Prognostic Molecular Markers of Response to Radiotherapy in Rectal Cancer

William S. Taylor

A thesis submitted for the degree of

Master of Science

at the University of Otago, Dunedin,

New Zealand

December 20th 2020

ABSTRACT

Colorectal cancer (CRC) is the second most deadly cancer globally, and 30% of these cancers occur in the rectum. The primary treatment for CRC is surgery, often radiotherapy with adjuvant chemotherapy is used before or following surgical resection.

Treatment carries with it a high cost and side effect burden while response rates remain unpredictable. Approximately 20% of patients have total tumour regression post chemoradiotherapy; however, most patients receive only partial or no benefit from treatment. The ability to predict which patients would benefit from standard treatment and those who should be directed to an alternative treatment or an accelerated pathway to surgery would potentially avoid lengthy and costly treatments that may only cause side effects for patients, improving survival rates and quality of life.

In this study, the microbiome, immune cells and patient gene expression were evaluated for their use as predictive biomarkers for response to chemoradiotherapy in rectal cancer patients. Tumour and adjacent normal tissue biopsies were taken before treatment and had DNA and RNA extracted and sequenced.

First, the methodology for analysing microbiomes via shotgun sequencing data was evaluated and improved, increasing taxonomic assignment accuracy by 11% and potentially decreasing analysis time more than nine-fold. Secondly, the sequencing technologies, Oxford Nanopore, 16S rRNA and RNA-sequencing, were evaluated for their ability to assess the microbiome. The results demonstrated that platforms had concordance with one another; however, this was reduced at the species level.

Third, microbial transcription was used to assess rectal cancer microbiomes, correlating them with response rates. The results showed that microbial diversity did not contribute to radiotherapy response, but that individual microbes may influence response. It was hypothesised that species such as *Hungatella hathewayi*, *Fusobacterium nucleatum*, *Butyricimonas faecalis*, *Alistipes finegoldii*, *Bacteroides thetaiotaomicron*, and *B. fragilis* may contribute to tumour regression by modulating metabolism and immune responses.

Third, the abundance of infiltrating immune cells was predicted using RNA-Sequencing data. Analysis indicated that the abundance of M1 macrophages and resting mast cells were correlated with response, while microbial transcription was correlated with the abundance of allergic and anti-tumour effector cells, as well as antigen-presenting cells. It was hypothesised that the microbiome might modulate anti-tumour immune responses directly, and indirectly by altering the tumour microenvironment. Microbes may help maintain a population of anti-tumour effector and antigen-presenting cells for tumour-antigen presentation during tumour cell death and neo-antigen uptake, which may be otherwise exhausted by targeting aspects of the inflammatory tumour microenvironment (i.e., lipid phagocytosis, anti-bacterial and allergic responses).

Lastly, machine-learning was employed to establish a panel of molecular biomarkers predictive of response, including microbial transcription, immune cells and gene expression. The final model demonstrated the ability to predict response with a 7% overall error rate, and that predicting response relied mostly on normal and tumour tissue gene expression, and tumour infiltrating immune cells.

This study provides a panel of prognostic biomarkers which could be utilised to predict patient response. Additionally, it provides evidence for microbial-immune interactions that could be manipulated to enhance treatment and increase response rates.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Colorectal cancer	CRC
Rectal cancer	RC
Radiotherapy	RT
Chemoradiotherapy	CRT
Human papilloma virus	HPV
Estrogen receptor	ER
Hereditary nonpolyposis CRC	HNPCC
Mismatch repair	MMR
Familial Adenomatous Polyposis	FAP
Faecal occult blood test	FOBT
Faecal immunochemical test	FIT
Gastrointestinal	GI
Computerised tomography	CT
Magnetic resonance imaging	MRI
Positron emission tomography	PET
Tumour-lymph node-metastasis	TNM
Mesorectal fascia positive	MRF+
Irritable bowel syndrome	IBS
Inflammatory bowel disease	IBD
Short-chain fatty acids	SCFA
Gastrointestinal	GI
Bacteroides fragilis toxin	bft
T cell transcription factor	TCF
Regulatory T cells	Tregs
Horizontal gene transfer	HGT
Neoadjuvant chemoradiotherapy	nCRT
Tumour regression grade	TRG
American Joint Committee on Cancer	AJCC
5-fluorouracil	5-FU
Thymidylate synthase	TS
Fluro-deoxyuridine monophosphate	FdUMP
Flurouradine triphosphate	FUTP
Capecitabine and cisplatin	CAPOX
leucovorin, 5-FU and oxaliplatin	FOLFOX
leucovorin, 5-FU, irinotecan	FOLFIRI
Pathological complete response	pCR
Linear energy transfer	LET
Reactive oxygen species	ROS
Carcinoembryonic antigen	CEA
Total mesorectal excision	TME
Consensus molecular subtype	CMS
Disease-free survival	DFS
Dendritic cells	DCs
Natural killer	NK
Next-Generation Sequencing	NGS
Bacterial artificial chromosome	BAC
Oxford Nanopore Technologies	ONT
Polymerase chain reaction	PCR
Homo-polymer compression	HPC
Messenger RNA	mRNA
16S ribosomal RNA	16S rRNA
Gene set expression analysis	GSEA
Gene ontology	GO
Kyoto Encyclopedia of Genes and Genomes	KEGG

Analysis of variance	ANOVA
Burrows-Wheeler Alignment	BWA
Operational taxonomic Unit	OTU
Internal transcribed spacer	ITS
Greengenes	GG
Long-course chemoradiotherapy	LCCRT
Short-course chemoradiotherapy	SCCRT
RNA-sequencing	RNA-Seq
Sequence read archive	SRA
Base pairs	bp
Bacterial database	Bac_DB
Database of all relevant genomes	All_DB
Christchurch	CHCH
Peter MacCallum	PM
Gray unit	Gy
Rectal Cancer ONT data	Rec-ONT
Coloretal cancer RNA-seq data	CRC-RNA
Principal components of analysis	PCoA
Male	M
Female	F
Not available	NA
False discovery rate	FDR
Data Integration Analysis for Biomarker discovery using a Latent cOmponents	DIABLO
Synthetic	Synth
Confidence interval	CI
non-metric multi-dimensional scaling	NMDS
Quality control	QC
16S rRNA	16S
Bray–Curtis	BC
Jensen–Shannon divergence	JSD
Jaccard index	JI
Standard deviation	SD
Tumour associated macrophages	TAMs
Lipopolysaccharide	LPS
Support vector regression	SVR
Linear model	lm
Number	*n*
T regulatory cells	Tregs
Bonferroni–Hochberg	BH
Follicular T helper	Fth
Activated dendritic cells	aDCs
Resting dendritic cells	rDCs
Antigen presenting cells	APCs
sparse partial least squares discriminant analysis	sPLS-DA
Centred log-ratio transform	CLR
Sparse partial least squares	sPLS
Least Absolute Shrinkage and Selection Operator	LASSO
Leave-one-out	loo
Normal tissue	N
Tumour tissue	T
Error rate	ER
Balanced error rate	BER
Receiver operating characteristic	ROC
Area under the curve	AUC
Long non-coding RNA	lncRNA

1 CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Colorectal cancer (CRC) has the third-highest incidence, and this frequency leads to the second-highest mortality of any cancer. Thirty percent of cases occur in the rectum [1]. In 2018 alone, 1.8 million new cases and 881,000 deaths were recorded globally, New Zealand and Australia had the second-highest rates of colon and rectal cancers (RC) [2]. There is increasing evidence that the microbiome plays a role in the development of CRC [3]. *Fusobacterium nucleatum* [4], enterotoxigenic *Bacteroides fragilis* [5], strains of *Escherichia coli* [6], and species from the oral cavity have been reported to produce potentially genotoxic compounds [7].

RC is considered distinct from colon cancer and requires a different treatment strategy [1]. The greatest contributor to RC survival rates is early detection, while the disease is still localised [8]. Early detection allows for less invasive surgical resection, while higher stage tumours require more radical interventions, such as anterior resection and total mesorectal excision (TME), which involve defunctioning of the bowel before surgery, and often result in a temporary or permanent colostomy [1]. Although early detection is increasingly common in countries with nationalised bowel screening programmes, late-stage detection still frequently occurs.

1.1.1 Epidemiology

New Zealand and Australia have the second-highest RC incidence, at a rate of 15.6 and 8.6 per 100,000 in males and females, respectively, and the third-highest mortality, globally [2]. RC accounted for 704,376 (3.9%) new cases and 310,394 (3.2%) cancer-related deaths in 2018, globally. Incidence and mortality rate is related to regional development, with increases in both incidence and mortality in Eastern European nations, China and Latin America. There is increasing incidence but reducing mortality in Canada, the U.K., Denmark and Singapore, while both mortality and incidence are reducing in the U.S., Japan and France [2]. These trends reflect increased survival as nations adopt screening programmes and improved treatment [9].

1.1.2 Risk factors

The risk factors for RC are poorly understood and less well studied than colon cancers, partly due to studies combining patients with rectal and colon cancers into generalised CRC cohorts [10]. As sporadic disease accounts for the majority of reported cases (~75%) and is characterised by having no known underlying genetic predisposition [11], the involvement of lifestyle factors (e.g., smoking, alcohol/drug use), diet (e.g., excessive processed and red meat and inadequate fibre intake), exercise and the microbiome [12] may partly explain underlying causes. Human papillomavirus (HPV) has been implicated in CRC risk [13]; however, this is not a universal finding, with some studies suggesting HPV is not involved in RC [14], or CRC carcinogenesis [15]. A meta-analysis of 2630 adenocarcinomas showed CRC HPV prevalence at 11.2%, with different regions having substantive differences, and adenomas having a lower prevalence (5.1%) [16]. Further investigation into the role of HPV in CRC is required, as it may depend on the viral strain, geography or an individual’s genetics.

In a comparative study of colon and rectal cancers, only sex and age correlated with RC [17]; however, there is a developing trend of increasing RC incidence in younger patients [18]. The incidence per 100,000 is 23.6 for males and 16.3 for females globally, and 41.7 and 32.1 in Australia/New Zealand, respectively [19]. There are differences in incidence between pre- and post-menopausal women, with individuals that have hormone replacement therapy post-menopause have a lower incidence of CRC, indicating that oestrogen may play a role in CRC aetiology [20]. However, the menopause-related risk may be adipose dependant, as lean women have been reported to have higher risk after menopause, and obese patients have a higher risk before menopause [20, 21]. The frequent occurrence of *KRAS* and *BRAF* mutations in CRC, as *KRAS* expression can be mediated by estrogen receptor (ER) to allow escape from cellular senescence [22] and *BRAF* overexpression often co-occurs with ER inactivation, and is associated with worse prognosis [23]. This evidence suggests a role for oestrogen playing a role in carcinogenesis and CRC treatment outcomes [24-26]. Environmental exposure to oestrogen and pseudo-oestrogens may partly explain age and sex associations with CRC risk [27-29].

There is evidence for fewer factors protective of RC than for colon cancers [30]. Red meat and processed food consumption are thought to increase colon cancer incidence but have weaker associations with RC [31]. The protective effects of exercise have been shown in CRC, but are still debated in RC [32, 33], with meta-analyses providing conflicting results [34, 35]. The differences in the effect of exercise may be due to the underlying cause of disease. A mouse model was used to show that tumour occurrence could be reduced with exercise, but only when occurring before carcinogen exposure [36]. Due to the rectum's role as a storage organ, it is subject to continuous potential carcinogen exposure, which may make pre-empting carcinogen exposure with exercise less effective [17].

1.2 CRCs

CRCs can be separated into hereditary or sporadic. Hereditary CRCs are a result of one or more mutations in multiple inherited genes. Well defined inherited syndromes are responsible for 2%–5% of CRCs, 25%–30% of cases have a family history of CRC, indicating a yet to be discovered link, and ~75% are sporadic (i.e., no identified inherited risk). Details of the incidence and conferred risk of hereditary CRC are summarised in Table 1.1 [37-43]. Hereditary conditions are conferred by autosomal dominant mutations, making familial history and genetic testing useful for early detection and potential risk for developing CRC [44].

Table 1.1. Hereditary CRC syndromes

Disease	**Clinical features**	**Gene involved**	**Lifetime CRC Risk**	**Percent of CRC Incidence**

Familial adenomatous polyposis (FAP)	Hundreds to thousands of adenomatous polyps in colon and rectum	*APC*	100%	1%
Hereditary non-polyposis colorectal cancer (HNPCC)	Family history of CRC and other cancers	*hMLH1, hMSH2, hMSH6, hMSH3, hPMS2*	37.6%–78.9%	2%–3%
Peutz–Jegher's syndrome	Small bowel hamartomatous polyps in and peri-oral pigmentation	*LKB1, STK11*	39%–57%	1%
Juvenile polyposis	Hamartomatous polyps in stomach and large bowel	*SMAD4, PTEN, BMPR1A*	39%–68%	1%

Details from [45], [46] and [47].

Sporadic cancers are caused by somatic mutations occurring due to DNA damage from carcinogens and environmental factors, such as microbes capable of sulphate and nitrate reduction, genotoxins, smoking and sedentary lifestyles [48]. When a tumour develops below the sigmoid colon, it is termed RC, with adenocarcinomas being the most common form of sporadic cancers in the large intestine, originating from a gland-type cell.

1.2.1 Diagnosis

CRCs typically present with bloody stools and rectal bleeding, changes in bowel habits and abdominal/anal pain, which may prompt further investigation by a general practitioner, such as physical examination, faecal occult blood test (FOBT) or faecal immunochemical test (FIT) [49]. Approximately 70% of RCs are identified with a physical *per rectum* examination; however, additional testing is required such as sigmoidoscopy or colonoscopy, biopsies and radiological imaging to confirm the diagnosis [50]. Screening programmes have been developed for symptomatic individuals and those over 60 in New Zealand [51]. However, there is an increasing incidence in patients under 40 [18, 52], and many patients are asymptomatic [53, 54].

1.2.2 Histology and staging

Histological grading of CRC is generally based on the proportion of glandular differentiation, although using cell cluster and glandular structure during grading is also common [55]. Grading systems are still debated and updated regularly; however, there are variations in clinical interpretation and differences in institutional and national guidelines [52, 53].

Staging is performed via rectoscopy, trans-rectal ultrasound, and computer-aided tomography (CT) scan, while magnetic resonance imaging (MRI) is the most powerful imaging tool for staging, and for informing treatment and surgical options [56][50]. As the lower rectum's venous drainage bypasses the liver, lung nodules are more common than liver metastasis as a primary metastatic site, and chest CT or positronic electron emission (PET) scans are needed to detect suspected metastases. Accurate staging is crucial for discovering which patients are candidates for chemoradiotherapy before surgery [57]. CRC staging is performed using tumour stage (T), lymph node invasion (N) and metastatic status (M). T1 and T2 tumours are limited to the bowel wall, while T3 is characterised by infiltration into the mesorectal fat. T4 tumours have penetrated other structures, with T4a indicating invasion of the peritoneal reflection, while T4b indicates invasion into adjacent organs [51].

Five-year survival rates in the USA for CRC are 92% for stage I, 65% and 87% for stage IIA and B, respectively, and 90%, 72% and 53% for stage IIIA, B and C, respectively, while for metastatic and stage IV CRC five-year survival rates are 12% [19]. For RC, the five-year survival rates are for stage I: 88%; stage IIA/B: 81% and 50%, respectively; stage IIIA/B/C: 83%, 72% and 58%, respectively; and stage IV: 13% [19]. In NZ, colon cancer rates are as follows, stage I: 80%, stage II: 71%, stage III 63%–50% (depending on nodal status), stage IV: 6%, and for RC are 65% and 10%, if metastatic [58]. The Australian Institute of Health and Welfare does not provide a distinction between colon and rectal cancers [59]; however for CRC, for stages I–IV the five year survival rates are 89%, 74.5%, 44.5% and 12.6%, respectively [60].

1.2.3 The rectum

The rectum is the terminal point of the colon before the sphincter muscles. The rectum is in contact with luminal contents for more extended periods than other parts of the bowel; its primary role is temporary storage for faeces before excretion. Therefore it is subject to high levels of metabolites and waste products from digestion and from the microbiome [61], which have been associated with carcinogenesis [62]. The rectum is supplied with blood from the interior, middle and superior rectal arteries. It is drained by the superior, middle and inferior rectal veins from the internal and external haemorrhoidal plexus which bypass the liver [63, 64].

1.2.4 Prognostic markers for RC

Prognostic markers used for RC survival are tumour stage, location, morphology, differentiation, lymph-node involvement, vascular or mesorectal invasion, and metastases. Advanced stage, poorly differentiated [65], mucinous, and signet-ring cell cancers [66] are associated with poor prognosis, while a greater distance from the anal verge is a positive prognostic indicator for surgical outcomes [67].

1.3 The microbiome

A microbiome is the community of microorganisms living in a given environment [68]. The gut microbiome has the highest microbial abundance in the human body, with hundreds of species being present in individuals from a pool of potentially thousands across the human population. It contains more than 30 times the number of human genes and usually exists in a symbiotic relationship with the host [69]. Microbes can affect host health in many ways, from metabolic activities [70], to altering nutrient absorption [71], and immune functionality [72, 73]. The gut microbiome can be influenced by host genetics [74], diet [75], lifestyle [76] and medical interventions [77]. A pathological imbalance in the gut microbiome is called dysbiosis. Dysbiosis can have consequences for health, being associated with irritable bowel syndrome (IBS) [78], malnutrition [79], mental health [80] and cancer [81]. Microbes can produce a variety of beneficial metabolites in the gut [79, 82]. Of the most well studied are short-chain fatty acids (SCFA), such as butyrate, propionate and acetate, produced from metabolising dietary fibre [83]. Summary data for faecal levels of SCFAs can be found in Table 1.2. SCFAs can have immunomodulatory, neuronal, and microbial regulatory effects [84, 85], and butyrate can be used as an energy source for colonocytes [86, 87].

Table 1.2. Short-chain fatty acids in different populations

Study	Population and Diet	Acetate	Propionate	Butyrate
		(mg/g faeces)		
Fleming et al. [88]	US; self-selected, low and high fibre	2.75	1.22	1.64
		(mM faeces)		
Muir et al. [89]	Simulated Australian	56.1	15	18.4
	Simulated Chinese	39.9	12.8	12.2
Takahashi et al. [90]	Japanese; self-selected	36	22.3	17.5
	Japanese; controlled	45.2	23.6	19
		(mmol/kg faeces)		
Høverstad et al. [91]	Norwegian; self-selected	37.3	12.5	12.4

Table adapted from [92].

1.3.1 The microbiome and CRC

The majority of CRCs are thought to be sporadic and lifestyle and environmental factors (such as microbiomes), likely play a role in their initiation [11]. Bacterial species such as *Fusobacterium nucleatum*, *Bacteroides fragilis* and *Helicobacter pylori* have been shown to affect carcinogenesis [93-95].

1.4 Microbial carcinogenesis

Members and activities of the microbiome may not be sufficient for carcinogenesis alone, as the Knudson hypothesis suggests that an underlying predisposition may be required [96], such as gene mutations or a diet that can be metabolised into carcinogenic compounds [97-100]. The gastrointestinal tract acts as selective environments with differing shearing forces, oxygen levels, and nutrients that lead to differences in the microbiomes between proximal and distal CRCs [101]. Nutrient, metabolite and oxygen levels change throughout the gut and are spatially distributed, with oxygen levels being higher closer to the lumen wall [102]. Simple carbohydrates are taken up rapidly by the host and the microbiome, the latter of which can convert simple nutrients into complex metabolic products [103]. In this sense, the gut can be thought of as a complex Winogradsky column, with species cross-feeding on the preceding communities' metabolic products [104].

The tumour microenvironment is separate from the lumen and is generally hypoxic, due to unrestricted cell growth, which in turn forces anaerobic glycolysis and selects for anaerobic microbes [105, 106]. Facultative anaerobes may also maintain the hypoxic tumour environment, impacting treatment efficacy [107].

A significant increase in the diversity in tumours compared to controls was noted in a study of RC tumour microbiomes, and the samples were clustered into two enterotypes [108]. The majority of samples were those dominated by *Bacteroides* and *Dorea* genera, and another group had elevated amounts of *Pseudomonas* and *Brevundimonas* [108]. Higher levels of anaerobic bacteria were found in tissues, which is consistent with other CRC tumour microbiome studies [109], while ectosymbiont Parcubacteria and the Planctomycetes phyla, were identified as possible biomarkers for RC [108].

The differences in microbes in CRC tumours could also be due to each area of the intestine expressing different genes, and primary tumour location is known to impact treatment efficacy [110-112], with left-sided tumours having better survival rates than right-sided tumours [113]. The colonisation of tissues by microbes is reflective of the selective microenvironment, and therefore, could be used as a prognostic marker [93, 109].

1.4.1 Models of microbiome involvement in CRC

There are three primary models for how the microbiome can contribute to cancer: the driver-passenger model, the keystone species model and the oral origin model.

The ‘driver-passenger’ model suggested by Tjalsma et al. states that toxigenic species, known as drivers, are causative of the disease, providing the “first hit”. Those microbes that are selected for by the subsequent tumour microenvironment are termed passengers [114]. However, drivers in one context may be passengers in another, and the presence of passenger species may be indicative of later-stage cancers [115].

The keystone species model describes an organism which can alter its environment, allowing the colonisation of subsequent species which capitalise on the altered environment and cause carcinogenic changes in the host [116]. The keystone species model is similar to the driver-passenger model in that particular microbes produce alterations to the microbiome. However, it differs in that driver species may become displaced by passengers as the environment changes, whereas a keystone species might not be displaced.

The oral origin model suggests that oral species are potential initiators of carcinogenesis [117], e.g. *Fusobacterium nucleatum* and *Porphyromonas gingivalis, Parvimonas micra, Peptostreptococcus stomatis, Gemella morbillorum, Leptotrichia trevisanii* [118], *Selenomonas sputigena* [109], *Lachnospiracea intertie sedis* [119], *Treponema denticola* and *Tannerella forsythia* [93], originate from the oral cavity and have been implicated in carcinogenesis. These species may require an already dysbiotic environment to establish residence in the gut and generate a genotoxic effect, requiring biofilm generation that is prevented in a healthy gut environment [120, 121].

One model could explain carcinogenesis in one individual, while another may explain carcinogenesis in another. These models are not mutually exclusive, as an oral microbe may be a ‘driver’ or ‘keystone species’ when established in the gastrointestinal (GI) tract.

1.4.2 Mechanisms of microbiome associated CRC development

Microbial species can contribute to carcinogenesis by altering their environment, causing inflammation, increasing cancer risk. Besides, microbes can act directly on cells in the gut to cause carcinogenesis in various ways.

E-cadherin is a part of cellular junctions that maintain cellular adhesion and epithelial barrier functionality and can have immune-modulating effects [122]. It is a common cellular target for microbial pathogens, allowing bacteria to adhere and enter cells, and manipulate host cell signalling [123, 124].

Enterotoxigenic *Bacteroides fragilis* can produce the *Bacteroides fragilis* toxin (BFT), a metalloprotease that leads to cleavage E-cadherin [125, 126], leading to the accumulation of free β-catenin that acts as a co-activator of T cell factor transcription factor (TCF), and the transcription of genes in the Wnt signalling pathway, which are involved in cell growth and proliferation [127]. Additionally, *Fusobacterium nucleatum* produces FadA, an adhesion protein that also interacts with E-cadherin, modulating β-catenin signalling [128, 129]. Other microbes that are known to target E-cadherin are *Campylobacter jejuni* [130], *Escherichia coli* [131], *Shigella flexneri* [132], *Helicobacter pylori* [133], *Pseudomonas aeruginosa*, *Serratia marcescens* [134], C*lostridium perfringins* [135] and *Enterococcus facium* [136]. Reduced E-cadherin membrane expression and increased cytosolic E-cadherin are associated with higher vascular endothelial growth factor-A levels and predict poor survival [137].

Other carcinogenic compounds are known to be produced by bacteria. Colibactin from *E. coli* can cause double-strand breaks in DNA [138], and sulfidogenic compounds, such as hydrogen sulphide produced by *Desulfobacter, Desulfobulbus, Desulfotomaculum,* and *Desulfovibrio* genera and *Bilophila wadsworthia* [139], have also been shown to be genotoxic [140].

1.4.3 Inflammation, oxygen and microbiota

Mucous is the primary physical protection of the intestinal epithelium. If this mucous barrier is compromised, the gut's epithelial lining can be directly affected by microbial metabolites and toxins. Some Bacteroidetes and Clostridia can limit inflammation in the gut by producing SCFAs, which can inhibit pro-inflammatory neutrophil activity, stimulate regulatory T cells (Tregs), and enhance the intestinal barrier via regulating tight-junctions [141-143]. Additionally, Clostridia can stimulate mucous production, protecting the intestine from inflammatory stimuli, and decreasing gut permeability [144].

Inflammation in the gut due to dysbiosis can increase blood flow to the area, which increases oxygen in the lumen. The sudden influx of oxygen can further dysbiosis via overgrowth of bacteria capable of aerobic respiration such as *Enterobacteriaceae,* which can rapidly outcompete anaerobic commensal species which cannot compete in an oxygen-rich environment [145]. Additionally, tumour hypoxia is frequently encountered in the tumour microenvironment and may play a role in microbial selection [102, 105, 146].

Continued inflammation allows microbes and microbial compounds to interact directly with the epithelial surface causing damage and increased inflammatory responses [147]. The damaged epithelial cells provide a source of phospholipids, that when metabolised, the carcinogens ammonia and acetal aldehyde are produced as byproducts [148, 149]. This reduced defence against lumen content, microbial compounds, and pathogen colonisation increase CRC risk [150].

Biofilms are structures comprised of a matrix of polymeric substances such as polysaccharides that give bacterial communities microenvironments in which they are protected from shearing forces, immune responses, and antimicrobial compounds, enhancing their survival and growth by concentrating metabolites [151, 152]. In the oral cavity, biofilms can cover teeth and sequester acids that allow the microbes to survive, but also cause tooth decay and carcinogenesis [153], as they can concentrate carcinogenic compounds, concentrating their effects in a localised area [154]. When inside a biofilm, microbes exhibit different phenotypes that can quickly change, leading to diversity among otherwise homogenous communities [155], as biofilms can facilitate horizontal gene transfer (HGT) from one species to another [156], and concentrating quorum sensing homoserine lactones can trigger pathogenic gene expression [157].

Receptor interactions between bacterial cells allow the co-aggregation of multiple species within biofilms [158]. Biofilm-producing species from the oral microbiome appear in the gut and may have a role in the early-stage carcinogenesis [159], such as the cancer-associated *Porphyromonas gingivalis*, which has been identified in CRC tissue samples [93]. Other transient oral species such as *Treponema denticola, F. nucleatum* and *Tanerrella forsythia* can survive in the gut by forming these proxy-oral communities, sheltered from the general gut environment [160]. Additionally, *Campylobacter showae*, strain CC57C, has been shown to co-aggregate with a carcinogenic strain of *F. nucleatum* via adhesion proteins [109].

Increased biofilms have been seen on CRC tumours in comparison to adjacent and healthy tissues, particularly on right-sided tumours, which are associated with poor prognosis [161]. These associations may be not causative of CRC but may simply be more prominent in the CRC environment.

1.4.4 The microbiome of tumour tissues

Healthy gut microbiome communities average ~160 species per person from a pool of more than 1150 characterised species [162]. Studies investigating the gut microbiome routinely use faecal samples; however, faecal microbiomes are not representative of tissue microbiomes [163, 164]. Therefore, studies using tissue samples may be more relevant for the study of treatment and pathogenesis, as many bacterial toxins rely on interactions with host-cell components, such as E-cadherin, to have pathogenic outcomes [123]. Despite this potential for radical diversity, the microbial communities of tumour tissues are dominated by four main phyla, Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria [165]. Therefore, the discriminatory power of communities based on lower taxonomic classifications such as phylum is lacking, and distinctions need to be identified at the species, strain and functional metabolic level for microbiome-based phenotype classifications. Furthermore, certain microbial products are only produced by certain species within a genus, and strains within a species, i.e., *Escherichia coli* are common microbiome constituents; however, some strains are benign while others are capable of producing deadly toxins [166].

1.4.4.1 Diversity

Alpha diversity is the diversity (i.e., number of different members) of members within an individual sample. Tumours have been found to have higher levels of alpha diversity compared to controls [108, 167-169]; however, some studies have found the opposite [170], or find the differences to be insignificant [119, 139, 171, 172]. Alpha diversity has been reported as being higher in CRC than in adenomas [173], and lower alpha diversity has been reported in survivors of CRC, compared to non-survivors [174]. The significance of alpha diversity in different studies may reflect the cohorts used, study design and chosen diversity measures [175].

Differences in tumour diversity may be more significant in different countries/regions [172]. Another factor is biopsy location, as demonstrated in one experiment, which took adjacent tissue from 5 cm and 10 cm distant from tumours, and found greater diversity in the 10 cm samples, although this was not statistically significant [119].

Taking the studies above into account, and according to a recent meta-analysis, alpha-diversity cannot reliably be used to distinguish between disease states [176]. Instead, distinct taxa within communities should be considered the relevant factors [176].

Beta diversity as a measure of differences in composition between groups. Beta diversity distinguishing between differences in taxa between individuals, rather than comparing the level of different taxa within individuals as with alpha diversity. For example, beta diversity allows for the categorisation of samples into enterotypes, or defined microbial communities, such as was performed by Arumugam et al. in 2011 [177], placing subjects into one of three broad clusters, based on 16S analysis of faecal samples. Enterotyping was also utilised by Thomas et al. [108] in which the majority of RC patients were separated from healthy patients into an enterotype categorised by higher *Dorea* and *Bacteroides* [108]. Additionally, a study by Sobhani et al. found significant beta diversity differences between healthy and CRC patients, and between healthy patients with and without high methylation of genes associated with CRC [178]. Beta diversity is a more useful measure of evaluating differences between the microbial communities of disease states, as it takes into account inter-sample variability.

1.4.4.2 Composition of the CRC tumour microbiome

1.4.4.3 Phyla

The most dominant phyla in tissues reported are Firmicutes, Bacteroidetes and Proteobacteria [93, 108, 167, 170, 172, 174, 175]. Fusobacteria is also found to be a predominantly abundant phylum [93, 108, 167, 171, 173-175], although not as frequently. Additionally, some studies find Synergistetes [170], Actinobacteria [108, 172], Verrucomicrobia, and Parcubacteria (candidate phyla OD1) [108] as highly abundant phyla in tumour tissues. Lower abundance of Firmicutes [167, 171, 172, 179], Bacteroidetes [121, 167, 171], Clostridiales [171], and Bacilli [121] have been reported in tumours compared to normal tissue, while Proteobacteria have been reported as both scarce [175] and enriched [167], in different instances.

1.4.4.4 Genera

Across studies, more than 80 genera are reported as enriched in colorectal tumours, with varied support. *Fusobacterium* [93, 109, 119, 164, 167, 170-175, 179-182], and *Bacteroides* [93, 108, 119, 164, 170, 175, 181, 183] are the most widely reported. Genera found to be enriched and depleted in tumour tissues (compared to normal tissues) can be found in Table 1.3.

There is often an overlap with the genera found to be enriched in tumours in some studies and depleted in others, such as *Bacteroides* [118, 167, 171, 173], *Faecalibacterium*, *Kluyvera* [179], *Blautia*, *Alistipes* [173], *Parabacteroides*, *Ruminococcus* [168], *Sutterella,* and *Collinsella* [173]. Taxa found to be decreased in tumours, without contradiction appear in the literature more rarely, implying that the tumour microenvironment contains substantial heterogeneity across populations, selecting for a variety of microbiomes in different circumstances. However, this may also indicate technical limitations of microbiome studies, for instance, where different taxa are omitted with different cut-offs for rarity or sampling differences. Genera reported to be enriched consistently, and those with conflicting reports of enrichment will likely provide the greatest insight into differences between tumour microbiomes.

1.4.4.5 Species

Some species have been identified as being enriched in tumour tissue, although due to difficulties in resolving to the species level generally, and particularly using 16S rRNA, these reports are less frequent in the literature and should be considered to be less accurate assessments. Species found to be enriched and depleted in tumour tissues can be found in Table 1.4.

Species associated with the oral microbiome are reported to be enriched in tumour tissues. They tend to be anaerobic, have biofilm production and modification capability, are implicated in dental plaques and oral disease, such as periodontitis

[184, 185]. Oral species found in tumours include *F. nucleatum* [109, 118, 171, 174, 181, 182] *F. periodonticum* [118, 181], *Parvimonas micra*, *Peptostreptococcus stomatis*, *Gemella morbillorum*, *Leptotrichia trevisanii* [118], *Selenomonas sputigena* [109], *Lachnospiracea intertie sedis* [119], *Porphyromonas gingivalis*, *Treponema denticola* and *Tannerella forsythia* [93]. *Faecalibacterium prausnitzii*, is commonly associated with good gut health due to its ability to produce butyrate [186]; however, it has been reported as being both enriched [174, 181], and decreased [118, 169, 173] in tumours.

Table 1.3. Genera reported as enriched and depleted in tumour tissues.

Enriched (E)				**Depleted (D)**		**Conflicting Reports**	
Genera	**Ref.**	**Genera**	**Ref.**	**Genera**	**Ref.**	**Genera**	**Ref.**
Bulleida	[172, 175]	*Paraprevotella*	[119, 175]	*Anoxybacillus*	[109]	*Alistipes*	E: [139, 164]; D: [173]
Campylobacter	[109, 172, 175]	*Parvimonas*	[164, 172, 173, 175, 180]	*Bacilli*	[121]	*Bacteroides*	E: [93, 108, 119, 164, 170, 175, 181, 183]; D: [118, 167, 171, 173]
Clostridium	[93, 108, 121, 164, 175]	*Peptostreptococcus*	[164, 170, 175, 180]	*Citrobacter*	[179]	*Blautia*	E: [93, 119, 164, 169, 181]; D: [173]
Coriobacterium	[172, 179]	*Porphyromonas*	[93, 115, 139, 164, 170, 172, 175]	*Cronobacter*	[179]	*Faecalibacterium*	E: [93, 164, 174, 181]; D: [179]
Desulfovibrio	[108, 175]	*Prevotella*	[93, 119, 169, 170, 175]	*Enterobacteria*	[179]	*Kluyvera*	E: [179, 183]; D: [179]
Dorea	[108, 181]	*Roseburia*	[93, 164, 168, 172, 179]	*Holdemania*	[109]	*Parabacteroides*	E: [108, 139]; D: [168]
Gemella	[170, 172, 173, 180]	*Selenomonas*	[93, 109]	*Microbacterium*	[109]	*Ruminococcus*	E: [121, 139, 164]; D: [168]
Granulicatella	[170, 180]	*Shewanella*	[164, 174]	*Pseudoflavonifractor*	[109]		
Haemophilus	[170, 180]	*Staphylococcus*	[139, 169]	*Serratia*	[179]		
Leptotrichia	[109, 173]	*Streptococcus*	[115, 119, 170-172, 175, 180]	*Shigella*	[179]		
Odoribacter	[108, 139, 175]	*Treponema*	[93, 175]	*Sutterella*	[173]		
Fusobacterium	[93, 109, 119, 164, 167, 170-175, 179-182]	*Veillonella*	[115, 170, 180]	*Collinsella*	[173]		

Table 1.4. Species reported as enriched and depleted in tumour tissues.

Enriched (E)				**Depleted (D)**		**Conflicting Reports**	
Species	**Ref.**	**Species**	**Ref.**	**Species**	**Ref.**	**Species**	**Ref.**
Aggregatibacter aphrophilus	[171]	*Fusobacterium necrophorum*	[118, 181]	*Acinetobacter baumannii*	[171]	*Faecalibacterium prausnitzii*	E: [174, 181]; D: [118, 169, 173]
Akkermansia muciniphila	[169]	*Fusobacterium nucleatum*	[109, 118, 171, 174, 181, 182]	*Acinetobacter* sp.	[171]		
Bacteroides fragilis	[108, 118, 174, 181]	*Fusobacterium periodonticum*	[118, 181]	*Alistipes putredinis*	[173]		
Bacteroides massiliensis	[181]	*Gemella morbillorum*	[118]	*Bacteroides dorei*	[118]		
Bacteroides uniformis	[108]	*Hafnia alvei*	[183]	*Bacteroides stercoris*	[118]		
Bilophila sp.	[108]	*Lachnospiracea intertie sedis*	[119]	*Bacteroides vulgatus*	[118]		
Blautia coccoides	[169]	*Leptotrichia hofstadii*	[109]	*Collinsella aerofaciens*	[173]		
Blautia sp. Marseille	[181]	*Leptotrichia trevisanii*	[118]	*Enterobacter cloacae*	[171]		
Campylobacter showae	[109]	*Methylobacterium suomiens*	[174]	*Fusobacterium mortiferum*	[118]		
Citrobacter freundii	[183]	*Parvimonas micra*	[118]	*Fusobacterium necrogenes*	[118]		
Clostridium sensu strictu	[119]	*Peptostreptococcus stomatis*	[118]	*Fusobacterium ulcerans*	[118]		
Comamonadaceae acidovrax spp.	[119]	*Porphyromonas gingivalis*	[93]	*Fusobacterium varium*	[118]		
Coprococcus comes	[181]	*Selenomonas sputigena*	[109]				
Dorea longicatena	[181]	*Tannerella forsythia*	[93]				
Escherichia coli	[183]	*Treponema denticola*	[93]				
Fusobacterium hwasooki	[93]						

1.4.5 Neoadjuvant treatment of RC

The predominant treatment for RC is surgical resection, often with neoadjuvant chemoradiotherapy (nCRT). Neoadjuvant treatments are administered before surgery to shrink more advanced tumors in order to improve outcomes. These include chemotherapy, radiation therapy or hormone therapies. As well as the localised primary tumour effects, neoadjuvant treatments are used to control metastatic disease [187]. The main goal of neoadjuvant treatment in RCs is to reduce tumour mass before surgery to reduce the volume of tissue removed.

1.4.6 Tumour regression grading

Tumour regression grading (TRG) is a method of categorising the level of tumour regression after cytotoxic treatment, based on the relative amount of tumour remaining after therapy or level of fibrosis induced relative to the level of residual tumour. There are multiple systems for TRG, such as the American Joint Committee on Cancer (AJCC), Dworak, Mandard, and Ryan systems, as well as modified systems that take into account lymph nodes and primary tumour regression [188].

Dworak TRG has five grades of regression: Four: complete regression where no residual tumour cells are seen, termed pathological complete response (pCR); three: near-complete regression where very few tumour cells are seen; two: moderate regression where significant fibrotic changes are detected with few tumour cells or groups of cells; one: minimal regression where tumour cells are dominant with some fibrosis; and zero: where no regression is seen [189]. As Dworak scoring is the main method of regression scoring at Christchurch hospital, it was used for all patients throughout this thesis.

1.4.7 Adjuvant chemotherapy

Capecitabine is a chemotherapy drug which is the currently prescribed prodrug of 5-fluorouracil (5-FU). 5-FU is a uracil analogue with an attached fluorine atom, which once converted from the prodrug capecitabine, can be converted into three downstream metabolites that poison the available pools of uracil and thymine, which interrupts proper DNA and RNA synthesis. The incorporation of these analogous metabolites causes double-stranded breaks and improper DNA replication and RNA transcription, interrupting the cell cycle and leading to apoptosis.

Normally, uracil is modified with a methyl group by thymidylate synthase (TS) to be converted into thymine. Fluro-deoxyuridine monophosphate (FdUMP) inhibits TS, preventing thymine conversion from uracil. 5-FU is converted into fluorodeoxyuridine triphosphate (FdUTP) and interferes with DNA synthesis when used in place of thymine. Additionally, conversion into fluorouridine triphosphate (FUTP) is used in place of uracil to inhibit RNA formation. 5-FU may cause hepatic toxicity; however, approximately 80% of the conversion process occurs in the liver, which is the primary location for colorectal metastases [190]. Despite this, even with severe liver dysfunction, it is considered effective and safe compared to other therapies [191].

Capecitabine has mostly replaced intravenous 5-FU due to its improved safety and efficacy profile [192]. It is the mainline drug used in nCRT for CRC treatment in NZ, and it can be used as an adjuvant monotherapy therapy and radiosensitiser, with tablets being taken on the day of radiotherapy treatment. The standard dose is 1250mg/m^2 and is associated with several adverse events such as nausea, diarrhoea, vomiting, stomatitis (inflammation of the mouth and lips, which may result in ulceration), hand-foot syndrome, which results in tingling, numbness, and broken skin and ulceration of the hands and feet. Additionally, polymorphisms and copy number variation of the TS gene can lead to tumour resistance and patient oversensitivity [190].

Leucovorin is often included in a 5-FU therapy to mitigate the negative impact 5-FU has on healthy tissue by acting as a folate supplement, a precursor for thymine. Leucovorin has also been shown to improve treatment and survival outcomes by increasing the pool of tetrahydrofolate, which can be utilised for TS inhibition and 5-FU metabolite synthesis [193]. The frequent lack of tolerance to the drug can often lead to patients ceasing their treatment early. Combination chemotherapies include FOLFOX (leucovorin, 5-FU and oxaliplatin) and FOLFIRI (leucovorin, 5-FU, irinotecan). Post initial treatment and surgery, a combination chemotherapy regimen of capecitabine and cisplatin (CAPOX) is often used to prevent or treat recurrence, with or without additional radiotherapy [194, 195].

1.4.8 Radiation therapy

Radiation therapy involves using ionising radiation via linear energy transfer (LET) to kill cancer cells and has been used for this purpose since 1896, soon after x-rays were first discovered, although the mechanism of action was not understood at the time [196]. Approximately 50% of all cancer patients will receive radiation therapy during their treatment, particularly for non-operable cancers and palliative treatment [197].

CRC patients receive short-course radiation over five days or long-course radiation over several weeks. The benefit of short-course radiation is that it can be delivered quickly in palliative care scenarios, or for patients with low tolerance to the treatment. A retrospective study of 28,193 non-metastatic RC patients receiving short- or long-course radiation found no statistically significant benefit of long-course over short-course radiation; however, it was found that a longer interval between therapy and surgery correlated with higher pathological complete response (pCR) rates [198].

The primary mechanism of action for RT is causing damage to the cells and DNA of tumour tissue via direct damage to the DNA, or through LET ionising water molecules in the cell, generating reactive oxygen species (ROS), which in turn damage DNA or proteins required for critical cell functions, resulting in cell cycle cessation, and cell death [197].

Paradoxically, therefore, a lack of water in combination with tumour hypoxia, which would otherwise hinder the growth of tumour cells, are significant factors that can reduce the effectiveness of radiotherapy, and is thus an area of significant research interest [199-205]. Hypoxic conditions occur due to a lack of proper blood supply and increased metabolism and cell division which rapidly depletes available oxygen [105].

As with many cancer therapies, radiotherapy impacts both tumour cells and healthy tissues of the patient, causing adverse side effects such as fatigue, mucositis, intestinal breakdown, gastrointestinal symptoms, bleeding, and changes in appetite, which are all associated with poor outcomes and quality of life reductions [206, 207].

1.4.9 Surgery

The surgical approaches post RT are usually limited resection of the malignant tissue or total mesorectal excision (TME) for advanced and invasive cancers [208]. Rates of good surgical outcomes and avoidance of temporary and permanent stoma differ between the sexes, principally due to anatomical differences between the musculature of respective pelvic floors [209, 210]. Additionally, the female colon is longer on average, with shorter anal canal and rectum than males, which may contribute to more significant surgical complications during resection and increased rates of stoma in older females [211, 212]. Stoma implementation can cause psychological issues and sexual dysfunction in both sexes [213, 214].

1.4.10 Prognostic markers of response to neoadjuvant treatment

The optimal result of neoadjuvant treatment is a pCR or complete regression. However, this outcome occurs only for approximately 10-25% of patients [215], with some studies reporting rates lower than 1% [216].

Predictive markers for pCR is a field of research attracting significant attention. Clinical features also have predictive value for neoadjuvant treatment outcomes, such as tumour size, distance from the anal verge, nodal involvement, the time between nCRT and surgery, pre-treatment serum carcinoembryonic antigen (CEA) levels, differentiation, and macroscopic ulceration [217, 218].

1.4.11 Molecular prognostic markers

There are no predictive molecular biomarkers for nCRT response that have been validated for clinical use [219]. However, potential molecular markers for response to therapy include mutations and copy number of oncogenes. The roles of *PT53, KRAS, BRAF* and *PIK3CA* as prognostic markers in nCRT is not definitive, with some studies showing no association [220-222], while other researchers have shown positive associations with response to nCRT [223], such as wild-type *PT53* gene being associated with a more significant response to nCRT [224]. It is speculated that these results may be due to the location of mutations in these genes, indicating that taking a personalised approach to research and treatment may provide better outcomes for some patients [225, 226].

Groupings of tumours with common factors, such as the consensus molecular subtypes (CMS) [227] may be more informative [228-230]. Subtypes such as those exhibiting hypermethylation of DNA and chromosomal instability [231, 232] have proved the most informative in RCs. Gaedcke et al. found ten differentially methylated regions which were predictive of disease-free survival (DFS) [233] while Murcia et al. found higher genetic instability associated with a greater response to nCRT [232]. Alternatively, as suggested by Ma et al., continuous subtypes may provide more reproducible results in terms of prognosis, stage, and grade using transcription data than the discrete subtyping CMSs provide [234].

Many differentially expressed genes have been associated with response to therapy, a summary of which can be found in Table 1.5. An increase in *XRCC3* (x-ray repair cross-complementing protein), a gene involved in DNA repair, has been associated with nCRT resistance [235, 236]. Karagkounis et al. showed that between responders and non-responders, decreased neuronal pentraxin-2 (*NPTX2*) expression in tumours was associated with an increased response to nCRT and DFS [237]. Differential expression of the zinc finger protein ZNF160, Helicase For Meiosis 1 (HFM1), additional sex combs-like protein 2 (ASXL2), aldo-keto reductase family 1 member C3 (AKR1C3), C-X-C motif chemokine ligands (CXCL9–11), indoleamine dioxygenase-1 (IDO1) and matrix metalloproteinase-12 (MMP12) have also been used to discriminate between response groups, with varying accuracy [238]. However, these genes are not used clinically for routine screening.

Table 1.5. Summary of genes thought to be involved in therapy response.

Gene	Name	Function	Ref.
XRCC3	X-ray repair cross-complementing 3	Member of the RecA/Rad51-related protein family, participate in homologous recombination to maintain chromosome stability and repair DNA damage.	[235, 236]
NPTX2	Neuronal pentraxin 2	Involved in excitatory synapse formation and plays a role in the clustering of alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA)-type glutamate receptors at established synapses, resulting in non-apoptotic cell death of dopaminergic nerve cells.	[237]
ZNF160	Zinc finger protein 160	Zing-finger protein and may function in transcription regulation.	[238]
HFM1	Helicase for meiosis 1	ATP-dependant DNA helicase expressed mainly in germ-line cells.	[238]

ASXL2	ASXL transcriptional regulator 2	Epigenetic regulator that binds histone-modifying enzymes involved in assembly of transcription factors. Mutations in this gene have been associated with cancers across several tissue types. Has a role in neurodevelopment, cardiac function, adipogenesis and osteoclastogenesis.	[238]
AKR1C3	Aldo-keto reductase family 1 member C3	Member of the aldo/keto reductase superfamily. Catalyses reduction of prostaglandins, phenanthrenequinone and oxidation of 9-alpha,11-beta-PGF2. May play a role in allergic diseases and controlling cell growth and differentiation.	[238]
CXCL9	C-X-C motif chemokine ligand 9	Antimicrobial gene involved in T cell trafficking. Is a chemoattractant for lymphocytes.	[238]
CXCL10	C-X-C motif chemokine ligand 10	Antimicrobial gene, stimulates monocytes, NK and T cell migration and modulation of adhesion molecule expression.	[238]
CXCL11	C-X-C motif chemokine ligand 11	Antimicrobial gene, induces a chemotactic response in activated T cells, and is the dominant ligand for CXCLR3 and is induced by INF-γ.	[238]
IDO1	Indoleamine 2,3-dioxygenase 1	Catalyses the rate-limiting step in tryptophan catabolism. Thought to play roles in antimicrobial and anti-tumour processes, neuropathology, immunoregulation, and antioxidant activity. Expressed in dendritic cells, monocytes and macrophages and modulates T cell behaviour.	[238]
MMP12	Matrix metallopeptidase 12	Involved in the breakdown of extracellular matrix. Degrades soluble and insoluble elastin and may play a role in aneurysm formation, and mutations in the gene are associated with chronic obstructive pulmonary disease.	[238]

1.5 Microbial and immune interactions with neoadjuvant treatment

When treating cancers, the microbiome is being increasingly shown to have consequences for cancer treatment. The microbiome has been shown to affect immunotherapy, chemotherapy, and radiation therapy [239-241], particularly in CRCs [242].

1.5.1 Chemotherapy

During 5-FU treatment, folate metabolism is a crucial factor due to its role in thymine and uracil synthesis [243, 244]. The gut microbiome is known to impact the uptake of metabolites [71] and research using a *Caenorhabditis elegans* model has demonstrated folate uptake is modulated by *E. coli* and can alter the lifespan of the host, independent of supplementation [245, 246]. The role of microbiota in the treatment of cancer with 5-FU was investigated with live bacteria, and different strains of live *E. coli* were found to impact the efficacy of 5-FU [211]. Many different strains of *E. coli* are prominent in the gut microbiome and have been shown within CRC tumour tissues [109, 183, 247].

Irinotecan, another drug used for CRC treatment, is deactivated by the liver via glucuronidation. It has been shown that when re-entering the intestine through the bile duct as waste, it can be reactivated by microbial β-glucuronidases [248]. This conversion increases the drug's toxicity and exacerbates side effects like diarrhoea, which affect the majority of CRC patients [249].

1.5.2 Radiation therapy

Radiation kills cells by using large amounts of energy in a targeted area to damage membranes, proteins and DNA by generating free radicals and ROS. As cells die, they release pro-inflammatory and immunostimulatory molecules which attract more immune cells to the site. The abscopal effect can also be induced, which produces a systemic immune response to distant malignancies with radiotherapy, in addition to localised effects [250]. The microbiome's role in radiation therapy outcomes and related immune interactions has been gaining attention in the research space [242].

One of the main side effects of nCRT therapy, particularly when used in areas containing mucosae, such as the oral cavity or pelvic region, is mucositis. Mucositis symptoms include pain, ulceration, bleeding, nausea vomiting diarrhoea and constipation [251], which leads to reduced quality of life and poorer outcomes for patients [252]. Mucositis has been associated with upregulated expression of pro-inflammatory cytokines IL-1β, IL-6, and tumour necrosis factor (TNF) [253]. The microbiome has been implicated in increasing severity of mucositis, with reported increases of Enterobacteriaceae and *Bacteroides,* and decreases in health-associated *Bifidobacterium* and *F. prausnitzii* [254].

As mentioned in a previous section, E-cadherin is a common target for numerous microbes and may have an immune-modulating effect [123]. The association of E-cadherin expression with survival may be immunologically mediated, as it has been demonstrated that E-cadherin adhesion disruption causes dendritic cells (DCs) to mature to a regulatory phenotype, rather than an antigen-presenting effector phenotype [122], leading to enhanced tumour progression, lack of T cell proliferation, and microbial persistence [255].

