most appropriate tools to use.

As a result, I developed a computational pipeline not only tailored to the needs of this study, but also to address critical gaps in microbial sequencing studies. Firstly, I looked at the need for methods to ascribe function to microbiome datasets and associate these functions to taxa in the community. While there are tools available to profile functions in microbiome data, most rely on only limited databases, and/or do not link functions to specific taxa. To identify microbial taxonomic IDs and provide protein translation matches of our reads in the form of accession numbers I used

Kaiju (Menzel et al., 2016). I utilized its protein translation results to mine Gene Ontologies annotating these accessions. I hypothesized that Kaiju, with its more inclusive database, its ability to link protein products to taxonomic IDs, and relatively faster computational turn out, would offer optimal accuracy, speed, and precision for microbial and functional identification. In addition, providing this pipeline as a single workflow simplifies and streamlines analyses of microbial datasets, with integrated steps of quality control, human gene mapping, microbial taxa and function identification, differential expression or differential abundance analysis, human gene set enrichment analysis, and correlation analyses to test for possible dependence of microbiome content and human gene abundances. Here, I show how the pipeline is applied to data from different subtypes of colorectal cancer and different responders to radiotherapy in rectal cancer patients.

1.5.2 Comprehensive Studies on Microbiota Contributions

Instead of a core set of taxa defining a healthy microbiome, it is thought that it may be more useful to define a healthy functional core, even if these functions are performed by different organisms in different people (Lloyd-Price et al., 2016). This underlines the importance of function in microbial contributions to the host.

Most studies of CRC either catalog microbiome composition, or more recently, function, of microbiome communities in patients, or thoroughly investigate a particular microbe and their singular contributions to CRC progression (Alexander et al., 2018). However, these studies show that no singular species may be universally

attributed to CRC development and specific interactions between microbiomes and their respectives host may come into play. While there are published studies looking into microbial composition and function in combined meta-analyses of microbiomes (Dai et al., 2018; Thomas et al., 2019; Wirbel et al., 2019), these do not take host genetics into account, which may be an important part of the complex interactions taking place in the human gut of CRC patients. Other efforts had been made to identify network correlations between enriched and depleted bacteria in CRC (Dai et al., 2018), and characterize tumor-microbe interactions in disease stages (Nakatsu et al., 2015) and tumor-associated mutations (Burns et al., 2018), but how these interactions affect distinct CRC pathways are not clear. I hypothesized that the microbiome content might affect CRC progression differently among molecular subtypes. Gaining quantitative insight into this would augment our understanding of how CRC could progress through these pathways. I aimed to expand on results of a previous study (Purcell, Visnovska, et al., 2017) associating specific microbiomes with CRC progression by determining enriched members of the microbial communities to the molecular profiles of the CMS subtypes. Furthermore, I also wished to explore the contributions of host genetics and microbial composition and function to CRC therapeutics, focusing on rectal cancer and radiotherapy.

1.6 Study Objectives

The overarching goal of this thesis is to investigate how microbes contribute to differing characteristics in colorectal cancer. I aimed to answer how the microbes, host genes, and CRC subtypes or response to therapy relate to one another. To

address these issues, I designed a computational pipeline, called MetaFunc, that can facilitate host gene analysis, identify microbe taxonomies, and perform a survey of microbiome function. The creation of the pipeline, along with documentation for its public use is described in Chapter 2 of this document. I used this pipeline for the analysis of host gene expression in different subtypes of CRC, and the possible roles of their respective microbiomes in the different subtypes' characteristics. The results of these analyses are described and discussed in Chapter 3. There, I focused on CMS1 and CMS4 as two immune-rich subsets of CRC, and how different bacteria in each of the subtypes could contribute to the different characteristics seen in the two subtypes. Finally, in Chapter 4, I used MetaFunc on a rectal cancer cohort of CRC to investigate how host genes and microbes differ in different responders to radiotherapy. It is the aim of this thesis to contribute to knowledge of how microbes in CRC affect the tumors' heterogeneous characteristics seen in patients. In Chapter 5, I briefly summarize the novel findings and contributions of this thesis, outline the study limitations, and give recommendations for future research.

Chapter 2:

MetaFunc: Taxonomic and Functional Analyses of High Throughput Sequencing for Microbiomes

Sulit, A.K.[1], Kolisnik, T[1]., Frizelle, F.A. [2], Purcell, R. [2],Schmeier, S. [1,3]

[1]School of Natural Sciences, Massey University, Auckland, New Zealand
[2]Department of Surgery, University of Otago, Christchurch, New Zealand
[3]Evotec SE, Hamburg, Germany

Article in preprint (doi: 10.1101/2020.09.02.271098)

Author contributions:

AKS and **SS** developed and co-wrote the pipeline which ultimately led to MetaFunc, and were involved with the majority of the design. **TK** developed the shiny application that is integrated in the pipeline. **AKS** wrote the manuscript with editorial input from **TK**, **SS** and **RP**. **RP** further contributed to the design of the pipeline. **FAF** provided guidance about all clinical aspects of the manuscript.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Arielle Kae L. Sulit
Name/title of Primary Supervisor:	Dr. Olin Silander
In which chapter is the manuscript /published work:	2

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

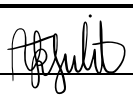
○ The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

◉ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	
Date:	16-Nov-2021
Primary Supervisor's Signature:	
Date:	16 Nov 2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

2.1 Abstract

Background: The identification of functional processes taking place in microbiome communities augment traditional microbiome taxonomic studies, giving a more complete picture of interactions taking place within the community. While there are applications that perform functional annotation on metagenomes or metatranscriptomes, very few of these are able to link taxonomic identity to function and are limited by their input types or databases used.

Results: Here we present MetaFunc, a workflow which takes input reads, and from these 1) identifies species present in the microbiome sample and 2) provides gene ontology (GO) annotations associated with the species identified. MetaFunc can also provide a differential abundance analysis step comparing species between sample conditions. In addition, MetaFunc allows mapping of reads to a host genome, and separates these reads, before proceeding with the microbiome analyses. From the host reads, MetaFunc is able to identify host genes, perform differential abundance analysis, and gene-set enrichment analysis. A final correlation analysis between microbial species and host genes can also be performed. Finally, MetaFunc builds an R shiny application that allows users to view and interact with the microbiome results. In this paper we showed how MetaFunc can be applied to metatranscriptomic datasets of colorectal cancer.

Conclusion: MetaFunc is a one-stop shop microbiome analysis pipeline that can identify taxonomies and their respective functional contributions in a microbiome

sample through GO annotations. It can also analyze host reads in a microbiome sample, providing information on host gene expression, and allowing for correlations between the microbiome and host genes. MetaFunc comes with a user-friendly R shiny application that allows for easier visualization and exploration of its results. MetaFunc is freely available through https://gitlab.com/schmeierlab/workflows/metafunc.git.

Keywords: *metatranscriptomics, microbiome, functional annotation, host correlation*

2.2 Background

Metagenomic or metatranscriptomic studies of microbiome communities allow for characterization of functional contributions as well as taxonomic load, by allowing the identification and quantification of genes possibly contributed by the microbial community. The ability to identify functional processes from the microbiome gives a more complete picture of microbe-microbe and/or microbe-host interactions that drive community dynamics (Langille, 2018).

There are existing bioinformatics programs (Nayfach et al., 2015; Sharma et al., 2015; Silva et al., 2016) that perform functional annotation on metagenomes and metatranscriptomes, but most of these are unable to link taxonomies (the microbes under study) to their respective functional processes. Existing packages with this capacity include PICRUSt and PICRUSt2 (Douglas et al., 2019; Langille et al., 2013), and HUMAnN2 (Franzosa et al., 2018). PICRUSt and PICRUSt2 predict metagenome function by inferring genes present in OTUs based on their phylogenetic similarities

to other OTUs with known gene content (Douglas et al., 2019; Langille et al., 2013). However, they do not directly measure the genes involved, but rather rely on 16S rRNA gene marker sequences, which, being highly conserved, are useful for the identification of bacterial genera (Bashiardes et al., 2016; Ternes et al., 2020) and are not present in other microbes aside from Bacteria and Archaea (S. H. Ye et al., 2019). Thus, 16S based taxonomic identification, and subsequent functional predictions, may be unsuitable for species level identification, and for recognizing other microbes aside from Bacteria and Archaea. HUMAnN2's taxonomic profiling, meanwhile, is reliant on MetaPhlAn2 (Segata et al., 2012; Truong et al., 2015), which uses clade-specific marker genes from reference genomes. Benchmarking efforts by Ye et al., 2019 highlights the limitations of using the MetaPhlAn2 package that includes non-customizable databases and marker gene-based databases, which results in relatively lower precision and recall in its classification.

To augment such meta-omic studies, we present here a simple, straight-forward pipeline named MetaFunc, a snakemake workflow (Köster & Rahmann, 2012) that maps function to a microbiome (and optionally host) sample. MetaFunc uses Kaiju (Menzel et al., 2016) as its main taxonomic classifier. Kaiju uses protein translations of input reads to generate taxonomic profiles. By generating protein-based classifications using metatranscriptomic reads, MetaFunc is able to determine which microbes are more metabolically active, allowing more focus on the functional contributions of microbes. MetaFunc then uses protein accession numbers from Kaiju results to obtain the set of Gene Ontology (GO) terms associated with the microbiome community. Furthermore, Kaiju outputs provide a direct protein –

taxonomy ID relationship that makes it possible for MetaFunc to establish which organisms are contributing to the functional GO terms. MetaFunc also has options for pre-processing of reads before running Kaiju: trimming of input reads with fastp (S. Chen et al., 2018) can be performed in addition to pre-mapping to a host genome (e.g. human) using STAR (Dobin et al., 2013). The unmapped reads following STAR processing are the input used by MetaFunc for microbe identification, while host gene expression information can be obtained from STAR-mapped reads. Thus, MetaFunc allows simultaneous investigation of host and microbe community active functional processes, as well as active host genes and microbes.

2.3 Implementation

2.3.1 Workflow

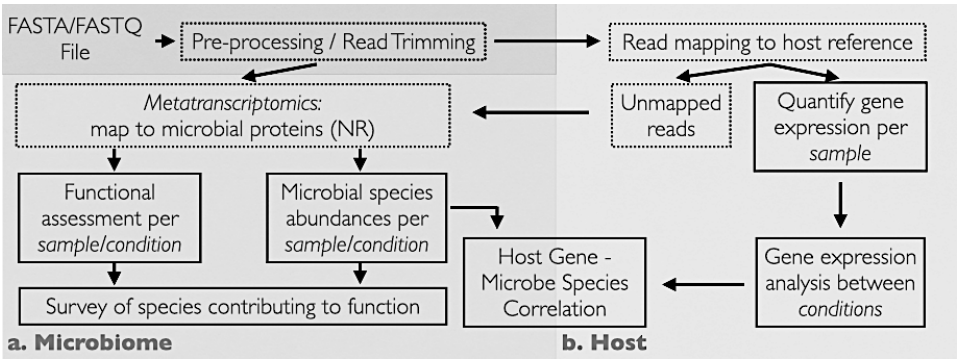


Figure 2.1. Illustration of the MetaFunc workflow. The workflow uses FASTQ or FASTA as input and processes reads through the **(a)** microbiome pipeline to give microbial abundance and function and/or (**b**) host gene analysis pipeline which will first map reads to a host before sending unmapped reads to the microbiome pipeline. Applying host read analysis will give gene expression analysis results as well as host gene-microbial species correlation. Reads may also be pre-processed for quality control before the rest of the analyses. Solid boxes indicate steps with an output while dotted boxes indicate intermediate steps in the pipeline.

NR: NCBI Blast nr database

Figure 2.1 shows the workflow that takes place within MetaFunc. Paired-end and/or single-end sequencing reads are used as input in fasta or fastq format. If trimming and mapping are not enabled, reads are used as input to Kaiju and subsequent microbiome analyses (**Figure 2.1a**). If trimming is enabled, reads are trimmed for adapters and quality controls using fastp. If mapping is enabled, either the trimmed reads or raw input reads are first mapped to a designated host genome using STAR. Unmapped reads after host mapping are then used as input to Kaiju. STAR results are then used to obtain host gene information (**Figure 2.1b**).

2.3.2 Microbiome Analysis (**Figure 2.1a**)

MetaFunc parses through Kaiju results and gathers taxonomy IDs of species for taxonomic characterization per sample and their corresponding protein accession numbers, which are subsequently annotated with Gene Ontology terms.

2.3.2.1 Taxonomic Identification

Each read matches to a taxonomy ID in Kaiju. MetaFunc gathers the species level matches and adds up the raw reads matching to each species taxonomy ID. In cases of strain level identification, MetaFunc adds this count to its parent species. It also obtains scaled read counts in percentages by dividing the final read count of each taxonomy ID by the total reads that have mapped to species level taxonomies. For a dataset, the pipeline removes any taxonomy ID that is less than 0.001% in abundance in all samples of the dataset; this filter removes thousands of species that are likely

to be false positives while retaining more confident classifications. Any remaining false classifications are thought not to affect downstream analyses, as the levels would be too low to impact true abundance (S. H. Ye et al., 2019), however, this value can be adjusted by the user. The taxonomy IDs that have passed the cutoff are then used in subsequent analyses. It should be noted that the pipeline still uses the original scaled percent abundances even after filtering. The pipeline would also include the lineage of the taxonomies using TaxonKit (Shen & Ren, 2021) in its output.

For a dataset, the MetaFunc pipeline outputs two tables containing species as rows and samples as columns with values being raw read counts or percent abundance for each species in the samples. If the user wishes to compare groups or conditions (e.g. disease state vs control), the pipeline calculates the average percent abundance of species among samples belonging to a group and this table is also given as an output. Differential abundance of microbes between groups is also carried out in MetaFunc using edgeR (McCarthy et al., 2012; Robinson et al., 2010). Raw read count tables are first filtered using the function filterbyExpr with threshold of 1 which is user-adjustable, and normalization factors are calculated by calcNormFactors with default settings. exactTest is then applied to calculate differential abundance with p-values adjusted using Benjamini & Hochberg correction or False Discovery Rate (FDR).

2.3.2.2 Protein Profiling

Kaiju outputs the accession number(s) of the protein match(es) with the highest BLOSUM62 alignment score of the read after translation into 6 open reading frames (ORF). It is possible to have more than one best protein match if two or more protein matches have equal scores in Kaiju. In order to account for this, we use proportional read counts per protein accession number where one read is divided by the number of best protein matches it has. Similar to that for taxonomy IDs, the pipeline adds up the proportional read counts per protein accession number of a species. Scaled reads as percent abundances are obtained by dividing the proportional count of each accession number by the total read counts that have mapped to a species.

2.3.2.3 Gene Ontology: Database Construction

Metafunc relies on Kaiju's nr_euk database for its taxonomic identification and corresponding protein matches. The nr_euk database is built on a subset from NCBI BLAST *nr* database containing Archaea, Bacteria, Fungi, Viruses, and other Microbial Eukaryotes (see https://raw.githubusercontent.com/bioinformatics-centre/kaiju/master/util/kaiju-taxonlistEuk.tsv). Identical sequences in the *nr* database are compiled into 1 entry and Kaiju only outputs the first protein accession number of an entry that has multiple identical sequences (Menzel et al., 2016). Thus we needed to construct the protein-to-GO database such that all functional terms of any protein compiled in 1 *nr* entry are considered.

To facilitate Gene Ontology annotations, we constructed an sqlite database in which GO annotations of a protein accession number from Kaiju can be looked up. We first gathered relevant NCBI *nr* database entries, converted all of the proteins of an *nr* entry into UniProt (Huang et al., 2011; The UniProt Consortium, 2017) entries, and then gathered corresponding GO annotations using the Gene Ontology Annotation (GOA) database for all those proteins (Camon et al., 2004). All GO annotations of one *nr* entry are then linked to the first protein of that entry in an sqlite database, which is used to annotate Kaiju protein accession matches with GO IDs. For more detailed information, please see the Notes section of the pipeline's documentation page (https://metafunc.readthedocs.io/en/latest/notes.html). For MetaFunc, we provide pre-made databases for download (A. K. Sulit et al., 2021a, 2021b) but users can make their own updated databases following instructions from https://gitlab.com/schmeierlab/metafunc/metafunc-nrgo.git.

2.3.2.4 Gene Ontology: Protein Annotation

For each sample, the pipeline obtains only the proteins that are from taxonomy IDs that passed cutoffs in the *Taxonomic Identification* section described above. Their scaled proportional read counts, as in the *Protein Profiling* section above, are still scaled against the total number of reads that mapped to a species. In order to compare groups or conditions, the pipeline first calculates the average of the corresponding proportional reads and scaled proportional reads of a protein accession number among samples of a group. It then searches for the GO terms annotating the (*nr*) protein using the created sqlite database described in *Gene*

Ontology: Database Construction. Each GO term set annotating an accession number is then updated by accessing parent terms related to the GO terms by *'is_a'* or *'part_of'* using *GOATOOLS* (Klopfenstein et al., 2018). Note that this update takes the entire set of GOs annotating the accession number into consideration such that no GO terms or path/s to the top of the GO directed acyclic graph (DAG) is doubled. GOATOOLS also parses other information regarding the GO term such as description, namespace, and depth through the go-basic.obo file (Ashburner et al., 2000). The proportional and scaled read counts are then added to all GO terms annotating a protein, including updated terms. Finally, the percentage of reads covering a GO term within a namespace (Biological Process, Molecular Function, and Cellular Component) is calculated by dividing the scaled read count of a GO term by the total scaled read counts covering a namespace and multiplying by 100. The final output table of the pipeline is a contingency table with GO IDs of all namespaces as rows and samples or groups as columns, with percentage within a namespace as values.

2.3.2.5 Visualization of Data

To facilitate exploration of results from MetaFunc, MetaFunc automatically builds an R shiny application, such that users can view and interact with the taxonomy and gene ontology tables. The application allows users to select GO terms and identify the species whose proteins are annotated with the searched for term. Conversely, users may search for a species and obtain all GO terms associated with the searched for species. See the pipeline's documentation page for more information (https://metafunc.readthedocs.io/en/latest/rshiny.html).

2.3.3 Host Analyses (**Figure 2.1b**)

Many microbiome communities are often associated with a host genome. Reads belonging to the host genome have the capacity to misclassify as microbiome (S. H. Ye et al., 2019) and filtering of host reads has been a part of many microbiome studies, either prior to sequencing or *in silico* (Hugerth & Andersson, 2017; Macklaim & Gloor, 2018; Y. Xia et al., 2018). The MetaFunc pipeline offers the option of mapping reads to a host genome using the program STAR and using the unmapped reads from this step as input to Kaiju for the microbiome analysis.

MetaFunc also allows additional analyses of host reads after STAR mapping. Host genes are quantified using featureCounts (Liao et al., 2014) of the subread package. If comparisons between groups are indicated, edgeR is used to perform differential gene expression analysis (DGEA). Additionally, supplying a gene matrix transposed (*.gmt*) file from e.g. the molecular signatures database (*GSEA*, n.d.; Liberzon et al., 2011; Subramanian et al., 2005) allows for gene set enrichment analysis (GSEA) of host genes using the clusterProfiler package (G. Yu et al., 2012).

2.3.3.1 Host Gene – Microbe Species Correlation

When a comparison between groups is specified, the pipeline also performs Spearman correlation analysis between the top most significant differentially expressed genes (DEGs) and top most significant differentially abundant (DA) microbes. Results of these correlations are summarized in a matrix on which hierarchical clustering is performed and a heatmap is generated using

Clustergrammer (Fernandez et al., 2017). Through this heatmap and table, a user can investigate the strength of correlation (*rho*) between a DA microbe and a DEG, and which microbes and genes have similar patterns of correlations.

2.3.4 Tutorial/Manual

For a more detailed description of the workflow, usage instructions, and results, documentation of the MetaFunc pipeline may be found in https://metafunc.readthedocs.io/en/latest/index.html.

2.4 Results (Usage Example)

2.4.1 Dataset PRJNA413956: Matched Colorectal Cancer (CRC) and Adjacent Non-Tumor Tissue

In order to demonstrate the utility of the MetaFunc pipeline, we obtained publicly available transcriptomics data from the study of Li et al., (2018) consisting of 10 tumor and corresponding adjacent non-tumor colorectal tissue samples. Raw sequencing data was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104836 and input to the pipeline and the full workflow carried out, generating data for host, microbiome, and host-microbiome correlation.

2.4.1.1 Microbiome Results

2.4.1.1.1 Identification and Comparison of Taxa

The MetaFunc pipeline outputs a table of percent abundances of species that are identified in each sample and an average of these abundances across members of the same group if a grouping condition is applied. We ran the pipeline with the intent of comparing microbiome species and function between colon cancer samples and non-tumor matched samples.

Previous studies have already established that certain microbes associate more with colorectal cancer samples compared to healthy controls. We searched for *Fusobacterium nucleatum*, *Parvimonas micra,* and *Porphyromonas asaccharolytica* in the averaged group results. These microbes have previously been found to be more abundant in colorectal cancer cohorts in meta-analyses of several datasets (Dai et al., 2018; Thomas et al., 2019). We also searched for the *Bifidobacterium* species, *B. bifidum* and *B. longum*; *Bifidobacteria* are thought to confer protection from colorectal cancer (Wei et al., 2018).

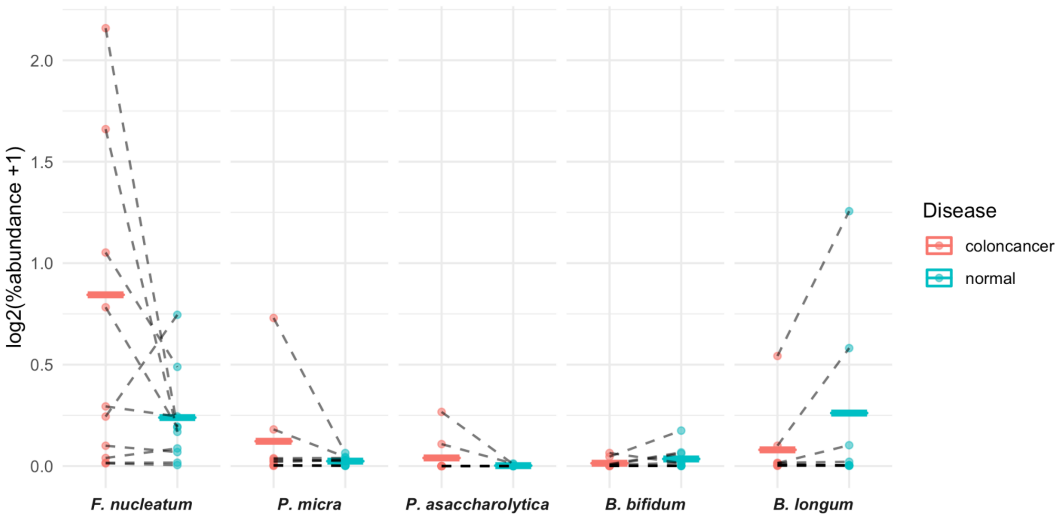


Figure 2.2. Average percent abundance of selected bacterial species in CRC tissue compared to matched non-tumor (normal) samples. From MetaFunc tabulated results, we plotted the percent abundances of selected bacteria in CRC and matched normal samples. Raw values were first $\log_2$ transformed, with prior addition of 1 as a pseudocount to account for 0 values. Individual points represent per sample transformed values in red (CRC) and blue (Normal). Per group means are represented by the horizontal lines.

The horizontal lines in **Figure 2.2** show the average percent abundance of the species between samples from tumor and matched non-tumor tissue as identified through MetaFunc. As MetaFunc provides a per sample data, we are also able to plot individual values of CRC (red) and matched normal (blue) samples.

As seen in **Figure 2.2**, MetaFunc identified *F. nucleatum, P. micra,* and *P. asaccharolytica* as being relatively more abundant in the CRC group while the *Bifidobacterium* species are relatively more abundant in the normal group.

MetaFunc also has a step that utilizes edgeR to perform differential abundance on per sample species read counts, stratified according to CRC and non-tumor grouping.

This resulted in a total of 117 species that were significantly different between the groups (FDR < 0.05). There are 59 species upregulated and 58 downregulated in colon cancer samples. Through the MetaFunc results, we identified *Tannerella forsythia* as the most prominent enriched species in the colon cancer cohort with a $\log_2$ fold change ($\log_2$ FC) = 7.40. *T. forsythia* is a known oral pathogen, thought to be part of the so-called Red complex of periodontal pathogens, along with *Porphyromonas gingivalis*, and *Treponema denticola* (Malinowski et al., 2019). Members of this Red Complex have been found to be enriched in subtype CMS1 of colorectal cancers (Purcell, Visnovska, et al., 2017), the subtype most associated with immune process activation in CRC (Dienstmann et al., 2017; Guinney et al., 2015; Inamura, 2018).

2.4.1.1.2 Functional Profiling and Comparison

MetaFunc is intended to enable comparisons of the functional potential of the microbiome between groups. MetaFunc uses gene ontology annotations of protein matches from Kaiju. To demonstrate, we focused on polyamine biosynthetic processes GO terms. Polyamines (PAs) are polycations found to play important biological functions in cell growth. These molecules have been found to be associated with tumor progression and growth (Gerner & Meyskens, 2004; Soda, 2011; Tofalo et al., 2019). Although cells are able to biosynthesize polyamines and even export them, a large source of cellular polyamines comes from uptake from their surroundings and, importantly, the microbiota is thought to be an essential

source (Soda, 2011; Thomas et al., 2019; Tofalo et al., 2019) with spermidine and putrescine being the most common of bacterial PAs (Tofalo et al., 2019).

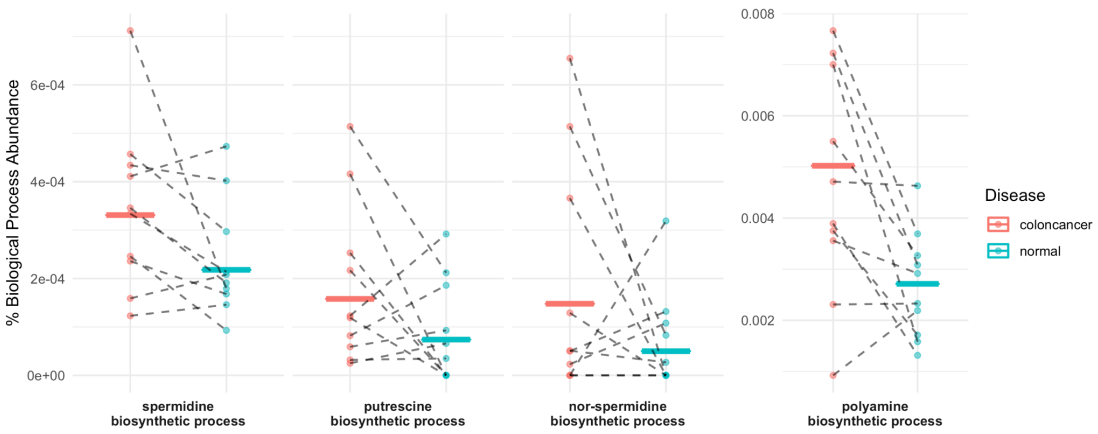
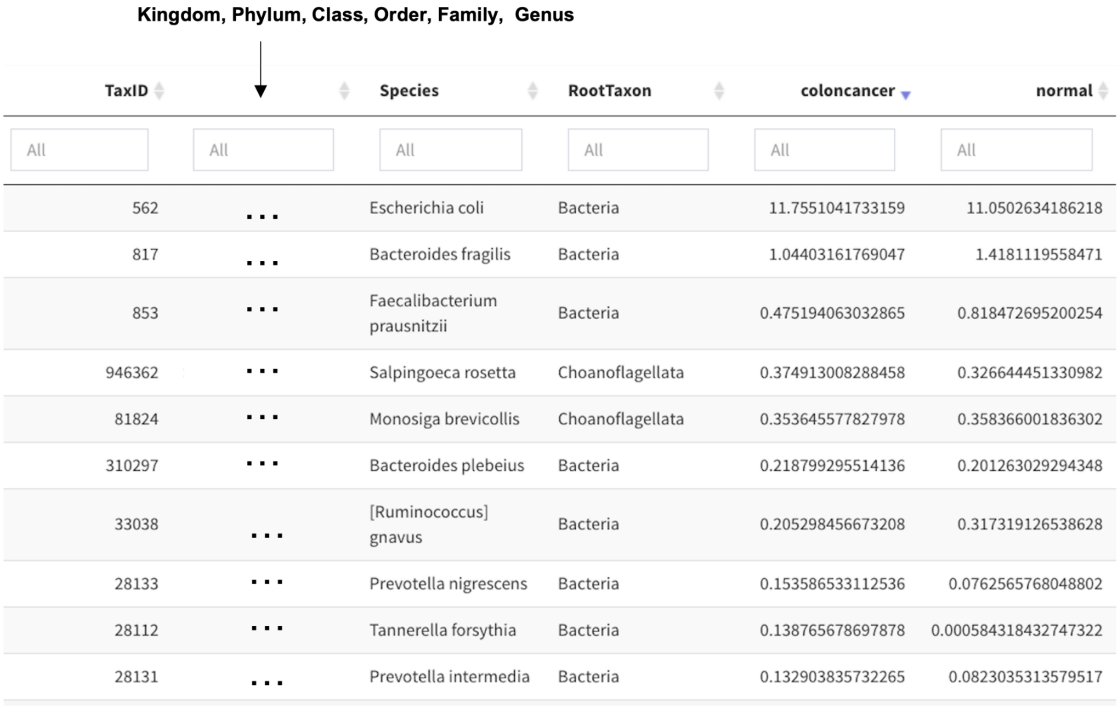


Figure 2.3. Percent abundance of specific polyamine biosynthetic process GO terms among all biological process GOs in a sample/group compared between CRC (red) and normal (blue) samples. Values were calculated as described in section *2.3.2.4 Gene Ontology: Protein Annotation,* and output in MetaFunc tables or in the R Shiny application. These values were plotted, overlaying group means (horizontal lines) and individual values (data points).

The horizontal lines in **Figure 2.3** show the percentage of reads among biological process GOs covering PA biosynthetic processes in the Colorectal Cancer and Normal conditions, superimposed with the individual values of samples from the CRC (red) and Normal (blue) groups. From **Figure 2.3**, we saw that several of the polyamine biosynthetic processes were relatively more abundant in the colorectal cancer cohort compared to the normal cohort, using protein annotations.

We used the built-in MetaFunc shiny application to facilitate an inquiry into the microbes species that may contribute to polyamine synthesis. To illustrate, we searched for *'polyamine biosynthetic process'* in the 'GO to TaxIDs' tab of the

application, and obtained a total of 126 TaxIDs contributing to the GO term in both colorectal cancer and normal samples. Of these TaxIDs, we identified *E.coli* and *B. fragilis* to be most abundant in both cohorts. However, differences in relative abundance of some microbial species can be identified between cancer and normal cohorts, notably several of which are oral pathogens from the genus *Prevotella*. A striking difference in abundance was seen in *Tannerella forsythia,* which was previously found to be significantly more abundant in the colorectal cancer cohort via edgeR (**Figure 2.4**). These data suggest that *T. forsythia* represents one of the bacterial species that most contributes to increased polyamine synthesis in CRC samples in this cohort.



TaxID		Species	RootTaxon	coloncancer	normal
All	All	All	All	All	All
562	...	Escherichia coli	Bacteria	11.7551041733159	11.0502634186218
817	...	Bacteroides fragilis	Bacteria	1.04403161769047	1.4181119558471
853	...	Faecalibacterium prausnitzii	Bacteria	0.475194063032865	0.818472695200254
946362	...	Salpingoeca rosetta	Choanoflagellata	0.374913008288458	0.326644451330982
81824	...	Monosiga brevicollis	Choanoflagellata	0.353645577827978	0.358366001836302
310297	...	Bacteroides plebeius	Bacteria	0.218799295514136	0.201263029294348
33038	...	[Ruminococcus] gnavus	Bacteria	0.205298456673208	0.317319126538628
28133	...	Prevotella nigrescens	Bacteria	0.153586533112536	0.0762565768048802
28112	...	Tannerella forsythia	Bacteria	0.138765678697878	0.000584318432747322
28131	...	Prevotella intermedia	Bacteria	0.132903835732265	0.0823035313579517

Figure 2.4. Screenshot from MetaFunc R shiny application. This view shows the first 10 species with proteins contributing to the GO *Polyamine Biosynthetic Process*. The R Shiny application columns include a URL (not shown in screenshot), which is linked to the NCBI's Taxonomy Browser, the Species Taxonomy ID, Lineage (indicated as `...` in screenshot), Root Taxon, and percent abundances of the species in the two groups being compared: CRC and normal samples. Note that percent abundances refer to the total abundance of the species in question, not just the proteins contributing to the GO term. Results can be sorted based on any column from highest to lowest percent abundance in the colon cancer cohort.

2.4.1.2 Host Results

The dataset we used for this study was from a total RNA transcriptomics run aiming to identify long non-coding RNAs (lncRNAs) and mRNAs in colorectal cancer samples (M. Li et al., 2018). Therefore, we first mapped the reads to the human genome using the STAR mapping utility of the pipeline, subsequently using only the unmapped reads for the microbiome analyses. From the reads mapped to the human genome, MetaFunc was able to obtain counts of reads covering human genes and using these,

obtained differentially expressed genes between CRC and matched normal samples through edgeR. MetaFunc results showed a total of 1476 differentially expressed genes with an FDR < 0.05 and $|\log_2$ fold change$|>2$, compared to the 3221 differentially abundant mRNAs found in the source publication. From these, we found all the top 5 upregulated and top 5 downregulated genes as reported in the source publication (M. Li et al., 2018), as well as all the genes they had randomly selected for expression confirmation via qPCR. **Figure 2.5** shows their fold change as found through MetaFunc.

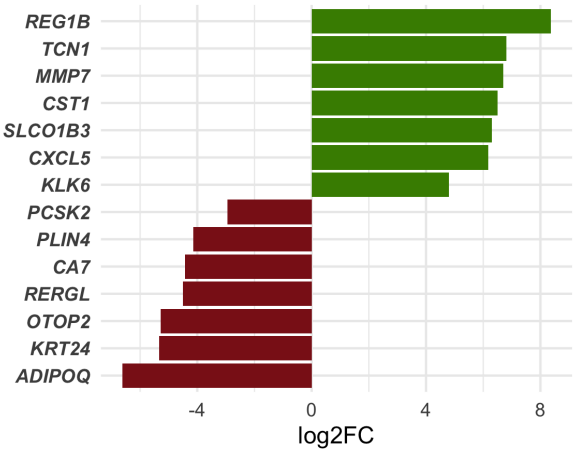


Figure 2.5. Log_2 fold changes (log_2FC) of representative upregulated and downregulated human genes (as enumerated in M. Li et al., 2018) between CRC and matched normal samples in this study. Fold change values were obtained from the edgeR results of the pipeline. All these genes are significant (FDR < 0.05) in both this study and the source publication.

MetaFunc is also able to perform host gene set enrichment analysis using the differentially expressed genes. Significant gene sets (adjusted p-value < 0.05) with the highest normalized positive enrichment scores (NES) included such terms as

ribosome biogenesis, DNA replication, mitotic nuclear division, and condensed chromosome (see **Supplementary Table 2.1**), many of which appear to be related to cell division or replication, consistent with the findings of Li et al., (2018) that the upregulated lncRNAs they found were involved in mitosis, cell cycle process, and mitotic cell cycle.

2.4.1.3 Host – Microbiome Correlations

We set MetaFunc's default abundance cutoff for microbial identification to 0.001% to remove most probable contaminants and so as not to lose any other meaningful taxonomies. It has been shown in a prior study (S. H. Ye et al., 2019), however, that most classifiers call false positives at below 0.01% abundance. We therefore applied this 0.01% cutoff in looking at the host-microbiome correlations in this dataset to narrow our focus on microbes that are more likely to be involved in our test case.

In using the 0.01% cutoff, MetaFunc was able to only identify 19 differentially abundant microbes. Their correlations with the top 100 significantly abundant genes can be seen at the URL: http://amp.pharm.mssm.edu/clustergrammer/viz/5f02a49e8ec9bb33170b865c/cor.deg-tax.matrix.tsv. **Table 2.1** highlights some notable correlations between differentially abundant microbes and differentially expressed human genes. *T. forsythia*, although significantly abundant in CRC samples, do not correlate significantly with any DEGs in CRC. Among its highest correlations however included the gene Colorectal Neoplasia Differentially Expressed (*CRNDE*).

Conversely, we investigated which species correlated with *CRNDE*. The highest correlations were with microbes *C. lusitaniae, C. necator,* and *S. pyogenes.* All correlations were determined to be significant. The same species were among the highest correlations of *TCN1*, and *WNT2*. *TCN1* was among the top DEGs in cancer identified in this study as well as in M. Li et al., 2018. *WNT2* meanwhile is part of the Wnt/β-catenin pathway, which has roles in cell proliferation, cell migration, and cell differentiation. WNT2 is responsible for hyperactivation of β-catenin and is known to be upregulated in CRC (Jung et al., 2015).

Table 2.1. Spearman Correlation Between Differentially Abundant Microbes and Differentially Expressed Genes in CRC

Gene Name	Gene ID	TaxID	Species	rho	p-value
CRNDE	ENSG00000245694.10	28112	*Tannerella forsythia*	0.29	0.22
		36911	*Clavispora lusitaniae*	0.70	0.00063
		106590	*Cupriavidus necator*	0.65	0.0019
		1314	*Streptococcus pyogenes*	0.63	0.0027
TCN1	ENSG00000134827.8	106590	*Cupriavidus necator*	0.71	0.00042
		36911	*Clavispora lusitaniae*	0.61	0.0045
		1314	*Streptococcus pyogenes*	0.60	0.0048
WNT2	ENSG00000105989.10	1314	*Streptococcus pyogenes*	0.84	4.07E-06
		106590	*Cupriavidus necator*	0.75	0.00015
		36911	*Clavispora lusitaniae*	0.75	0.00016

2.4.2 Dataset PRJNA4040030: Consensus Molecular Subtypes (CMS) of CRC Samples

To illustrate MetaFunc's capacity to compare more than two sample groups, we used MetaFunc to analyze transcriptome reads from the study of Purcell et al., (2017) (raw

reads may be accessed at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA404030), which are grouped into four CRC consensus molecular subtypes (CMS). A total of 33 samples were collected during surgical resection of tumors, and sample preparation for RNA sequencing was carried out using the Illumina TruSeq Stranded Total RNA Library preparation kit. For this sample SolexaQA++ (Cox et al., 2010) was used to trim reads, which were then run through Salmon (Patro et al., 2017) to quantify transcript expression. The publicly available CRC CMS classifier (Guinney et al., 2015) was used to categorize samples into one of four CMSs. Of the 33 samples, only 27 were classified into a CMS and of these, only one sample was classified into CMS4. This sample was also removed from the dataset for lack of replicates leaving a total of 26 samples – 7 samples in CMS1, 11 in CMS2, and 8 in CMS3. Metafunc was used with default parameters, except for the following options: trimming was set to false, and featureCounts with reverse stranded option was used.

2.4.2.1 Microbiome Results

2.4.2.1.1 Taxonomic Identification and Comparison

MetaFunc performed pairwise differential abundance analysis on the three groups using edgeR. From MetaFunc's results, we considered a species to be significantly abundant in a subtype if it is significantly abundant compared to both of the other subtypes. For instance, a significantly abundant species in CMS1 must be significantly abundant in the CMS1 vs CMS2 and CMS1 vs CMS3 comparisons. Using this definition, only CMS1 had species that were significantly abundant (FDR < 0.05) compared to both CMS2 and CMS3. **Figure 2.6** shows the False Discovery Rate (FDR;

point size), and $\log_2$ fold change of the species in CMS1 compared to CMS2 (purple) and CMS3 (yellow).

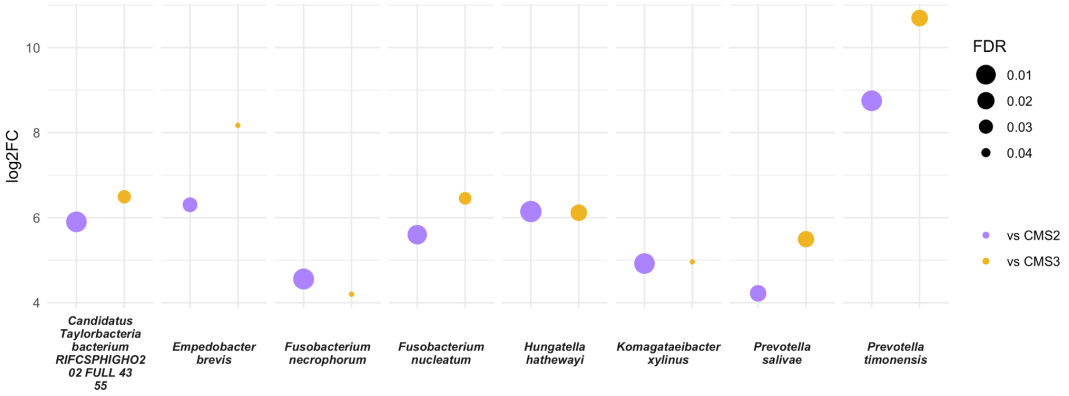


Figure 2.6. Microbes that are significantly more abundant (FDR < 0.05) in CMS1 compared to CMS2 (purple) and CMS3 (yellow). Microbes are considered differentially abundant (DA) in CMS1 if it is identified through edgeR as DA in both CMS1 vs CMS2 and CMS1 vs CMS3 comparisons. Log_2FC (y-axis) is the $\log_2$ of the fold-change between CMS1 and the other subtypes (e.g. CMS1/CMS2); FDR (point sizes) is the false discovery rate adjusted p-values.