E-cadherin mediated maturation of dendritic cells is impactful, as DCs are antigen-presenting cells that act as a central regulator of immune activity. DCs are capable of uptaking and presenting antigens for CD8 and natural killer (NK) cells to target. This factor is exacerbated by the immunosuppressive effects of radiation on dendritic cells, causing continued release IL-10 and becoming less effective at priming T-cells [256].

1.6 Microbiome sequencing and analysis

Sequence analysis pipelines involve data processing (quality control and sequencing trimming), analysis and statistical validation. With advances in Next-Generation Sequencing (NGS), throughput has increased dramatically, and the cost has decreased considerably (Figure 1.1).

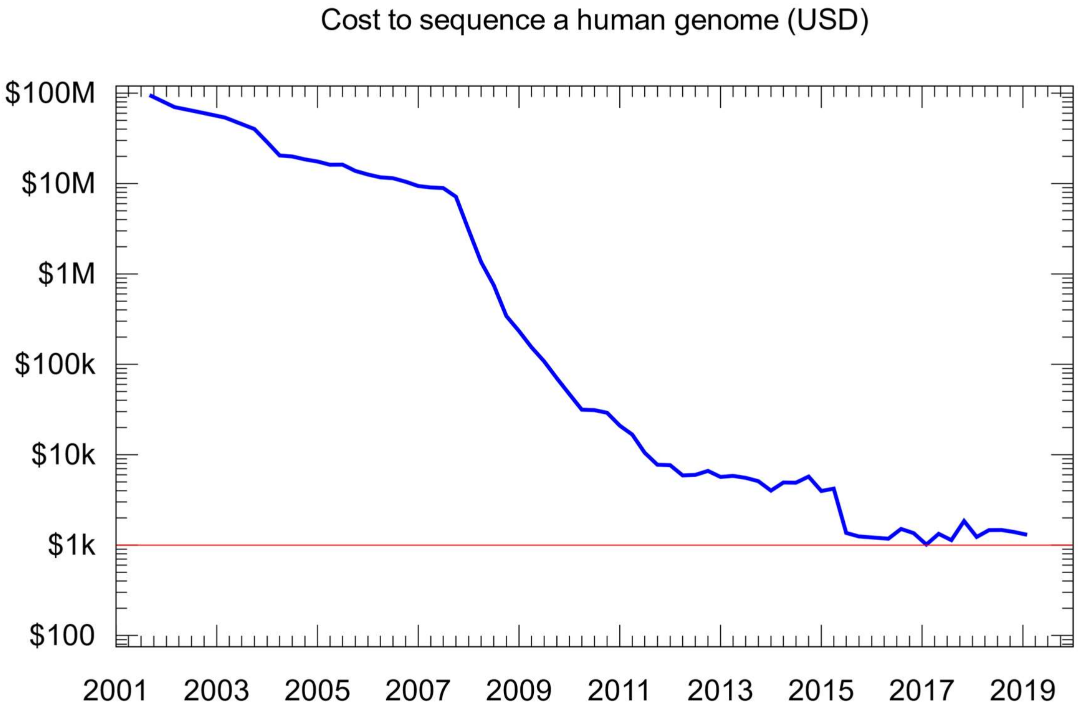

Figure 1.1. The Carlson Curve. Demonstrating the cost of sequencing over time. Image sourced from National Human Genome Research Institute (www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost)

The first sequencing experiments involved fragmenting sequences and transforming them into a bacterial artificial chromosome (BAC) to be amplified as the bacterial population that contains it multiplies [257]. Alternatively, Sanger sequencing is a low throughput method, that involves binding fluorescently tagged nucleotides to a template sequence, which are then used to generate a digital sequencing read via fluorescence detection [258].

Using modern NGS techniques, higher throughput options are available. With millions of sequence reads being produced from template sequences directly with synthesis-based methods such as bridge amplification used in Illumina sequencers, or non-synthesis-based methods like Oxford Nanopore Technologies (ONT) long-read sequencing.

In this thesis, RNA-Seq was performed using Illumina sequencing, 16S rRNA sequencing was performed using amplicon sequencing on an Illumina platform, and ONT sequencing was carried out on the GridION 5X. RNA-Seq was used for host and microbiome analysis of transcribed genes, while 16S rRNA and ONT sequencing were used only to analyse the microbiome.

1.6.1 Illumina sequencing

Illumina sequencing has evolved considerably since its inception, with the MiSeq, HiSeq and most recently, NextSeq and NovoSeq platforms now available. The principal difference between Illumina sequencing and other technologies is the employment of bridge amplification [259]. This process involves binding primers to nucleotide sequences, which are then attached to adapters that allow binding to a flow cell. The sequences are separated into a single strand, and the bound sequences are then supplied with raw nucleotides with fluorescent tags, which produce signals when incorporated during synthesis. These signals can then be interpreted and recorded as the supplied strand is sequenced, which provides a digital sequencing read. The synthesised double strand is then separated again, rebound, and sequenced again, thus amplifying the original sequence repeatedly.

1.6.2 Oxford nanopore long-read sequencing

ONT sequencing is a relatively new, non-synthesis-based method [260]. The reads can be up to a million base pairs long; however, they carry an inherently high error rate, which is decreasing over time with the introduction of new tools and technology [261].

New tools for aligning sequences that incorporate ONT reads efficiently are becoming more widely available [262-264]. These tools often work a lot faster than traditional sequence aligners as the query sequence tends to be much larger unbroken sequence and can thus match to fewer places on a reference genome. The errors in ONT reads are most common in homopolymeric repeat regions. Algorithms can be designed to compress homopolymeric regions into homopolymer compressed mers (HPC), allowing the algorithm to map long reads to a reference very rapidly; however, HPCs reduce ONT read sensitivity for some applications [265].

1.7 Sequencing techniques and methods

1.7.1 Amplicon sequencing

Amplicon sequencing is a process by which a region or sequence of interest is amplified exponentially using polymerase chain reaction (PCR) and sequenced at high depth, allowing for greater discriminative analysis of a region of interest, such as universally conserved genes like 16S rRNA or *cpn60* for identifying bacteria and assessing phylogenetic relationships [266], or for genotyping analysis [267]. High read depth is required for variation analyses to avoid spurious conclusions from sequencing errors, with the recommended depth to be approximately 100–300x depending on the sample type and the study design. Read depth is essential when analysing single nucleotide polymorphisms, where many reads of the same region are required to identify single-base differences and separate these from potential sequencing errors [268].

1.7.2 Shotgun sequencing

Shotgun sequencing is a short read sequencing technique that refers to sequencing all available nucleotides in a query sample. As an agnostic method of sequencing, it does not use any selection method or selective discrimination, sequencing everything in a given sample, thus reducing bias in downstream analyses. However, a downside of this method is that the resulting sequencing is less targeted and does not provide control over which reads are sequenced, resulting in low read number and depth of regions of interest, and a less informative analysis [269]. Shotgun sequencing is used less often than amplicon sequencing for microbiomes due to increased computational complexity, reduced microbial specificity (leading to a reduced ability to measure abundance accurately), and experiments investigating tissue microbiomes (i.e., tumour biopsies) the resulting sequencing data contains mostly host sequences.

1.7.3 Sequence alignment

A primary step in genetic analysis of all kinds is alignment. The generated sequence reads are mapped or aligned to a reference genome, allowing comparison of the sequenced reads and the reference, giving insight into changes between the reference and subject for variation analysis. Alignment software can take into account the presence of insertions, deletions and fusions with tools like TopHat2 [270], or be built for speed, with efficient splice aware alignment such as with STAR which allows RNA to be aligned to genomes taking into account the location of intergenic regions [271].

ONT sequencing data requires specialised algorithms. One such tool is MiniMap2 which can leverage the longer read length to quickly map reads to a reference genome and tolerate the ~15% error rate that is common in ONT reads, this is because longer reads and increased query sizes give the ability to skip over repetitive homopolymeric regions more easily [263].

1.8 Transcriptomics

RNA sequencing is used to research active gene expression or transcription, a field known as transcriptomics. By analysing the RNA present in a sample, researchers can discern which genes are being transcribed or are in active use in the sample under certain conditions. A standard method of performing gene expression experiments is subjecting cells or isolates to different stimuli to determine which genes are expressed and in what amount; the method is often used to determine the effects of treatments in both in vivo experiments samples taken from case and control patients. Transcriptomics requires annotated sequences, assignment of a gene name to identify it and a function, predicted or known via experimentation; RNA function is predicted based on what is known about similar sequences [272].

The advantage of transcriptomics in studying the microbiome is that it can be used to analyse microbial activity and not just abundance; however, this provides additional challenges. Bacterial genomes are not always well-annotated and frequently contain errors [273]. There is genetic similarity between species and differences within species [274], as well as the complicating factor of horizontal gene transfer, which allows microbes to express genes that are not originally part of their genome [275], thus making discriminating between microbial species using transcriptomics or assigning function to transcripts accurately, particularly difficult.

1.8.1 Gene set expression analysis

Gene set expression analysis (GSEA) can computationally determine the statistical significance of differences between different biological states or phenotypes. Profiles built from gene expression data are compared to a gene set database or genome annotations to determine which genes are being expressed and at what level. Many tools and methods are available for gene expression analysis, such as the GSEA tools built by the Broad Institute which utilises the molecular signatures database (MSigDB) [276] containing more than 22,500 annotated human genes in the latest release. Other commonly utilised databases include gene ontology (GO) [277], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [278], and MetaCyc [279].

GSEA software first defines a baseline to compare genes by their level of expression, then uses increases or decreases from this point to determine the increase or decrease in expression of other genes [276]. Differentially expressed genes are given an enrichment score and ranked based on how distant from the centre they are.

Hypothesis testing by Student's t-tests, Welch's t-test, paired t-tests, analysis of variance (ANOVA), or linear models are carried out on the results to determine the results' local significance. Local significance is a test of the strength of gene associations with the phenotype or state, while other methods additionally use global tests, which are used to compare the calculated associations, such as Wilcoxon rank-sum, Fisher's Exact and Pearson's Chi-squared tests [280].

1.8.2 Immune infiltration

Tumour heterogeneity affects not just the genetics of cancer cells but also the cells which comprise a tumour. Other cells in the tumour microenvironment include fibroblasts, vascular cells, stem cells, adipocytes, pericytes and immune cells [281].

Transcriptomics can be used to identify cell types by their associated gene expression, which allows the determination of the proportion of tumour cells compared to stromal or immune cells in a tumour. Tools such as ESTIMATE provide a score for the proportion of different cell types in the tumour microenvironment using gene expression data [282].

Additionally, the different immune cells present can be further scrutinised by profiling individual immune cell gene expression to determine the subpopulations of immune cells and their level of activity or stage of maturation. Software like xCell [283] and CIBERSORT [284] use gene expression data and compare it to databases of gene expression profiles associated with different cells, allowing discrimination between different immune cell subpopulations, such as M0, M1 or M2 macrophages.

1.8.3 Transcriptomic analysis tools

Similar to DNA sequencing, RNA can be aligned to a reference genome or transcriptome. RNA alignment can be done effectively for single organisms with Bowtie [285], STAR [271], or with BWA (Burrows-Wheeler Alignment) [286], or gene quantification software can be used, such as Salmon [287]. Software specifically designed for RNA sequencing is required for alignment, as algorithms need to consider features unique to RNA-seq such as splicing, coding regions, and uracil nucleotides, which are not present in genomic sequences [288]. Splicing in human RNA is common, but not in prokaryotes. Splicing allows a single transcribed RNA to be processed into many different forms depending on which introns are excised and thus altering their function. Splice aware aligners such as STAR are built to handle the RNA transcripts generated from the genome and map them accurately to the reference genome. Typically, rRNA depletion is used before sequencing to prevent excessive rRNAs sequencing, which would otherwise comprise a substantial proportion of sequencing reads than are not phenotypically informative, reducing read depth of informative genes. However, this is often an incomplete process and can have drawbacks when applying RNA-Seq for microbial identification.

1.9 Metagenomics

Metagenomics, is an interdisciplinary field of study for the investigation of microbial communities using genetic sequencing, involving methods and concepts from immunology [73, 289, 290], microbiology [291], genetics [292, 293], ecological [294] and computer science [295, 296]. Many species have proven difficult to culture due to undiscovered metabolic and environmental needs, the limitations of culturing for microbial identification, lead to research focusing on genetic analysis [297-300]. Sequencing the DNA or RNA of microorganisms is a method for studying microorganisms culture-free. With advancements in sequencing techniques, such as 16S rRNA analysis, transcriptomics and whole-genome sequencing, sequencing has become a standard bacterial classification method. Additionally, metabolomics is often used to detect and study the metabolites produced in a sample, typically using liquid or gas chromatography, to augment metagenomic studies [301].

Metagenomic microbiome analyses use the extracted and sequenced nucleotides present in a sample to assess the microbial community present [302]. One of the most significant hurdles in the field is the data analysis bottleneck, as data is being produced much faster than it can be analysed [303].

Amplicon sequencing discrete genomic regions can improve the speed of analysis and reduce computational requirements by restricting the analyses to the most biologically conserved or informative data [303]. Marker genes like 16S rRNA in bacterial and archaeal analysis and internal transcribed spacer (ITS) regions for analysis of fungi are most often used for taxonomic assignment, as they are well conserved across species [304]. Heavily conserved genes can be used to construct phylogenetic relationships between organisms using the slowly acquired variability between them, which results in a reliable phylum to genus level taxonomy. However, variability in conserved genes is often less apparent between species that are more recently evolutionarily separated, making precise species identification within metagenomic samples challenging.

1.9.1 16S rRNA amplicon sequencing

The 16S rRNA gene is the most commonly sequenced marker in bacterial metagenomics, used to measure the abundance of identified microbes. The transcribed gene is part of the small ribosomal subunit, the 16S denoting a Svedberg number, indicating how quickly it would sediment during ultracentrifugation. Ribosomes are present in all species and critical for protein synthesis, making their loss or rapid alteration over evolutionary time, unlikely.

The 16S rRNA gene is useful as it often exists in multiple copies in the genome, containing conserved regions critical for function, which flank nine variable regions that can tolerate minor changes that can differentiate between organisms based on evolutionary divergence. A significant advantage of 16S rRNA sequencing in microbiome metagenomic analysis is that primers are specific to prokaryotes, reducing the potential for host sequencing contamination.

Initially, a technique to separate the 5S rRNA gene by separating it from mixed samples with electrophoresis was utilised in the 1980s; a phylogenetic analysis was performed to determine the evolutionary distance between sequences. This method proved useful for analysing unculturable species from extreme environments, such as hydrothermal vents [305]. Using the 120bp region of 5S rRNA had its limitations, as electrophoresis was needed to separate the gene from other molecules, limiting the process to low complexity samples.

It was suggested that larger genes could be used to increase the fidelity of the technique, such as the 23S rRNA gene; however, at approximately 3000bp, it would be time-consuming to analyse [306]. The 16S rRNA gene is approximately half the length and was easier to analyse. The initial method involved making a DNA library in bacteriophage, allowing them to replicate, and using a 16S rRNA specific probe to select clones, which could then be phylogenetically analysed [307]. With PCR, this labour-intensive process could be sped up by amplifying only the variable regions of interest using known priming sites in flanking conserved regions [308, 309].

1.9.2 16S rRNA analysis

The most commonly used tools and pipelines for 16s rRNA analysis are mothur [310], UPARSE [311], DADA2 [312], and QIIME [313]. QIIME remains the most common despite it being no longer supported by the developer, while its successor QIIME2 is available with a graphical user interface, it is more computationally intensive still lacks some of the functionality of the original [313, 314].

R packages can be easily combined with these pipelines, such as data visualisation tools, included as part of a pipeline or by linking into other R packages such as ggplot

[315], phyloseq [316], and vegan [317]. More recently, Kraken2 has gained support for 16S rRNA analysis, with the authors suggesting it is faster and more accurate than alternatives [318]; however, this has not been independently tested.

1.9.3 16S classification strategies

Using the software mentioned above, the primary method for assigning taxonomy to 16S rRNA data is to cluster variant genomic regions into operational taxonomic units (OTUs). These can be constructed de-novo or against a closed reference. An issue with de-novo OTU picking is that the resulting OTUs are clustered relative to the others in the sample (with a similarity threshold of 3%), further limiting species-level identification and making them incomparable between sample sets. Alternatively, closed reference OTU picking involves comparing sample sequences to a reference database. This approach's downside is that the taxa being investigated must be present in the database, and the same reference database used between studies for results to be accurately compared [319]. However, the choice of a reference database for comparative purposes may lead to the use of inaccurate or out of date databases, such as the widely utilised GreenGenes (GG) database, last updated in 2013 [320-323].

Amplicon sequence variants (ASVs), produced by DADA2 and qiime2 [324] were developed to overcome the shortcomings of OTU picking [319]. They use no arbitrary dissimilarity threshold. They infer the biological reality in the sample before sequencing and take into account amplification errors. Studies show that ASVs can be more sensitive than OTU picking and better discriminate between ecological patterns in the environment [312, 325]. ASVs provide a higher resolution taxonomy assignment due to distinguishing between single nucleotides, rather than by overall sequence similarity [326], allowing more reliable assigning of sequences to the species level [327]. ASVs can also be used in other analysis pipelines, such as QIIME, where feature tables can be populated with ASVs in place of OTUs.

1.10 Metagenomic shotgun sequencing

Metagenomic shotgun sequencing has advantages over 16S rRNA amplicon sequencing in that it can produce longer reads that are more useful for species detection and gene prediction [328]. Also, host DNA in a sample can be used to genotype the host at the same time as the microbial component, allowing for the simultaneous assessment of host-microbiome combinatorial phenotypes [329, 330], assuming appropriate read depth can be achieved.

Colonisation potential, drug interactions, the presence of pathogens and polymicrobial signatures can be investigated, which can inform diagnosis, treatment, probiotic and prebiotic applicability. Additionally, discriminating between host and microbial reads bioinformatically is preferable to enriching the microbial community and remove host sequences before sequencing, which may introduce bias to the analysis [329, 331-333].

1.10.1 Taxonomic assignment with shotgun sequencing data

With massive datasets being produced more routinely, more efficient methods for processing data are needed. Utilising k-mers makes it possible to use data containing sequences from multiple organisms and rapidly classify them. Many k-mer based classification tools are available for metagenomics, including 16s rRNA data [334] and single-threaded options [335]. One of the most popular options is Kraken [336]. While the Kraken successor, Centrifuge, was faster and required lower computational resources due to database compression, it had lower accuracy [337] and was ultimately superseded by Kraken2 [338].

The major hurdle with k-mer based metagenomic classification is constructing databases, as publicly available databases are often generalised or severely out of date. Construction of a custom database requires more computational resources than are required to utilise them; this requirement increases with the number of taxa used in an index, leaving them out of reach for many researchers [339].

1.10.2Meta-transcriptomics

16S rRNA amplicon sequencing is used to answer questions of microbial abundance, 'How represented are microbes in the community?', while meta-transcriptomics is used to measure the activity of microbes, 'What are these microbes doing?'. Transcriptomics is used to infer what proteins are produced in a cell, in different environments or conditions [272]. Protein-coding genes are well enough conserved to allow for high-resolution taxonomic assignment with RNA-Seq data [340].

Software that can be used for this classification type are DIAMOND [338] and Kaiju [339], rapidly assigning sequences to taxa using translated protein databases. For instance, Kaiju has the advantage of providing accession numbers for the gene/protein of the aligned sequence, if the associated reference has been annotated, allowing for gene expression analysis at a community level. However, annotation of prokaryote genomes is incomplete and sometimes unreliable due to lineage trends [341], leading to homology-based automated annotation using software such as Prokka resulting in significant hypothetical proteins and unannotated regions [342, 343].

1.11 Thesis hypotheses and aims

This thesis's main aims were to investigate the bacterial and host factors associated with chemoradiotherapy outcomes and determine which could be used as predictive prognostic indicators.

My hypotheses are that:

(i) Residual host reads interfere with the accuracy of taxonomic assignment in tissue microbiome analysis.

(ii) Microbiomes influence host response to chemoradiotherapy; the with microbiomes differing between response groups.

(iii) Bacteria are differentially abundant and have differential gene expression in different response groups, and their and gene expression impacts therapeutic outcomes.

(iv) Immune cell infiltration significantly contributes to therapeutic outcomes, and the microbiome influences this immune infiltration.

(v) In combination with host gene expression, immune cells and bacterial expression can be used as prognostic and predictive biomarkers of chemoradiotherapy response.

To test these hypotheses, using a cohort of rectal tumours and adjacent normal tissue samples collected before chemoradiation therapy, this thesis aims to:

(i) Improve methods for microbial assignment and database construction.

(ii) Compare sequencing technologies for their applicability in microbiome analysis.

(iii) Examine microbial gene expression using RNA-sequencing to identify taxa that may influence response to CRT in RC.

(iv) Investigate immune cell infiltration in rectal tumours and relate this to the microbial abundance, microbial gene expression, and therapeutic response.

(v) Investigate the utility of the above factors in combination with patient gene expression to discover potential biomarkers of response to CRT.

2 CHAPTER 2: METHODS

2.1 Summary

First, to validate method alterations, a synthetic dataset with known contents was used, and clinical RNA-Seq and Oxford Nanopore Technology (ONT) datasets were then used to determine the alterations' impact in a real-world context. Second, sequencing samples' comparative ability to assess the microbiome was assessed by sequencing samples RC patient tumour and normal tissue biopsies. The concordance of the relative bacterial abundance of taxonomies from the 16S rRNA, ONT and RNA-Sequencing platforms were assessed and measured using the altered methods.

Third, the RNA-Seq data from the RC patient samples were used to evaluate and correlate bacterial transcription with chemoradiotherapy response. Fourth, RNA-Seq gene expression data was used to estimate the abundance of immune cells in patient biopsies. Then, the predicted immune cell abundance was correlated with response to radiotherapy and bacterial transcription.

Finally, a panel of biomarkers was established by using sparse partial least squares regression. The biomarkers were refined during the development of a machine learning model to predict response to radiotherapy using bacterial transcription, immune cell abundance and gene expression.

2.2 Ethics

Informed written consent was given for the collection of tissues, and this study adheres to the relevant guidelines and regulations of the Health and Disability Ethics Committee (HDEC) and the University of Otago Human Ethics Committee (ethics approval number: 18/STH/40), Māori consultation took place between Dr Purcell (primary supervisor) and the Māori Research Advisor, Karen Keelan, through the University of Otago, Christchurch Māori consultation process. A letter of support for the project was provided by Karen Keelan.

2.3 Synthetic dataset

The synthetic dataset consisted of RNA-sequencing (RNA-Seq) data from the NCBI read sequencing archive. Bioproject PRJNA588285, SRA: SRR10417449 and Bioproject PRJNA589694, SRA: SRR10445802 were accessed for both a human cell line (TPC-1, human papillary thyroid carcinoma) and *Pseudomonas fluorescens* sequencing reads, respectively. Both projects utilised Illumina 150 bp paired-end RNA-sequencing (RNA-Seq) on the HiSeq X Ten platform. The human reads were concatenated with the *P. fluorescens* reads yielding the synthetic dataset, containing 31,233,071 and 9,590,255 human and bacterial reads (23.49% bacterial and 76.5% human). The synthetic dataset was built by concatenating reads between the two read sets using the cat command.

2.4 CRC cohort

The CRC RNA-seq dataset (CRC-RNA) was taken from the repository of Visnovska et al., SRA: SRP117763, Bioproject PRJNA404030. An Illumina HiSeq 2500 was used to produce 125 bp paired-end RNA-seq reads from 33 CRC patients [344].

2.5 RC cohort

Rectal tumour samples and adjacent normal tissue were collected from 20 RC patients at Christchurch Hospital, New Zealand (CHCH). An additional cohort of 20 matched patient samples came from the Peter MacCallum Cancer Institute of Melbourne (PM), Australia. The PM cohort consisted of RNA samples extracted from patient tumour and adjacent tissue samples and is referred to as the PM cohort. All biopsies were collected prior to the commencement of treatment via colonoscopy or rectoscopy.

2.5.1 Metadata and cohort details

Patient data were collected from the CHCH cohort medical records. All data was anonymised while age, sex, and information on disease status were collected (Table 2.1); however, only post-operative pathology reports were available for the PM cohort.

Table 2.1. Demographics and tumour regression of combined Rectal Cancer (RC) cohort

	Patients (*n*)
Age	
32–86 years	
Mean = 63.6 years	
Sex	
Female	12
Male	28
Differentiation	
Well	2
Moderate	28
Poor	4
NA	6
Histology	
Mucinous	3
Lymphovascular invasion	13
Dworak Score	
Four	5
Three	6
Two	22
One	7

Patient medical records for the CHCH cohort were available (Table 2.2); however, the supplied metadata for the PM cohort was less comprehensive (Table 2.3). Patients received 5-FU based neoadjuvant chemoradiotherapy over the weeks before surgery.

Table 2.2. Christchurch cohort patient details

Patient #	Treatment	RT dose	Dworak	Sex	Age	Tumour Size	Metastasis	Staging	Biopsy date	Differentiation
1	LCCRT (capecitabine)	50.4 gy	Four	M	74	40 mm	N	T3N0M0	14/3/2018	Moderate
2	LCCRT (capecitabine)	50.4 gy	Three	M	64	53 mm	N	T3dN2M0	2/07/2018	Well
3	LCCRT (capecitabine)	50.4 gy	One	M	70	40 mm	N	T3N0M0	3/02/2018	Poor
4	LCCRT (capecitabine)	50.4 gy	Two	M	59	32 mm	Liver	T4aN1M1	30/4/2018	Moderate
5	LCCRT (capecitabine)	50.4 gy	Four	F	78	47 mm	Lung	T3N1M1	6/11/2018	Moderate
6	LCCRT (capecitabine)	50.4 gy	Four	M	63	66 mm	N	T4aN0M0	6/11/2018	Moderate
7	LCCRT (capecitabine)	50.4 gy	Three	M	65	53 mm	N	T3N2M0	19/3/2018	Well
8	LCCRT (capecitabine)	50.4 gy	Two	F	50	50 mm	N	T3bN0M0	23/7/2018	Moderate
9	LCCRT (capecitabine)	50.4 gy	Two	M	67	27 mm	N	T2N0M0	30/7/2018	Moderate
10	LCCRT (capecitabine)	50.4 gy	Two	F	61	45 mm	N	T3bN1M0	8/01/2018	Moderate
11	LCCRT (capecitabine)	50.4 gy	Two	M	75	75 mm	N	T4aN0M0	9/07/2018	NA
12	LCCRT (capecitabine)	50.4 gy	Two	F	66	51 mm	N	T4N1M0	24/9/2018	Moderate
13	LCCRT (capecitabine)	50.4 gy	Four	M	57	54 mm	N	T3N0M0	24/9/2018	NA
14	RT	40 gy	One	M	80	42 mm	N	T3aN0M0	28/9/2018	NA
15	LCCRT (capecitabine)	50.4 gy	Three	M	58	50 mm	N	T3N0M0	7/11/2018	Moderate
16	LCCRT (capecitabine)	50.4 gy	One	F	73	60 mm	N	T3N1M0	13/12/2018	Moderate
17	LCCRT (capecitabine)	50.4 gy	One	M	74	40 mm	N	T2N1M0	4/02/2019	Moderate
18	LCCRT (capecitabine)	50.4 gy	Two	M	66	60 mm	N	T3N2M0	30/5/2019	Moderate
19	LCCRT (capecitabine)	50.4 gy	One	M	76	80 mm	N	T4aN2M0	7/07/2019	Moderate
20	LCCRT (capecitabine)	50.4 gy	Two	F	86	60 mm	N	T4bN2M0	9/08/2019	Moderate

LCCRT: Long course chemoradiation therapy; RT: Radiation therapy; gy: gray unit; N: No; NA: Not Available; M: Male; F: Female

Table 2.3. Peter MacCallum cohort patient details

Patient #	Treatment	RT dose	Dworak	Sex	Age	Differentiation
21	LCCRT (capecitabine)	50.4 gy	Three	M	58	Moderate
22	LCCRT (5-FU)	50.4 gy	Two	M	84	Moderate
23	LCCRT (5-FU)	50.4 gy	Four	M	60	Moderate
24	LCCRT (FOLFIRI)	50.4 gy	Two	M	62	Moderate
25	LCCRT (NA)	50.4 gy	Two	M	39	Poor
26	LCCRT (capecitabine)	50.4 gy	Two	M	62	Moderate
27	LCCRT (capecitabine)	50.4 gy	One	M	44	Moderate
28	LCCRT (capecitabine)	50.4 gy	Two	F	61	Moderate
29	LCCRT (capecitabine)	50.4 gy	One	F	34	Moderate
30	LCCRT (NA)	50.4 gy	Two	F	35	Moderate/Poor
31	LCCRT (5-FU)	50.4 gy	Two	M	65	Moderate
32	LCCRT (FOLFOX)	50.4 gy	Two	M	81	Moderate
33	LCCRT(NA)	50.4 gy	Two	M	85	Moderate
34	LCCRT (capecitabine)	50.4 gy	Two	F	53	Moderate
35	LCCRT (NA)	50.4 gy	Two	M	32	Moderate
36	SCCRT (NA)	25 gy	Two	M	69	NA
37	LCCRT (capecitabine)	50.4 gy	Two	M	75	Moderate
38	LCCRT (FOLFIRI)	50.4 gy	Two	F	53	Moderate
39	LCCRT (NA)	50.4 gy	Three	M	62	NA
40	LCCRT (capecitabine)	50.4 gy	Three	F	69	NA

LCCRT: Long course chemoradiation therapy; SCCRT: Short course chemoradiation RT: Radiation therapy; gy: gray unit; N: No; NA: Not Available; M: Male; F: Female

The majority of patients receiving 50.4 gy of radiation throughout treatment, except for one patient receiving palliative therapy (40 gy) and another receiving short course (25 gy). The age of patients ranged from 32–86 (mean = 63.5, median = 64.5), the cohort was 30% female.

2.6 Nucleic acid extraction and sequencing

2.6.1 Nucleic acid extraction

The CHCH patient samples had both DNA and RNA extracted from tissue samples. The nucleotides were extracted from approximately 20 mg of tissue. Each of the tumour and matched normal biopsies were homogenised in a Precellys Evolution Homogenizer (Bertin Instruments, Montigny-le-Bretonneux, France), using zirconium beads and lysis buffer (Buffer RLT Plus, QIAGEN, Hilden, Germany). DNA and RNA were extracted using a QIAGEN Allprep DNA/RNA Mini Kit. Resulting DNA was quantified using a Qubit 2.0 instrument (Invitrogen, Carlsbad, CA, USA), and the DNA and RNA were quantified and Nanodrop 2000c spectrophotometer (Thermo Scientific, Asheville, NC, USA). The PM patient samples were previously extracted from tissue samples using the same method.

2.6.2 GridION sequencing

Size selection to 400 bp was performed on each of the samples, using a 0.45x volume of MAGBIO HighPrep magnetic beads (Gaithersburg, Maryland, USA). The Oxford Nanopore protocol (RBK_9054_v2_revD_23Jan2018) was followed for DNA sequencing using the SQK_RBK004 rapid kit. For each sample, triplicates of 400 ng genomic DNA were used, each had the volume adjusted to 7.5 µl with nuclease-free water, and 2.5 µl of barcode fragmentation mix added. The samples were incubated in a thermal cycler at 30 °C for 1 minute and 80 °C for 1 minute. The barcoded samples were then pooled, and DNA was purified using AMPure XP beads and resuspended in 10 µl of 10 mM Tris-HCl pH 7.5 with 50 mM NaCl. Then, 1 µl of RAP (Rapid sequencing AdaPtor) was added to the barcoded DNA. The resulting libraries were loaded onto R9.4.1 (106) flow cells in groups of four and sequenced for 48 hrs. Base-calling was carried out using Guppy v3.0.3 (Oxford Nanopore Technology developer access required). Porechop v0.2.3

(https://github.com/rrwick/Porechop) was used for demultiplexing, and barcode and adaptor removal.

2.6.3 16S rRNA sequencing

For each sample, 10 ng of DNA was used to prepare libraries that were sent for 16S rRNA amplicon sequencing by the Massey Genome Service (Massey University, New Zealand). The V3 to V4 regions of the 16S rRNA gene were amplified flanking primers: 16SF_V3: 5′-TATG GTAATTGGCCTACGGGAGGCAGCAG-3′ and 16SR_V4: 5′-AGTCAGTCAGCCGGACTACHVGGGTWTCTAAT -3′). Libraries were prepared using the Illumina MiSeq 500 cycle Kit (V2), and sequencing was performed with PhiX control sequences.

2.6.4 RNA sequencing

RNA-sequencing was performed on the NovoSeq 6000 platform by Novogene (Singapore) using the Illumina V2 library prep. Ribo-ZeroTM Magnetic Kit (Illumina) was used to deplete rRNA from the samples. Unstranded libraries were created using the following primers:
5' Adapter: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3', 3' Adapter: 5'-GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG-3′.

2.7 Bioinformatics

2.7.1 Quality control

Sequencing data were quality controlled with FastQC [345] and trimmed using bbduk2, part of the BBTools suite [346]. For RNA-seq and 16S rRNA amplicon data, only reads with a length < 50bp with a quality of 20 or greater and reads that had matching pairs were kept for further processing while adapters and PhiX sequences were removed. Oxford nanopore reads were filtered for lengths > 400bp using FiltLong v0.2.0 (https://github.com/rrwick/Filtlong).

2.7.2 Host mapping

RNA-Seq and GridION data were mapped to the GRC38p12 human genome using STAR v2.6.1 [271] and Minimap v2.16-r922 [347], respectively. Resulting sam files were converted to bam files, separated into human mapped and unmapped reads, and sorted using samtools v1.9 [348]. For taxonomic assignment analysis, only unmapped reads were retained unless otherwise specified, and additionally for RNA-seq and 16S rRNA, only matched paired-reads were retained. Bam files were converted into fastq files with bedtools v 2.26.0 [349].

2.7.3 Microbiome analysis

Taxonomic assignment was performed using Kraken2 [350] using customised databases. For the taxonomic assignments, a database containing all complete assemblies for bacteria, fungi, protists and archaea, as well as the human genome (GRCh38p12) was used, termed All_DB (database contents can be found in Supplementary Table S2.1). Additionally, the database contained several taxa, regardless of the level of assembly completion, associated with CRC. The bacterial database (Bac_DB) contained only bacterial genomes, including those associated with CRC (database contents can be found in Supplementary Table S2.2). Taxonomic databases were constructed on 19 July 2019. Kraken reports were analysed using Pavian [351] and transformed into biom format using kraken-biom (https://github.com/smdabdoub/kraken-biom). Biom files were analysed using phyloseq v1.32 [316], vegan v2.5 [352], and ape v5.4 [353] R packages, and visualised with ggplot2 v3.3.2 [315] and ggpubr v0.4 packages.

2.7.4 Gene quantification and expression data

Gene quantification data were generated using Salmon. Reads were mapped to the human transcriptome using Salmon v1.2.1 [287]. First, a decoy transcriptome was constructed from the human transcriptome (GRCh38p12) to reduce low confidence mapping of reads to unannotated loci with sequence similarity to annotated regions [354]. An index of the transcriptome was constructed from the transcriptome and the decoy information, using an auxiliary k-mer hash over k-mer length of 31. The index was used for quasi-mapping of RNA-Seq paired-end reads to the human transcriptome with a mapping validation score of 30, based on fragment lengths and their level of direct mapping to a region; reads with lower mapping scores were discarded. Gene count results were visualised with the Integrative Genomic Viewer (IGV) [355] and ggplot2 [315].

2.7.5 Cellular estimation, prediction and subtyping

The ESTIMATE [282] package was used to generate predictive tumour purity, immune and stromal scores from RNA-Seq gene expression data. CIBERSORT [284] was used to estimate the level of immune cell infiltration from RNA-Seq gene expression data.

2.7.6 Functional and differential analysis

Reads assigned to bacterial species using Kraken2 were extracted using the extract_kraken_reads script in the KrakenTools GitHub repository (https://github.com/jenniferlu717/KrakenTools). Extracted reads from assigned bacterial taxonomies were aligned to respective genome assemblies using STAR v2.6.1 [271] without splice aware alignment. Differential analysis was performed using edgeR [356]. Patient identifiers were used as blocking factors in a generalised linear model with likelihood testing.

2.7.7 Correlations

Correlations were performed using Spearman's rank correlation, and point biserial correlation using the cor, cor.test, ltm v1.1 [357], and Hmsic v4.4 R packages, using *p*-value and false discovery rate correction (FDR) cuttoffs of 0.05, while Benjamini and Hochberg [358] was used for false discovery rate (FDR) correction.

2.7.8 Biomarker analysis

For the discovery of prognostic biomarkers, microbial taxonomy, immune cell infiltrates, and gene expression data were used. A predictive model was built to assess the data points most informative of therapeutic outcomes utilising a multi-block discriminant analysis, using the mixOmics package v6.12.2 DIABLO (**D**ata **I**ntegration **A**nalysis for **B**iomarker discovery using a **L**atent c**O**mponents) framework [359]. The model's inputs included data from both the normal and tumour tissue gene expression, immune cell infiltration, and microbiome data, and a pairwise patient blocked study design to mitigate interpersonal variation [360]. The cohort was split into training and test datasets, and leave-one-out validation was used to assess error rates.

2.8 Code availability

All scripts, supplementary files and R code used can be found on the GitHub repository: https://github.com/William-S-Taylor/MSc

3 CHAPTER 3: METHODOLOGY AND PLATFORM COMPARISONS

3.1 Introduction

The gut microbiome and its relationship to human health and disease is an area of increasing research. Microbiome composition has been associated with diarrhoea [361], developmental disorders [362], immune system changes [290], Crohn's disease [363], psychological disorders [364], irritable bowel disease [365], and CRC [366] the latter of which has a high mortality (9.2% of all cancer deaths in 2018) and increasing global incidence (10.2% of all diagnosed cancers in 2018) [367, 368]. Despite this, the assessment of microbiomes in a clinical setting is not widely practised due to the lack of access to sequencing technology and clinical training for interpreting microbiome data [369]. Sampling, library preparation, and sequencing can be expensive and time-consuming, reducing the feasibility of using the microbiome in clinical settings [303, 370].

When performing metagenomic studies using tissue samples without a bacterial selection step, host reads may be misassigned as microbial. To counteract this, reads are first mapped to the host genome, and unmapped reads are then classified using a bacterial taxonomic database [93, 371]. Mapping the sequencing datasets to the host genome before classification is the most time-consuming and computationally intensive step of the taxonomic analysis, particularly with increasingly large datasets and host organism genomes.

This chapter consists of two analyses. First, the efficacy of host genome mapping was investigated using three different datasets: Oxford Nanopore Technology (ONT) data from the CHCH cohort, RNA-Seq data from a CRC cohort from [371], and a synthetic RNA-Seq dataset. Taxonomic assignment was performed with Kraken2 and the taxonomic databases listed in Section 2.6.3, the Bac_DB and All_DB. Secondly, the three platform datasets from the CHCH cohort (16S rRNA, ONT and RNA-Seq) were compared using the methods employed in prior CRC tissue studies [93, 371], with alterations from the results of Analysis 1.

My hypotheses for this chapter were as follows:

- If a taxonomic classifier can be relied on to discriminate between different bacteria, it can also be relied upon to classify the host, thereby reducing the steps required for microbiome analysis of patient samples while giving proportional information of the sample's microbial content.

- The host mapping process may lead to increased type 1 classification errors, as residual host sequences may be misassigned as microbial.
- Appropriately classifying residual host reads will improve inter-platform concordance.

3.2 Methods

Taxonomic assignment was performed using 0.1 confidence scores in Kraken2 v 2.0.7 unless otherwise specified. The CRC-RNA dataset was comprised of CRC patient data (*n* = 33) from a previous study as described in Section 2.4 [344]. The Rec-ONT dataset was comprised of the CHCH cohort samples (*n* = 20) GridION sequencing data. Filtering was performed on the CRC-RNA and Rec-ONT datasets, and to remove any assignments with two or fewer counts. ONT, 16S rRNA and RNA sequencing data were prepared, and quality checked, as described in Sections 2.5 and 2.6. The R packages Pavian v1.0 [351], phyloseq v1.28 [316], VEGAN v 2.5 [352], ggplot2 v3.2.1 [315], stats [372] and Venny v2.1 [373] were used to evaluate and visualise the results. Scripts used for the analysis can be found here: https://github.com/William-S-Taylor/MSc.

3.2.1 Statistics

Statistical operations were performed using R version 3.6. The statistical significance between the differences in databases and mapping methodologies were tested using Wilcoxon signed-rank tests. The beta-diversity was tested using the adonis function in the VEGAN package. Concordance between platforms was investigated using the methodology used in [93] and [371], where Spearman's rank correlation was utilized for comparing the level of concordance between different platforms. The stats package was used to assess the significance of platform concordance.

3.3 Analysis 1: mapping methodology

3.3.1 Synthetic database

To establish the effect of host mapping on taxonomic assignment, a large synthetic dataset with known bacterial and human content was created (the Synth dataset). The Synth dataset consisted of 31 million human and 9.5 million bacterial RNA-Seq paired-end reads, generated using the same Illumina HiSeq X Ten sequencing platform (see Section 2.2). The resulting database was 23.49% bacterial, and 76.5% human (Figure 3.1a). The reads were either mapped or not mapped against the human genome, and subsequently assigned taxonomy using the bacteria only database (Bac_DB) and a comprehensive database containing the human genome (All_DB).

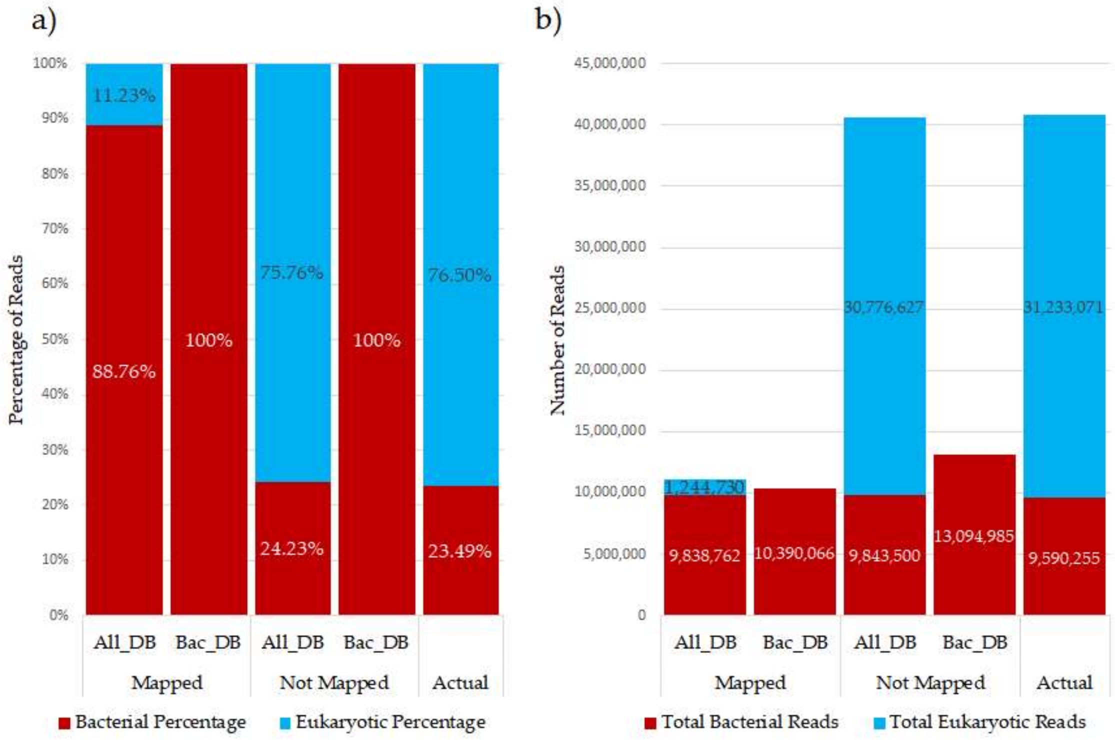

Figure 3.1. Classification of bacterial and human reads in the synthetic dataset. a) percentage of reads assigned, b) total numbers of reads assigned. All_DB, human genome containing database; Bac_DB, non-human genome containing database.

Using the All_DB with prior mapping to the human genome was the most accurate in its taxonomic assignment, assigning 2.52% more reads as bacterial than were present, followed by using the All_DB without prior mapping at 2.57%, while residual human reads not removed by mapping accounted for 11.23% of the post-mapped dataset. Using the All_DB and no prior host mapping, the eukaryotic and bacterial proportions were accurately assigned with a 0.74% difference between the assigned and expected values (Figure 3.1b). Using the Bac_DB, the number of assigned bacterial reads increased compared to using the All_DB by 5.6% and 24.8% with and without mapping, respectively. Using the Bac_DB and host mapping, there were 7.69% more bacterial assignments than existed in the sample, and 26.76% more without host mapping (Figure 3.1b).

Assigning taxonomy with the Bac_DB with prior mapping had a three-fold lower accuracy compared to the All_DB. Based on the findings from the Synth dataset, it was theorized that using the All_DB on other datasets with prior mapping would be the most accurate, deviating by ~2.52% and ~2.57%, with and without host mapping, respectively.

3.3.2 Analysis time

Using the Synth dataset, mapping to the human genome, removing host reads using samtools and converting the resulting bam files into fastq files required 1 hour, 10 min and 10 seconds of compute time, and subsequent taxonomic classification of those unmapped reads with the Bac_DB took 1 min and 15 seconds; totalling 1 hour, 11 minutes and 25 seconds of compute time. Using the All_DB for taxonomic classification with no prior mapping, required 7 min and 27 seconds of compute time. Classification and mapping were carried out on a server with an Intel® Xeon® CPU E5-2683 v4 @ 2.10GHz CPU with 32 cores and 250 GB RAM, running Ubuntu 18.04.3 LTS.

Classification time was higher when using the All_DB and no host mapping; however, when employing host mapping and the Bac_DB, compute time increased significantly, by 9.53-fold (Wilcoxon test *p*-value = 0.01, 95% confidence intervals (CI): 75.94–1441.64).

3.3.3 Clinical Datasets

Using CRC-RNA and Rec-ONT datasets from clinical tissue biopsies, taxonomic assignment was compared using the All_DB and the Bac_DB, with and without host mapping.

Using the CRC-RNA dataset, there were a higher number of bacterial assignments in each sample using the Bac_DB database than with the All_DB (Table 3.1). There was a total bacterial read increase of more than 40% when using the Bac_DB database compared to using the All_DB database, with a difference of more than 41,000 bacterial reads, regardless of prior host mapping (Wilcoxon *p*-value = 0.0001, CI: 718–1702).

Table 3.1. CRC-RNA taxonomic assignment

	Mapped		**Not mapped**
	All_DB	**Bac_DB**	**All_DB**
Bacterial Percentage	13.05%	100%	2.17%
Eukaryotic Percentage	86.95%	0%	97.83%
Total Bacterial Reads	102,450	144,449	102,694
Total Eukaryotic Reads	682,865	0	4,628,480

Using the All_DB, there were 244 more bacterial reads when host mapping was not used; however, this was not statistically significant (Wilcoxon *p*-value = 0.7275, CI: -3.31×10^6–5.65×10^5). The difference in bacterial reads assigned with the All_DB compared to the Bac_DB was over 40,000, and was significant, with (Wilcoxon *p*-value $< 2.2 \times 10^{16}$, CI: 3.99–4.00) or without host mapping (Wilcoxon *p*-value $< 2.2 \times 10^{16}$, CI: 3.99–4.00).

The Rec-ONT DNA dataset had more samples with lower read counts. When using the All_DB, there were 545 more bacterial reads when host mapping was not used (Figure 3.2); however, the difference was not statistically significant (Wilcoxon *p*-value = 0.9132, CI: -3.83 × 10^5–4.39 × 10^5).

Table 3.2. Rec-ONT dataset taxonomic assignment

	Mapped		**Not Mapped**
	All_DB	**Bac_DB**	**All_DB**
Bacterial Percentage	3.74%	100%	0.25%
Eukaryotic Percentage	96.26%	0%	99.75%
Total Bacterial Reads	34,159	253,172	34,704
Total Eukaryotic Reads	879,766	0	13,818,571

The difference in bacterial reads assigned with the All_DB compared to the Bac_DB was greater than 218,000, and was significant, with (Wilcoxon *p*-value < 2.2 × 10^{16}, CI: 8.00–8.99) or without host mapping (Wilcoxon *p*-value < 2.2 × 10^{16}, CI: 7.99–8.99). When assigning taxonomy to the RNA-CRC and Rec-ONT datasets with the All_DB, the number of bacterial reads differed by 244 (0.24%), and 545 (1.57%), when prior-host mapping was and was not performed, respectively. Additionally, reads assigned to Eukaryotes provided information on the samples' relative host content and the proportion of residual reads.

3.3.3.1 *Alpha and beta diversity*

The bacterial diversity of the CRC-RNA and Rec-ONT datasets was assessed to investigate the impact of host mapping. All reads not assigned bacterial taxonomy, or taxa with fewer than two reads assigned were removed before analysis.

In CRC, the distribution of microbiota in the colon differs depending on location [374]. Using the CRC-RNA dataset, diversity was compared between left- and right-sided tumours using observed diversity (richness of the sample), and the Shannon and Simpson diversity indexes (measures of evenness and dominance, respectively).

When using the All_DB, alpha diversity did not change significantly regardless of host mapping (Figure 3.2). When using the Bac_DB, alpha diversity was higher compared to using the All_DB, with more than double the observed mean diversity when using prior host mapping.

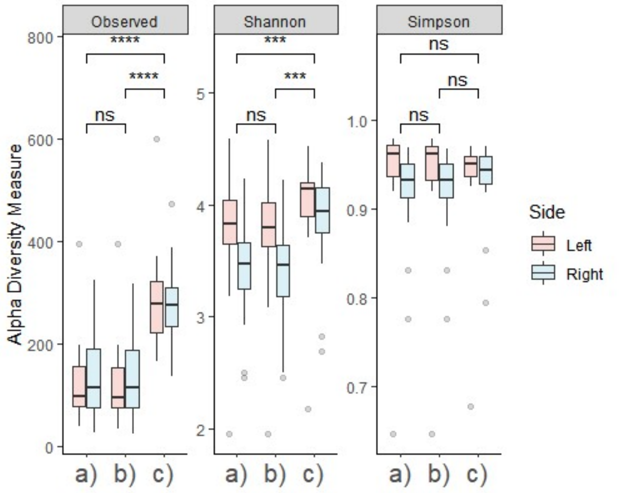

*Figure 3.2. Alpha diversity measures for comparison of left- and right-sided tumours in the CRC-RNA dataset: a) using the All_DB and no prior host mapping, b) using the All_DB and prior host mapping, and c) using the Bac_DB and prior host mapping. Outliers are represented in grey. ns: not significant; ***: p <= 0.001; ****: p <= 0.0001.*

There was no statistically significant difference in the mean microbial diversity when using the All_DB regardless of prior host mapping (Table 3.3). Differences in mean diversity were statistically significant when using the All_DB compared to the Bac_DB, except for the Simpson diversity index.

Table 3.3. CRC-RNA diversity Wilcoxon signed-rank test results

	Not Mapped All_DB vs Mapped Bac_DB		**Mapped All_DB vs Mapped Bac_DB**		**Mapped vs not Mapped All_DB**	
	***p*-value**	**95% CI**	***p*-value**	**95% CI**	***p*-value**	**95% CI**
Observed	2.11×10^{9}	129–189	2.39×10^{9}	127–186	0.717	-27–23

Shannon	6.30×10^{5}	0.231–0.639	1.50×10^{4}	0.21–0.611	0.674	-0.258–0.205
Simpson	0.382	-0.007–0.022	0.500	-0.008–0.02	0.704	-0.017–0.015

95% CI: 95% confidence interval.

When comparing the alpha diversity differences between left- and right-sided tumours within the CRC-RNA dataset, no measure was statistically significant (Figure 3.3).

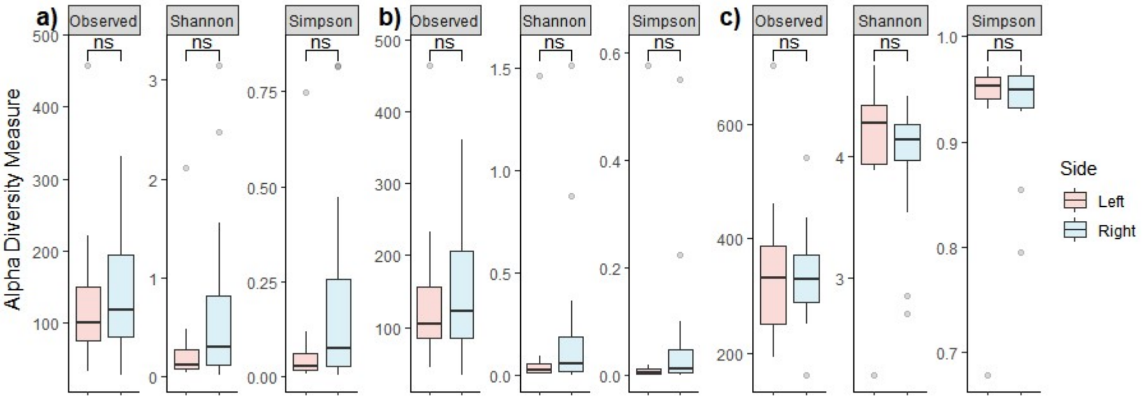

Figure 3.3. Alpha diversity measures for comparison of left and right side tumours within the CRC-RNA dataset: a) using the All_DB and no prior host mapping, b) using the All_DB and prior host mapping, and c) using the Bac_DB and prior host mapping. Outliers are represented in grey. ns: not significant.

Using the Rec-ONT dataset, the impact of mapping on the diversity between tumour and matched normal tissue microbiomes was investigated.

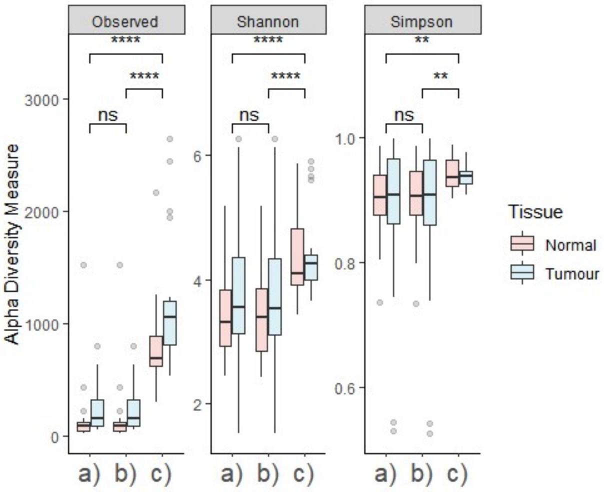

*Figure 3.4. Alpha diversity measures for comparison of tumour and normal tissues in the Rec-ONT dataset: a) using the All_DB and no prior host mapping, b) using the All_DB and prior host mapping, and c) using the Bac_DB and prior host mapping. Outliers are represented in grey. ns: not significant; **: p <= 0.001; ***: p <= 0.001; ****: p <= 0.0001.*

Host mapping when using the All_DB had little impact; however, there was a substantial difference between using the All_DB and Bac_DB (Figure 3.4). There was no statistically significant difference in any diversity measure if prior host mapping was or was not performed when using the All_DB; however, the differences between using the All_DB and Bac_DB were statistically significant (Table 3.4).

Table 3.4. Rec-ONT diversity Wilcoxon signed-rank test results

	Not Mapped All_DB vs Mapped Bac_DB		**Mapped All_DB vs Mapped Bac_DB**		**Mapped vs not Mapped All_DB**	
	***p*-value**	**95% CI**	***p*-value**	**95% CI**	***p*-value**	**95% CI**
Observed	9.13×10^{14}	635–849	9.80×10^{14}	635–848	0.873	-27–25
Shannon	9.74×10^{9}	0.764–1.473	1.12×10^{8}	0.793–1.44	0.969	-0.415–0.390
Simpson	0.0003	0.025–0.0836	0.0001	0.0255–0.0813	0.892	-0.035–-0.035

When comparing the alpha diversity of tumour and normal tissues, only observed diversity differences were statistically significant ($p < 0.05$), regardless of mapping or the database used (Figure 3.5). However, the *p*-value was six-fold lower when using the Bac_DB.

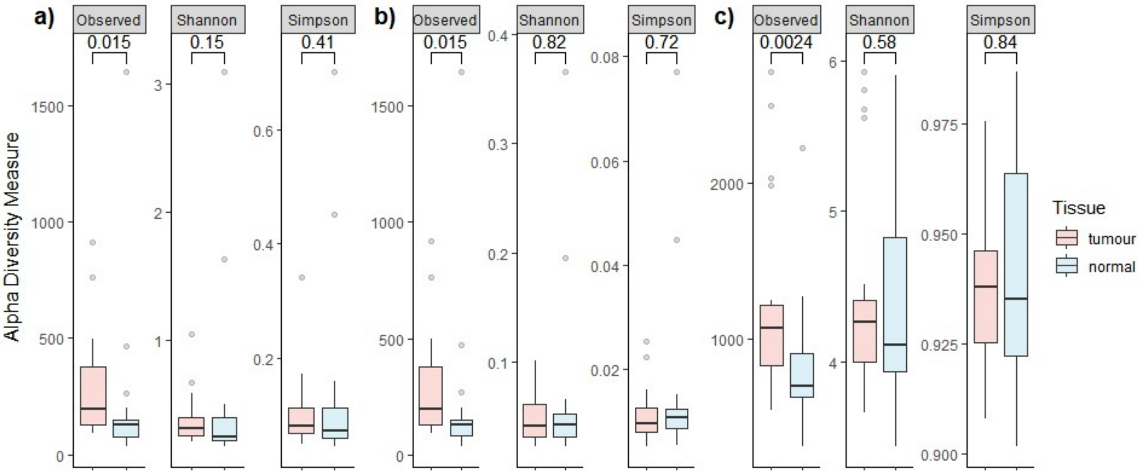

Figure 3.5. Alpha diversity measures for comparison of tumour and normal tissue within the Rec-ONT dataset: a) using the All_DB and no prior host mapping, b) using the All_DB and prior host mapping, and c) using the Bac_DB and prior host mapping. P-values < 0.05 are considered significant.

Beta-diversity was investigated using non-metric multi-dimensional scaling (NMDS) of Bray–Curtis distances for the CRC-RNA and Rec-ONT datasets. The side and tissue effect on the clustering of samples was tested using the adonis2 function in the VEGAN package.

When using the All_DB, the impact of host mapping on the clustering of left- and right-sided tumours in CRC-RNA dataset was insignificant (Figure 3.6a,b); however, using the Bac_DB resulted in a tight clustering of left-sided tumours (Figure 3.6c).

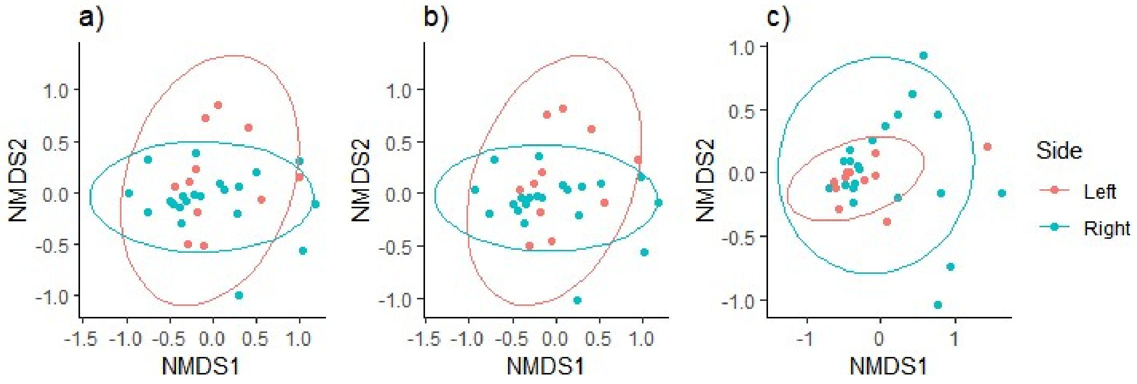

Figure 3.6. Bray–Curtis clustering of left- and right-sided tumours in the CRC-RNA dataset: a) using the All_DB and no prior host mapping, b) using the All_DB and prior host mapping, and c) using the Bac_DB and prior host mapping.

The effect of side on the clustering of samples was statistically significant when using the All_DB and not when using the Bac_DB while the R2 value was similar regardless of the database used (Table 3.5).

Table 3.5. Effect of side on CRC-RNA dataset

Side	All_DB mapped	All_DB not mapped	Bac_DB mapped
R2	0.049	0.050	0.052
Residuals	0.950	0.949	0.947
***p*-value**	0.037	0.043	0.085

When using the Rec-ONT dataset, the impact of host mapping on the clustering of samples by tissue type was minor; however, tumour samples were more closely clustered than normal samples when using the Bac_DB and host mapping, which also resulted in substantive outliers (Figure 3.7).

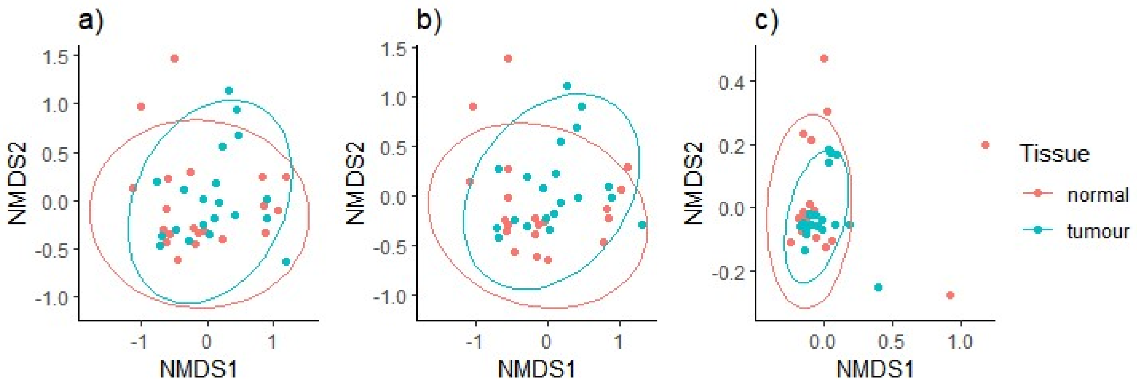

Figure 3.7. Bray–Curtis clustering of normal and tumour tissues in the Rec-ONT dataset: a) using the All_DB and no prior host mapping, b) using the All_DB and prior host mapping, and c) using the Bac_DB and prior host mapping.

Although the effect was slightly higher using the Bac_DB, the effect of tissue type was not statistically significant in the Rec-ONT dataset (Table 3.6).

Table 3.6. Effect of tissue in Rec-ONT dataset

Tissue	**All_DB mapped**	**All_DB not mapped**	**Bac_DB mapped**
R2	0.0241	0.024	0.023
Residuals	0.975	0.975	0.976
***p*-value**	0.482	0.391	0.491

3.3.3.2 *Sample taxa compositions*

There was only a 0.1% difference in phyla assignment in the CRC-RNA dataset between mapped and unmapped samples when using the All_DB. The difference between using the Bac_DB and the All_DB with host mapping was substantial; there were 12.7% fewer Bacteroidetes, 2.3% fewer Fusobacteria, 2.6% fewer Firmicutes, and 15.6% more Proteobacteria when the Bac_DB was used (Table 3.7).

Table 3.7. CRC-RNA dataset total phyla composition

Phyla	**All_DB Mapped**	**All_DB No Mapping**	**Bac_DB Mapped**

Bacteroidetes	59.5%	59.4%	46.7%
Proteobacteria	4.4%	4.5%	20.1%
Firmicutes	26.7%	26.7%	24.1%
Fusobacteria	8.0%	8.0%	6.3%

The genus-level differences between using the All_DB with or without host mapping were minor and did not exceed 0.1% between the top 15 genera (Table 3.8). Of the 15 top genera when using the Bac_DB, six did not appear when using the All_DB. *Klebsiella* and *Pasteurella* were the third and second most abundant when using the Bac_DB, respectively, along with other genera such as *Bacillus*, *Staphylococcus*, *Enterobacter* and *Ralstonia*. Additionally, *Bacteroides*, *Faecalibacterium* and *Fusobacterium* were detected in lower proportions when using the Bac_DB. Those genera detected within the 15 most abundant when using the All_DB were *Hungatella*, *Campylobacter*, *Eubacterium*, *Lachnoanaerobaculum*, *Leptotrichia* and *Alistipes*.

Table 3.8. CRC-RNA top 15 genera

All_DB Mapped		**All_DB Unmapped**		**Bac_DB Mapped**	
Genus	**%**	**Genus**	**%**	**Genus**	**%**
Bacteroides	57.0	*Bacteroides*	56.9	*Bacteroides*	44.0
Fusobacterium	8.7	*Fusobacterium*	8.7	*Fusobacterium*	6.9
Prevotella	3.9	*Prevotella*	3.8	*Klebsiella*	4.3
Porphyromonas	3.2	*Porphyromonas*	3.1	*Pasteurella*	3.3
Faecalibacterium	3.1	*Faecalibacterium*	3.1	*Lachnoclostridium*	3.3
Hungatella	2.5	*Hungatella*	2.5	*Porphyromonas*	2.8
Roseburia	2.5	*Roseburia*	2.5	*Prevotella*	2.6
Lachnoclostridium	1.9	*Clostridium*	2.0	*Faecalibacterium*	2.5
Clostridium	1.9	*Lachnoclostridium*	1.9	*Staphylococcus*	2.3
Blautia	0.9	*Blautia*	0.9	*Clostridium*	1.9

Campylobacter	0.8	*Campylobacter*	0.8	*Enterobacter*	1.8
Eubacterium	0.7	*Eubacterium*	0.6	*Ralstonia*	1.4
Lachnoanaerobaculum	0.6	*Lachnoanaerobaculum*	0.6	*Roseburia*	1.3
Leptotrichia	0.6	*Leptotrichia*	0.6	*Bacillus*	0.9
Alistipes	0.5	*Alistipes*	0.5	*Blautia*	0.7

The Rec-ONT dataset differences at the phylum level when using the All_DB were minor, regardless of host mapping (Table 3.9). However, compared to using the Bac_DB, the differences were more considerable; Bacteroidetes were 30% higher while Proteobacteria and Fusobacteria were 30% and 1.5% lower, respectively. Firmicutes were detected at similar proportions regardless of database and host mapping. Using the Bac_DB, the differences were the highest for Proteobacteria and Bacteroidetes, with the former comprising >50% of the sample totals, and the latter less than 10%.

Table 3.9. Rec-ONT total phyla

Phyla	All_DB Mapped	All_DB No Mapping	Bac_DB Mapped
Bacteroidetes	40.1%	39.6%	9.6%
Proteobacteria	26.3%	26.9%	57.6%
Firmicutes	23.5%	23.4%	23.4%
Fusobacteria	2.4%	2.4%	0.9%

At the genera level, the differences between host mapping and no host mapping using the All_DB were less than 1% for the three most abundant genera; however, *Salmonella* and *Campylobacter* were absent with and without mapping, respectively (Table 3.10). *Porphyromonas* and *Fusobacterium* abundance were most impacted by host mapping, which increased by 1.2% when host mapping was used with the All_DB.

Staphylococcus, Pasteurella, Klebsiella, Candidatus Portiera, Ralstonia, Yersinia, Enterobacter, Mycoplasma and *Burkholderia* appeared in the top 15 most abundant genera when the Bac_DB was used (Table 3.10). *Bacteroides* was the most abundant genus when using the All_DB at >26% of sample totals; however, using the Bac_DB resulted in the abundance of *Bacteroides* being detected at less than 4% of sample totals.

Table 3.10. Rec-ONT top 15 genera

All_DB Mapped		**All_DB Unmapped**		**Bac_DB Mapped**	
Genus	**%**	**Genus**	**%**	**Genus**	**%**
Bacteroides	29.1	*Bacteroides*	26.7	*Staphylococcus*	13.6
Escherichia	9.8	*Escherichia*	9.8	*Pasteurella*	13.3
Porphyromonas	6.7	*Porphyromonas*	5.5	*Klebsiella*	8.2
Faecalibacterium	3.1	*Faecalibacterium*	3.3	*Candidatus Portiera*	5.7
Fusobacterium	2.6	*Alistipes*	2.4	*Escherichia*	5.6
Alistipes	2.2	*Hungatella*	2.1	*Bacteroides*	3.6
Hungatella	2.0	*Oscillibacter*	1.9	*Ralstonia*	3.5
Oscillibacter	1.6	*Pseudomonas*	1.6	*Yersinia*	3.3
Prevotella	1.3	*Fusobacterium*	1.4	*Enterobacter*	3.2
Campylobacter	1.3	*Prevotella*	1.4	*Mycoplasma*	2.4
Lachnoclostridium	1.3	*Streptomyces*	1.4	*Bacillus*	2.0
Clostridium	1.3	*Salmonella*	1.4	*Clostridium*	1.9
Streptomyces	1.2	*Clostridium*	1.4	*Burkholderia*	1.2
Pseudomonas	1.2	*Bacillus*	1.3	*Streptomyces*	0.9
Bacillus	1.1	*Lachnoclostridium*	1.2	*Porphyromonas*	0.9

3.4 Analysis 2: RC dataset platform comparison

Once the appropriate methodology had been established for microbial taxonomic assignment of metagenomic sequencing reads, the comparative methodology from Section 1 was applied to the Christchurch (CHCH) cohort.

The dataset contained sequencing data from the same three platform datasets, described in Section 2.6 (ONT, 16S rRNA and RNA-Seq). The cohort was comprised of 20 patients, with two samples per patient (tumour and adjacent non-malignant tissue).

The taxonomy of each patient sample and the cross-platform correlations were carried out using the same methodology as in [93, 371]; with changes based on the results of Analysis 1; utilising a broad taxonomic database containing the host genome, and using the same taxonomic database and assignment software for all platform datasets.

3.4.1 Data processing and information

3.4.1.1 Quality control and read counts

16S rRNA read length ranged from 52–247 bp, with per-sequence Phred scores of 28–38 (Figure 3.8a) and per-base mean Phred scores of 33.8–38.8 (Figure 3.8b).

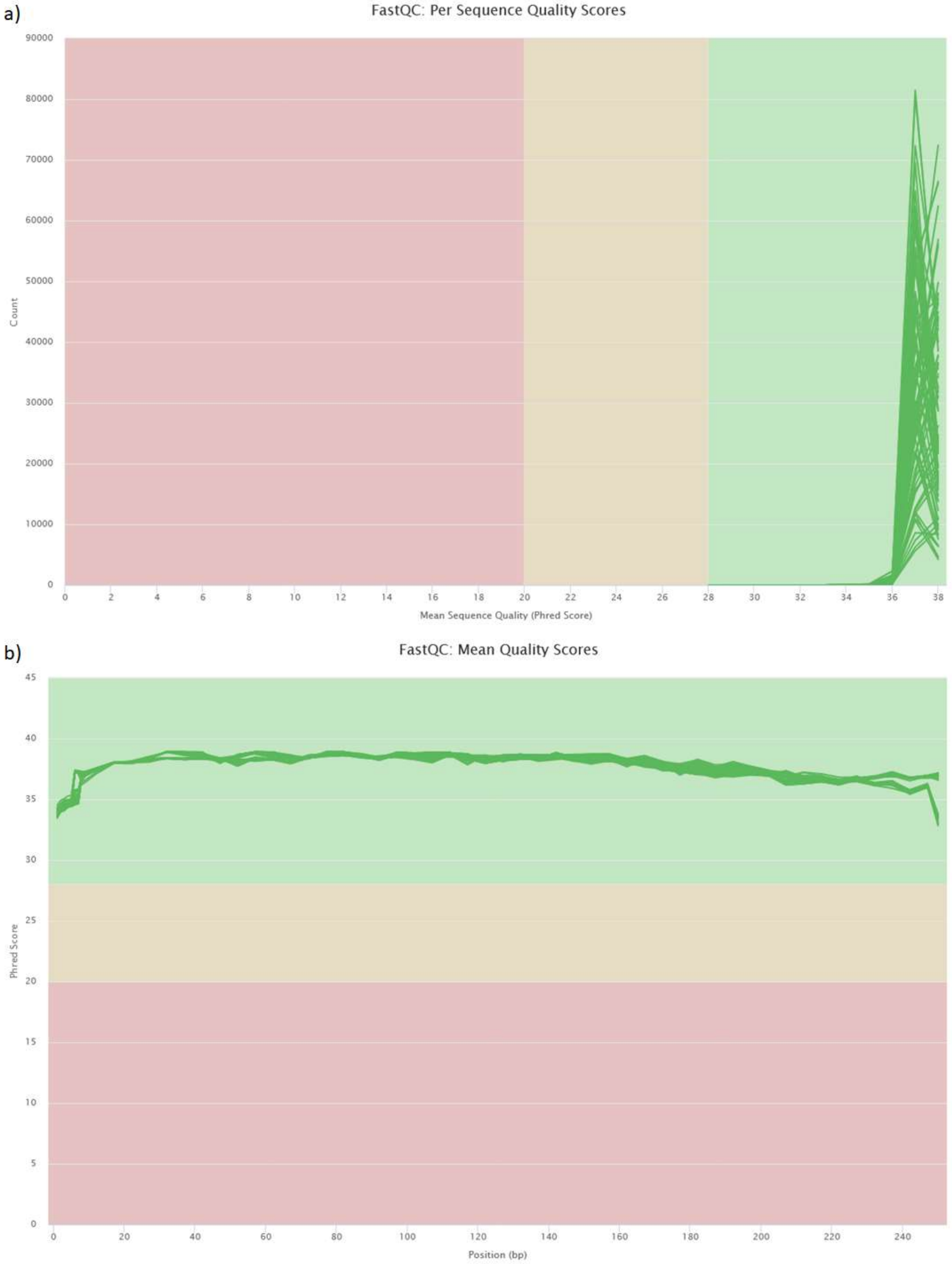

Figure 3.8. 16S rRNA quality scores. a) Per-sequence quality scores, b) mean quality scores

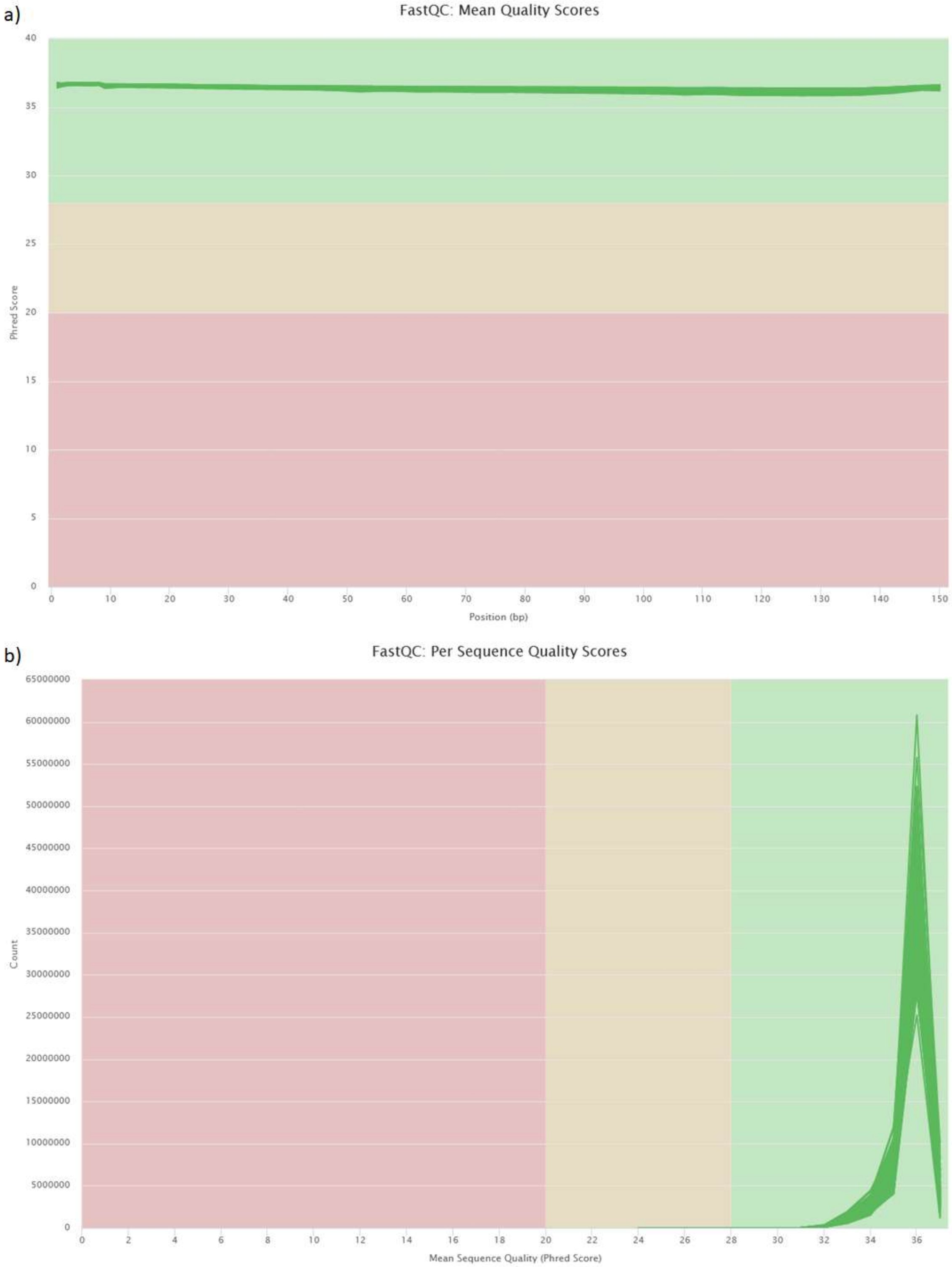

Figure 3.9. RNA-Seq quality scores. a) Per-sequence quality scores, b) mean quality scores

RNA-Seq reads ranged in length from 52bp–150 bp, with per-sequence Phred scores of 24–37 (Figure 3.9a), and per-base mean Phred scores of 36.1–36.7 (Figure 3.9b).

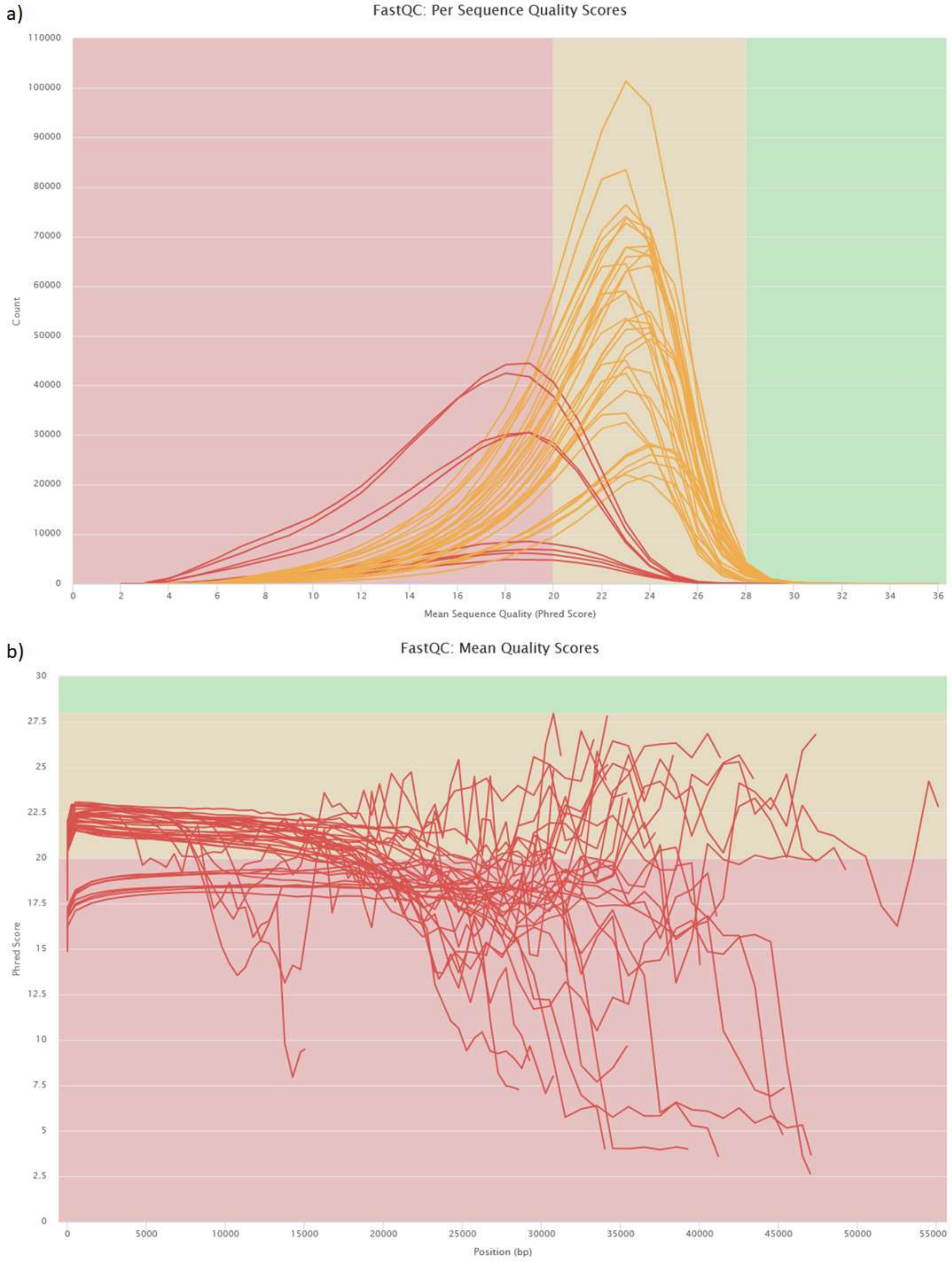

Figure 3.10. ONT quality scores. a) Per-sequence quality scores, b) mean quality scores.

ONT reads had a mean length of 2600, ranging from 499–54,595 bp, with per-sequence Phred scores of 2–36 (Figure 3.10a), and per-base mean Phred scores of 2.6–27.8 (Figure 3.10b). The mean number of raw reads per-sample for 16S rRNA, RNA-Seq and ONT were 75,137, 56,672,929 and 358,996, respectively (Table 3.11).

Table 3.11. Number of reads per-sample

	Raw Reads			**Processed/Unmapped**			**Assigned Bacterial Taxonomy**		
Sample	**16S**	**RNA**	**ONT**	**16S**	**RNA**	**ONT**	**16S**	**RNA**	**ONT**
RT1N	47801	65797848	460355	41733	759910	29235	41645	130487	23
RT1T	72039	50629572	316304	63027	381566	20635	62917	53451	102
RT2N	64573	58393905	420266	56880	329306	24440	56825	39096	6
RT2T	65109	67848414	379022	58546	290747	25578	55826	70508	6
RT3N	146127	58274719	240536	121577	2340228	27625	121410	1653462	2889
RT3T	51603	68019961	471354	44482	3032153	29651	44417	2330350	565
RT4N	103582	55679261	318958	85448	357961	20016	85336	32318	195
RT4T	148125	53838505	174438	121740	5063581	17674	121510	4923353	137
RT5N	74483	69511002	223418	63375	392221	18646	63324	61397	194
RT5T	102337	55610236	405361	89884	454105	23709	89837	183595	298
RT6N	73255	44112606	355170	64091	311903	20377	62426	60026	12
RT6T	85457	64679678	480529	70208	1177014	29912	70186	923117	2732
RT7N	71763	44370270	151240	62667	243397	7885	62610	30539	5
RT7T	33824	75592926	488667	29413	961405	20486	29276	352667	90
RT8N	92311	45567364	217400	79901	289829	14187	79761	51956	11
RT8T	35565	76497189	577258	31486	4033741	27791	31441	3717615	82
RT9N	83232	42691787	487433	72977	327650	25694	72893	46956	11
RT9T	16849	70245212	376484	15013	470682	23091	14985	19162	46
RT10N	44944	48311893	362386	38256	292579	20410	38101	21634	2
RT10T	42297	55048172	483918	37831	468270	24006	37747	64878	42

RT11N	97007	74481282	192418	82776	557978	17994	82757	168629	1577
RT11T	43566	51185511	315209	36290	779596	35373	36253	316816	723
RT12N	21696	53866573	201183	17493	518465	13572	17182	75546	30
RT12T	18947	51619614	551584	16519	341611	32887	16478	36449	59
RT13N	37515	42435289	229239	32781	319737	20156	32715	48729	4
RT13T	57997	74446240	470690	50836	508365	25345	50751	57156	32
RT14N	82895	41915825	562818	70257	363086	29696	70213	44746	618
RT14T	132916	48035590	430829	116631	370048	28358	116350	62063	464
RT15N	85852	51753938	284966	73510	241447	17462	73377	35652	179
RT15T	105534	45461331	592632	91832	364479	34997	91699	44542	272
RT16N	99157	52662204	749205	85535	416219	43843	85463	60039	992
RT16T	93180	53381132	590318	79794	283392	28118	79741	34855	89
RT17N	79290	50559600	94358	70685	200144	21259	70490	11568	6
RT17T	87192	56401397	308720	76826	416337	102428	76742	67920	20
RT18N	99273	45275146	75278	89013	664194	15774	88949	310182	2
RT18T	110449	52601165	293778	92579	313101	109087	92420	78856	80
RT19N	64078	47015510	55087	55196	214147	12505	55149	38604	28
RT19T	91902	68820033	447806	78645	580995	56856	78539	126030	65
RT20N	120428	69006835	70049	103244	590887	16026	103175	298105	59
RT20T	21369	65272435	453176	18738	627297	68762	18714	173641	46
Mean	75137	56672929	358996	64692	766244	29538	64490	421417	319
Median	76886	53852539	369435	67149	404220	24223	66755	61730	62
Std	33762	10467244	164660	28509	1037031	20958	28508	1020002	659

RT: Rectal tissue; N: normal tissue; T: tumour tissue; 16S: 16S rRNA sequencing; RNA: RNA-Seq; ONT: Oxford Nanopore Technology sequencing.

After mapping and processing, the taxonomy assignment rate was highest for 16S rRNA, at 85.8% of raw-reads, while RNA-Seq and ONT reads were assigned taxonomy at magnitude lower rates, at 0.007% and 0.001%, respectively (Table 3.12). An amplification bias was found in the 16S rRNA dataset at the genus level, resulting in excessive levels of the *Burkholderia* genus, which was not reflected in the other platforms, and were removed before relative abundance calculations.

Table 3.12. Percentage of reads retained and assigned taxonomy

%	Post-processing	Assigned taxonomy	
		Of total	Of post-processed
16S rRNA	86.1%	85.8%	99.7%
RNA-Seq	1.4%	0.07%	55%
ONT	8.2%	0.01%	1.1%

3.4.2 Platform concordance

The correlation between taxa at the phylum level between platforms was found to be highest between RNA and 16S rRNA sequencing, and lowest was between ONT and RNA sequencing (Figure 3.11), with one sample being negatively correlated; however, this was not statistically significant (r_s = -0.067, *p*-value = 0.8987). The mean correlation between 16S rRNA and ONT sequencing was higher than between RNA and ONT sequencing (Table 3.13).

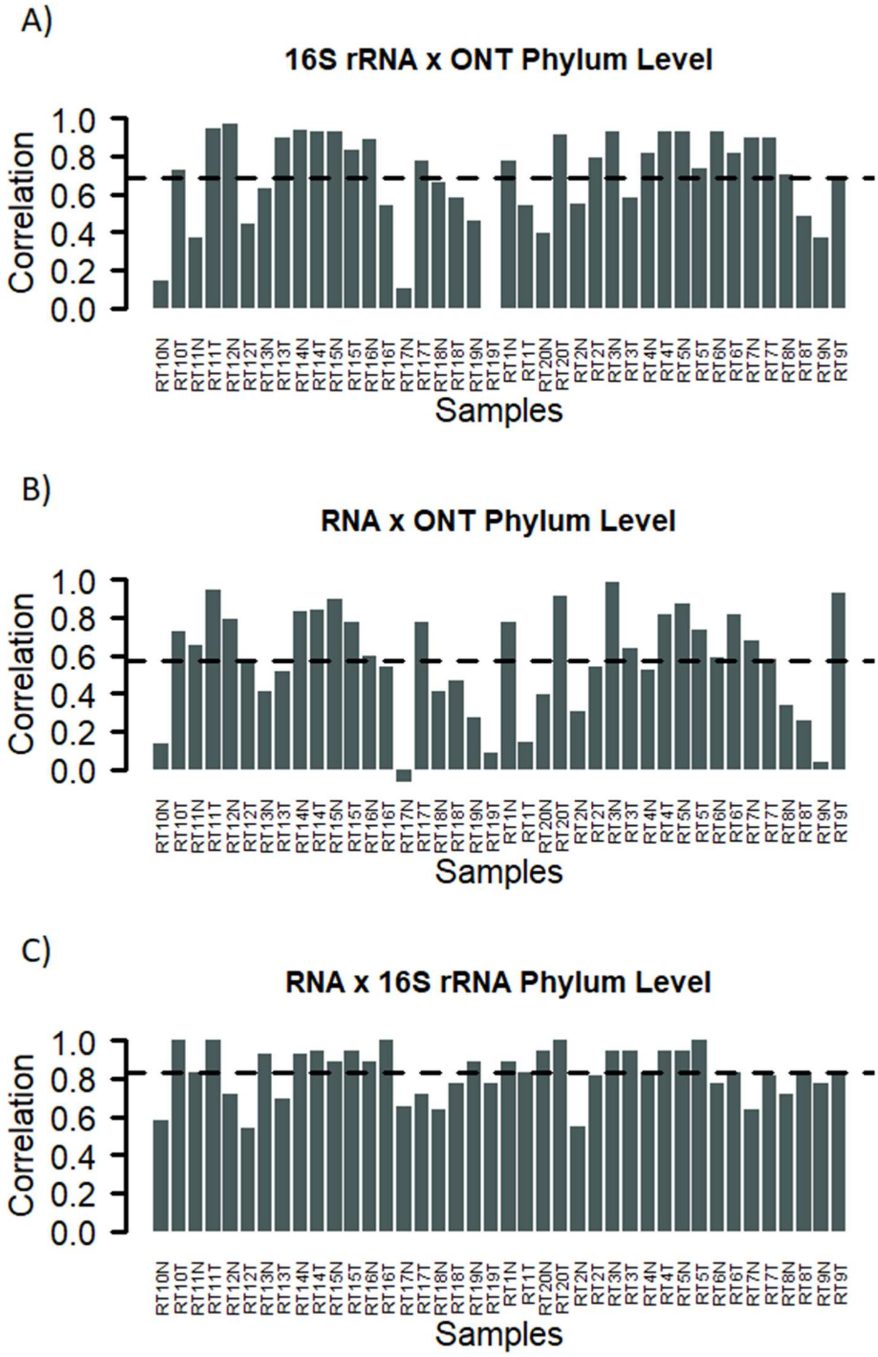

Figure 3.11. Platform comparisons of the RC cohort at the phylum level, A) 16S rRNA vs ONT, B) RNA-Seq vs ONT and C) RNA-Seq vs 16S rRNA. The dashed line indicates the sample mean.

At the genus level (Figure 3.12), there was a more consistent concordance level than at the phylum level; however, the mean correlation ranged from 0.48–0.52 for ONT and 16S rRNA comparisons with RNA-Seq, respectively (Table 3.13).

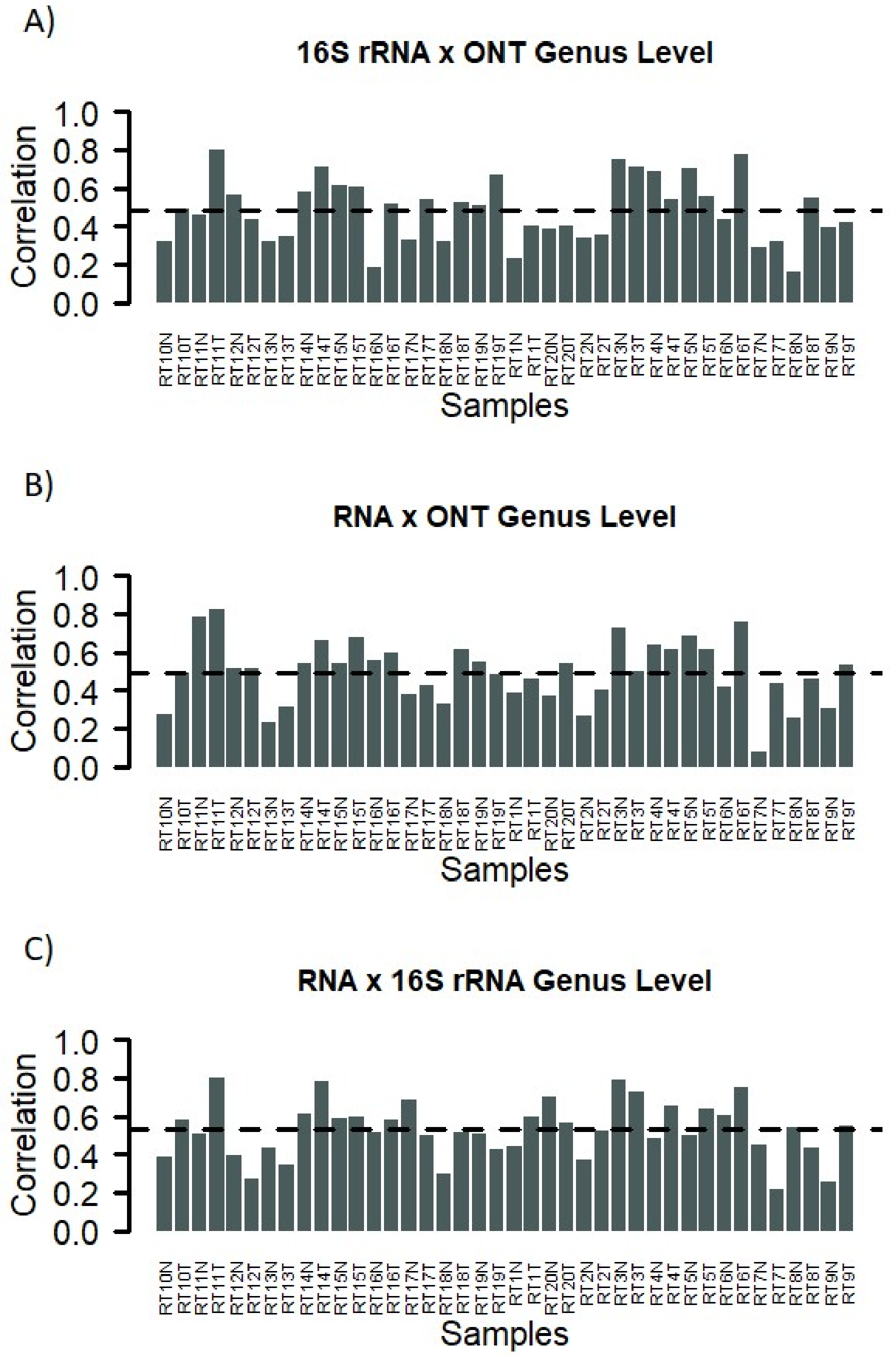

Figure 3.12. Platform comparisons of the RC cohort at the genera level, A) 16S rRNA vs ONT, B) RNA-Seq vs ONT and C) RNA-Seq vs 16S rRNA. The dashed line indicates the sample mean.

At the species level, the concordance was higher than at the genus level (Figure 3.13), with both 16S rRNA and RNA sequencing had a mean correlation with ONT sequencing of 0.55 (Table 3.13), and between 16S rRNA and RNA-Seq, it was 0.62. One sample (RT13N) had a non-statistically significant negative correlation when comparing RNA and ONT sequencing ($\varrho = -0.06$, *p*-value = 0.4375 (Figure 3.13B).

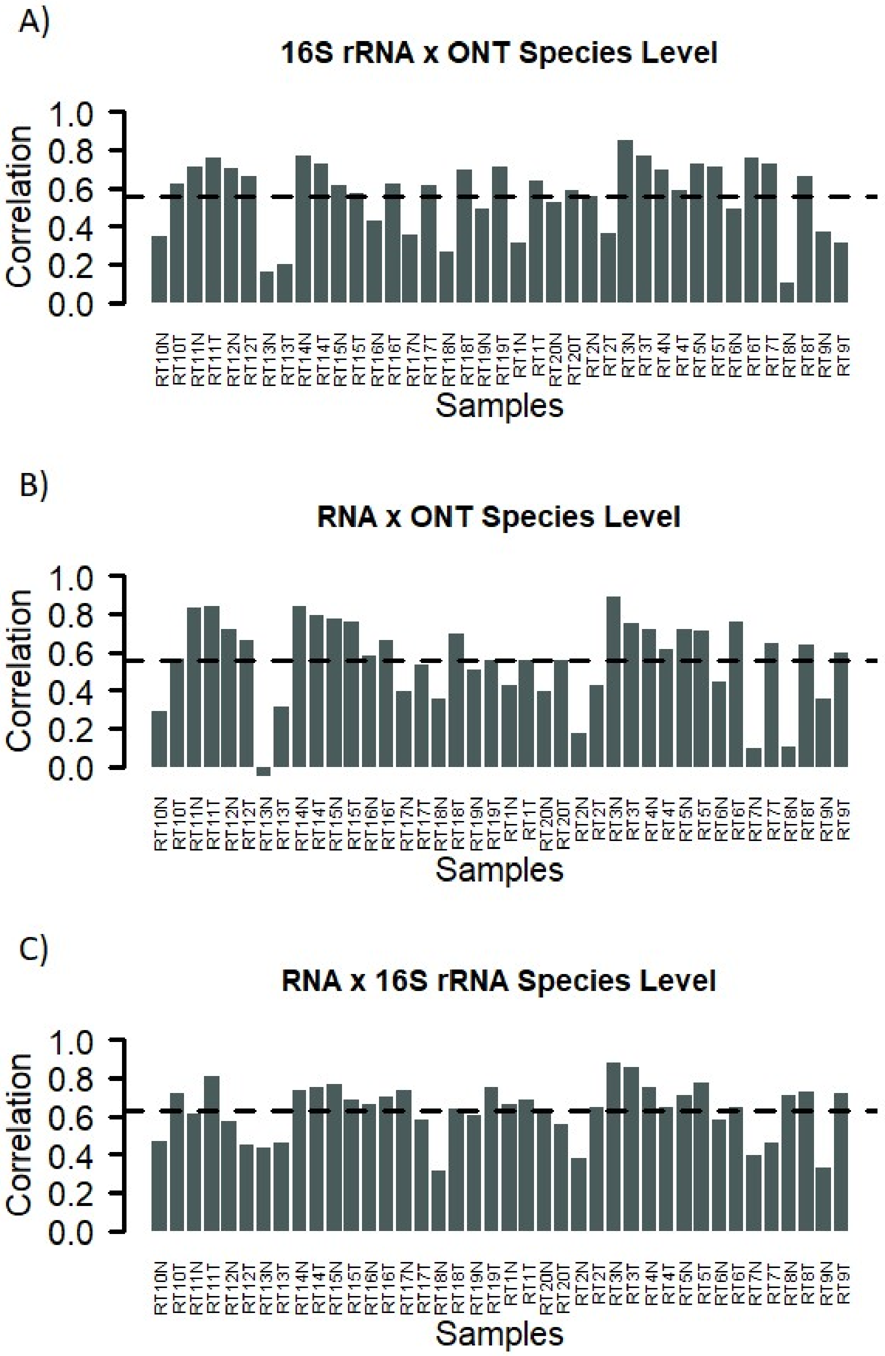

Figure 3.13. Platform comparisons of the RC cohort at the species level, A) 16S rRNA vs ONT, B) RNA-Seq vs ONT and C) RNA-Seq vs 16S rRNA. The dashed line indicates the sample mean.

Table 3.13. Mean correlation between platforms and rectal dataset

Taxonomic Level	**16S rRNA vs. ONT**		**16S rRNA vs. RNA**		**ONT vs. RNA**	
	ϱ	***p*-value**	ϱ	***p*-value**	ϱ	***p*-value**
Phylum	0.68	$<2.2 \times 10^{16}$	0.83	$<2.2 \times 10^{16}$	0.58	$<2.2 \times 10^{16}$
Genus	0.49	$<2.2 \times 10^{16}$	0.53	$<2.2 \times 10^{16}$	0.48	$<2.2 \times 10^{16}$
Species	0.56	$<2.2 \times 10^{16}$	0.63	$<2.2 \times 10^{16}$	0.55	$<2.2 \times 10^{16}$

In contrast to our published CRC study [371], the overall concordance was higher between 16S rRNA and both ONT and RNA sequencing at each taxonomic level (Table 3.14), with the most considerable improvement being at the species level. The concordance between ONT and RNA-Seq decreased at the phyla and genera levels; however, there was an improvement seen at the species level. Across all taxonomic levels, the highest concordance was seen between 16S rRNA and RNA-Seq.

Table 3.14. Changes in mean correlation between platforms compared to Taylor et al. 2020

Taxonomic Level	**16S rRNA vs. ONT**	**16S rRNA vs. RNA-Seq**	**ONT vs. RNA-Seq**
Phyla	+0.009	+0.023	-0.092
Genus	+0.135	+0.161	-0.035
Species	+0.363	+0.440	+0.204

Each platform's taxa identification rate was investigated in terms of the number of raw-reads per species identified (Table 3.15). Read efficiency was evaluated without taxonomic filtering, as filtering cut-offs are often arbitrary, and the same cut-off would not be applicable cross-platform.