We took note of species in the genera *Prevotella* and *Fusobacterium,* which have previously been associated with colorectal cancer. *Fusobacterium nucleatum* in particular has strong evidence of an association with CRC (Dai et al., 2018; Gao et al., 2015; X. Ye et al., 2017). Most of these are also members of the oral microbiota, which have also previously been associated with cancer development particularly through inflammatory processes (Whitmore & Lamont, 2014). We found no species that were significantly abundant in CMS2 or CMS3 using the given criteria.

2.4.2.1.2 Functional Profiling and Comparison

Through the microbiome functional results of MetaFunc, we then investigated if processes relating to pathogen-associated molecular patterns (PAMPs) were contributed by the microbial communities, considering that CMS1 is characterized by immune responses, which are usually triggered when the human immune system recognizes such molecules. We used the MetaFunc R shiny application to search for terms *'lipopolysaccharide biosynthetic process', 'lipid A biosynthetic process',* and *'peptidoglycan biosynthetic process',* and their relative abundances. Unsurprisingly, all PAMPs were relatively more abundant in CMS1 (**Figure 2.7**).

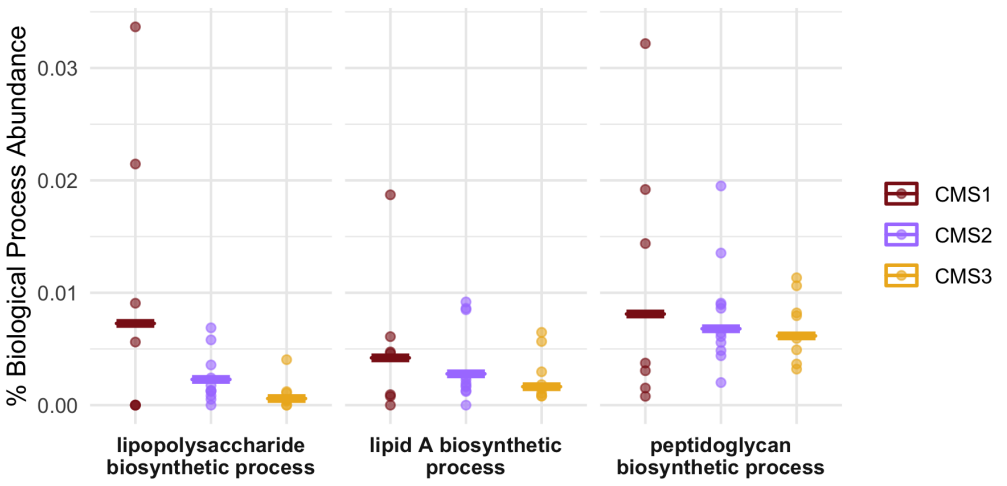
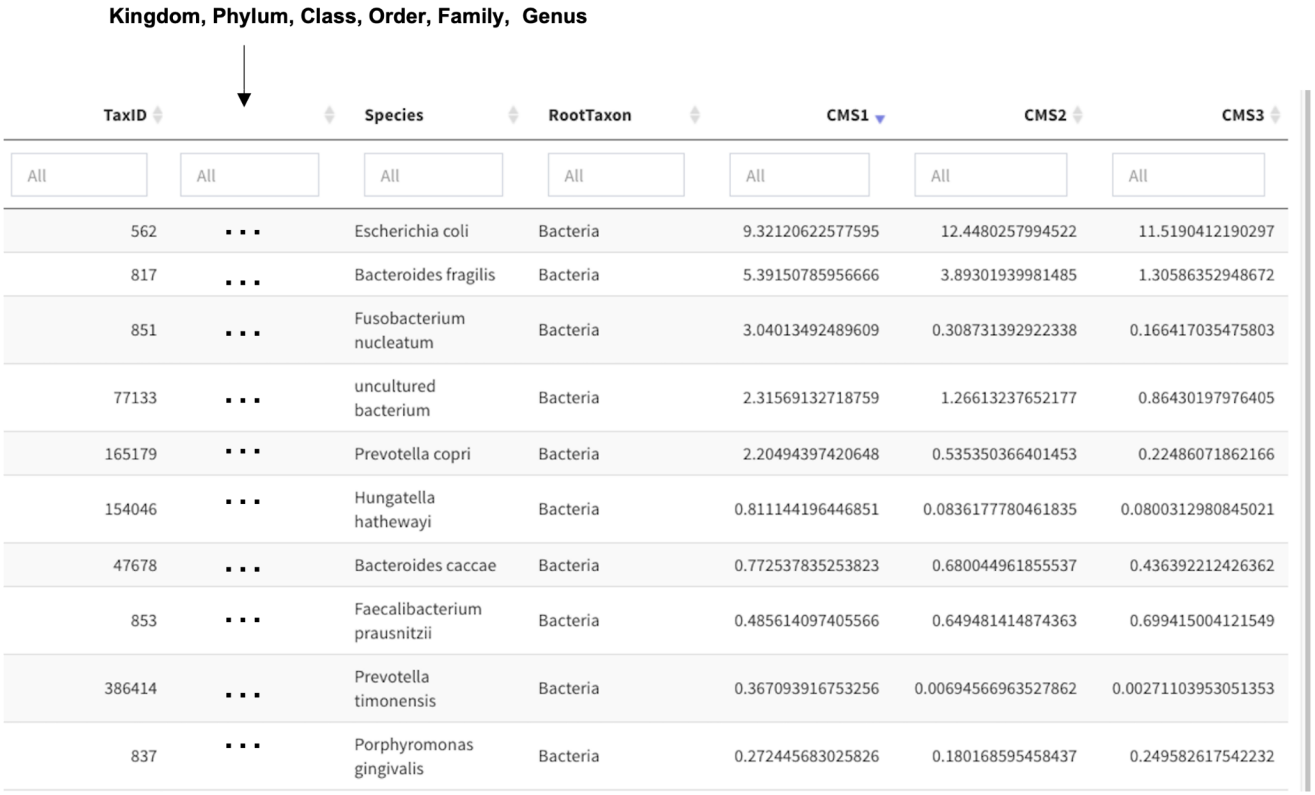


Figure 2.7. Percent abundance of specific PAMPs biosynthetic process GO terms among all biological process GOs in a sample/group compared between CRC subtypes, CMS1 (red) , CMS2 (purple), and CMS3 (yellow). Values were calculated as described in section *2.3.2.4 Gene Ontology: Protein Annotation,* and output in MetaFunc tables or in the R Shiny application. These values were plotted, overlaying group means (horizontal lines) and individual values (data points).

Using the MetaFunc R shiny application, we also searched for which species might be contributing to the above terms. **Figure 2.8** is a screenshot of the application showing the species contributing to any of the terms in **Figure 2.7**. **Figure 2.8** is arranged from highest to lowest relative abundance in CMS1 and we observed microbes that were among those identified to be significantly abundant in CMS1 such as *Fusobacterium nucleatum, Prevotella timonensis* , and *Hungatella hathewayi*.



TaxID		Species	RootTaxon	CMS1	CMS2	CMS3
All	All	All	All	All	All	All
562	...	Escherichia coli	Bacteria	9.32120622577595	12.4480257994522	11.5190412190297
817	...	Bacteroides fragilis	Bacteria	5.39150785956666	3.89301939981485	1.30586352948672
851	...	Fusobacterium nucleatum	Bacteria	3.04013492489609	0.308731392922338	0.166417035475803
77133	...	uncultured bacterium	Bacteria	2.31569132718759	1.26613237652177	0.86430197976405
165179	...	Prevotella copri	Bacteria	2.20494397420648	0.535350366401453	0.22486071862166
154046	...	Hungatella hathewayi	Bacteria	0.811144196446851	0.0836177780461835	0.0800312980845021
47678	...	Bacteroides caccae	Bacteria	0.772537835253823	0.680044961855537	0.436392212426362
853	...	Faecalibacterium prausnitzii	Bacteria	0.485614097405566	0.649481414874363	0.699415004121549
386414	...	Prevotella timonensis	Bacteria	0.367093916753256	0.00694566963527862	0.00271103953051353
837	...	Porphyromonas gingivalis	Bacteria	0.272445683025826	0.180168595458437	0.249582617542232

Figure 2.8. Screenshot of R shiny application showing the relative abundances of species associated with PAMPs biosynthetic processes compared among CMS1, CMS2, and CMS3. This view shows the first 10 species, with highest abundances in CMS1, with proteins contributing to any of the PAMPs biosynthetic processes described above. The application columns show a URL (not shown in screenshot), which is linked to the NCBI's Taxonomy Browser, the Species Taxonomy ID, Lineage (shown as `...` in screenshot), Root Taxon, and percent abundances of the species in the three groups being compared: CMS1, CMS2, and CMS3. Note that percent abundances refer to the total abundance of the species in question, not just the proteins contributing to the GO term. Results shown are sorted from highest to lowest percent abundance in the CMS1 group.

2.4.2.2 Host Results

2.4.2.2.1 Gene Set Expression Analysis

MetaFunc calculated differentially expressed genes between subtypes in a pairwise manner (i.e. CMS1 vs CMS2, CMS1 vs CMS3, CMS2 vs CMS3). From the DEGs of the results, MetaFunc was also able to calculate enriched gene sets for each comparison. Similar to identifying DA microbes, we obtained a final set of enriched gene sets for a subtype if it showed enrichment compared to both other subtypes (adjusted p-value < 0.05). Unsurprisingly, we observed several GO terms involved in immune response enriched in CMS1, including regulation of innate immune response, response to interferon gamma, and positive regulation of cytokine production among others. Enriched GOs in CMS2 are involved in the cell cycle and and ribosome biogenesis, with terms such as tRNA metabolic process, ribosomal large subunit biogenesis, and DNA replication initiation, while GOs enriched in CMS3 involve metabolic processes, e.g. primary xenobiotic metabolic process, flavonoid metabolic process, and lipid catabolic process. These results are consistent with the description of these three CRC subtypes in the original CMS study (Guinney et al., 2015). The top enriched gene sets for each subtype can be found in **Supplementary Tables 2.2-2.7**.

2.4.2.3 Host – Microbiome Results

Next, using correlation results from MetaFunc, we investigated which of the top significant differentially expressed genes correlated with the significantly abundant

microbes in CMS1. We obtained the following statistically significant correlations between host and microbiome abundances shown in **Table 2.2**.

Table 2.2. Spearman Correlation Between Differentially Abundant Microbes in CMS1 and Differentially Expressed Genes in CMS1

Gene Name	Gene ID	TaxID	Species	*rho*	p-value
WARS1	ENSG00000140105.18	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.59	0.0015
WARS1	ENSG00000140105.18	**851**	*Fusobacterium nucleatum*	0.55	0.0035
RNF213	ENSG00000173821.19	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.54	0.0048
ICAM1	ENSG00000090339.9	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.50	0.01
RNF213	ENSG00000173821.19	**851**	*Fusobacterium nucleatum*	0.50	0.01
PARP14	ENSG00000173193.15	**851**	*Fusobacterium nucleatum*	0.47	0.02
PARP14	ENSG00000173193.15	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.47	0.02
PARP9	ENSG00000138496.16	**851**	*Fusobacterium nucleatum*	0.46	0.02
ICAM1	ENSG00000090339.9	**851**	*Fusobacterium nucleatum*	0.46	0.02
SLC15A3	ENSG00000110446.11	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.46	0.02
STAT1	ENSG00000115415.19	**386414**	*Prevotella timonensis*	0.44	0.02
CD163	ENSG00000177575.12	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.42	0.03
PARP14	ENSG00000173193.15	**386414**	*Prevotella timonensis*	0.42	0.03
CD163	ENSG00000177575.12	**386414**	*Prevotella timonensis*	0.42	0.03
ICAM1	ENSG00000090339.9	**386414**	*Prevotella timonensis*	0.41	0.04
PARP9	ENSG00000138496.16	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.41	0.04
CD163	ENSG00000177575.12	**851**	*Fusobacterium nucleatum*	0.41	0.04
SLC15A3	ENSG00000110446.11	**851**	*Fusobacterium nucleatum*	0.40	0.04
STAT1	ENSG00000115415.19	**851**	*Fusobacterium nucleatum*	0.40	0.04
PML	ENSG00000140464.20	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.39	0.05
GBP1	ENSG00000117228.10	**28448**	*Komagataeibacter xylinus*	-0.39	0.05
CEBPA	ENSG00000245848.3	**154046**	*Hungatella hathewayi*	-0.41	0.04
GNLY	ENSG00000115523.16	**28448**	*Komagataeibacter xylinus*	-0.43	0.03

Some of these correlations may be found in http://maayanlab.cloud/clustergrammer/viz/610d8b3c97f268000ea37f41/cor.deg-tax.matrix.tsv. This is the hierarchical cluster obtained when correlating top DA microbes and top DEGs in CMS1 compared to CMS2. It is to be noted that there may be correlations in this clustering that are not found in CMS1 compared to CMS3 and are therefore not reported in **Table 2.2**.

The Spearman correlations (*rho*) between DA microbes and DEGs were quite small in value (the highest value being ~ |0.59| between WARS1 and *Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*). Nevertheless, several of the genes appeared to have relevant function with regards to CRC and immune responses. **Table 2.3** shows information for genes that correlated with *Fusobacteria* and *Prevotella* species in our analyses. These two microorganisms have previously been associated with CRC.

Table 2.3. Gene Information of Differentially Expressed Genes correlated with Differentially Abundant Microbes in CMS1.

Gene Name	Protein Name	Relevant Protein/Gene Function	Association with CRC and/or Inflammation	Sources
ICAM1	Intercellular adhesion molecule 1	Mediates cell adhesion of cytotoxic T lymphocytes and natural killer cells	upregulation of ICAM1 inhibits tumor growth and metastasis; a soluble form (sICAM1) is increased in CRC tissues compared to normal, and is associated with an inflammatory tumor microenvironment	Sánchez-Rovira et al., 1998; Schellerer et al., 2019; Tachimori et al., 2005
SLC15A3	Solute carrier (SLC) 15A3	Membrane transporter; highly expressed in macrophage populations	Upregulated by LPS via NF-κB pathway; influences pro-inflammatory cytokine production triggered by TLR-4	Wang et al., 2014 Song et al., 2018
CD163	CD163 receptor	M2 Macrophage marker	M2 macrophages are anti-inflammatory macrophages and CD163+ tumor-associated macrophages are with mesenchymal transition and poor prognosis in CRC; is correlated with CCL4	Argyle and Kitamura, 2018; Bayoumi et al., 2016; De la Fuente López et al., 2018; Pinto et al., 2019
STAT1	Signal transducer and activator of transcription 1	Transcription factor for IFN signaling	upregulated in CRCs; correlated with PD-L1 and PD1 immune checkpoint inhibitors; pro-oncogenic in MSI CRCs	Leon-Cabrera et al., 2018; Tanaka et al., 2020
PARP 9	Poly(ADP-ribose) polymerase family member 9	Involved in cell migration	possible role in metastasis	Vyas and Chang, 2014
PARP 14	Poly(ADP-ribose) polymerase family member 14	Involved in IL-4 signaling and cell migration	involved in anti-apoptotic effects	Vyas and Chang, 2014
RNF213	Ring Finger Protein 213	Involved in PI3K-AKT pathway for cell growth	involved in endothelial angiogenesis	Ohkubo et al., 2015
WARS1	Tryptophanyl-TRNA Synthetase 1	Inhibitor of angiogenesis; Involved in IFN-γ signaling	involved in immune responses; cleaved form potentially inhibits angiogenesis; increased levels indicate better CRC survival	Ghanipour et al., 2009; Jin, 2019

2.4.3 Comparison of Kaiju Results to HUMAnN2

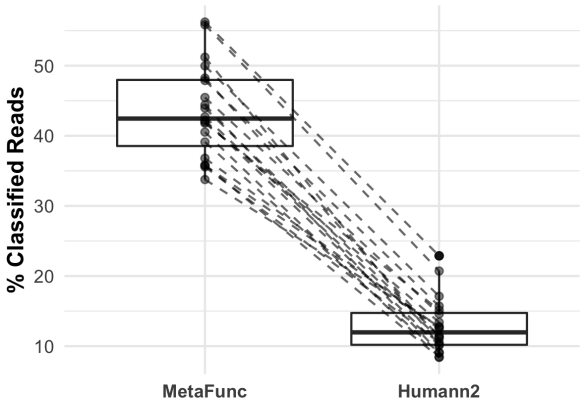


Figure 2.9. Percentage (%) of reads classified using HUMAnN2 and MetaFunc pipeline using Kaiju as distributed across 20 samples. This data shows that many more reads are classified to a microbiome taxonomy using MetaFunc's Kaiju compared to HUMAnN2's MetaPhlan2.

HUMAnN2 (Franzosa et al., 2018) is one of the packages most frequently used to assess functional pathways of the microbiome, and to ascertain which organisms are contributing to the functional pathways. HUMAnN2 works by using MetaPhlAn2 to pre-screen for taxonomic classifications on which the database for gene hits will be based. MetaPhlAn2 has a rather limited database for pre-screening of organisms (S. H. Ye et al., 2019), resulting in a high level of unmapped reads and a limited number of organisms identified.

We ran the same sequencing reads from the study PRJNA413956 (M. Li et al., 2018) through HUMAnN2, first trimming with fastp and removing human-mapped reads using the same conditions as for the MetaFunc pipeline. To be more comparable, we changed the pre-screen threshold of HUMAnN2 to 0.001% of mapped reads. Results

show that for the 20 samples analyzed, 8.4% – 22.9% of reads mapped after the nucleotide and protein alignment steps. In contrast, using Kaiju in the MetaFunc pipeline resulted in 33.8% – 56.2% reads mapped to microbial species through protein matches (**Figure 2.9**).

We also detected only 87 species across the 20 samples using HUMAnN2, compared with a total of 4267 species using Kaiju in the MetaFunc pipeline.

As raw read counts that are scaled to species-classified reads as percent abundance are used in MetaFunc, while HUMAnN2 uses reads-per-kilobase (RPK), it is difficult to compare results from HUMAnN2 and MetaFunc directly.

2.5 Discussion

MetaFunc allowed us to investigate the relative abundances of known CRC – associated bacteria between CRC samples and matched normal tissues using the PRJNA413956 dataset. MetaFunc results showed that microbes known to contribute to CRC progression are relatively more abundant in cancer samples while those protective in CRC are relatively more abundant in the normal samples. Through MetaFunc, we also identified that *Tannerella forsythia*, a known oral pathogen and part of the Red Complex that causes periodontal diseases (Malinowski et al., 2019), is significantly more abundant in CRC tissues than in normal tissues. Oral pathogens have previously been seen to associate with CRC samples (Flemer et al., 2018; Koliarakis et al., 2019; Thomas et al., 2019; Whitmore & Lamont, 2014). By

investigating the R shiny application from MetaFunc, we also found that *T. forsythia,* along with bacteria in the *Prevotella* genera, contributed to polyamine biosynthetic processes indicating that some oral pathogens contribute to cancer progression by producing polyamines that could be taken up by the surrounding cells.

MetaFunc was also able to replicate the differentially expressed host genes enumerated in the source publication. A comparison between an overlap of the 3221 mRNAs the publication found as differentially expressed, and our 1476 DEGs could not be made as their 3221 genes were not named. This difference in the number of DEGs could have stemmed from different analysis parameters such as quality control packages (fastp in MetaFunc versus Trimmomatic in the source publication), mapping tool (STAR in MetaFunc vs TopHat), differential expression analysis tool (edgeR in Metafunc vs DESeq2), or even the difference in host genome reference (Grch38 in MetaFunc's vs. hg19). Although the difference in number is large, the complete overlap of their top differentially expressed genes and our analyses suggest that we are obtaining equivalence in the most impactful results. Indeed, we found three out of five of the publication's top upregulated genes, and two out of five of the publication's top downregulated genes, in MetaFunc's top five upregulated and downregulated genes, respectively, when arranged by fold-change.

Furthermore, we were able to identify known bacteria in the MSI-Immune subset of CRCs by identifying the DA microbes in CMS1 compared to both CMS2 and CMS3 subtypes, as identified by MetaFunc's edgeR step. *Fusobacteria* have long been associated with colorectal cancer development (Dai et al., 2018; Gao et al., 2015;

Thomas et al., 2019; X. Ye et al., 2017) while *Prevotella* includes species that inhabit the oral cavity; there have also been *Prevotella* species that were found to be abundant in CRC cohorts (Dai et al., 2018; Flemer et al., 2018; Gao et al., 2015). In line with this, PAMPs were also found to be relatively more abundant in the CMS1 cohort upon investigation through MetaFunc's R shiny application. The involvement of these bacteria in CMS1, as well as a relatively higher abundance of proteins contributing to biosynthesis of PAMPs in CMS1, indicates a role for microorganisms in the immune responses that drive the development of CRC in these tumors. This is further supported by correlation with host genes involved in inflammation and/or CRC development as found using MetaFunc's Spearman correlation step. The lack of significantly abundant microorganisms in CMS2 and CMS3 may reflect that the CRC development in these subtypes are not as dependent on immune dysregulation.

We created MetaFunc with the aim of identifying microbes and their functional contribution in a microbiome environment. One of the most widely used packages for this is HUMAnN2 (Franzosa et al., 2018) but we find the taxonomic identification generated by HUMAnN2 to be limited, because of its reliance on marker genes. We do acknowledge that, especially at the 0.001% abundance cutoff, some of these species we are seeing could be false positives, or that these could be contaminants from sequencing and processing kits used (Goffau et al., 2018; Salter et al., 2014). We would caution users in interpreting data from microbes of very low abundances and would recommend following Salter and Colleagues' (2014) advice of including negative control samples in sequencing. Indeed we could be seeing these effects upon looking at the microbes correlating with significantly abundant host genes in

CRC samples from PRJNA413956. While *C. lusitaniae* is an opportunistic pathogen causing candidemia (Desnos-Ollivier et al., 2011; Krcmery et al., 1999) possibly exploiting the lowered immune responses in cancer patients (Aslani et al., 2018), and some *Streptococcus* species have previously been implicated in CRC (Kumar et al., 2017; X. Xia et al., 2020), with *S. pyogenes* having been known to cause invasive infections in humans (Parks et al., 2015), *Cupriavidus necator* (formerly known as *Ralstonia eutropha* (Reinecke & Steinbüchel, 2009), is a soil bacterium that may be a sequencing contaminant in this dataset. *Cupriavidus* and *Ralstonia* species have been previously identified as common contaminants in meta-omics studies (M. Guo et al., 2019; Salter et al., 2014).

MetaFunc analyzes host and microbiome reads, providing a user-friendly, interactive R shiny application to investigate results, most useful for those with candidate microbes and function in mind, or for exploratory analyses of the characteristics of a user's dataset. Downstream analysis, such as differential abundance of microbes, can also facilitate parsing of tables in the shiny application. Its results also provide potential starting points for more in-depth analyses or hypothesis generation for experimental procedures. Results in *'.tsv'* formats are also provided for use in other downstream bioinformatics applications. For instance, while we initially included edgeR as a tool to explore differential expression/abundance analysis, as it had been described by Thorsen et al., (2016) as able to distinguish true differentially abundant taxa in 16S datasets, we used DESeq2 for subsequent analyses in Chapters 3 and 4 of this thesis, as we found DESeq2 more amenable for the design formulas we needed

to carry out our analyses. DESeq2 was also found to be consistent in false positive rates, concordance, power, and computational time (Calgaro et al., 2020).

While this method was developed for a metatranscriptomic dataset, it is also suitable for metagenomic data input for the microbiome analysis portion of the pipeline. As Kaiju (Menzel et al., 2016) identifies a single best protein match (or multiple matches with equal scores) of a read, we recommend its usage for short-read datasets. An exception could be made for long read sets in which the user is certain an input read will only span one protein.

We used the MetaFunc pipeline to compare genes and microbes between or among groups, but exploratory analyses of datasets from single groups can also be carried out.

2.6 Conclusion

Here we presented MetaFunc, a pipeline that is a one-stop shop for analyzing host and microbiome sequencing reads and their relationships. We found that we identified more microbes in our test datasets using MetaFunc compared to HUMAnN2. We have used MetaFunc to determine that microbes previously known to have associations with CRC are indeed relatively more abundant in CRC samples compared to normal samples. Furthermore, we were able to use MetaFunc to highlight that these microorganisms could contribute to CRC progression through polyamine production.

For a dataset with more than two groups, we have also used MetaFunc to identify abundant bacteria in a CRC subtype associated with immune responses, while conversely, we have not been able to identify significant microbes in the other CRC subtypes. MetaFunc's Spearman correlation step showed that the significant bacteria correlate with human DEGs that function in immune responses and CRC progression. We showed that MetaFunc was able to identify candidate microorganisms that differentiate sample groups and provide insight on the functional capacities of these candidates.

2.7 Supplementary Data

Supplementary Table 2.1. Top 25 Gene Sets Enriched in CRC Samples from PRJNA413956 Dataset as Measured by Normalized Enrichment Scores.

GO ID	GO Term	|NES|	p.adjust
GO:0042254	GO_RIBOSOME_BIOGENESIS	2.54	0.0021
GO:0000779	GO_CONDENSED_CHROMOSOME_CENTROMERIC_REGION	2.50	0.0021
GO:0034660	GO_NCRNA_METABOLIC_PROCESS	2.49	0.0021
GO:0000793	GO_CONDENSED_CHROMOSOME	2.47	0.0021
GO:0006260	GO_DNA_REPLICATION	2.47	0.0021
GO:0098687	GO_CHROMOSOMAL_REGION	2.47	0.0021
GO:0140014	GO_MITOTIC_NUCLEAR_DIVISION	2.46	0.0021
GO:0022613	GO_RIBONUCLEOPROTEIN_COMPLEX_BIOGENESIS	2.46	0.0021
GO:0034470	GO_NCRNA_PROCESSING	2.46	0.0021
GO:0016072	GO_RRNA_METABOLIC_PROCESS	2.45	0.0021
GO:0006261	GO_DNA_DEPENDENT_DNA_REPLICATION	2.45	0.0021
GO:0007059	GO_CHROMOSOME_SEGREGATION	2.45	0.0021
GO:0000819	GO_SISTER_CHROMATID_SEGREGATION	2.45	0.0021
GO:0000775	GO_CHROMOSOME_CENTROMERIC_REGION	2.44	0.0021
GO:0000070	GO_MITOTIC_SISTER_CHROMATID_SEGREGATION	2.44	0.0021
GO:0071103	GO_DNA_CONFORMATION_CHANGE	2.43	0.0021
GO:0098813	GO_NUCLEAR_CHROMOSOME_SEGREGATION	2.39	0.0021
GO:0000776	GO_KINETOCHORE	2.38	0.0021
GO:0030684	GO_PRERIBOSOME	2.37	0.0021
GO:0051983	GO_REGULATION_OF_CHROMOSOME_SEGREGATION	2.37	0.0021
GO:0048285	GO_ORGANELLE_FISSION	2.37	0.0021
GO:0006405	GO_RNA_EXPORT_FROM_NUCLEUS	2.36	0.0021
GO:0000075	GO_CELL_CYCLE_CHECKPOINT	2.36	0.0021
GO:0051783	GO_REGULATION_OF_NUCLEAR_DIVISION	2.35	0.0021
GO:0034728	GO_NUCLEOSOME_ORGANIZATION	2.34	0.0021

Supplementary Table 2.2. Top 25 Gene Sets Enriched in CMS1 Dataset as Measured by Normalized Enrichment Scores against CMS2.			
GO ID	**GO Term**	**|NES|**	**p.adjust**
GO:0034341	GO_RESPONSE_TO_INTERFERON_GAMMA	2.62	0.0022
GO:0060333	GO_INTERFERON_GAMMA_MEDIATED_SIGNALING_PATHWAY	2.53	0.0022
GO:0002250	GO_ADAPTIVE_IMMUNE_RESPONSE	2.49	0.0022
GO:0007159	GO_LEUKOCYTE_CELL_CELL_ADHESION	2.40	0.0022
GO:0051607	GO_DEFENSE_RESPONSE_TO_VIRUS	2.39	0.0022
GO:0042110	GO_T_CELL_ACTIVATION	2.38	0.0022
GO:0045088	GO_REGULATION_OF_INNATE_IMMUNE_RESPONSE	2.37	0.0022
GO:0001909	GO_LEUKOCYTE_MEDIATED_CYTOTOXICITY	2.37	0.0022
GO:0001906	GO_CELL_KILLING	2.36	0.0022
GO:0034340	GO_RESPONSE_TO_TYPE_I_INTERFERON	2.36	0.0022
GO:0031341	GO_REGULATION_OF_CELL_KILLING	2.36	0.0022
GO:0050863	GO_REGULATION_OF_T_CELL_ACTIVATION	2.36	0.0022
GO:0002449	GO_LYMPHOCYTE_MEDIATED_IMMUNITY	2.35	0.0022
GO:0002228	GO_NATURAL_KILLER_CELL_MEDIATED_IMMUNITY	2.34	0.0022
GO:1903039	GO_POSITIVE_REGULATION_OF_LEUKOCYTE_CELL_CELL_ADHESION	2.33	0.0022
GO:0002703	GO_REGULATION_OF_LEUKOCYTE_MEDIATED_IMMUNITY	2.32	0.0022
GO:0042098	GO_T_CELL_PROLIFERATION	2.32	0.0022
GO:0002697	GO_REGULATION_OF_IMMUNE_EFFECTOR_PROCESS	2.32	0.0022
GO:0032609	GO_INTERFERON_GAMMA_PRODUCTION	2.32	0.0022
GO:0009615	GO_RESPONSE_TO_VIRUS	2.31	0.0022
GO:0001818	GO_NEGATIVE_REGULATION_OF_CYTOKINE_PRODUCTION	2.30	0.0022
GO:0098542	GO_DEFENSE_RESPONSE_TO_OTHER_ORGANISM	2.30	0.0022
GO:0002237	GO_RESPONSE_TO_MOLECULE_OF_BACTERIAL_ORIGIN	2.30	0.0022
GO:0050852	GO_T_CELL_RECEPTOR_SIGNALING_PATHWAY	2.30	0.0022
GO:0001819	GO_POSITIVE_REGULATION_OF_CYTOKINE_PRODUCTION	2.30	0.0022

Supplementary Table 2.3. Top 25 Gene Sets Enriched in CMS1 Dataset as Measured by Normalized Enrichment Scores against CMS3.

GO ID	GO Term	|NES|	p.adjust
GO:0051607	GO_DEFENSE_RESPONSE_TO_VIRUS	2.38	0.0024
GO:0034341	GO_RESPONSE_TO_INTERFERON_GAMMA	2.34	0.0024
GO:0009615	GO_RESPONSE_TO_VIRUS	2.32	0.0024
GO:0098542	GO_DEFENSE_RESPONSE_TO_OTHER_ORGANISM	2.28	0.0024
GO:0071216	GO_CELLULAR_RESPONSE_TO_BIOTIC_STIMULUS	2.28	0.0024
GO:0045088	GO_REGULATION_OF_INNATE_IMMUNE_RESPONSE	2.26	0.0024
GO:0048002	GO_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_ANTIGEN	2.25	0.0024
GO:0034340	GO_RESPONSE_TO_TYPE_I_INTERFERON	2.25	0.0024
GO:0030199	GO_COLLAGEN_FIBRIL_ORGANIZATION	2.24	0.0024
GO:0032611	GO_INTERLEUKIN_1_BETA_PRODUCTION	2.23	0.0024
GO:0019882	GO_ANTIGEN_PROCESSING_AND_PRESENTATION	2.22	0.0024
GO:0042590	GO_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_EXOGENOUS_PEPTIDE_ANTIGEN_VIA_MHC_CLASS_I	2.22	0.0024
GO:0032612	GO_INTERLEUKIN_1_PRODUCTION	2.21	0.0024
GO:0060333	GO_INTERFERON_GAMMA_MEDIATED_SIGNALING_PATHWAY	2.20	0.0024
GO:0002474	GO_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_ANTIGEN_VIA_MHC_CLASS_I	2.18	0.0024
GO:0005201	GO_EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT	2.18	0.0024
GO:0043062	GO_EXTRACELLULAR_STRUCTURE_ORGANIZATION	2.18	0.0024
GO:0001819	GO_POSITIVE_REGULATION_OF_CYTOKINE_PRODUCTION	2.18	0.0024
GO:0050663	GO_CYTOKINE_SECRETION	2.18	0.0024
GO:0005581	GO_COLLAGEN_TRIMER	2.17	0.0024
GO:0032635	GO_INTERLEUKIN_6_PRODUCTION	2.17	0.0024
GO:0070555	GO_RESPONSE_TO_INTERLEUKIN_1	2.17	0.0024
GO:0071706	GO_TUMOR_NECROSIS_FACTOR_SUPERFAMILY_CYTOKINE_PRODUCTION	2.15	0.0024
GO:0071887	GO_LEUKOCYTE_APOPTOTIC_PROCESS	2.15	0.0024
GO:0031349	GO_POSITIVE_REGULATION_OF_DEFENSE_RESPONSE	2.12	0.0024

Supplementary Table 2.4. Top 25 Gene Sets Enriched in CMS2 Dataset as Measured by Normalized Enrichment Scores against CMS1.			
GO ID	**GO Term**	**|NES|**	**p.adjust**
GO:0042273	GO_RIBOSOMAL_LARGE_SUBUNIT_BIOGENESIS	2.10	0.0042
GO:0140053	GO_MITOCHONDRIAL_GENE_EXPRESSION	2.04	0.0056
GO:0032543	GO_MITOCHONDRIAL_TRANSLATION	1.97	0.0052
GO:0042788	GO_POLYSOMAL_RIBOSOME	1.93	0.0035
GO:0008033	GO_TRNA_PROCESSING	1.93	0.0052
GO:0044391	GO_RIBOSOMAL_SUBUNIT	1.92	0.006
GO:0042254	GO_RIBOSOME_BIOGENESIS	1.92	0.0079
GO:0000462	GO_MATURATION_OF_SSU_RRNA_FROM_TRICISTRONIC_RRNA_TRANSCRIPT_SSU_RRNA_5_8S_RRNA_LSU_RRNA	1.92	0.0037
GO:0030490	GO_MATURATION_OF_SSU_RRNA	1.90	0.0063
GO:0034660	GO_NCRNA_METABOLIC_PROCESS	1.90	0.0116
GO:0006400	GO_TRNA_MODIFICATION	1.90	0.0044
GO:0016072	GO_RRNA_METABOLIC_PROCESS	1.90	0.0063
GO:0006399	GO_TRNA_METABOLIC_PROCESS	1.89	0.0059
GO:0000184	GO_NUCLEAR_TRANSCRIBED_MRNA_CATABOLIC_PROCESS_NONSENSE_MEDIATED_DECAY	1.88	0.0051
GO:0006270	GO_DNA_REPLICATION_INITIATION	1.88	0.0060
GO:0003735	GO_STRUCTURAL_CONSTITUENT_OF_RIBOSOME	1.88	0.0056
GO:0030684	GO_PRERIBOSOME	1.88	0.0042
GO:0042274	GO_RIBOSOMAL_SMALL_SUBUNIT_BIOGENESIS	1.88	0.004
GO:0070129	GO_REGULATION_OF_MITOCHONDRIAL_TRANSLATION	1.87	0.0097
GO:0034470	GO_NCRNA_PROCESSING	1.85	0.0097
GO:0005736	GO_RNA_POLYMERASE_I_COMPLEX	1.84	0.0087
GO:0001510	GO_RNA_METHYLATION	1.84	0.0043
GO:0098798	GO_MITOCHONDRIAL_PROTEIN_COMPLEX	1.84	0.0073
GO:0015934	GO_LARGE_RIBOSOMAL_SUBUNIT	1.82	0.005
GO:0022626	GO_CYTOSOLIC_RIBOSOME	1.81	0.005

Supplementary Table 2.5.Top 25 Gene Sets Enriched in CMS2 Dataset as Measured by Normalized Enrichment Scores against CMS3.			
GO ID	**GO Term**	**|NES|**	**p.adjust**
GO:0034660	GO_NCRNA_METABOLIC_PROCESS	2.48	0.0029
GO:0042254	GO_RIBOSOME_BIOGENESIS	2.46	0.0029
GO:0034470	GO_NCRNA_PROCESSING	2.45	0.0029
GO:0022613	GO_RIBONUCLEOPROTEIN_COMPLEX_BIOGENESIS	2.44	0.0029
GO:0016072	GO_RRNA_METABOLIC_PROCESS	2.43	0.0029
GO:0030684	GO_PRERIBOSOME	2.41	0.0029
GO:0006270	GO_DNA_REPLICATION_INITIATION	2.40	0.0029
GO:0032543	GO_MITOCHONDRIAL_TRANSLATION	2.31	0.0029
GO:0006399	GO_TRNA_METABOLIC_PROCESS	2.29	0.0029
GO:0032040	GO_SMALL_SUBUNIT_PROCESSOME	2.28	0.0029
GO:0140053	GO_MITOCHONDRIAL_GENE_EXPRESSION	2.28	0.0029
GO:0042273	GO_RIBOSOMAL_LARGE_SUBUNIT_BIOGENESIS	2.21	0.0029
GO:0008033	GO_TRNA_PROCESSING	2.20	0.0029
GO:0006415	GO_TRANSLATIONAL_TERMINATION	2.16	0.0029
GO:0007143	GO_FEMALE_MEIOTIC_NUCLEAR_DIVISION	2.14	0.0029
GO:0120114	GO_SM_LIKE_PROTEIN_FAMILY_COMPLEX	2.11	0.0029
GO:0000313	GO_ORGANELLAR_RIBOSOME	2.09	0.0029
GO:0044391	GO_RIBOSOMAL_SUBUNIT	2.08	0.0029
GO:0003688	GO_DNA_REPLICATION_ORIGIN_BINDING	2.08	0.0029
GO:0070126	GO_MITOCHONDRIAL_TRANSLATIONAL_TERMINATION	2.08	0.0029
GO:0140101	GO_CATALYTIC_ACTIVITY_ACTING_ON_A_TRNA	2.08	0.0029
GO:0071826	GO_RIBONUCLEOPROTEIN_COMPLEX_SUBUNIT_ORGANIZATION	2.02	0.0029
GO:0000375	GO_RNA_SPLICING_VIA_TRANSESTERIFICATION_REACTIONS	2.02	0.0029
GO:0015934	GO_LARGE_RIBOSOMAL_SUBUNIT	2.02	0.0029
GO:0000184	GO_NUCLEAR_TRANSCRIBED_MRNA_CATABOLIC_PROCESS_NONSENSE_MEDIATED_DECAY	2.02	0.0029

Supplementary Table 2.6. Top 25 Gene Sets Enriched in CMS3 Dataset as Measured by Normalized Enrichment Scores against CMS1.

GO ID	GO Term	|NES|	p.adjust
GO:0009812	GO_FLAVONOID_METABOLIC_PROCESS	2.22	0.0032
GO:0072329	GO_MONOCARBOXYLIC_ACID_CATABOLIC_PROCESS	2.22	0.0037
GO:0009062	GO_FATTY_ACID_CATABOLIC_PROCESS	2.19	0.0037
GO:0006805	GO_XENOBIOTIC_METABOLIC_PROCESS	2.17	0.0037
GO:0034440	GO_LIPID_OXIDATION	2.16	0.0037
GO:0006635	GO_FATTY_ACID_BETA_OXIDATION	2.14	0.0035
GO:0003707	GO_STEROID_HORMONE_RECEPTOR_ACTIVITY	2.11	0.0034
GO:0006631	GO_FATTY_ACID_METABOLIC_PROCESS	2.07	0.0047
GO:0044242	GO_CELLULAR_LIPID_CATABOLIC_PROCESS	2.06	0.0043
GO:0006063	GO_URONIC_ACID_METABOLIC_PROCESS	2.04	0.0032
GO:0034308	GO_PRIMARY_ALCOHOL_METABOLIC_PROCESS	2.04	0.0035
GO:0016408	GO_C_ACYLTRANSFERASE_ACTIVITY	1.99	0.0032
GO:0005903	GO_BRUSH_BORDER	1.99	0.0037
GO:0045277	GO_RESPIRATORY_CHAIN_COMPLEX_IV	1.99	0.0032
GO:0001972	GO_RETINOIC_ACID_BINDING	1.93	0.0064
GO:0016614	GO_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_CH_OH_GROUP_OF_ DONORS	1.92	0.0037
GO:0034754	GO_CELLULAR_HORMONE_METABOLIC_PROCESS	1.92	0.0036
GO:0032787	GO_MONOCARBOXYLIC_ACID_METABOLIC_PROCESS	1.91	0.0060
GO:0071280	GO_CELLULAR_RESPONSE_TO_COPPER_ION	1.91	0.0081
GO:0004879	GO_NUCLEAR_RECEPTOR_ACTIVITY	1.90	0.0052
GO:0016042	GO_LIPID_CATABOLIC_PROCESS	1.90	0.0050
GO:0033559	GO_UNSATURATED_FATTY_ACID_METABOLIC_PROCESS	1.90	0.0035
GO:0015701	GO_BICARBONATE_TRANSPORT	1.89	0.0068
GO:0033293	GO_MONOCARBOXYLIC_ACID_BINDING	1.89	0.0055
GO:0033540	GO_FATTY_ACID_BETA_OXIDATION_USING_ACYL_COA_OXIDASE	1.88	0.0093

Supplementary Table 2.7. Top 25 Gene Sets Enriched in CMS3 Dataset as Measured by Normalized Enrichment Scores against CMS2.