Table 3.15. Number of different bacterial taxa detected using each sequencing platform

	RNA-Seq	**16S rRNA**	**ONT**
Phyla detected	33	21	7

Genera detected	900	314	70
Species detected	2512	459	129
Unique phyla	13	1	0
Unique genera	621	39	0
Unique species	2111	95	0
Raw reads per-species	902,435	6548	111,316

For every 900,000 reads, RNA-Seq could identify a species, while 16S rRNA was more efficient as reads were amplified from bacterial specific genes and were with 6548 reads per species identified, despite 16S rRNA data being known to be less reliable at distinguishing between taxa at the species level. ONT data was the least efficient, with substantially fewer taxa being detected overall, and even fewer being detected on a per-raw read basis. Furthermore, no unique taxa were detected when compared to other platforms (Table 3.15).

3.4.3 Platform composition

Figure 3.14 shows the relative bacterial abundance in the CHCH cohort using each platform. The proportions of each phylum varied between platforms (Figure 3.14A), with RNA-Seq and 16S rRNA sequencing having more comparable levels of Bacteroidetes and Proteobacteria; ONT sequencing detected a comparatively higher proportion of Bacteroidetes and Firmicutes.

The proportional differences between each platform at the genera level were smaller, with each displaying more consistent levels of each taxon (Figure 3.14B); however, ONT sequencing detected fewer genera overall. In particular, ONT sequencing detected fewer *Prevotella, Campylobacter* and *Streptococcus*, compared to 16S rRNA and RNA sequencing.

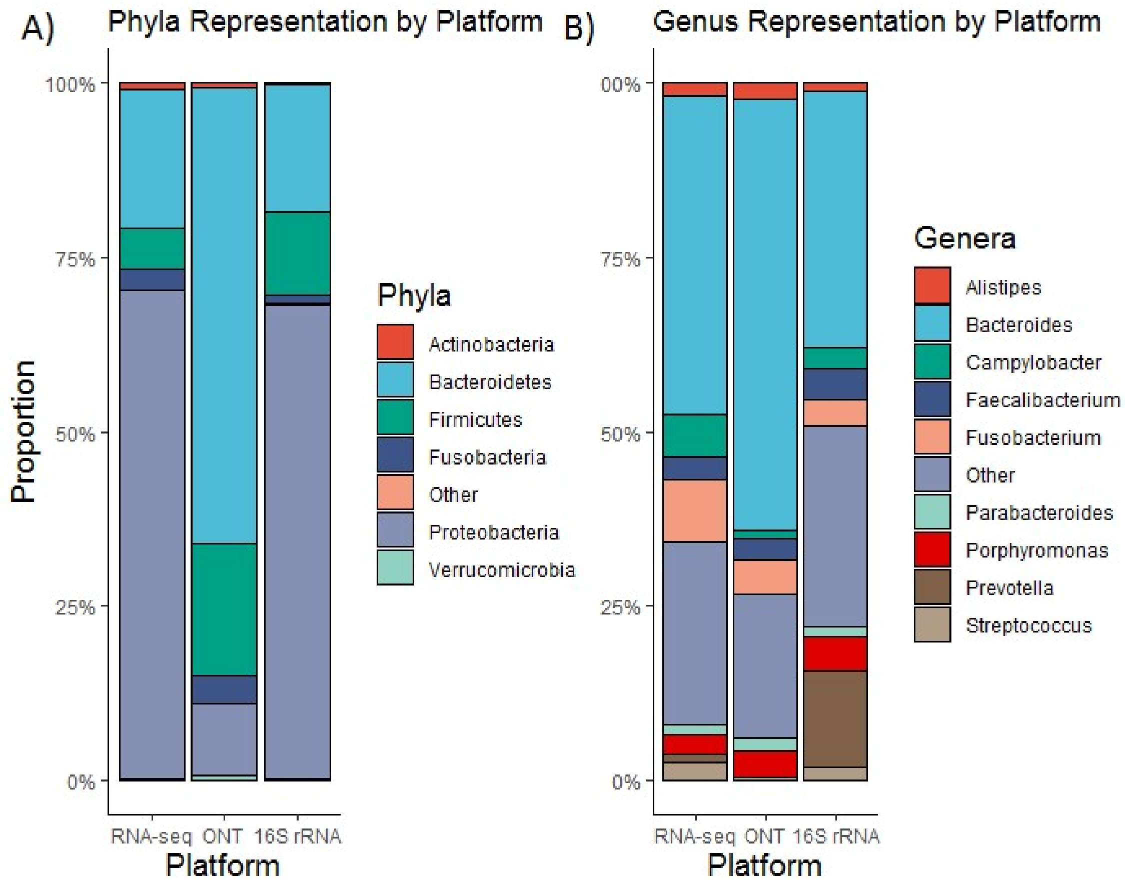

Figure 3.14. Comparison of relative abundance in Rectal samples between sequencing platforms. A) The phyla level and B), the genus level.

When comparing the number of species uniquely detected by different platforms, ONT sequencing did not detect any unique species, while RNA-Seq and 16S rRNA sequencing detected 2548 and 66 unique species, respectively (Figure 3.15).

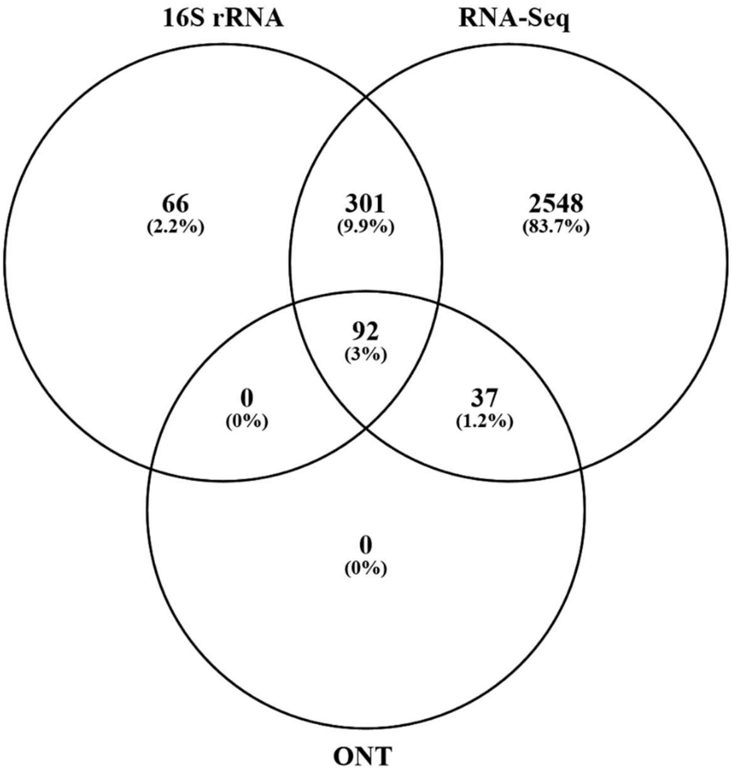

Figure 3.15. Comparison of bacterial species detection between each sequencing platform.

3.4.4 Community standard evaluation

To assess ONT sequencing's accuracy and to determine the effect of lower read counts and quality, a microbial community standard had DNA extracted, was sequenced and had taxonomy assigned using the same methodology as other samples. The standard contained ten species: two yeast, three gram-negative bacteria and five gram-positive bacteria.

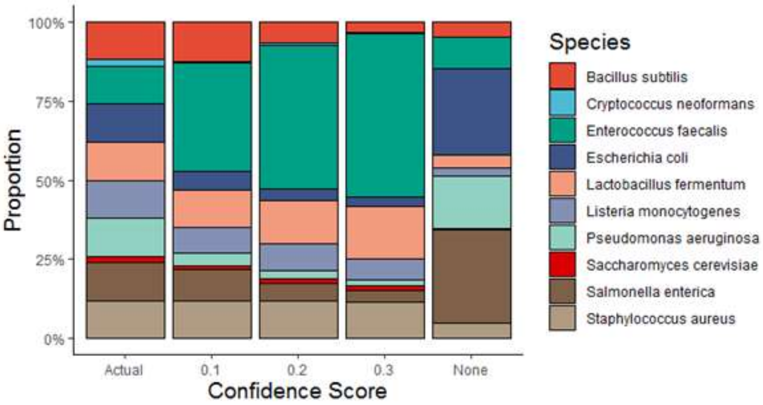

Figure 3.16. ONT sequencing of microbial community standard using increasing Kraken2 confidence scores. Actual represents the proportion of each species advertised in the community standard.

Yeast species were under detected regardless of the confidence score used. A confidence score of 0.2, detected yeast more accurately; however, this was at the expense of reduced bacterial detection accuracy. A confidence score of 0.1 gave the most consistent and accurate results; however, there was an over-estimation of *Enterococcus faecalis* and reduced detection of *Salmonella enterica*, *Escherichia coli* and *Pseudomonas aeruginosa* (Figure 3.16).

Table 3.16. ONT sequencing of community standard using increasing Kraken2 confidence scores.

	Kraken2 Confidence Level					
Species	**0.1**	**0.2**	**0.3**	**None**	**Community standard abundance**	**Gram**
Enterococcus faecalis	33.87%	45.17%	51.29%	9.46%	12	+
Bacillus subtilis	12.46%	6.83%	3.35%	4.63%	12	+
Lactobacillus fermentum	11.87%	13.79%	16.36%	3.77%	12	+
Staphylococcus aureus	11.72%	12.03%	11.45%	4.93%	12	+
Salmonella enterica	10.13%	5.45%	3.85%	28.48%	12	-
Listeria monocytogenes	8.04%	8.39%	6.82%	2.54%	12	+
Escherichia coli	5.93%	3.53%	3.15%	26.44%	12	-
Pseudomonas aeruginosa	3.86%	2.621%	1.87%	15.99%	12	-

Saccharomyces cerevisiae	1.05%	1.273%	1.20%	0.31%	2	Yeast
Cryptococcus neoformans	0.51%	0.55%	0.42%	0.15%	2	Yeast

Additionally, microbial compositions contained 0.198%, 0.065% and 0.043%, species not in the community standard using 0.1, 0.2 and 0.3 confidence levels, respectively. These results show that ONT sequencing is adequate for sequencing and detecting bacterial species.

3.5 Discussion

3.5.1 Analysis 1: mapping methodology

The tests on the effect of host mapping in different scenarios were performed to test the hypothesis that the taxonomic assignment software, Kraken2, could classify reads into both prokaryotic and eukaryotic categories, negating the need for host genome mapping before assignment.

Using the Synth dataset, human reads were not wholly removed via mapping, with approximately 11% of the resulting data being classified as eukaryotic, and if not accounted for were classified as bacterial in origin. The results show that non-bacterial reads can be classified if the appropriate reference is included in the database, and that host mapping is an imperfect process, with residual host reads leading to higher type 1 error rates, if not binned by a secondary filter such as during taxonomic classification. Using host mapping improved accuracy by 0.05% when using the All_DB; however, compute time increased more than a nine-fold with the addition of host-mapping.

Using the two clinical datasets (CRC-RNA and Rec-ONT) and host mapping, alpha diversity was significantly different when using host mapping and the Bac_DB for taxonomic assignment compared to the All_DB. There were no statistically significant differences in either dataset when using the All_DB with or without host mapping.

In the CRC-RNA dataset, differences in mean diversity were statistically significant when using the All_DB compared to the Bac_DB, except for the Simpson diversity index; implying that taxonomic dominance is not influenced by database choice or prior host mapping in the CRC-RNA dataset. In the Rec-ONT dataset differences in Simpson diversity were statistically significant when comparing the use of the All_DB to the Bac_DB, which may reflect the higher inter-sample read number differences not present in the CRC-RNA dataset.

The results demonstrate that using a database containing only bacterial genomes can alter the results of metagenomic analyses. Misassigned residual reads after host mapping may result in inflation of microbial diversity or misreporting the presence and abundance of clinically relevant taxa.

Using the All_DB, Bacteroidetes were increased by 10% and more than 30% in the CRC-RNA and Rec-ONT datasets, respectively, compared to using the Bac_DB. Proteobacteria were decreased using the All_DB by four-fold in the CRC-RNA dataset, and by more than 50% in the Rec-ONT dataset. Additionally, the abundance of Fusobacteria was increased when using the All_DB, while regardless of the database or dataset used, the relative levels of Firmicutes remained consistent. The differences in detection could be clinically relevant as, for example, Bacteroidetes contain several taxa associated with health and disease, such as *B.fragilis*, and health-associated *F. prausnitzii* and *Lactobacillus* species. Additionally, Proteobacteria contains important pathogens like *Shigella flexneri* and cancer-associated microbes such as *Helicobacter pylori* and *Escherichia coli* [375].

There were also differences at the genus level depending on the taxonomic database used. Important and potentially pathogenic genera, such as *Klebsiella*, *Pasteurella*, *Mycoplasma*, *Yersinia* and *Staphylococcus* were present in the most abundant genera when the Bac_DB was used. Several consequential taxa were found to be among the most abundant when the All_DB was used, such as oral microbes *Leptotrichia and Lachnoanaerobaculum* [376] [377], the foodborne pathogen *Campylobacter* [378], *Hungatella,* which was recently associated with brain aneurysms [379], SCFA-producing *Eubacterium* [380], and *Alistipes* which has been implicated in cancer, inflammatory disease and mental health [381].

Metagenomic data from tissue samples has been used to study microbiome composition for classification or grouping of disease states, such as IBD [382], or CRC signatures [93, 109], and potentially for CRC diagnosis and predicting survival [180, 383]; however, without appropriate measures to remove host sequences, the results and interpretation of data in these and other studies may be incorrect.

Overall, the results of the first analysis demonstrate that not only is mapping to the host genome before bacterial taxonomic assignment insufficient for removal of host reads, but that host mapping may be largely redundant when the host genome is included in the taxonomic assignment database. Adopting a broad taxonomic classification database that includes the host genome is shown to increase the accuracy and speed of metagenomic analyses.

3.5.2 Analysis 2: RC dataset platform comparison

In Analysis 2, ONT sequencing was compared to 16S rRNA sequencing and RNA-Seq in their ability to analyse RC samples' tissue microbiomes using the methods implemented in previous CRC studies [93, 371]. The analysis was carried out using host contaminated samples extracted from human tissue. No attempt was made to reduce the amount of host genomic material during extraction methods; only 16S rRNA sequencing specifically selected microbial DNA via selective PCR primers.

For ONT sequencing, samples were barcoded and sequenced as multiplexed libraries. Due, in part, to lacking an amplification step, the number of bacterial reads were several orders of magnitude lower than the other platforms. In addition, the synthesis-free nanopore sequencing method can only read sequences once as they pass through a pore, in contrast to the bridge amplification used in Illumina sequencers which can sequence the same strand multiple times. Another factor in the low number of post-mapping ONT reads was that almost half of the ONT reads were not barcoded, which may introduce sampling bias. The loss of barcodes is a common factor in ONT sequencing generally, as the high molecular weight DNA becomes prone to breakage as it is purified, and stabilising proteins are lost.

Based on the initial comparative study, concordance results between the sequencing platforms was initially promising [371]. Concordance of phyla, genera and species assignment between 16S rRNA and ONT sequencing was 67.6%, 35.8% and 19.5%; between 16S rRNA and RNA-Seq, 80.5%, 36.7% and 18.9%; and between ONT and RNA-Seq, 66.7%, 51.5% and 35%. Despite the low numbers of reads acquired using ONT sequencing, more than a 1300 species could be taxonomically assigned, most of which could also be detected using RNA-Seq data. It has been theorised that long ONT reads might compensate for a lower number of reads by being more efficient, as it is possible to discriminate between species using larger query sequences; Wommack et al. [384] showed that long reads could detect 72% more hits than short read lengths of up to 400bp at twice the read depth. Although ONT sequencing is known to have an inherently high error rate, which was also observed in this study (Figure 3.10), this can be compensated for by using longer reads [385]. Additionally, ONT data is known to suffer at shorter read lengths (<1000 bp) in terms of taxonomic classification, and this is particularly the case when using Kraken2 [386].

Implementing the methodological changes from Analysis 1, the concordance between platforms was increased compared to the initial study, likely due to the methodological changes; using the All_DB for all taxonomic assignment and a taxonomic confidence score. Removal of excessive *Burkholderia* levels from the 16S rRNA dataset from amplification bias was justified in that it resulted in better concordance (Figure 3.12) and preserved genus representation between platforms (Figure 3.14), which was improved between both RNA-Seq and ONT compared to the CRC study [371]. Concordance increased at the species level when comparing ONT and RNA-Seq; however, concordance was reduced at both the phyla and genera levels (Table 3.14), which was likely due to residual host sequences no longer being misassigned to the same taxa. The highest concordance was between RNA-Seq and 16S rRNA in line with the previous CRC study, while the lowest was between ONT and RNA-Seq.

As Kraken2 has been recently shown to be an accurate method of 16S rRNA taxonomic classification [387], the benefit of utilising the same taxonomic database for all comparisons likely had the largest impact on concordance compared to the prior CRC study [371]. The DADA2 and the SILVA132 databases previously used for 16S rRNA taxonomic assignment may have contained naming differences compared to the Bac_DB used for the ONT and RNA-Seq reads, and that taxa were absent from one database were present in the other, and vice versa. Additionally, removing residual human and non-bacterial reads may explain the increased concordance between other platforms and 16S rRNA sequencing in this study.

When comparing ONT with RNA-Seq, some samples had low or negative correlations, due to some samples having few reads after mapping which could be assigned taxonomy and the effect of the relative transform on the dataset. For example, the RNA-Seq taxonomy for sample RT13N contained 122 species ranging in abundance from 0.00328 to 0.2547; however, the ONT taxonomy for the same sample contained a single species (*Bacteroides cellulosilyticus*), giving it an abundance of 1, while it appeared in the RNA-Seq dataset an abundance of 0.000328.

Using the All_DB also reduced the detection efficiency of all three platforms compared to the previous CRC study, with almost 21-fold more reads being required per species identified for RNA-Seq, 124-fold more for ONT reads, and 1.75-fold more 16S rRNA reads. It should be noted that 16S rRNA is known to discriminate between species poorly; however, the efficiency decrease was magnitudes smaller than with other platforms when using the All_DB, likely due to the lack of impact from host reads. Despite the reduction in species detection, ONT sequencing still outperformed RNA-Seq on a raw-read per species detection basis, requiring eight-fold fewer reads per species detected.

ONT sequencing had high error rates, lower concordance with other platforms, and detected species less efficiently than in the previous CRC study. To demonstrate that ONT sequencing data could be used for microbial community evaluation, a community standard was sequenced using the same methodology as the CHCH cohort ONT samples. The results showed that ONT sequencing was capable of sequencing bacteria close to the proportions of a given sample; however, additional species not included in the community standard were classified using different confidence scores. The additional species identified did not exceed 0.2% of the compositional total when using the most accurate confidence score of 0.1. These additional species may be due to bleed over from other multiplexed samples or indicate the limitations that Kraken2 has in identifying species. Overall, based on the community standard analysis results, a confidence score of 0.1 and a species abundance cut-off of 0.2% per sample was utilised for analyses moving forward. However, this may not be sufficient for 16S rRNA, and RNA-Seq analyses as the community standard was not used with these platforms, and there may be additional variance due to sequence quality and read numbers.

3.6 Conclusions

This chapter's results show that using a broad taxonomic assignment database that contains the human genome improves the accuracy and reduces the time required for metagenomic studies while providing proportional information on the host-microbe content of a sample. Additionally, the increased accuracy and reduced interference from host reads increases inter-platform concordance; however, this improvement did not apply when comparatively few reads were supplied, such as the case for ONT sequencing concerning RNA-Seq.

4 CHAPTER 4: RC MICROBIOME

4.1 Introduction

The microbiome has been implicated in sporadic CRC [388]; however, the relationship between the microbiome and chemoradiotherapy outcomes has yet to be thoroughly investigated. It has been shown that the microbiome can modify the pharmacokinetics of anti-cancer drugs, for example, *Fusobacterium nucleatum* can promote resistance to 5-FU and platinum-based drugs used to treat CRC [389-392]. Additionally, the role of the microbiome in radiotherapy side effects, such as postradiotherapy diarrhoea, mucositis, fatigue, and other gastrointestinal side-effects has been well studied [393-396], as well as the effect of radiotherapy in altering the microbiome [397-399]. However, a link between radiotherapy outcomes and the microbiome has yet to be proven. In terms of direct microbiome-radiotherapy interactions, evidence points toward an immunomodulatory effect [242], such as gram-positive bacterial depletion with vancomycin leading to enhanced anti-tumour immune responses dependant on dendritic cell antigen presentation [400]. In contrast, other studies point toward microbial metabolites, such as vitamin D metabolism by microbiota contributing to radioresistance [401].

This chapter aimed to investigate the microbiomes of RC tissues, using RNA-Seq data using Kraken2, as described in Chapters 2 and 3. In the first section, alpha diversity was compared between tissues and response groups using Observed diversity, the total transcription of taxa; Shannon diversity index, a measure of evenness within the sample; and the Simpson index, used to measure the level of dominance within a sample [402]. Microbiome composition, differential transcription, and beta-diversity were measured across response groups and patient tissues. The effect size of beta-diversity measures by the response and metadata variables on the microbiome was calculated and tested. Transcriptional activity was correlated with response to find which taxa were most associated with response in each tissue type. Lastly, transcriptional alignment was used to determine the presence of enterotoxigenic *Bacteroides fragilis*.

4.2 Methods

Taxonomies were produced using host mapped reads from the sequencing data described in Chapter 3, and assigned taxonomy using Kraken2 v2.0.7 and the All_DB with a confidence score of 0.1, as described in Chapter 2. Diversity and composition were analysed and visualised using the phyloseq v1.28 [316], ape v5.4 [353], Pavian v1.0 [403] and VEGAN 2.5 [352] R packages. Samples were rarefied to 90% of the sample sum of the lowest sample before alpha diversity calculations. The statistical significance of alpha diversity differences between tumour and normal tissues and response groups was tested using Wilcoxon signed-rank tests, and *p*-values were adjusted for false discovery rate (FDR) using Benjamini and Hochberg (BH) correction [358].

Filtering was done after alpha-diversity analysis as singletons are rare taxa are used for calculating diversity indices. Additionally, there is no gold standard for taxa filtering, and thresholds are mostly arbitrary, with choices varying between researchers and the data used [404]. Low abundance taxa with a read count lower than 30 in 20% of samples were removed after alpha diversity analysis for compositional barplots and differential transcription. For beta-diversity analyses, taxa with more than five reads in 20% of samples were retained to preserve greater sample heterogeneity. Significance of differences between tissues and response groups in taxa boxplots was measured using Wilcoxon signed-rank tests, and *p*-values were adjusted for FDR with Benjamini and Hochberg [358].

Effect sizes were calculated using adonis function, and homogeneity of variable dispersions was calculated with betadisper, both part of the VEGAN [352] R package. Differential microbial transcription was analysed with edgeR [356], relative log expression, a likelihood ratio test, and *p*-values were adjusted for false discovery rate (FDR) with Benjamini and Hochberg [358]**.** Correlations between relative abundance and response were calculated using Spearman's rank correlation, and *p*-values were adjusted for FDR with Benjamini and Hochberg [358]. Extracted reads assigned to *B. fragilis* were aligned to a reference genome using STAR v2.6.1 [271] without splice aware alignment. Scripts and code used for the analyses can be found at https://github.com/William-S-Taylor/MSc.

4.3 Results

First, each platform compared in the previous chapter was tested for sample depth and appropriateness for community analysis, then alpha diversity, composition, beta diversity and effect sizes were analysed. Finally, RNA-Seq data was used to determine if the most transcriptionally active taxa were expressing a particular gene.

4.3.1 Rarefaction

Rarefaction curves were calculated from the taxonomies of ONT (Figure 4.1a), 16S rRNA (Figure 4.1b) (microbial abundance) and RNA-Seq (Figure 4.1c) (transcriptomic activity). The 16S rRNA and ONT datasets contained data from the Christchurch (CHCH) cohort (*n* = 20), while the RNA-Seq dataset contained data from both he CHCH and Peter MacCallum cohort (PM). The ONT dataset sample sums ranged from 2714 to 2 with a standard deviation of 635, and half of the sample sums were below 60;

therefore, rarefaction could not be performed without reducing the data considerably. Additionally, the 16S rRNA dataset was also adversely affected by rarefaction (reducing total taxa from 671 to 489); however, the rarefied RNA-Seq dataset contained the greatest number of taxa (1916) and samples and therefore had the greatest likelihood of detecting significant response group variability.

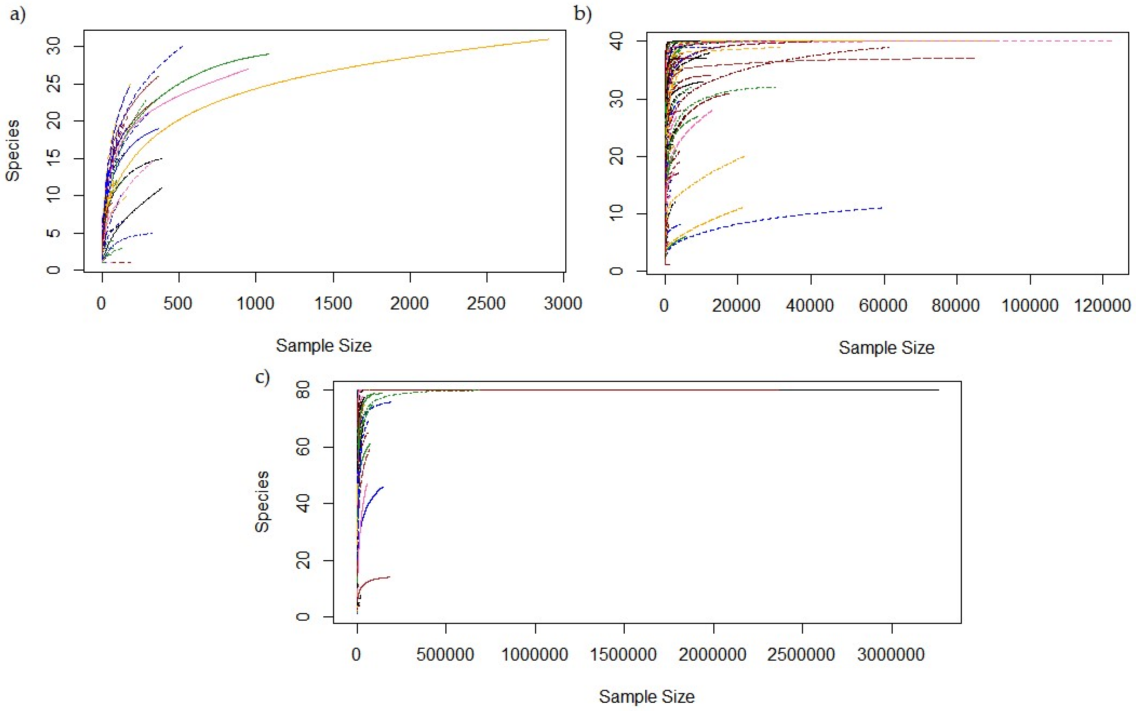

Figure 4.1. Dataset rarefaction curves. a) ONT; b) 16S rRNA, c) RNA-Seq.

Due to the impact of rarefaction and the uneven and lower sample sizes of the other taxonomies, only the RNA-Seq taxonomy was utilised for microbiome analysis. Rarefaction reduced the number of taxa (Figure 4.2) in the RNA-Seq dataset by 46% (3534 to 1916). Post-rarefaction, each sample was normalised to taxa sample sums of 4216 to give a balanced comparison between samples.

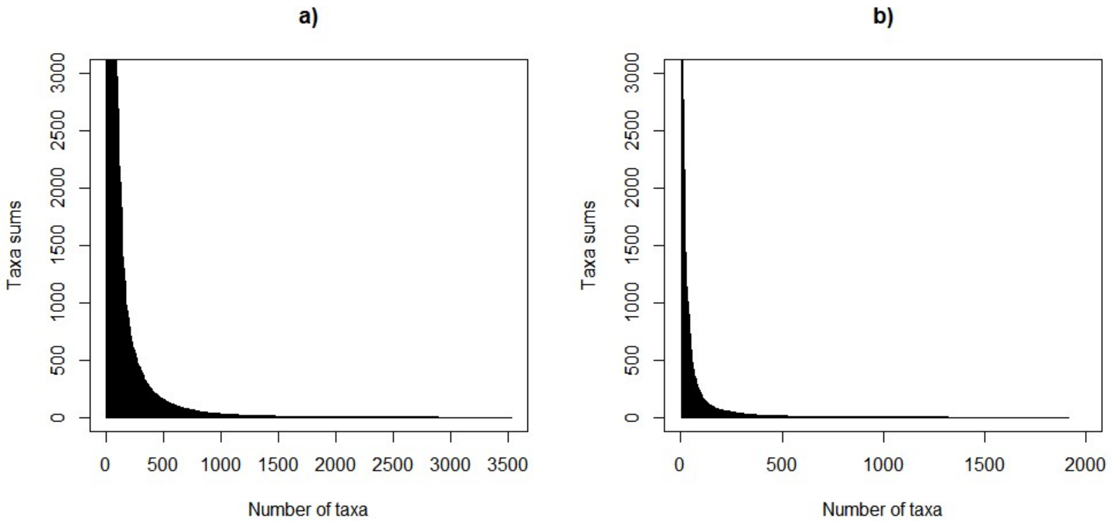

Figure 4.2. Effect of rarefaction on the RNA-seq taxonomy. a) RNA-Seq pre-rarefaction, b) RNA-Seq post-rarefaction.

4.3.2 Alpha diversity

Alpha diversity differences were tested for statistical significance ($p > 0.05$) using Wilcoxon signed-rank tests, and FDR correction was performed with Benjamini and Hochberg. Comparisons were made between tissue types, tissues of Dworak score groups, and the differences between tissues when grouped into high and low response grades (Dworak Four and Three, and Dworak Two and One, respectively) (Figure 4.3). The differences between the Shannon diversity scores of Dworak One and Two normal tissues, and the Simpson diversity scores between of Dworak Two tumour and normal tissues were found to be statistically significant ($p = 0.047$ and $p = 0.049$, respectively) (Figure 4.3b); however, no other alpha diversity comparisons were statistically significant.

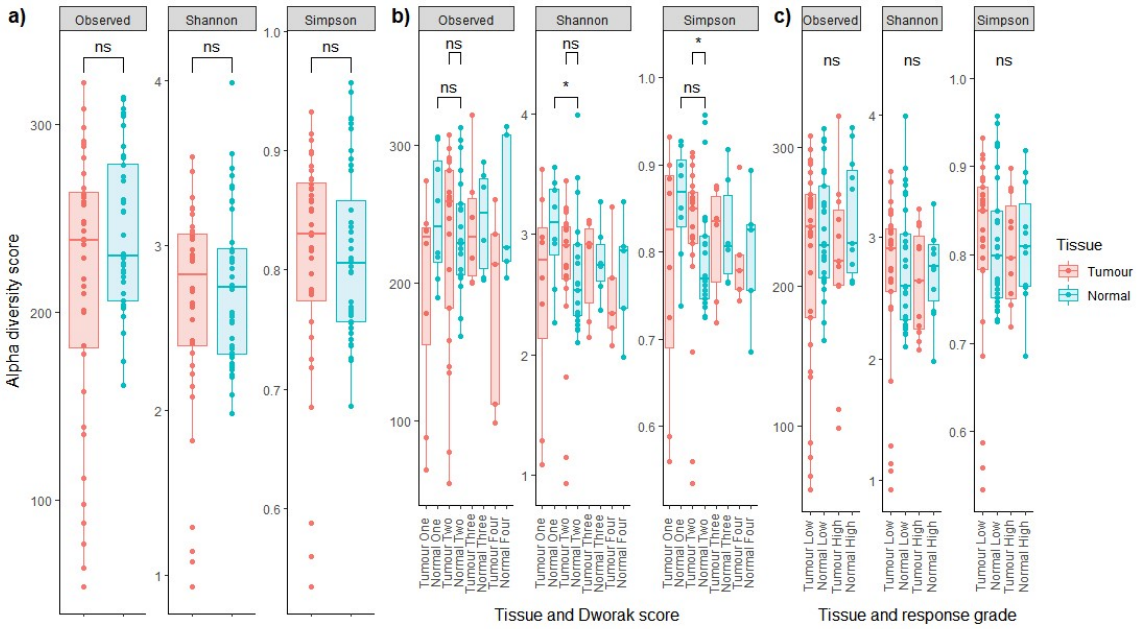

*Figure 4.3. Alpha diversity of bacterial transcription. a) Tissue comparison, b) Dworak score tissue comparison, c) response grade tissue comparison. *: p < 0.05; ns: not statistically significant. Comparisons that were not statistically significant in at least one instances are not shown.*

4.4 Microbiome compositions and differential analysis

The relative level of bacterial transcription within samples was hypothesised to be associated with response. These levels would theoretically be different between the tumour and normal tissues of different response groups. Therefore, the relative transcriptional activity was plotted to compare the phyla and genus expression of different tissues and response groups, followed by differential expression analysis to determine these differences' statistical significance. Samples were grouped to assess the differences between high (Dworak Four and Three) and low (Dworak Two and One) response groups.

4.4.1 Compositions and differential transcription

At the phylum level, samples were mainly dominated by Proteobacteria, with 25 samples being comprised of more than 90% Proteobacteria. In comparison, the taxonomies of three samples were comprised of mainly Bacteroidetes, the normal tissue of a Dworak One patient, the normal and tumour tissue of a Dworak Two patient, the normal tissue of a Dworak Three patient, and the tumour tissue of a Dworak Four patient (Figure 4.4). On average, the top three most transcriptional active phyla were Proteobacteria, Bacteroidetes and Firmicutes, followed by Fusobacteria and Actinobacteria. Fusobacteria transcription was the highest in those samples with a Dworak score of Two. Fusobacteria transcription was detected in the normal samples of complete responders; however, at less than 1% of total transcriptional abundance. Cyanobacteria and Tenericutes had low transcriptional activity. Verrucomicrobia had the highest activity in the normal and tumour tissues of a Dworak Three patient (9.5% and 2.2%, respectively), followed by the normal and tumour tissues of a Dworak Four patient (8.3% and 2.3%, respectively). The Dworak Three patient with the highest Verrucomicrobia activity levels also had considerable levels of Spirochates in their normal and tumour tissues (38% and 0.5%, respectively) which were 163 and two times higher, respectively, than the sample with the next highest level.

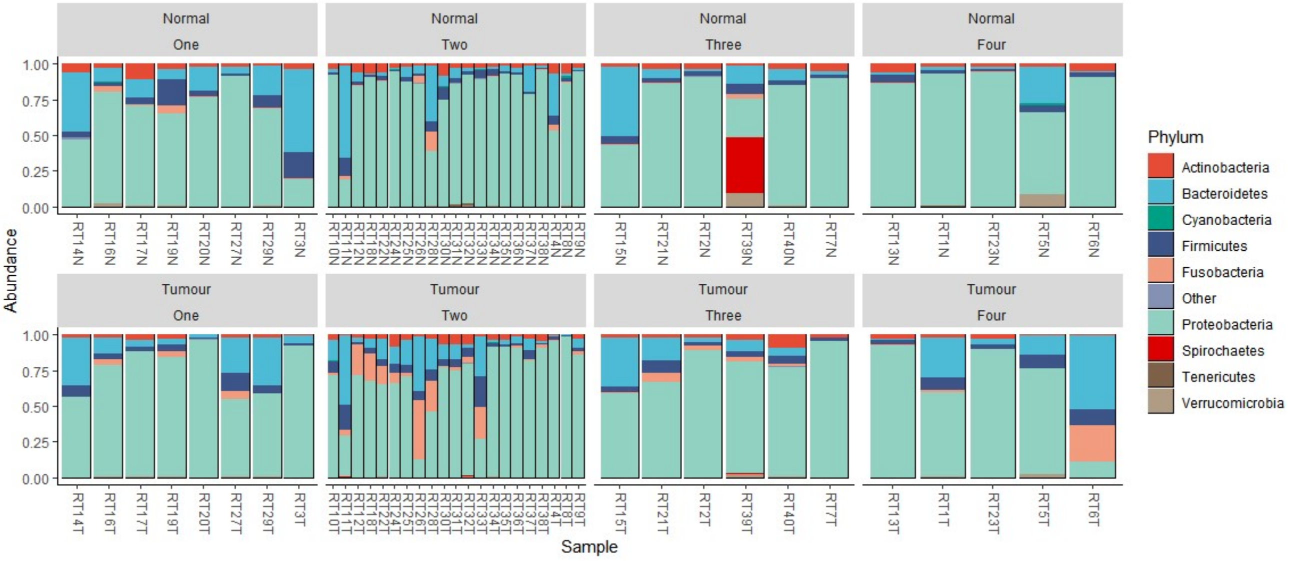

Figure 4.4. Relative transcriptional phylum activity ordered by tissue type and Dworak score.

Analysis of the differential expression of phyla showed that Cyanobacteria had statistically significant (FDR corrected p-value < 0.05) differential transcription between tissues (Figure 4.5). Overall, there was 1.55-fold more Cyanobacteria transcription in tumour tissues than normal tissues (Figure 4.5a), while in the high and low response groups, there was 2.27-fold and 1.33-fold higher Cyanobacteria transcription in tumour tissues (Figure 4.5a,b). No other comparison yielded phyla with statistically significant fold-change differences.

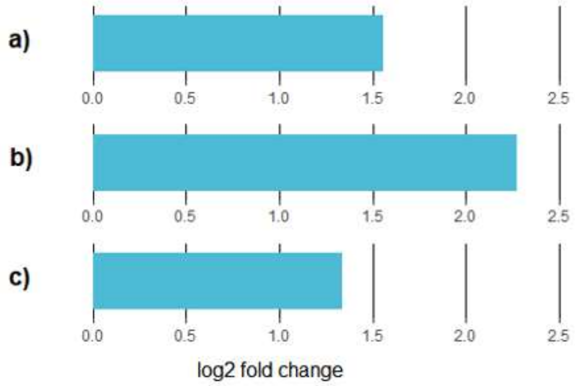

Figure 4.5. Differential transcription of Cyanobacteria between normal and tumour tissues. a) tumour vs normal, b) high responder tumour vs normal, c) low responder tumour vs normal.

At the genus level, samples were dominated mainly by *Bacteroides*, the ‘Other’ category, and *Pseudomonas* (Figure 4.6). *Fusobacterium* transcription was highest in Dworak Two patients' tissues, while the least was found in Dworak Four patients, with the majority occurring tumour tissues. *Bacteroides* were most transcriptionally active in Dworak One samples, while a Dworak Four patient had the highest transcription levels of *Faecalibacterium* (Figure 4.6).

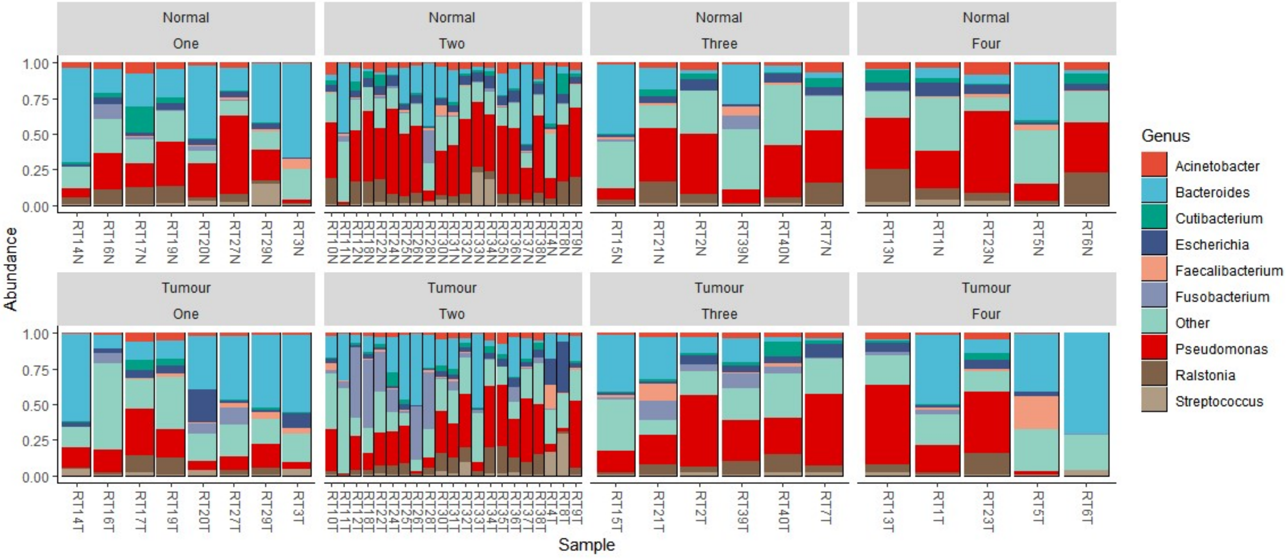

Figure 4.6. Relative genera transcriptional activity, ordered by tissue type and Dworak score.

After correcting for FDR, analysis of differential expression at the genus level revealed that only *Corynebacterium* had lower expression in tumours compared to normal tissues (-0.46), while six genera had statistically significant ($p < 0.05$) log2 fold change differences > 1 (Figure 4.7a). *Campylobacter* and *Porphyromonas* had the highest levels of differential transcription between tumour and normal tissues (4.93-fold; $p = 1.28 \times 10^{-10}$ and 3.13-fold; $p = 2.59 \times 10^{-5}$, respectively), followed by *Streptococcus*, *Clostridium*, *Bacteroides* and *Collinsella* (1.59-fold; $p = 0.00014$, 1.31-fold; $p = 0.00015$, 1.24-fold; $p = 0.012$, and 1.68-fold; $p = 0.016$, respectively) (Figure 4.7a).

In contrast, in the high responder group, 12 genera had higher transcriptional activity in tumours than normal tissues with a log2 fold-change higher than one (Figure 4.7b), while in the low responder group three genera had higher transcriptional activity (Figure 4.7c). *Campylobacter* had similar fold-differences between tumour and normal tissues in both high and low response groups (4.7-fold and 5.03-fold, respectively). When assessing the log2 fold changes differences between these groups (fold change differences of fold change differences between tumour and normal samples), it was found that *Hungatella*, *Butrycimonas*, *Flavonifractor* and *Oscillibacter* were more transcriptionally active in the high responder compared to the low responder group (Figure 4.7d).

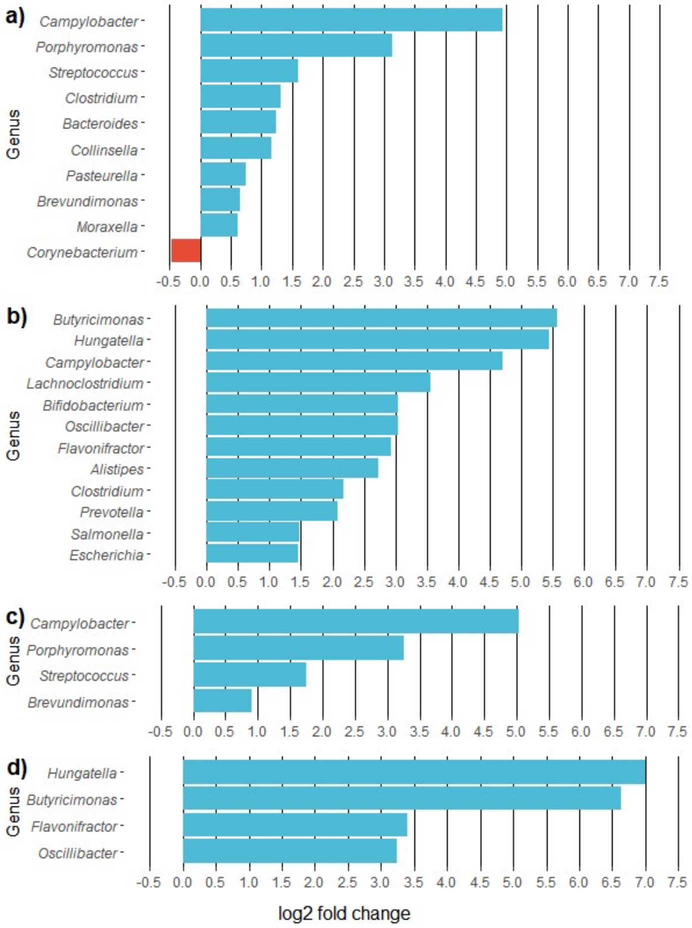

Figure 4.7. Differentially expressed genera between tissues. a) tumour vs normal, b) high responder tumour vs normal, c) low responder tumour vs normal, d) high and low responder tumour vs normal tissue differences.

The differences between tissue types of each response group were then tested, showing a statistically significant ($p < 0.05$) log2 fold difference (>1), in the transcriptional activity of 21 genera in the tumour tissues of high and low responders (Figure 4.8a). The greatest differences in transcriptional activity between high and low responder tumour tissues were those of *Butyricimonas*, *Flavonifractor*, *Odoribacter* and *Alistipes*, all of which had higher transcriptional activity in high responders (4.92-fold, 2.60-fold, 2.51-fold and 2.19-fold, respectively) (Figure 4.8a). Genera with higher transcriptional activity in the tumour tissue of low responders were *Veillonella*, *Staphylococcus*, *Brevundimonas*, *Cutibacterium* and *Stenotrophomonas* (1.60-fold, 1.32-fold, 1.23-fold, 1.17-fold and 1.01-fold, respectively) (Figure 4.8a). Between the normal tissues of high and low responders, Hungatella were found to have statistically significant higher transcription in low responder normal tissues (5.44-fold) (Figure 4.8b).

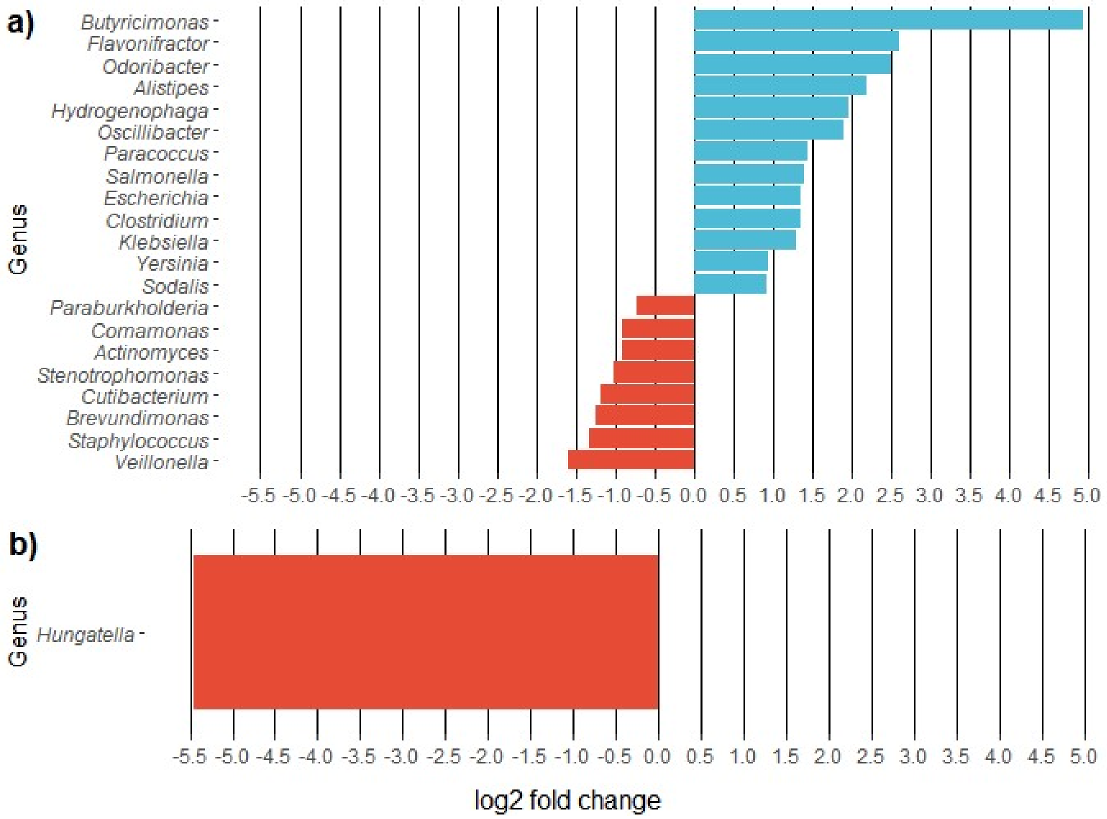

Figure 4.8. Differential transcription of genera between tissues. a) high responder vs low responder tumour tissue, b) low responder normal vs normal tissue.

Differential relative transcriptional activity of eight species was found between tumour and normal tissues (Figure 4.9a). The greatest differences were higher levels in tumour tissue of *Campylobacter ureolyticus, Lachnospiraceae bacterium* GAM79, and *Porphyromonas asaccharolytica* (4.00-fold, 2.29-fold and 1.91-fold, respectively), and lower levels of *Hungatella hathewayi*, *Fusobacterium nucleatum* and *Bacteroides thetaiotaomicron* (2.37-fold, 2.16-fold and 1.76-fold, respectively), in comparison to normal tissues (Figure 4.9a).

Between the tumour and normal tissues of high responders, six species were significantly higher in tumour tissues, *Butyricimonas faecalis*, *Eubacterium rectale*, *Alistipes finegoldii*, *Streptococcus pyogenes, Bifidobacterium longum* and *Clostridium saccharobutylicum* (4.17-fold, 3.16-fold, 2.95-fold, 2.92-fold, 1.75-fold and 1.98-fold, respectively) (Figure 4.9b). In contrast, within low responders, two species had significantly higher levels of transcription in tumours (*C. ureolyticus*: 4.79-fold and *L. bacterium* GAM79: 2.12-fold), while three had significantly higher levels in normal tissues (*F. nucleatum*: 3.11-fold, *H. hathewayi*: 2.93-fold and *B. thetaiotaomicron*: 2.16-fold) (Figure 4.9c). When assessing the differences in the log2 fold changes between response group tissues, it was found that the differences in transcription of *H. hathewayi*, *B. faecalis*, *A. finegoldii* and *E. rectale* were higher between the tumour and normal tissues of high compared to those of low responders (Figure 4.9d).

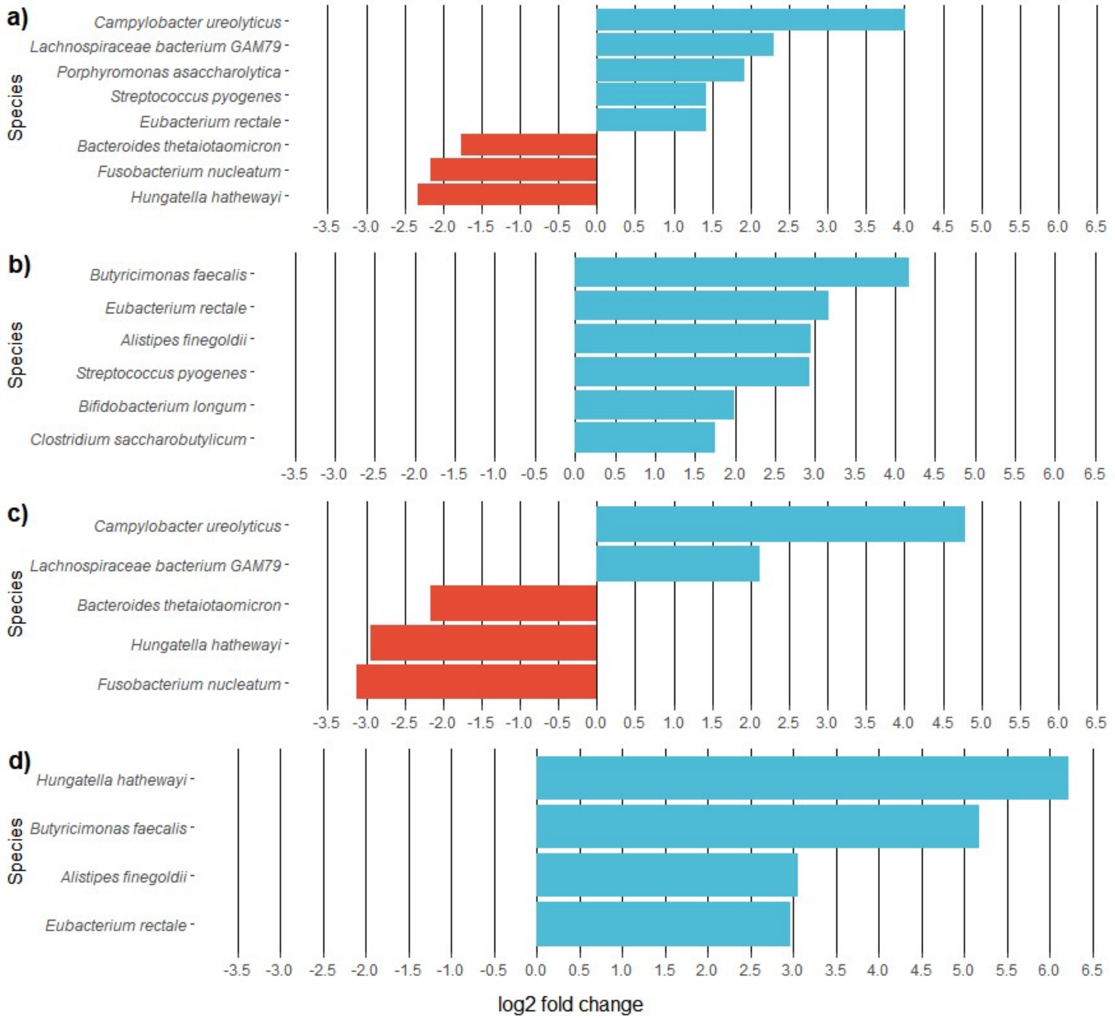

Figure 4.9. Differential transcription of species between tissues. a) tumour vs normal, b) high responder tumour vs normal tissue, c) low responder tumour vs normal tissue, d) high and low responder tumour vs normal tissue differences.

Between the tumour tissues of high and low responders, 15 species were significantly differentially transcribed (Figure 4.10a). Of those with higher transcription in high responder tumour tissues, *B. faecalis*, *E. rectale*, *S. pyogenes* and *A. finegoldii* had the greatest differences compared to the tumour tissue of low responders (3.33-fold, 2.35-fold, 1.98-fold and 1.80-fold, respectively). *C. saccharobutylicum* was also significantly higher in high responder tumour tissue (1.50-fold) (Figure 4.10a), as when compared to respective normal tissues (Figure 4.10b).

The species with the greatest differential transcription between the tumour tissues of response groups were *Actinomyces oris*, *Cutibacterium acnes*, *Stentrophomonas maltophilla*, *Ralstonia pickettii*, *R. insidiosa* and *Cupiavidus metallidurans* 1.77-fold, 1.75-fold, 1.67-fold, 1.37-fold, 1.33-fold, and 1.32-fold, respectively) (Figure 4.10a). *H. hathewayi* was found to be 5.56-fold lower in the normal tissue of the high response group compared to the low response group (Figure 4.10b).

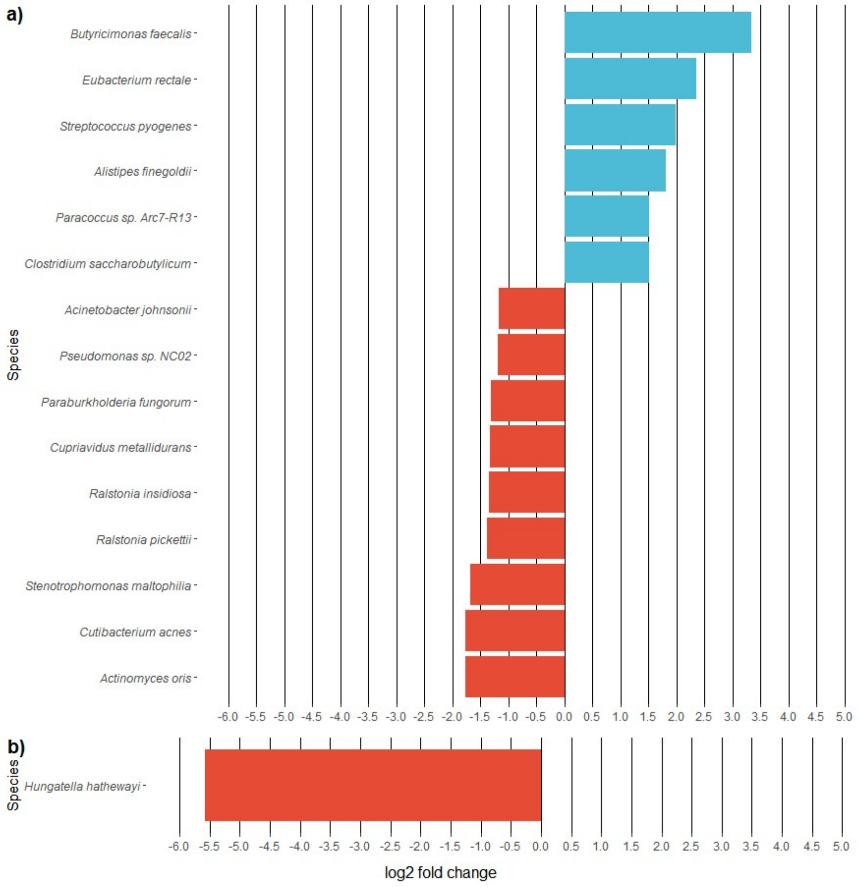

Figure 4.10. Differential transcription of species within tissues. a) high responder vs low responder tumour tissue, b) low responder normal vs normal tissue.

Those species found to have significant differential transcription in at least one comparison were used in comparisons between Dworak scores and tissue types. Significance testing (FDR adjusted $p < 0.05$) showed nine species to be differentially transcribed between different Dworak response group tissues (Figure 4.11). The two species with the highest relative transcription were *C. ureolyticus*, *F. nucleatum* and *H. hathewayi*. *C. ureolyticus* transcription levels were significantly different between the tumour and normal tissues of Dworak four patients (Figure 11a), and *F. nucleatum* between Dworak Two tumour and normal tissue, and Dworak Three and Four normal tissues (Figure 4.11b).

H. hathewayi had significantly different transcription levels between Dworak Four tumour and normal tissues, Dworak Four and Dworak Three tumour tissues and Dworak Four and Dworak Two and One tumour tissues (Figure 4.11c). Less transcriptionally active species with more than one significant difference were *Paraburkholderia fungorum* (Dworak One and Two tumour tissues; Dworak Two tumour and normal tissues) (Figure 4.11g), *Pseudomonas* sp. NC02 (Dworak Two tumour and normal tissues; Dworak Two and Dworak One and Three tumour tissues) (Figure 4.11h), and *Ralstonia insidiosa* (Dworak One tumour and normal tissues, Dworak Two tumour and normal tissues, and between tumour tissues of Dworak One and Two) (Figure 4.11i).

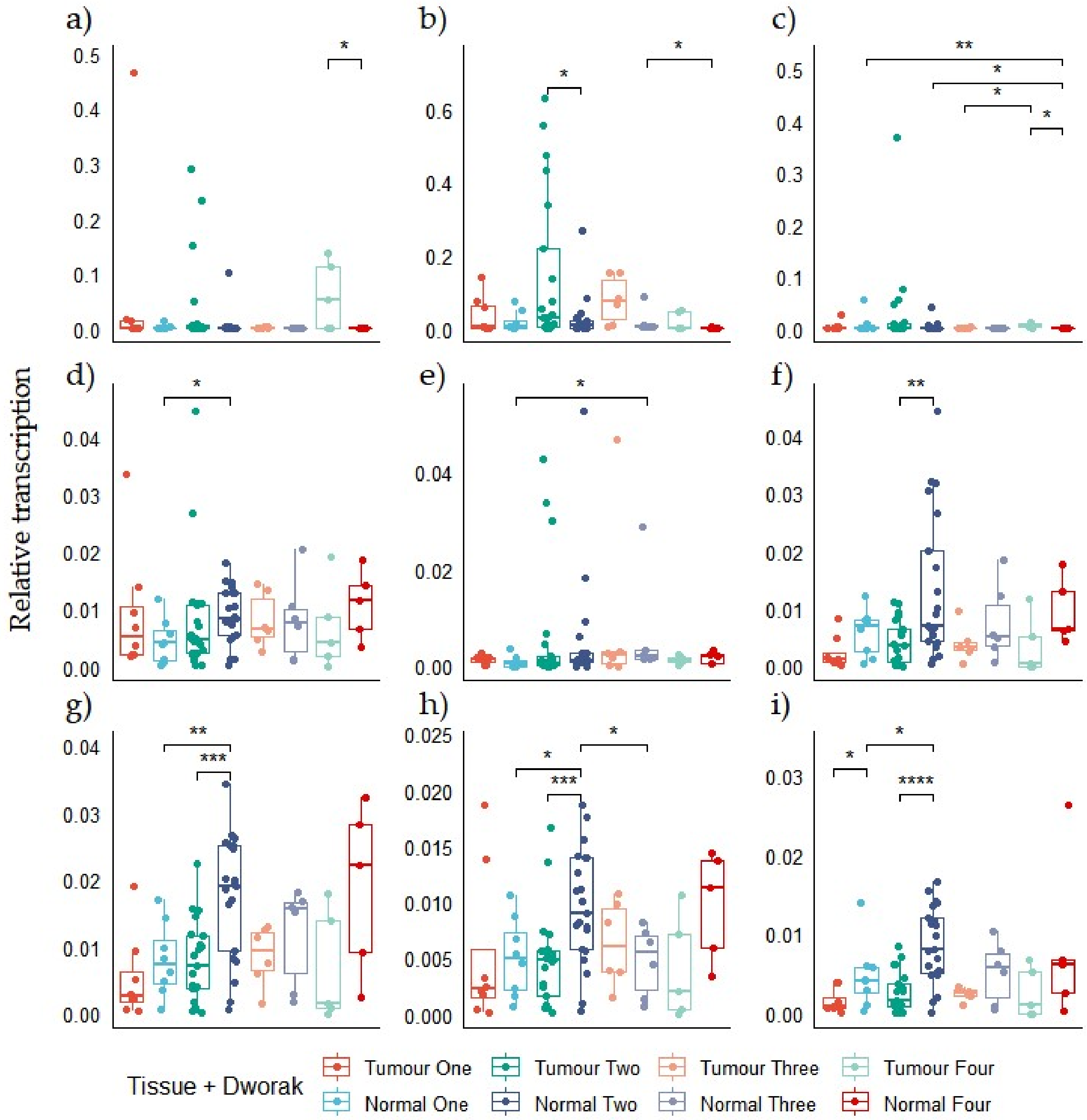

*Figure 4.11. Relative transcription levels of differentially transcribed species with atleast one statistically significant differences between Dworak score tissue groups. a) Campylobacter ureolyticus, b) Fusobacterium nucleatum, c) Hungetalla hathewayi, d) Clostridium saccharobutylicum, e) Bifidobacterium longum, f) Stenotrophomonas maltophila, g), Paraburkholderia fungorum, h) Pseudomonas sp. NC02, i) Ralstonia insidiosa. *: p < 0.05; **: p <0.01; ***: p < 0.001; **** p < 0.0001.*

4.5 Effect size and beta-diversity

Beta diversity was investigated between Dworak score groups and tissues using non-metric dimensional scaling (NMDS) plots and distance metrics: Bray–Curtis dissimilarity (compositional dissimilarity), Jensen–Shannon divergence (similarity between probability distributions) and the Jaccard index (similarity between unique members). Bray–Curtis (BC) takes taxa abundance into account (in this case, the level of transcription) to measure dissimilarity, while Jensen–Shannon divergence (JSD) measures similarity and has been used to establish enterotypes [177]. The Jaccard index (JI) does not take abundance into account and can be seen as a difference in the presence and absence of taxa in respective samples.

The effect size of gender, age, cohort, Dworak score and tissue on the homogeneity of dispersion within nested groups was assessed using adonis R^2 values. The homogeneity of each variable was tested using betadisper and the distance from centroids.

4.5.1 Beta-diversity

Using NMDS plots of calculated distance metrics, little clustering and high 95% confidence interval overlap could be seen within Dworak group tumour tissue samples (Figure 4.12a–c) or normal tissue samples (Figure 4.12d–f), indicating little difference between the groups. Using BC and the JI resulted in very similar clustering of samples. Additionally, when clustering samples by tissue types, substantial interpersonal variability was seen with each distance metric, with tumour tissues tending to be more distant from the centre (Figure 4.12g–i).

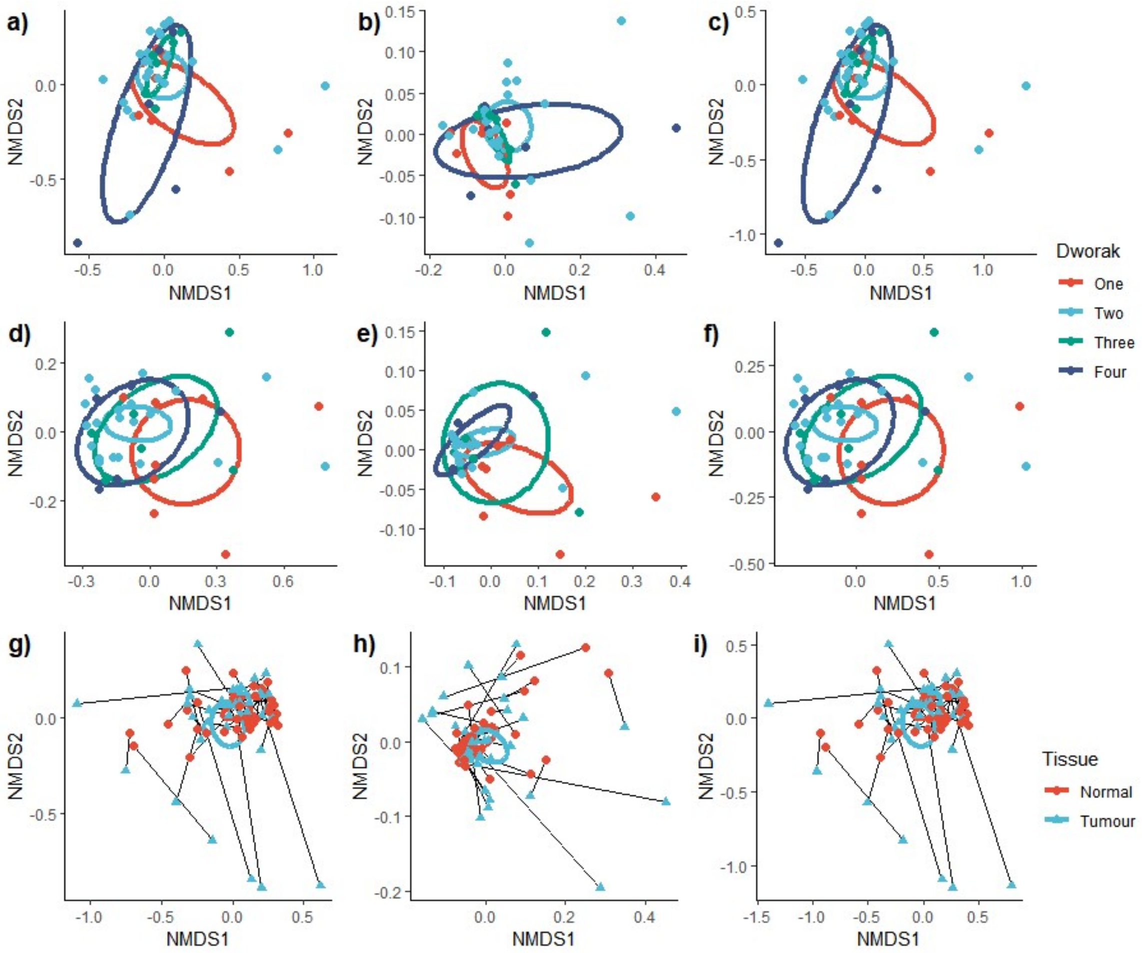

Figure 4.12. Beta-diversity non-metric dimensional scaling (NMDS) plots of bacterial transcription in tissues and Dworak groups. a) Bray–Curtis tumour tissue, b) Jenson–Shannon tumour tissue, c) Jaccard index tumour tissue, d) Bray–Curtis normal tissue, e) Jenson–Shannon normal tissue, f) Jaccard index normal tissue. Lines in the following indicate sample pairs; circles indicate tumour tissue and triangles are normal tissues. g) Bray–Curtis tumour and normal tissue, h) Jenson–Shannon tumour and normal tissue, i) Jaccard index tumour and normal tissue. Ellipses represent 95% confidence intervals.

4.5.2 Effect sizes

In testing individual variables, the effect of Dworak as a variable had the largest single effect size (0.032–0.038) with JSD and JD displaying slightly higher effect sizes than BC, followed by cohort (0.019–0.023) and tissue (0.014–0.02) (Figure 4.13). However, these effect sizes were not found to be statistically significant.

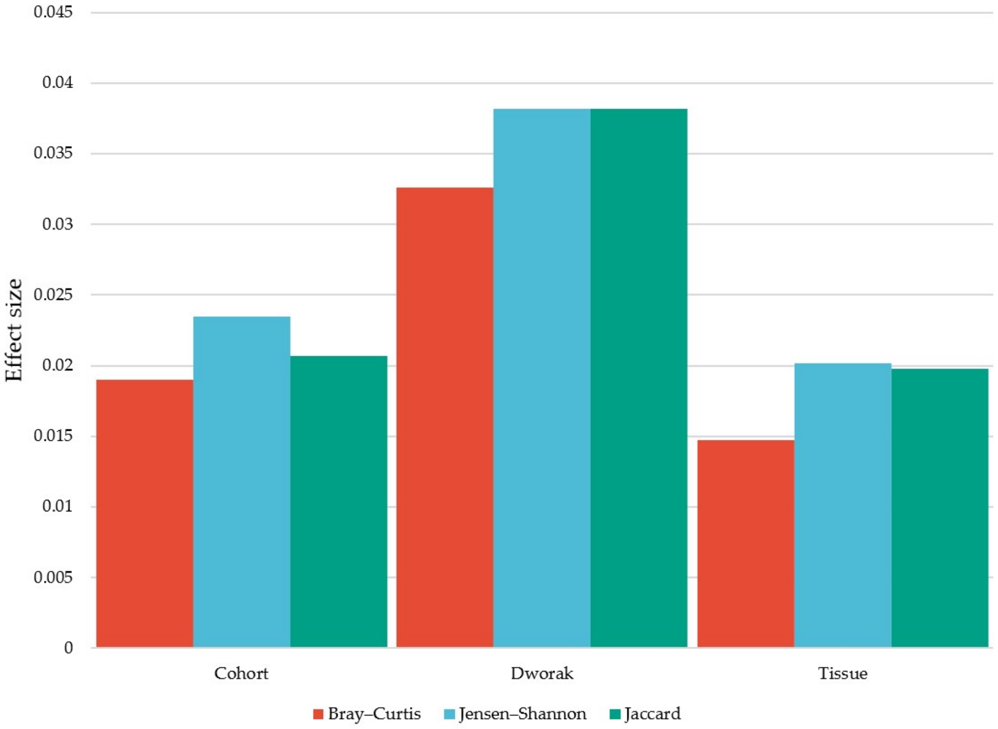

Figure 4.13. Variable effect sizes

As homogeneity of dispersion within variables used is a condition for the use of adonis, homogeneity of dispersion was tested using betadisper. It can be seen that there was significant dispersion within the cohort variable (Table 4.1), so the lack of significance of the above effect size of this variable may not be due to real differences in centroids.

Table 4.1. Variable homogeneity test results

Distance		**Cohort**	**Dworak**	**Tissue**
Bray–Curtis	sum of squares	0.128	0.038	0.055
	mean squares	0.128	0.013	0.055
	p-value	0.020	0.690	0.127
Jensen–Shannon	sum of squares	0.128	0.020	0.018
	mean squares	0.128	0.007	0.018
	p-value	0.019	0.468	0.124
Jaccard	sum of squares	0.147	0.037	0.088

	mean squares	0.147	0.012	0.088
	p-value	0.022	0.694	0.062

4.6 Correlation with response

Species abundance and relative transcription were correlated with response grades to determine which species potentially have a role in treatment outcomes. Taxa were filtered to those with at least 30 reads in 20% of samples, separated into phylum, family, genus and species groups. Their relative abundances were correlated with Dworak scores in respective tissues using Spearman correlation with Benjamini and Hochberg FDR correction.

4.6.1 Results

Within tumour tissues, although no taxa were correlated with Dworak scores overall; *Bacteroides caccae* was positively correlated with a poor response (Dworak One) (Figure 3.14). Although not statistically significant after FDR correction, the phyla Fusobacteria and the family Fusobacteriaceae were negatively correlated with a poor response (Dworak one), while the genera *Escherichia* and the species *B. vulgatus* were positively correlated with a poor response with unadjusted p-values < 0.05 (Figure 3.14).

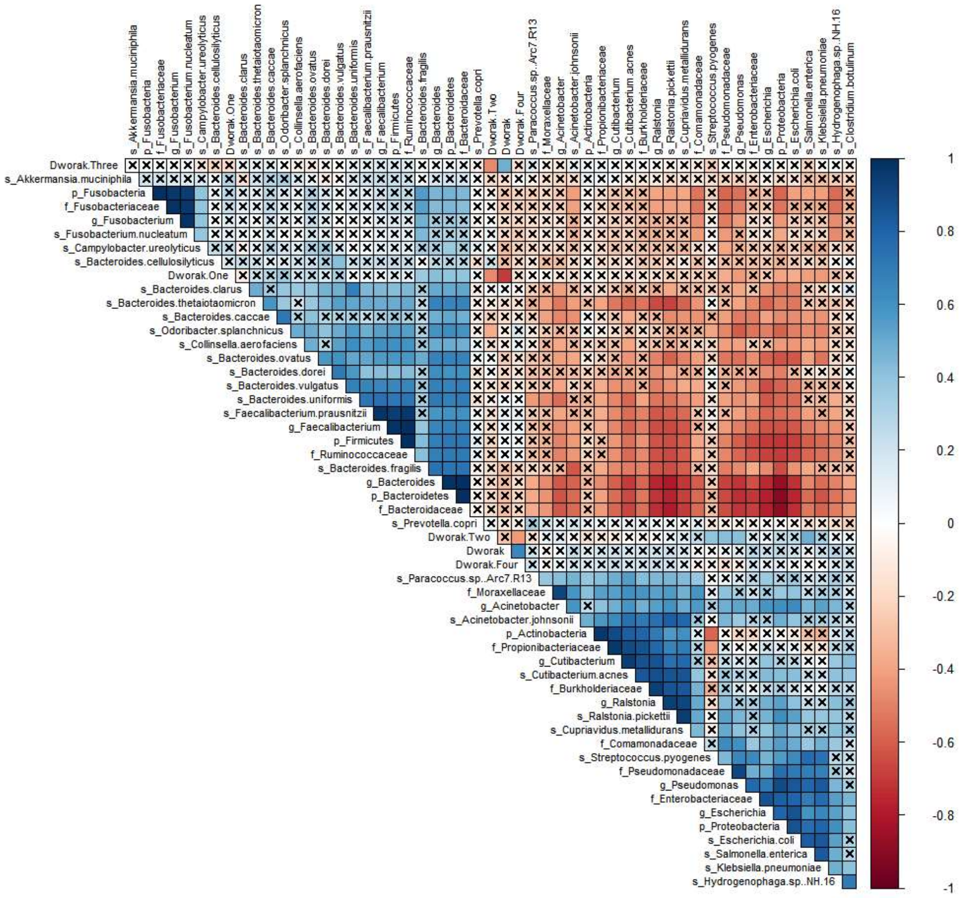

Figure 4.14. Correlation heatmap of tumour tissue taxa and Dworak scores. s_: species; g_: genus; f_family; p_ phylum. X symbols indicate no statistical significant correlation (FDR > 0.05).

Of those taxa with correlations with a poor response (Dworak One), it was seen via scatter plots (Figure 4.15) that the correlation of any taxa with response was only slight and not statistically significant, and that considerable variation existed within response groups, particularly in regards to Dworak Two samples.

Additionally, as Fusobacteria and Fusobacteriaceae were significantly correlated with a poor response, the most transcriptionally active species in the Fusobacterium genus (*F. nucleatum*) was investigated (Figure 4.15f); however, it was not found to have a significant association with response.

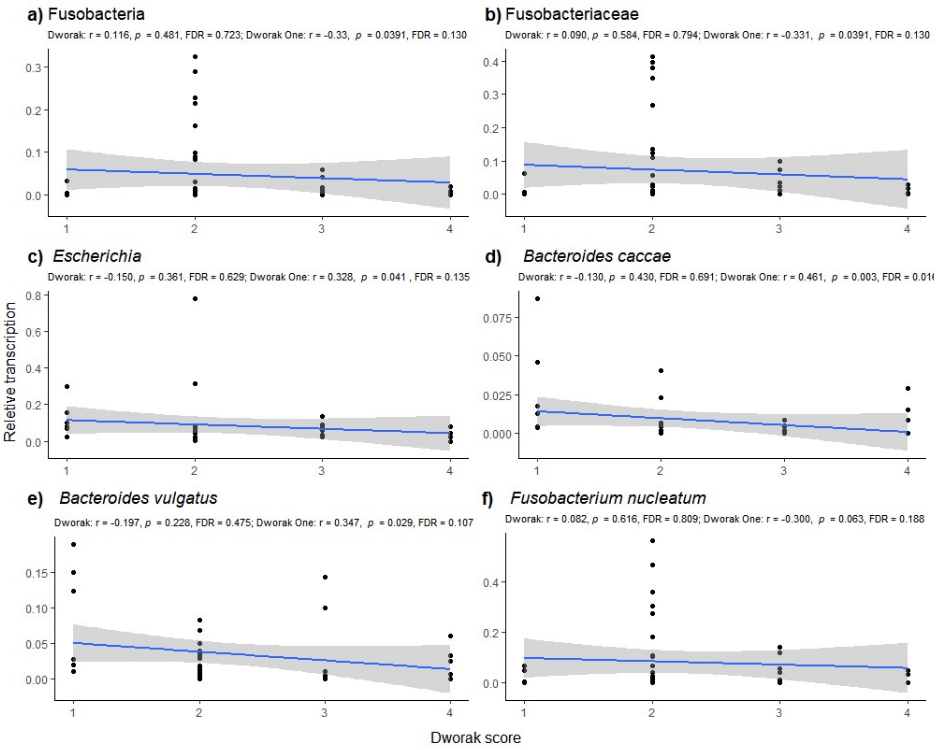

Figure 4.15. Taxa correlated with response in tumour tissues. Blue Line = regression line; grey bars: 95% confidence interval.

C. ureolyticus correlated negatively with Dworak scores within normal tissues and 13 other taxa correlated with poor response. However, after FDR correction, ten taxa correlated with poor response were statistically significant (Figure 4.16).

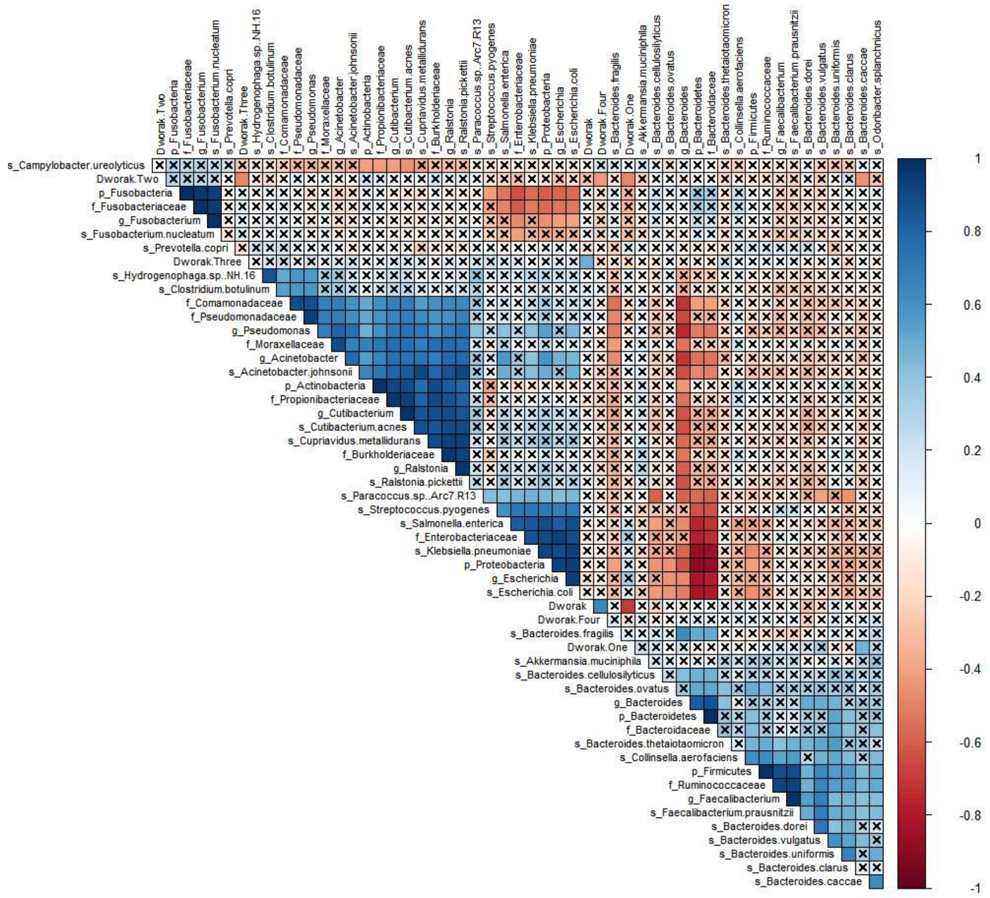

Figure 4.16. Correlation heatmap of normal tissue taxa and Dworak scores.

At the Phylum level, Bacteroidetes and Proteobacteria were positively and negatively correlated with a poor response (0.39 and –0.39, respectively) (Figure 4.17a,b). The families Pseudomonadaceae and Bacteroidaceae correlated with poor response (–0.36 and 0.39, respectively) (Figure 4.17c,d); Enterobacteriaceae was also negatively correlated with poor response, but this was not statistically significant after FDR correction (FDR = 0.068). The *Pseudomonas* and *Bacteroides* genera were negatively (–0.42) and positively correlated (0.41) with poor response, respectively (Figure 4.17e,f).

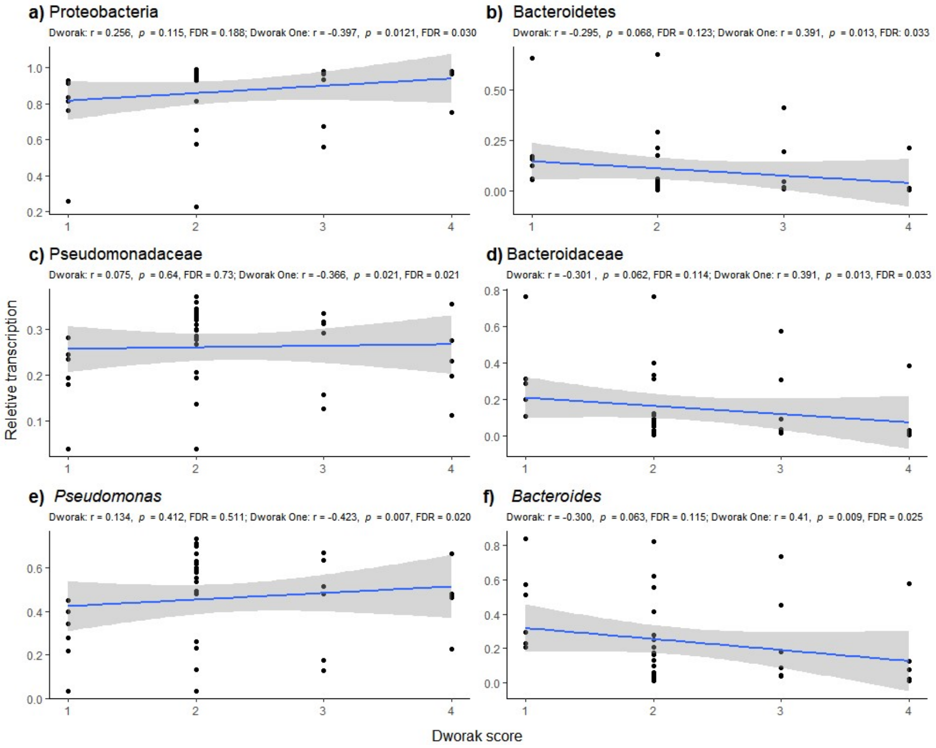

Figure 4.17. Phyla, families and genera correlated with response in normal tissues.

The negative correlations of *E. coli*, *Klebsiella pneumoniae* and *Salmonella enterica* with poor response (–0.36, –0.42 and –0.41, respectively) were statistically significant (Figure 4.18a,b,c). *B. fragilis* and *Odoribacter splanchicus* were positively correlated with a poor response (0.36 and 0.35, respectively); however, the later was not statistically significant after FDR correction (Figure 4.18d,e). *C. ureolyticus*, which was negatively correlated with response had very low transcription levels across samples except for one Dworak Two sample in which considerably higher levels were seen (Figure 4.18e), and after FDR correction, was not found to be statistically significant.

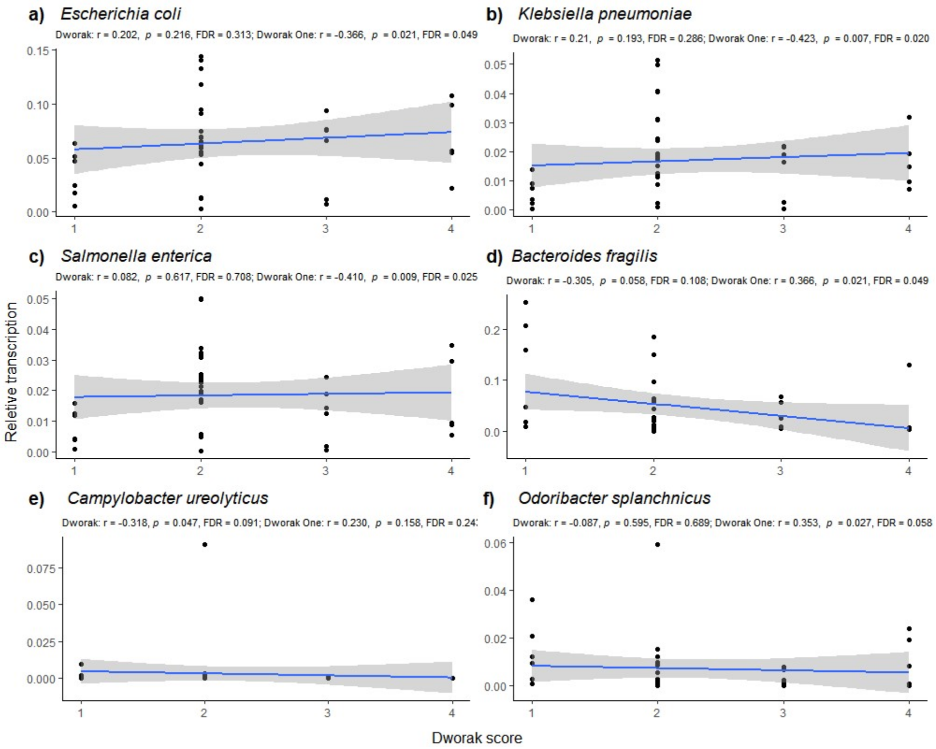

Figure 4.18. Species correlated with response in normal tissue.

4.7 Alignment with bacterial genomes

B. fragilis was the most consistently transcriptionally active species across all samples (RNA-Seq reads: mean = 8557, median = 183, max = 346490, min = 2, SD= 43402). Due to its correlation with poor response, reads assigned to the species were aligned with a *B. fragilis* reference genome to determine the genes being transcribed which may explain its role in response. However, insufficient RNA-Seq reads were present for accurate differential expression analysis [405], with only seven samples having more than 10,000 reads per sample before mapping to the reference genome. All tumour and normal tissue reads were aligned to the *B. fragilis* genome to determine if the *bft* toxin was being expressed. As can be seen in Figure 4.19, no expression of the *bft* gene was detected at the time of sampling.

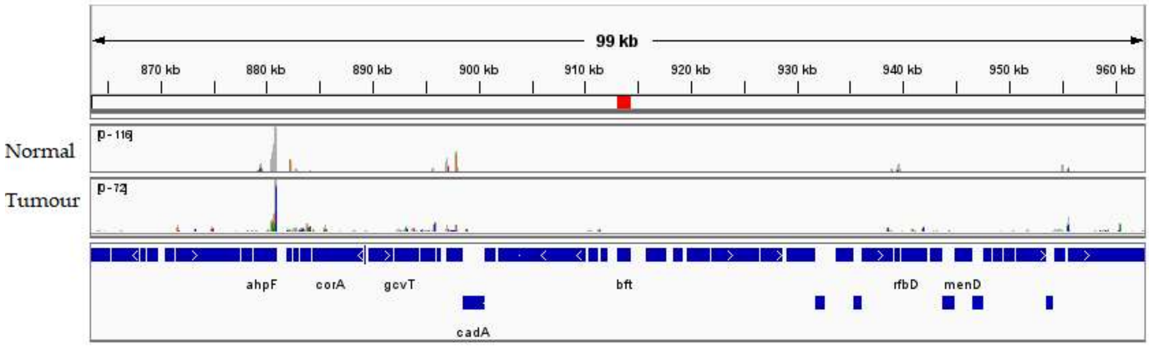

Figure 4.19. Alignment of all Bacteroides fragilis aligned RNA-Seq reads. Red bar indicates the locus of the bft gene.

4.8 Discussion

4.8.1 Alpha-diversity

Alpha-diversity between tissues and response groups was not statistically significant in most instances. However, Shannon diversity differences were statistically significant between Dworak One and Two normal tissues, which could be due to two factors: a) the comparative differences in the number of samples; b) the differences found between the poorest responders (Dworak One) and moderate responders (Dworak Two) may be due to the more selective microenvironment compared to higher responders (which did not have statistically significant differences). Of these, the evidence provided would suggest a), as the impact of the microenvironment on diversity could not be confirmed with such small and uneven sampling. Additionally, when grouping into high and low response groups, no statistically significant difference was seen, supporting the idea that the difference in sample sizes contributed to the outcome.

There was a statistically significant Simpson diversity difference between the tumour and normal tissues of Dworak Two patients; again, the difference may be due to the higher number of samples in the Dworak Two group (22 of 40, 55%) allowing greater discrimination between respective tissues. However, as with Shannon diversity, when combined with Dworak One patients into the low response group, there was no statistically significant difference between the low and high responders, again indicating that the statistically significant differences resulted from the insufficient sample sizes in this study. Alternatively, the difference between Dworak One and Two patient tissues' selective microenvironments may make grouping them inappropriate.

However, under the assumption that the statistically significant findings between Dworak Two tissues and between Dworak Two and One normal tissues are valid, the results would indicate that there may be increased intrapersonal tissue differences in moderate responders. Simultaneously, the increased diversity in the normal tissues of the poorest responders (Dworak One) may contain additional taxa from the more selective radioresistant tumour environment that does not occur in less radioresistant tumours.

4.8.2 Filtering

For correlation and differential expression analyses, taxa were filtered to those represented by 30 reads in at least 20% of samples, while for beta-diversity analysis, taxa with read counts of five or higher in at least 20% of samples were retained. The less stringent filtering for beta-diversity analysis was done to retain more variability between sample groups, while in correlation and expression analyses, only the most accurately identified taxa were retained.

The filtering threshold was chosen to eliminate taxa that would be uninformative, while still retaining some rarer taxa, as excessive filtering would leave only the core microbiome. Filtering thresholds have implications for microbiome studies overall, as to how researchers decide to group, filter and display sample data can impact results and their interpretation. As such, caution is advised when interpreting study results utilising less than a few hundred samples per group, as small sample sizes (as used here) can contribute to the loss of taxa from filtering that would otherwise be informative to study outcomes [406].

4.8.3 Composition and differential transcription

Proteobacteria transcription dominated most samples, the abundance of which has been used as a marker of dysbiosis and disease in the gut [407]; however, this pertains mainly to studies performed in faecal samples, the results of which are not directly applicable to tissue studies [164]. As many pathogenic taxa are found in the Proteobacteria phylum, the compromised and inflammatory gut-tissue environment of cancer patients may allow the expansion of such taxa, allowing discrimination between response groups microbiome variation between individuals can often be attributed to interpersonal differences between Proteobacteria taxa [408].

Cyanobacteria were the only phyla to have statistically significant higher transcriptional activity in tumours, and these differences were higher in high responders than low responders. Cyanobacterial toxins have been identified as contributing to modulation of the innate immune system in the mucosa, and some have been identified as producing carcinogenic compounds [409]. Additionally, species of Cyanobacteria used in food and supplements such as *Arthrospira* sp. are known to have radioresistant properties [410]; however, how this might contribute to tumour radiosensitivity is unknown.

When comparing differentially transcriptionally active genera, *Campylobacter* was found to be five-fold higher in tumour than normal tissues, across response grades; however, this is likely due to contributing to and therefore survivability in inflammatory environments [411], and maybe an indicator of tissue inflammation more generally. Additionally, the *Campylobacter* species most prominently transcribed in this study, *C. ureolyticus*, was not evenly transcribed across samples, indicating that a few samples may drive this result.

More interestingly, when comparing the differences between the tumour differences between high and low responders, it was found that *Hungatella* and *Butyricimonas* were 6–7-fold and *Flavonifrator* and *Oscillibacter* 3-fold more transcriptionally active in high responder tumour tissues compared to normal tissues, compared to the differences between tissues of low responders. These results indicate that relative differences between the transcription of these genera between tumour and normal tissue biopsies could be used prognostically to indicate tumour regression outcomes.
Hungatella are anaerobic bacteria that have been associated with CRC in the past, particularly in consensus molecular subtype 1 (CMS1), characterised by immunogenicity and hypermutation [93]. Additionally, in a study by Xia et al., *Hungatella hathewayi* colonisation in mice promoted epithelial cell proliferation, immunoreactivity, and the methylation of *SOX11*, *THBD*, *SFRP2*, *GATA5*, and *ESR1* tumour suppressor genes, and could be contributing to hypermethylation more generally [412]. In addition to *Hungatella hathewayi*, Xia et al. also found that and *Fusobacterium nucleatum* was associated with increased methylation of *MTSS1*, *RBM38*, *PKD1*, and *PTPRT* [412]. The above may have implications for response to tumour regression.

Indeed, these two species were identified in this study, along with *Bacteroides thetaiotaomicron*, to have lower differential expression in tumour tissues compared to normal tissues in low responders. However, *B. thetaiotaomicron* has been reported to reduce inflammation in irritable bowel disease cell and animal models [413]. Additionally, *H. hathewayi* was seen to be more than 5.5-fold lower in the normal tissues of high responders compared to low responders, indicating that they may contribute to diverting increased immune activity away from the tumour site in low responders. However, this would require further investigation.

In this study, high responder tumours contained more than three-fold higher *B. faecalis* compared to low responders and were found to have higher transcription in high responder tumours relative to normal tissues, and these differences being five-fold higher compared to low responders. *Butyricimonas*, are anaerobic butyrate producers, of which one species has been implicated in causing septic shock in a CRC patient [414], indicating that they may have high immunogenicity in addition to their butyrate production. Butyrate can upregulate PPARγ signalling maintains epithelial hypoxia by driving mitochondria toward β-oxidation of fatty acids and can act as an energy source for colonocytes and other bacterial species, such as *Salmonella enterica* [415].

Flavonifractor has been reported as more abundant in healthy controls when studying CRC microbiomes [416, 417], and its higher transcription was identified here as being more than three-fold higher in tumour compared to normal tissues of high responders compared to low responders. The above indicates that the tumour microenvironment of high responder tumour tissues may be more similar to healthy tissues than poor responders. The effect of *Flavonifractor* on response may involve an immune regulatory function, as oral supplementation in of *F. plautii* has been found to suppress Th2 pro-inflammatory immune responses [418].

A. finegoldii was the most prominently transcribed species in the *Alistipes* genus and was found to have higher differential transcription in the tumours of complete responders, three-fold higher transcription in tumour than normal tissues in high responders, compared to the differences of tumour and normal tissues of low responders. *A. finegoldii* is a propionate producer [419], a SCFA that that can stimulate PYY and GLP-1 in colonic cells, reducing energy intake [420] and can utilise saturated fatty acids for membrane production [421]. *Oscillibacter* transcription was detected in high responders' tumours, which has been correlated with high-fat diets and reduced intestinal barrier function [422]. In conjunction, the above species gives further evidence for the benefit of high-fat diets on radiotherapy outcomes, as has been demonstrated in randomised controlled studies on ketogenic and high-fat diets in radiotherapy of RC [423], having benefits to quality of life and body mass in gastrointestinal cancers [424]. The results here indicate a microbial metabolism component that may impact outcomes in the tumour microenvironment; however, this requires further investigation as the taxa may be selected for by a beneficial diet, rather than having an impact on response themselves.

The above evidence may suggest that the contribution to CRC by *H. hathewayi* and *F. nucleatum* via methylation of tumour suppressor genes may result in increased radiosensitivity due to enhanced immunogenicity and methylation; however, this requires further investigation, particularly with the similarly associated *Flavonifractor*, and *B. thetaiotaomicron* having a role in ameliorating inflammation. Additionally, it may be possible that high-fat diets may encourage the colonisation of tissues with *Oscilliobacter* and *A. finegoldii*, altering tumour metabolism in conjunction with butyrate produced by *B. faecalis*, pushing tumour metabolism toward ketogenic metabolism, and starving glucose dependant tumours, reducing cellular proliferation [425]. However, although ketogenic diets are generally well tolerated, further investigation is required on the interactions between radiotherapy, the microbiome and ketogenic diets in RCs [426].

4.8.4 Beta diversity and effect sizes

Beta diversity was measured using Bray–Curtis (BC) Jensen–Shannon distance (JSD), and the Jaccard index (JI). It was not possible to use UniFrac measures as they rely on phylogenetic relationships between assigned reads. Due to the random nature of total RNA-sequencing from tissue samples, the lack of phylogenetic relationship between different reads from the same species would result in heavily distorted results.

There was little difference between the BC distance and the JI clustering of tumour and normal tissues or by Dworak score, indicating that incorporating transcriptional abundance does not substantially affect clustering of samples beyond that established by the presence or absence of unique taxa. All beta-diversity analysis showed that patient tumour tissues might be more diverse than normal tissues; however, the level of interpersonal variability and cluster overlap suggests this is not always the case.

Effect sizes showed that using beta diversity, Dworak scores and tissue variables explain little of patients' microbial transcription variability. However, this leaves the majority of variability to be explained by other unknown factors, of which genetic dissimilarity and differences in environmental exposure between patients are likely the major contributing factors.

4.8.5 Correlations with response

Only *Campylobacter ureolyticus* transcription was correlated with Dworak score in normal tissues; however, this was not found to be statistically significant and was likely driven by two samples with higher than normal levels. Although no statistically significant taxa were found to be correlated with Dworak scores, several were seen to be correlated with a poor response (Dworak One), but not conversely negatively correlated with a complete response (Dworak Four). The phylum Fusobacteria and family Fusobacteriaceae were both correlated with poor response, but not *F. nucleatum*, the most transcriptional active member species. Similarly, the Escherichia genus was positively correlated with poor response, but not *E. coli*, its most transcriptionally active member species within tumour tissues; however, it was within normal tissues. This effect may be due to a species-level assignment limitation as multiple closely related species were included in the database. As most reads assigned to Fusobacteria were also assigned to Fusobacteriaceae, only some reads likely had species-specific content; most reads assigned to a family or genus would be largely homologous between member species, which indicates that only a minority of reads produced from species-specific regions can discriminate between species.

Two Bacteroides species, *B. vulgatus* and *B. caccae,* were also positively correlated with a poor response. *B. vulgatus* is known to reduce inflammation by reducing microbial lipopolysaccharide production [427], while *B. caccae* has been implicated in appendicitis, intra-abdominal infections [428] and ulcerative collitis [429], indicating it may contribute to anti-bacterial inflammatory immune responses. The Bacteroidetes phylum, the Bacteroidaeceae family, and the Bacteroides genus were positively correlated with a poor response within normal tissues. *B. fragilis* was the only species in the Bacteroides genus that was significantly correlated with a poor response. *B. fragilis* strains have been implicated in CRC due to bft enterotoxin expression [430], while other strains have been suggested as a probiotic and have been shown to enhance phagocytosis and polarisation of macrophages to the M1 phenotype [431], which would improve tumour clearance. As the most abundant species with consequential strain differences, the reads assigned to *B. fragilis* were extracted and aligned to a representative genome to determine the differential expression occurring in different tissues and response groups. Although the number of reads was too small to identify differentially expressed genes, it was possible to determine that no bft expression was occurring in any samples, indicating that if enterotoxigenic strains were present, they could not actively express the bft toxin. Due to the limitation in the read depth and sampling consistency across samples, it cannot be said what strains of *B. fragilis* are involved or what role they have in radiotherapy outcomes and should therefore be further investigated.

Salmonella enterica was negatively correlated with a poor response; however, there was high interpersonal variability. The species *S. enterica*, similar to *B. fragilis*, has several subdivisions which may have different roles in environments [432], and assignment could not be resolved to the subspecies level. However, considering the previously discussed ability to feed on butyrate, one hypothesis for its association with response is that in poor responders, it can metabolise available butyrate, depriving host tissues of its effects, while in higher responders, it may be selected for higher levels of butyrate, but its metabolism does not deplete the pool available for host uptake. Additionally, *Salmonella* species have been suggested as a novel anti-cancer therapy, showing some success in animal models alone and in conjunction with radiotherapy, through a direct growth inhibition of tumours and immune mediation, inhibiting Tregs and increasing CD8 T cell populations [433, 434]. Similarly, *E. coli* were identified as being negatively correlated with poor response and had substantial amounts of variability between different strains, which makes forming a hypothesis around the mechanism of effect on response difficult. However, as all *E. coli* are facultative anaerobes, they may reduce the available oxygen in the microenvironment. Additionally, as they are one of the fastest replicating prokaryotes [435], their relatively high transcription may result from the temporary environment of the colonoscopy pre-treatment, after which the biopsies were collected.