GO ID	GO Term	|NES|	p.adjust
GO:0006063	GO_URONIC_ACID_METABOLIC_PROCESS	2.09	0.0029
GO:0071294	GO_CELLULAR_RESPONSE_TO_ZINC_ION	2.07	0.0029
GO:0015020	GO_GLUCURONOSYLTRANSFERASE_ACTIVITY	2.06	0.0029
GO:0009812	GO_FLAVONOID_METABOLIC_PROCESS	2.05	0.0029
GO:0071276	GO_CELLULAR_RESPONSE_TO_CADMIUM_ION	1.98	0.0029
GO:0010043	GO_RESPONSE_TO_ZINC_ION	1.96	0.0029
GO:0006805	GO_XENOBIOTIC_METABOLIC_PROCESS	1.96	0.0029
GO:0034754	GO_CELLULAR_HORMONE_METABOLIC_PROCESS	1.96	0.0029
GO:0004745	GO_RETINOL_DEHYDROGENASE_ACTIVITY	1.94	0.0029
GO:0005496	GO_STEROID_BINDING	1.92	0.0029
GO:0030258	GO_LIPID_MODIFICATION	1.91	0.0029
GO:0016042	GO_LIPID_CATABOLIC_PROCESS	1.91	0.0029
GO:0006635	GO_FATTY_ACID_BETA_OXIDATION	1.91	0.0029
GO:0034440	GO_LIPID_OXIDATION	1.89	0.0029
GO:0006654	GO_PHOSPHATIDIC_ACID_BIOSYNTHETIC_PROCESS	1.87	0.0029
GO:0007586	GO_DIGESTION	1.86	0.0029
GO:0006066	GO_ALCOHOL_METABOLIC_PROCESS	1.85	0.0029
GO:0042572	GO_RETINOL_METABOLIC_PROCESS	1.83	0.0047
GO:0001972	GO_RETINOIC_ACID_BINDING	1.82	0.0047
GO:0016408	GO_C_ACYLTRANSFERASE_ACTIVITY	1.81	0.0078
GO:0032411	GO_POSITIVE_REGULATION_OF_TRANSPORTER_ACTIVITY	1.81	0.0029
GO:0005319	GO_LIPID_TRANSPORTER_ACTIVITY	1.81	0.0029
GO:0044242	GO_CELLULAR_LIPID_CATABOLIC_PROCESS	1.80	0.0029
GO:0008028	GO_MONOCARBOXYLIC_ACID_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	1.79	0.0062
GO:0042445	GO_HORMONE_METABOLIC_PROCESS	1.79	0.0029

Chapter 3:

Lipopolysaccharide from Microbes Associated with Consensus Molecular Subtypes of Colorectal Cancer have Antagonistic Effects on Cytokine Production in Peripheral Blood Mononuclear Cells

Sulit, A.K.[1], Daigneault, M.[3], Allen-Vercoe, E.[3], Silander, O.K.[1], Hock, B.[4], McKenzie, J.[4], Pearson, J.[5], Frizelle, F.A.[2], Schmeier, S.[1,6], Purcell, R.[2]

[1]School of Natural Sciences, Massey University, Auckland, New Zealand
[2]Department of Surgery, University of Otago, Christchurch, New Zealand
[3]Department of Molecular and Cellular Biology, University of Guelph, Ontario, Canada
[4]Haematology Research Group, University of Otago, Christchurch, New Zealand
[5]Biostatistics and Computational Biology Unit, University of Otago, Christchurch, New Zealand
[6]Evotec SE, Hamburg, Germany

Article in preprint (doi: 10.1101/2022.04.26.489473)

Author contributions:

AKS performed the computational analyses with advice and input from **OKS, SS,** and **RP.** AKS carried out the experimental work in the study with materials provided by **MD** and **EA-V,** and experimental advice from **BH** and **JM**. AKS carried out the statistical analyses with advice from **JP** and **BH. AKS** wrote the manuscript with editorial comments by **RP, EA-V, SS, OKS, BH,** and **FAF. FAF** provided guidance on all clinical aspects of the manuscript.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Arielle Kae L. Sulit
Name/title of Primary Supervisor:	Dr. Olin Silander
In which chapter is the manuscript /published work:	3

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

○ The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

S

◉ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	
Date:	16-Nov-2021
Primary Supervisor's Signature:	
Date:	16 Nov 2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

3.1 Background

Colorectal cancer (CRC) tumors exist in a complex microenvironment whose composition greatly affects the tumor's progression, prognostics, and therapy response (Angell et al., 2020; Colangelo et al., 2017). These microenvironment components include cancer associated fibroblasts (CAF), responsible for normal maintenance of the basement membrane; endothelial cells, pericytes, and platelets, which have roles in development and function of blood and lymph vessels, and when perturbed could lead to angiogenesis; and immune cells. Microbes may also affect the balance in this tumor microenvironment (Colangelo et al., 2017).

Immune-cell infiltration has vital, but conflicting key roles in CRC. Tumor-specific antigens can prime the immune system for tumor cell killing by tumor infiltrating lymphocytes (TILs) and natural killer (NK) cells. Within the microenvironment, these various signals may also contribute to chronic inflammation or immune evasion and escape, as well as signals for cell growth and vascular changes (Colangelo et al., 2017; Markman & Shiao, 2015). Cytokines have also been shown to have opposing effects on cancer progression, with cytokines IL-6, IL-17, and IL-1β generally viewed as tumor promoting and IL-12, IFN-γ, and IL-18 as tumor suppressive (Mager et al., 2016; West et al., 2015). Some, such as IL-10, may have opposing effects, being a regulatory cytokine (Mager et al., 2016). T-cells, whose presence are largely favorable in cancer, can also promote the production of growth factors and tumor promoting cytokines (Waldner et al., 2006).

The importance of immune responses in CRC is reflected in its subtypes. CRC subtypes are categories ascribed to tumors with the same set of defining characteristics. Different studies have proposed different subtyping systems based on a variety of characteristics (De Sousa E Melo et al., 2013; Guinney et al., 2015; Marisa et al., 2013; Rodriguez-Salas et al., 2017; Roepman et al., 2014; Sadanandam et al., 2013), but common among these is one subset characterized by microsatellite instability (MSI) and immune activation, as well as a subset with angiogenic and mesenchymal characteristics (Rodriguez-Salas et al., 2017). A meta-analysis has identified four consensus molecular subtypes (CMS1 - CMS4) (Guinney et al., 2015). Subsequent studies on the immune responses in the molecular subtypes have shown that while CMS1 (MSI-Immune) has increased numbers of cells associated with immune activation, CMS4 (mesenchymal) has an inflamed phenotype and is characterized by immunosuppressive cell populations (Becht et al., 2016; Dienstmann et al., 2017; Karpinski et al., 2017). CMS4 is associated with the lowest overall survival rate (Guinney et al., 2015), while CMS1 has the lowest survival rate after relapse. MSI tumors generally have favorable overall prognosis (Becht et al., 2016; Hale et al., 2018). Understanding the complicated balance involved in immune responses and CRC progression is important for improving disease prognosis and therapy responses.

Our understanding of the role immune responses have in CRC prognoses cannot be complete without an awareness of the influence of the microbiome. This is evidenced by the decreased tumor susceptibility in germ-free rats compared to conventional rats upon carcinogen introduction (Janney et al., 2020; Reddy et al.,

1974); differences in tumor susceptibility between mice populations with different microbiomes (A. I. Yu et al., 2020, p. 8); findings that specific gut microbiota improves responses to CTLA-4 blockade immunotherapy (Vétizou et al., 2015); and differences in the microbiomes of cancers with deficient versus proficient mismatch repair functions (Hale et al., 2018).

The differences in the immune landscapes of CMS1 and CMS4 tumors provide an opportunity for us to investigate how microbes in these tumor subtypes may contribute to the immune responses leading to either anti-tumor activity, or tumor progression. We therefore sought to determine how microbes from the CMS1 and CMS4 subtypes could potentially mediate the immune environments characteristic of these respective tumor subtypes.

3.2 Methodology

3.2.1 Sample Collection and Handling

We collected tumor samples during surgical resection of colorectal tumors from 308 patients who had not received chemotherapy prior to surgery. Patients with diagnosed hereditary nonpolyposis colorectal cancer (HNPCC) or familial adenomatous polyposis (FAP) were excluded. All participants provided written and informed consent, and the study was approved by the University of Otago Human Ethics Committee with approval number ***H16/037.*** During surgery, samples were collected, frozen in liquid nitrogen, and stored at -80^{0}C. Before RNA extraction,

samples were first transferred to RNAlater ICE™ (Qiagen) and equilibrated at-20^{0}C for at least 48 hours (Visnovska et al., 2019).

RNEasy Plus Mini Kit (Qiagen) was used to extract RNA from 15-20mg of tissue which had been disrupted using a Retsch Mixer Mill, including a DNAse treatment step in the procedure. We quantified purified RNA using a NanoDrop 2000c spectrophotometer (Thermo Scientific, Asheville, NC, USA) and subsequently stored the RNA at -80^{0}C.

3.2.2 RNA Sequencing

Library preparation for RNA sequencing was carried out using the Illumina TruSeq Stranded Total RNA Library preparation kit (Illumina), with ribosomal RNA depletion using Ribo-Zero Gold. We used the Illumina Hi-Seq 2500 V4 platform to carry out RNA sequencing, producing 125 bp paired-end reads. Each sample library was split equally into two lanes to avoid technical bias, and were later merged during the data processing phase. Merged data can be found under Bioproject ID PRJNA788974 in the NCBI SRA database.

3.2.3 Consensus Molecular Subtype Classification

SolexaQA++ (Cox et al., 2010) was used to trim reads, which were then run through Salmon (Patro et al., 2017) to quantify transcript expression. The publicly available

CRC CMS classifier (Guinney et al., 2015) was used to categorize samples into one of four CMSs as described in Purcell et al. (2019).

3.2.4 Bioinformatics Analysis

After trimming with SolexaQA++, we ran the 260 samples through the MetaFunc pipeline (A. K. L. Sulit et al., 2020) using default settings, unless otherwise stated. Briefly, MetaFunc uses STAR (Dobin et al., 2013) to map reads to a human genome reference (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_33/GRCh38.primary_assembly.genome.fa.gz), which are then quantified using featureCounts (reverse stranded option) (Liao et al., 2014) from the subread package and organized into per sample gene expression matrices for downstream analysis using custom scripts within the MetaFunc pipeline. Unmapped reads from the mapping to the human genome are then parsed into Kaiju (Menzel et al., 2016) for microbial identification using NCBI's *nr* database. Gene Ontology annotations of the protein matches from Kaiju are also output by custom scripts in the MetaFunc pipeline. To be included in downstream analysis, a species was required to have at least 0.01% abundance in at least one of the 260 samples. Bacteria, Archaea, Fungi, and Viruses were selected in the `*TaxChoices*` section of the configuration file. Databases used were those provided in https://metafunc.readthedocs.io/en/latest/usage.html#databases.

3.2.5 Differential Expression and Gene Set Enrichment Analysis in Host

From the results of the MetaFunc analysis, we gathered the host-gene expression raw counts table into a matrix for input into DESeq2 (Love et al., 2014) with metadata information on their respective CMS. Using the 260 samples that had been classified into a subtype, DESeq2 was used to obtain differentially expressed genes (DEGs) in one CMS compared to the average of the other three subtypes. Using the p-values and log-fold change gathered through DESeq2 for ranking and sign information, respectively, we performed gene-set enrichment analysis on the ranked genes resulting from DESeq2, using the clusterProfiler package (G. Yu et al., 2012) with the C5 Ontology Gene Sets collection (version 7) from the molecular signatures database (MSigDB) (Liberzon et al., 2011; Subramanian et al., 2005). Specifically, we ranked the genes using the formula:

$$\text{rank} = - \log10(\text{p-value}) * \text{sign}(\log_2\text{FoldChange}),$$

with sign being directionality (+ or -) of the $\log_2$ fold change value, and actual value was not used. This ranking places the genes with lowest p-values and positive $\log_2$ fold change at the top of the list, and the genes with lowest p-values and negative $\log_2$ fold change at the bottom of the list. Genes at the top of the list contribute to gene sets with positive enrichment scores and genes at the bottom contribute to gene sets with negative enrichment scores.

We considered a gene set enriched in a subtype if it had an adjusted p-value of <0.05 and positive Normalized Enrichment Score (NES). We interrogated the enriched gene sets in the CMS subtypes using immune-related keywords as follows: "IMMUN", "T CELL", "INTERFERON", "CYTOKINE", "TOLL LIKE", "LYMPHOCYTE", "LEUKOCYTE",

"PATTERN RECOGNITION", "LIPOPOLYSACCHARIDE", "MHC", "INFLAMMATORY", "ANTIGEN", "INTERLEUKIN", and "BACTERIA".

Detailed host analysis may be found at https://gitlab.com/alsulit08/2021_uoc_massey_lps-crc/-/tree/master/Bioinformatics commit 8dbc30e3.

3.2.6 Differential Abundance of Microbes in the Microbiomes of CRC Subtypes

From the results of the MetaFunc analysis, we gathered raw counts of microbe taxonomies into a Phyloseq object (McMurdie & Holmes, 2013), with metadata information on their respective CMS. We used DESeq2 to identify differentially abundant microbes in either CMS1 or CMS4 compared to the average of the other three subtypes. We considered microbes to be differentially abundant in a CMS if it had a $\log_2$ fold change > 0 and adjusted p-value < 0.05.

Detailed analysis may be found at https://gitlab.com/alsulit08/2021_uoc_massey_lps-crc/-/tree/master/Bioinformatics commit 8dbc30e3.

3.2.7 Lipopolysaccharide-Associated Bacteria

MetaFunc produces a table that indicates which Bacterial Taxonomy IDs have proteins that are annotated with gene ontology terms. We searched for “lipopolysaccharide biosynthetic process” or “lipid A biosynthetic process” terms,

obtaining all Bacterial species with proteins annotated with these terms in any CMS1 or CMS4 sample. We cross-referenced these with our differentially abundant microbes in CMS1 or CMS4, respectively, obtaining differentially abundant bacteria that have proteins annotated with LPS-related processes.

Detailed analysis may be found at https://gitlab.com/alsulit08/2021_uoc_massey_lps-crc/-/tree/master/Bioinformatics commit 8dbc30e3.

3.2.8 Lipopolysaccharide from Bacterial Strains

We extracted LPS from strains of *Fusobacterium periodonticum* (*1/1/54 (D10), 2/1/31,* and *1/1/41 FAA*), *Bacteroides fragilis* (*3/2/5, 2/1/16,* and *2/1/56 FAA*), and *Porphyromonas asaccharolytica* (*CC44 001F, CC1/6* F2) using Bacterial Lipopolysaccharides (LPS) Extraction Kit (Alpha Diagnostic International, Catalog # 1000-100-LPS) as per the manufacturer's instructions, resulting in a final yield of 30 µg/mL of LPS. *F. periodonticum* and *B. fragilis* strains are part of the Human Microbiome Project reference strains collection (Huttenhower et al., 2012; Methé et al., 2012), and *P. asaccharolytica* strains were previously isolated under British Columbia (BC) Cancer Institutional Review Board approval as part of a larger project to recover oncomicrobial species from colorectal cancer biopsy specimens.

3.2.9 PBMC Treatment with LPS from Bacterial Species

PBMCs from nitrogen stocks were washed in GIBCO DPBS solution, counted, and again resuspended in PBMC medium (RPMI with 10%FCS, 1% glutamine, 0.2% penicillin/streptomycin). In 96 well plates, we incubated cultures of PBMCs ($2x10^5$ cells) with LPS preparations of varying concentrations (600 ng/mL, 60 ng/mL or 6 ng/mL) from *B. fragilis* (strains *3/2/5, 2/1/16,* and *2/1/56 FAA*), *F. periodonticum* (strains *1/1/54 (D10), 2/1/31,* and *1/1/41 FAA*), or *P. asaccharolytica* (strains *CC44 001F, CC1/6* F2) overnight. For co-incubation tests, we used *B. fragilis* strain *2/1/16, F. periodonticum* strain *2/1/31*, and *P. asaccharolytica* strain *CC1/6 F2*. We treated the PBMCs with 6 ng/mL of *F. periodonticum* (strain *2/1/31*) LPS and 600 ng/mL of *B. fragilis* (strain *2/1/*16) or *P. asaccharolytica* (strain *CC1/6 F2*) LPS. As no-treatment controls, PBMC medium (RPMI, 10%FCS, 1% glutamine, 0.2% Penicillin/Streptomycin) or RPMI alone was added to the initial culture of PBMCs.

Co-incubation experiments were conducted at least three times. For each repeated experiment, PBMCs were obtained from a different individual.

3.2.10 Measurement of Cytokine Production and Statistical Analysis

We measured secreted cytokine expression following the manufacturer's instructions for LegendPlex Human Inflammation Panel 1 (Cat no. 740809) on a Beckman Coulter Cytomics FC500 Flow Cytometry Analyzer. LegendPlex Human Inflammation Panel 1 measures IL-1β, IFN-α2, IFN-γ, TNF-α, MCP-1, IL-6, IL-8, IL-12p70, IL-10, IL-17A, IL-18, IL-23, and IL-33. For all runs, baseline values of the cytokines from untreated PBMCs

were obtained. We used the Legendplex Software (Windows version 8 or MacOS version 7.1, using the 5-parameter curve fitting model) to assess final concentrations of the cytokines tested. Final cytokine concentrations were obtained from at least two technical replicates per run as calculated by the software. Paired student's t-tests were used to test for differences in cytokine production between a) PBMC baseline and *F. periodonticum* alone; b) *F. periodonticum* alone and *F. periodonticum* with *B. fragilis;* and c) *F. periodonticum* alone and *F. perioconticum* with *P. asaccharolytica*. Effect Sizes were obtained using Cohen's d, with Hedges correction set to true to account for small sample sizes. Statistical analyses were carried out using the rstatix (Kassambara, 2021) package in R 3.6.1 (R Core Team, 2019) and details may be found at https://gitlab.com/alsulit08/2021_uoc_massey_lps-crc/-/tree/master/LPS_Experiments commit 463b5446.

3.3 Results

3.3.1 Patient Cohort Characteristics

The cohort comprised 308 colorectal cancers, all taken prior to chemotherapy. Subtyping of the 308 samples classified 260 samples, with 60 samples in CMS1, 145 in CMS2, 38 in CMS3, and 17 in CMS4. For subsequent analysis, we focused on the 260 classified samples.

The mean age of the 260 patients was 71.78 years. There were 139 females and 121 males. The majority of the tumors were from the colon (81.15%) while the rest were

from the rectum (18.85%). Of the colon tumors, 113 were right-sided (proximal tumors) while 98 were left-sided. Together, the left-sided and rectal tumors are categorized as distal tumors (147 in total). **Table 3.1** summarizes these characteristics, and gives more details of demographics per CMS group.

Table 3.1 Cohort Characteristics by Consensus Molecular Subtype (CMS)

	All (n=260)	CMS1 (n=60)	CMS2 (n=145)	CMS3 (n=38)	CMS4 (n=17)
Age, mean ± SD	71.78 ± 11.39	76.2 ± 10.00	70.11 ± 11.18	71.03 ± 12.99	72.06 ± 10.71
Sex					
Male	121 (46.54%)	16 (27.67%)	83 (57.24%)	14 (36.84%)	8 (47.06%)
Female	139 (53.46%)	44 (73.33%)	62 (42.76%)	24 (63.16%)	9 (52.94%)
Site					
Colon	211 (81.15%)	59 (98.33%)	110 (75.86%)	32 (84.21%)	11 (64.71%)
Rectum	49 (18.85%)	1 (1.67%)	35 (24.14%)	6 (15.79%)	6 (35.29%)
Side					
Proximal	113 (43.46%)	48 (80.00%)	42 (28.97%)	19 (50.00%)	4 (23.53%)
Distal*	147 (56.54%)	12 (20%)	103 (71.03%)	19 (50.00%)	13 (76.47%)
Stage					
1	48 (18.46%)	10 (16.67%)	25 (17.24%)	11 (28.95%)	2 (11.76%)
2	112 (43.08%)	33 (55.00%)	65 (44.83)	10 (26.32%)	4 (23.53%)
3	83 (31.92%)	16 (26.67%)	43 (29.66%)	16 (42.10%)	8 (47.06%)
4	17 (6.54%)	1 (1.67%)	12 (8.28%)	1 (2.63%)	3 (17.65%)

n = number of patients
***Distal = contains both left-sided and rectal cancers**

3.3.2 CMS1 and CMS4 have Enriched Gene Sets Involved in Immune Response

In order to compare which gene sets were enriched in different subtypes of CRC, we first used DESeq2, to compare gene expression in CMS1 samples with the average of the CMS2, CMS3, and CMS4 samples. There were 4736 genes that were significantly over expressed (adjusted p-value < 0.05) in CMS1 compared to the other subtypes. We carried out gene-set enrichment analysis (GSEA) on the list of genes as described in the methods section, and obtained those with positive normalized enrichment scores (NES) as gene sets enriched in CMS1. We obtained a total of 329 gene sets that had positive NES in CMS1, with adjusted p-value < 0.05. To provide an overview of which functional clusters were present in CMS1 samples, we plotted the top gene sets by NES and p-values, in an enrichment map (**Supplementary Figure 3.1**). We found that several of the most enriched gene sets were related to immune responses (cytokines, antigen processing and presentation, and cell killing processes) as well as nuclear organization and replication processes. We therefore subset the enriched gene sets in CMS1 using immune-related keywords (see **Methods**). As it has been theorized that microbes might affect the balance of immune responses in the tumor microenvironment (TME), we sought to identify if response to microbes is captured among the enriched gene sets of our subtypes by including "BACTERIA" in the keywords.

We performed the same analyses for CMS2, CMS3, and CMS4, against the average of the other three subsets. We obtained no enriched gene sets using the

immune-related keywords above in CMS2 and CMS3, consistent with previous studies describing these as “immune-neglected” (Fidelle et al., 2020). The top enriched gene sets for CMS4 primarily comprised terms corroborating its epithelial-mesenchymal-transition (EMT) and angiogenic characteristics (**Supplementary Figure 3.2**). As studies have mentioned a role for immune cells in immunosuppression in CMS4 (Becht et al., 2016; Dienstmann et al., 2017; Karpinski et al., 2017), we examined the 1186 enriched gene sets in CMS4, and found 61 enriched gene sets that contained immune-related keywords.

We found an overlap of enriched immune-related gene sets between CMS1 and CMS4 (**Figure 3.1**). These included production of interleukin 6, cytokine secretion, and T-cell activation, all of which could lead to immune-induced cytotoxic activity that could destroy cancer cells, or chronic inflammation and escape in favor of cancer progression (Fisher et al., 2014; Mager et al., 2016; Markman & Shiao, 2015; West et al., 2015). Some gene sets had prominently higher enrichment scores in CMS1 compared to CMS4; among these was the gene set for T-cell activation. T-cell infiltration in CRC has been associated with better survival (Galon et al., 2006; Ganesh et al., 2019). We also found *Response to Molecule of Bacterial Origin* among these enriched gene sets, indicating the role microbes likely play in the characteristics of these subtypes.

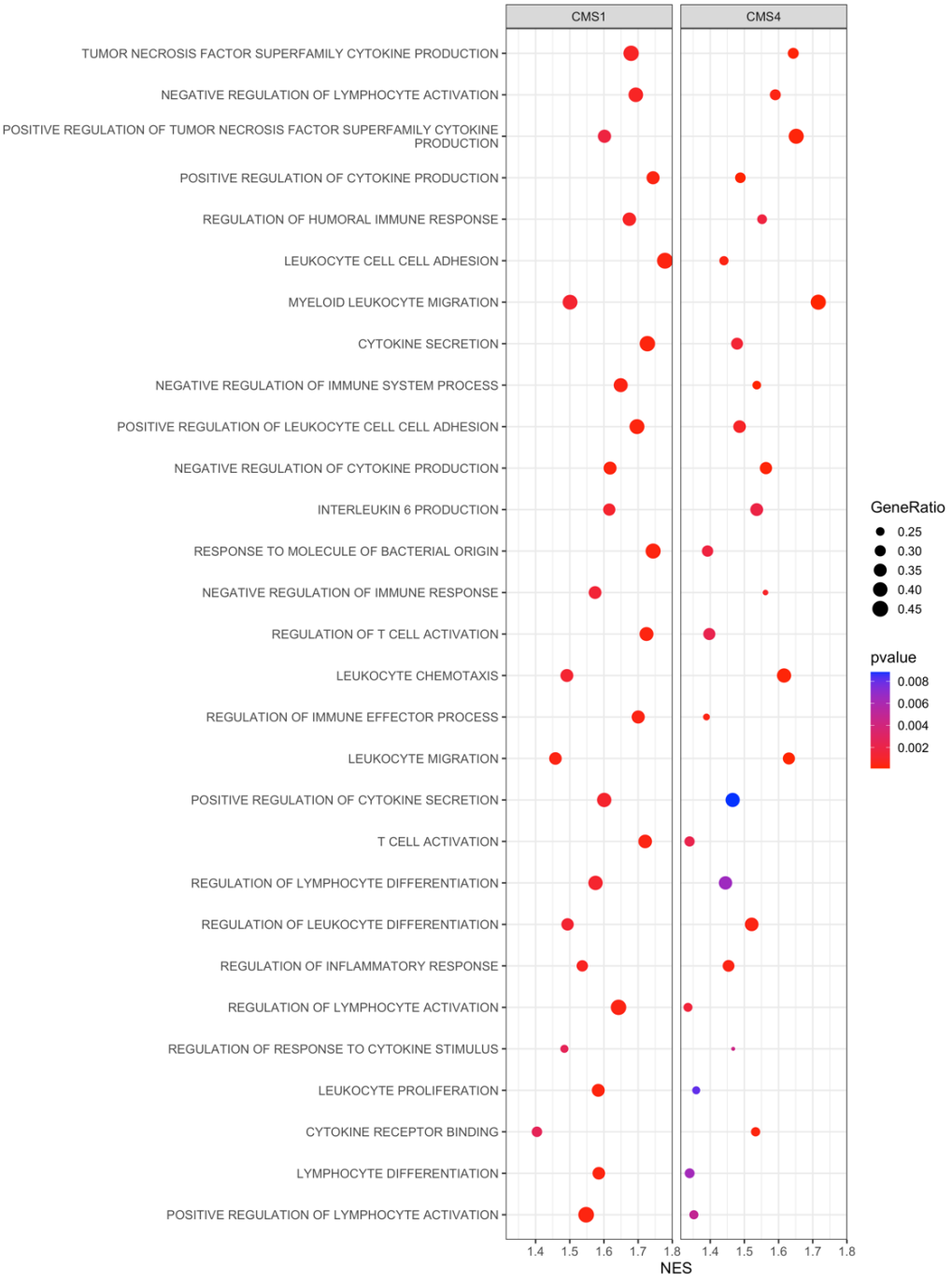


Figure 3.1. Common immune-related enriched gene sets in CMS1 and CMS4. CMS1 and CMS4 were found to be enriched in gene sets relating to immune responses. A number of these were found in both subtypes. Of interest is the gene set Response to Molecule of Bacterial Origin, indicating a possible role of microbes in these subtypes. All adjusted p-values < 0.05.

NES: normalized enrichment score

Gene Ratio: number of genes contributing to enrichment of the gene set from the dataset divided by total number of genes in the gene set in question

We found 81 immune-related gene sets unique to CMS1 (**Figure 3.2**). Several of these (e.g. antigen processing and presentation, MHC protein) indicate increased levels of antigen presentation, an early critical process in the induction of antitumor responses (Reeves & James, 2017).

Gene sets indicating activity of natural killer cells, and production and response to interleukins, including interleukin-12 and interleukin-1 beta, were also unique to CMS1, in addition to a wide array of functions indicating regulation and homeostasis of immune responses, including regulation of apoptosis of leukocytes, lymphocytes, and T-cells. The apoptosis of T-cells has been implicated in tumor immune resistance and escape (Huber et al., 2005), and can act as mechanisms of immune checkpoints (Zhu et al., 2019). Regulation (both positive and negative) of these processes may affect how CMS1 tumors progress. Genes associated with Toll-like receptor activity and lipopolysaccharide signaling pathways were also identified in CMS1, again suggesting that bacteria may play a role in the responses observed.

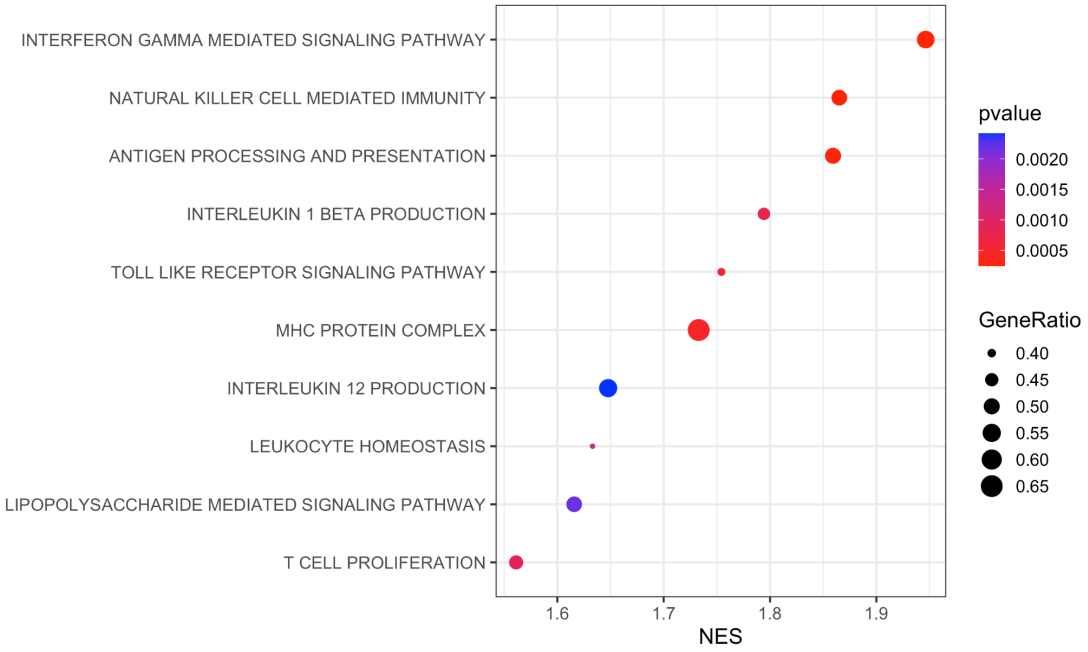


Figure 3.2. Representative subset of immune-related enriched gene sets unique to CMS1. Immune-related gene sets unique to CMS1 include those involved in antigen presentation, Natural Killer Cell activity, production of interleukins and homeostasis in immune responses. The gene set for lipopolysaccharide-mediated signaling pathway is also enriched, indicating a response to LPS in microbes. All adjusted p-values < 0.05.

NES: normalized enrichment score

Gene Ratio: genes contributing to enrichment of the gene set from the dataset divided by total set size of the gene set in question

CMS4 had 32 unique immune related enriched gene sets (**Figure 3.3** and **Supplementary Figure 3.3**), several of which are associated with negative regulation of T-cells and other immune responses that suggest an association with immunosuppression, consistent with previous studies (Becht et al., 2016; Colangelo et al., 2017). Lipopolysaccharide binding was also enriched in CMS4, suggesting a link to bacterial regulation, as seen in CMS1.

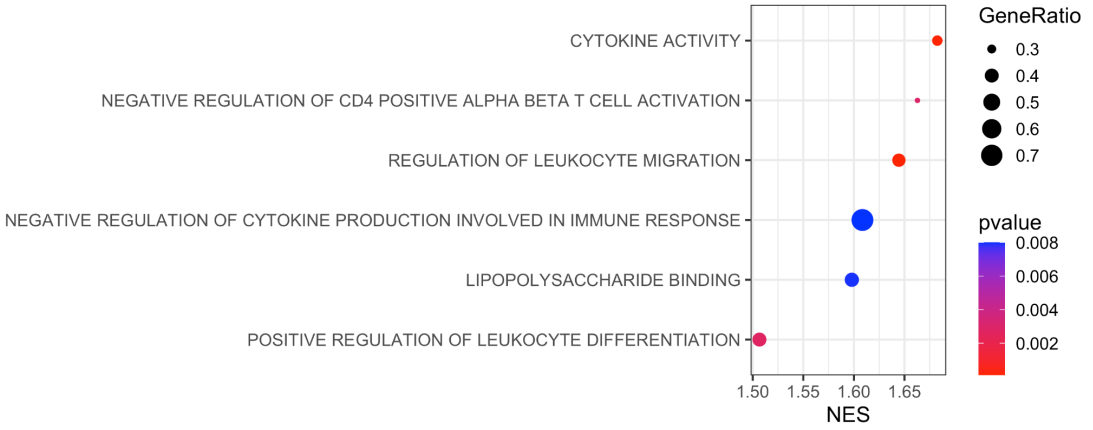


Figure 3.3. Representative subset of immune-related enriched gene sets unique to CMS4. There are fewer immune-related gene sets enriched in CMS4, which seem to be associated with immunosuppression. Among the enriched gene sets is lipopolysaccharide binding, indicating a response to LPS in microbes. All adjusted p-values < 0.05.

NES: normalized enrichment score

Gene Ratio: genes contributing to enrichment of the gene set from the dataset divided by total set size of the gene set in question

Taken together, our findings show that CMS1 has more immune-related enriched gene sets that include functionalities involving natural killer cell activity, T-cell infiltration, adaptive immune responses (e.g. antigen processing, MHC protein complex), and Toll-like receptor functions, as well as active regulation of immune processes. This is consistent with previous work showing that the CMS1 immune landscape is anti-tumorigenic, adaptive, and has a favorable prognosis (Becht et al., 2016; Dienstmann et al., 2017; Karpinski et al., 2017). In contrast, we see fewer immune-related enriched gene sets in CMS4 and several gene sets indicating a negative regulation of T-cell activation, which may serve to dampen the effects of T-cells, allowing tumor tolerance.

3.3.3 *Fusobacterium* and *Bacteroides fragilis* species Contribute to LPS Biosynthetic Processes in CMS1

We found 295 microbial species that were differentially abundant (DA) in CMS1 compared to the other CMS subtypes (adjusted p-value <0.05). As a bridge between the DEGs in our host dataset and these DA microbes, we selected for bacteria that have proteins annotated with the Gene Ontology terms "Lipid A Biosynthetic Process" or "Lipopolysaccharide Biosynthetic Process" (see **Methods**), which identified 20 bacterial species out of the 295 microbes (**Figure 3.4a**). Notably, we identified *Fusobacterium* and *Bacteroides* as among the abundant bacteria with LPS processes. These two genera have previously been implicated in the progression of CRC (Dai et al., 2018; Goodwin et al., 2011; Wu et al., 2009; X. Ye et al., 2017). We focused on *Fusobacterium periodonticum* and *Bacteroides fragilis* (**Figure 3.4b**). Enterotoxigenic *B. fragilis* (ETBF) is associated with immune activation, and the *B. fragilis* toxin (bft) causes an increase in reactive oxygen species (ROS) production and DNA damage, as well as E-cadherin cleavage leading to cell proliferation (Goodwin et al., 2011; Wu et al., 2009). *B. fragilis* is also among the most enriched bacterial species in a multi-cohort analysis of CRC (Dai et al., 2018). *F. periodonticum* was reported to be enriched in deficient MMR (dMMR) tumors compared to proficient MMR (pMMR tumors) (Hale et al., 2018) and found in CRC tissues in several studies reviewed for the role of *Fusobacteria* in CRC (Hussan et al., 2017). We used these bacteria to investigate the potential interaction between the *F. periodonticum* and *B. fragilis* LPS and immune responses in host cells.

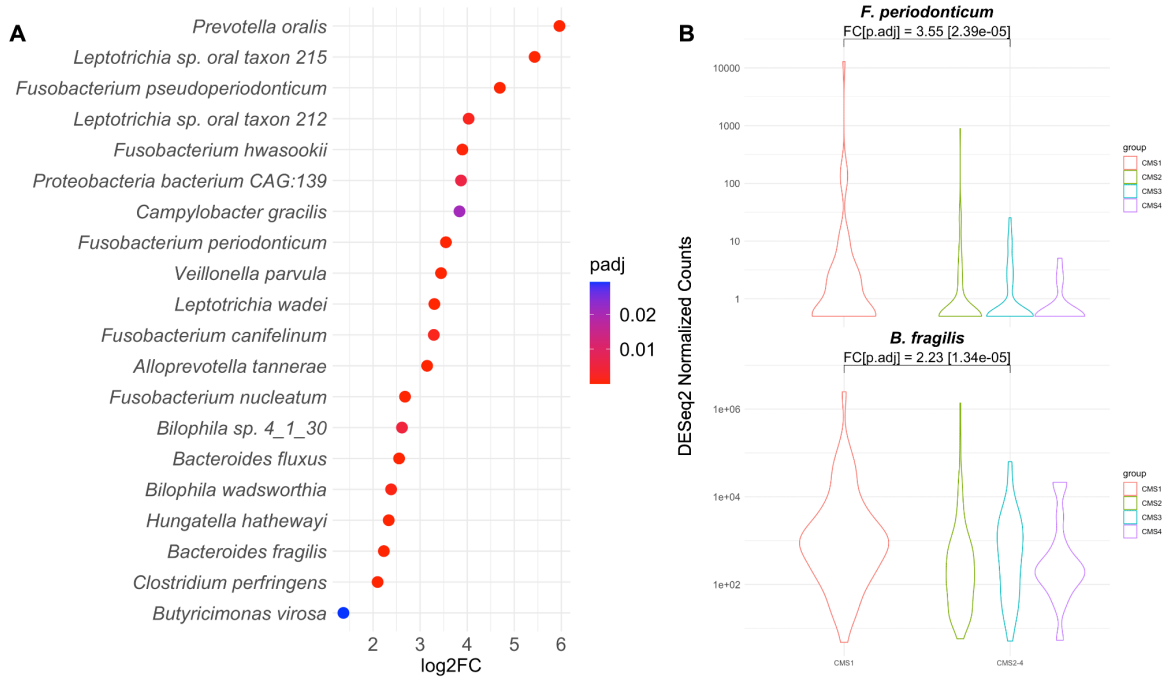


Figure 3.4. Differentially abundant microbes in CMS1. A. Differentially abundant (DA) Bacteria in CMS1 with lipopolysaccharide or lipid-A biosynthetic process annotations. Among the DA microbes found in CMS1 compared to the other three subtypes, we identified microbes that are annotated with LPS-related processes. **B.** *F. periodonticum* and *B. fragilis* DESeq2 normalized counts in CMS1 compared to other subtypes. We focused on *F. periodonticum* and *B. fragilis.* These plots show their count distributions in CMS1 compared to their count distributions in the other three subtypes.

log2FC=$\log_2$ fold change, *FC*=$\log_2$ fold change, *p.adj*=adjusted p-value

3.3.4 CMS4 has Fewer Differentially Abundant Bacteria with LPS Processes

In contrast to the 295 microbes that we found to be differentially abundant in CMS1, only 127 microbes were found to be differentially more abundant in CMS4 compared to the other three subtypes. Of these, there were only three bacteria with proteins annotated with LPS or Lipid-A biosynthetic processes (**Figure 3.5A**). One of these, *P. asaccharolytica* (**Figure 3.5B**), has previously been identified as a CRC marker in

multi-cohort analyses (Dai et al., 2018; Thomas et al., 2019), and we selected it to further investigate its effects on the immune response using *in vitro* experiments.

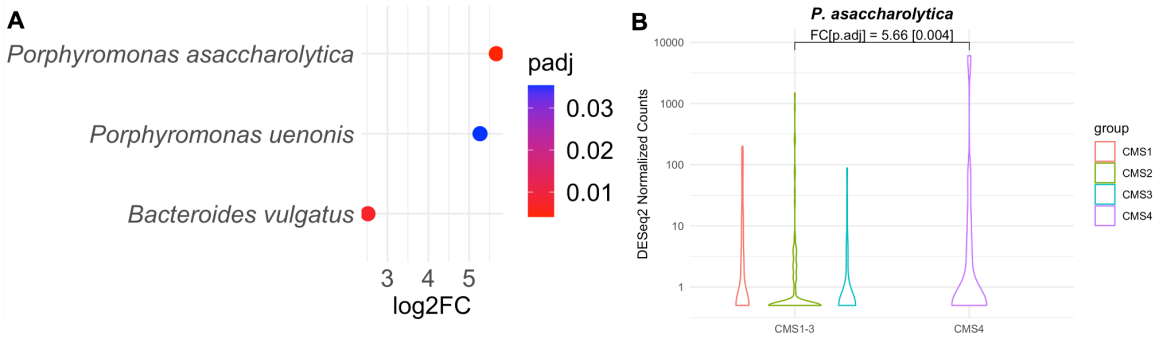


Figure 3.5. Differentially abundant microbes in CMS4. A. Differentially abundant (DA) Bacteria in CMS4 with lipopolysaccharide or lipid-A biosynthetic process annotations. Among the DA microbes found in CMS4 compared to the other three subtypes, we identified microbes that are annotated with LPS-related processes. **B.** *P. asaccharolytica* normalized counts in CMS4 compared to other subtypes. We chose to focus on *P. asaccharolytica.* This plot shows its count distribution in CMS1 compared to its count distribution in the other three subtypes.
log2FC=$\log_2$ fold change, *FC*=$\log_2$ fold change, *p.adj*=adjusted p-value

3.3.5 LPS from Different Bacterial Species have Different Effects on Cytokine Release

We wanted to determine the effects of the LPS from the three selected bacteria on immune cells, specifically on the release of cytokines. We therefore incubated stock cultures of PBMCs with increasing concentrations of LPS, extracted from *F. periodonticum (*strains *1/1/54 (D10), 2/1/31, and 1/1/41 FAA) , B. fragilis (*strains *3/2/5, 2/1/16, and 2/1/56 FAA),* or *P. asaccharolytica (*strains *CC44 001F,* and *CC1/6 F2)* overnight. We focused on the release of IFN-γ, IL-6, IL-10, IL-12p70, IL-1β, and IL-18, as these were prominent cytokines seen in our gene set enrichment analysis; we included IL-18 because it has been found to synergise with IL-12 to increase

production of IFN-γ in T-cells (Tominaga et al., 2000), and IL-10 because it is known as a regulatory cytokine (J. Li et al., 2020; Mager et al., 2016; West et al., 2015). We found that all strains within a species had similar effects on cytokine production.