Although both the Pseudomondaceae family and the *Pseudomonas* genus were negatively correlated with poor response, no *Pseudomonas* species was correlated with response. *Pseudomonas* sp. NC02 transcription was significantly higher in low responders' tumours compared to those of high responders; however, given that *P.* sp. NC02 is a soil bacteria with no evidence of human gut colonisation, and that the database used for taxonomic assignment contained 127 species and strains of the *Pseudomonas* genus, it is possible that this species was likely misidentified, largely due to its 99% similarity to both *P. yamanorum* and *P. fluorescens* [436]. *Odoribacter splanchnicus* was also negatively correlated with poor response. However, due to its high inter-sample transcriptional variability (mean reads assigned per sample: 340; SD:

1353) and that it is closely related to other significantly transcriptionally active species in the genus *Butyricimonas* [437], *O. splanchnicus* may have also been misidentified. Although identification of *O. splanchnicus* is more likely than *P.* sp. NC02, as it has been reported in the human gut [438].

Overall, since taxa were correlated with a poor response but not inversely correlated with a complete response or Dworak scores, it may be possible that taxa can contribute to or are selected by the tissue microenvironment of poor responders, but not of higher responder groups. It could be that the selective environment of poor responders is more extreme, with the more subtle selective pressures of other responder groups not being assessed with the small sample sizes employed in this study.

4.8.6 Limitations

A strong limitation of this study was the unbalanced design, using two low powered cohorts, and more samples having a Dworak Score of two than any other group. Furthermore, the inability to perform differential gene expression on individual taxa or to resolve them beyond a species designation leaves only speculative functional contributions to tumour regression. Transcriptional sequencing and taxa assignment error may play a role in the results of this study. Finally, the speculative cohort effect may have further affected the data and subsequent interpretation of results.

4.9 Conclusions

Based on the results of this study, microbial alpha or beta diversity does not contribute to enhanced or decreased response rates to radiotherapy in RCs. However, the increased transcription of individual species such as *Hungatella hathewayi*, *Fusobacterium nucleatum*, *Butyricimonas faecalis*, *Alistipes finegoldii*, *Bacteroides thetaiotaomicron*, and *B. fragilis* may contribute to response by modulating metabolism and anti-tumour immune responses. Additional research is required with higher sample sizes, more per-response group power and higher bacterial read depth; to resolve contributing species to strain designations and to evaluate the contribution of differential gene transcription within those strains to radiotherapy outcomes.

5 CHAPTER 5: IMMUNE CELL INFILTRATION

5.1 Introduction

The immune system's role in response to therapy is widely accepted [439], with the field of immunotherapy taking advantage of this fact [440]. Neo-antigens produced in damage-associated molecular patterns (DAMPs) are thought to trigger immune responses against tumour tissues, with natural killer T cells and CD8 T cells being most implicated in cytotoxicity induced tumour regression [441].

More recently, macrophages and dendritic cells have also been implicated, with different phenotypes having alternative responses [442]. So-called tumour tolerant cells such as D0 dendritic cells are implicated in preventing the immune system from taking action against tumour cells [443]. M2 tumour-associated macrophages (TAMs) have paradoxical roles in increased tumour growth and fibrosis, while M1 or tumour-sensitive macrophages have been identified as having a tumour clearing role [444, 445]. The latter action of these macrophages may be linked to their role in tissue remodelling actions [446], such as differentiating from macrophages to osteoclasts [447]. Evidence for this role comes from the lipopolysaccharide (LPS) and RANK-L induced differentiation to osteoclastic phenotypes, as well as the association of high osteonectin with higher responses to radiotherapy [448] [449]. A study in mice has shown that osteonectin deletion decreased M1 macrophage levels and fibrosis in heart tissues. Overall, the implications of this remain to be thoroughly investigated [450].

The immune system's interaction with the microbiome has been a source of much contention in the research community over the last decade. Irritable bowel syndrome (IBS), inflammatory bowel disease (IBD) and other inflammatory diseases are associated with dysbiosis, which can cause inappropriate immune responses, resulting in pathologies [451]; suggesting that the microbiome can modulate immune activity in the gut [289]. Additionally, different cytokine levels have been associated with disruptions to the microbiome, e.g., IL-6 and IL-10, implicated in autoimmune reactions such as colitis and Crohn's disease [452].

In this chapter, the aim was to measure the abundance of different immune cells within RC biopsies of tumour and normal tissues and test if they are correlated with response. Immune cells correlated with response to therapy were then correlated with the abundance and transcription of identified bacterial species within the respective tissues. The hypotheses of this chapter were:

- Particular immune cells present in tumour tissues modulate the response to chemoradiotherapy (CRT) in RCs.
- The transcription of different bacteria partly regulates relevant immune cells.

5.2 Methods

The Salmon tool [287] was used to quantify host gene expression using RNA-Seq data from the RC cohort (*n* = 40), using the GRCH38p12 transcriptome. The relative levels of immune cells in each sample were predicted using 500 permutations of the digital cytometry tool, CIBERSORT, which utilises support vector regression (SVR) to compute the proportion of immune cells based on gene expression data [284]. The ESTIMATE R package v1.0.13 was used on RC cohort RNA-Seq gene expression data to provide immune and stromal scores, while the ESTIMATE score was used to assess tumour purity [282]. Tumour purity was calculated using the formula: Tumour purity = cos

(0.6049872018+0.0001467884 × ESTIMATE score). The significance of differences between tissue and response groups was measured using Wilcoxon signed-rank tests with Benjamini and Hochberg False discovery rate (FDR) correction [453] and Kruskal–Wallis tests.

Spearman rank sum correlation with Benjamini and Hochberg correction was used to establish a correlation between the level of relative immune cells and response. Bacterial transcription levels were evaluated as described in the Methods chapter (Section 2.6), with microbial RNA expression being assigned taxonomy with Kraken 2 v2.0.7[338]. Taxa were filtered for the most transcriptionally active bacteria with a cut-off of at least 30 reads in 20% of samples, which were converted to relative levels. Spearman correlation was used to ascertain the relationships between taxonomic abundance and transcription levels using Benjamini and Hochberg for FDR correction. Scatterplot regression lines were built using the linear-model (lm) method in the geom_smooth() function of ggplot2 v3.3.2 [315].

Scripts and code used for the analyses and supplementary tables can be found at https://github.com/William-S-Taylor/MSc.

5.3 Results

5.3.1 Immune cell profiling

Between tumour and normal tissues, there were significantly lower levels of B memory cells (4.67-fold), CD8 T cells, (2.77-fold), plasma cells (1.86-fold) detected in tumour tissues. In comparison, there were significantly higher levels of activated mast cells (2.3-fold), M0 macrophages (2.6-fold), neutrophils (3.32-fold) and M1 macrophages (8.76-fold) (Figure 5.1, Supplementary Table 5.1).

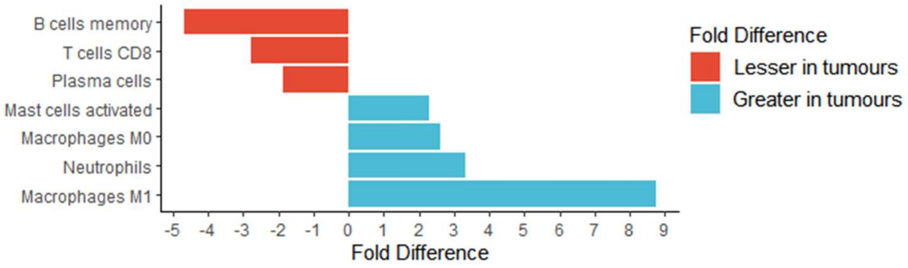

Figure 5.1. Statistically significant fold-change differences between immune cells in tumour tissues and normal tissues.

There were statistically significant differences between tumour and normal tissues in the abundance of B memory cells, plasma cells, CD8 T cells, activated mast cells, M0 macrophages, neutrophils and M1 macrophages (Figure 5.2a–g). The abundance of plasma cells, activated mast cells and CD8 T cells had the greatest overlap between tissues; however, in terms of the abundance of B memory cells, activated mast cells and neutrophils, there were several outliers.

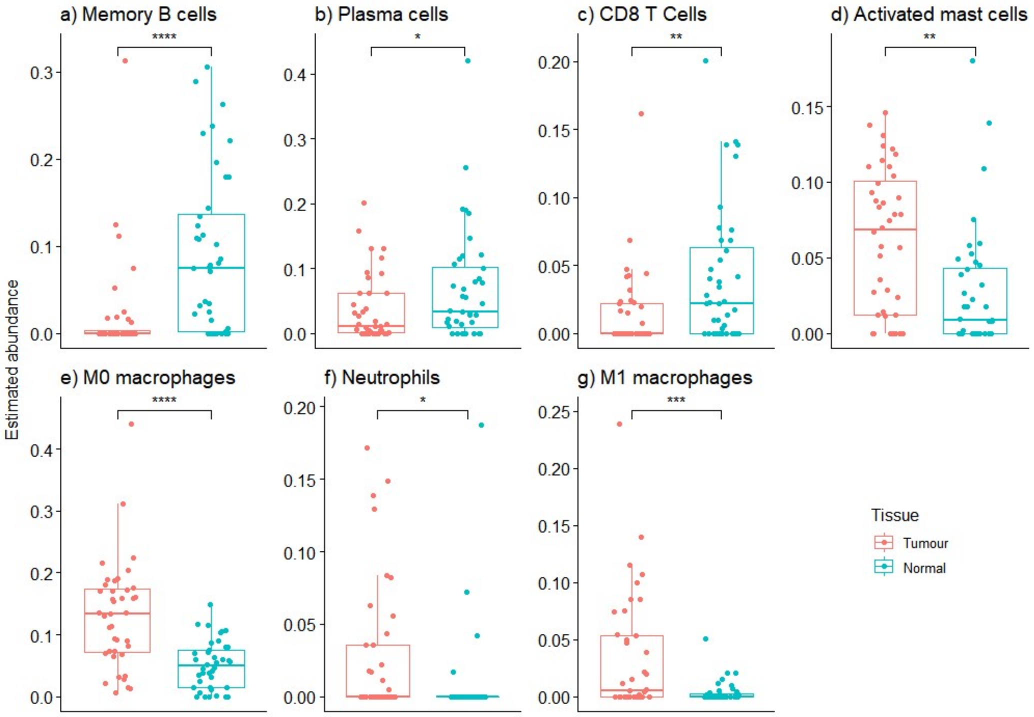

*Figure 5.2. Abundance of immune cells significantly differentially abundant between normal and tissues. a) memory B cells; b) plasma cells; c) CD8 T cells; d) activated mast cells; e) M0 macrophages; f) neutrophils; g) M1 macrophages. * p <= 0.05, ** p <= 0.01, *** p <= 0.001, **** p <= 0.0001.*

Within Dworak groups, the differences between tissue types were not consistent. The largest group, Dworak Two (*n* = 22), had significant differences between tissues for the same immune cells as the entire cohort; however, no other Dworak group had statistically significant differences between tissues.

There were significant differences between detected immune cells in the tumour tissues of complete responders (Dworak Four) and the average abundance of incomplete responders (Dworak One–Three) in the abundance of Tregs, resting and activated mast cells (**Error! Reference source not found.**). The highest differences were seen for M1 macrophages (**Error! Reference source not found.**). There were no statistically significant differences between normal tissues (Supplementary Table S5.2).

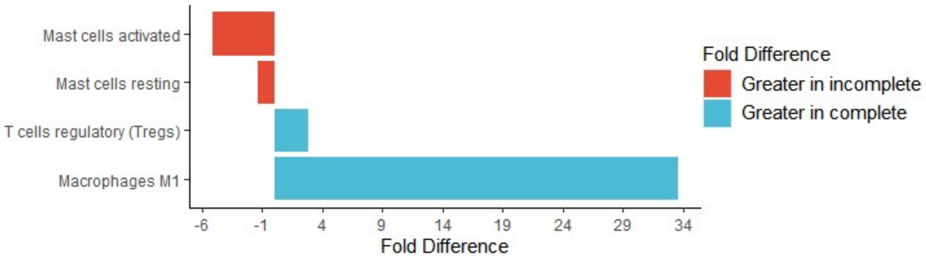

Figure 5.3. Significant fold differences between the detected immune cells in the tumour tissues of complete and incomplete responders.

Dworak scores were grouped into a high response group (Dworak Four and Three) and a low response group (Dworak Two and One). No detected immune cells were significantly different between the tumour or normal tissues of the two groups, after FDR correction.

However, there were significant differences in the abundances of resting and activated mast cells (Figure 5.4a,b), Tregs (Figure 5.4c) and M1 macrophages (Figure 5.4d) between the tumour tissues of complete responders (Dworak Four) and incomplete responders (Dworak One–Three).

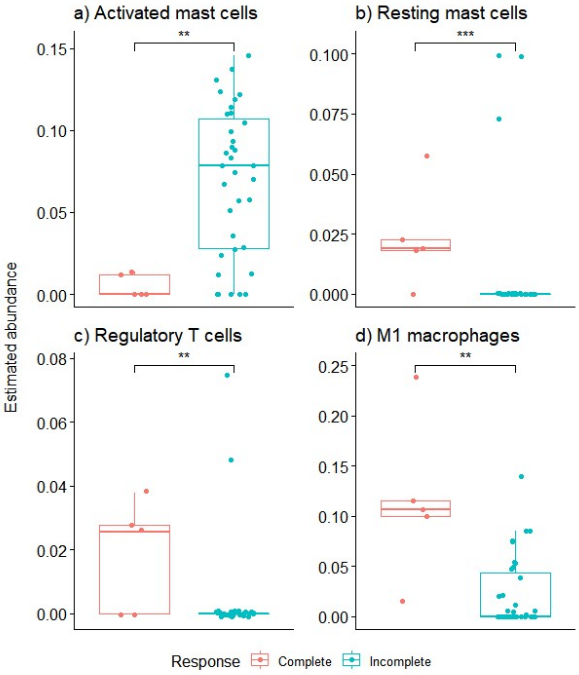

*Figure 5.4. The abundance of immune cells significantly differentially detected between tumours of complete and incomplete responders. a) activated mast cells; b) resting mast cells; c) T regulatory cells (Tregs); d) M1 macrophages. ** $p < 0.01$; *** $p < 0.001$.*

5.3.2 ESTIMATE scores

ESTIMATE was used to establish tumour purity within tumour samples, giving stromal and immune scores, and an ESTIMATE score used to measure of tumour purity. However, there was no statistically significant correlation between tumour purity, stromal or immune scores and Dworak scores, before or after FDR correction (Supplementary Table S5.2).

5.3.3 Immune cell correlations with response

After FDR correction, no immune cells were significantly correlated with Dworak scores in normal tissues; however, within tumour tissues, M1 macrophages and resting mast cells had statistically significant positive correlations with Dworak scores of 0.46 and 0.51, respectively (Figure 5.5).

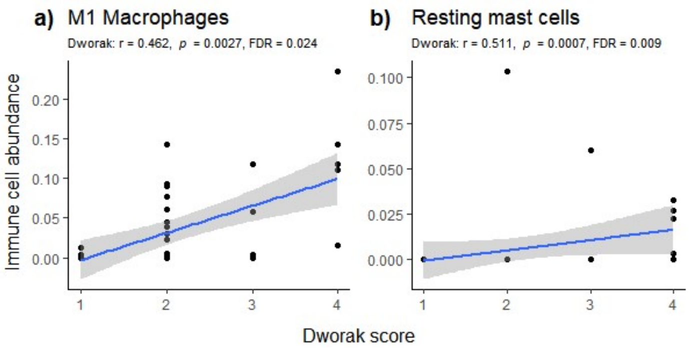

Figure 5.5. Immune cell abundance vs Dworak scores in tumour tissues. a) relative abundance of resting mast cells, a) M1 macrophages, b) resting mast cells. Blue line: regression line; grey bars: 95% confidence interval.

The significance of differences between immune cells significantly correlated with Dworak scores in tumours was tested between response and tissue types (Figure 5.6).

There were statistically significant differences between the M1 macrophage levels in the tumour and normal tissues of Dworak Four and Dworak Two patients and between Dworak Four and Dworak One–Three tumour tissues (Figure 5.6a). Comparisons between relative levels of resting mast cells in tumour tissues showed statistically significant differences between the tumour and normal tissues of Dworak One and Dworak Two patients, and between Dworak Four and Dworak One and Two tumours (Figure 5.6b).

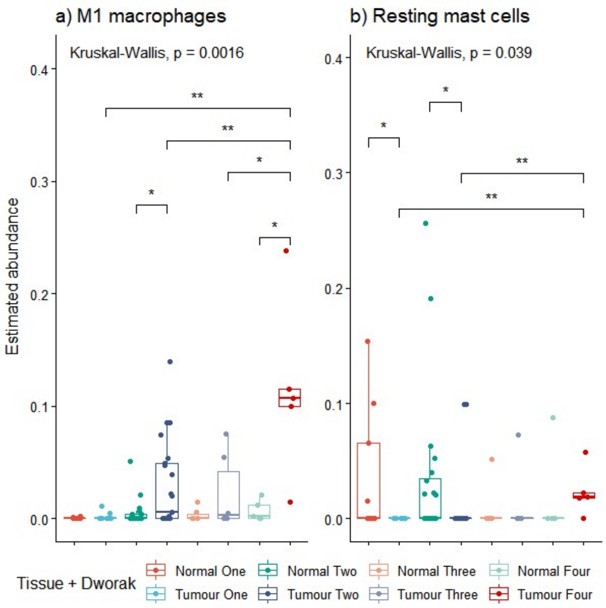

*Figure 5.6. Abundance of immune cells correlated with response in tumour cells. a) M1 macrophages, b) resting mast cells. * p < 0.05; ** p < 0.01.*

5.3.4 Correlation with the microbiome

In Chapter 4, the microbiome was correlated with chemoradiotherapy response, in a possibly immune-mediated manner. The relative transcription of microbial taxa was correlated with the abundance of immune cells to investigate possible interactions between the microbiome and the immune system. Spearman correlation with Benjamini and Hochberg FDR correction was used to find statistically significant correlations between bacterial transcription with immune cell abundance. Taxa with more than 30 reads across at least 20% of samples were retained for the correlation analysis, with a focus on taxa correlating with response in Chapter 4.

5.3.5 Bacteria–immune correlations

M1 macrophages correlated with Dworak scores within tumour tissues but did not have any statistically significant correlation with bacterial transcription. M1 macrophages were positively correlated with *Campylobacter ureolyticus*, Campylobacteraceae, and negatively with Cyanobacteria, *Bacteroides vulgatus*, *Desulfovibrio*, *Alistipes finegoldii* and

Alistipes; however, these correlations were not statistically significant after FDR correction (Supplementary Figure S5.1).

M1 macrophages were negatively correlated with Pasteurellaceae in normal tissues (r = -0.439) (Figure 5.7a). Additionally, *Bacteroides vulgatus* was negatively correlated with resting dendritic cells within tumour tissues (Figure 5.7b). In contrast, activated dendritic cells were correlated with *Clostridium saccharobutylicum*, *Salmonella enterica* and *Escherichia coli* and its parent genus were positively correlated with activated dendritic cells (Figure 5.7c–f). Activated mast cell levels in tumour tissues were negatively correlated with Dworak scores (r = 0.303, *p* = 0.04); however, were not statistically significant after FDR correction (FDR adjusted *p*-value = 0.15).

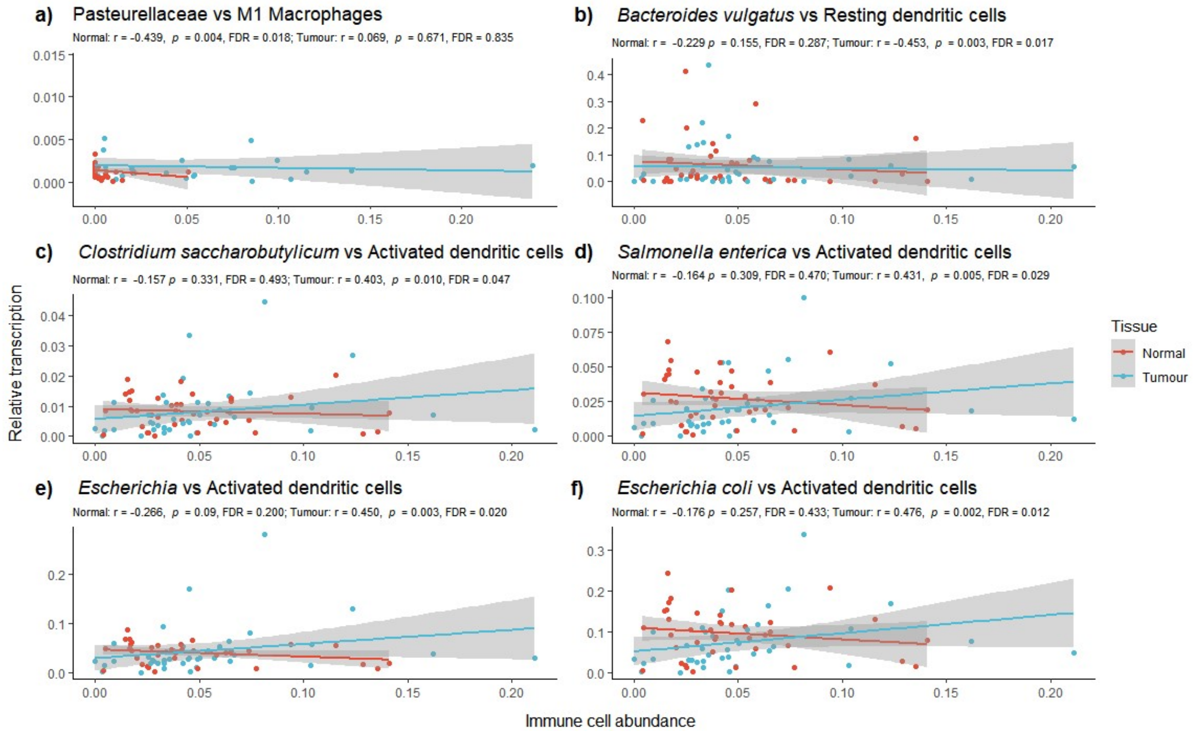

Figure 5.7. Scatterplot of M1 macrophages, dendritic cells and correlated bacterial transcription. a) Pasteurellaceae vs M1 macrophages; b) Bacteroides vulgatus vs Resting dendritic cells, c) Clostridium saccharobutylicum vs Activated dendritic cells; d) Salmonella enterica vs Activated dendritic cells; e) Escherichia vs Activated dendritic cells; f) Escherichia coli vs Activated

dendritic cells. r: Spearman's Rho; FDR: false discovery rate adjusted p-value. Lines of regression are coloured by tissue type; grey bars indicate 95% confidence intervals.

Activated mast cells were positively correlated with *Bacteroides dorei* transcription (Figure 5.8a), while activated mast cells were positively correlated with resting mast cells (Figure 5.8b). Bifidobacteriaceae was negatively correlated with resting mast cells, but in normal tissues (Figure 5.8c).

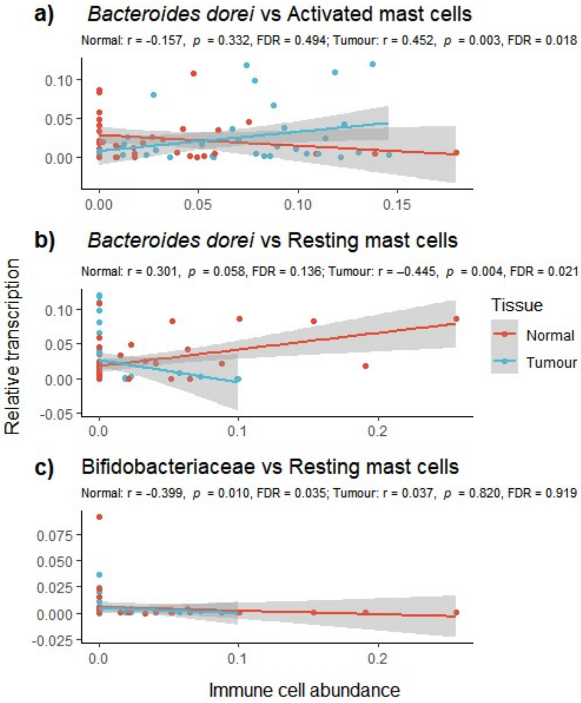

Figure 5.8. Scatterplot of mast cells and correlated bacterial transcription. a) Bacteroides dorei vs Activated mast cells, b) B. dorei vs resting mast cells, c) Bifidobacteriaceae vs Resting mast cells. r: Spearman's Rho; FDR: false discovery rate adjusted p-value. Lines of regression are coloured by tissue type; grey bars indicate 95% confidence intervals.

The transcription of *Bacteroides fragilis* was correlated with naïve B cells and CD8 T cells in tumour tissues (Figure 5.9a,b), while *Eubacterium rectale* was correlated with eosinophils and resting natural killer (NK) cells (Figure 5.9c,d). *Streptococcus pyogenes* transcription was positively correlated with follicular T helper cells in normal tissues

(Figure 5.9e) while the transcription of the family Flavobacteriaceae was in tumour tissues (Figure 5.9f).

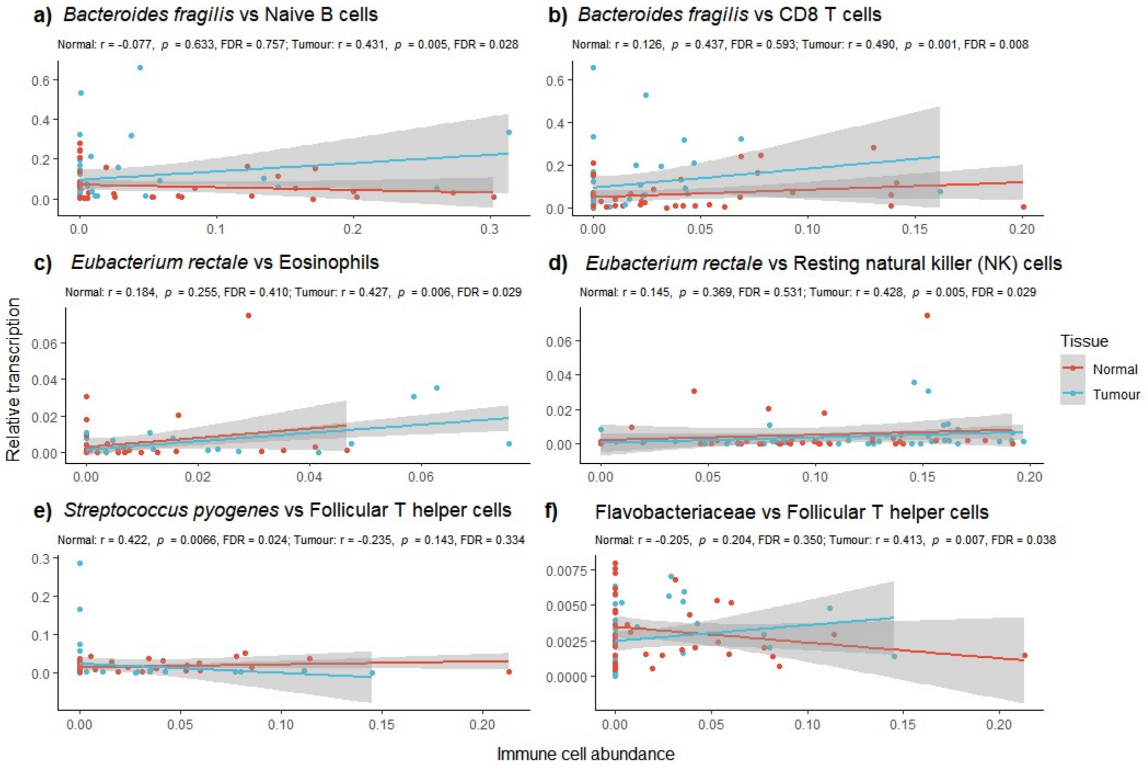

Figure 5.9. Scatterplot of B, T and eosinophils and correlated bacterial transcription. a) Bacteroides Fragilis vs Naïve B cells; b) B. fragilis vs CD8 T cells; c) Eubacterium rectale vs Eosinophils; d) Eubacterium vs Resting natural killer cells; e) Streptococcus pyogenes vs Follicular T helper cells, f) Flavobacteriaceae vs Follicular T helper cells. r: Spearman's Rho; FDR: false discovery rate adjusted p-value. Lines of regression are coloured by tissue type; grey bars indicate 95% confidence intervals.

Neutrophils were the immune cell which most correlated with bacterial transcription in normal tissues. Although neutrophils had a small negative correlation with Dworak scores in normal and tumour tissues (r = -0.191 and -0.032, respectively), this was not statistically significant (*p* = 0.236, FDR = 0.533; *p* = 0.841, FDR = 0.906, respectively). In normal tissue, the transcription of the phyla Proteobacteria and Bacteroidetes were negatively and positively correlated with the abundance of neutrophils, respectively (Figure 5.10a,b). *E. coli* transcription and that of the parent genus *Escherichia* were negatively correlated with neutrophils in normal tissue (Figure 5.10c,d), as well as *Pseudomonas* sp. NC02 and its parent genus *Pseudomonas* (Figure 5.10e,f).

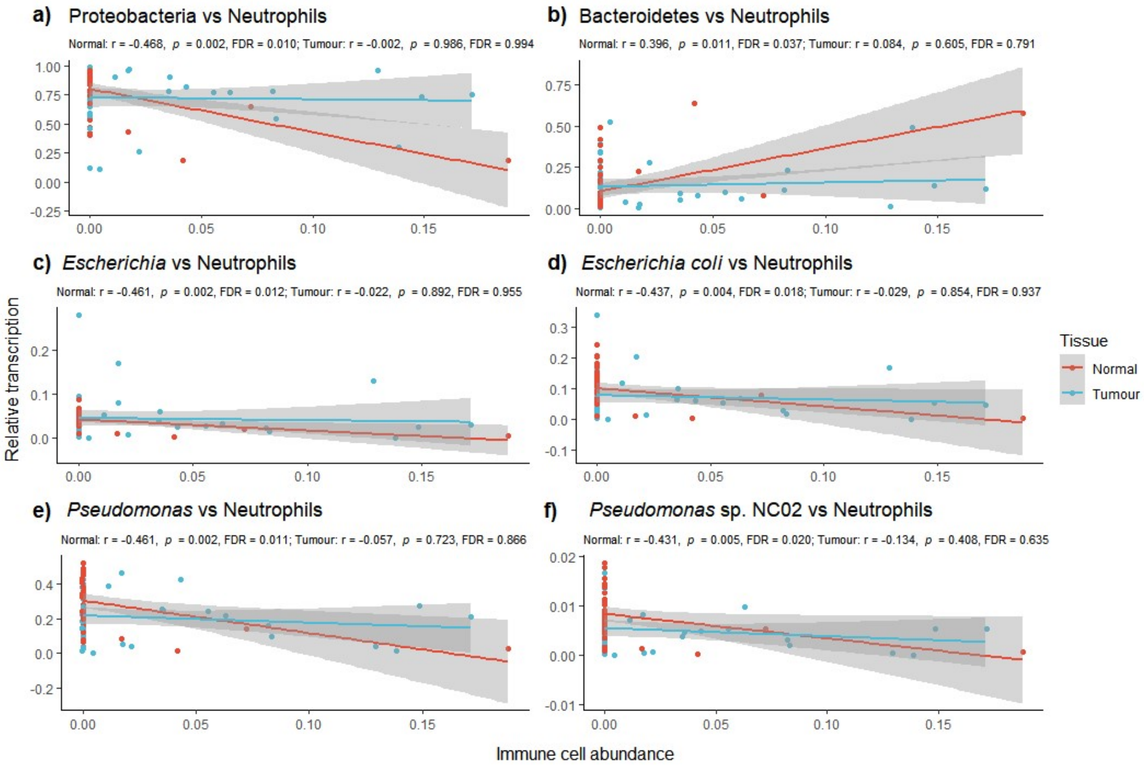

Figure 5.10. Phyla, genera and species transcription correlated with neutrophils. r: Spearman's Rho; FDR: false discovery rate adjusted p-value. Lines of regression are coloured by tissue type; grey bars indicate 95% confidence intervals.

Additionally, *Hungatella hathewayi* transcription and that of its parent genus *Hungatella* were positively correlated with neutrophil levels in normal tissues (Figure 5.11b,a), along with *Butyricimonas* faecalis and its parent genus *Butyricimonas* (Figure 5.11d,c).

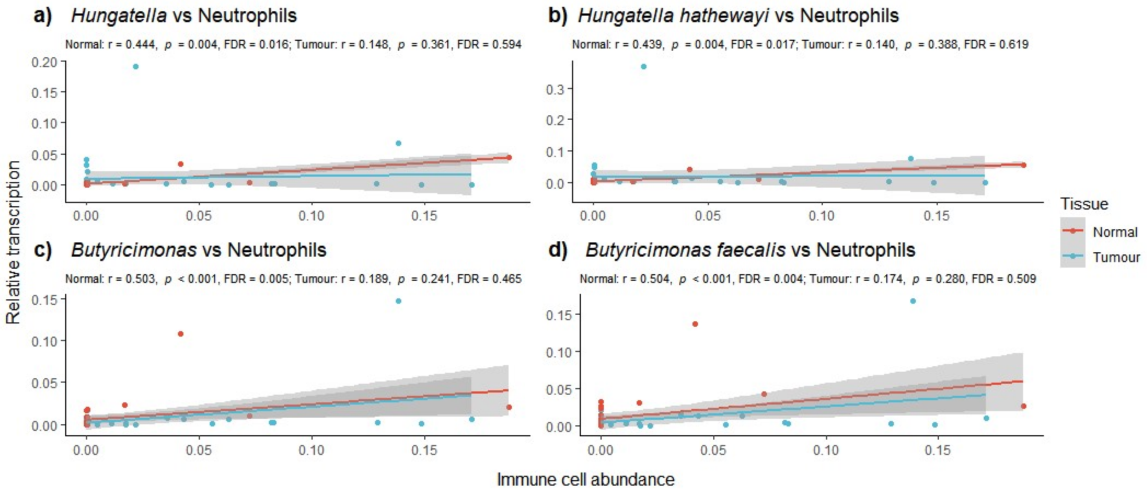

Figure 5.11. Genera and species correlated with neutrophils. r: Spearman's Rho; FDR: false discovery rate adjusted p-value. Lines of regression are coloured by tissue type; grey bars indicate 95% confidence intervals.

Transcription of the genus *Alistipes* and the species' *Flavonifractor plautii*, *B. fragilis and Odoribacter splanchnicus* were positively correlated with neutrophil abundance in normal tissues (Figure 5.12a,c,e,f). *B. fragilis* was also negatively correlated with neutrophil abundance in tumour tissues, but this was not statistically significant after FDR correction (Figure 5.12e). *Clostridium saccharobutylicum* and *Salmonella enterica* transcription were negatively correlated with neutrophil abundance in normal tissues (Figure 5.12b,d)

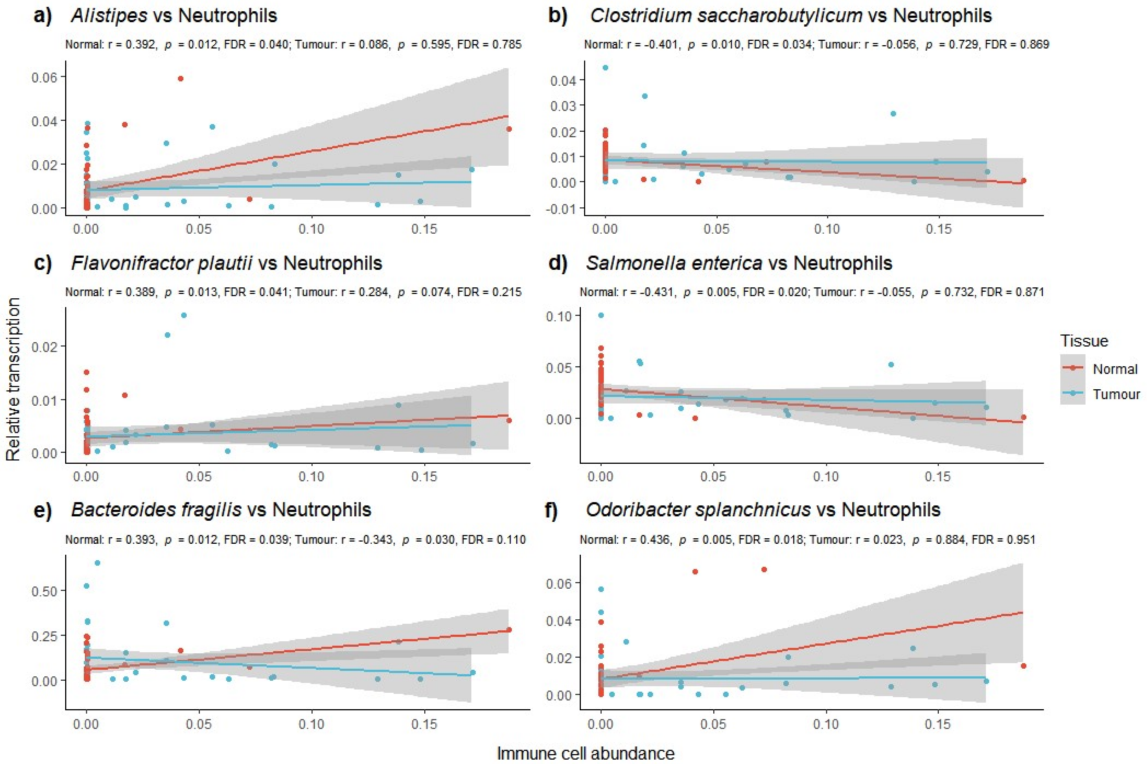

Figure 5.12. Additional taxa transcription correlated with neutrophil abundance. r: Spearman's Rho; FDR: false discovery rate adjusted p-value. Lines of regression are coloured by tissue type; grey bars indicate 95% confidence intervals.

M1 macrophage abundance was significantly different between Dworak Two and Four tumour and normal tissues, and the tumour tissues of Dworak Four patients and those of Dworak Three–One patients, with the highest levels being in Dworak Four tumour tissues (Figure 5.13a). Resting mast cell abundance was significantly different between tumour and normal tissues of Dworak One and Two patients, as well as between the tumour tissues of Dworak Four patients and those of Dworak One and Two patients (Figure 5.13b). Additionally, the tumour tissues of Dworak Four patients were the only tumour samples to have a consistent abundance of resting mast cells (Figure 5.13b).

The abundance of activated mast cells was higher across samples than resting mast cells, with statistically significant differences between the normal and tumour tissues of Dworak One and Two tumour patients, and between Dworak Four, Three, Two and One tumour tissues (Figure 5.13c). Eosinophils were most abundant in Dworak Two, and Three tumour and normal tissues, respectively, and significant differences were found between the normal tissues of Dwoark Four and Three patients, between Dworak Three tumour and normal tissues, and between Dworak Three and One normal tissues (Figure 5.13h). Neutrophil abundance was significantly different between Dworak Two tumour and normal tissues and was in low abundance in normal tissues generally, and Dworak Four tumour tissues (Figure 5.13j). The abundance of resting dendritic cells was only evident in four samples, one normal sample, and three tumour samples, all from Dworak Two patients (Figure 5.13d). In contrast, activated dendritic cells were found in abundance across all Dworak groups and tissues (Figure 5.13e); additionally, and along with the abundance of naïve B cells (Figure 5.13f) and follicular T helper cells (Figure 5.13i), were not significantly different between the tissues or Dworak groups.

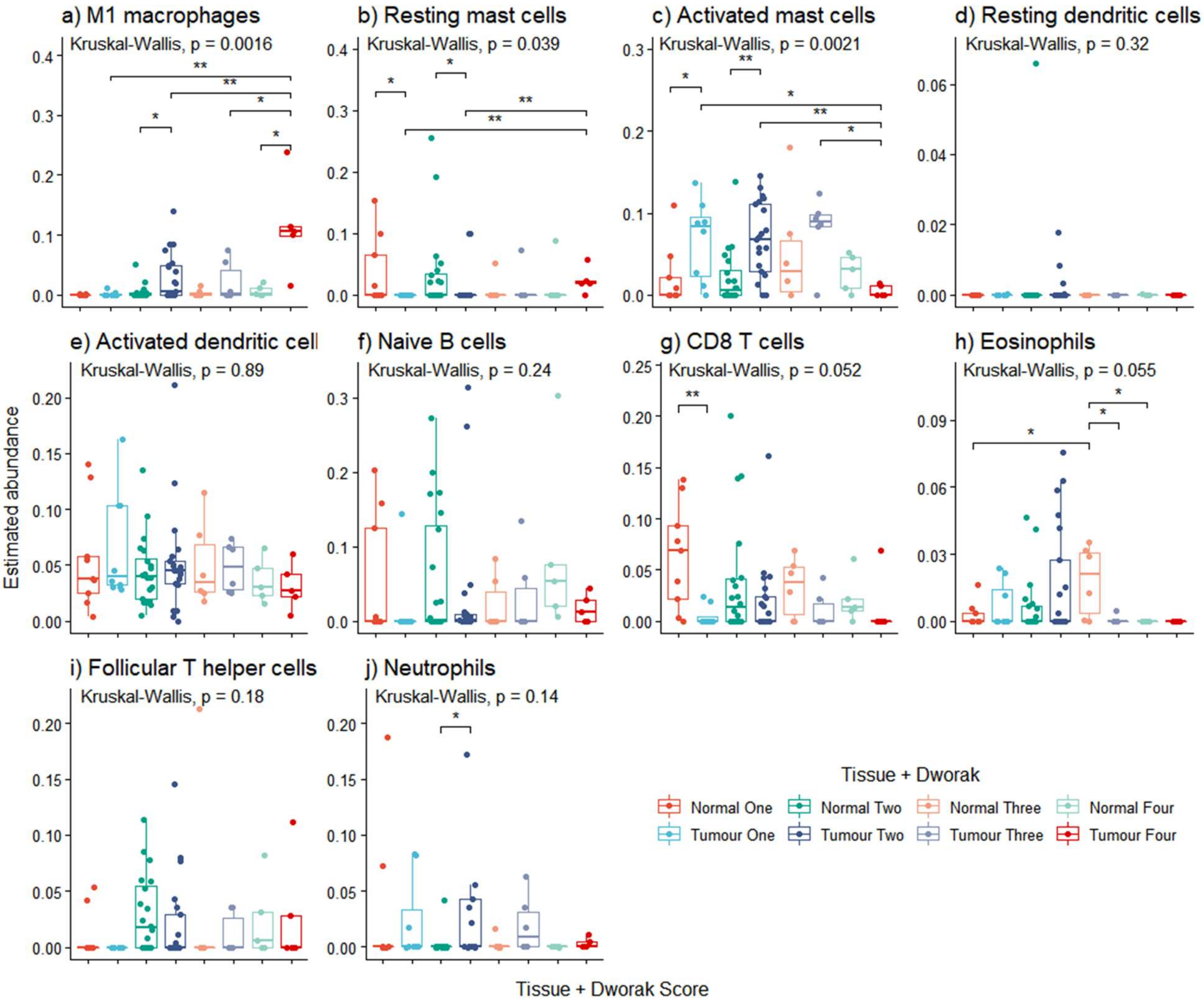

*Figure 5.13. Immune cell levels correlated with bacterial transcription. a) M1 macrophages; b) resting mast cells; c) activated mast cells; d) resting dendritic cells; e) activated dendritic cells; f) naïve B cells; g) CD8 T cells; h) eosinophils; i) follicular helper T cells; j) neutrophils. * indicate significance of FDR adjusted Wilcoxon signed-rank tests; *: p < 0.05; **: p < 0.01. Not statistically significant comparisons not shown.*

5.4 Discussion

5.4.1 Myeloid lineage most correlated with radiotherapy response

In this study, the common myeloid progenitor lineage was of most consequence for response correlations, with mast cells and M1 macrophages being correlated with radiotherapy response in tumour tissues; indicating that interference with myeloid differentiation is the major contributor to the production of effector cells contributing to tumour regression. M1 macrophage correlations with response were likely due to the

34-fold higher levels in complete responders compared to other groups. Resting mast cells were also positively correlated with response, indicating that the mitigation of an allergic response may be prognostically relevant. Macrophages are well understood to have a role in radiotherapy and radiation-induced cellular injury where they play a role in cellular clearance and fibrotic wound responses; a characteristic of tumour regression [450].

5.4.2 M1 macrophages associations with bacterial transcription

Although not statistically significant after FDR correction, the negative correlations of Campylobacteraceae and *C. ureolyticus*, and positive correlations of Cyanobacteria, *B. vulgatus, Desulfovibrio*, *Alistipes* and *A. finegoldii* transcription with M1 macrophages should be further investigated. It is known that Campylobacter can infect macrophages, and can direct M1 macrophages to target Campylobacter cells [454], which may interfere with their anti-tumour role. Cyanobacteria toxins are known to have immunomodulatory and carcinogenic effects [409] [455]; however, the direct interaction with M1 macrophages is unknown. [427]. *B. vulgatus* are known to modulate the immune system via lipopolysaccharide production [427], which may decrease M1 macrophage levels by reducing LPS induced cytokines. *Desulfovibrio* species have been shown to infect macrophages and alter their gene expression [456], and their LPS has been shown to stimulate TNF-a secretion more so than other bacterial LPS [457]. *Alistipes* are bile tolerant [458] and in obesity, studies have been shown to alter polarised macrophage ratios [459]; additionally, they are produced of short-chain fatty acids (SCFAs) and can reduce glucose uptake, and incorporate saturated fatty acids into their membranes [419, 421]. Cholesterol uptake by macrophages is the principal cause of atherosclerosis, maintaining a pro-inflammatory response [460]. The most compelling evidence for M1 macrophage CRC microbiome interactions comes from studies into oral bacteria causing differential polarisation of macrophages, particularly in regards to

Porphyromonas gingivalis [461, 462] [463]; however, this was not detected in this study, likely due to the limitations of sample sizes and sequencing depth. Within normal tissues, Pasturellaceae were negatively correlated with M1 macrophages, which may be due to toxin production causing potential macrophage differentiation to an osteoclastic lineage [464].

Taking the above into account, it may be that the microbiota in high responders act in a lipid-rich environment and limit tumour glucose utilisation. By preventing monocytes from being 'distracted' by other microenvironment factors such as fat and resident microbiota, they can differentiate and polarise to anti-tumour M1 macrophages.

5.4.3 Mast cells and bacterial transcription

Mast cells have a role in allergen response, pathogen recognition, and wound healing [465], and have been shown to dysregulate T cell regulation in CRC [466]. Due to the varied roles of mast cells, their abundance alone does not indicate any particular pathway for their interaction with tumour regression, which may differ between patients. However, their pathogen recognition role may be essential for their involvement in tumour regression; indeed, *B. dorei* transcription was negatively and positively correlated with resting and activated mast cells, respectively. Studies have demonstrated that gut commensals and *Bacteroides* species specifically can suppress mast cell degranulation and reduce reactivity [467, 468]; however, these studies did not focus on *B. dorei*. *B. dorei* may have a role in immune modulation more directly, as has been shown by a study which found high levels of the species preceding autoimmunity onset [469]. Unfortunately, the species *B. vulgatus*, *B. dorei* and *B. xylanisolvens* are very closely related and can be challenging to differentiate from one another [470]. Additionally, the negative correlation of Bifidobacteriaceae with resting mast cells conflicts with the literature, with *Bifidobacteria* species being shown to have suppressive effects on mast cells [471]. *B. longum* transcription was two-fold higher in tumours compared to normal tissues in high responders, which may indicate that its role is microenvironment dependant, as otherwise, it would have a similar correlation in tumour tissues, where it was found to be more transcriptionally active. Therefore, the potential interaction of *B. dorei* and Bifidobacteraceae species on mast cells in the tumour microenvironment should be further investigated.

Small sample sizes likely influenced the correlation of resting mast cells with radiotherapy response, although they were detected mainly within the tumour tissues of complete responders (Dworak Four). However, the low abundance of resting mast cells indicates that a minority of samples may have driven possible correlations.

5.4.4 Bacterial transcription and immune cells not correlated with response

Several other immune cells were correlated with bacterial transcription that was previously implicated in response in Chapter Four. The correlation of *B. fragilis* transcription with both naïve B cells and CD8 T cell abundances was of particular note, as their interaction with these cells, and the immune system more generally has been well studied [472], and tumour regression has been suggested to be dependent on CD8 T cells [473]. *B.fragilis* capsular polysaccharide A (PSA) is the most immunomodulatory aspect of this species, as when binding to B cells, it can result in CD8 T cell anti-inflammatory IL-10 secretion [474]. The maintenance of a population of naïve B cells, is of particular note, as it could indicate that once radiation therapy commences, they can uptake and present tumour antigens. In contrast, previously activated B cells would present antigens from the tumour microenvironment. Indeed, studies suggest that infiltrating B cells are a possible prognostic indicator for radiation therapy [475], which be reduced during chemotherapy [476]. The above may indicate that *B. fragilis* transcription plays a role in maintaining the anti-tumoral potential of the immune system before radiation therapy, and prevents non-tumour antigen presentation, which would not increase until radiation-induced tumour cell death, and thus, increased tumour neo-antigen availability [477].

Eubacterium rectale is a well-studied and abundant bacteria in the gut; however, maintains many subspecies which are more prevalent in different parts of the world, with carbohydrate metabolism, exopolysaccharide and motility operons contributing most to their diversity [478]. The correlation of *E. rectale* transcription with eosinophils is interesting, as it was found at much higher levels in tumour tissues of high responders, eosinophilia is a rare but known complication of radiation therapy [479], and eosinophil levels have been suggested as prognostic markers for several cancers [480-482]; being involved in radiation-induced fibrosis in intestinal tissues [483]. NK cells are known to play a major role in anti-tumour activity in radiation therapy [484], and *E. rectale* has been associated with a decrease in NK activity in the past [485]. Taken together, it may indicate that *E. rectale* may act before radiotherapy to preserve anti-tumour immunogenic potential to be activated during therapy generation of tumour antigens via tumour ablation [486].

Follicular T helper cells are required for the generation of high-affinity antibodies and memory B cell formation, which are critical anti-tumour immunity, and mitigation of relapse [486, 487]. Flavobacteriaceae and *S. pyogenes* transcription was correlated with follicular T helper (Fth) cells, in normal and tumour tissues, respectively. Fth cells have been associated with reduced tumour growth and higher survival rates in some cancers [488]. Flavobacteriaceae is the parent family of Flavonifractor, which had more than 2.5-fold higher transcription in the tumour tissues of high responders compared to low responders. As was suggested in the discussion of Chapter Four, Flavonifractor may confer a benefit to tumour regression rates in an immune-modulatory way, particularly by suppressing Th2 pro-inflammatory responses [418]. Fths play a role in lowering immune reactions to commensal microbiota in the gut by modulating IgA secretion [489] and Fth differentiation from the common Th2, and Fth progenitor cell is a STAT3 dependent process [490, 491], which has been suggested as a therapeutic target in cancer therapies as it a known promoter of cellular proliferation [492, 493]. Taking these pieces of evidence into account, it may be that the correlation of Flavobacteriaceae with Fth, and its child genus' association with high responder tumour tissue, may indicate that they are associated with STAT3 expression, which has been associated with both tumour progression inhibition [494] and accelerated tumour growth [495], in a microbiota dependent manner.