LPS from both *B. fragilis* and *P. asaccharolytica* inhibited the release of the measured cytokines in a concentration dependent manner, with the highest degree of inhibition observed at 600 ng/mL (**Figure 3.6A and 3.6B**, respectively; **Supplementary Figure 3.4**). Conversely, LPS from *F. periodonticum* strains showed stimulatory effects on the cytokines of interest, with low concentrations of LPS (6 ng/mL) causing an increase in their secretion compared to baseline PBMC levels (**Figure 3.6C.; Supplementary Figure 3.4**). No further increases were observed at higher LPS levels.

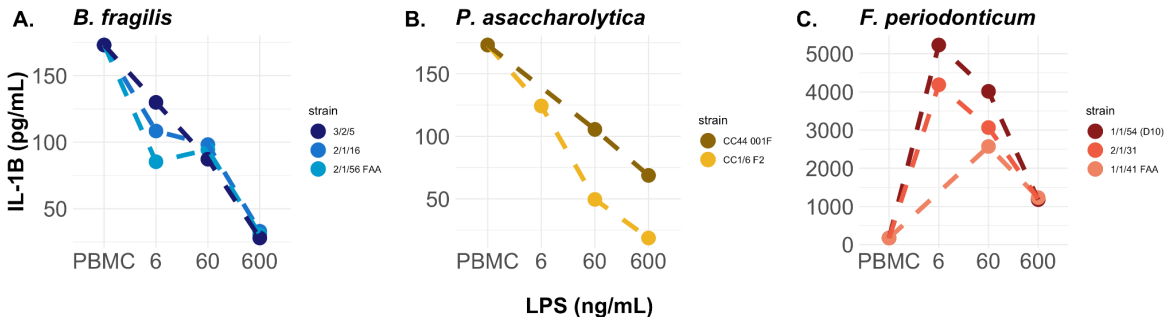


Figure 3.6. Secreted IL-1β concentrations (pg/mL) after overnight incubation of PBMCs with different LPS concentrations (6 ng/mL, 60 ng/mL, 600 ng/mL) from different strains of A.) *B. fragilis,* B.) *F. periodonticum*, and C.) *P. asaccharolytica* compared to PBMC baseline (no treatment). Dashed lines connect data points from the same strains. Colors differentiate the strains of each species. Similar trends were seen for other cytokines of interest (Supplementary Figure 3.4).

As all strains exhibited the same effects on cytokine release, we chose a single strain from each species, *F. periodonticum 2/1/31, B. fragilis 2/1/16,* and *P. asaccharolytica CC1/6 F2,* to use in subsequent experiments.

3.3.6 *B. fragilis* and *P. asaccharolytica* LPS Inhibit Stimulatory Effects of LPS from *F. periodonticum* in Co-cultures

The findings described above suggest that *B. fragilis* and *P. asaccharolytica* LPS can inhibit cytokine release, and *F. periodonticum* LPS can stimulate it. To further investigate the immune modulatory effects of different LPS, we tested whether LPS from *B. fragilis 2/1/16* or *P. asaccharolytica CC1/6 F2* could retain their inhibitory properties on PBMCs when co-cultured with LPS from immunostimulatory *F. periodonticum 2/1/31.* For these experiments we used the lowest concentration of *F. periodonticum* LPS (6 ng/mL) and the highest concentration of LPS from *B. fragilis* or *P. asaccharolytica* (600 ng/mL), as these respective concentrations had the largest effects on cytokine production in earlier experiments (**Figure 3.6, Supplementary Figure 3.4**).

PBMCs were incubated with either *F. periodonticum* LPS alone, or in combination with either *B. fragilis* or *P. asaccharolytica* LPS, and cytokine production was measured. We found that the increased cytokine production observed in the presence of *F. periodonticum* LPS alone was attenuated by co-incubation with LPS from *B. fragilis* or *P. asaccharolytica* (**Figure 3.7**). *B. fragilis* LPS, when co-cultured with LPS from *F. periodonticum,* reduced cytokine secretion towards baseline levels

and this reduction was significant for IL-1β, IL-18, and IL-6 (Cohen's d greater than |0.8| for all; **Figure 3.7; Supplementary Table 3.1**). We found a similar trend with *P. asaccharolytica* LPS, where its addition caused a reduction of cytokine production towards baseline PBMC levels, and was significant for IL-10, IL-1β, IL-18, and IL-6 (Cohen's d greater than |0.8| for all; **Figure 3.7; Supplementary Table 3.2**).

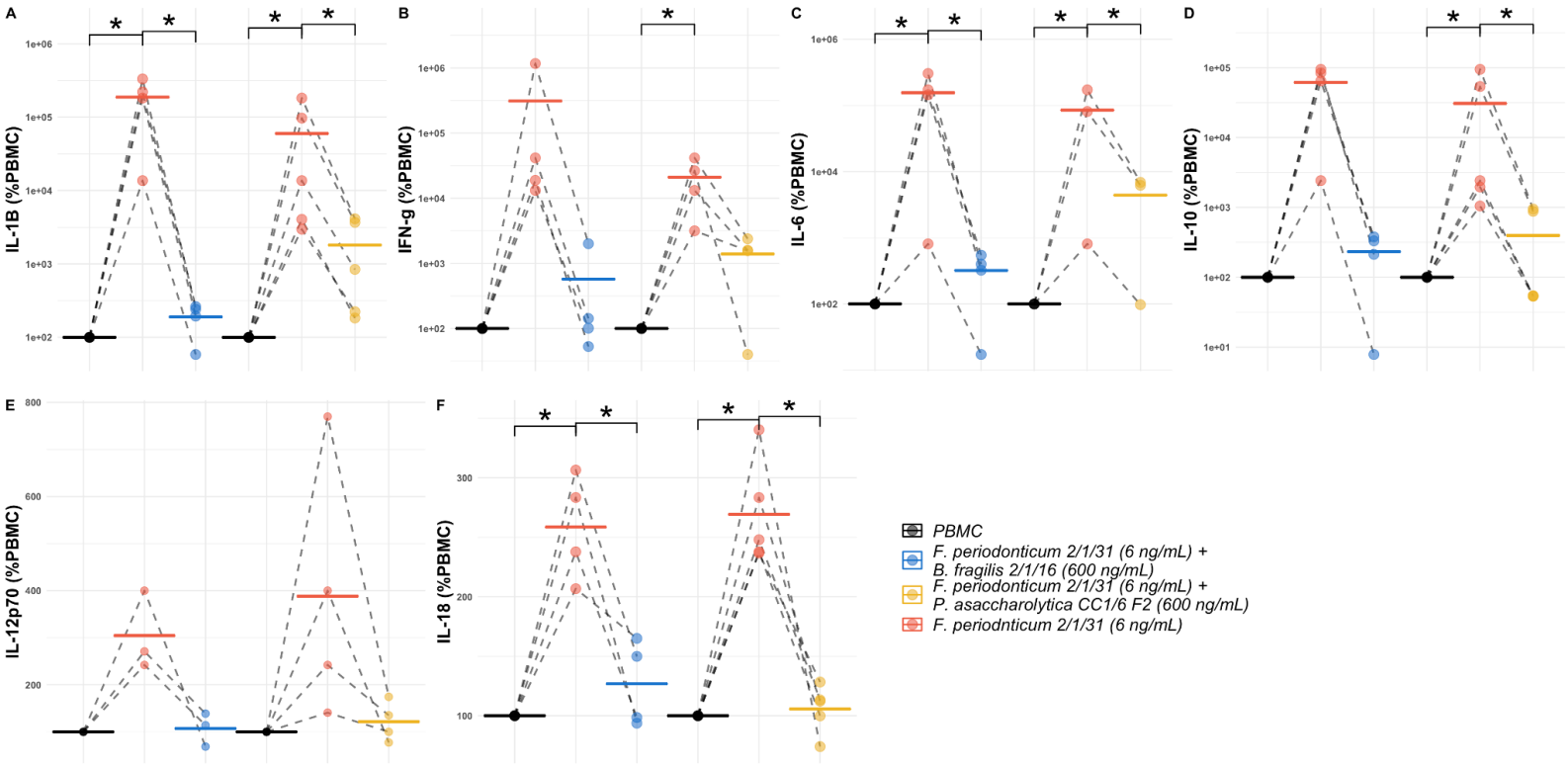


Figure 3.7. Changes in cytokine expression in peripheral blood mononuclear cells (PBMCs) following treatment with *F. periodonticum* alone (red) or in combination with *B. fragilis* (blue) or *P. asaccharolytica* (yellow). We show results for cytokines of interest **A.** IL-1β, **B.** IFN-γ, **C.** IL-6, **D.** IL-10, **E.** IL-12p70, and **F.** IL-18. *F. periodonticum 2/1/31* LPS used was at a concentration of 6 ng/mL**.** *B. fragilis 2/1/16* LPS and *P. asaccharolytica CC1/6 F2* LPS used were at a concentration of 600ng/mL. Values are shown as percentages of PBMC baseline secretion, which is set at 100%. Dashed lines indicate a single experimental run. Horizontal lines represent the mean of repeat experiments. Y-axes of A, B, C, and D are in $\log_{10}$ scale, while Y-axes of E and F are in linear scale. *PBMC=PBMC baseline; (*) = t-test p-value < 0.05*

3.4 Discussion

Studies of immune response in CRC have reported conflicting findings, where they can either induce tumor regression and cell death, or lead to cancer progression through chronic inflammation and even influence angiogenic signaling. This delicate balance between pro- and anti-tumor effects is reflected in the consensus molecular subtypes (CMS) of CRC (Guinney et al 2015), where CMS1 and CMS4 have immune infiltrates of differing compositions (Becht et al., 2016; Dienstmann et al., 2017; Fidelle et al., 2020; Karpinski et al., 2017). The two subtypes have contrasting prognoses, with CMS1 tumors, characterized with MSI, having favorable prognosis and CMS4 having the worst relapse-free survival. These phenomena highlight the complexity and heterogeneity of CRC, which leads to difficulties in clinical management, prognosis, and treatment. Subtyping efforts have paved the way to stratify CRC into clinically useful categories, but our understanding of the molecular mechanisms underpinning tumor heterogeneity is far from complete. While most subtyping efforts are focused on molecular differences between subtypes, the microbiome, whose contributions to CRC progression are now widely accepted, has largely been omitted from the subtype discussion. The interplay between microbiome and host molecular pathways, and the overall tumor microenvironment dynamics is, we believe, an important facet of CRC heterogeneity, with repercussions for prognosis and therapy response.

Previous studies have shown that microbes affect inflammation-induced tumorigenesis (Brennan & Garrett, 2016; J. Chen et al., 2017; Reddy et al., 1974; A. I.

Yu et al., 2020, p. 8). An increase in immune activation in response to several CRC-associated bacteria has also been reported: Enterotoxigenic *Bacteroides fragilis* (ETBF) has been shown to induce IL-17 and IL-23 responses leading to colon tumorigenesis in mice (Wu et al., 2009); *Fusobacterium nucleatum* has been shown increase the expression of the cytokines IL-6, IL-8, and IL-18 in CRC cell lines and xenograft mice (Rubinstein et al., 2013); *F. nucleatum, B. fragilis,* and *E. coli* were further found to stimulate production of chemokines that favor recruitment of beneficial T- cells in CRC cell lines and mouse models (Cremonesi et al., 2018).

In this study, we combined metatranscriptomics and human transcriptomics data to interrogate how the microbiome could affect the conflicting effects of immune responses reported in CRC, informing a mechanistic hypothesis that could be tested in the laboratory. Most studies attributing function to the microbiome have largely been associative (Dai et al., 2018; Thomas et al., 2019; Wirbel et al., 2019), and while important analytical results regarding functional contributions of the microbiome had been made, these were seldom linked to host gene expression and functional contributions.

Our analysis of host gene sets confirmed previously published reports that CMS1 and CMS4 are immune infiltrated and inflamed, respectively, while CMS2 and CMS3 have little immune activation. The enriched gene sets in CMS1 and CMS4, relating to response to bacteria and to LPS, indicated a role for the microbiome in the characteristics of these subtypes, and are more prominent in CMS1. Among the

bacteria that had detectable LPS biosynthetic processes, we found *F. periodonticum* and *B. fragilis* to be abundant in CMS1, and *P. asaccharolytica* in CMS4.

Due to the previously reported characteristics of CMS1 and CMS4, we hypothesized that *F. periodonticum* and *B. fragilis* would increase tumor-fighting cytokines IFN-γ, IL-12, and IL-18, and decrease tumor promoting cytokines IL-1β and IL-6 while the opposite might happen with *P. asaccharolytica* from CMS4 tumors. However, our results indicated that *F. periodonticum* LPS increased the production of all cytokines of interest, including IL-10, while both *P. asaccharolytica* and *B. fragilis* decreased the production of all of these. Moreover, LPS from *P. asaccharolytica* and *B. fragilis* attenuated the cytokine stimulatory activity of LPS from *F. periodonticum,* indicating that interactions between these microbes may occur in the tumor microenvironment, further adding to the complexity of immune responses.

Reviews summarizing the roles of cytokines in CRC progression categorize IL-1β and IL-6 as tumor-promoting, due to their pro-inflammatory characteristics and induction of cell proliferation, and IFN-γ, IL-12, and IL-18 as anti-tumorigenic, due to their abilities to induce cytotoxic immune-cell responses, such as that of Natural Killer and $CD8^+$ T-cells; IL-10, considered a regulatory cytokine, could tilt the balance either way (J. Li et al., 2020; Mager et al., 2016; West et al., 2015). Although these general attributes may be ascribed to cytokines, many of these still display pleiotropic functions that may have opposing effects in CRC. For instance, decreased levels of IL-1β, along with IL-18, are correlated with increased colitis-associated cancer (CAC) in an inflammasome context (Allen et al., 2010; Baker et al., 2019); IL-6 has

anti-tumorigenic properties in the form of priming effector T-cells (Fisher et al., 2014); and unchecked IFN-γ could compromise the colonic epithelial barrier (Ferrier et al., 2003; Mager et al., 2016) allowing an influx of microbiota that could influence cancer progression. This seeming pleiotropy and paradoxical role of cytokines suggests that balance and control in cytokine production is vital in determining whether it is pro- or anti-tumorigenic and may be reflected in the interactions we see in response to the bacterial LPS we have tested.

Although our results with *B. fragilis* were the opposite of our hypothesis and initially unexpected, previous work has shown that LPS activity, conserved among the *Bacteroidales* order, can be immunosuppressive by promoting tolerance to the high microbial load that comes with hosting a complex microbial ecosystem (d'Hennezel et al., 2017). Indeed, some species of the *Bacteroides* genus, including *B. fragilis*, have been found to have immunoregulatory properties (Di Lorenzo et al., 2020; Tan et al., 2019; Yoshida et al., 2018). Further, it was identified that ETBF could promote colonic tumors and induce inflammatory processes but not its NTBF (non-toxigenic) counterpart (Wu et al., 2009), indicating that it is the toxin that is necessary for its inflammatory role and not LPS. We were unable to identify whether the *B. fragilis* in our CRC samples were toxigenic or non-toxigenic, or whether the toxigenic strains were expressing the BFT toxin; the BFT toxin may also only be produced within a certain timeframe during carcinogenesis (Purcell, Pearson, et al., 2017). There is also a possibility of a mixture of NTBF and ETBF in our tumor samples, and this combination may indicate a nuanced balance between LPS and *B. fragilis* toxin, along with the more immunogenic LPS of other microbes such as *F. periodonticum*. These

studies, together with our findings in this study, indicate that while LPS from *B. fragilis* may be immunosuppressive, interaction with immunogenic molecules produced by the same or other microbes add a layer of complexity to immune responses in carcinogenesis. Indeed, a balance is necessary to keep immune responses in check. In CMS1, we postulate that this balance skews towards immune activation, contributing to anti-tumor effects and positively affecting prognosis, while the immunosuppressive LPS of *P. asaccharolytica* may contribute to immune evasion and escape of tumors in CMS4.

The cytokine-release inhibiting capabilities of LPS from *B. fragilis* and *P. asaccharolytica* are notable as previous studies emphasize the pro-inflammatory activities of CRC-associated microbiota (J. Chen et al., 2017; Cremonesi et al., 2018; Janney et al., 2020; Rubinstein et al., 2013; Wu et al., 2009). While we acknowledge that these events also occur within our tumor samples, we also suggest that microorganisms play a role in immunosuppression, either by aiding the progression of tumorigenesis through immune evasion and escape, or contributing to homeostasis of immunogenic processes.

The limitations of the study include the use of PBMCs as proxies for immune cells in the tumor microenvironment. PBMCs may not adequately reflect the effects of LPS on tumor-infiltrating lymphocytes. *In vitro* cultures of single cell types do not allow for crosstalk between different cell populations, which would be expected in the complex tumor microenvironment. Furthermore, we acknowledge that while we have tested the effects of LPS from single species and pairs, colorectal tumors may

harbor up to hundreds of different species, and each may elicit an effect dependent on LPS structure and absolute bacterial counts, which could contribute to the nuanced immune-modulation within the TME. In addition, the effects of LPS may be countered or exacerbated by other known bacterial mechanisms, e.g. bacterial toxins, or as yet undiscovered interactions.

3.5 Conclusion

In this study we have assigned CRC patient samples into one of the four Consensus Molecular Subtypes, and have focused on understanding the interplay of the microbiome in the immune-infiltrated CMS1 and CMS4 subtypes. We confirmed that these two subtypes have enriched gene sets that are related to immune response processes, whereas CMS2 and CMS3 do not. Gene sets that indicate immune response to microbial triggers were found among the enriched sets of both CMS1 and CMS4 prompting an investigation into the microbes differentially abundant in CMS1 and CMS4, and containing LPS biosynthetic processes. We found that *F. periodonticum* LPS, found in CMS1 tumors, increased production of cytokines IL-1β, IFN-γ, IL-18, IL-10, IL-6, and IL-12p70 in an *in vitro* model, while LPS from *B. fragilis*, found in CMS1 tumors, and *P. asaccharolytica* LPS, found mainly in CMS4 tumors, decreased production of these cytokines and could also attenuate the immunogenic effect of *F. periodonticum* LPS. Where most previous studies have focused on the inflammation-inducing capabilities of CRC-associated microorganisms, our results indicated that their immunosuppressive potential should not be overlooked and adds another layer of complexity in immune responses in CRC. These immunosuppressive

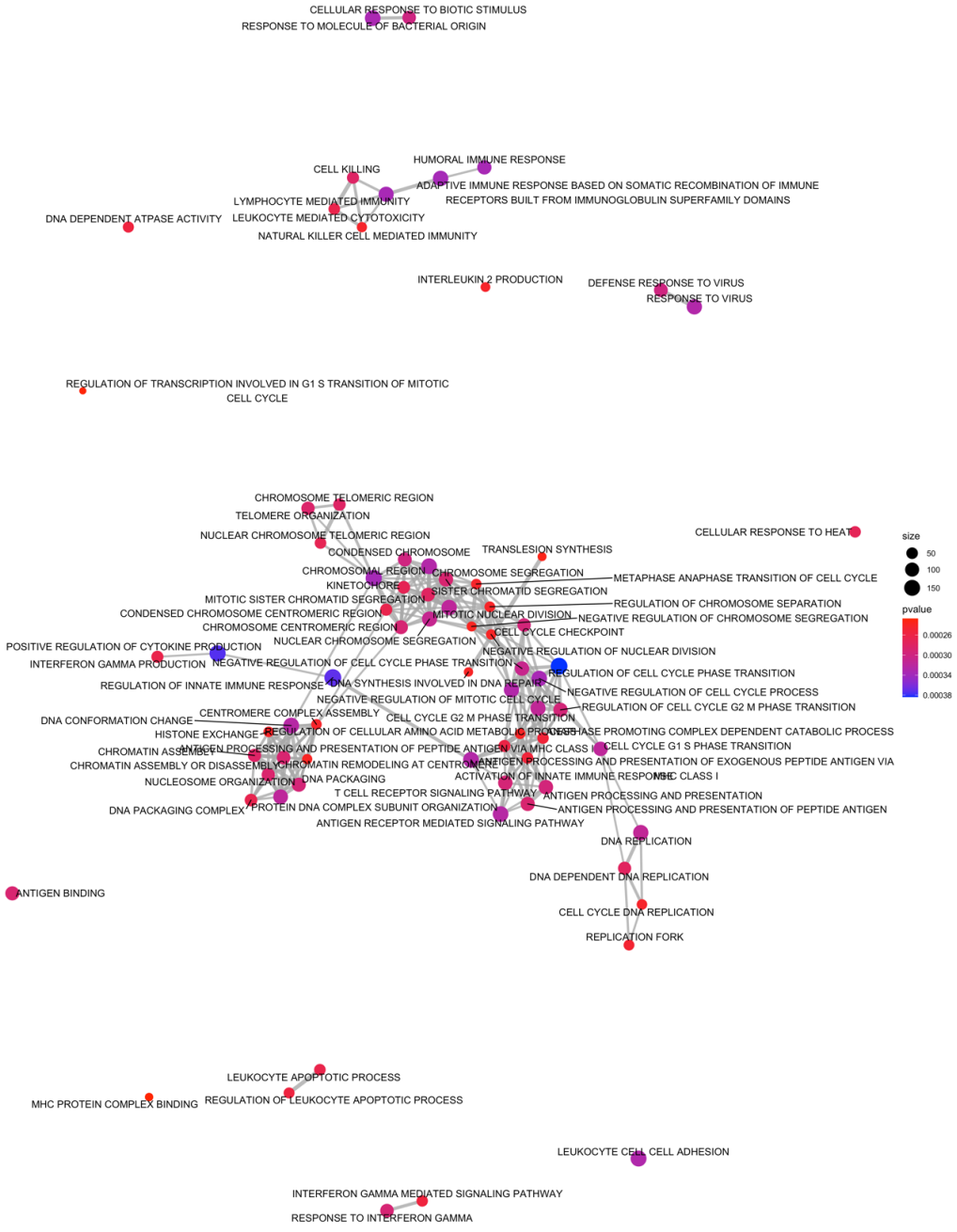
effects could work towards immune-response homeostasis, ensuring an anti-tumorigenic response is attained without overflowing to chronic inflammation, or on the other extreme, contributing to immune tolerance, evasion and escape which could lead to metastasis. We also emphasize that these effects occur as a result of stimulation with LPS from different bacterial species, and other microbial molecules from the same species could elicit a different immune response.

3.6 Supplementary Material

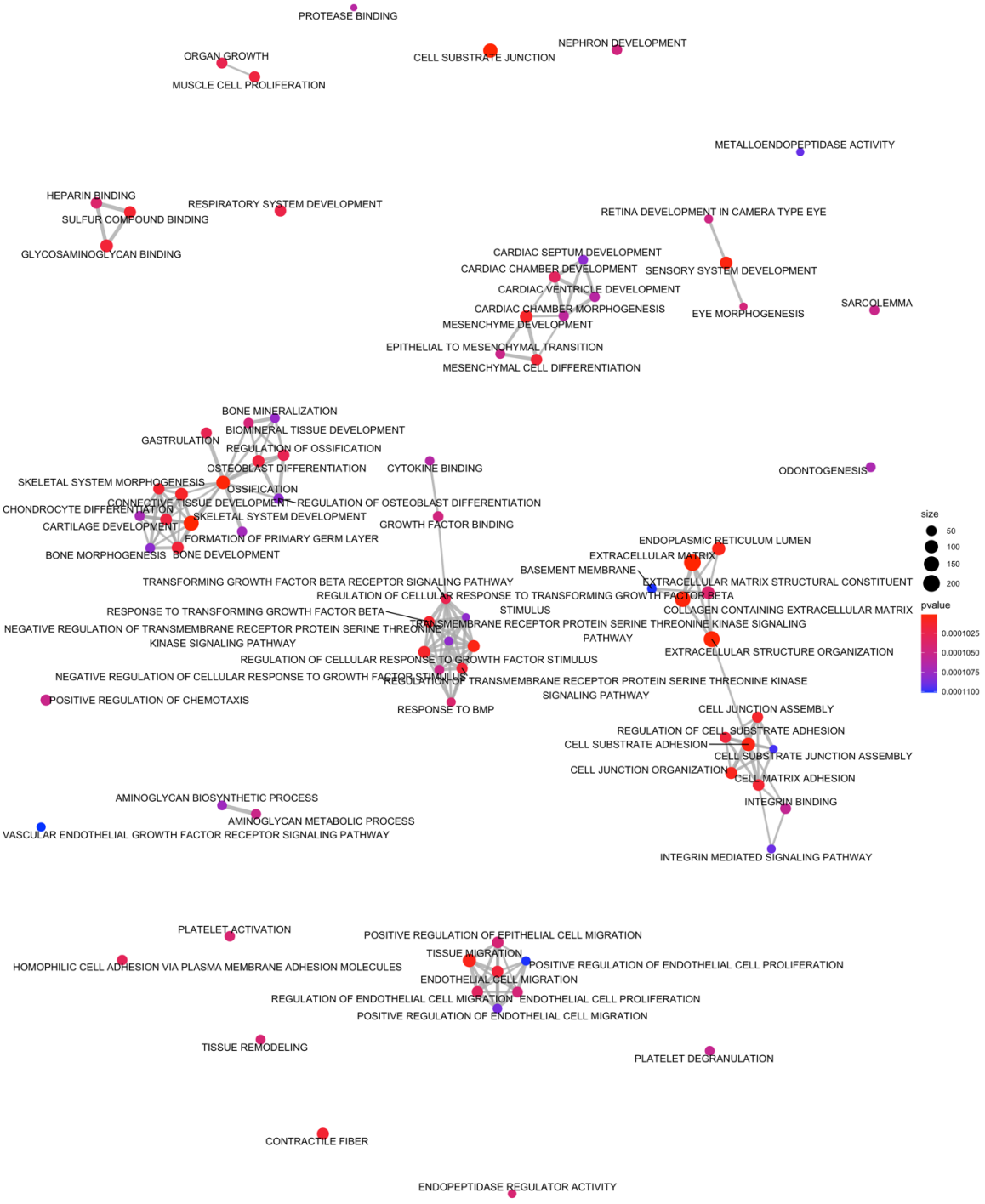
3.6.1 Supplementary Methods

For co-incubation tests, we tested *B. fragilis* strain *2/1/16, F. periodonticum* strain *2/1/31*, and *P. asaccharolytica* strain *CC1/6 F2*. We treated the PBMCs with 6 ng/mL of *F. periodonticum* (strain *2/1/31*) LPS and 3 concentrations of *B. fragilis* (strain *2/1/*16) or *P. asaccharolytica* (strain *CC1/6 F2*) LPS (600 ng/mL, 60 ng/mL, 6 ng/mL). As no-treatment controls, PBMC medium (RPMI, 10%FCS, 1% glutamine, 0.2% Penicillin/Streptomycin) or RPMI alone was added to the initial culture of PBMCs. As we saw that the highest degree of effect was seen in the highest concentrations (600 ng/mL) of *B. fragilis* or *P. asaccharolytica* (**Supplementary Figures 3.5 and 3.6**), we focused on these results in the results and discussion parts of this chapter.

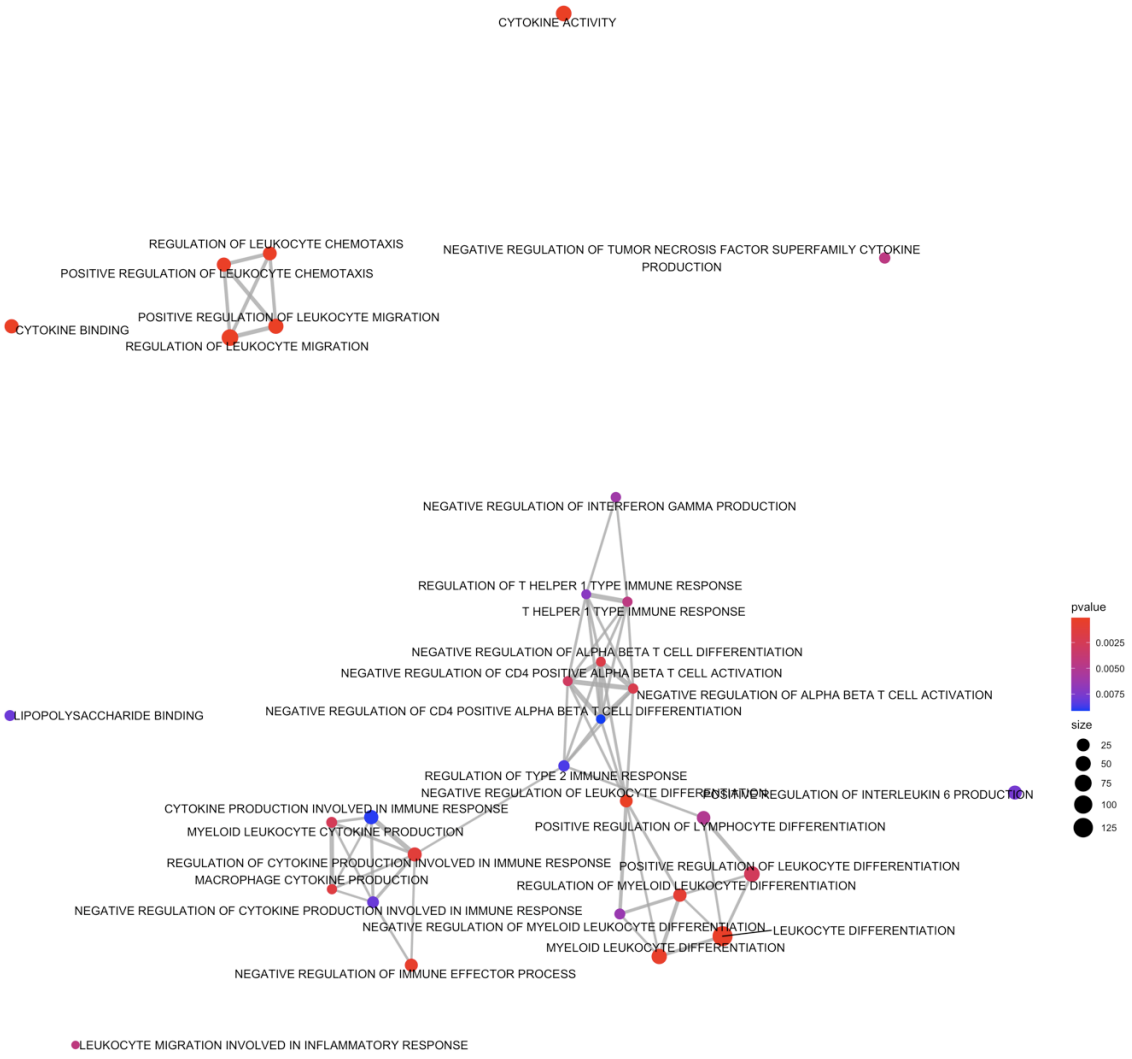
3.6.2 Supplementary Figures



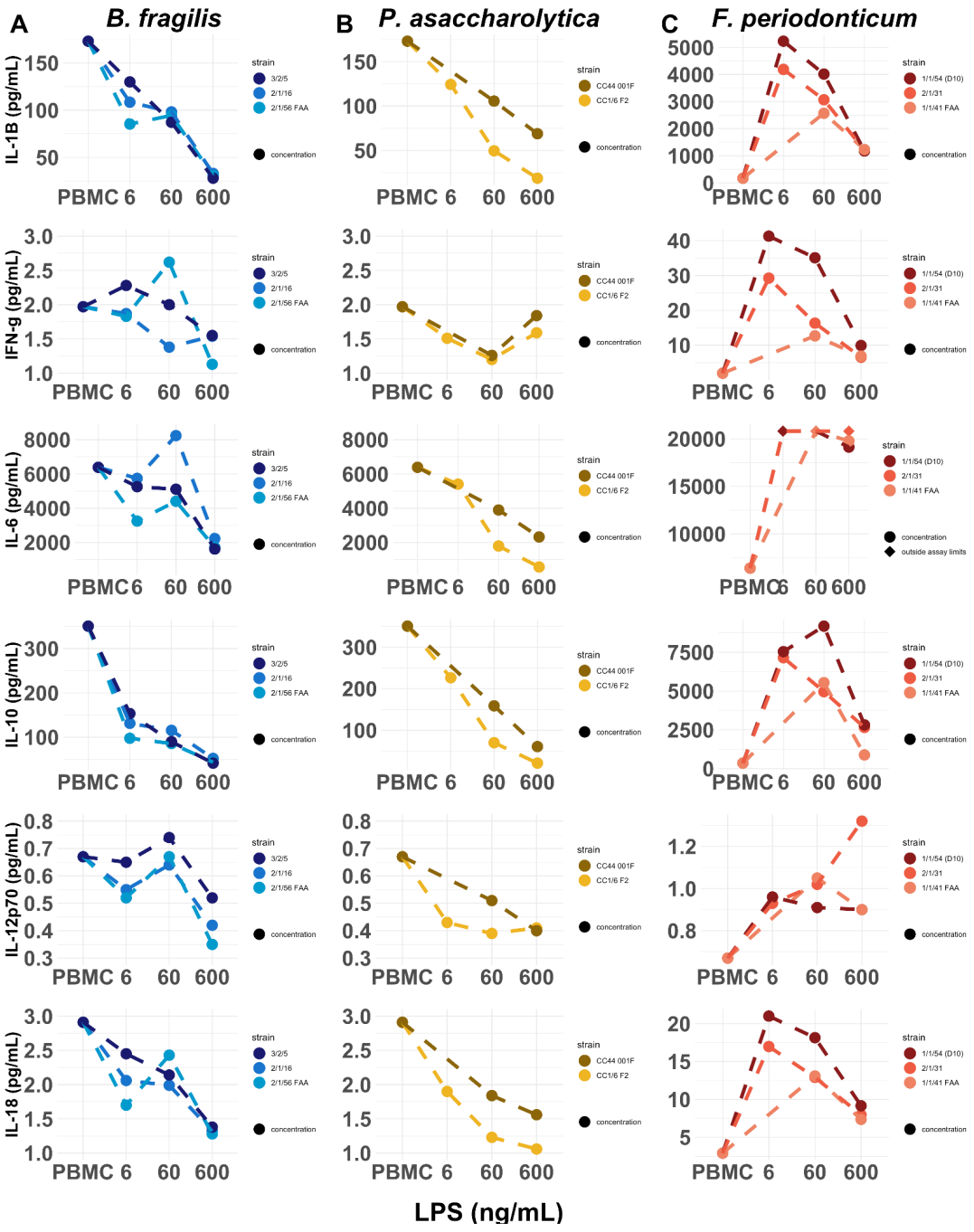
Supplementary Figure 3.1. Enrichment map of the top enriched gene sets in CMS1 by normalized enrichment score (NES) and p-values. Enrichment maps show which gene sets have shared genes between them (gene sets with connecting lines share genes). CMS1 enriched gene sets have functional clusters involved in immune responses, as well as nuclear organization and DNA replication processes. All adjusted p-values < 0.05.



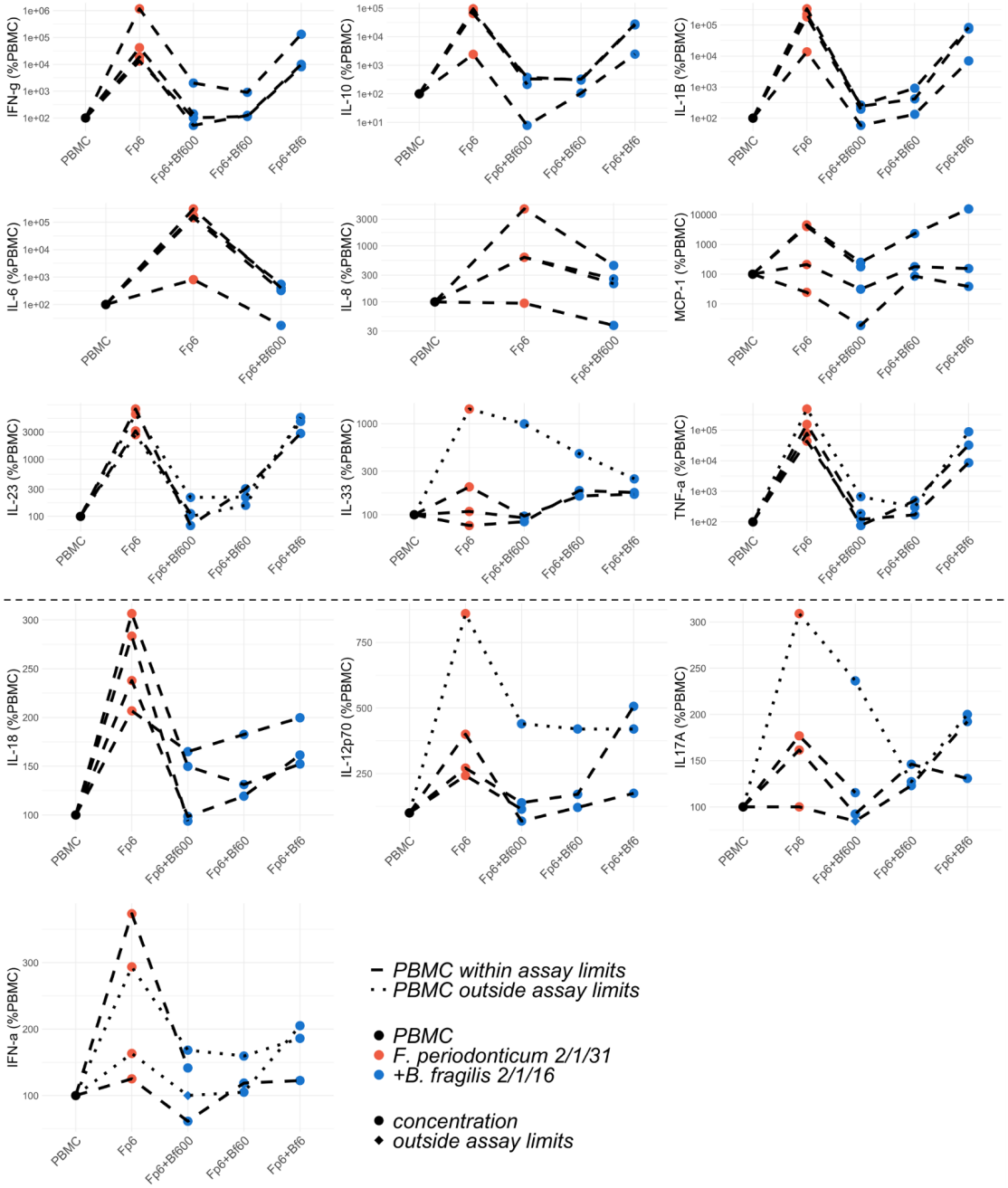
Supplementary Figure 3.2. Enrichment map of the top enriched gene sets in CM4 by normalized enrichment score (NES) and p-values. Enrichment maps show which gene sets have shared genes between them (gene sets with connecting lines share genes). CMS4 enriched gene sets have functional clusters involved in epithelial-mesenchymal transition (EMT), development, cell adhesion, proliferation and migration, and growth factors. All adjusted p-values < 0.05.



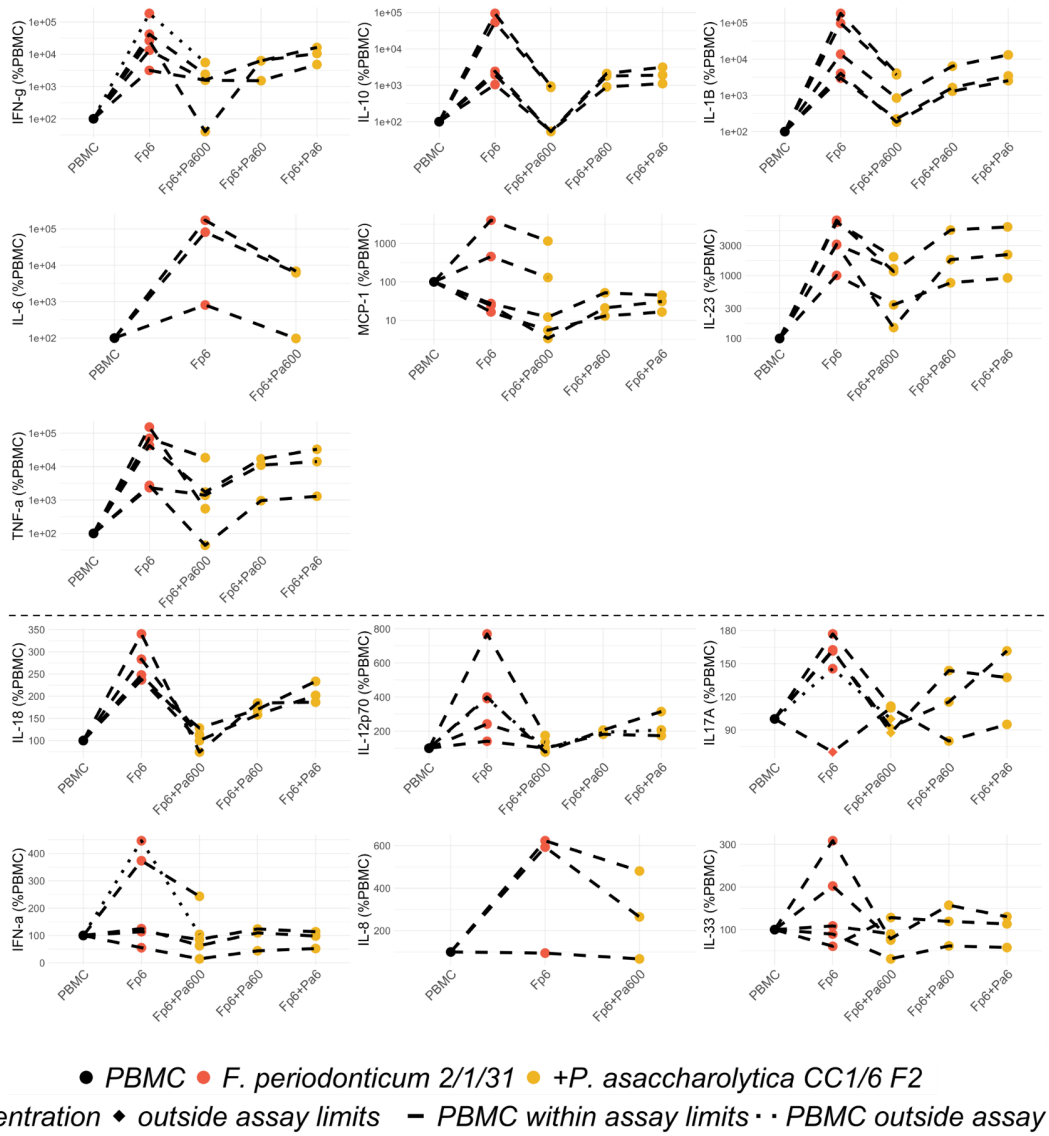
Supplementary Figure 3.3. Enrichment map of immune-related enriched gene sets unique to CMS4. Enrichment maps show which gene sets have shared genes between them (gene sets with connecting lines share genes). All adjusted p-values < 0.05.



Supplementary Figure 3.4. Secreted cytokine concentrations (pg/mL) after overnight incubation of PBMCs with different LPS concentrations (6 ng/mL, 60 ng/mL, 600 ng/mL) from different strains of A. *B. fragilis* B. *F. periodonticum*, and C. *P. asaccharolytica* compared to PBMC baseline (no treatment). Dashed lines connect data points from the same strains. Colors differentiate the strains of each species. Where values are outside of assay limits, we show the assay limit, and denote the point with '◆'.



Supplementary Figure 3.5. Changes in cytokine expression in peripheral blood mononuclear cells (PBMCs) following treatment with *F. periodonticum* alone or in combination with *B. fragilis*. *F. periodonticum 2/1/31* LPS used at a concentration of 6 ng/mL (Fp6, red). *B. fragilis 2/1/16* LPS used at a concentration of 600 ng/mL (Bf600, blue), 60 ng/mL (Bf60, blue), and 6 ng/mL (Bf6, blue). Values are shown as percentages of PBMC baseline secretion, which is set at 100%. Connecting lines indicate a single experimental run. Where concentrations are outside assay limits, we show the limit concentration and denote this with a '◆'. When the PBMC (black) as baseline was outside the assay limit, we used the limit concentration as baseline, and denote the entire replicate as dotted lines. Y-axes of plots above the dashed midline are in $\log_{10}$ scale; Y-axes of plots below it are in linear scale.



Supplementary Figure 3.6. Changes in cytokine expression in peripheral blood mononuclear cells (PBMCs) following treatment with *F. periodonticum* alone or in combination with *P. asaccharolytica*. *F. periodonticum 2/1/31* LPS used at a concentration of 6 ng/mL (Fp6, red)**.** *P. asaccharolytica 2/1/16* LPS used at a concentration of 600ng/mL (Bf600, yellow), 60 ng/mL (Bf60, yellow), and 6 ng/mL (Bf6, yellow). Values are shown as percentages of PBMC baseline secretion, which is set at 100%. Connecting lines indicate a single experimental run. Where concentrations are outside assay limits, we show the limit concentration and denote this with a '◆'. When the PBMC (black) as baseline was outside the assay limit, we used the limit concentration as baseline, and we denote the entire replicate as dotted lines. Y-axes of plots above the dashed midline are in $\log_{10}$ scale; Y- axes of plots below it are in linear scale.

3.6.3 Supplementary Tables

Supplementary Table 3.1. Tests of Differences between Cytokine Concentrations Released after PBMC Treatment using *F. periodonticum and B. fragilis* LPS							
Cytokine	**Reference Group**	**Group 2**	**N1**	**N2**	**p-val**	**Cohen's d**	**Magnitude**
IL-12p70	PBMC	Fp6	3	3	0.16	-0.72	moderate
	Fp6	Fp6 + Bf600	3	3	0.20	0.62	moderate
IL-10	PBMC	Fp6	4	4	0.057	-1.10	large
	Fp6	Fp6+Bf600	4	4	0.055	1.11	large
IFN-γ	PBMC	Fp6	4	4	0.090	-0.90	large
	Fp6	Fp6+Bf600	4	4	0.090	0.90	large
IL-1β	PBMC	Fp6	4	4	0.041	-1.26	large
	Fp6	Fp6+Bf600	4	4	0.041	1.26	large
IL-18	PBMC	Fp6	4	4	0.017	-1.77	large
	Fp6	Fp6+Bf600	4	4	0.050	1.163	large
IL-6	PBMC	Fp6	4	4	0.025	-1.52	large
	Fp6	Fp6+Bf600	4	4	0.023	1.57	large
Legend: Fp6=*F. periodonticum 2/1/31* (6 ng/mL) alone; Fp6+Bf600=*F. periodonticum 2/1/31* (6 ng/mL) + *B. fragilis 2/1/16* (600 ng/mL) co-incubation; N=number of repeats used in statistical calculations; p-val=paired t-test p-value; d=cohen's d; Magnitude=interpretation of cohen's d							

Supplementary Table 3.2. Test of Differences between Cytokine Concentrations Released after PBMC Treatment using *F. periodonticum* and *P. asaccharolytica* LPS							
Cytokine	**Reference Group**	**Group 2**	**N1**	**N2**	**p-val**	**Cohen's d**	**Magnitude**
IL-12p70	PBMC	Fp6	4	4	0.060	-1.07	large
	Fp6	Fp6+Pa600	4	4	0.080	0.95	large
IL-10	PBMC	Fp6	5	5	0.017	-1.41	large
	Fp6	Fp6+Pa600	5	5	0.016	1.45	large
IFN-γ	PBMC	Fp6	4	4	0.035	-1.34	large
	Fp6	Fp6+Pa600	4	4	0.060	1.07	large
IL-1β	PBMC	Fp6	5	5	0.010	-1.63	large
	Fp6	Fp6+Pa600	5	5	0.011	1.61	large
IL-18	PBMC	Fp6	5	5	0.020	-1.35	large
	Fp6	Fp6+Pa600	5	5	0.021	1.33	large
IL-6	PBMC	Fp6	3	3	0.017	-2.52	large
	Fp6	Fp6+Pa600	3	3	0.015	2.69	large
Legend: Fp6=*F. periodonticum 2/1/31* (6 ng/mL) alone; Fp6+Pa600=*F. periodonticum 2/1/31* (6 ng/mL) + *P. asaccharolytica CC1/6 F2*(600 ng/mL) co-incubation; N=number of repeats used in statistical calculations; p-val=paired t-test p-value; d=cohen's d; Magnitude=interpretation of cohen's d							

Chapter 4:

Host and Microbiome Contributions to Response to Radiotherapy in Rectal Cancer

Sulit, A.K.[1,2], Pearson, J.[3], Liu, W.[2], Silander, O.K.[1], Schmeier, S. [1,5], Wilson, K.[4], Heriot, A.[4], Frizelle, F.A.[2], Purcell, R.[2]

[1]School of Natural Sciences, Massey University, Auckland, New Zealand
[2]Department of Surgery, University of Otago, Christchurch, New Zealand
[3]Biostatistics and Computational Biology Unit, University of Otago, Christchurch, New Zealand
[4]Cancer Surgery, Peter MacCallum Cancer Centre, Melbourne, Australia
[5]Evotec SE, Hamburg, Germany

Article in preparation

Author contributions:

AKS performed the computational analyses and statistical analyses with input from **WL, OKS, SS**, **JP,** and **RP. AKS** wrote the manuscript with input from **OKS, RP, JP** and **KW. FAF** provided guidance about clinical aspects of the manuscript. **AH** is the project clinical lead at Peter MacCallum Cancer Centre, Melbourne, Australia. **AH** and **KW** were involved with sample collection. **KW** was involved in clinical data collection and RNA extraction from samples.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Arielle Kae L. Sulit
Name/title of Primary Supervisor:	Dr. Olin Silander
In which chapter is the manuscript /published work:	4

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

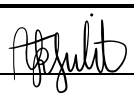
○ The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

◉ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	
Date:	16-Nov-2021
Primary Supervisor's Signature:	
Date:	16 Nov 2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

4.1 Background

The incidence of young onset (< 50 years) colorectal cancer (CRC) has been steadily increasing over the past two decades in the US, Oceania, and Canada. This increase is partly driven by an increased incidence in rectal cancer (Saad El Din et al., 2020). Rectal cancers comprise about 30% of CRCs, and its primary location usually calls for a more aggressive local treatment (Tamas et al., 2015). The treatment of rectal cancer is based predominantly on pre-operative staging, with upfront surgical resection reserved for early-stage tumors (T1-T2). Tumors that are locally advanced or have concerning features on pre-operative MRI require neo-adjuvant chemoradiotherapy (nCRT). Neo-adjuvant long-course chemoradiotherapy (LCCRT) involves the delivery of 50.4Gy of radiotherapy in combination with a radio sensitizing agent (5-FU or capecitabine), followed by an interval to surgery of 8–12 weeks (Terzi et al., 2020). Alternatively, short-course radiotherapy (25 Gy in 5 fractions) can be administered with a reduced interval to surgery. The response to neo-adjuvant therapy varies; up to 20% of patients achieve a pathological complete response (pCR), up to 60% demonstrate partial response, and another 20% display resistance to nCRT (Dayde et al., 2017). A validated method to accurately predict how an individual patient will respond to therapy is lacking.

A pCR is defined as the absence of residual viable tumor cells in the resected specimen. The reliance of pCR on pathological confirmation often fails to identify patients that may benefit from organ conservation and be spared major surgery. Habr-Gama (Habr-Gama et al., 2004) has pioneered the definition and adoption of a

clinical complete response (cCR) as a surrogate to a pCR. cCR is the absence of detectable tumors upon clinical examination and endoscopy (Feeney et al., 2019) - allowing for the avoidance of surgery. Patients that achieve a cCR are subsequently managed with a “watch and wait” approach, thereby avoiding the morbidity associated with major resectional surgery. Although 25% of these patients show tumor regrowth by two years; the majority are amenable to salvage procedures.

Extensive work has been carried out to identify biological markers for response to radiotherapy in rectal cancer (Dayde et al., 2017; Garland et al., 2014; Ryan et al., 2016). However, no reliable biomarkers have been validated for clinical use. A robust biomarker that selects patients likely to achieve a pCR to neo-adjuvant therapy, would allow for increased confidence in selecting patients amenable to non-operative management. Chemoradiotherapy is associated with significant localized and systemic side effects and has been demonstrated to negatively impact quality of life (Herman et al., 2013). Thus, it would also be useful to pre-select patients unlikely to benefit from conventional CRT, as they may be spared the associated side effects, and could be considered for novel treatment strategies.

Due to the largely sporadic nature of colon and rectal cancer, environmental factors are likely to play a critical role in the development of the disease, and recent international data points to the importance of the microbiome in its development and progression (Ahn et al., 2013; Gao et al., 2015; Marchesi et al., 2011). Recent reports have also demonstrated that systemic effects of the gut microbiome may contribute to treatment response in other cancer types (Gopalakrishnan et al., 2018;

Routy et al., 2018) and could be predictors of a favorable response to immunotherapy. In addition, gut microbiota have been shown to locally influence the treatment efficacy of irinotecan for colorectal cancer (Guthrie et al., 2017). However, although some studies on the protective effect of the microbiome on radiotherapy-induced toxicity have been carried out (Touchefeu et al., 2014; Vanhoecke et al., 2016), very little is known about whether or how the gut microbiome may regulate the tumor response to radiotherapy.

Here we present host and microbiome gene expression data from a unique cohort of pre-RT rectal cancer tumors and their matched normal mucosa samples. Having matched tumors and normal samples accounts for interpersonal variation in gene expression and allows for a robust identification of tumor and microbial genes and molecular pathways associated with response to RT in rectal cancer.

4.2 Methods

4.2.1 Patient Cohort and Characteristics

A total of 40 patients from two prospective cohorts of rectal cancer patients were used in this study: 20 patients from Christchurch Hospital, New Zealand, and 20 patients from the Peter MacCallum Cancer Centre, Melbourne, Australia. Biopsies of tumor tissue and adjacent, visually normal tissue (>10cm from tumor) were taken at colonoscopy, prior to treatment. Patient data, including staging, recurrence, metastases, treatment and histology was collected.

Patients who had received prior chemotherapy or radiation therapy for treatment of their rectal tumor were excluded from the study. This study was undertaken with ethical approval from the Health and Disability Ethics Committee of New Zealand (ethics approval number: ***18/STH/40***) and the Human Research Ethics Committees of Australia (ethics approval number: ***HREC 14/85***), and patients provided written, informed consent.

4.2.2 RNA Extraction

Tumor and normal tissue biopsies were taken at colonoscopy and immediately frozen in liquid nitrogen and stored at -80°C. RNA extraction was carried out as detailed previously (Purcell, Visnovska, et al., 2017). Briefly, RNA was extracted from approximately 20 mg of tissue using RNEasy Plus Mini Kit (Qiagen, Hilden, Germany), including DNAse treatment, following tissue disruption using a Retsch Mixer Mill. Purified RNA was quantified using a NanoDrop 2000c spectrophotometer (Thermo Scientific, Asheville, NC, USA), and stored at -80°C.

4.2.3 RNA Sequencing

RNA-sequencing was performed to produce 150bp paired end reads, as previously described (Purcell, Visnovska, et al., 2017). Ribo-Zero™ Magnetic Kit (Human & Bacteria) was used for ribosomal RNA depletion, and libraries were prepared using NEBNext® Ultra™ RNA Library Prep Kit (New England BioLabs Inc.®). Approximately

50 million reads (15Gb raw data) were produced per sample on an Illumina NovaSeq 6000 instrument. Raw sequencing reads were deposited at NCBI SRA under BioProject ID PRJNA815861.

4.2.4 RNA Sequencing Data Processing

Raw sequencing data was parsed through the Metafunc pipeline (A. K. L. Sulit et al., 2020), which performs read pre-processing, host gene mapping, and microbiome species identification. Read pre-processing was performed using fastp (S. Chen et al., 2018) with default parameters. Cleaned reads were then mapped to the human host genome (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_33/GRCh38.primary_assembly.genome.fa.gz) using STAR (Dobin et al., 2013). Host genes were then quantified using featureCounts (Liao et al., 2014) from the subread package and organized into per-sample gene expression matrices for downstream analysis using custom scripts within the MetaFunc pipeline.

The pipeline redirects reads that did not map to the human genome to Kaiju (Menzel et al., 2016) for microbial species identification. We used Kaiju's `nr_euk` option built from NCBI BLAST *nr* database obtained on January 27, 2020. The final output is a summary table of taxonomy read counts per sample.

4.2.5 Differential Human Gene Expression Analysis

We generated a gene expression per sample table using the MetaFunc pipeline and used DESeq2 to perform differential gene expression analysis, with this table as input. In order to identify differentially expressed genes in tumor samples compared to normal samples that were specific to complete responders compared to other responders, we analyzed the dataset as having group-specific condition effects as outlined in DESeq2's vignette (*http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#group-specific-condition-effects-individuals-nested-within-groups*), where groups refer to our sample response type, and condition refers to tumor or normal tissue type.

4.2.6 Gene Set Enrichment Analysis

From the results of this DESeq2 comparison, we generated a pre-ranked list of all resulting genes based on p-values and $\log_2$ fold-change, and these genes were used as input to GSEA analysis using clusterProfiler (G. Yu et al., 2012) with the C5 Ontology Gene Sets collection (version 7) from the molecular signatures database (MSigDB) (Liberzon et al., 2011; Subramanian et al., 2005). Specifically, we ranked the genes using the formula:

rank = - log10(p-value) * sign(log2FoldChange),

with sign being directionality (+ or -) of the $\log_2$ fold change value, and actual value was not used. This ranking places the genes with lowest p-values and positive $\log_2$ fold change at the top of the list, and the genes with lowest p-values and negative $\log_2$ fold change at the bottom of the list. Genes at the top of the list contribute to

gene sets with positive enrichment scores and genes at the bottom contribute to gene sets with negative enrichment scores.

For more details on host gene analyses, we provide R scripts and result files in https://gitlab.com/alsulit08/uoc_response_rectalca/-/tree/main/Human commit affc1e24.

4.2.7 Differential Metatranscriptome Analysis

For the microbiome dataset, we gathered raw counts of microbe taxonomies into a Phyloseq object (McMurdie & Holmes, 2013), with metadata information on their response, and tumor or normal status. We then used the same model for group-specific condition effects as specified in DESeq2, to obtain differentially abundant microbes in tumor samples compared to matched normal samples, specific to complete responders compared to other responders. We performed pre-filtering of our species before the analysis, only including those within the Bacterial Kingdom, and those with at least 10 reads in 20% of our samples (**Supplementary Table 4.1**).

We initially obtained 26 differentially abundant bacteria from complete responders. We further narrowed down this list to 10 species through manual curation (see ***Supplementary Methods*** section).

4.2.8 Correlation between Differentially Expressed Genes and Bacteria in Rectal Cancer

We included the 87 identified DEGs and the 10 DA bacterial species we identified. Using the DESEq2 *rlog* transformed values for gene expression and microbial abundance, we performed Spearman correlation analysis between each gene and species, correcting the final p-values using Benjamini-Hochberg (BH) adjustment.

For more details on microbiome analyses, we provide R scripts and result files in https://gitlab.com/alsulit08/uoc_response_rectalca/-/tree/main/Microbiome commit affc1e24.

4.2.9 Microbial Diversity

The sample set was rarefied to 90% of the smallest sample size in the dataset. We obtained alpha diversities (diversity within a community) per sample using Phyloseq's (McMurdie & Holmes, 2013) *estimate_richness* function for Observed (richness) and Shannon (richness and evenness) measures, and compared differences of the alpha diversity measures by Wilcoxon test. We visualized microbiome community differences between groups using Phyloseq's *ordinate* function on the rarefied dataset, using NMDS, and Bray-Curtis distances. We compared alpha diversities (Observed and Shannon), and performed NMDS ordination among the following iterations of groupings for our dataset: 1) Tumor vs Normal, 2) Complete Responders vs Non-responders, 3) Complete vs Non-responders in tumors and normal separately, and 4) Tumors vs Normals in Complete responders and other responders separately.

For more details on microbial diversity, we provide R scripts and result files in https://gitlab.com/alsulit08/uoc_response_rectalca/-/tree/main/Microbiome/Diversity commit affc1e24.

4.3 Results

4.3.1 Rectal Cancer Cohort

The cohort comprised 39 patients with diagnosed rectal cancer who were subsequently treated with chemo-radiotherapy (**Table 4.1**). The majority (n = 36) were treated with long course CRT (LCCRT), with either capecitabine, FOLFIRI (folinic acid, fluorouracil, and irinotecan), or 5FU (fluorouracil). Two patients did not complete LCCRT due to the development of grade 3 toxicity. One patient received short-course RT, while the remaining two patients received sandwich CRT (FOLFOX (folinic acid, fluorouracil, and oxaliplatin)). There were 12 females and 27 males, who ranged in age from 29–86 years (mean age, 62 years). Response to LCCRT was assessed histologically from surgical resection specimens, and reported using Dworak grading (Christchurch cohort) or the American Joint Committee on Cancer (AJCC) grading (Melbourne cohort). Response groups were designated as complete responders (Dworak 4/AJCC 0), near-complete responders (Dworak 3/AJCC 1), incomplete responders (Dworak 2/AJCC 2), and non-responders (Dworak 1/AJCC 3). There were five patients with complete response to LCCRT, five patients with near-complete response, 18 patients with incomplete response and eight patients

who did not respond to LCCRT. In addition, three patients developed progressive disease, or died of disease, during the course of therapy, and these patients were also designated non-responders. Six patients died of the disease during the follow-up period of 24 months.

Table 4.1 Characteristics of the Rectal Cancer Patient Cohort

	All (n=39)	Complete Response (n=5)	Near-Complete Response (n=5)	Incomplete Response (n=18)	No Response (n=11)
Age, mean ± SD	62.05 ± 14.55	66.2 ± 9.31	61.20 ± 4.82	62.44 ± 14.83	69.91 ± 19.29
Sex:					
Male	26 (66.67%)	4 (80%)	4 (80%)	13 (72.22%)	5 (45.45%)
Female	13 (22.22%)	1 (20%)	1 (20%)	5 (27.78%)	6 (54.55%)
Died of disease:	5 (12.82%)	0 (0.00%)	0 (0.00%)	3 (16.67%)	2 (18.18%)
Treatment:					
LCCRT (no other data)	4 (10.26%)	0 (0.00%)	0 (0.00%)	3 (16.67%)	1 (9.09%)
LCCRT (capecitabine)	26 (66.67%)	5 (100%)	3 (60%)	9 (50%)	9 (81.82%)
LCCRT (5FU)	3 (7.69%)	0 (0.00%)	0 (0.00%)	3 (16.67%)	0 (0.00%)
LCCRT (FOLFIRI)	2 (5.13%)	0 (0.00%)	0 (0.00%)	2 (11.11%)	0 (0.00%)
Sandwich CRT	2 (5.13%)	0 (0.00%)	2 (40%)	0 (0.00%)	0 (0.00%)
SCRT	1 (2.56%)	0 (0.00%)	0 (0.00%)	1 (5.56%)	0 (0.00%)
Palliative	1 (2.56%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (9.09%)

LCCRT: long course chemoradiotherapy **FOLFIRI:** folinic acid, fluorouracil and irinotecan
SCRT: short course radiotherapy
5FU: fluorouracil

In our analysis, we compared complete responders against all other responders grouped together, and conversely, we also analyzed non-responders compared to all other response groups as one condition.

4.3.2 Differential Gene Expression Between Radiotherapy Response Groups

We analyzed differentially expressed genes (DEGs) between matched pairs of tumor and normal tissues, to account for interpersonal variation in gene expression, and then identified which of these DEGs were associated with a response group, i.e., if DEGs were specific to complete responders compared to other responders, and *vice versa*. Accounting for gene expression differences unique to an individual greatly reduces noise and provides a more robust identification of gene expression that is truly different between response groups. We found 87 genes that were differentially expressed between tumor and normal samples (adjusted p-value < 0.1), specific to complete responders. Interestingly, the majority of these genes were associated with immunoglobulin chains with 75 out of 87 genes having prefixes of IGH-, IGL-, or IGK- (**Figure 4.1**).

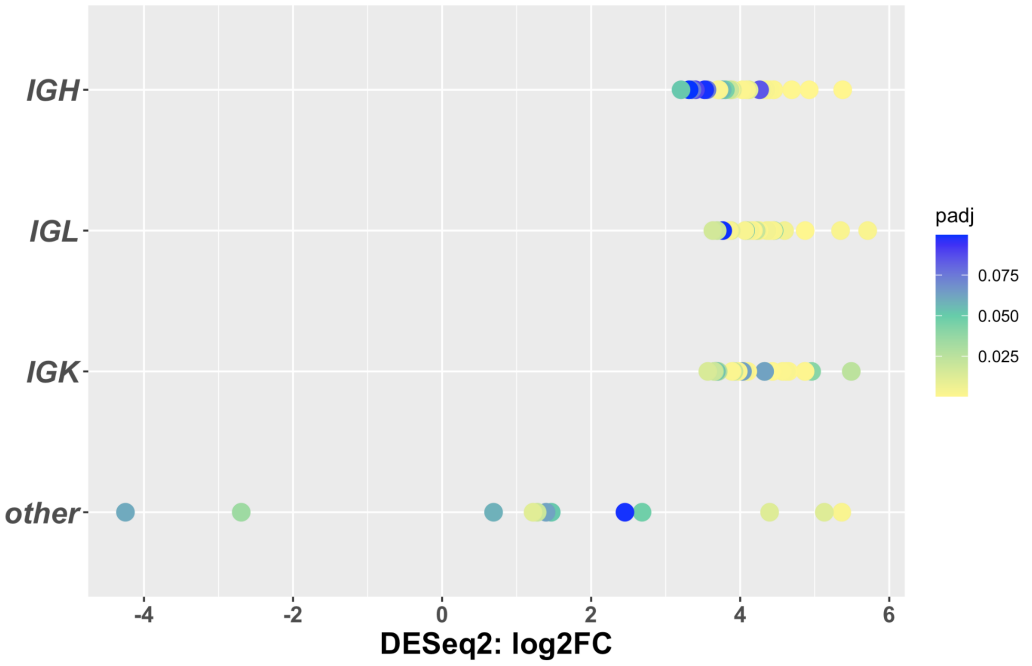


Figure 4.1. Differentially expressed genes in tumors vs matched normal samples specific to complete responders compared to other responders. The DEGs (individual points) are classified into immunoglobulin- related genes (*IGH-, IGL-, and IGK-* prefixes), and other non-Ig genes, as indicated by the y-axis. The x-axis shows their $\log_2$ fold changes, and point colors show their adjusted p-values.

We plotted tumor vs normal *rlog* transformed counts per sample of representatives from these genes and found that, for most of these genes, complete responders cluster at high tumor-low normal values (upper left quadrant) indicating that in the complete responder group, these genes are more highly expressed in tumors compared to normal tissues (**Supplementary Figure 4.1**). We show *IGKC* as representative of these genes in **Figure 4.2**.

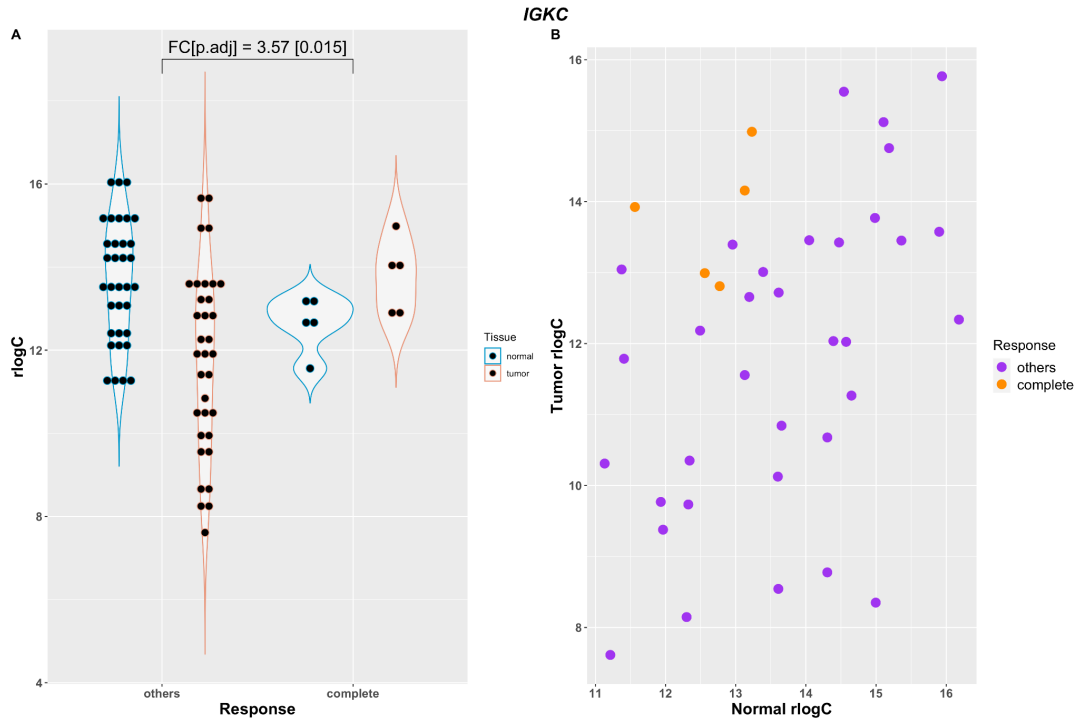


Figure 4.2. rlog transformed counts of Tumor/Normal of complete responders compared to other responders of a representative DEG, *IGKC*. A. Distribution of the values in Tumor (red) vs Normal (blue) for each response category (*FC*=$\log_2$ Fold Change, *p.adj*=adjusted p-value). **B.** Scatter plot of the tumor values and their corresponding normal values per sample, colored by response (complete: orange, others: purple).

When differential gene expression analysis (DGEA) was carried out comparing non-responders to all other responders, we found no differentially expressed genes between the two groups.

We found that the majority of the significant differentially expressed genes were up-regulated in tumors compared to normal samples. The top 10 DEGs (all immunoglobulin-related) in addition to 12 non-immunoglobulin-related DEGs robustly distinguished complete responders from other responders, as illustrated in **Figure 4.3**.

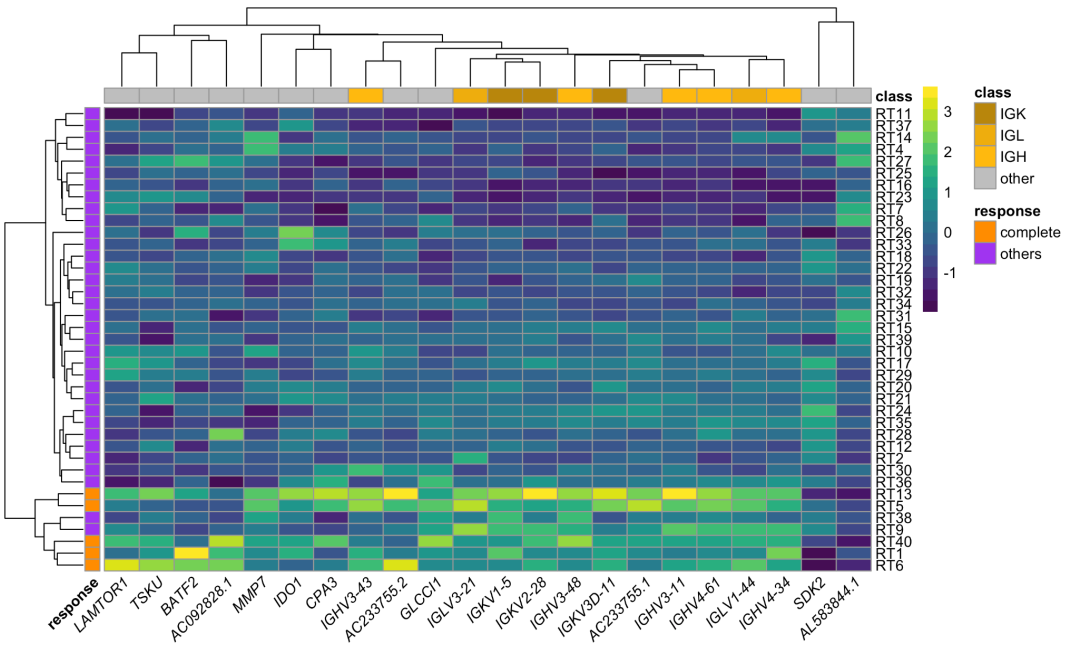


Figure 4.3. The top 10 DEGs (all Ig-related genes), and the other 12 non-Ig DEGs separate complete responders from others. Genes are ranked by adjusted p-values. We obtained the tumor vs normal (T/N) ratios of these genes using rlog transformed counts per patient sample, and scaled these values as z-scores per gene (heatmap colors). Heatmap and clusters were generated by the *pheatmap* (Kolde, 2019) function and library in R.

4.3.3 Host Gene Set Enrichment Analysis

We also performed Gene Set Enrichment Analysis (GSEA) of the genes, ranked according to DESeq2 p-value and $\log_2$ fold-change (-/+) as described in the methods section. Investigation of the top 50 (lowest p-values, highest normalized enrichment scores) enriched gene sets showed that immune responses constituted the majority of these 50 gene sets (**Figure 4.4**).

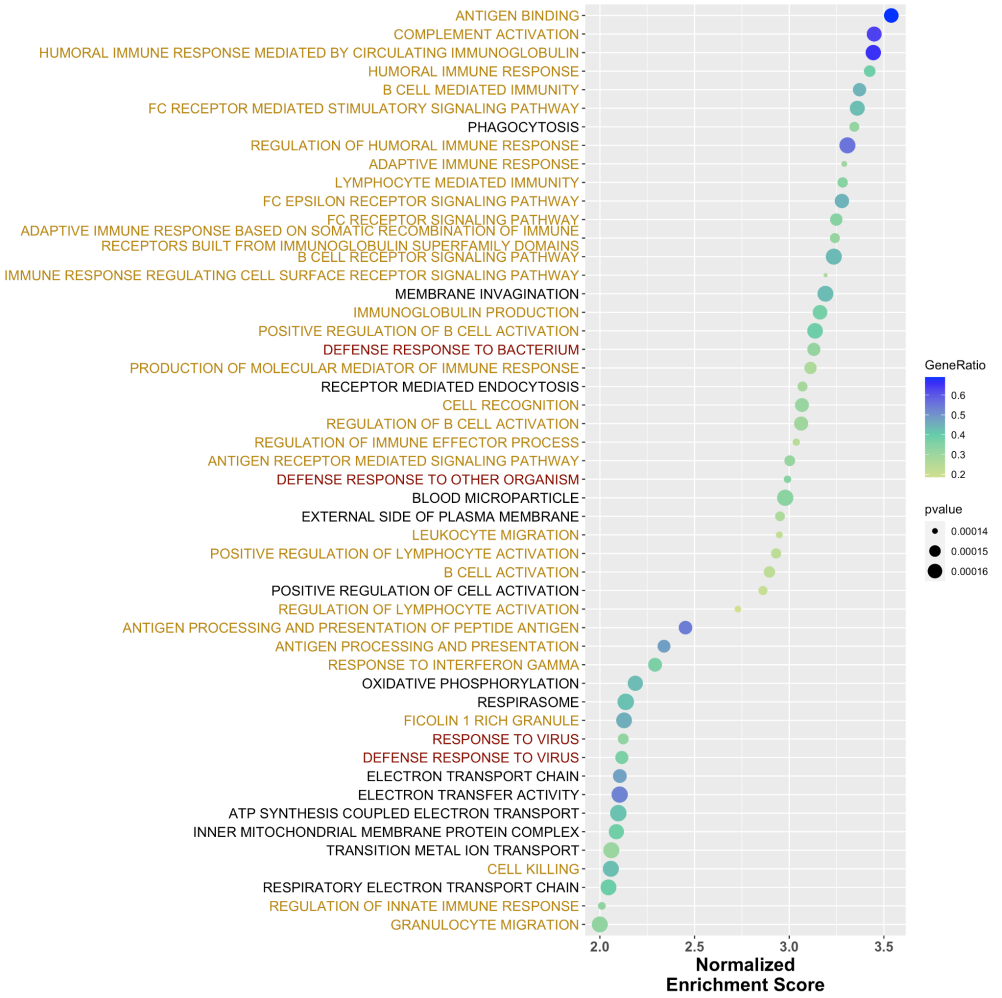


Figure 4.4. Top 50 positively enriched gene sets in Tumors vs Normal Samples of complete Responders. All have adjusted p-values of 0.0093 and dot sizes represent their respective individual p-values. Dot colors are gene ratios which is the ratio of core enrichment genes to the gene set size. Plot is ranked by Normalized Enrichment Score (NES). Labels colored yellow are immune-related gene sets. Labels colored red are gene sets that refer to a response to microbial input.

Among the gene sets with the highest enrichment scores were *Complement Activation* and *B-Cell Mediated Immunity*, which was not surprising as the majority of the 87 differentially expressed genes in complete responders are related to immunoglobulins. The enrichment map of these top gene sets (**Supplementary Figure 4.2**) also showed the functional clusters with shared genes, identifying a large

cluster of gene sets involved in immune responses. Notable among these gene sets was “defense response to bacterium”, indicating a potential role of the microbiome.

4.3.4 Diversity Analysis of Microbiome of Rectal Tumors

Overall, we found no significant differences in observed or Shannon measures of alpha diversity and no distinct group separations by NMDS (**Figure 4.5**).

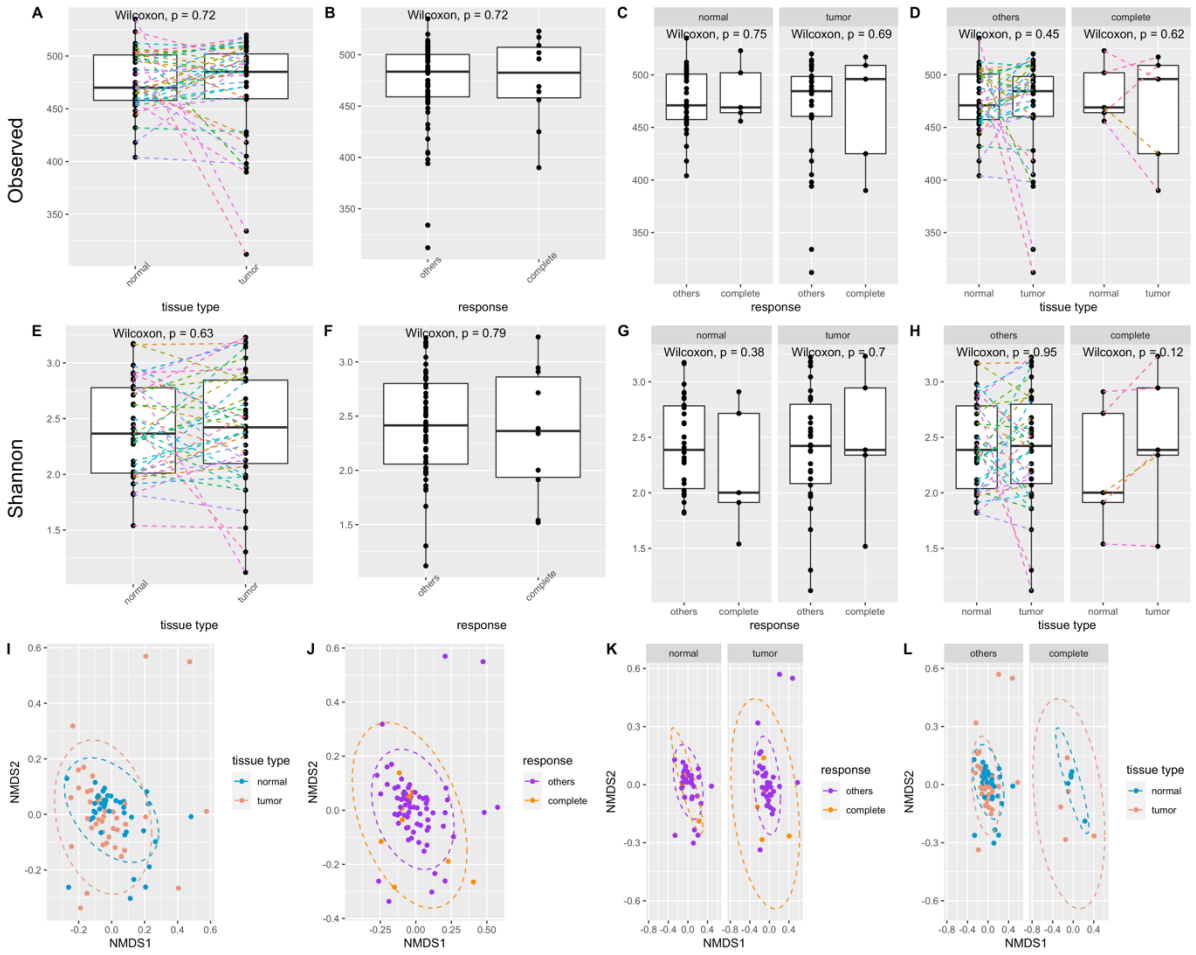


Figure 4.5. Diversity analyses on different grouping combinations of Tumor, Normal, and Response Groups in Rectal Cancer. A-D: Observed Alpha Diversity. **A.** between normal and tumor, **B.** between response groups, **C.** between response groups within tissue type, **D.** between tissue type within response groups. **E-H**. Shannon Alpha Diversity. **E.** between normal and tumor, **F.** between response groups, **G.** between response groups within tissue type, **H.** between tissue type within response groups. **I-L**. NMDS ordination of Bray-Curtis distances between samples, grouped by: **I.** between normal and tumors, **J.** between response groups, **K.** between response groups within tissue types, **L.** between tissue type within response groups.

4.3.5 Differential Abundance Analysis of Bacterial Species between Radiotherapy Response Groups

A similar approach was used to identify bacterial transcripts that were differentially abundant in tumor versus matched normal samples, and specific to complete responders. Analysis of differences in the tumor microbiome between different response groups identified 10 bacteria that are differentially abundant (adjusted p-value < 0.1) in tumor tissue compared to matched normal tissue in complete responders (**Figure 4.6**).