5.4.4.1 Bacterial transcription associations with dendritic cells

Activated dendritic cells (aDCs) in tumour tissues were negatively correlated with Dworak scores, but this was not statistically significant. Additionally, the transcription of *E. coli* and *B. vulgatus* were correlated with poor response, while the former was positively correlated with aDCs, the latter was negatively correlated with resting dendritic cells (rDCs). The transcription of these species could contribute to a poor response by generating antigen presentation against them, rather than the tumour. Indeed, studies on *E. coli* antigen presentation by DCs shows that it occurs readily, to the extent that has been suggested that *E. coli* antigen uptake by DCs could be used in anti-tumour immunotherapy [496, 497]. However, a recent study showed that *E. coli* but not *B. vulgatus* was able to attenuate intestinal inflammatory in a colitis model [498]. Therefore, these taxa may interact differently with the immune system, and this is likely to be strain-dependent. The transcription of *C. saccharobutylicum* and *S. enterica* were also implicated in response, and they were both correlated with aDCs. *C. saccharobutylicum* is used in the generation of biobutanol in the biofuel industry;

therefore its species assignment may be speculative. Furthermore, the correlation of DCs and response was not statistically significant, and the precise role of DCs in radiotherapy response is not well understood [499]; however, they play a critical role in tumoural-immunity via antigen presentation. Therefore, further research is needed into the role of DCs and their relationship with tumour infiltrating bacteria and radiotherapy response; as similar to other immune cells, the preservation of rDCs until after radiotherapy may be critical for the uptake of tumoural, rather than bacterial or microenvironmental antigens [477].

5.4.4.2 Neutrophils have the most microbial associations

Neutrophils are a known contributor to radiation therapy responses, contributing to radioresistance in some cancers [500], and their abundance has been suggested as a biomarker for radiotherapy outcomes, depending on their phenotype (anti-tumour N1, pro-tumour N2) [501]. The different phenotypes of neutrophils, similar to the polarisation of macrophages, is likely to be dependant on the tumour microenvironment [502]. However, CIBERSORT does not differentiate between neutrophil phenotypes, so it remains unknown if N1 or N2 was more abundant in the study population. It may be that no statistically significant correlation was found as neutrophils generally are not correlated with response, but N1 or N2 phenotypes may.

Neutrophil abundance was highest in tumour tissues, and Dworak Four tumours had the lowest levels, which may be due to the low sample size (Dworak Four n = 5*)*; it also could indicate that neutrophil levels are prognostically relevant for a complete response. Despite this, no bacterial transcription was correlated with neutrophil abundance in tumour tissues, with all significant correlations being found in normal tissues. As the normal tissue biopsies were collected adjacent to tumour tissue, the abundance of neutrophils in these tissues could provide a reservoir from which to migrate to the tumour microenvironment during radiotherapy as part of the early wound repair response [503, 504].

The transcription of the Proteobacteria phylum generally and the contained species *E. coli*, *Pseudomonas* sp. NC02, *S. enterica*, as well as *C. saccharobutylicum* in the Firmicutes phylum, were negatively correlated with neutrophil levels in normal tissues. Low neutrophil levels may be due to strains of *E. coli* being able to cause apoptosis and necrosis in neutrophils [505, 506]. Neutrophils also react in different ways to *Pseudomonas* species, with most studies on cystic fibrosis patients showing decreased clearance abilities of neutrophils against *Pseudomonas* [507], and dying and dead neutrophils have been shown to facilitate *Pseudomonas* biofilm production [508]. *Salmonella* species have been shown to use flagellar motility to agonise neutrophil ROS production [509], while some *Clostridium* toxins have been shown to inhibit neutrophil proliferation [510]. The above evidence may indicate that the taxa can neutralise neutrophils via killing and disrupting their proliferation and bacterial clearance actions, triggering biofilm formation that shields the taxa from further neutrophil attention. However, as much of this evidence is not specific to the species identified here, further investigation is required.

The transcription of the Bacteroidetes phyla and the contained species *Odoribacter splanchnicus*, *B. fragilis*, *Butyricimonas faecalis* and *Alistipes*, as well as the Firmicutes *Hungatella hathewayi* and *Flavonifractor plautii*, were correlated with neutrophil abundance in normal tissues. Consistent with the literature, *Alistipes* are known to induce and thrive in inflamed environments [381, 511]. At the same time, *Hungatella* has been found in greater abundance in CRC tissues of the immunogenic subtype [93], and *H. hathwayi* specifically is immunoreactive [412]; however, their interaction with neutrophils specifically has not been investigated. Additionally, the oral administration of *F. plautii* has been demonstrated to lower the Th2 immune response [418], in which neutrophils have an immunomodulatory role [512] and is thought to be required for some neutrophil associated inflammation [513]. *O. splanchnicus* has been shown to reduce neutrophil attraction into mucosa via IL-8 inhibition, while *B. fragilis* PSA has similar anti-inflammatory properties [514]. *B. faecalis* is a butryate producer, which is known to stimulate colonic Tregs, which, in turn, are known to limit neutrophil responses [515]. Taken together, it is likely that in comparison to the negatively correlated taxa in normal tissues, taxa which were positively correlated with neutrophil levels can be explained by immuno-tolerance activities. Therefore they likely do not stimulate neutrophil abundance but also do not inhibit their proliferation or migration.

Overall, the immune system can interact with the complex tumour microenvironment in different ways. An allergic pro-inflammatory response can be triggered, causing myeloid cells to differentiate via the mast cell lineage. Alternatively, myeloblasts can be differentiated to neutrophils, eosinophils or the monocyte/macrophage lineage. The monocyte/macrophage lineage can result in anti-microbial effector cells via LPS stimulated cytokines, an anti-lipid response similar to that seen in atherosclerosis, or an anti-tumour response.

Based on the correlations with response and with bacterial transcription, the immune anti-tumour response may only be achieved when macrophages and other immune cells are free to exhibit anti-tumour activity and are not distracted by aspects of the tumour microenvironment. Non-tumour immune activity may occur via direct infection by microbes and their antigens; generating an antimicrobial response, or effector cells may react to other aspects of the tumour microenvironment (i.e., lipids, cellular debris), in the absence of potent immunogenic targets and tumour antigens. Furthermore, the suppression of antigen presenting cells (APCs) before radiotherapy may be critical, preserving naïve and resting APC populations to uptake and display tumour neo-antigens as tumour cells die, rather than displaying antigens of the tumour microenvironment.

5.4.5 Study limitations

It is possible that CIBERSORT's digital cell cytometry for evaluating immune cell abundance may not be completely accurate [516]; therefore, real-world cytometry should be used in future to confirm these associations, particularly in regards to the phenotypic subtypes of immune cells (i.e., neutrophils). The immune cell abundance (as seen in Figure 5.13) showed that immune cells were not evenly represented across samples, indicating that the identified assocaitions may be due to outlier samples, which was particularly evident for less abundant immune cells, and in Dworak Two patients. Additionally, the small sample sizes, particularly in regards to the limited numbers of Dworak One, Four and Three response groups, reduces the power of this study to determine associations of immune cells with radiotherapy response properly. Furthermore, Kraken2 may have misclassified some of the bacterial taxa, particularly at the species level, and the filtering process may have removed consequential taxa which may have been retained with larger sample sizes. Additionally, species-level designations do not fully elucidate the capabilities of the taxa identified, and strain and subspecies level designations are required for a more accurate analysis. Finally, some species correlations, particularly *B. dorei* with mast cells, *B. frag* with naïve B and CD8 T cells, Flavobacteriaceae with follicular T helper cells, and correlations with neutrophils need to be investigated further, as correlations may have been due to the low number of samples with an abundance of these immune cells.

5.5 Conclusions

Overall, the abundance of resting mast cells and M1 macrophage levels in tumour tissues were the most robust indicator for chemoradiotherapy response. The results indicated that interference with myeloid differentiation might be the major contributor to the production of effector cells contributing to tumour regression. The positive role of microbial transcription in immune modulation before radiotherapy likely involves suppressing the APC and effector cell populations before radiation ablation occurs, allowing them to be available for rapid uptake of and action toward tumour antigens. Additionally, beneficial microbes prevent the immune system from establishing a non-tumour response in reaction to elements of the microenvironment (i.e., allergic, anti-fat, anti-microbial), which may persist during or inhibit anti-tumour responses once radiotherapy commences.

The association and specific mechanisms and roles of microbial species on immune-mediated response to radiotherapy require further investigation with larger samples sizes, and both in vitro and in vivo laboratory experiments.

6 CHAPTER 6: BIOMARKER ANALYSIS

6.1 Introduction

Prognostic markers for CRC radiotherapy treatment have not been well elucidated, with 10% to 25% of cases resulting in complete response. Few biomarkers for radiotherapy response have been discovered and may not be universally applicable [517-523]. The microbiome has been investigated in terms of response to radiotherapy, with most studies focusing on the role of the microbiome in radiotherapy side effects such as mucositis, radiation-induced diarrhoea and other gastrointestinal symptoms [393, 394, 396, 399], and few focusing on the prognostic value of the microbiome [524].

The role of the immune system in terms of differential responses to therapy in RC has been increasingly implicated, particularly in regards to antigen-presentation inhibition and depletion of tumour infiltrating leukocytes, as well as the abscopal effect of distant metastatic tumour regression post-radiotherapy [250, 439-443].

In this chapter, the aim was to investigate possible biomarkers based on pre-therapy biopsies, from tumour and adjacent normal tissues via block supervised sparse partial least squares discriminant analysis (sPLS-DA). The microbiome, gene expression, and estimated immune cell infiltrates were the blocks used as putative markers for response to therapy, in an attempt to classify patients into response groups before the commencement of therapy. By finding biomarkers indicative of response to therapy, it would be possible to direct patients to alternative treatments or accelerate them to surgery; thus, increasing disease-free survival in RC patients, decreasing unnecessary side effects and reducing treatment costs.

6.2 Methods

Microbial species relative abundance and relative transcriptional abundance was determined as previously described (Section 2.6). Immune cell abundance was estimated as described in Section 5.2. The level of gene expression was quantified as described in Section 2.6 using SALMON v1.2.1 [287], using the quasi-alignment method, which quantifies gene expression based on the reference transcriptome (GRCH38p12) and imported into R using the tximport package v1.16.1. The three datasets were input to DIABLO (**D**ata **I**ntegration **A**nalysis for **B**iomarker discovery using **L**atent c**O**mponents) [359], as part of the mixOmics data integration project R package v 6.12.2 [525].

The microbiome species data from the RNA-Seq experiments were utilised, as well as the immune dataset. The gene expression dataset was adjusted before analysis using the filterByExpr function in edgeR [356], removing any genes with a count of less than 100 in all samples in each tissue group (*n* = 40). Gene expression filtering left 13,613 genes out of the original 36,622 for normal and tissue samples. Additionally, it was found that pre-sequencing ribodepletion was not complete, with substantive levels of RNA subunit genes, particularly RNA28S and 45S, remaining, which were removed. The final gene expression dataset contained 13,587 genes.

Tumour and normal tissue data were normalised separately via centred log-ratio transform (CLR), to avoid overcorrection and loss of treatment effects. The CLR transform maps a composition in Aitchison-simplex to euclidean vector subspace, allowing consequent matrices to be singular, and allow the use of classical multivariate dataset analysis [526]. Any potential batch/cohort effect was removed using the removeBatchEffect function in limma [527].

Sparse partial least squares regression (sPLS) was used for variable selection [528], which includes Least Absolute Shrinkage and Selection Operator (LASSO) penalisations on pairs of loading vectors, allowing minor coefficients to be forced to zero values and permitting varied coefficient sizes in the same model [528]. Raw inputs were used to establish the categorisation ability using variables which in the immune and bacterial transcription data met an R^2 threshold of ≥ 0.2 or ≤–0.2 and within the gene expression dataset of ≥ 0.01 or ≤–0.01.

The data was split into training and test sets (1), and the training set was reduced manually, variable by variable, to the point where further removal reduced the predictive ability of the variable sets, as assessed by 5-fold leave-one-out (loo) cross-validation centroid distribution in the mixOmics R package (v6.12.2) [525]. Weighted vote error rates were used to establish response group classification overall error rates and overall balanced error rates; the latter being more accurate due to the uneven number of individuals in each response group.

Table 6.1. Training and test datasets

Response	Dworak	Train	Test
Complete	Four	3	2
Intermediate	Three	4	2
	Two	17	5
Poor	One	4	2

The goal was to iteratively remove variables to find the most discriminatory set, using the lowest number of inputs, and without disrupting the ability to predict response groups. Once a reasonable error rate had been reached, the test set response grade was established based on the training set output.

It was determined that Dworak scoring may not be a true continuous variable due to the nature of its allocation (different pathologists may assign different scores based on different thresholds) and that there may be differences between a potential poor responder with an unusually high response, and a potential complete responder with an unusually low response. Therefore, it was decided to change the model into one with a more simple logical model of three groups, complete (Dworak Four), intermediate (Dworak Two and Three) and poor (Dworak one), allowing for simpler categorisation. Heatmaps were generating using the cimDIABLO command, a method similar to classical hierarchical clustering.

Gene roles were taken from data acquired from the Human Gene Database: https://www.genecards.org/, and Gene Ontology information was taken from https://biit.cs.ut.ee/gprofiler/. Supplementary tables and R code used can be found at https://github.com/William-S-Taylor/MSc.

6.3 Results

6.3.1 Batch/cohort effect testing and removal

Post-CLR-normalisation of datasets, the cohort effect was visualised using sPLS-DA (Figure 6.1). There was found to be a cohort effect, which was most evident in gene transcription data.

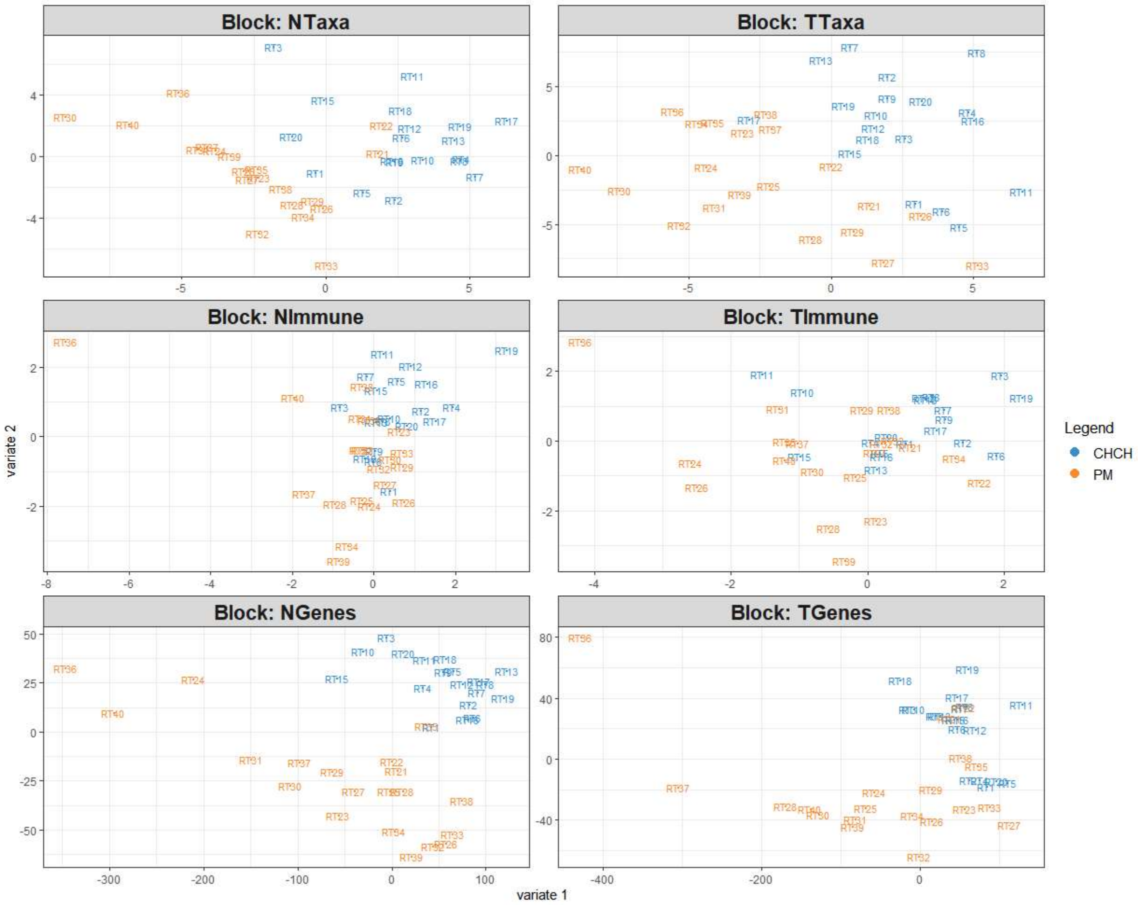

Figure 6.1. sPLS-DA plot of the cohort effect on data. N: Normal tissue; T: Tumour tissue; CHCH: Christchurch cohort; PM: Peter MacCallum cohort; RT: RC patient.

Batch correction effectively reduced inter-cohort variability, as shown in the plot in Figure 6.2.

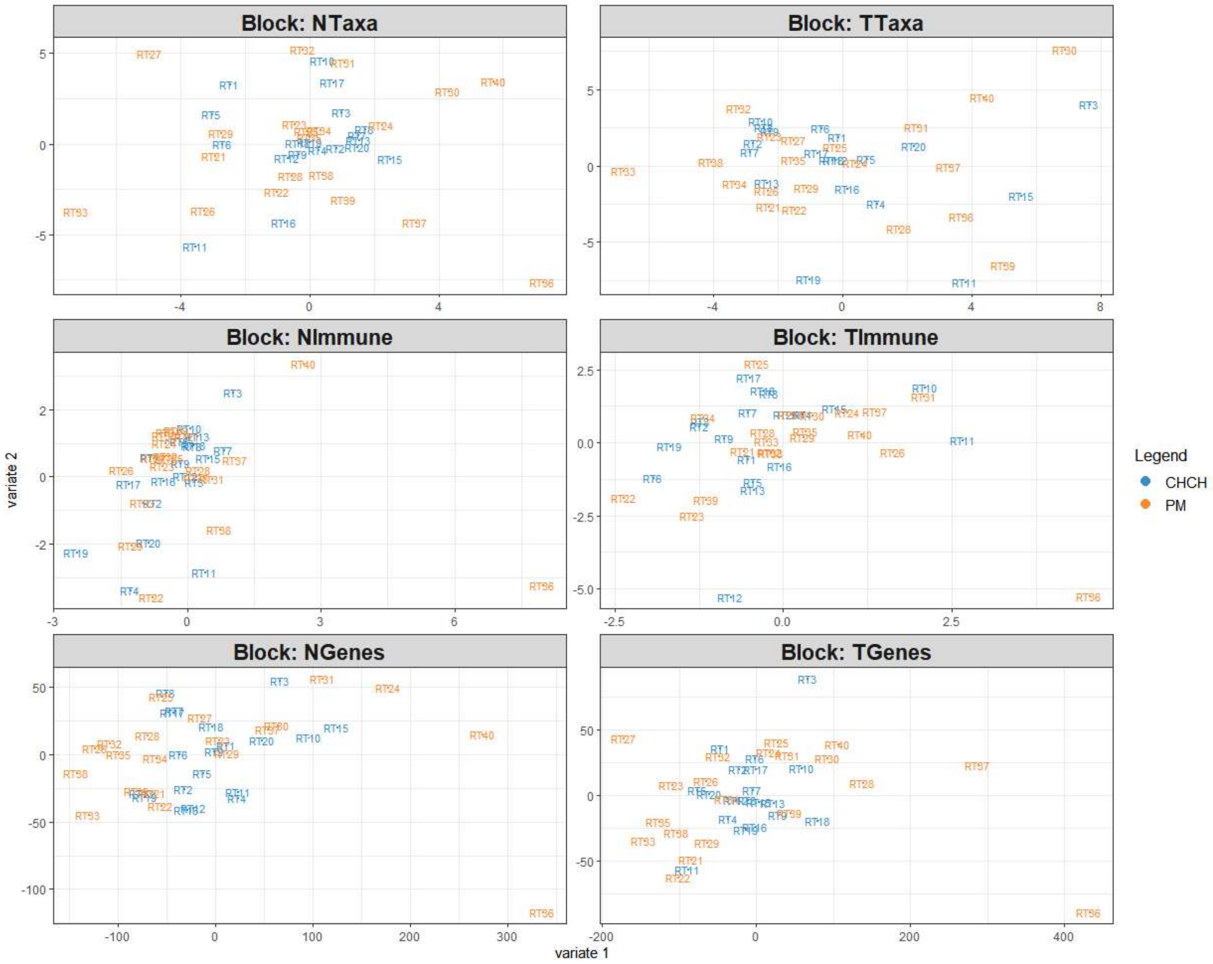

Figure 6.2. sPLS-DA plot of batch effect adjusted data.

6.3.2 Initial categorisation

Once the data had been normalised and adjusted for batch effects, the datasets were used for discriminant analysis, using the expression of 13,587 genes, 22 immune cell types and 160 bacterial taxa in tumour and normal adjacent tissues (Figure 6.3).

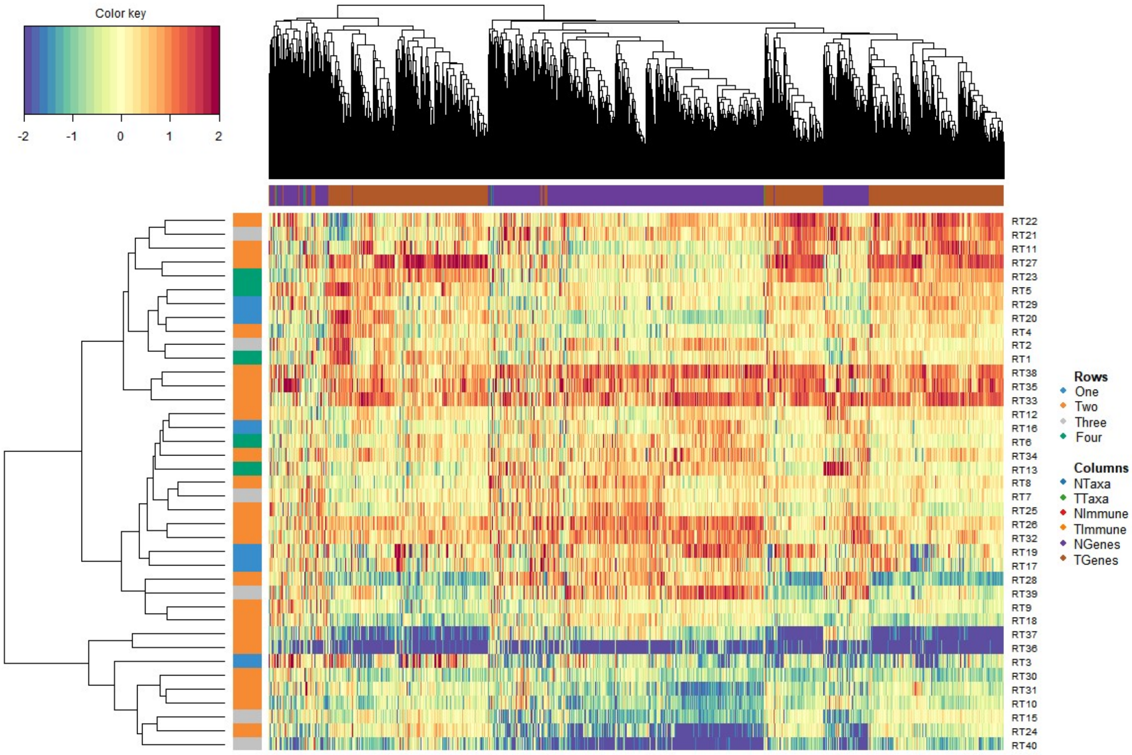

Figure 6.3. Heatmap of clustering by Dworak scores using unadjusted data.

The lack of clustering also impacted the ability to predict response based on the input variables. The lowest error rates were found in predicting Dworak Four patients (0.4, Table 6.2). Overall error and balanced error rates were high, at 0.74–0.79 and 0.64–0.74, respectively, depending on the component used.

Table 6.2. Classification error rates for Dworak scores using unadjusted data.

Dworak	comp1	comp2	comp3	comp4	comp5	comp6
One	1.00	0.67	0.67	0.67	0.83	0.67
Two	0.73	0.86	0.82	0.91	0.82	0.77
Three	0.83	0.83	0.83	0.83	0.83	0.83
Four	0.40	0.20	0.40	0.40	0.40	0.60
Overall.ER	0.74	0.74	0.74	0.79	0.77	0.74

Overall.BER	0.74	0.64	0.68	0.70	0.72	0.72

ER: Error rate; BER: Balanced error rate.

To maximise the differences between response groups, patients were then grouped into response grades, complete (Dworak Four), intermediate (Dworak Two and Three) and poor, (Dworak One). There was a similar lack of clustering using the input variables using response grades (Figure 6.4), as was seen when using Dworak scores. However, this allowed for more straightforward variable selection.

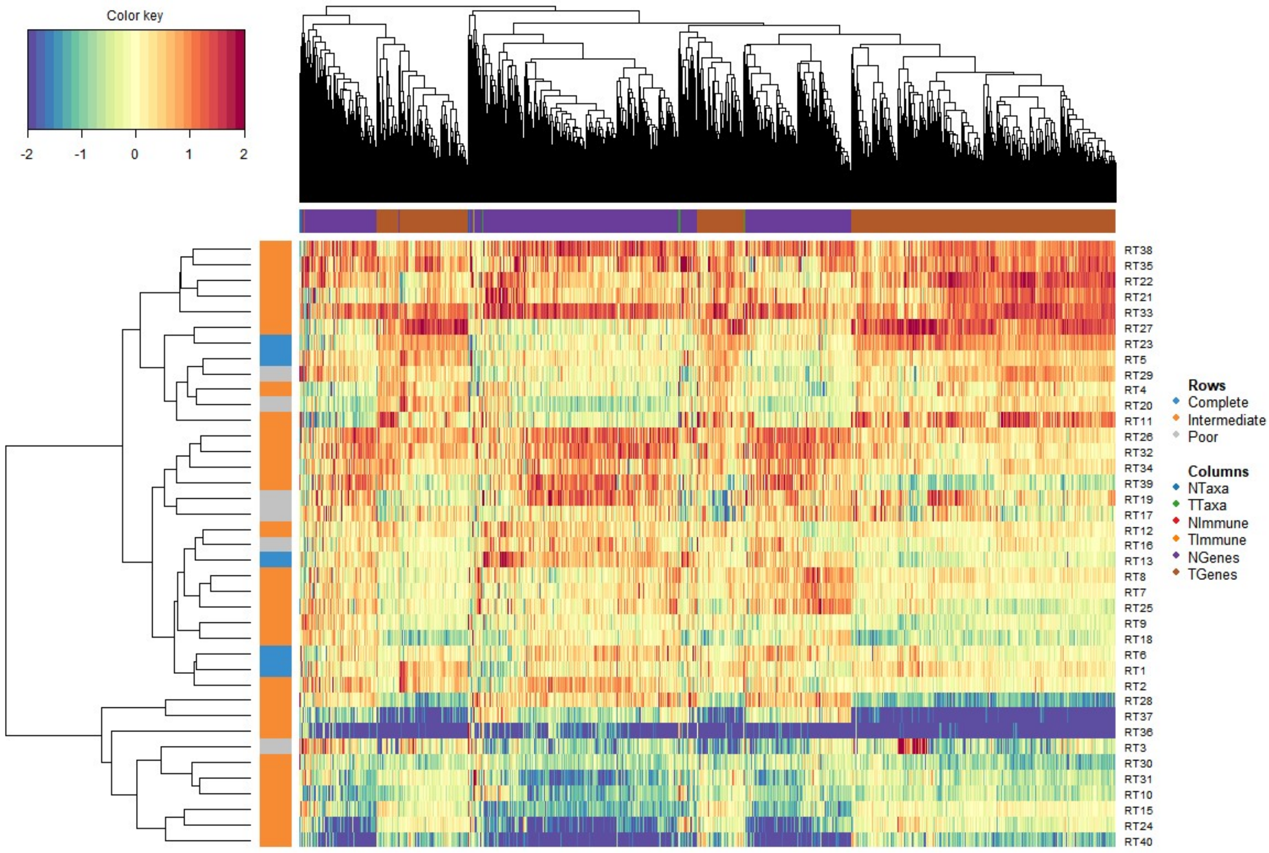

Figure 6.4. Heat map of clustering by response grade using unadjusted data.

Characterisation of patients by response grade reduced the error rates substantively, as distinguishing between three categories is a lower resolution process. The overall and balanced error rates decreased by 0.1–0.25 and 0.02–0.22, respectively (Table 6.3).

Table 6.3. Classification error rates for response grades using unadjusted data.

Response	comp1	comp2	comp3	comp4	comp5	comp6
Complete	0.60	0.40	0.20	0.40	0.60	0.80
Intermediate	0.57	0.54	0.50	0.54	0.46	0.43
Poor	1.00	0.83	0.67	0.67	0.83	0.83
Overall.ER	0.64	0.56	0.49	0.54	0.54	0.54
Overall.BER	0.72	0.59	0.46	0.53	0.63	0.69

ER: Error rate; BER: Balanced error rate.

6.3.3 Highly correlated variables

Sparse partial least squares (sPLS) regression was utilised to identify meaningful variables for categorising samples. Each dataset (x) was computed against the CIP response grades (y), and variables were retained if the immune and bacterial transcription data R^2 was ≥ 0.025 or ≤–0.025, and for gene expression, ≥ 0.01 or ≤–0.01. The approach yielded 38 and 42 taxa, 13 and 14 immune cells, and 152 and 101 genes for tumour and normal tissues, respectively. Compared to the unadjusted dataset, the correlated data could characterise patients more effectively (Figure 6.5).

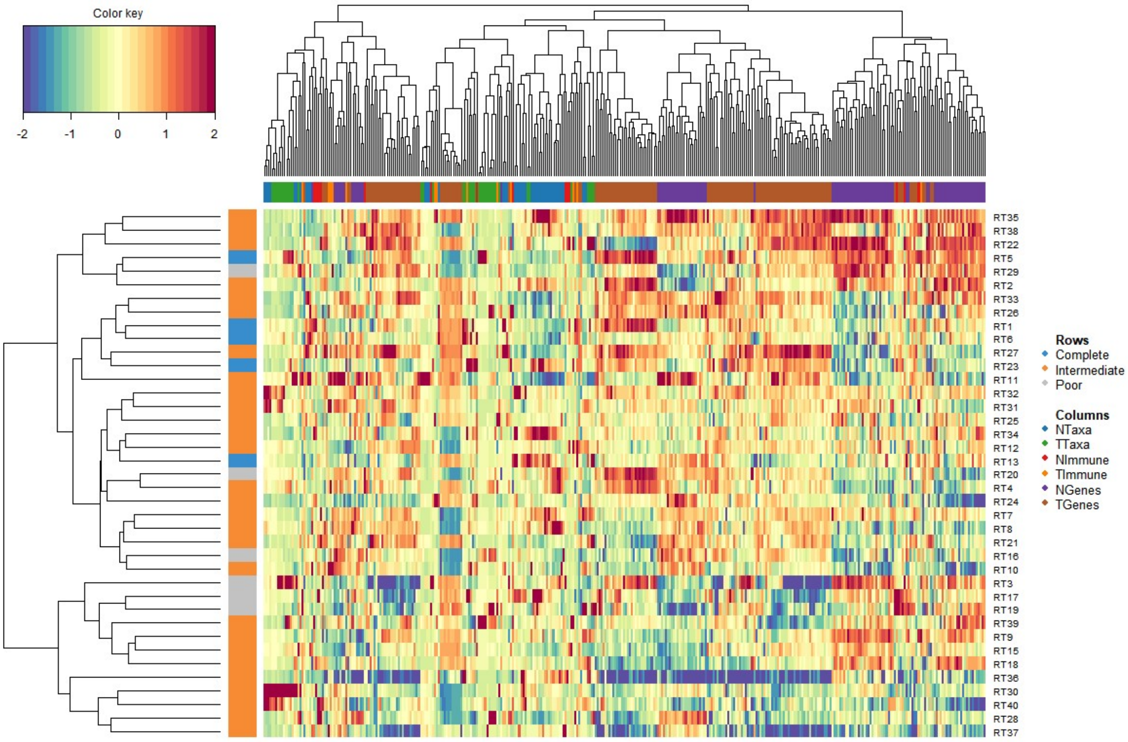

Figure 6.5. Heat map of clustering of regression grades post-variable regression.

The reduced datasets provided improved error rates as low as 0.46 and balanced error rates as low as 0.45, depending on the component used (Table 6.4).

Table 6.4. Error rates post variable regression

Response	comp1	comp2	comp3	comp4	comp5	comp6
Complete	0.40	1.00	0.80	0.80	0.60	0.80
Intermediate	0.46	0.43	0.50	0.39	0.43	0.43
Poor	0.50	0.50	0.50	0.50	0.50	0.50
Overall.ER	0.46	0.51	0.54	0.46	0.46	0.49
Overall.BER	0.45	0.64	0.60	0.56	0.51	0.58

At this point, the dataset was split into training and test sets. Within the training set, the number of variables was iteratively reduced by one and its impact on clustering and error rate evaluated. If removal of the variable reduced accuracy and clustering, it was restored.

The final iterated datasets contained five and 14 taxa, six and eight immune cells, and 21 and 18 genes in tumour and normal tissues, respectively. This allowed for the clustering of three of four poor responders within the training set, and two of three complete responders (Figure 6.6).

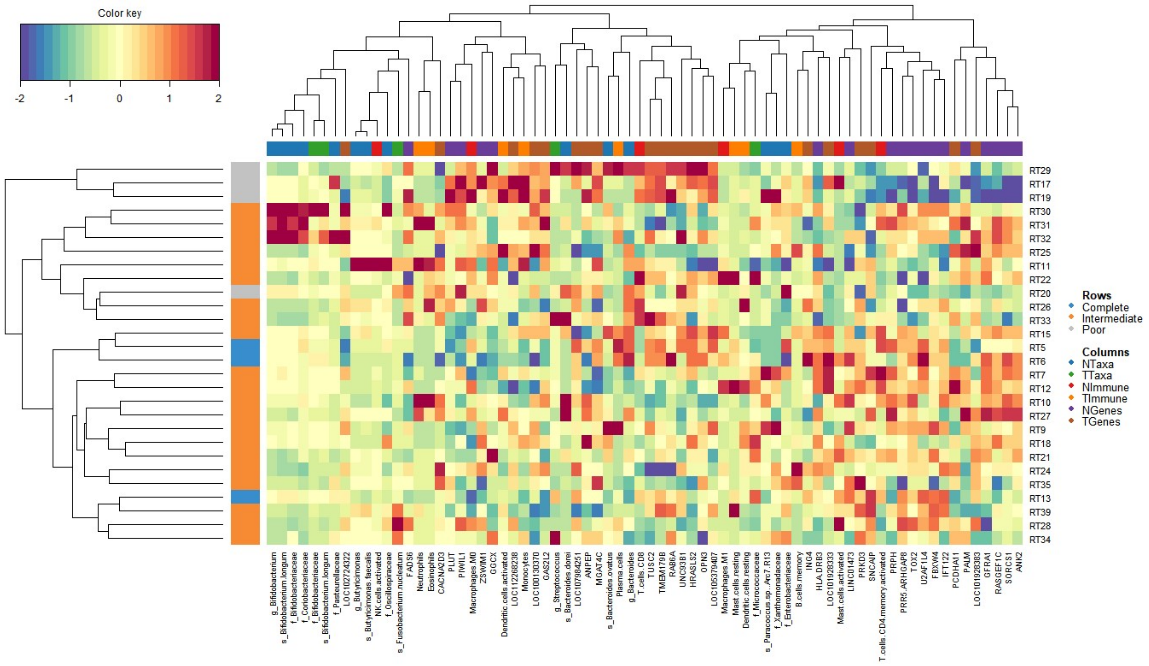

Figure 6.6. Heatmap of clustering by response grade using the final model.

When investigating sPLS-DA plots of variable sets used in the classification, tumour and normal genes were the most discriminatory, followed by normal immune cells. The taxa dataset had a considerable overlap between them in comparison (Figure 6.7).

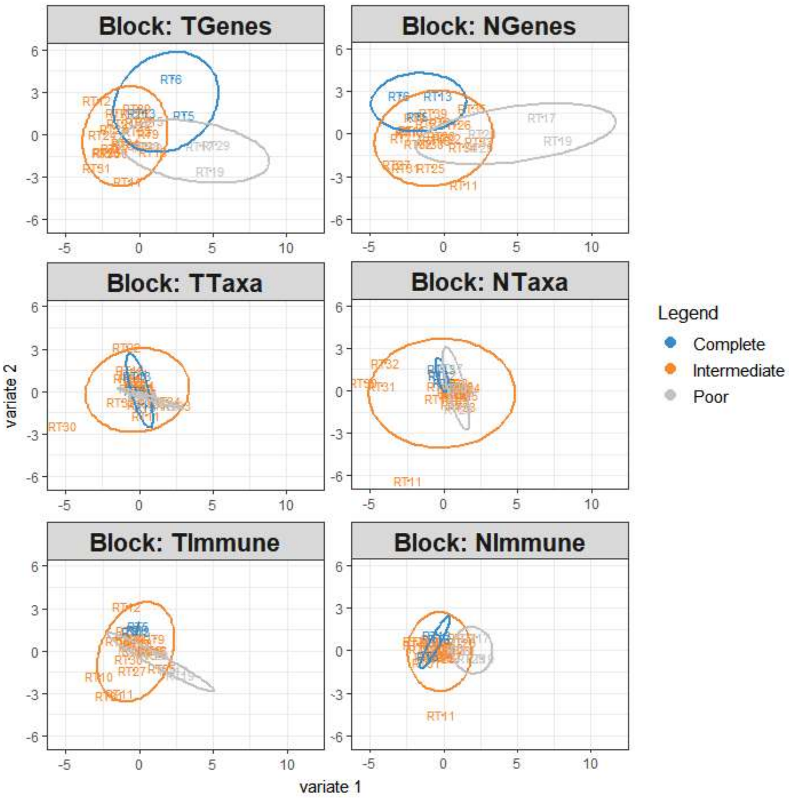

Figure 6.7. Variable sPLS-DA plots

Validation of the final model showed overall error rates between 0.07–0.43 and 0.1–0.52 for balanced error rates (Table 6.5). Using component two gave the lowest error and balanced error rates at 0.07 and 0.1, respectively.

Table 6.5. Final model error rates

Response	comp1	comp2	comp3	comp4	comp5
Complete	0.67	0.00	0.33	0.33	0.67
Intermediate	0.38	0.05	0.10	0.10	0.14
Poor	0.50	0.25	0.50	0.50	0.50
Overall.ER	0.43	0.07	0.18	0.18	0.25
Overall.BER	0.52	0.10	0.31	0.31	0.44

However, using component five gave the test set the greatest prediction accuracy, with only one intermediate patient being misclassified as a complete responder (Table 6.6).

Table 6.6. Prediction accuracy of test set with component five.

		Prediction		
		Complete	**Intermediate**	**Poor**
Truth	**Complete**	2	0	0
	Intermediate	1	6	0
	Poor	0	0	2

Receiver operating characteristic curve calculation showed that the highest contributions to accuracy came from gene expression data, followed by immune and bacterial transcription (Figure 6.8).

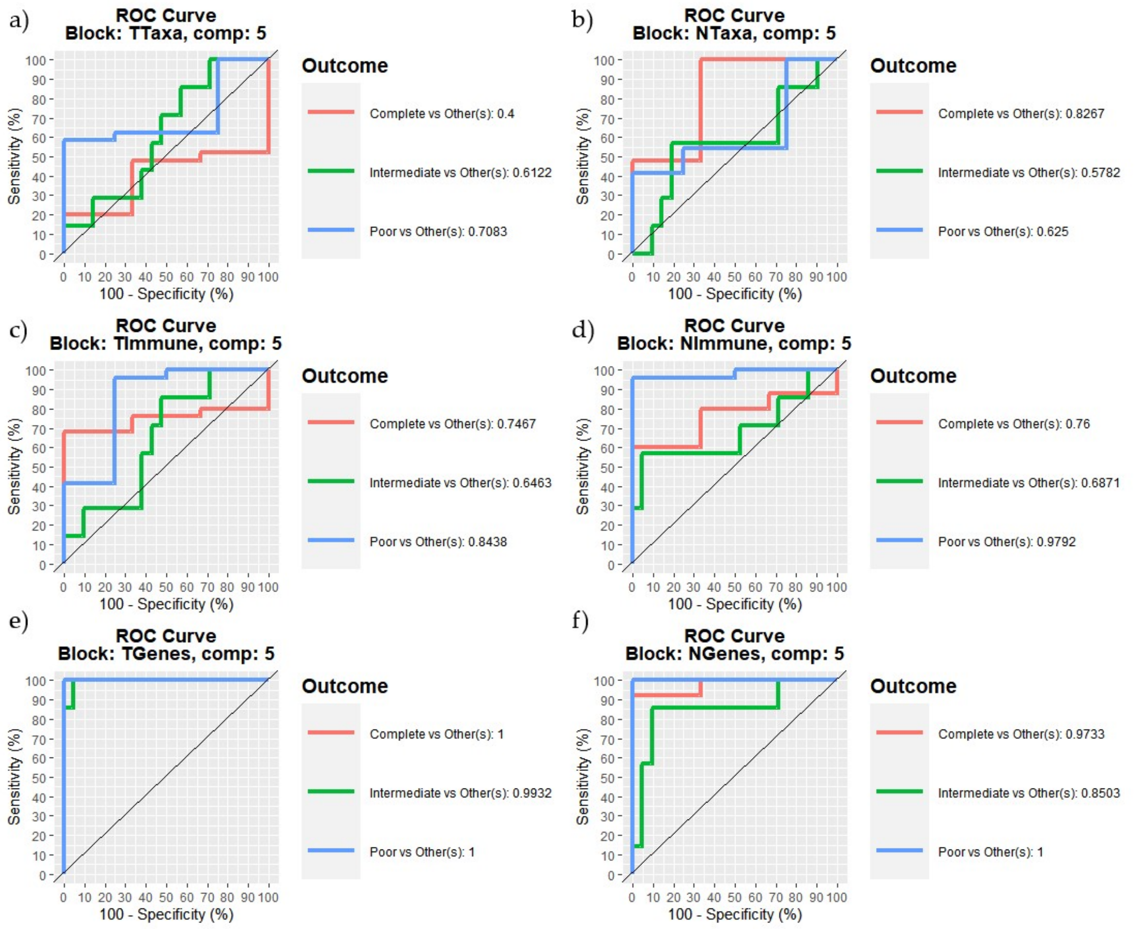

Figure 6.8. ROC curves of each dataset in the final model using component five. ROC: receiver operating characteristic; T: tumour tissue; N: normal tissue.

According to the area under the curve (AUC) statistics, gene expression data was the most statistically significant measure for predicting response (Table 6.7). The only other variable set that was statistically significant was tumour immune cells, which could discriminate between poor responders and other patients.

Table 6.7. Area under the curve (AUC) response grade classification data using component five

		Response Score	**Complete vs Others**	**Intermediate vs Others**	**Poor vs Others**
Taxonomy	**Normal**	**AUC**	0.827	0.578	0.625
		***p*-value**	0.069	0.542	0.431
	Tumour	**AUC**	0.400	0.612	0.708

		***p*-value**	0.578	0.381	0.189
Immune Cells	**Normal**	**AUC**	0.771	0.623	0.722
		***p*-value**	0.053	0.236	0.087
	Tumour	**AUC**	0.760	0.687	0.979
		***p*-value**	0.148	0.145	0.003
Gene Expression	**Normal**	**AUC**	0.973	0.850	1.000
		***p*-value**	0.008	0.006	0.002
	Tumour	**AUC**	1.000	0.993	1.000
		***p*-value**	0.005	<0.001	0.002

6.3.4 Biomarker investigation

6.3.4.1 Genes

Based on the AUC statistics, gene expression data was the strongest and most statistically significant for predicting patient responses to radiotherapy. In both tumour and normal tissue, no other dataset was more capable of predicting outcomes.

Most genes had negative loading weights on poor responders in tumours, while the greatest loading was from *ING4*, a tumour suppressor, and *LOC101928333*, a *GRM8* anti-sense lncRNA. Of genes included in response to therapy in tumour tissues, seven were lncRNAs, of which five remain uncharacterised. However, besides *LOC101928333*, *LOC112268238* is known to protect formylmethionine-tRNA from hydrolysis (Figure 6.9a). Tumour genes predictive of complete response had roles as tumour suppressors, growth regulation, and glutamate receptor silencing. GO terms related to response in tumour tissues were related to immune-related biological processes, including immune effector process (GO:0002252), neutrophil-mediated immunity (GO:0002446), myeloid leukocyte mediated immunity (GO:0002444), immune system process (GO:0002376), leukocyte mediated immunity (GO:0002443), leukocyte activation (GO:0045321); however, no GO term was statistically significant.

The genes expressed in normal tissues that were retained in the model and had highest loading weights toward complete response were for the *PRR5-ARHGAP8* read-through transcript, of which the role has yet to be determined, *PLUT* is a lncRNA and has a role in transcription upregulation and is associated with diabetes (Figure 6.9b). *HLA-DRB3* is typically expressed in antigen-presenting cells for recognition by CD4 T cells, while *FBXW4* is involved in ubiquitination and may have a role in WNT signalling.

The Gene Ontology (GO) terms most related to genes retained in the model in normal tissues were response to stimulus (GO:0050896), cell communication (GO:007154), signal transduction (GO: 007165), cellular response to stimulus (GO: 0051716), signalling (GO: 0023052) and macromolecule metabolic process (GO:0043170); however, these were not statistically significant.

A full list of genes and their roles can be found in Supplementary Table S6.1, while a list of GO terms can be found in Supplementary Table S6.2.

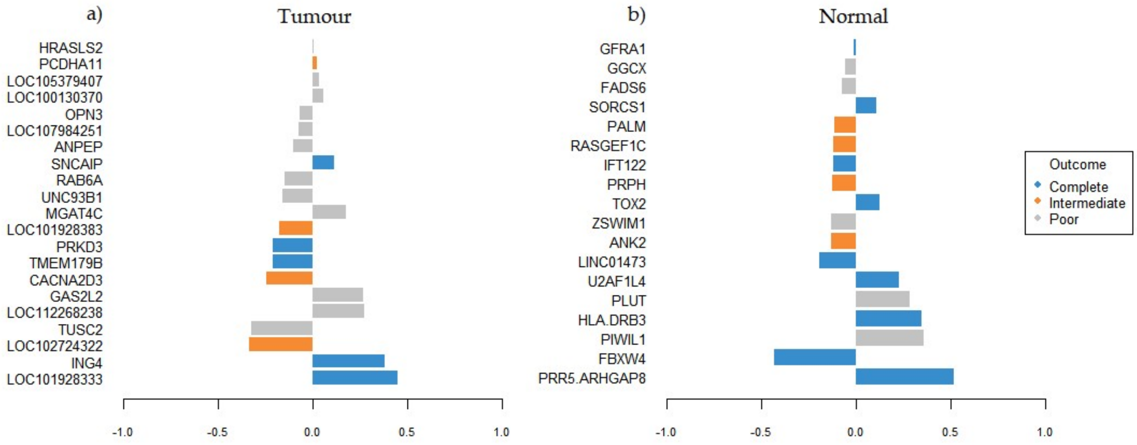

Figure 6.9. Weight loadings on response from genes on component five. a) tumour genes; b) normal genes.

6.3.4.2 Immune cells and bacterial transcription

According to the AUC statistics, the only statistically significant discriminatory power of the immune cell dataset was the abundance of tumour immune cells in differentiating poor responders from others.

Indeed the cells with the highest discriminatory power within tumour tissues were B memory cells loading on an intermediate response, while eosinophils and resting dendritic cells loaded against an intermediate response. Activated dendritic cells weighted toward a poor response (Figure 6.10).

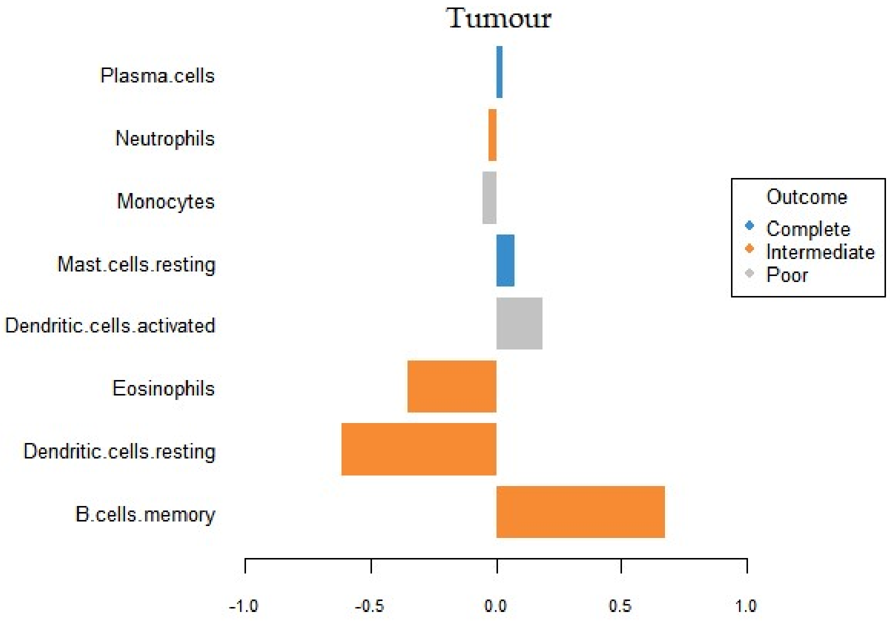

Figure 6.10. Loadings weights on response from tumour immune cells on component five.

According to the AUC statistics, no bacterial transcription in either tumour or normal tissues was statistically significant in terms of their contribution to response. A full list of all bacterial transcription, immune cells and gene transcription included in the final model can be found in Supplementary Table S6.3.

6.4 Discussion

In the final model, the expression of 21 and 18 genes, the transcription of five and 14 bacteria, the abundance of eight and six immune cells, in tumour and normal tissues, respectively, were used to classify patients into different response groups with 94.1% accuracy, with 7.14% overall and 9.92% balanced error rates.

6.4.1 Biomarkers

It could be seen in the sPLS-DA plots that the level of overlap between response groups indicate that different blocks may compensate for the lack of predictive ability of others.

6.4.1.1 Bacterial transcription and immune cells

According to the AUC statistics, there were no taxa that were found to be statistically significant in their predictive ability, which was not unsurprising, as in Chapter Four it was found that bacterial transcription accounted for little variation between response groups. Bacterial transcription may have some role in radiotherapy response; however, this is likely indirect and immune dependant. As suggested previously, the microbiota's contributing role in complete response may be one of support and non-interference, as excessive anti-microbial stimulation before therapy commencement may reduce the number of cells available for neo-antigen presentation and differentiation into anti-tumour effector cells [529].

Overall, this may be due to the lesser importance of immune activation before radiotherapy, which is thought to stimulate anti-tumour responses. It may be the case that immune cells are useful for classifying poor responders as anti-tumour activity is less likely to occur than in intermediate and complete responders before therapy. However, although tumour immune cells were relevant for predicting poor outcomes, this may also be an erroneous conclusion, as their abundance was predicted from gene expression data, so it may be the case that gene expression data effectively inputted to the model twice.