Plotting Tumor vs Normal *rlog* transformed counts per sample of these microbes showed that there was a separation between the complete responders compared to other responders although not as pronounced as that seen in human DEGs (**Supplementary Figure 4.3**). We found bacteria including *Ruminococcaceae bacterium*, *Hungatella hathewayi*, *Bacteroides thetaiotaomicron, and Clostridium* species, which were previously reported to have a role in CRC or cancer.

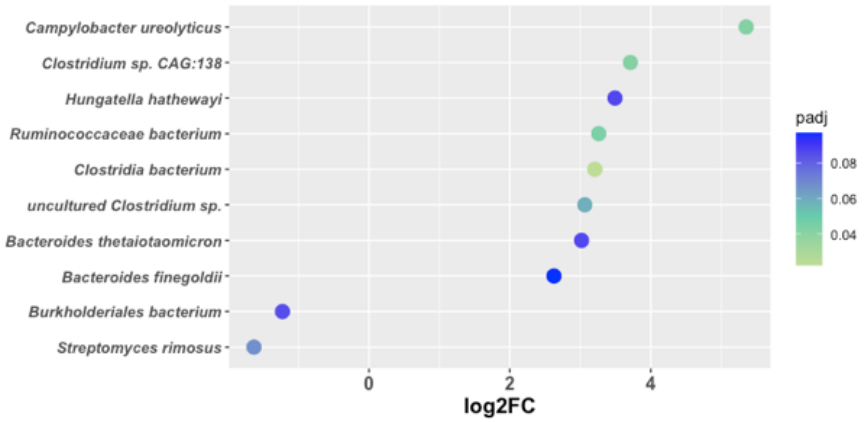


Figure 4.6. Differentially abundant bacteria in tumor samples compared to matched normal tissue, specific to complete responders. Plot of the 10 DA bacteria showing their $\log_2$ fold changes (x-axis) and adjusted p-values by color.

4.3.6 Correlation between Host Gene Expression and Microbial Abundances

We found significant correlations between differentially expressed genes and bacterial abundances (**Figure 4.7A)**. Among DEGs that positively correlated with the microbial abundances, the *BATF2* gene was notable. We found positive correlations (*Spearman coefficient: 0.355 – 0.549; BH: 0.0002 - 0.077*) of this gene with several of our DA bacterial species, previously linked to colorectal cancer, including *Ruminococaccae bacterium*, and *Bacteroides thetaiotaomicron* (**Figure 4.7B**).

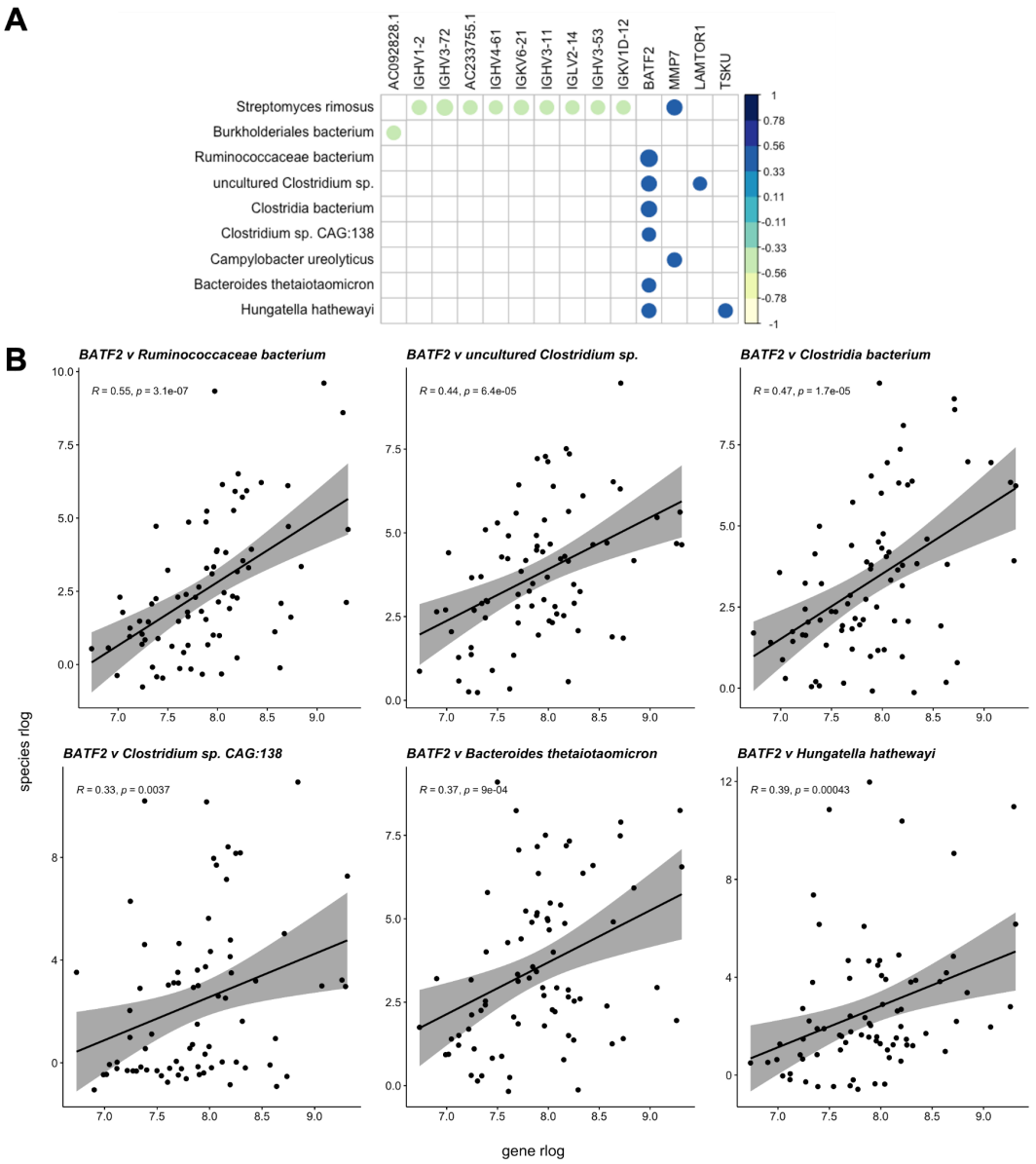


Figure 4.7. Correlations between differentially expressed genes and differentially abundant microbes. A. Correlation plot showing bacteria – gene correlations with at least 1 significant (*BH < 0.1*) correlation. Spearman correlation coefficients are represented by the colorbar; blank spaces represent correlations that are not significant. **B.** We focused on *BATF2* and its positive correlations with our DA bacteria. We removed influential points (*species rlog > 12*) for the Spearman correlation calculation and scatter plots. Only *Clostridium sp. CAG:138* had any influential points.
R=Spearman coefficient; p=p-value calculated.

4.4 Discussion

Neoadjuvant chemoradiotherapy aims to downstage rectal tumors and reduce rates of local recurrence (Dayde et al., 2017; Feeney et al., 2019). A proportion of patients will have an excellent response to this therapy with complete tumor sterilization (Habr-Gama et al., 2004). There has been a reliance on post-surgical examination of post-surgical specimens to confirm this response. However, surgical resection may offer no survival benefit to patients with a pCR, while exposing them to potential morbidity and mortality. A means of predicting complete response to neo-adjuvant therapy is therefore required to allow accurate assessment of those that may benefit from organ conservation.

Previous studies have attempted to identify markers of complete response to chemoradiotherapy, and while clinicopathological and radiological features have been identified, they are limited in sensitivity and specificity (Dayde et al., 2017). Intratumoral heterogeneity contributes to lack of correlation between molecular biomarker studies, and as such, no biomarker is currently in clinical use (Dayde et al., 2017; Ryan et al., 2016). In order to identify a universal biomarker of response for rectal tumors a mechanistic link to tumor biology is essential.

Ionizing radiation results in damaged cell molecules being presented to the immune system as an antigen (Lhuillier et al., 2019; Regal et al., 2016). Antigen presentation is important in the activation of T-cell responses that result in anti-tumor immunity, activating a response akin to viral infection responses. Ionizing radiation causes the

production of reactive oxygen species that causes damage to the cells' DNA. Damaged double stranded (ds) DNA accumulation in the cell's cytosol, can trigger an inflammatory response. Furthermore, mutated DNA can be translated into peptides that are presented on MHC-1 molecules and elicit a cytotoxic immune response (Lhuillier et al., 2019).

Complete responders in our cohort have higher expression of genes responsible for complement activation and B-cell related functions in their tumors compared to normal cells. This observation supports published data about the mechanism of radiotherapy. Immunoglobulins may recognize radiotherapy induced neo-antigens resulting in complement activation. A study by Surace and colleagues (Surace et al., 2015) determined that radiotherapy-induced necrosis can result in IgM-mediated binding to necrotic cells, with subsequent complement activation. This results in anaphylatoxin production, activating dendritic cells that mediate CD8+ T-cell responses. This systemic immune response is important to attain abscopal effects, distant from radiation sites.

We also observed enriched gene sets related to viral response in complete responders, which may be explained by radiotherapy induced viral mimicry, where the proteome induced by the ionizing radiation gets presented as antigens to MHC-I molecules, similar to viral antigens (Lhuillier et al., 2019). An important player in antiviral immune responses are type 1 interferons (IFN-I) that recruit dendritic cells specialized in antigen presentation to $CD8^+$ T-cells (Lhuillier et al., 2019). Type-1 interferons play a part in the stimulator of interferon genes (STING) pathway that

activates adaptive and immune system responses (Lhuillier et al., 2019; Y. Wang et al., 2018).

Type-1 interferons also activate BATF2 which has recently been shown to be depleted in CRC tumors compared to normal samples, and correlates with poor prognosis (Liu et al., 2015). *BATF2* was over-expressed in complete responders in our cohort. As *BATF2* is thought to be induced by IFN-I, it was hypothesized to play a role in viral infections (Guler et al., 2015) and contribute to radiotherapy-induced responses that mimic viral infection. BATF2 was found to have an anti-tumor effect through promoting IL-12 p40 expression in tumor-associated macrophages (TAMs), which induces $CD8^+$ T-cells (Kanemaru et al., 2017).

Anti-tumor immunity as a consequence of radiotherapy indicates a possible interaction with immunotherapy-related processes (Y. Wang et al., 2018). This could be further aided by the presence of gut microbes that have been shown to be beneficial in immunotherapy applications, examples of which include *Bacteroides thetaiotaomicron* (Vétizou et al., 2015) and the *Ruminococcaceae* family (Gopalakrishnan et al., 2018).

Our finding of an enriched gene set for 'defense response to bacterium' also implies a response to an invading organism. Consistent with previous studies, we found no differences in microbiome diversity metrics between response groups (Shi et al., 2020). However, we identified several bacterial species that were differentially abundant in tumors of complete responders, several of which had previously been

implicated in CRC carcinogenesis and prognosis. *Hungatella hathewayi* has been identified as among the most significant bacterial markers that are differentially abundant in CRC (Wirbel et al., 2019), and has been reported to drive methylation of tumor suppressor genes (X. Xia et al., 2020). *Ruminococcaceae* have been linked to the expression of T-cell-recruiting chemokines (Cremonesi et al., 2018), indicating a role in tumor-killing immune activation and a low abundance has been associated with CRC (Burns et al., 2015; L. C.-H. Yu et al., 2018). Similarly, *Clostridium* species had been associated with a protective effect against CRC, due to their ability to synthesize short-chain fatty acids, which inhibit inflammation and carcinogenesis (Zou et al., 2018). *Bacteroides thetaiotaomicron*, meanwhile, was shown to augment efficacy of CTLA-4 blockade immunotherapy due to their ability to activate T-cell responses (Gao et al., 2015).

Recent studies have highlighted the effect of the gut microbiome on side effects of radiotherapy (Shi et al., 2020; Touchefeu et al., 2014), and, conversely, the effect of radiotherapy on the microbiome (Fan et al., 2021). On review of the literature, one study has previously investigated differences in the microbiome in pre-radiotherapy samples and their association with response to radiotherapy (Vétizou et al., 2015). Similar to our findings, this group reported no difference in alpha diversity between responders and non-responders, but did report a difference in beta-diversity. In contrast to our study of pre-treatment tissue samples, this study used amplicon sequencing of patient stool samples, which is an effective way to measure large taxonomic differences in microbiome samples, but performs poorly at discriminating between species or strains. The importance of the gut microbiome in response to

immunotherapy has garnered much attention in recent years, with landmark publications leading to clinical trials of fecal microbiome transplants to improve efficacy of immunotherapy (Fan et al., 2021; Jang et al., 2020; Shi et al., 2020). Given the known immune mechanisms at play in radiotherapy response, e.g. CD8+ T-cell activation, an overlap between the established systemic effect of the microbiome on immunotherapy efficacy and that of radiotherapy is likely. *Ruminococcaceae* and *B. thetaiotaomicron*, both enriched in complete responders in our cohort, had been found to augment immunotherapy effects and this mechanism may also contribute to radiotherapy-induced immune clearance of cancer cells. Taken together with the correlations seen between differentially abundant bacteria and *BATF2* expression and the increased abundance of commensal bacteria, the activation of immune pathways in the tumor microenvironment of complete responders indicates a potential role of the tumor microenvironment in radiotherapy response in rectal cancer.

We summarize a hypothetical mechanism of complete response in radiotherapy in **Figure 4.8.**

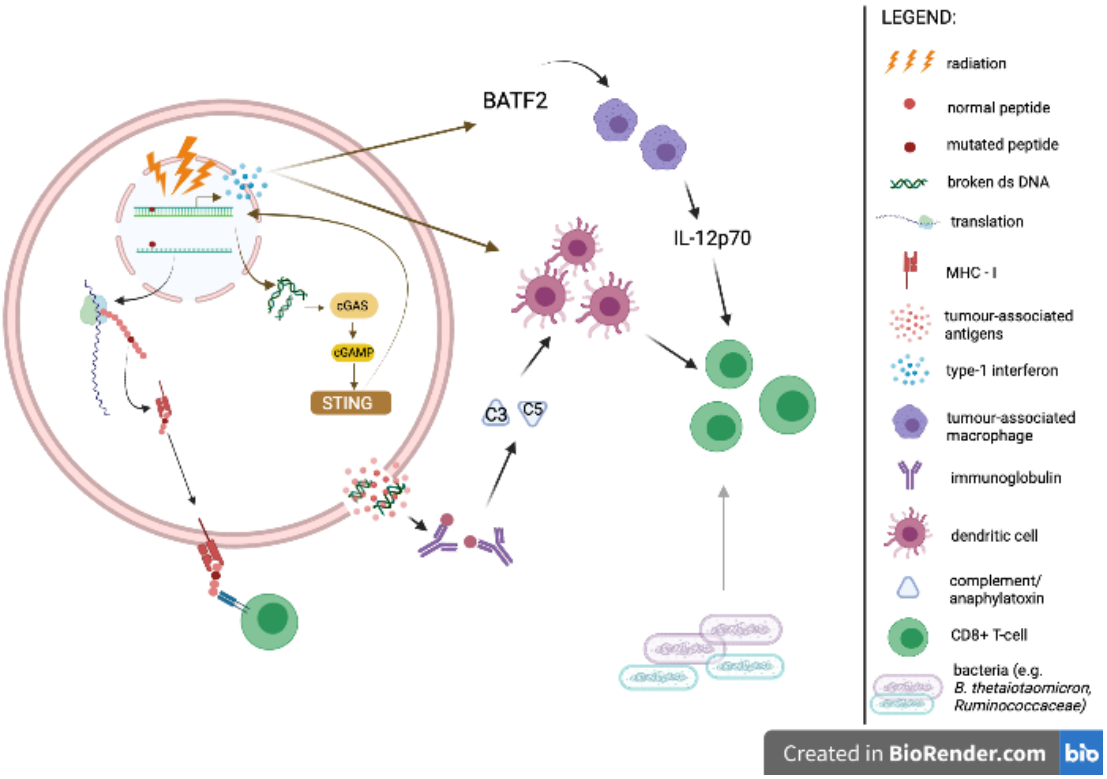


Figure 4.8. Hypothetical mechanism of complete response in radiotherapy. Radiotherapy results in DNA mutations that may be translated into mutated peptides that get presented by MHC molecules to recruit CD8+ T-cells. CD8+ T-cells include cytotoxic cells important in killing cancerous cells. Radiotherapy also results in double stranded DNA breaks that activate the STING pathway that induces the production of Type-1 interferons. Type-1 interferons, in turn, leads to increased *BATF2* production that activates tumor-associated macrophages, which recruits CD8+ T-cells via the IL-12 cytokine. Type-1 interferons can also activate dendritic cells. Dendritic cells may also be activated through anaphylatoxins C3 and C5 of the complement system when they are in turn activated by immunoglobulins that detect broken double stranded DNA. Dendritic cells in turn recruit CD8+ T-cells. Finally, bacteria such as *B. thetaiotaomicron* and *Ruminococcaceae* could also influence the recruitment of CD8+ T-cells. Image created with BioRender.com. Parts of the figure were adapted from Lhuillier et al., 2019 which is licensed for reuse under https://creativecommons.org/licenses/by/4.0/.

4.5 Conclusion

Our results suggest that an increase in the expression of genes contributing to immune activation in tumors compared to normal samples contributes to radiosensitivity. We hypothesize that this primes a microenvironment that may activate anti-tumor responses when the effects of radiotherapy take place. These

anti-tumor responses may occur through complement activation by immunoglobulins, when irradiated cells produce mutated proteins as antigens in a manner similar to viral infection, as evidenced by the increase in viral response gene sets in our tumors compared to normal samples in complete responders. Expression of the *BATF2* gene, activated by interferons, may also contribute to differences in radiotherapy response. Its expression is also mildly correlated with microbes that are associated with CRC. Bacteria enriched in complete responders have been associated with prognosis in colorectal cancer and improved efficacy of immunotherapy. These data provide future targets for biomarker validation and provide direction to investigate the mechanisms of radiotherapy response in rectal cancer.

4.6 Supplementary Material

4.6.1 Supplementary Methods

4.6.1.1 Manual Curation of Differentially Abundant Bacteria

Analysis of differences in the tumor microbiome between different response groups identified 26 microbes to be significantly associated (adjusted p-value <0.1) with complete responders compared to all other patients (**Supplementary Figure 4.4A**). Initial Spearman correlations between our microbial abundances and the 87 DEGs show several bacteria amongst these 26 identified that correlated highly (reaching *rho* > 0.8) with our DEGs (**Supplementary Figure 4.4B**). Several of these bacteria however were not expected to be found in the human gut, so we decided to look more closely at the protein matches of these bacteria. For several of these, we have

only a number of protein (accession number) matches (**Supplementary Table 4.2**) and closer inspection of these matches indicate that these proteins contain immunoglobulin domains and are highly similar to human immunoglobulin domains, leading us to believe that these could have been contaminants from the host RNA that were not identified in the human genome reference. As Ig genes had a strong signature in our T/N comparison between different radiotherapy responders, it stands to reason that if these were human Ig genes spuriously mapping to microbes, they would correlate very highly to the Ig genes identified as DEGs in our dataset. We therefore set a cutoff of protein matches >=100, and further investigated the remaining bacteria to remove those with possible Ig-domains from human contamination, leaving us with the 10 bacteria discussed in the paper.

4.6.1.2 Differentially Abundant Gene Ontologies

We used prot2go.py from the metafunc custom scripts to obtain Gene Ontologies from the 557 bacterial species that passed the 10 reads in 20% of samples cutoff we used for metatranscriptome analysis. This obtained a Gene ontology to sample table for Gene Ontologies annotating the proteins of the 557 bacterial species. We used the proportional read counts attributed to the gene ontologies for differential abundance analysis (see https://metafunc.readthedocs.io/en/latest/results.html directory-source-go). From this data, we again subset for Gene Ontologies with 10 reads in 20% of samples. Those that passed filters were processed through DESeq2 using the same design as the DGEA and bacterial DA analyses. DESeq2 rounds off the proportional read counts to integers.

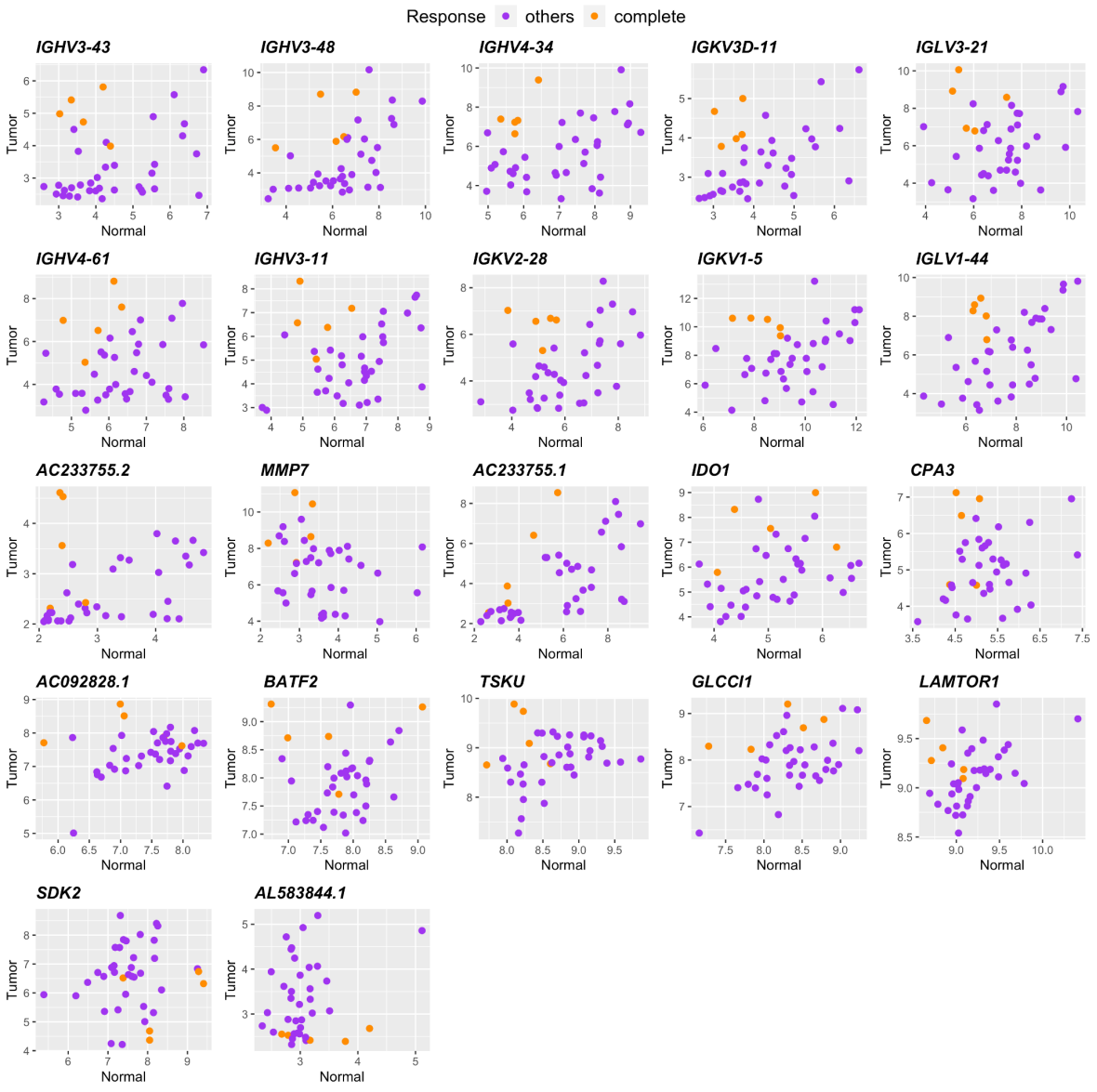
With this method, we only found one differentially abundant GO (*GO:0004674, protein serine/threonine kinase activity*) that had a $\log_2$ fold change of -2.856, and adjusted p-value of 0.096, which we deemed insignificant for the purposes of the paper.

4.6.1.3 Comparison of Gene Ontologies of 10 DA Bacteria between Complete and Other Responders

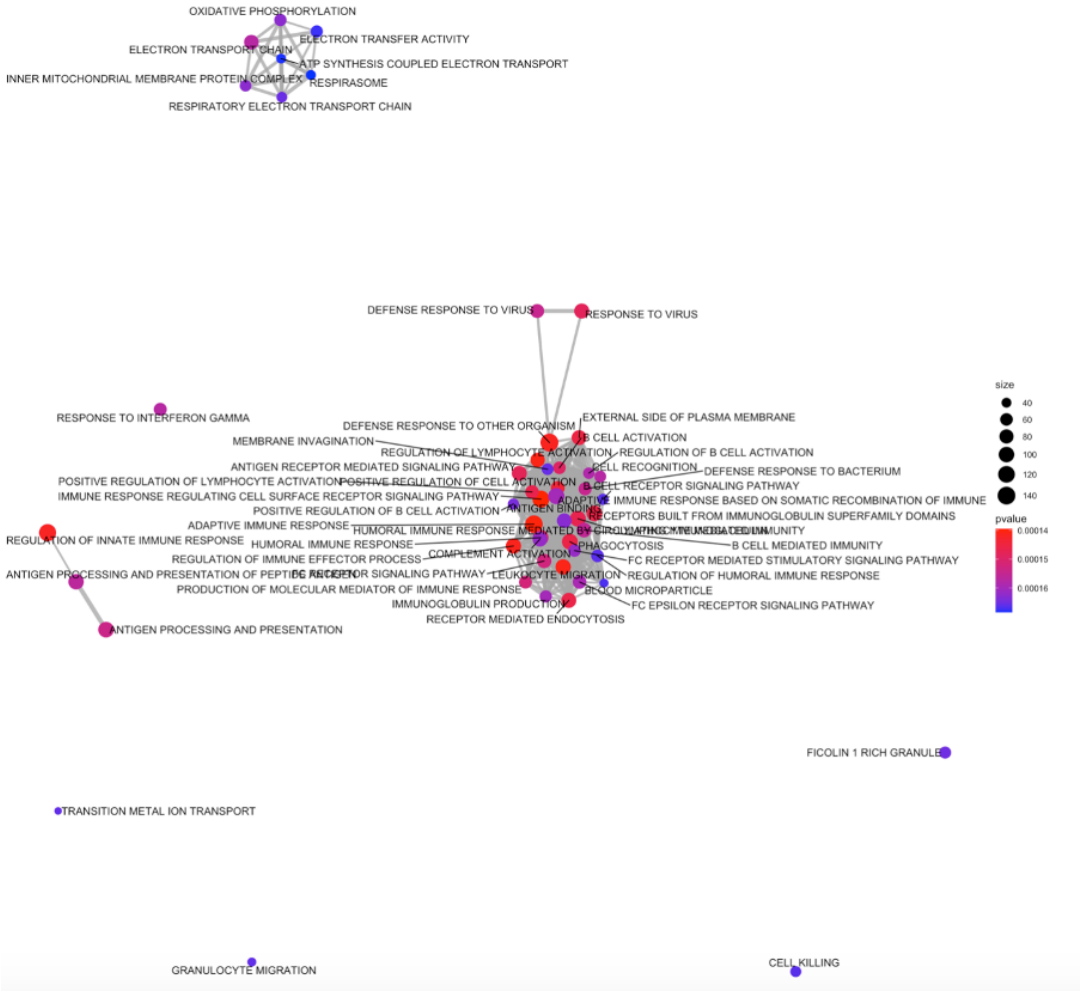
We used MetaFunc's prot2go.py custom script on the 10 differentially abundant bacteria we identified, getting the cumulative gene ontologies from only these 10 DA bacteria per sample. We only used gene ontologies with proportional read counts of greater than or equal to 10 in at least 10% of samples. We used the 'sp_percent' value from Metafunc's output (see https://metafunc.readthedocs.io/en/latest/results.html#directory-source-go) and obtained the fold change of tumor/normal per sample of this value. We show representative heatmaps of these values in **Supplementary Figures 4.5-4.8**, categorized by depth less than 5 (most general GO terms) and more than 12 (most specific GO terms). Depth is defined as the maximum path of the GO term from the root term. We also categorize the gene ontologies by namespace of Biological Process (BP) and Molecular Function (MF). For a more detailed analysis, see https://gitlab.com/alsulit08/uoc_response_rectalca/-/tree/main/Microbiome/GO commit affc1e24.

We did not find any significant differences between the values of these Gene Ontologies for the responders.

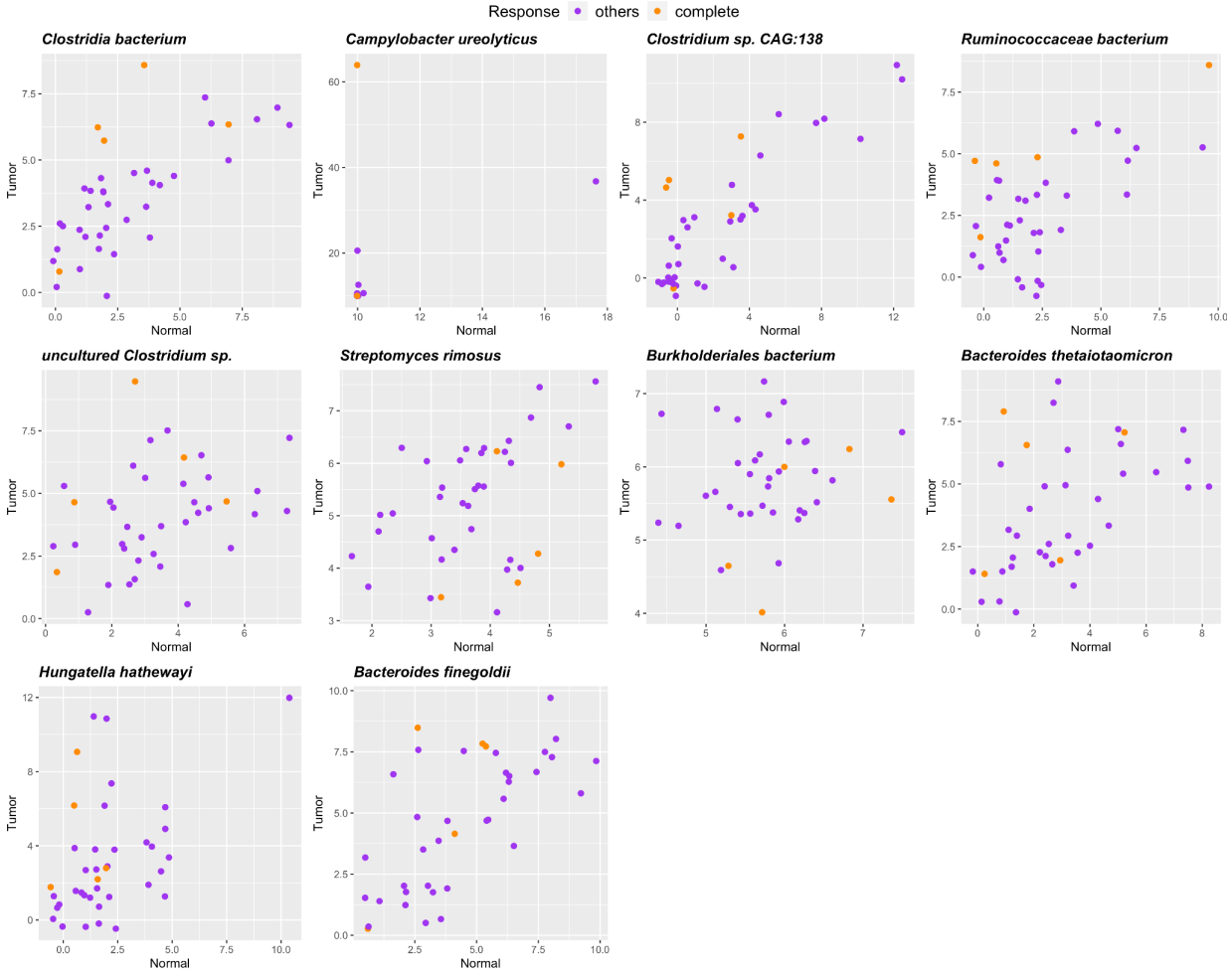
4.6.2 Supplementary Figures



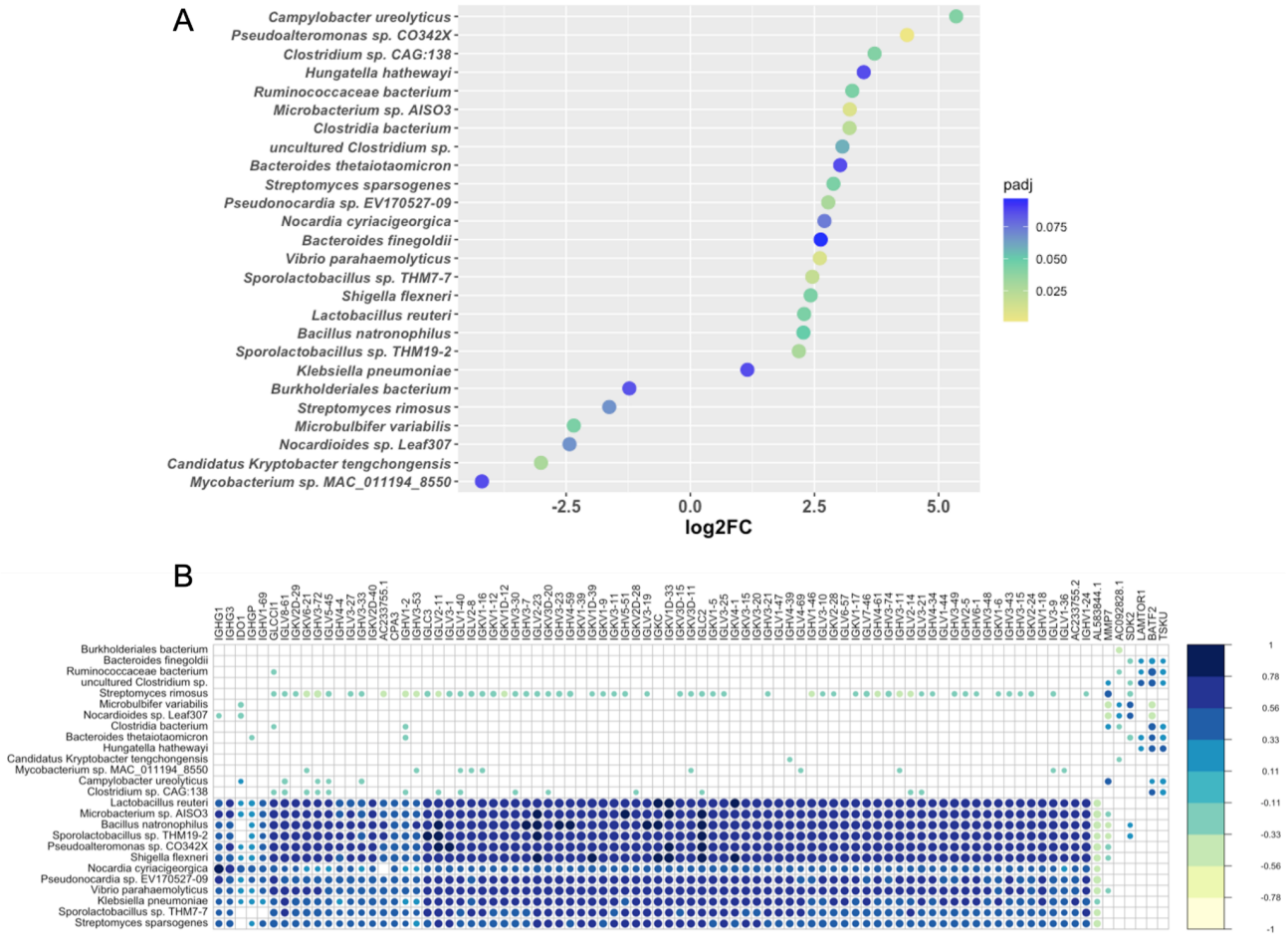
Supplementary Figure 4.1. Scatterplot of DEGs rlog values between tumor samples and their corresponding matched normal samples. We see that for many of these genes, complete responders (orange) cluster at the upper-left quadrant of the plots distinct from other responders (purple), except for SDK2 and AL583844.1, which have negative DESeq2 log fold changes.



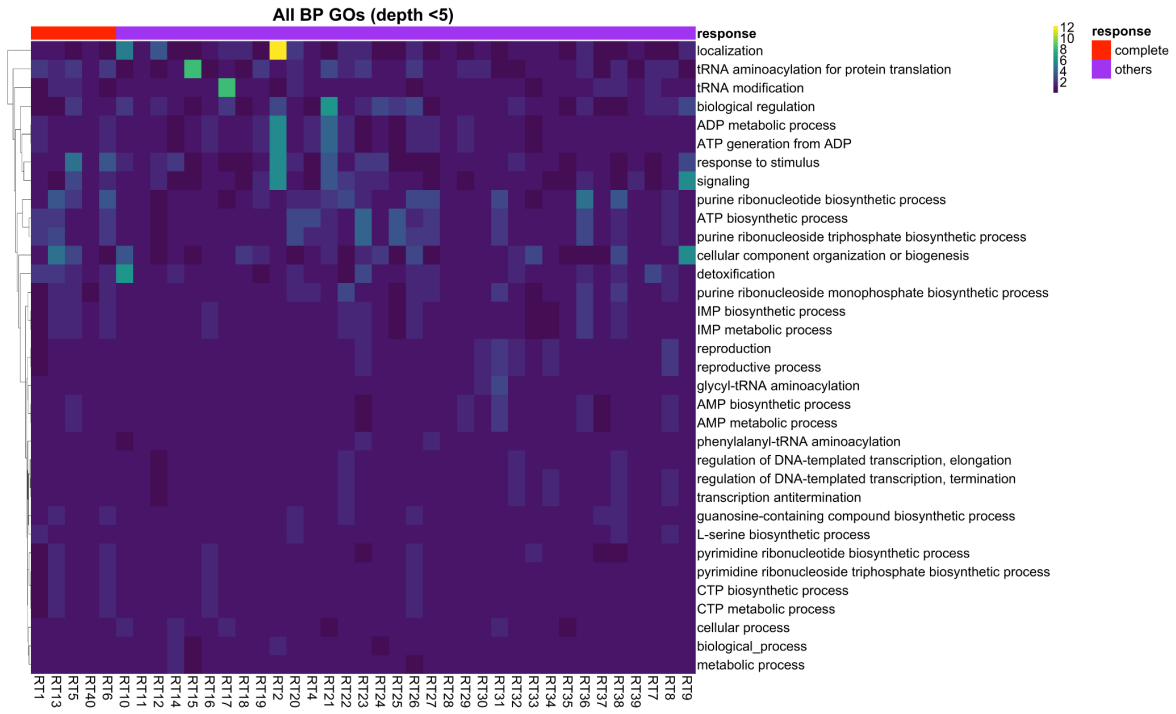
Supplementary Figure 4.2. Enrichment Map of the top 50 enriched gene sets. Enrichment maps show which gene sets have shared genes between them (gene sets connected by lines share genes among them). We see a large cluster of gene sets involved in immune responses.



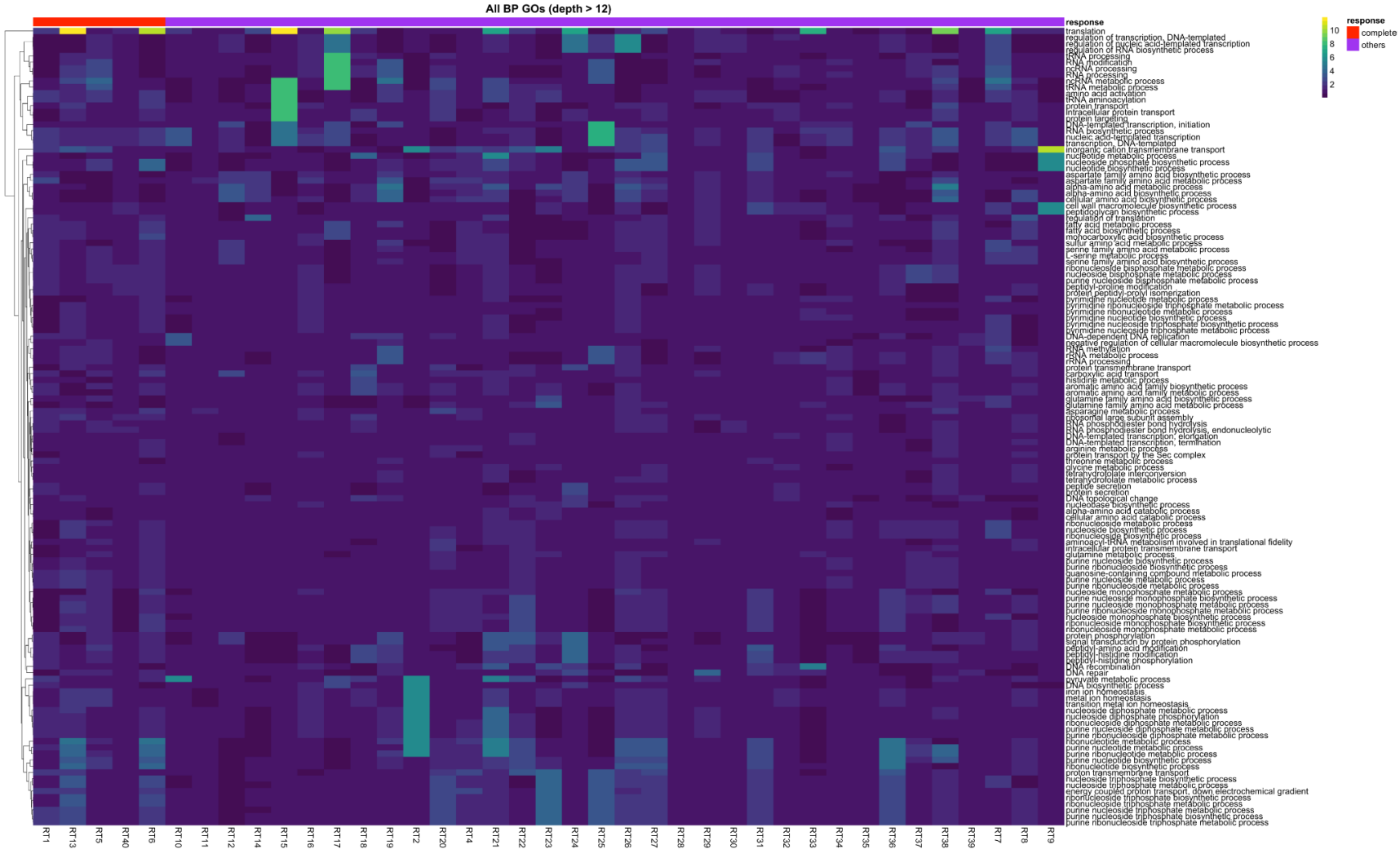
Supplementary Figure 4.3. Differentially abundant bacteria in tumor samples compared to matched normal tissue, specific to complete responders. Scatter plot between rlog values of the bacterial species in tumor samples and their corresponding matched normal samples.



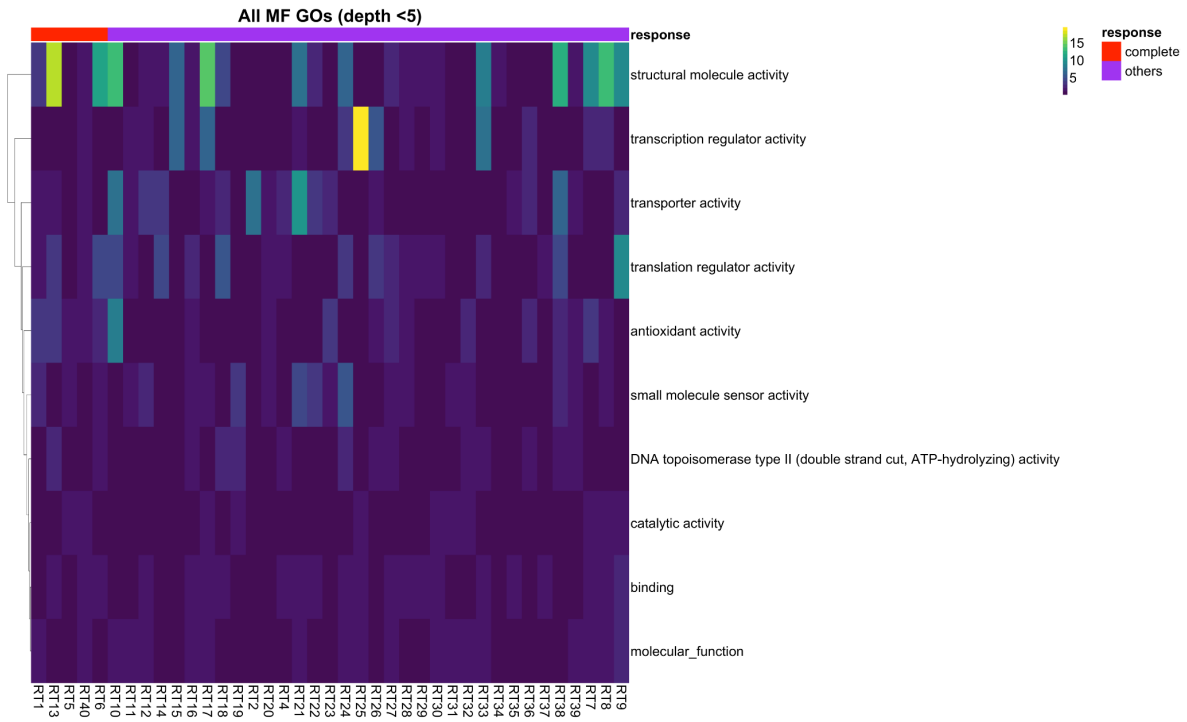
Supplementary Figure 4.4. Original 26 bacteria identified as differentially abundant (DA) in Complete Responders. **A.** Log_2 Fold Changes and adjusted p-values of the 26 DA Bacteria. **B**. Spearman correlation between the rlog values of the 26 DA bacteria and 87 DEGs. Blank spaces represent non-significant correlations. The colorbar represents Spearman correlation coefficients. We see a number of the 26 DA bacteria correlating highly with the Ig DEGs from the host (cor > 0.5).



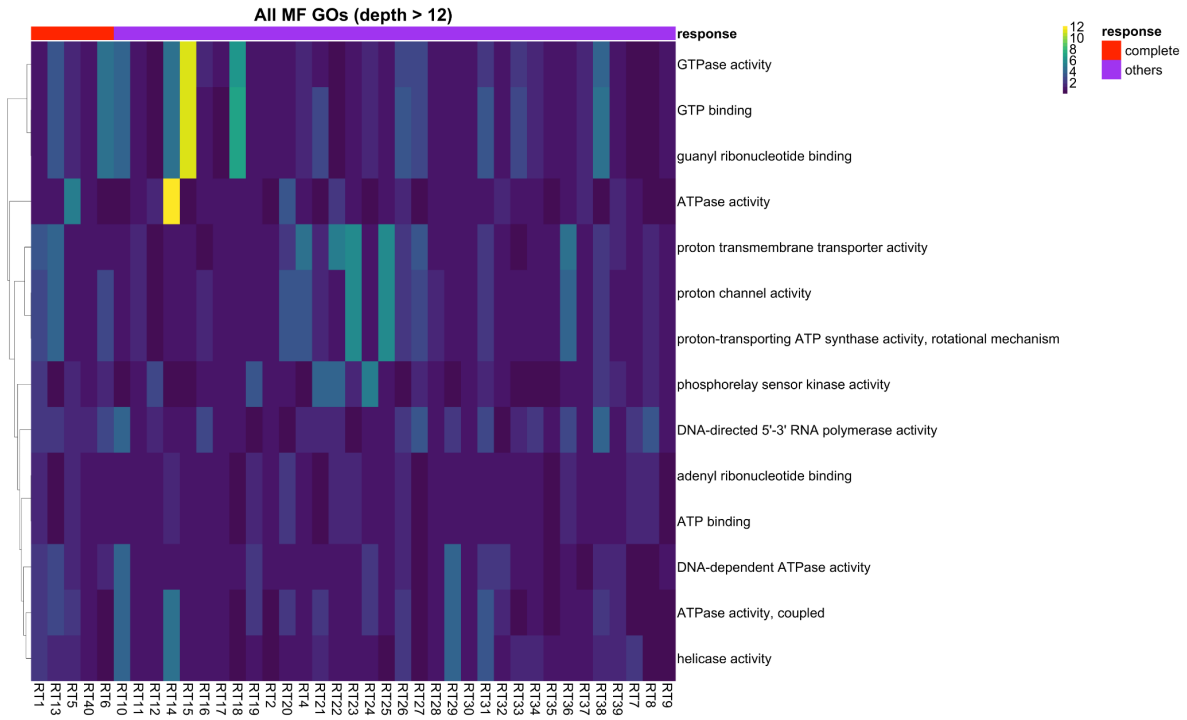
Supplementary Figure 4.5. Biological Process Gene Ontologies, Complete vs Other Responders. Depth < 5. Colorbar represents tumor/normal fold change of `*sp_percent*` values per sample.



Supplementary Figure 4.6. Biological Process Gene Ontologies, Complete vs Other Responders. Depth > 12. Colorbar represents tumor/normal fold change of `*sp_percent*` values per sample.



Supplementary Figure 4.7. Molecular Function Gene Ontologies, Complete vs Other Responders. Depth < 5. Colorbar represents tumor/normal fold change of `*sp_percent*` values per sample.



Supplementary Figure 4.8. Molecular Function Gene Ontologies, Complete vs Other Responders. Depth >12. Colorbar represents tumor/normal fold change of `*sp_percent*` values per sample.

4.6.3 Supplementary Tables

Supplementary Table 4.1. Sample read counts at each stage of filters applied

Samples	All Species	Bacterial Species	Bacterial species (≥10 reads in ≥ 20% samples)	Samples	All Species	Bacterial Species	Bacterial species (≥10 reads in ≥ 20% samples)
1	367426	334300	326851	**42**	479193	448343	443052
2	317406	278701	271299	**43**	601569	564140	554126
4	5931870	5882950	5845561	**44**	452421	395261	385673
5	585255	541121	530142	**45**	410728	365669	328088
6	875245	835945	679145	**46**	513392	476225	467308
7	866635	804903	787767	**47**	480272	443574	432136
8	4860390	4802410	4787394	**48**	383540	351465	341460
9	601895	533753	526513	**49**	551243	515149	508789
10	551296	504847	494214	**50**	699032	631147	619988
11	975784	908424	735936	**51**	425052	376982	364515
12	375162	343675	336541	**52**	320249	281107	261935
13	578534	538725	530001	**53**	537872	495479	487088
14	367454	335204	328846	**54**	453227	430320	425300
15	447437	411601	406028	**55**	2875935	2810923	2777574
16	285138	252577	243678	**56**	562694	519930	509384
17	458340	411431	401716	**57**	295674	267930	261254
18	332139	300591	291267	**58**	473651	432885	423353
19	643355	578661	560692	**59**	1350459	1293165	1282557
20	822941	730470	723644	**60**	415056	384240	376299
21	806288	752946	744397	**61**	1114509	1072929	1046429
22	362957	320352	313032	**62**	699917	647675	639164
24	390506	344275	331246	**63**	505533	461366	456405
25	415956	368846	357479	**64**	958532	912709	900185
26	324131	294920	280149	**65**	652663	599250	591965
27	253146	223906	217432	**66**	1237887	1175208	1163791
28	313007	287574	277655	**67**	1016942	951309	877339
29	360823	338036	331357	**68**	309833	277843	268831
30	348255	310552	304588	**69**	779138	720907	708018
31	682507	647605	599747	**70**	599765	553194	543786
32	557969	520187	508076	**71**	1076467	996567	978575
33	357295	332291	324029	**72**	598850	540178	528115
34	354420	319445	306087	**73**	842029	775016	763428
35	301498	272226	263884	**74**	458221	425640	418649
36	434425	401450	389622	**75**	651332	588586	582518
37	231866	201032	195343	**76**	697161	641296	636027
38	619188	518821	495002	**77**	305206	278644	273270
39	216192	196204	184911	**78**	1466012	1415040	869519
40	423583	387599	376912	**79**	886148	838698	823192
41	552440	521903	514376	**80**	526245	487721	469950

Supplementary Table 4.2. Number of Unique Proteins (as Accession Numbers) Identified as Belonging to the 26 Differentially Abundant Bacteria

Species	Number of Proteins
Pseudoalteromonas sp. CO342X	2
Mycobacterium sp. MAC_011194_8550	4
Microbacterium sp. AISO3	5
Microbulbifer variabilis	7
Sporolactobacillus sp. THM19-2	9
Bacillus natronophilus	9
Sporolactobacillus sp. THM7-7	11
Nocardioides sp. Leaf307	18
Pseudonocardia sp. EV170527-09	24
Candidatus Kryptobacter tengchongensis	39
Streptomyces sparsogenes	104
Lactobacillus reuteri	127
Nocardia cyriacigeorgica	182
Streptomyces rimosus	204
Shigella flexneri	252
Bacteroides finegoldii	333
Vibrio parahaemolyticus	526
Clostridium sp. CAG:138	1531
Burkholderiales bacterium	1697
uncultured Clostridium sp.	2079
Klebsiella pneumoniae	2270
Ruminococcaceae bacterium	2746
Clostridia bacterium	4085
Bacteroides thetaiotaomicron	4161
Campylobacter ureolyticus	7329
Hungatella hathewayi	10489

Chapter 5:

Summary

5.1 Summary of Findings

In this dissertation, I aimed to identify the contributions and interactions of the microbiome with host genes in colorectal cancer development and therapy. In Chapter 2, I described MetaFunc, a pipeline I designed to streamline the computational analysis of microbiome data from raw reads. From raw reads, users can map reads to the host genome with feature identification and quantitation. Reads that do not map are treated as microbial reads, and are processed for microbiome species identification and functional annotation. With this pipeline, I fulfilled the need for a method that would ascribe function to microbiome datasets, as existing methods either do not link the function to the taxonomic identification, or are limited by the databases they use. MetaFunc also offered a way to analyze host and microbiome together in a single workflow, and provided users with initial statistical analysis to explore their data. Lastly, MetaFunc provided users a way to visualize the abundance tables results with an *R* shiny application. I used MetaFunc to obtain gene and microbiome information for Chapters 3 and 4.

Colorectal cancer is a highly heterogeneous disease that differs from patient to patient. This causes difficulties in disease management, diagnosis, prognosis, and

therapy as there exists no single solution for all patients. A thorough understanding of CRC is therefore necessary to offer the best management to patients. CRC is well studied, being among the top causes of cancer mortality worldwide, and categorization of CRC tumors into subtypes has been done to facilitate characterization of similar types of tumors. The microbiome, however, whose impact on CRC development is undeniable, has been largely left out of the subtype discussion. Furthermore, it is unexplored if the type of microbes affect or correlate with the different characteristics of CRC subtypes.

In Chapter 3, I aimed to address this knowledge gap by focusing on CMS1 and CMS4 subtypes. Previous literature described these two subtypes as immune-infiltrated. However, these two subtypes differ in patient prognosis, with CMS1 being generally favorable, and CMS4 having the worst overall survival. Immune response is an important driving force in CRC tumors. The seemingly opposing prognoses in CMS1 and CMS4 with different immune infiltrates is therefore intriguing. I first confirmed, using host gene information from MetaFunc, that CMS1 and CMS4 have an enrichment of immune responses, and that these gene sets are largely immunogenic in CMS1 and regulatory in CMS4. I also confirmed that these gene sets were not present in the CMS2 and CMS3 subtypes. Interestingly, I found indications of response to lipopolysaccharides among the gene sets of CMS1 and CMS4. Lipopolysaccharides are pathogen-associated molecules that can trigger a defensive immune response in humans, and finding up-regulated responses to these molecules in CMS1 and CMS4 is indicative of a role for bacteria in the development of these immune-infiltrated subtypes. I therefore identified differentially abundant bacteria in

CMS1 and CMS4 that have LPS biosynthetic processes annotations, with the aim of testing how these bacteria could affect immune responses *in vitro*. I treated PBMCs with LPS from *F. periodonticum* and *B. fragilis* identified from CMS1, and *P. asaccharolytica* identified in CMS4. These results indicated that while *F. periodonticum LPS* led to an increased release of cytokines from PBMCs, LPS from the other two bacteria had the opposite effect. Moreover, co-incubating *F. periodonticum* LPS with LPS from either of the other bacteria resulted in attenuation of cytokine production by *F. periodonticum* LPS. This showed that synergistic or antagonistic interactions between microbes in the gut are important to consider in CRC development. I propose that different molecules from even a single species could have very different effects in CRC, as seen in *B. fragilis*, whose LPS was shown to attenuate cytokine release in PBMCs, but whose toxin is known to promote immunogenicity in CRC. I also showed that attenuation of immune responses is an important factor to consider in CRC, as many previous studies have only emphasized the immunogenicity related to bacteria in CRC. Here, I showed that attenuation of immune response by bacterial LPS could contribute to CRC development.

The importance of immune responses in CRC extends to therapeutic responses. Indeed, it has been shown previously that although immunotherapy is an important achievement in the field of cancer therapeutics, only a subset of CRCs with microsatellite instability (MSI) responds to it. Studies have also shown that microbes, such as *B. thetaiotaomicron* and *Ruminococcaceae,* contribute to the positive responses in immunotherapy. Immune responses are also important in the efficacy of radiotherapeutics in rectal cancers. Rectal cancers account for around 30% of CRC

cases. Because of its location, the standard treatment for rectal cancers is total mesorectal excision preceded by neoadjuvant chemoradiotherapy (nCRT) to downsize and downgrade tumors. A subset of patients, however, responds completely to nCRT, negating the need for subsequent surgery. As surgery and nCRT both have their own respective side-effects and potential morbidities, a way to predict response to nCRT could possibly allow patients to avoid going through the unnecessary side effects of either procedure. I used MetaFunc to facilitate analysis of paired tumor vs normal (T/N) tissues in complete responders compared to other responders to nCRT in Chapter 4. I found that enriched genes in T/N of complete responders were mostly related to immunoglobulins, and respective enriched gene sets included complement activation and B-cell activation, consistent with the enriched genes I identified. I also identified enriched gene sets related to response to viruses and bacteria, indicating a role for the microbiome in radiotherapy responses. I subsequently identified 10 microbes as enriched in T/N of complete responders, including *B. thetaiotaomicron* and *Ruminococcaceae* bacteria, and their abundances correlated with the abundance of *BATF2*, an enriched gene in T/N of complete responders.

These results reiterate the importance of immune responses in CRC therapeutics. Previous studies have described radiotherapy as similar to an *in situ* vaccine, where ionizing radiation causes cells to die off, presenting mutated peptides and fragmented DNA as antigens that activate immune responses. These immune responses can lead to cytotoxicity that can kill cancer cells. This is akin to a viral infection, which probably explains the enriched gene sets relating to response to

viruses that were found in T/N of complete responders. This also results in activation of complement that is important in the recruitment of dendritic cells and CD8+ cells to combat tumor cells. Furthermore, this could result in BATF2 activation. *BATF2* is an abundantly expressed gene in T/N of complete responders in our cohort. It has been found to have anti-tumor properties by virtue of activating tumor-associated macrophages, which in turn produce IL-12 that recruits CD8+ T-cells. The effects of radiotherapy indicate an overlap with immunotherapeutic responses, which possibly aids in the complete response to radiotherapy. I found *B. thetaiotaomicron* and *Ruminococcaceae* as among the differentially abundant microbes in T/N of complete responders and these two microbes have previous evidence of associations with response to immunotherapy. Possibly, the presence of these microbes influence the immunotherapeutic responses induced by radiotherapy.

5.2 Conclusions

This thesis highlights and reiterates the importance of immune responses in colorectal cancer development and treatment, and that these immune responses are affected by the human gut microbiome. I showed in two chapters, Chapters 3 and 4, that immune responses, possibly affected by the microbiome, result in different molecular subtypes of CRC (CMS1 vs CMS4), and different therapy responses (complete vs incomplete response in radiotherapy). I came to this conclusion using a computational pipeline I developed in Chapter 2. Immune infiltration is recognized as an important factor in cancer development, and the type of infiltrate can dictate how cancer can progress. In this work I showed that LPS from different microbes could

interact and affect immune responses in PBMCs. This informs us that it is possible that LPS from different microbes in the gut could also interact, and their synergy or antagonism could affect the immune responses in CRC development. Moreover, my results also imply that different bacterial molecules could possibly have opposing effects on immune responses as seen from our results of *B. fragilis* LPS and our previous knowledge on its toxin. This further reiterates the complexity of molecular mechanisms influencing this disease. These LPS properties also offer further opportunity to explore how these molecules could be used in CRC immunotherapy.

Immunotherapy harnesses the body's immune system to work against its cancer cells. I showed in Chapter 4 of this thesis that such processes could potentially be harnessed as well in complete responders to radiotherapy. Enriched genes and gene sets in tumors of complete responders indicated that a viral infection-like response could be triggered by radiotherapy that leads to activation of cells that could kill off the tumors. I also found differentially abundant bacteria in these samples that have previously been shown to associate with good response to immunotherapy.

These results further emphasize the complexity of immune responses in CRC development. I used computational methods to mine host and microbiome in CRC subtypes and different responders to radiotherapy. This work opens up new avenues to explore regarding host immune responses, how the gut microbiome could affect them, and how these relationships influence CRC development and therapeutics.

5.3 Limitations and Future Directions

Most of this work was computational, setting the stage for experimentation in the laboratory. I first started with a computational pipeline to analyze host and microbiome reads. Important aspects of this analysis are microbiome identification and differential abundance analysis of these microbes between groups. Having agreement among two or more tools will improve the accuracy of results, and future features to the pipeline can include adding Kraken2 along with Kaiju as a taxonomy identification tool, and I expect concordant results between different methods to give a more truthful picture of microbiome taxonomies. Furthermore, I aim to include the option of using other tools for differential abundance, such as DESeq2 which I used extensively in Chapters 3 and 4 of this thesis. While I mentioned that edgeR is sensitive in finding differentially abundant genes or microbes, it could be prone to false positives and an additional tool for this analysis could give more accurate results.

Furthermore, it is prudent to augment computational results with experimental ones. In Chapter 3, I provided an initial *in vitro* perspective of how LPS from different bacteria affect immune responses in PBMCs. Indeed, because of these results, a provisional patent has been filed for the use of *F. periodonticum* as an adjuvant to immunotherapy. The next crucial steps are to determine how these could affect cancer cell lines, and determine their *in vivo* effects in mouse models. This future research aims to characterize how these LPS could affect immune responses by direct treatment of CRC cell lines, and by treating the cell lines with supernatants from

PBMCs after incubation with LPS. I expect molecules, like the cytokines I measured, produced by PBMCs after LPS treatment, will affect immune responses in cancer cells as well. Furthermore, it would be good to test our hypothesis that in *B. fragilis*, the enterotoxin is responsible for upregulation of immune response and if so, how it could interact with *B. fragilis* LPS and LPS from our other microbes. In mouse models, I also aim to see if *F. periodonticum* LPS could augment immunotherapy, and how it, along with *B. fragilis* and *P. asaccharolytica* LPSs, could affect CRC progression. Lastly, I aim to characterize the LPS from each of these microbes, identifying their respective active sites.

Meanwhile, I have provided a mechanistic hypothesis for complete response in radiotherapy in rectal cancer. I do emphasize that the results from Chapter 4 are wholly computational, and further confirmation of findings would be necessary. Quantitative real-time PCR or profiling using nCounter analysis systems (NanoString, n.d.) could be used to quantitatively confirm the DEGs and DA microbes I found in the tumors of complete responders. Furthermore, the nCounter NanoString panels allow for identification of populations of tumor-infiltrating lymphocytes thereby allowing us to confirm our theory on immune responses in complete responders. Lastly, additional cohorts of rectal cancer biopsies will be useful. This study uses 39 pairs of normal and tumor tissues, with only five of these classified as complete responders. Future work with larger cohorts would reinforce the results of this work.

Bibliography

Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., Goedert, J. J., Hayes, R. B., & Yang, L. (2013). Human Gut Microbiome and Risk for Colorectal Cancer. *JNCI: Journal of the National Cancer Institute*, *105*(24), 1907–1911. https://doi.org/10.1093/jnci/djt300

Alexander, J. L., Scott, A. J., Pouncey, A. L., Marchesi, J., Kinross, J., & Teare, J. (2018). Colorectal carcinogenesis: An archetype of gut microbiota–host interaction. *Ecancermedicalscience*, *12*. https://doi.org/10.3332/ecancer.2018.865

Allen, I. C., TeKippe, E. M., Woodford, R.-M. T., Uronis, J. M., Holl, E. K., Rogers, A. B., Herfarth, H. H., Jobin, C., & Ting, J. P.-Y. (2010). The NLRP3 inflammasome functions as a negative regulator of tumorigenesis during colitis-associated cancer. *The Journal of Experimental Medicine*, *207*(5), 1045–1056. https://doi.org/10.1084/jem.20100050

Angell, H. K., Bruni, D., Barrett, J. C., Herbst, R., & Galon, J. (2020). The Immunoscore: Colon Cancer and Beyond. *Clinical Cancer Research*, *26*(2), 332–339. https://doi.org/10.1158/1078-0432.CCR-18-1851

Argyle, D., & Kitamura, T. (2018). Targeting Macrophage-Recruiting Chemokines as a Novel Therapeutic Strategy to Prevent the Progression of Solid Tumors. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.02629

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Aslani, N., Janbabaei, G., Abastabar, M., Meis, J. F., Babaeian, M., Khodavaisy, S., Boekhout, T., & Badali, H. (2018). Identification of uncommon oral yeasts from cancer patients by MALDI-TOF mass spectrometry. *BMC Infectious Diseases*, *18*(1). https://doi.org/10.1186/s12879-017-2916-5

Baker, K. J., Houston, A., & Brint, E. (2019). IL-1 Family Members in Cancer; Two Sides to Every Story. *Frontiers in Immunology*, *10*. https://doi.org/10.3389/fimmu.2019.01197

Bashiardes, S., Zilberman-Schapira, G., & Elinav, E. (2016). Use of Metatranscriptomics in Microbiome Research. *Bioinformatics and Biology Insights*, *10*, 19–25. https://doi.org/10.4137/BBI.S34610

Bayoumi, A., Sayed, A., Broskova, Z., Teoh, J.-P., Wilson, J., Su, H., Tang, Y.-L., & Kim, I. (2016). Crosstalk between Long Noncoding RNAs and MicroRNAs in Health and Disease. *International Journal of Molecular Sciences*, *17*(3), 356. https://doi.org/10.3390/ijms17030356

Becht, E., de Reyniès, A., Giraldo, N. A., Pilati, C., Buttard, B., Lacroix, L., Selves, J., Sautès-Fridman, C., Laurent-Puig, P., & Fridman, W. H. (2016). Immune and Stromal Classification of Colorectal Cancer Is Associated with Molecular Subtypes and Relevant for Precision Immunotherapy. *Clinical Cancer Research*, *22*(16), 4057–4066. https://doi.org/10.1158/1078-0432.CCR-15-2879

Boland, C. R., & Goel, A. (2010). Microsatellite Instability in Colorectal Cancer. *Gastroenterology*, *138*(6), 2073-2087.e3. https://doi.org/10.1053/j.gastro.2009.12.064

Boland, C. R., Thibodeau, S. N., Hamilton, S. R., Sidransky, D., Eshleman, J. R., Burt, R. W., Meltzer, S. J., Rodriguez-Bigas, M. A., Fodde, R., Ranzani, G. N., & Srivastava, S. (1998). A National Cancer Institute Workshop on Microsatellite Instability for Cancer Detection and Familial Predisposition: Development of International Criteria for the Determination of Microsatellite Instability in Colorectal Cancer. *Cancer Research*, *58*(22), 5248–5257.

Bowel cancer. (2021, August 4). Ministry of Health NZ. https://www.health.govt.nz/your-health/conditions-and-treatments/diseases-and-illnesses/bowel-cancer

Brennan, C. A., & Garrett, W. S. (2016). Gut Microbiota, Inflammation, and Colorectal Cancer. *Annual Review of Microbiology*, *70*, 395–411. https://doi.org/10.1146/annurev-micro-102215-095513

Burkholder, B., Huang, R.-Y., Burgess, R., Luo, S., Jones, V. S., Zhang, W., Lv, Z.-Q., Gao, C.-Y., Wang, B.-L., Zhang, Y.-M., & Huang, R.-P. (2014). Tumor-induced perturbations of cytokines and immune cell networks. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1845*(2), 182–201. https://doi.org/10.1016/j.bbcan.2014.01.004

Burns, M. B., Lynch, J., Starr, T. K., Knights, D., & Blekhman, R. (2015). Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. *Genome Medicine*, *7*(1), 55. https://doi.org/10.1186/s13073-015-0177-8

Burns, M. B., Montassier, E., Abrahante, J., Priya, S., Niccum, D. E., Khoruts, A., Starr, T. K., Knights, D., & Blekhman, R. (2018). *Colorectal cancer mutational profiles correlate with defined microbial communities in the tumor microenvironment*. https://doi.org/10.1101/090795

Calgaro, M., Romualdi, C., Waldron, L., Risso, D., & Vitulo, N. (2020). Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biology*, *21*(1), 191. https://doi.org/10.1186/s13059-020-02104-1

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., & Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database: Sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, *32*(Database issue), D262–D266. https://doi.org/10.1093/nar/gkh021

Carethers, J. M., & Jung, B. H. (2015). Genetics and Genetic Biomarkers in Sporadic Colorectal Cancer. *Gastroenterology*, *149*(5), 1177-1190.e3. https://doi.org/10.1053/j.gastro.2015.06.047

Chen, J., Pitmon, E., & Wang, K. (2017). Microbiome, inflammation and colorectal cancer. *Seminars in Immunology*, *32*, 43–53. https://doi.org/10.1016/j.smim.2017.09.006

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Colangelo, T., Polcaro, G., Muccillo, L., D'Agostino, G., Rosato, V., Ziccardi, P., Lupo, A., Mazzoccoli, G., Sabatino, L., & Colantuoni, V. (2017). Friend or foe?: The tumour microenvironment dilemma in colorectal cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1867*(1), 1–18. https://doi.org/10.1016/j.bbcan.2016.11.001

Coleman, O. I., & Nunes, T. (2016). Role of the Microbiota in Colorectal Cancer: Updates on Microbial Associations and Therapeutic Implications. *BioResearch Open Access*, *5*(1), 279–288. https://doi.org/10.1089/biores.2016.0028

Colon cancer—Diagnosis and treatment—Mayo Clinic. (n.d.). Retrieved October 19, 2021, from https://www.mayoclinic.org/diseases-conditions/colon-cancer/diagnosis-treatment/drc-20353674

Colussi, D., Brandi, G., Bazzoli, F., & Ricciardiello, L. (2013). Molecular Pathways Involved in Colorectal Cancer: Implications for Disease Behavior and Prevention. *International Journal of Molecular Sciences*, *14*(8), 16365–16385. https://doi.org/10.3390/ijms140816365

Cox, M. P., Peterson, D. A., & Biggs, P. J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, *11*(1), 485. https://doi.org/10.1186/1471-2105-11-485

Cremonesi, E., Governa, V., Garzon, J. F. G., Mele, V., Amicarella, F., Muraro, M. G., Trella, E., Galati-Fournier, V., Oertli, D., Däster, S. R., Droeser, R. A., Weixler, B., Bolli, M., Rosso, R., Nitsche, U., Khanna, N., Egli, A., Keck, S., Slotta-Huspenina, J., … Iezzi, G. (2018). Gut microbiota modulate T cell trafficking into human colorectal cancer. *Gut*, gutjnl-2016-313498. https://doi.org/10.1136/gutjnl-2016-313498

Cuevas-Ramos, G., Petit, C. R., Marcq, I., Boury, M., Oswald, E., & Nougayrède, J.-P. (2010). Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(25), 11537–11542. https://doi.org/10.1073/pnas.1001261107

d'Hennezel, E., Abubucker, S., Murphy, L. O., & Cullen, T. W. (2017). Total Lipopolysaccharide from the Human Gut Microbiome Silences Toll-Like Receptor Signaling. *MSystems*, *2*(6). https://doi.org/10.1128/mSystems.00046-17

Dai, Z., Coker, O. O., Nakatsu, G., Wu, W. K. K., Zhao, L., Chen, Z., Chan, F. K. L., Kristiansen, K., Sung, J. J. Y., Wong, S. H., & Yu, J. (2018). Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome*, *6*(1). https://doi.org/10.1186/s40168-018-0451-2

Daillère, R., Derosa, L., Bonvalet, M., Segata, N., Routy, B., Gariboldi, M., Budinská, E., De Vries, I. J. M., Naccarati, A. G., Zitvogel, V., Caldas, C., Engstrand, L., Loilbl, S., Fieschi, J., Heinzerling, L., Kroemer, G., & Zitvogel, L. (2020). Trial watch: The gut microbiota as a tool to boost the clinical efficacy of anticancer immunotherapy. *OncoImmunology*, *9*(1), 1774298. https://doi.org/10.1080/2162402X.2020.1774298

Dayde, D., Tanaka, I., Jain, R., Tai, M. C., & Taguchi, A. (2017). Predictive and Prognostic Molecular Biomarkers for Response to Neoadjuvant Chemoradiation in Rectal Cancer. *International Journal of Molecular Sciences*, *18*(3), 573. https://doi.org/10.3390/ijms18030573

De Almeida, C. V., de Camargo, M. R., Russo, E., & Amedei, A. (2019). Role of diet and gut microbiota on colorectal cancer immunomodulation. *World Journal of Gastroenterology*, *25*(2), 151–162. https://doi.org/10.3748/wjg.v25.i2.151

De la Fuente López, M., Landskron, G., Parada, D., Dubois-Camacho, K., Simian, D., Martinez, M., Romero, D., Roa, J. C., Chahuán, I., Gutiérrez, R., Lopez-K, F., Alvarez, K., Kronberg, U., López, S., Sanguinetti, A., Moreno, N., Abedrapo, M., González, M.-J., Quera, R., & Hermoso-R, M. A. (2018). The relationship between chemokines CCL2, CCL3, and CCL4 with the tumor microenvironment and tumor-associated macrophage markers in colorectal cancer. *Tumor Biology*, *40*(11), 1010428318810059. https://doi.org/10.1177/101042831881005

De Sousa E Melo, F., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L. P. M. H., de Jong, J. H., de Boer, O. J., van Leersum, R., Bijlsma, M. F., Rodermond, H., van der Heijden, M., van Noesel, C. J. M., Tuynman, J. B., Dekker, E., Markowetz, F., Medema, J. P., & Vermeulen, L. (2013). Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Medicine*, *19*(5), 614–618. https://doi.org/10.1038/nm.3174

Desnos-Ollivier, M., Moquet, O., Chouaki, T., Guérin, A.-M., & Dromer, F. (2011). Development of Echinocandin Resistance in Clavispora lusitaniae during Caspofungin Treatment. *Journal of Clinical Microbiology*, *49*(6), 2304–2306. https://doi.org/10.1128/JCM.00325-11

Di Lorenzo, F., Pither, M. D., Martufi, M., Scarinci, I., Guzmán-Caldentey, J., Łakomiec, E., Jachymek, W., Bruijns, S. C. M., Santamaría, S. M., Frick, J.-S., van Kooyk, Y., Chiodo, F., Silipo, A., Bernardini, M. L., & Molinaro, A. (2020). Pairing Bacteroides vulgatus LPS Structure with Its Immunomodulatory Effects on Human Cellular Models. *ACS Central Science*, *6*(9), 1602–1616. https://doi.org/10.1021/acscentsci.0c00791

Dienstmann, R., Vermeulen, L., Guinney, J., Kopetz, S., Tejpar, S., & Tabernero, J. (2017). Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer*, *17*(2), 79–92. https://doi.org/10.1038/nrc.2016.126

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Douglas, G. M., Maffei, V. J., Zaneveld, J., Yurgel, S. N., Brown, J. R., Taylor, C. M., Huttenhower, C., & Langille, M. G. I. (2019). PICRUSt2: An improved and extensible approach for metagenome inference. *BioRxiv*, 672295. https://doi.org/10.1101/672295

Eisenstein, M. (2020). The hunt for a healthy microbiome. *Nature*, *577*(7792), S6–S8. https://doi.org/10.1038/d41586-020-00193-3

Evrard, C., Tachon, G., Randrian, V., Karayan-Tapon, L., & Tougeron, D. (2019). Microsatellite Instability: Diagnosis, Heterogeneity, Discordance, and Clinical Impact in Colorectal Cancer. *Cancers*, *11*(10), 1567. https://doi.org/10.3390/cancers11101567

Fan, Q., Shang, F., Chen, C., Zhou, H., Fan, J., Yang, M., Nie, X., Liu, L., Cai, K., & Liu, H. (2021). Microbial Characteristics of Locally Advanced Rectal Cancer Patients After Neoadjuvant Chemoradiation Therapy According to Pathologic Response. *Cancer Management and Research*, *13*, 2655–2667. https://doi.org/10.2147/CMAR.S294936

Fearon, E. R., & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, *61*(5), 759–767. https://doi.org/10.1016/0092-8674(90)90186-I

Feeney, G., Sehgal, R., Sheehan, M., Hogan, A., Regan, M., Joyce, M., & Kerin, M. (2019). Neoadjuvant radiotherapy for rectal cancer management. *World Journal of Gastroenterology*, *25*(33), 4850–4869. https://doi.org/10.3748/wjg.v25.i33.4850

Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., & Gloor, G. B. (2013). ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLoS ONE*, *8*(7), e67019. https://doi.org/10.1371/journal.pone.0067019

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, *2*(1), 15. https://doi.org/10.1186/2049-2618-2-15

Fernandez, N. F., Gundersen, G. W., Rahman, A., Grimes, M. L., Rikova, K., Hornbeck, P., & Ma'ayan, A. (2017). Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific Data*, *4*(1), 1–12. https://doi.org/10.1038/sdata.2017.151

Ferrier, L., Mazelin, L., Cenac, N., Desreumaux, P., Janin, A., Emilie, D., Colombel, J.-F., Garcia-Villar, R., Fioramonti, J., & Bueno, L. (2003). Stress-induced disruption of colonic epithelial barrier: Role of interferon-γ and myosin light chain kinase in mice. *Gastroenterology*, *125*(3), 795–804. https://doi.org/10.1016/S0016-5085(03)01057-6

Fidelle, M., Yonekura, S., Picard, M., Cogdill, A., Hollebecque, A., Roberti, M. P., & Zitvogel, L. (2020). Resolving the Paradox of Colon Cancer Through the Integration of Genetics, Immunology, and the Microbiota. *Frontiers in Immunology*, *11*. https://doi.org/10.3389/fimmu.2020.600886

Fisher, D. T., Appenheimer, M. M., & Evans, S. S. (2014). The Two Faces of IL-6 in the Tumor Microenvironment. *Seminars in Immunology*, *26*(1), 38–47. https://doi.org/10.1016/j.smim.2014.01.008

Flemer, B., Warren, R. D., Barrett, M. P., Cisek, K., Das, A., Jeffery, I. B., Hurley, E., O'Riordain, M., Shanahan, F., & O'Toole, P. W. (2018). The oral microbiota in colorectal cancer is distinctive and predictive. *Gut*, *67*(8), 1454–1463. https://doi.org/10.1136/gutjnl-2017-314814

Fontana, E., Eason, K., Cervantes, A., Salazar, R., & Sadanandam, A. (2019). Context matters—Consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. *Annals of Oncology*, *30*(4), 520–527. https://doi.org/10.1093/annonc/mdz052

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., & Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, *15*(11), 962. https://doi.org/10.1038/s41592-018-0176-y

Fulbright, L. E., Ellermann, M., & Arthur, J. C. (2017). The microbiome and the hallmarks of cancer. *PLOS Pathogens*, *13*(9), e1006480. https://doi.org/10.1371/journal.ppat.1006480

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., Wind, P., Zinzindohoué, F., Bruneval, P., Cugnenc, P.-H., Trajanoski, Z., Fridman, W.-H., & Pagès, F. (2006). Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. *Science*, *313*(5795), 1960–1964. https://doi.org/10.1126/science.1129139

Ganesh, K., Stadler, Z. K., Cercek, A., Mendelsohn, R. B., Shia, J., Segal, N. H., & Diaz, L. A. (2019). Immunotherapy in colorectal cancer: Rationale, challenges and potential. *Nature Reviews Gastroenterology & Hepatology*, *16*(6), 361–375. https://doi.org/10.1038/s41575-019-0126-x

Gao, Z., Guo, B., Gao, R., Zhu, Q., & Qin, H. (2015). Microbiota disbiosis is associated with colorectal cancer. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.00020

Garland, M. L., Vather, R., Bunkley, N., Pearse, M., & Bissett, I. P. (2014). Clinical tumour size and nodal status predict pathologic complete response following neoadjuvant chemoradiotherapy for rectal cancer. *International Journal of Colorectal Disease*, *29*(3), 301–307. https://doi.org/10.1007/s00384-013-1821-7

Gerner, E. W., & Meyskens, F. L. (2004). Polyamines and cancer: Old molecules, new understanding. *Nature Reviews. Cancer*, *4*(10), 781–792. https://doi.org/10.1038/nrc1454

Ghanipour, A., Jirström, K., Pontén, F., Glimelius, B., Påhlman, L., & Birgisson, H. (2009). The Prognostic Significance of Tryptophanyl-tRNA Synthetase in Colorectal Cancer. *Cancer Epidemiology and Prevention Biomarkers*, *18*(11), 2949–2956. https://doi.org/10.1158/1055-9965.EPI-09-0456

Gloor, G. B., Macklaim, J. M., & Fernandes, A. D. (2016). Displaying Variation in Large Datasets: Plotting a Visual Summary of Effect Sizes. *Journal of Computational and Graphical Statistics*, *25*(3), 971–979. https://doi.org/10.1080/10618600.2015.1131161

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.02224

Gloor, G. B., & Reid, G. (2016). Compositional analysis: A valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, *62*(8), 692–703. https://doi.org/10.1139/cjm-2015-0821

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., & Egozcue, J. J. (2016). It's all relative: Analyzing microbiome data as compositions. *Annals of Epidemiology*, *26*(5), 322–329. https://doi.org/10.1016/j.annepidem.2016.03.003

Goffau, M. C. de, Lager, S., Salter, S. J., Wagner, J., Kronbichler, A., Charnock-Jones, D. S., Peacock, S. J., Smith, G. C. S., & Parkhill, J. (2018). Recognizing the reagent microbiome. *Nature Microbiology*, *3*(8), 851–853. https://doi.org/10.1038/s41564-018-0202-y

Goodwin, A. C., Shields, C. E. D., Wu, S., Huso, D. L., Wu, X., Murray-Stewart, T. R., Hacker-Prietz, A., Rabizadeh, S., Woster, P. M., Sears, C. L., & Casero, R. A. (2011). Polyamine catabolism contributes to enterotoxigenic Bacteroides fragilis-induced colon tumorigenesis. *Proceedings of the National Academy of Sciences*, *108*(37), 15354–15359. https://doi.org/10.1073/pnas.1010203108

Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., Prieto, P. A., Vicente, D., Hoffman, K., Wei, S. C., Cogdill, A. P., Zhao, L., Hudgens, C. W., Hutchinson, D. S., Manzo, T., Macedo, M. P. de, Cotechini, T., Kumar, T., Chen, W. S., … Wargo, J. A. (2018). Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science*, *359*(6371), 97–103. https://doi.org/10.1126/science.aan4236

Granados-Romero, J. J., Valderrama-Treviño, A. I., Contreras-Flores, E. H., Barrera-Mera, B., Herrera Enríquez, M., Uriarte-Ruíz, K., Ceballos-Villalba, J. C., Estrada-Mata, A. G., Alvarado Rodríguez, C., & Arauz-Peña, G. (2017). Colorectal cancer: A review. *International Journal of Research in Medical Sciences*, *5*(11), 4667. https://doi.org/10.18203/2320-6012.ijrms20174914

GSEA. (n.d.). Retrieved May 20, 2020, from https://www.gsea-msigdb.org/gsea/index.jsp

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B. M., Morris, J. S., Simon, I. M., Gerster, S., Fessler, E., de Sousa e Melo, F., Missiaglia, E., Ramay, H., Barras, D., … Tejpar, S. (2015). The Consensus Molecular Subtypes of Colorectal Cancer. *Nature Medicine*, *21*(11), 1350–1356. https://doi.org/10.1038/nm.3967

Guler, R., Roy, S., Suzuki, H., & Brombacher, F. (2015). Targeting Batf2 for infectious diseases and cancer. *Oncotarget*, *6*(29), 26575–26582.

Guo, M., Xu, E., & Ai, D. (2019). Inferring Bacterial Infiltration in Primary Colorectal Tumors From Host Whole Genome Sequencing Data. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00213

Guo, Z. S. (2018). The 2018 Nobel Prize in medicine goes to cancer immunotherapy (editorial for BMC cancer). *BMC Cancer*, *18*. https://doi.org/10.1186/s12885-018-5020-3

Guthrie, L., Gupta, S., Daily, J., & Kelly, L. (2017). Human microbiome signatures of differential colorectal cancer drug metabolism. *Npj Biofilms and Microbiomes*, *3*(1), 1–8. https://doi.org/10.1038/s41522-017-0034-1

Habr-Gama, A., Perez, R. O., Nadalin, W., Sabbaga, J., Ribeiro, U., Silva e Sousa, A. H., Campos, F. G., Kiss, D. R., & Gama-Rodrigues, J. (2004). Operative Versus Nonoperative Treatment for Stage 0 Distal Rectal Cancer Following Chemoradiation Therapy. *Annals of Surgery*, *240*(4), 711–718. https://doi.org/10.1097/01.sla.0000141194.27992.32

Hagland, H. R., Berg, M., Jolma, I. W., Carlsen, A., & Søreide, K. (2013). Molecular Pathways and Cellular Metabolism in Colorectal Cancer. *Digestive Surgery*, *30*(1), 12–25. https://doi.org/10.1159/000347166

Hale, V. L., Jeraldo, P., Chen, J., Mundy, M., Yao, J., Priya, S., Keeney, G., Lyke, K., Ridlon, J., White, B. A., French, A. J., Thibodeau, S. N., Diener, C., Resendis-Antonio, O., Gransee, J., Dutta, T., Petterson, X.-M., Sung, J., Blekhman, R., … Chia, N. (2018). Distinct microbes, metabolites, and ecologies define the microbiome in deficient and proficient mismatch repair colorectal cancers. *Genome Medicine*, *10*(1), 78. https://doi.org/10.1186/s13073-018-0586-6

Hardy, R. G. (2000). ABC of colorectal cancer: Molecular basis for risk factors. *BMJ*, *321*(7265), 886–889. https://doi.org/10.1136/bmj.321.7265.886

Herman, J. M., Narang, A. K., Griffith, K. A., Zalupski, M. M., Reese, J. B., Gearhart, S. L., Azad, N. S., Chan, J., Olsen, L., Efron, J. E., Lawrence, T. S., & Ben-Josef, E. (2013). The Quality-of-Life Effects of Neoadjuvant Chemoradiation in Locally Advanced Rectal Cancer. *International Journal of Radiation Oncology*Biology*Physics*, *85*(1), e15–e19. https://doi.org/10.1016/j.ijrobp.2012.09.006

Huang, H., McGarvey, P. B., Suzek, B. E., Mazumder, R., Zhang, J., Chen, Y., & Wu, C. H. (2011). A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics*, *27*(8), 1190–1191. https://doi.org/10.1093/bioinformatics/btr101

Huber, V., Fais, S., Iero, M., Lugini, L., Canese, P., Squarcina, P., Zaccheddu, A., Colone, M., Arancia, G., Gentile, M., Seregni, E., Valenti, R., Ballabio, G., Belli, F., Leo, E., Parmiani, G., & Rivoltini, L. (2005). Human colorectal cancer cells induce T-cell death through release of proapoptotic microvesicles: Role in immune escape. *Gastroenterology*, *128*(7), 1796–1804. https://doi.org/10.1053/j.gastro.2005.03.045

Hugerth, L. W., & Andersson, A. F. (2017). Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.01561

Hussan, H., Clinton, S. K., Roberts, K., & Bailey, M. T. (2017). Fusobacterium's link to colorectal neoplasia sequenced: A systematic review and future insights. *World Journal of Gastroenterology*, *23*(48), 8626–8650. https://doi.org/10.3748/wjg.v23.i48.8626

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., … The Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214. https://doi.org/10.1038/nature11234

Inamura, K. (2018). Colorectal Cancers: An Update on Their Molecular Pathology. *Cancers*, *10*(1), 26. https://doi.org/10.3390/cancers10010026

Jang, B.-S., Chang, J. H., Chie, E. K., Kim, K., Park, J. W., Kim, M. J., Song, E.-J., Nam, Y.-D., Kang, S. W., Jeong, S.-Y., & Kim, H. J. (2020). Gut Microbiome Composition Is Associated with a Pathologic Response After Preoperative Chemoradiation in Patients with Rectal Cancer. *International Journal of Radiation Oncology, Biology, Physics*, *107*(4), 736–746. https://doi.org/10.1016/j.ijrobp.2020.04.015

Janney, A., Powrie, F., & Mann, E. H. (2020). Host–microbiota maladaptation in colorectal cancer. *Nature*, *585*(7826), 509–517. https://doi.org/10.1038/s41586-020-2729-3

Jin, M. (2019). Unique roles of tryptophanyl-tRNA synthetase in immune control and its therapeutic implications. *Experimental & Molecular Medicine*, *51*(1). https://doi.org/10.1038/s12276-018-0196-9

Jung, Y.-S., Jun, S., Lee, S. H., Sharma, A., & Park, J.-I. (2015). Wnt2 complements Wnt/β-catenin signaling in colorectal cancer. *Oncotarget*, *6*(35), 37257–37268.

Kanemaru, H., Yamane, F., Fukushima, K., Matsuki, T., Kawasaki, T., Ebina, I., Kuniyoshi, K., Tanaka, H., Maruyama, K., Maeda, K., Satoh, T., & Akira, S. (2017). Antitumor effect of Batf2 through IL-12 p40 up-regulation in tumor-associated macrophages. *Proceedings of the National Academy of Sciences*, *114*(35), E7331–E7340. https://doi.org/10.1073/pnas.1708598114

Karamalegos, A., Vazquez-Prada, M., & Ezcurra, M. (2020). What Is a Healthy Microbiome? In J. Sholl & S. I. S. Rattan (Eds.), *Explaining Health Across the Sciences* (pp. 221–241). Springer International Publishing. https://doi.org/10.1007/978-3-030-52663-4_14

Karpinski, P., Rossowska, J., & Sasiadek, M. M. (2017). Immunological landscape of consensus clusters in colorectal cancer. *Oncotarget*, *8*(62), 105299–105311. https://doi.org/10.18632/oncotarget.22169

Kassambara, A. (2021). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests* (0.7.0) [Computer software]. https://CRAN.R-project.org/package=rstatix

Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., & Tang, H. (2018). GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-28948-z

Kohlmann, W., & Gruber, S. (2018, February). *Hnpcc-msi.pdf*. https://www.ncbi.nlm.nih.gov/books/NBK1211/bin/hnpcc-msi.pdf

Kolde, R. (2019). *pheatmap: Pretty Heatmaps* (1.0.12) [Computer software]. https://CRAN.R-project.org/package=pheatmap

Koliarakis, I., Messaritakis, I., Nikolouzakis, T. K., Hamilos, G., Souglakos, J., & Tsiaoussis, J. (2019). Oral Bacteria and Intestinal Dysbiosis in Colorectal Cancer. *International Journal of Molecular Sciences*, *20*(17). https://doi.org/10.3390/ijms20174146

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Krcmery, V., Mateicka, F., Grausova, S., Kunova, A., & Hanzen, J. (1999). Invasive infections due to *Clavispora lusitaniae*. *FEMS Immunology & Medical Microbiology*, *23*(1), 75–78. https://doi.org/10.1111/j.1574-695X.1999.tb01719.x

Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., van de Velde, C. J. H., & Watanabe, T. (2015). Colorectal cancer. *Nature Reviews Disease Primers*, *1*(1), 1–25. https://doi.org/10.1038/nrdp.2015.65

Kumar, R., Herold, J. L., Schady, D., Davis, J., Kopetz, S., Martinez-Moczygemba, M., Murray, B. E., Han, F., Li, Y., Callaway, E., Chapkin, R. S., Dashwood, W.-M., Dashwood, R. H., Berry, T., Mackenzie, C., & Xu, Y. (2017). Streptococcus gallolyticus subsp. Gallolyticus promotes colorectal tumor development. *PLOS Pathogens*, *13*(7), e1006440. https://doi.org/10.1371/journal.ppat.1006440

Langille, M. G. I. (2018). Exploring Linkages between Taxonomic and Functional Profiles of the Human Microbiome. *MSystems*, *3*(2), e00163-17. https://doi.org/10.1128/mSystems.00163-17

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, *31*(9), 814–821. https://doi.org/10.1038/nbt.2676

Leon-Cabrera, S., Vázquez-Sandoval, A., Molina-Guzman, E., Delgado-Ramirez, Y., Delgado-Buenrostro, N. L., Callejas, B. E., Chirino, Y. I., Pérez-Plasencia, C., Rodríguez-Sosa, M., Olguín, J. E., Salinas, C., Satoskar, A. R., & Terrazas, L. I. (2018). Deficiency in STAT1 Signaling Predisposes Gut Inflammation and Prompts Colorectal Cancer Development. *Cancers*, *10*(9). https://doi.org/10.3390/cancers10090341

Lhuillier, C., Rudqvist, N.-P., Elemento, O., Formenti, S. C., & Demaria, S. (2019). Radiation therapy and anti-tumor immunity: Exposing immunogenic mutations to the immune system. *Genome Medicine*, *11*(1), 40. https://doi.org/10.1186/s13073-019-0653-7

Li, J., Huang, L., Zhao, H., Yan, Y., & Lu, J. (2020). The Role of Interleukins in Colorectal Cancer. *International Journal of Biological Sciences*, *16*(13), 2323–2339. https://doi.org/10.7150/ijbs.46651

Li, M., Zhao, L., Li, S., Li, J., Gao, B., Wang, F., Wang, S., Hu, X., Cao, J., & Wang, G. (2018). Differentially expressed lncRNAs and mRNAs identified by NGS analysis in colorectal cancer patients. *Cancer Medicine*, *7*(9), 4650–4664. https://doi.org/10.1002/cam4.1696

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

Liu, Z., Wei, P., Yang, Y., Cui, W., Cao, B., Tan, C., Yu, B., Bi, R., Xia, K., Chen, W., Wang, Y., Zhang, Y., Du, X., & Zhou, X. (2015). BATF2 Deficiency Promotes Progression in Human Colorectal Cancer via Activation of HGF/MET Signaling: A Potential Rationale for Combining MET Inhibitors with IFNs. *Clinical Cancer Research*, *21*(7), 1752–1763. https://doi.org/10.1158/1078-0432.CCR-14-1564

Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome Medicine*, *8*, 51. https://doi.org/10.1186/s13073-016-0307-y

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., Lake, K., Landers, C. J., Mallick, H., Plichta, D. R., … Huttenhower, C. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, *569*(7758), 655–662. https://doi.org/10.1038/s41586-019-1237-9

Loftus, M., Hassouneh, S. A.-D., & Yooseph, S. (2021). Bacterial community structure alterations within the colorectal cancer gut microbiome. *BMC Microbiology*, *21*(1), 98. https://doi.org/10.1186/s12866-021-02153-x

Love, M. I., Huber, W., & Anders, S. (2014). *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. https://doi.org/10.1101/002832

Macklaim, J. M., & Gloor, G. B. (2018). From RNA-seq to Biological Inference: Using Compositional Data Analysis in Meta-Transcriptomics. In R. G. Beiko, W. Hsiao, & J. Parkinson (Eds.), *Microbiome Analysis: Methods and Protocols* (pp. 193–213). Springer New York. https://doi.org/10.1007/978-1-4939-8728-3_13

Mager, L. F., Wasmer, M.-H., Rau, T. T., & Krebs, P. (2016). Cytokine-Induced Modulation of Colorectal Cancer. *Frontiers in Oncology*, *6*. https://doi.org/10.3389/fonc.2016.00096

Malinowski, B., Węsierska, A., Zalewska, K., Sokołowska, M. M., Bursiewicz, W., Socha, M., Ozorowski, M., Pawlak-Osińska, K., & Wiciński, M. (2019). The role of Tannerella forsythia and Porphyromonas gingivalis in pathogenesis of esophageal cancer. *Infectious Agents and Cancer*, *14*(1), 3. https://doi.org/10.1186/s13027-019-0220-2

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microbial Ecology in Health & Disease*, *26*(0). https://doi.org/10.3402/mehd.v26.27663

Marchesi, J. R., Dutilh, B. E., Hall, N., Peters, W. H. M., Roelofs, R., Boleij, A., & Tjalsma, H. (2011). Towards the Human Colorectal Cancer Microbiome. *PLOS ONE*, *6*(5), e20447. https://doi.org/10.1371/journal.pone.0020447

Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M.-C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J.-F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., … Boige, V. (2013). Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *PLoS Medicine*, *10*(5), e1001453. https://doi.org/10.1371/journal.pmed.1001453

Markman, J. L., & Shiao, S. L. (2015). Impact of the immune system and immunotherapy in colorectal cancer. *Journal of Gastrointestinal Oncology*, *6*(2), 208–223. https://doi.org/10.3978/j.issn.2078-6891.2014.077

Matson, V., Fessler, J., Bao, R., Chongsuwat, T., Zha, Y., Alegre, M.-L., Luke, J. J., & Gajewski, T. F. (2018). The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. *Science*, *359*(6371), 104–108. https://doi.org/10.1126/science.aao3290

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297. https://doi.org/10.1093/nar/gks042

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*, 11257. https://doi.org/10.1038/ncomms11257

Methé, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., Gevers, D., Petrosino, J. F., Abubucker, S., Badger, J. H., Chinwalla, A. T., Earl, A. M., FitzGerald, M. G., Fulton, R. S., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., … The Human Microbiome Project Consortium. (2012). A framework for human microbiome research. *Nature*, *486*(7402), 215–221. https://doi.org/10.1038/nature11209

Nakatsu, G., Li, X., Zhou, H., Sheng, J., Wong, S. H., Wu, W. K. K., Ng, S. C., Tsoi, H., Dong, Y., Zhang, N., He, Y., Kang, Q., Cao, L., Wang, K., Zhang, J., Liang, Q., Yu, J., & Sung, J. J. Y. (2015). Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nature Communications*, *6*, 8727. https://doi.org/10.1038/ncomms9727

NanoString. (n.d.). *NCounter Systems Overview*. NanoString. Retrieved May 2, 2022, from https://nanostring.com/products/ncounter-analysis-system/ncounter-systems-overview/

Nayfach, S., Bradley, P. H., Wyman, S. K., Laurent, T. J., Williams, A., Eisen, J. A., Pollard, K. S., & Sharpton, T. J. (2015). Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. *PLOS Computational Biology*, *11*(11), e1004573. https://doi.org/10.1371/journal.pcbi.1004573

Ohkubo, K., Sakai, Y., Inoue, H., Akamine, S., Ishizaki, Y., Matsushita, Y., Sanefuji, M., Torisu, H., Ihara, K., Sardiello, M., & Hara, T. (2015). Moyamoya disease susceptibility gene RNF213 links inflammatory and angiogenic signals in endothelial cells. *Scientific Reports*, *5*(1), 13191. https://doi.org/10.1038/srep13191

Park, R., Umar, S., & Kasi, A. (2020). Immunotherapy in Colorectal Cancer: Potential of Fecal Transplant and Microbiota-Augmented Clinical Trials. *Current Colorectal Cancer Reports*, *16*(4), 81–88. https://doi.org/10.1007/s11888-020-00456-1

Parks, T., Barrett, L., & Jones, N. (2015). Invasive streptococcal disease: A review for clinicians. *British Medical Bulletin*, *115*(1), 77–89. https://doi.org/10.1093/bmb/ldv027

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Pinto, M. L., Rios, E., Durães, C., Ribeiro, R., Machado, J. C., Mantovani, A., Barbosa, M. A., Carneiro, F., & Oliveira, M. J. (2019). The Two Faces of Tumor-Associated Macrophages and Their Clinical Significance in Colorectal Cancer. *Frontiers in Immunology*, *10*. https://doi.org/10.3389/fimmu.2019.01875

Potter, J. D. (1999). Colorectal Cancer: Molecules and Populations. *JNCI: Journal of the National Cancer Institute*, *91*(11), 916–932. https://doi.org/10.1093/jnci/91.11.916

Purcell, R. V., Pearson, J., Aitchison, A., Dixon, L., Frizelle, F. A., & Keenan, J. I. (2017). Colonization with enterotoxigenic Bacteroides fragilis is associated with early-stage colorectal neoplasia. *PloS One*, *12*(2), e0171602. https://doi.org/10.1371/journal.pone.0171602

Purcell, R. V., Schmeier, S., Lau, Y. C., Pearson, J. F., & Frizelle, F. A. (2019). Molecular subtyping improves prognostication of Stage 2 colorectal cancer. *BMC Cancer*, *19*(1), 1155. https://doi.org/10.1186/s12885-019-6327-4

Purcell, R. V., Visnovska, M., Biggs, P. J., Schmeier, S., & Frizelle, F. A. (2017). Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-11237-6

R Core Team. (2019). *R: A language and environment for statistical computing*. R: A Language and Environment for Statistical Computing. https://www.r-project.org/

Reddy, B. S., Weisburger, J. H., Narisawa, T., & Wynder, E. L. (1974). Colon Carcinogenesis in Germ-free Rats with 1,2-Dimethylhydrazine and N-Methyl-N′-nitro-N-nitrosoguanidine. *Cancer Research*, *34*(9), 2368–2372.

Reeves, E., & James, E. (2017). Antigen processing and immune regulation in the response to tumours. *Immunology*, *150*(1), 16–24. https://doi.org/10.1111/imm.12675

Regal, J. F., Dornfeld, K. J., & Fleming, S. D. (2016). Radiotherapy: Killing with complement. *Annals of Translational Medicine*, *4*(5), 94. https://doi.org/10.21037/atm.2015.12.46

Reinecke, F., & Steinbüchel, A. (2009). Ralstonia eutropha strain H16 as model organism for PHA metabolism and for biotechnological production of technically interesting biopolymers. *Journal of Molecular Microbiology and Biotechnology*, *16*(1–2), 91–108. https://doi.org/10.1159/000142897

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rodriguez-Salas, N., Dominguez, G., Barderas, R., Mendiola, M., García-Albéniz, X., Maurel, J., & Batlle, J. F. (2017). Clinical relevance of colorectal cancer molecular subtypes. *Critical Reviews in Oncology/Hematology*, *109*, 9–19. https://doi.org/10.1016/j.critrevonc.2016.11.007

Roepman, P., Schlicker, A., Tabernero, J., Majewski, I., Tian, S., Moreno, V., Snel, M. H., Chresta, C. M., Rosenberg, R., Nitsche, U., Macarulla, T., Capella, G., Salazar, R., Orphanides, G., Wessels, L. F. A., Bernards, R., & Simon, I. M. (2014). Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *International Journal of Cancer*, *134*(3), 552–562. https://doi.org/10.1002/ijc.28387

Routy, B., Chatelier, E. L., Derosa, L., Duong, C. P. M., Alou, M. T., Daillère, R., Fluckiger, A., Messaoudene, M., Rauber, C., Roberti, M. P., Fidelle, M., Flament, C., Poirier-Colame, V., Opolon, P., Klein, C., Iribarren, K., Mondragón, L., Jacquelot, N., Qu, B., … Zitvogel, L. (2018). Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. *Science*, *359*(6371), 91–97. https://doi.org/10.1126/science.aan3706

Rubinstein, M. R., Wang, X., Liu, W., Hao, Y., Cai, G., & Han, Y. W. (2013). Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host & Microbe*, *14*(2), 195–206. https://doi.org/10.1016/j.chom.2013.07.012

Russel, J., Thorsen, J., Brejnrod, A. D., Bisgaard, H., Sørensen, S. J., & Burmølle, M. (2018). DAtest: A framework for choosing differential abundance or expression method. *BioRxiv*, 241802. https://doi.org/10.1101/241802

Ryan, J. E., Warrier, S. K., Lynch, A. C., Ramsay, R. G., Phillips, W. A., & Heriot, A. G. (2016). Predicting pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: A systematic review. *Colorectal Disease*, *18*(3), 234–246. https://doi.org/10.1111/codi.13207

Saad El Din, K., Loree, J. M., Sayre, E. C., Gill, S., Brown, C. J., Dau, H., & De Vera, M. A. (2020). Trends in the epidemiology of young-onset colorectal cancer: A worldwide systematic review. *BMC Cancer*, *20*(1), 288. https://doi.org/10.1186/s12885-020-06766-9

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., Ostos, L. C. G., Lannon, W. A., Grotzinger, C., Del Rio, M., Lhermitte, B., Olshen, A. B., Wiedenmann, B., Cantley, L. C., Gray, J. W., & Hanahan, D. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine*, *19*(5), 619–625. https://doi.org/10.1038/nm.3175

Salem, M. E., Weinberg, B. A., Xiu, J., El-Deiry, W. S., Hwang, J. J., Gatalica, Z., Philip, P. A., Shields, A. F., Lenz, H.-J., & Marshall, J. L. (2017). Comparative molecular analyses of left-sided colon, right-sided colon, and rectal cancers. *Oncotarget*, *8*(49), 86356–86368. https://doi.org/10.18632/oncotarget.21169

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*(1), 87. https://doi.org/10.1186/s12915-014-0087-z

Sánchez-Rovira, P., Jimenez, E., Carracedo, J., Barneto, I. C., Ramirez, R., & Aranda, E. (1998). Serum levels of intercellular adhesion molecule 1 (ICAM-1) in patients with colorectal cancer: Inhibitory effect on cytotoxicity. *European Journal of Cancer*, *34*(3), 394–398. https://doi.org/10.1016/S0959-8049(97)10033-8

Saus, E., Iraola-Guzmán, S., Willis, J. R., Brunet-Vega, A., & Gabaldón, T. (2019). Microbiome and colorectal cancer: Roles in carcinogenesis and clinical potential. *Molecular Aspects of Medicine*. https://doi.org/10.1016/j.mam.2019.05.001

Schellerer, V. S., Langheinrich, M. C., Zver, V., Grützmann, R., Stürzl, M., Gefeller, O., Naschberger, E., & Merkel, S. (2019). Soluble intercellular adhesion molecule-1 is a prognostic marker in colorectal carcinoma. *International Journal of Colorectal Disease*, *34*(2), 309–317. https://doi.org/10.1007/s00384-018-3198-0

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, *9*(8), 811–814. https://doi.org/10.1038/nmeth.2066

Sharma, A. K., Gupta, A., Kumar, S., Dhakan, D. B., & Sharma, V. K. (2015). Woods: A fast and accurate functional annotator and classifier of genomic and metagenomic sequences. *Genomics*, *106*(1), 1–6. https://doi.org/10.1016/j.ygeno.2015.04.001

Shen, W., & Ren, H. (2021). TaxonKit: A practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and Genomics*, *48*(9), 844–850. https://doi.org/10.1016/j.jgg.2021.03.006

Shi, W., Shen, L., Zou, W., Wang, J., Yang, J., Wang, Y., Liu, B., Xie, L., Zhu, J., & Zhang, Z. (2020). The Gut Microbiome Is Associated With Therapeutic Responses and Toxicities of Neoadjuvant Chemoradiotherapy in Rectal Cancer Patients—A Pilot Study. *Frontiers in Cellular and Infection Microbiology*, *10*. https://doi.org/10.3389/fcimb.2020.562463

Silva, G. G. Z., Green, K. T., Dutilh, B. E., & Edwards, R. A. (2016). SUPER-FOCUS: A tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics (Oxford, England)*, *32*(3), 354–361. https://doi.org/10.1093/bioinformatics/btv584

Singh, M. P., Rai, S., Pandey, A., Singh, N. K., & Srivastava, S. (2019). Molecular subtypes of colorectal cancer: An emerging therapeutic opportunity for personalized medicine. *Genes & Diseases*. https://doi.org/10.1016/j.gendis.2019.10.013

Soda, K. (2011). The mechanisms by which polyamines accelerate tumor spread. *Journal of Experimental & Clinical Cancer Research*, *30*(1), 95. https://doi.org/10.1186/1756-9966-30-95

Song, F., Yi, Y., Li, C., Hu, Y., Wang, J., Smith, D. E., & Jiang, H. (2018). Regulation and biological role of the peptide/histidine transporter SLC15A3 in Toll-like receptor-mediated inflammatory responses in macrophage. *Cell Death & Disease*, *9*(7), 1–15. https://doi.org/10.1038/s41419-018-0809-1

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sulit, A. K., Kolisnik, T., Frizelle, F. A., Purcell, R., & Schmeier, S. (2021a). *MetaFunc Databases: Kaiju database* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5602178

Sulit, A. K., Kolisnik, T., Frizelle, F. A., Purcell, R., & Schmeier, S. (2021b). *MetaFunc Databases: Nr-go database* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5602157

Sulit, A. K. L., Kolisnik, T., Frizelle, F. A., Purcell, R., & Schmeier, S. (2020). *MetaFunc: Taxonomic and Functional Analyses of High Throughput Sequencing for Microbiomes* [Preprint]. Bioinformatics. https://doi.org/10.1101/2020.09.02.271098

Surace, L., Lysenko, V., Fontana, A. O., Cecconi, V., Janssen, H., Bicvic, A., Okoniewski, M., Pruschy, M., Dummer, R., Neefjes, J., Knuth, A., Gupta, A., & van den Broek, M. (2015). Complement Is a Central Mediator of Radiotherapy-Induced Tumor-Specific Immunity and Clinical Response. *Immunity*, *42*(4), 767–777. https://doi.org/10.1016/j.immuni.2015.03.009

Tachimori, A., Yamada, N., Sakate, Y., Yashiro, M., Maeda, K., Ohira, M., Nishino, H., & Hirakawa, K. (2005). Up regulation of ICAM-1 gene expression inhibits tumour growth and liver metastasis in colorectal carcinoma. *European Journal of Cancer (Oxford, England: 1990)*, *41*(12), 1802–1810. https://doi.org/10.1016/j.ejca.2005.04.036

Tamas, K., Walenkamp, A. M. E., de Vries, E. G. E., van Vugt, M. A. T. M., Beets-Tan, R. G., van Etten, B., de Groot, D. J. A., & Hospers, G. A. P. (2015). Rectal and colon cancer: Not just a different anatomic site. *Cancer Treatment Reviews*, *41*(8), 671–679. https://doi.org/10.1016/j.ctrv.2015.06.007

Tan, H., Zhao, J., Zhang, H., Zhai, Q., & Chen, W. (2019). Novel strains of Bacteroides fragilis and Bacteroides ovatus alleviate the LPS-induced inflammation in mice. *Applied Microbiology and Biotechnology*, *103*(5), 2353–2365. https://doi.org/10.1007/s00253-019-09617-1

Tanaka, A., Zhou, Y., Ogawa, M., Shia, J., Klimstra, D. S., Wang, J. Y., & Roehrl, M. H. (2020). STAT1 as a potential prognosis marker for poor outcomes of early stage colorectal cancer with microsatellite instability. *PLOS ONE*, *15*(4), e0229252. https://doi.org/10.1371/journal.pone.0229252

Ternes, D., Karta, J., Tsenkova, M., Wilmes, P., Haan, S., & Letellier, E. (2020). Microbiome in Colorectal Cancer: How to Get from Meta-omics to Mechanism? *Trends in Microbiology*. https://doi.org/10.1016/j.tim.2020.01.001

Terzi, C., Bingul, M., Arslan, N. C., Ozturk, E., Canda, A. E., Isik, O., Yilmazlar, T., Obuz, F., Gorken, I. B., Kurt, M., Unlu, M., Ugras, N., Kanat, O., & Oztop, I. (2020). Randomized controlled trial of 8 weeks' vs 12 weeks' interval between neoadjuvant chemoradiotherapy and surgery for locally advanced rectal cancer. *Colorectal Disease*, *22*(3), 279–288. https://doi.org/10.1111/codi.14867

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., Gandini, S., Serrano, D., Tarallo, S., Francavilla, A., Gallo, G., Trompetto, M., Ferrero, G., Mizutani, S., Shiroma, H., … Segata, N. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine*, *25*(4), 667–678. https://doi.org/10.1038/s41591-019-0405-7

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., Sørensen, S., Bisgaard, H., & Waage, J. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome*, *4*(1). https://doi.org/10.1186/s40168-016-0208-8

Tjalsma, H., Boleij, A., Marchesi, J. R., & Dutilh, B. E. (2012). A bacterial driver–passenger model for colorectal cancer: Beyond the usual suspects. *Nature Reviews Microbiology*, *10*(8), 575–582. https://doi.org/10.1038/nrmicro2819

Tofalo, R., Cocchi, S., & Suzzi, G. (2019). Polyamines and Gut Microbiota. *Frontiers in Nutrition*, *6*. https://doi.org/10.3389/fnut.2019.00016

Tominaga, K., Yoshimoto, T., Torigoe, K., Kurimoto, M., Matsui, K., Hada, T., Okamura, H., & Nakanishi, K. (2000). IL-12 synergizes with IL-18 or IL-1β for IFN-γ production from human T cells. *International Immunology*, *12*(2), 151–160. https://doi.org/10.1093/intimm/12.2.151

Touchefeu, Y., Montassier, E., Nieman, K., Gastinne, T., Potel, G., Bruley des Varannes, S., Le Vacon, F., & de La Cochetière, M. F. (2014). Systematic review: The role of the gut microbiota in chemotherapy- or radiation-induced gastrointestinal mucositis - current evidence and potential clinical applications. *Alimentary Pharmacology & Therapeutics*, *40*(5), 409–421. https://doi.org/10.1111/apt.12878

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., & Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, *12*(10), 902–903. https://doi.org/10.1038/nmeth.3589

The UniProt Consortium. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099

Uribe-Herranz, M., Rafail, S., Beghi, S., Gil-de-Gómez, L., Verginadis, I., Bittinger, K., Pustylnikov, S., Pierini, S., Perales-Linares, R., Blair, I. A., Mesaros, C. A., Snyder, N. W., Bushman, F., Koumenis, C., & Facciabene, A. (2020). Gut microbiota modulate dendritic cell antigen presentation and radiotherapy-induced antitumor immune response. *The Journal of Clinical Investigation*, *130*(1), 466–479. https://doi.org/10.1172/JCI124332

Vanhoecke, B. W., De Ryck, T. R., De boel, K., Wiles, S., Boterberg, T., Van de Wiele, T., & Swift, S. (2016). Low-dose irradiation affects the functional behavior of oral microbiota in the context of mucositis. *Experimental Biology and Medicine*, *241*(1), 60–70. https://doi.org/10.1177/1535370215595467

Vétizou, M., Pitt, J. M., Daillère, R., Lepage, P., Waldschmitt, N., Flament, C., Rusakiewicz, S., Routy, B., Roberti, M. P., Duong, C. P. M., Poirier-Colame, V., Roux, A., Becharef, S., Formenti, S., Golden, E., Cording, S., Eberl, G., Schlitzer, A., Ginhoux, F., … Zitvogel, L. (2015). Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science (New York, N.Y.)*, *350*(6264), 1079–1084. https://doi.org/10.1126/science.aad1329

Visnovska, T., Biggs, P. J., Schmeier, S., Frizelle, F. A., & Purcell, R. V. (2019). Metagenomics and transcriptomics data from human colorectal cancer. *Scientific Data*, *6*(1), 116. https://doi.org/10.1038/s41597-019-0117-3

Vyas, S., & Chang, P. (2014). New PARP targets for cancer therapy. *Nature Reviews. Cancer*, *14*(7), 502–509. https://doi.org/10.1038/nrc3748

Waldner, M., Schimanski, C. C., & Neurath, M. F. (2006). Colon cancer and the immune system: The role of tumor invading T cells. *World Journal of Gastroenterology : WJG*, *12*(45), 7233–7238. https://doi.org/10.3748/wjg.v12.i45.7233

Wang, Q., Wang, K., Wu, W., Giannoulatou, E., Ho, J. W. K., & Li, L. (2019). Host and microbiome multi-omics integration: Applications and methodologies. *Biophysical Reviews*, *11*(1), 55–65. https://doi.org/10.1007/s12551-018-0491-7

Wang, W., Kandimalla, R., Huang, H., Zhu, L., Li, Y., Gao, F., Goel, A., & Wang, X. (2019). Molecular subtyping of colorectal cancer: Recent progress, new challenges and emerging opportunities. *Seminars in Cancer Biology*, *55*, 37–52. https://doi.org/10.1016/j.semcancer.2018.05.002

Wang, Y., Deng, W., Li, N., Neri, S., Sharma, A., Jiang, W., & Lin, S. H. (2018). Combining Immunotherapy and Radiotherapy for Cancer Treatment: Current Challenges and Future Directions. *Frontiers in Pharmacology*, *0*. https://doi.org/10.3389/fphar.2018.00185

Wang, Y., Sun, D., Song, F., Hu, Y., Smith, D. E., & Jiang, H. (2014). Expression and Regulation of the Proton-Coupled Oligopeptide Transporter PhT2 by LPS in Macrophages and Mouse Spleen. *Molecular Pharmaceutics*, *11*(6), 1880–1888. https://doi.org/10.1021/mp500014r

Wei, H., Chen, L., Lian, G., Yang, J., Li, F., Zou, Y., Lu, F., & Yin, Y. (2018). Antitumor mechanisms of bifidobacteria. *Oncology Letters*, *16*(1), 3–8. https://doi.org/10.3892/ol.2018.8692

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1). https://doi.org/10.1186/s40168-017-0237-y

West, N. R., McCuaig, S., Franchini, F., & Powrie, F. (2015). Emerging cytokine networks in colorectal cancer. *Nature Reviews Immunology*, *15*(10), 615–629. https://doi.org/10.1038/nri3896

Westreich, S. T., Treiber, M. L., Mills, D. A., Korf, I., & Lemay, D. G. (2018). SAMSA2: A standalone metatranscriptome analysis pipeline. *BMC Bioinformatics*, *19*(1), 175. https://doi.org/10.1186/s12859-018-2189-z

Whitmore, S. E., & Lamont, R. J. (2014). Oral Bacteria and Cancer. *PLoS Pathogens*, *10*(3), e1003933. https://doi.org/10.1371/journal.ppat.1003933

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J. S., Voigt, A. Y., Palleja, A., Ponnudurai, R., Sunagawa, S., Coelho, L. P., Schrotz-King, P., Vogtmann, E., Habermann, N., Niméus, E., Thomas, A. M., Manghi, P., Gandini, S., … Zeller, G. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine*, *25*(4), 679–689. https://doi.org/10.1038/s41591-019-0406-6

Wong, S. H., & Yu, J. (2019). Gut microbiota in colorectal cancer: Mechanisms of action and clinical applications. *Nature Reviews Gastroenterology & Hepatology*, *16*(11), 690–704. https://doi.org/10.1038/s41575-019-0209-8

Wong, S. H., Zhao, L., Zhang, X., Nakatsu, G., Han, J., Xu, W., Xiao, X., Kwong, T. N. Y., Tsoi, H., Wu, W. K. K., Zeng, B., Chan, F. K. L., Sung, J. J. Y., Wei, H., & Yu, J. (2017). Gavage of Fecal Samples From Patients With Colorectal Cancer Promotes Intestinal Carcinogenesis in Germ-Free and Conventional Mice. *Gastroenterology*, *153*(6), 1621-1633.e6. https://doi.org/10.1053/j.gastro.2017.08.022

Wu, S., Rhee, K.-J., Albesiano, E., Rabizadeh, S., Xinqun Wu, Hung-Rong Yen, Huso, D. L., Brancati, F. L., Wick, E., McAllister, F., Housseau, F., Pardoll, D. M., & Sears, C. L. (2009). A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nature Medicine*, *15*(9), 1016–1022. https://doi.org/10.1038/nm.2015

Xia, X., Wu, W. K. K., Wong, S. H., Liu, D., Kwong, T. N. Y., Nakatsu, G., Yan, P. S., Chuang, Y.-M., Chan, M. W.-Y., Coker, O. O., Chen, Z., Yeoh, Y. K., Zhao, L., Wang, X., Cheng, W. Y., Chan, M. T. V., Chan, P. K. S., Sung, J. J. Y., Wang, M. H., & Yu, J. (2020). Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer. *Microbiome*, *8*(1), 108. https://doi.org/10.1186/s40168-020-00847-4

Xia, Y., Sun, J., & Chen, D.-G. (2018). Bioinformatic Analysis of Microbiome Data. In Y. Xia, J. Sun, & D.-G. Chen (Eds.), *Statistical Analysis of Microbiome Data with R* (pp. 1–27). Springer Singapore. https://doi.org/10.1007/978-981-13-1534-3_1

Xie, Y.-H., Chen, Y.-X., & Fang, J.-Y. (2020). Comprehensive review of targeted therapy for colorectal cancer. *Signal Transduction and Targeted Therapy*, *5*(1), 1–30. https://doi.org/10.1038/s41392-020-0116-z

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, *178*(4), 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Ye, X., Wang, R., Bhattacharya, R., Boulbes, D. R., Fan, F., Xia, L., Adoni, H., Ajami, N. J., Wong, M. C., Smith, D. P., Petrosino, J. F., Venable, S., Qiao, W., Baladandayuthapani, V., Maru, D., & Ellis, L. M. (2017). *Fusobacterium Nucleatum* Subspecies *Animalis* Influences Proinflammatory Cytokine Expression and Monocyte Activation in Human Colorectal Tumors. *Cancer Prevention Research*, *10*(7), 398–409. https://doi.org/10.1158/1940-6207.CAPR-16-0178

Yoshida, N., Emoto, T., Yamashita, T., Watanabe, H., Hayashi, T., Tabata, T., Hoshi, N., Hatano, N., Ozawa, G., Sasaki, N., Mizoguchi, T., Amin, H. Z., Hirota, Y., Ogawa, W., Yamada, T., & Hirata, K. (2018). Bacteroides vulgatus and Bacteroides dorei Reduce Gut Microbial Lipopolysaccharide Production and Inhibit Atherosclerosis. *Circulation*, *138*(22), 2486–2498. https://doi.org/10.1161/CIRCULATIONAHA.118.033714

Yu, A. I., Zhao, L., Eaton, K. A., Ho, S., Chen, J., Poe, S., Becker, J., Gonzalez, A., McKinstry, D., Hasso, M., Mendoza-Castrejon, J., Whitfield, J., Koumpouras, C., Schloss, P. D., Martens, E. C., & Chen, G. Y. (2020). Gut Microbiota Modulate CD8 T Cell Responses to Influence Colitis-Associated Tumorigenesis. *Cell Reports*, *31*(1), 107471. https://doi.org/10.1016/j.celrep.2020.03.035

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS : A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Yu, L. C.-H., Wei, S.-C., & Ni, Y.-H. (2018). Impact of microbiota in colorectal carcinogenesis: Lessons from experimental models. *Intestinal Research*, *16*(3), 346–357. https://doi.org/10.5217/ir.2018.16.3.346

Zaborowski, A. M., Winter, D. C., & Lynch, L. (2021). The therapeutic and prognostic implications of immunobiology in colorectal cancer: A review. *British Journal of Cancer*, 1–9. https://doi.org/10.1038/s41416-021-01475-x

Zhu, J., Petit, P.-F., & Van den Eynde, B. J. (2019). Apoptosis of tumor-infiltrating T lymphocytes: A new immune checkpoint mechanism. *Cancer Immunology, Immunotherapy*, *68*(5), 835–847. https://doi.org/10.1007/s00262-018-2269-y

Zou, S., Fang, L., & Lee, M.-H. (2018). Dysbiosis of gut microbiota in promoting the development of colorectal cancer. *Gastroenterology Report*, *6*(1), 1–12. https://doi.org/10.1093/gastro/gox031