6.4.1.2 Genes

As normal genes were predicting response demonstrates that normal tissue biopsies can deliver untapped and important accessory information for research and clinical applications. Tumour genes predictive of response contained five uncharacterised long noncoding RNAs, and two characterised lncRNAs, one being an anti-sense silencer for a glutamate receptor and ribosomal binding assistance (*LOC101928333* and *LOC112268238*, respectively). In normal tissues, *LINC01473* remains uncharacterised. Overall, this may indicate that lncRNAs have an untapped role as prognostic indicators for radiotherapy response; however, they largely remain poorly understood and uncharacterised [530].

Normal tissue gene expression indicated that genes most predictive of response in normal tissues were related to enteric nerve stimulation and repair, immune regulation, protein digestion and transcription regulation. Enteric nerve-related genes may indicate the rectum's baseline non-pathological status, as it houses extensive enteric nerve conditions compared to the rest of the large intestine [531]. As radiation-induced nerve damage is a common complication of radiation therapy [532] and maintaining enteric nerve function is critical for intestinal health [533], they may provide a novel prognostic indicator for post-radiation quality of life.

ING4, *TUSC2*, *HRASLS2* were three tumour suppressor genes from tumour tissue in the final model that weighted loadings toward a complete response, and against a poor response, respectively, although the latter only slightly. The above may indicate that the gene is not operating the same way in all patients due to: a) The gene is mutated in some patients; b) the response element the gene targets is mutated or not expressed; c) there is an issue with being translated to protein via missing chaperone proteins or protective RNAs.

The inclusion of genes associated with GO terms related to macromolecular and protein metabolic processes is also of interest and could indicate that these tumours can take advantage of necrotic or burst cells in the microenvironment. Alternatively, they may have an increased capacity to degrade crosslinked proteins resulting from ionising radiation ROS generation, a known radioresistance mechanism in other species [534].

6.4.2 Study limitations

One limitation of this study was the low sample sizes from multiple study centres. Although typically a strength, multi-study centres generally increase a studies power; however, in this instance, it may contribute to a cohort effect, and adjusting for the cohort effect, the treatment effect may be interrupted or have its power otherwise reduced. Additionally, due to the inter-cohort differences (only one poor and one complete responder in the PM cohort), the predictive model's utility may be restricted to geographical regions, reducing their potential application across geographical borders.

Another issue is that Dworak scores should not be considered a true continuous variable. Each sample scoring is subjective and relies on the section analysed being representative of the area and differences between institutional or individual techniques and methodologies. The distinction between groups is based on the histopathologist's subjective view; a borderline Dworak Two maybe another pathologist’s Dworak Three, and vice versa. The effect may be exacerbated by the study's multi-centre design, as the translation from one grading system to another may lead a Dworak Two to be classified as Dworak Three in one instance, and not in another.

Ideally, a better way to classify tumour regression for future analyses would be the percentage of fibrosis over a set area, or tumour size reduction as a percentage of the original. Using tumour reduction or overall fibrosis would allow linear allocation of variables, rather than utilising classification methods which rely on data which may not be truly continuous. Unfortunately, it may not be that tumour regression is in reality, a continuous variable with universally defined characteristics. The variability within the studied cohort indicates that response to radiotherapy may be a 'complex trait' with multiple factors to be taken into account, not least of all, the expression of multiple genes.

There were more subtle differences between a moderate response and a near-complete response (Dworak Two and Dworak Three), along with the small number of Dworak Three samples making them challenging to characterise. Additionally, the molecular causation of 'near-complete' is likely less distinct, and not necessarily due to a lesser variable dose than that received by complete responders, which may be what led them to be grouped with other response groups more often.

Patients were grouped into response grades, complete (Dworak Four), intermediate (Dworak Two and Three) and poor, (Dworak One), to better characterise their response to therapy. The alternative grouping aimed to increase the signal for the two extreme responses, complete and poor; and created a third response category, intermediate, which was more easily characterised as not poor or complete. However, using a triplicate grouping also presents issues; primarily that a lack of a clear signal for intermediate patients may be that they are presenting an inhibited complete response or radiosensitisation of a poor response, which is difficult to gauge in the group due to the groups' non-uniformity, and the goal of reducing variables as much as possible. Generating a larger, more variable group for non-complete or non-poor responders is an imprecise method of characterisation, and if used in clinical settings, would not be a reliable way of determining the extent of tumour regression a patient could expect.

Overall, the largest issue was the small sample size, making the training and test sets even smaller (*n* = 22 and *n* = 17, respectively); therefore, the model testing should not be considered rigorous and may still contain overfitting. The final model should be tested against a larger cohort containing data from multiple geographical locations to determine its true prediction accuracy, as the results here may only be valid for the studied cohorts. Additionally, the sPLS-DA methodology employed here is one of many regression-based machine learning approaches to multi-omics datasets, and alternatives such as Random Forest may provide more robust results [535]. Finally, the potential biomarkers were derived purely *in silico* and remain to be validated in a biological setting.

6.4.3 Conclusions and future directions

The panel of biomarker genes identified here could be used in clinical investigations, as RT-qPCR and staining for gene expression and immune cell detection would be most appropriate and effective. Providing RNA-Seq for all patients undergoing prognostic analysis is not currently feasible; however, if costs and technical limitations are reduced, a pipeline utilising the methods in this study could be used to predict response in a single step, rather than with sequential staining or microarrays which may be complicated and costly with more than 20 targets.

The multi-omic design employed here benefited from allowing data blocks that were predictive of one response but not another to compensate for the non-discrimination of another, leading to a more refined signature from a single sequencing experiment. Additionally, the predictive model could not be improved by further removing of variables; this indicates that response to therapy is a complex and likely further involves genes and cells in the microenvironment unidentified by this study, both human and not.

7 CHAPTER 7: DISCUSSION

As described in Chapter 1, response to therapy for RC remains unpredictable, which carries a high-cost burden, both financially and in terms of patient survival rates and quality of life. Being able to predict which patients would respond to therapy and to what extent would be a valuable tool for patients and clinicians to determine which patients would likely have a high therapeutic response, which should be directed to alternatives and be given an accelerated path to surgery. Of the many aspects of the tumour microenvironments, it was discussed that the microbiome might play a role in carcinogenesis and these established microbial communities may, in turn, impact therapeutic response rates.

7.1 Methodological improvements

First, the typical methodology for analysing shotgun sequencing data was assessed and improved upon by including the human genome in the taxonomic database. The time-consuming process of aligning sequencing data to the human genome to remove host reads before taxonomic assignment was shown to be incomplete, leaving substantial residual host reads that were given taxonomic assignment, affecting the interpretation of study results. It was shown that there was a non-statistically significant difference in taxonomic assignment accuracy when host mapping was employed in conjunction with a taxonomic database containing the host genome, making the rate-limiting step in microbiome analysis of host contaminated shotgun sequencing data largely redundant. The result of this first study demonstrates that it would be possible to increase the speed at which microbiome analysis of shotgun sequencing is performed by more than nine-fold, while simultaneously increasing taxonomic assignment accuracy by approximately 11%. However, it remains possible that these results are only applicable to the taxonomic assignment and aligning algorithms used in this study.

7.2 Platform comparisons

The concordance between three sequencing platforms in their comparative ability to evaluate the microbiome: high throughput shotgun RNA-Sequencing; long-read, low depth Oxford Nanopore Technology (ONT) multiplexed sequencing; and 16S rRNA amplicon sequencing. It was found that ONT data produced the least reads; however, the resulting taxonomies remained mostly concordant with the other higher depth sequencing platforms at higher taxonomic levels. At the species level, the concordance between all platforms was decreased. Overall 16S rRNA sequencing was more concordant with RNA-Sequencing taxonomies than either was with those produced by ONT sequencing. This result may be due to 16S rRNA sequencing is usually assessed with dedicated algorithms and that Kraken2's ability to accurately assign taxonomies to 16S rRNA data is a recent development and may not yet be adequately benchmarked.

Additionally, the three platforms measure different things in fundamentally different ways. For instance, 16S rRNA requires amplification of the 16S rRNA markers gene, while the RNA-Sequencing relies on cDNA translation of transcripts and allows the assessment of actively transcribed genes, and finally, ONT sequencing is a direct DNA sequencing method with no amplification or translational steps. All of these techniques carry with them inherent flaws, with the low sequence output of ONT sequencing being the most severe. The inherent error rates of ONT sequencing and resulting low numbers of bacterial reads without the benefit of microbial DNA selection or higher throughput resulted in only two taxa identified some samples. It was decided that due to the highest and most consistent sample depth across the largest number of samples, RNA-Sequencing and bacterial transcription should be utilised for further analysis.

Additionally, RNA-Sequencing was performed on 40 patient samples, as opposed to the other two platforms, which were only performed on the CHCH cohort (*n* = 20).

7.3 Microbial transcription in RC

Assessing microbial transcription in RC tumour and normal tissues showed that alpha diversity was not a significant contributor to response and that individual taxa may contribute to chemoradiotherapy outcomes in RC. Overall, Proteobacteria transcription dominated across all samples and was correlated with a poor response, while Bacteroidetes were negatively correlated with poor response in normal tissues.

Subsequently, taxa within these phyla showed similar correlations, with the Pseudomonas and Bacteroides genera positive and negative correlations with response. The results showed that species such as *C. ureolyticus*, *B. caccae*, *B. vulgatus*, *B. fragilis*, *E. coli*, *K. pneumonia*, *S. enterica*, and *O. splanchicus* were most likely to be contributing to response to radiotherapy. Due to low read depths at the species level, it was not possible to properly investigate their differential gene expression to determine their direct contributions to response. However, it was possible to determine that *B. fragilis*, one of the most transcriptionally active species, was not producing the *bft* toxin, which is thought to contribute to carcinogenesis.

It was hypothesised that due to the number of correlations of microbial transcription occurring in normal tissues and what is known about the correlated species identified, it was possible that the microbiome impacts response to therapy by modulating the immune system directly and via changes in the microenvironment. These changes were speculated to be due to altering the metabolism of lipids, reducing allergic inflammatory responses and mitigating anti-bacterial responses, which would be concordant with the literature on the benefit of high-fat diets during cancer treatments;

however, the direct interaction with the microbiome and immune system requires additional investigation.

Additionally, due to most correlations being with poor response and not with response generally, the extreme nature of the microenvironment of the poorest responders may give them the strongest microbial selection pressure, and thus the strongest microbial signature.

However, not being able to determine the precise activities of putatively involved taxa in response rates, and that some bacterial species may have been misassigned means that further investigation should be carried out to validate these results and to investigate the generated hypotheses.

7.4 Immune infiltration

The presence of immune cells in the respective tissues of RC patients was assessed by using RNA-Sequencing data and a support vector regression algorithm, which resulted in predictions for the abundance of different immune cells and phenotypes within patient samples. Each immune cell population was correlated with response and bacterial transcription data.

Statistically significant correlations of M1 macrophages and resting mast cells within tumour tissues with response identified them as the strongest immune predictor of response. The reasons for this was determined to be the higher populations of these cell types within complete responders, with M1 macrophages being involved in tumour clearance and mast cells having allergic inflammatory and pathogen clearing roles. It was hypothesised that these immune cells are reacting to elements of the microenvironment, with the microbiome possibly modulating anti-tumour immune responses, preventing the inflammatory role of mast cells and allowing M1 macrophages to infiltrate and act upon tumour tissues.

The abundance of immune cells was correlated with bacterial transcription and was found to be associated with a variety of antigen-presenting and effector cells. It was hypothesised that bacterial cells allow for the preservation of antigen-presenting cells that would otherwise display bacterial antigens or those in the tumour microenvironment, instead of tumour neo-antigens that would be produced en masse after ablative ionising radiation therapy. The hypothesis was supported by the proportion of resting effector cells in conjunction with the transcription of bacteria associated with butyrate production, immune tolerance of gut microbiota, changes in tissue metabolic uptake of glucose and lipids and immune stimulation.

Overall, the evidence of the investigation of possible interactions between the microbiome and the immune system indicates that, although they may not have a direct effect on anti-tumour responses, they may modulate the immune system to prevent non-anti-tumour action and maintain the immune system for a rapid anti-tumour response upon the commencement of radiation treatment. However, the possibility remains that the immune cells identified in silico are not present in the predicted proportions and that the phenotypic plasticity of immune cells, particularly neutrophils, remains to be evaluated and may play a role in divergent radiotherapy response.

7.5 Biomarker analysis

In the final analysis, the bacterial transcription, immune cell abundance and gene expression data were used to determine if a model could be formed to predict response to radiotherapy. The aim was to build a predictive model using the least number of predictive markers. The number of discriminant variables was reduced one by one using a training dataset of 28 samples, in order to reduce the number of variables that would have to be investigated in a clinical setting while preserving the predictive ability of the model.

The final model contained 21 and 18 genes, the transcription of five and fourteen bacteria and the abundance of eight and six immune cells in tumour tissues, normal tissues, respectively. The model could accurately classify poor and complete responders in the test dataset, while one intermediate responder was incorrectly classified as a complete responder. The overall error rate of the final model, as assessed by cross-fold validation, was ~7%. The strongest predictive value came from the gene expression in tumour and normal tissues, and the infiltration of immune cells in tumour tissues, which was of particular note, as it shows that normal tissue biopsies have additional prognostic power which could be utilised in future assessments of RC patients. Additionally, this may be due to the normal tissue providing a baseline for gene expression, which was strengthened in the predictive model when combined with pathological tumour gene expression.

During the formation of the predictive model, it was determined that the variability in Dworak Two and Three patients made classifying them into their respective Dworak score groups too challenging. It was surmised that Dworak scores might not be a true continuous variable as some Dworak scores were translated from another scoring method, and different histopathologists, centres and countries were utilised which may add variability of the interpretation of different samples. For instance, a poor responder may be a moderate or poor responder depending on the histopathologists' interpretation, the section used and the time between surgery and completion of treatment. Additionally, a near-complete response may be due to an inhibited anti-tumour response that would have otherwise been complete, while another patient's near-complete response may be an enhanced anti-tumour response that would have otherwise been moderate or poor.

An alternative grouping strategy was adopted to determine the strongest predictors utilising the most extreme outcomes of poor and complete response, allowing the Dworak Two and Three patients to have their responses categorised as not poor or complete. The alternative strategy made finding predictive variables more straight forward; however, this came at the cost of more refined classifications. It was suggested that an alternative measure of tumour regression could be used which would make evaluating it as a true continuous variable more straightforward, such as percentage of fibrosis or comparative tumour and lesion sizes in comparison to the pre-treatment tumour.

Besides the categorisation strategy, the predictive biomarker analysis was hindered by the number of samples in the respective training and test sets with non-representative numbers of responders in each. Additionally, the model may, to an extent, be the result of overfitting, as the immune cell abundance was derived from gene expression data; it may have led to the inclusion of some variables in the model multiple times, albeit in an abstracted way.

7.6 Study limitations

This study suffered most from the limits of sample sizes overall and for respective response groups in each cohort, with the PM cohort only containing one complete responder and one poor responder. Additionally, the lack of full access to the PM cohort's medical history precluded the accounting for additional variables which may impact treatment response, such as ethnicity, comorbidities, and other medications taken during treatment. Furthermore, the number of samples in each response group were not representative, with the majority of samples being moderate responders (Dworak Two), while six or less were near complete, complete or poor responders. There was also a sex imbalance, with the majority of the patients being male, which may have contributed to erroneous study interpretations, particularly in the case of incorporating gene expression data into the biomarker model. Additionally, biopsies were collected by different clinicians from different centres, and interpersonal differences due to patient condition, tumour size and preference of the clinician may have changed the depth and size of biopsies, as well as the distance from the tumour of adjacent normal tissue collection.

7.7 Conclusions and future directions

This study demonstrated that microbiome assessment accuracy and speed can be improved with a simple methodological change, that taxonomies built from ONT sequencing are comparable to 16S rRNA amplicon and RNA-Sequencing if read numbers are high enough. The study also found a possible interaction between the microbiome and the immune system, directly via modulating stimulation and microbial immune tolerance, and indirectly by modulating the tumour microenvironment. It was hypothesised that the microbiome might influence response to therapy by preventing the immune system from acting on aspects of the microenvironment including bacteria, lipids or allergic inflammatory responses, and maintaining naïve effector and antigen-presenting cells to act on neo-antigens produced from radiotherapy-induced tumour cell death.

Finally, the biomarker analysis suggested possible prognostic molecular markers to predict response to radiotherapy in RC patients, with tumour infiltrating immune cells and tumour and normal tissue gene expression being the key determinants, indicating an untapped strength of dual biopsies in clinical practice.

Overall, this study provides methodological improvements and a panel of potential biomarkers that could be used in future validation studies to assess prediction of radiotherapy responses in RC patients. The panel of biomarkers should be further investigated in biological and clinical settings and with larger sample sizes to validate and potentially refine them.

REFERENCES

1. Kuipers, E.J., et al., *Colorectal cancer.* Nat Rev Dis Primers, 2015. **1**: p. 15065.
2. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
3. Ahmed, I. and S. Umar, *Microbiome and Colorectal Cancer.* Current Colorectal Cancer Reports, 2018. **14**(6): p. 217-225.
4. Shang, F.M. and H.L. Liu, *Fusobacterium nucleatum and colorectal cancer: A review.* World J Gastrointest Oncol, 2018. **10**(3): p. 71-81.
5. Sears, C.L., A.L. Geis, and F. Housseau, *Bacteroides fragilis subverts mucosal biology: from symbiont to colon carcinogenesis.* J Clin Invest, 2014. **124**(10): p. 4166-72.
6. Cougnoux, A., et al., *Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype.* Gut, 2014. **63**(12): p. 1932-42.
7. Karpinski, T.M., *Role of Oral Microbiota in Cancer Development.* Microorganisms, 2019. **7**(1): p. 20.
8. Keane, M.G. and G.J. Johnson, *Early diagnosis improves survival in colorectal cancer.* Practitioner, 2012. **256**(1753): p. 15-8, 2.
9. Glynne-Jones, R., et al., *Rectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up.* Ann Oncol, 2017. **28**(suppl_4): p. iv22-iv40.
10. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2016.* CA Cancer J Clin, 2016. **66**(1): p. 7-30.
11. Yamagishi, H., et al., *Molecular pathogenesis of sporadic colorectal cancers.* Chin J Cancer, 2016. **35**: p. 4.
12. Drewes, J.L., F. Housseau, and C.L. Sears, *Sporadic colorectal cancer: microbial contributors to disease prevention, development and therapy.* Br J Cancer, 2016. **115**(3): p. 273-80.
13. Bernabe-Dones, R.D., et al., *High Prevalence of Human Papillomavirus in Colorectal Cancer in Hispanics: A Case-Control Study.* Gastroenterol Res Pract, 2016. **2016**: p. 7896716.
14. Martins, S.F., et al., *Human papillomavirus (HPV) 16 infection is not detected in rectal carcinoma.* Infect Agent Cancer, 2020. **15**(1): p. 17.
15. Dalla Libera, L.S., et al., *Detection of Human papillomavirus and the role of p16INK4a in colorectal carcinomas.* PLoS One, 2020. **15**(6): p. e0235065.
16. Baandrup, L., et al., *The prevalence of human papillomavirus in colorectal adenomas and adenocarcinomas: a systematic review and meta-analysis.* Eur J Cancer, 2014. **50**(8): p. 1446-61.
17. Wei, E.K., et al., *Comparison of risk factors for colon and rectal cancer.* Int J Cancer, 2004. **108**(3): p. 433-42.
18. Meyer, J.E., et al., *Increasing incidence of rectal cancer in patients aged younger than 40 years: an analysis of the surveillance, epidemiology, and end results database.* Cancer, 2010. **116**(18): p. 4354-9.
19. Rawla, P., T. Sunkara, and A. Barsouk, *Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors.* Prz Gastroenterol, 2019. **14**(2): p. 89-103.
20. Franceschi, S., et al., *Menopause and colorectal cancer.* Br J Cancer, 2000. **82**(11): p. 1860-2.
21. Terry, P.D., A.B. Miller, and T.E. Rohan, *Obesity and colorectal cancer risk in women.* Gut, 2002. **51**(2): p. 191-4.
22. Gil Ferreira, C., et al., *KRAS mutations: variable incidences in a Brazilian cohort of 8,234 metastatic colorectal cancer patients.* BMC Gastroenterol, 2014. **14**(1): p. 73.
23. Liu, D., *Concomitant dysregulation of the estrogen receptor and BRAF/MEK signaling pathways is common in colorectal cancer and predicts a worse prognosis.* Cell Oncol (Dordr), 2019. **42**(2): p. 197-209.
24. Caiazza, F., et al., *Estrogen receptors and their implications in colorectal carcinogenesis.* Front Oncol, 2015. **5**: p. 19.
25. He, Y.Q., et al., *Estradiol regulates miR-135b and mismatch repair gene expressions via estrogen receptor-beta in colorectal cells.* Exp Mol Med, 2012. **44**(12): p. 723-32.
26. Ring, K.L., et al., *Endometrial Cancers With Activating KRas Mutations Have Activated Estrogen Signaling and Paradoxical Response to MEK Inhibition.* Int J Gynecol Cancer, 2017. **27**(5): p. 854-862.

27. Maruyama, K., T. Oshima, and K. Ohyama, *Exposure to exogenous estrogen through intake of commercial milk produced from pregnant cows.* Pediatr Int, 2010. **52**(1): p. 33-8.

28. Watson, C.S., G. Hu, and A.A. Paulucci-Holthauzen, *Rapid actions of xenoestrogens disrupt normal estrogenic signaling.* Steroids, 2014. **81**: p. 36-42.

29. Meyer, S.K., et al., *Environmental Xenoestrogens Super-Activate a Variant Murine ER Beta in Cholangiocytes.* Toxicol Sci, 2017. **156**(1): p. 54-71.

30. Crosara Teixeira, M., et al., *Primary prevention of colorectal cancer: myth or reality?* World J Gastroenterol, 2014. **20**(41): p. 15060-9.

31. Chan, D.S., et al., *Red and processed meat and colorectal cancer incidence: meta-analysis of prospective studies.* PLoS One, 2011. **6**(6): p. e20456.

32. Slattery, M.L., *Physical activity and colorectal cancer.* Sports Med, 2004. **34**(4): p. 239-52.

33. Harriss, D.J., et al., *Physical activity before and after diagnosis of colorectal cancer: disease risk, clinical outcomes, response pathways and biomarkers.* Sports Med, 2007. **37**(11): p. 947-60.

34. Harriss, D.J., et al., *Lifestyle factors and colorectal cancer risk (2): a systematic review and meta-analysis of associations with leisure-time physical activity.* Colorectal Dis, 2009. **11**(7): p. 689-701.

35. Moore, S.C., et al., *Association of Leisure-Time Physical Activity With Risk of 26 Types of Cancer in 1.44 Million Adults.* JAMA Intern Med, 2016. **176**(6): p. 816-25.

36. Kelly, S.A., et al., *Prevention of tumorigenesis in mice by exercise is dependent on strain background and timing relative to carcinogen exposure.* Sci Rep, 2017. **7**(1): p. 43086.

37. Steinke, V., et al., *Hereditary nonpolyposis colorectal cancer (HNPCC)/Lynch syndrome.* Dtsch Arztebl Int, 2013. **110**(3): p. 32-8.

38. Beggs, A.D., et al., *Peutz-Jeghers syndrome: a systematic review and recommendations for management.* Gut, 2010. **59**(7): p. 975-86.

39. Brosens, L.A., et al., *Juvenile polyposis syndrome.* World J Gastroenterol, 2011. **17**(44): p. 4839-44.

40. Jasperson, K.W., et al., *Hereditary and familial colon cancer.* Gastroenterology, 2010. **138**(6): p. 2044-58.

41. Half, E., D. Bercovich, and P. Rozen, *Familial adenomatous polyposis.* Orphanet J Rare Dis, 2009. **4**: p. 22.

42. Aaltonen, L.A., et al., *Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease.* N Engl J Med, 1998. **338**(21): p. 1481-7.

43. Burt, R. and D.W. Neklason, *Genetic testing for inherited colon cancer.* Gastroenterology, 2005. **128**(6): p. 1696-716.

44. Kastrinos, F. and S. Syngal, *Inherited colorectal cancer syndromes.* Cancer J, 2011. **17**(6): p. 405-15.

45. Fearnhead, N.S., J.L. Wilding, and W.F. Bodmer, *Genetics of colorectal cancer: hereditary aspects and overview of colorectal tumorigenesis.* Br Med Bull, 2002. **64**(1): p. 27-43.

46. Campos, F.G., M.N. Figueiredo, and C.A. Martinez, *Colorectal cancer risk in hamartomatous polyposis syndromes.* World J Gastrointest Surg, 2015. **7**(3): p. 25-32.

47. Grover, S. and S. Syngal, *Risk assessment, genetic testing, and management of Lynch syndrome.* J Natl Compr Canc Netw, 2010. **8**(1): p. 98-105.

48. Hullar, M.A., A.N. Burnett-Hartman, and J.W. Lampe, *Gut microbes, diet, and cancer.* Cancer Treat Res, 2014. **159**: p. 377-99.

49. Mousavinezhad, M., et al., *The effectiveness of FOBT vs. FIT: A meta-analysis on colorectal cancer screening test.* Med J Islam Repub Iran, 2016. **30**: p. 366.

50. Swiderska, M., et al., *The diagnostics of colorectal cancer.* Contemp Oncol (Pozn), 2014. **18**(1): p. 1-6.

51. NZ., M.o.H. *National Bowel Screening Programme*. 2014; Available from: https://www.health.govt.nz/our-work/preventative-health-wellness/screening/national-bowel-screening-programme.

52. Gandhi, J., et al., *Population-based study demonstrating an increase in colorectal cancer in young patients.* Br J Surg, 2017. **104**(8): p. 1063-1068.

53. Brandeau, M.L. and D.M. Eddy, *The workup of the asymptomatic patient with a positive fecal occult blood test.* Med Decis Making, 1987. **7**(1): p. 32-46.

54. Lieberman, D.A., et al., *Use of colonoscopy to screen asymptomatic adults for colorectal cancer. Veterans Affairs Cooperative Study Group 380.* N Engl J Med, 2000. **343**(3): p. 162-8.

55. Barresi, V., et al., *Histological grading in colorectal cancer: new insights and perspectives.* Histol Histopathol, 2015. **30**(9): p. 1059-67.

56. Bipat, S., et al., *Rectal cancer: local staging and assessment of lymph node involvement with endoluminal US, CT, and MR imaging--a meta-analysis.* Radiology, 2004. **232**(3): p. 773-83.
57. Pan, H.D., et al., *Pulmonary metastasis in rectal cancer: a retrospective study of clinicopathological characteristics of 404 patients in Chinese cohort.* BMJ Open, 2018. **8**(2): p. e019614.
58. Sharples, K.J., et al., *The New Zealand PIPER Project: colorectal cancer survival according to rurality, ethnicity and socioeconomic deprivation-results from a retrospective cohort study.* N Z Med J, 2018. **131**(1476): p. 24-39.
59. Welfare, A.I.o.H.a., *Cancer in Australia.* 2019.
60. Morris, M., B. Iacopetta, and C. Platell, *Comparing survival outcomes for patients with colorectal cancer treated in public and private hospitals.* Medical Journal of Australia, 2007. **186**(6): p. 296-300.
61. Shafik, A., et al., *Functional activity of the rectum: A conduit organ or a storage organ or both?* World J Gastroenterol, 2006. **12**(28): p. 4549-52.
62. Yoon, K. and N. Kim, *The Effect of Microbiota on Colon Carcinogenesis.* J Cancer Prev, 2018. **23**(3): p. 117-125.
63. van Hoogdalem, E., A.G. de Boer, and D.D. Breimer, *Pharmacokinetics of rectal drug administration, Part I. General considerations and clinical applications of centrally acting drugs.* Clin Pharmacokinet, 1991. **21**(1): p. 11-26.
64. Siddharth, P. and B. Ravo, *Colorectal neurovasculature and anal sphincter.* Surg Clin North Am, 1988. **68**(6): p. 1185-200.
65. Laohavinij, S., J. Maneechavakajorn, and P. Techatanol, *Prognostic factors for survival in colorectal cancer patients.* J Med Assoc Thai, 2010. **93**(10): p. 1156-66.
66. De Divitiis, C., et al., *Prognostic and predictive response factors in colorectal cancer patients: between hope and reality.* World J Gastroenterol, 2014. **20**(41): p. 15049-59.
67. Khan, M.A.S., et al., *The Impact of Tumour Distance From the Anal Verge on Clinical Management and Outcomes in Patients Having a Curative Resection for Rectal Cancer.* J Gastrointest Surg, 2017. **21**(12): p. 2056-2065.
68. Ursell, L.K., et al., *Defining the human microbiome.* Nutr Rev, 2012. **70 Suppl 1**(Suppl 1): p. S38-44.
69. Shreiner, A.B., J.Y. Kao, and V.B. Young, *The gut microbiome in health and in disease.* Curr Opin Gastroenterol, 2015. **31**(1): p. 69-75.
70. Sharon, G., et al., *Specialized metabolites from the microbiome in health and disease.* Cell Metab, 2014. **20**(5): p. 719-730.
71. Krajmalnik-Brown, R., et al., *Effects of gut microbes on nutrient absorption and energy regulation.* Nutr Clin Pract, 2012. **27**(2): p. 201-14.
72. Mathewson, N.D., et al., *Gut microbiome–derived metabolites modulate intestinal epithelial cell damage and mitigate graft-versus-host disease.* Nature Immunology, 2016. **17**: p. 505.
73. Round, J.L. and S.K. Mazmanian, *The gut microbiota shapes intestinal immune responses during health and disease.* Nat Rev Immunol, 2009. **9**(5): p. 313-23.
74. Bonder, M.J., et al., *The effect of host genetics on the gut microbiome.* Nat Genet, 2016. **48**(11): p. 1407-1412.
75. Singh, R.K., et al., *Influence of diet on the gut microbiome and implications for human health.* J Transl Med, 2017. **15**(1): p. 73.
76. Woods, J.A., et al., *Exercise alters the gut microbiome and microbial metabolites: Implications for colorectal cancer and inflammatory bowel disease.* Brain, Behavior, and Immunity, 2015. **49**: p. e7.
77. Lederer, A.K., et al., *Postoperative changes of the microbiome: are surgical complications related to the gut flora? A systematic review.* BMC Surg, 2017. **17**(1): p. 125.
78. Raskov, H., et al., *Irritable bowel syndrome, the microbiota and the gut-brain axis.* Gut Microbes, 2016. **7**(5): p. 365-83.
79. Magnusdottir, S., et al., *Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes.* Front Genet, 2015. **6**: p. 148.
80. Li, Q., et al., *The Gut Microbiota and Autism Spectrum Disorders.* Front Cell Neurosci, 2017. **11**: p. 120.
81. Schwabe, R.F. and C. Jobin, *The microbiome and cancer.* Nat Rev Cancer, 2013. **13**(11): p. 800-12.
82. Akbar, N., et al., *Gut bacteria of cockroaches are a potential source of antibacterial compound(s).* Lett Appl Microbiol, 2018. **66**(5): p. 416-426.
83. Morrison, D.J. and T. Preston, *Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism.* Gut Microbes, 2016. **7**(3): p. 189-200.
84. Wu, W., et al., *Microbiota metabolite short-chain fatty acid acetate promotes intestinal IgA response to microbiota which is mediated by GPR43.* Mucosal Immunol, 2017. **10**(4): p. 946-956.

85. Furusawa, Y., et al., *Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells.* Nature, 2013. **504**(7480): p. 446-50.

86. Velazquez, O.C., H.M. Lederer, and J.L. Rombeau, *Butyrate and the colonocyte. Production, absorption, metabolism, and therapeutic implications.* Adv Exp Med Biol, 1997. **427**: p. 123-34.

87. Bultman, S.J., *Butyrate consumption of differentiated colonocytes in the upper crypt promotes homeostatic proliferation of stem and progenitor cells near the crypt base.* Transl Cancer Res, 2016. **5**(Suppl 3): p. S526-S528.

88. Fleming, S.E., A.U. O'Donnell, and J.A. Perman, *Influence of frequent and long-term bean consumption on colonic function and fermentation.* Am J Clin Nutr, 1985. **41**(5): p. 909-18.

89. Muir, J.G., et al., *Modulation of fecal markers relevant to colon cancer risk: a high-starch Chinese diet did not generate expected beneficial changes relative to a Western-type diet.* Am J Clin Nutr, 1998. **68**(2): p. 372-9.

90. Takahashi, H., et al., *Effect of partially hydrolyzed guar gum on fecal output in human volunteers.* Nutrition Research, 1993. **13**(6): p. 649-657.

91. Hoverstad, T., et al., *Short-chain fatty acids in the normal human feces.* Scand J Gastroenterol, 1984. **19**(3): p. 375-81.

92. Topping, D.L. and P.M. Clifton, *Short-chain fatty acids and human colonic function: roles of resistant starch and nonstarch polysaccharides.* Physiol Rev, 2001. **81**(3): p. 1031-64.

93. Purcell, R.V., et al., *Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer.* Sci Rep, 2017. **7**(1): p. 11590.

94. Ito, M., et al., *Association of Fusobacterium nucleatum with clinical and molecular features in colorectal serrated pathway.* Int J Cancer, 2015. **137**(6): p. 1258-68.

95. Malfertheiner, P., et al., *Helicobacter pylori eradication has the potential to prevent gastric cancer: a state-of-the-art critique.* Am J Gastroenterol, 2005. **100**(9): p. 2100-15.

96. Rowan, A.J., et al., *APC mutations in sporadic colorectal tumors: A mutational "hotspot" and interdependence of the "two hits".* Proc Natl Acad Sci U S A, 2000. **97**(7): p. 3352-7.

97. Kumar, M., et al., *Negative regulation of the tumor suppressor p53 gene by microRNAs.* Oncogene, 2011. **30**(7): p. 843-53.

98. Segditsas, S., et al., *APC and the three-hit hypothesis.* Oncogene, 2009. **28**(1): p. 146-55.

99. Kobayashi, J., *Effect of diet and gut environment on the gastrointestinal formation of N-nitroso compounds: A review.* Nitric Oxide, 2018. **73**: p. 66-73.

100. Pritchard, C.C., et al., *Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer.* N Engl J Med, 2016. **375**(5): p. 443-53.

101. Shimada, Y., et al., *Genomic overview of right-sided and left-sided colorectal cancer using comprehensive genomic sequencing.* Journal of Clinical Oncology, 2017. **35**(15_suppl): p. e15101-e15101.

102. Vacca, I., *Microbiome: The microbiota maintains oxygen balance in the gut.* Nat Rev Microbiol, 2017. **15**(10): p. 574-575.

103. Esteban, D.J., B. Hysa, and C. Bartow-McKenney, *Temporal and Spatial Distribution of the Microbial Community of Winogradsky Columns.* PLoS One, 2015. **10**(8): p. e0134588.

104. Zoetendal, E.G., et al., *The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates.* ISME J, 2012. **6**(7): p. 1415-26.

105. Brown, J.M., *Tumor hypoxia in cancer therapy.* Methods Enzymol, 2007. **435**: p. 297-321.

106. Jiang, B., *Aerobic glycolysis and high level of lactate in cancer metabolism and microenvironment.* Genes Dis, 2017. **4**(1): p. 25-27.

107. Jeong, H., et al., *Radiation-induced immune responses: mechanisms and therapeutic perspectives.* Blood Res, 2016. **51**(3): p. 157-163.

108. Thomas, A.M., et al., *Tissue-Associated Bacterial Alterations in Rectal Carcinoma Patients Revealed by 16S rRNA Community Profiling.* Front Cell Infect Microbiol, 2016. **6**(DEC): p. 179.

109. Warren, R.L., et al., *Co-occurrence of anaerobic bacteria in colorectal carcinomas.* Microbiome, 2013. **1**(1): p. 16.

110. Administration, U.S.F.a.D., *FDA approves first cancer treatment for any solid tumor with a specific genetic feature.* 2017.

111. Loupakis, F., et al., *Impact of primary tumour location on efficacy of bevacizumab plus chemotherapy in metastatic colorectal cancer.* Br J Cancer, 2018. **119**(12): p. 1451-1455.

112. Ulivi, P., et al., *Right- vs. Left-Sided Metastatic Colorectal Cancer: Differences in Tumor Biology and Bevacizumab Efficacy.* Int J Mol Sci, 2017. **18**(6): p. 1240.

113. Christodoulidis, G., et al., *Clinicopathological differences between right- and left-sided colonic tumors and impact upon survival.* Tech Coloproctol, 2010. **14 Suppl 1**(S1): p. S45-7.

114. Tjalsma, H., et al., *A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects.* Nat Rev Microbiol, 2012. **10**(8): p. 575-82.

115. Geng, J., et al., *Co-occurrence of driver and passenger bacteria in human colorectal cancer.* Gut Pathog, 2014. **6**: p. 26.

116. Hajishengallis, G., R.P. Darveau, and M.A. Curtis, *The keystone-pathogen hypothesis.* Nat Rev Microbiol, 2012. **10**(10): p. 717-25.

117. Flynn, K.J., N.T. Baxter, and P.D. Schloss, *Metabolic and Community Synergy of Oral Bacteria in Colorectal Cancer.* mSphere, 2016. **1**(3).

118. Drewes, J.L., et al., *High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia.* NPJ Biofilms Microbiomes, 2017. **3**: p. 34.

119. Kinross, J., et al., *A prospective analysis of mucosal microbiome-metabonome interactions in colorectal cancer using a combined MAS 1HNMR and metataxonomic strategy.* Sci Rep, 2017. **7**(1): p. 8979.

120. Swidsinski, A., et al., *Comparative study of the intestinal mucus barrier in normal and inflamed colon.* Gut, 2007. **56**(3): p. 343-50.

121. Dejea, C.M., et al., *Microbiota organization is a distinct feature of proximal colorectal cancers.* Proc Natl Acad Sci U S A, 2014. **111**(51): p. 18321-6.

122. Jiang, A., et al., *Disruption of E-cadherin-mediated adhesion induces a functionally distinct pathway of dendritic cell maturation.* Immunity, 2007. **27**(4): p. 610-24.

123. Costa, A.M., et al., *Adherens junctions as targets of microorganisms: a focus on Helicobacter pylori.* FEBS Lett, 2013. **587**(3): p. 259-65.

124. Klezovitch, O. and V. Vasioukhin, *Cadherin signaling: keeping cells in touch.* F1000Res, 2015. **4**(F1000 Faculty Rev): p. 550.

125. Wu, S., et al., *Bacteroides fragilis enterotoxin cleaves the zonula adherens protein, E-cadherin.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14979-84.

126. Sears, C.L., *The toxins of Bacteroides fragilis.* Toxicon, 2001. **39**(11): p. 1737-46.

127. Clevers, H. and R. Nusse, *Wnt/beta-catenin signaling and disease.* Cell, 2012. **149**(6): p. 1192-205.

128. Rubinstein, M.R., et al., *Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/beta-catenin signaling via its FadA adhesin.* Cell Host Microbe, 2013. **14**(2): p. 195-206.

129. Fardini, Y., et al., *Fusobacterium nucleatum adhesin FadA binds vascular endothelial cadherin and alters endothelial integrity.* Mol Microbiol, 2011. **82**(6): p. 1468-80.

130. Heimesaat, M.M., et al., *The role of serine protease HtrA in acute ulcerative enterocolitis and extra-intestinal immune responses during Campylobacter jejuni infection of gnotobiotic IL-10 deficient mice.* Front Cell Infect Microbiol, 2014. **4**: p. 77.

131. Hoy, B., et al., *Distinct roles of secreted HtrA proteases from gram-negative pathogens in cleaving the junctional protein and tumor suppressor E-cadherin.* J Biol Chem, 2012. **287**(13): p. 10115-20.

132. Purdy, G.E., C.R. Fisher, and S.M. Payne, *IcsA surface presentation in Shigella flexneri requires the periplasmic chaperones DegP, Skp, and SurA.* J Bacteriol, 2007. **189**(15): p. 5566-73.

133. Zhang, Y., et al., *The Role of E-cadherin in Helicobacter pylori-Related Gastric Diseases.* Curr Drug Metab, 2019. **20**(1): p. 23-28.

134. Reboud, E., et al., *Pseudomonas aeruginosa ExlA and Serratia marcescens ShlA trigger cadherin cleavage by promoting calcium influx and ADAM10 activation.* PLoS Pathog, 2017. **13**(8): p. e1006579.

135. Seike, S., et al., *Clostridium perfringens Delta-Toxin Damages the Mouse Small Intestine.* Toxins (Basel), 2019. **11**(4): p. 232.

136. Steck, N., et al., *Enterococcus faecalis metalloprotease compromises epithelial barrier and contributes to intestinal inflammation.* Gastroenterology, 2011. **141**(3): p. 959-71.

137. Bendardaf, R., et al., *Cytoplasmic E-Cadherin Expression Is Associated With Higher Tumour Level of VEGFA, Lower Response Rate to Irinotecan-based Treatment and Poorer Prognosis in Patients With Metastatic Colorectal Cancer.* Anticancer Res, 2019. **39**(4): p. 1953-1957.

138. Wilson, M.R., et al., *The human gut bacterial genotoxin colibactin alkylates DNA.* Science, 2019. **363**(6428): p. eaar7785.

139. Yazici, C., et al., *Race-dependent association of sulfidogenic bacteria with colorectal cancer.* Gut, 2017. **66**(11): p. 1983-1994.

140. Attene-Ramos, M.S., et al., *Evidence that hydrogen sulfide is a genotoxic agent.* Mol Cancer Res, 2006. **4**(1): p. 9-14.

141. Pickard, J.M., et al., *Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease.* Immunol Rev, 2017. **279**(1): p. 70-89.

142. Lee, N. and W.U. Kim, *Microbiota in T-cell homeostasis and inflammatory diseases.* Exp Mol Med, 2017. **49**(5): p. e340.

143. Peng, L., et al., *Butyrate enhances the intestinal barrier by facilitating tight junction assembly via activation of AMP-activated protein kinase in Caco-2 cell monolayers.* J Nutr, 2009. **139**(9): p. 1619-25.

144. Rodriguez-Pineiro, A.M. and M.E. Johansson, *The colonic mucus protection depends on the microbiota.* Gut Microbes, 2015. **6**(5): p. 326-30.

145. Zechner, E.L., *Inflammatory disease caused by intestinal pathobionts.* Curr Opin Microbiol, 2017. **35**: p. 64-69.

146. Tegtmeier, D., et al., *Oxygen Affects Gut Bacterial Colonization and Metabolic Activities in a Gnotobiotic Cockroach Model.* Appl Environ Microbiol, 2016. **82**(4): p. 1080-1089.

147. Sun, J. and I. Kato, *Gut microbiota, inflammation and colorectal cancer.* Genes Dis, 2016. **3**(2): p. 130-143.

148. Salaspuro, M., *Acetaldehyde as a common denominator and cumulative carcinogen in digestive tract cancers.* Scand J Gastroenterol, 2009. **44**(8): p. 912-25.

149. Garsin, D.A., *Ethanolamine utilization in bacterial pathogens: roles and regulation.* Nat Rev Microbiol, 2010. **8**(4): p. 290-5.

150. Armstrong, H., et al., *The Complex Interplay between Chronic Inflammation, the Microbiome, and Cancer: Understanding Disease Progression and What We Can Do to Prevent It.* Cancers (Basel), 2018. **10**(3).

151. Flemer, B., et al., *The oral microbiota in colorectal cancer is distinctive and predictive.* Gut, 2017.

152. Srivastava, A., et al., *Gut biofilm forming bacteria in inflammatory bowel disease.* Microb Pathog, 2017. **112**: p. 5-14.

153. Washio, J. and N. Takahashi, *Metabolomic Studies of Oral Biofilm, Oral Cancer, and Beyond.* Int J Mol Sci, 2016. **17**(6).

154. Zeller, G., et al., *Potential of fecal microbiota for early-stage detection of colorectal cancer.* Mol Syst Biol, 2014. **10**(11): p. 766.

155. Boles, B.R., M. Thoendel, and P.K. Singh, *Self-generated diversity produces "insurance effects" in biofilm communities.* Proc Natl Acad Sci U S A, 2004. **101**(47): p. 16630-5.

156. Merod, R.T. and S. Wuertz, *Extracellular polymeric substance architecture influences natural genetic transformation of Acinetobacter baylyi in biofilms.* Appl Environ Microbiol, 2014. **80**(24): p. 7752-7.

157. Donlan, R.M., *Biofilms: microbial life on surfaces.* Emerg Infect Dis, 2002. **8**(9): p. 881-90.

158. Huang, R., M. Li, and R.L. Gregory, *Bacterial interactions in dental biofilm.* Virulence, 2011. **2**(5): p. 435-44.

159. Zhao, H., et al., *Variations in oral microbiota associated with oral cancer.* Sci Rep, 2017. **7**(1): p. 11773.

160. Flemming, H.C., et al., *Biofilms: an emergent form of bacterial life.* Nat Rev Microbiol, 2016. **14**(9): p. 563-75.

161. Baran, B., et al., *Difference Between Left-Sided and Right-Sided Colorectal Cancer: A Focused Review of Literature.* Gastroenterology Res, 2018. **11**(4): p. 264-273.

162. Lloyd-Price, J., G. Abu-Ali, and C. Huttenhower, *The healthy human microbiome.* Genome Med, 2016. **8**(1): p. 51.

163. Brown, D.G., et al., *Metabolomics and metabolic pathway networks from human colorectal cancers, adjacent mucosa, and stool.* Cancer Metab, 2016. **4**(1): p. 11.

164. Flemer, B., et al., *Tumour-associated and non-tumour-associated microbiota in colorectal cancer.* Gut, 2017. **66**(4): p. 633-643.

165. Ley, R.E., et al., *Evolution of mammals and their gut microbes.* Science, 2008. **320**(5883): p. 1647-51.

166. Kaper, J.B., J.P. Nataro, and H.L. Mobley, *Pathogenic Escherichia coli.* Nat Rev Microbiol, 2004. **2**(2): p. 123-40.

167. Burns, M.B., et al., *Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment.* Genome Med, 2015. **7**(1): p. 55.

168. Geng, J., et al., *Diversified pattern of the human colorectal cancer microbiome.* Gut Pathog, 2013. **5**(1): p. 2.

169. Mira-Pascual, L., et al., *Microbial mucosal colonic shifts associated with the development of colorectal cancer reveal the presence of different bacterial and archaeal biomarkers.* J Gastroenterol, 2015. **50**(2): p. 167-79.

170. Chen, W., et al., *Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer.* PLoS One, 2012. **7**(6): p. e39743.

171. Kostic, A.D., et al., *Genomic analysis identifies association of Fusobacterium with colorectal carcinoma.* Genome Res, 2012. **22**(2): p. 292-8.

172. Allali, I., et al., *Gut microbiome compositional and functional differences between tumor and non-tumor adjacent tissues from cohorts from the US and Spain.* Gut Microbes, 2015. **6**(3): p. 161-72.

173. Nakatsu, G., et al., *Gut mucosal microbiome across stages of colorectal carcinogenesis.* Nat Commun, 2015. **6**: p. 8727.

174. Wei, Z., et al., *Could gut microbiota serve as prognostic biomarker associated with colorectal cancer patients' survival? A pilot study on relevant mechanism.* Oncotarget, 2016. **7**(29): p. 46158-46172.

175. Xu, K. and B. Jiang, *Analysis of Mucosa-Associated Microbiota in Colorectal Cancer.* Med Sci Monit, 2017. **23**: p. 4422-4430.

176. Duvallet, C., et al., *Meta-analysis of gut microbiome studies identifies disease-specific and shared responses.* Nat Commun, 2017. **8**(1): p. 1784.

177. Arumugam, M., et al., *Enterotypes of the human gut microbiome.* Nature, 2011. **473**(7346): p. 174-80.

178. Sobhani, I., et al., *Colorectal cancer-associated microbiota contributes to oncogenic epigenetic signatures.* Proc Natl Acad Sci U S A, 2019. **116**(48): p. 24285-24295.

179. Marchesi, J.R., et al., *Towards the human colorectal cancer microbiome.* PLoS One, 2011. **6**(5): p. e20447.

180. Flemer, B., et al., *The oral microbiota in colorectal cancer is distinctive and predictive.* Gut, 2018. **67**(8): p. 1454-1463.

181. Hale, V.L., et al., *Distinct microbes, metabolites, and ecologies define the microbiome in deficient and proficient mismatch repair colorectal cancers.* Genome Med, 2018. **10**(1): p. 78.

182. Castellarin, M., et al., *Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma.* Genome Res, 2012. **22**(2): p. 299-306.

183. Swidsinski, A., et al., *Association between intraepithelial Escherichia coli and colorectal cancer.* Gastroenterology, 1998. **115**(2): p. 281-286.

184. Popova, C., V. Dosseva-Panova, and V. Panov, *Microbiology of Periodontal Diseases. A Review.* Biotechnology & Biotechnological Equipment, 2014. **27**(3): p. 3754-3759.

185. Thurnheer, T., et al., *Fusobacterium Species and Subspecies Differentially Affect the Composition and Architecture of Supra- and Subgingival Biofilms Models.* Front Microbiol, 2019. **10**(1716): p. 1716.

186. Parada Venegas, D., et al., *Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases.* Front Immunol, 2019. **10**(277): p. 277.

187. Trimble, E.L., et al., *Neoadjuvant therapy in cancer treatment.* Cancer, 1993. **72**(11 Suppl): p. 3515-24.

188. Thies, S. and R. Langer, *Tumor regression grading of gastrointestinal carcinomas after neoadjuvant treatment.* Front Oncol, 2013. **3**: p. 262.

189. Kim, S.H., et al., *What Is the Ideal Tumor Regression Grading System in Rectal Cancer Patients after Preoperative Chemoradiotherapy?* Cancer Res Treat, 2016. **48**(3): p. 998-1009.

190. Longley, D.B., D.P. Harkin, and P.G. Johnston, *5-fluorouracil: mechanisms of action and clinical strategies.* Nat Rev Cancer, 2003. **3**(5): p. 330-8.

191. Faugeras, L., et al., *Treatment options for metastatic colorectal cancer in patients with liver dysfunction due to malignancy.* Crit Rev Oncol Hematol, 2017. **115**: p. 59-66.

192. Mikhail, S.E., J.F. Sun, and J.L. Marshall, *Safety of capecitabine: a review.* Expert Opin Drug Saf, 2010. **9**(5): p. 831-41.

193. Danenberg, P.V., et al., *Folates as adjuvants to anticancer agents: Chemical rationale and mechanism of action.* Crit Rev Oncol Hematol, 2016. **106**: p. 118-31.

194. Dasari, S. and P.B. Tchounwou, *Cisplatin in cancer therapy: molecular mechanisms of action.* Eur J Pharmacol, 2014. **740**: p. 364-78.

195. Dong, Y., et al., *Chemoradiation Cancer Therapy: Molecular Mechanisms of Cisplatin Radiosensitization.* The Journal of Physical Chemistry C, 2017. **121**(32): p. 17505-17513.

196. Holsti, L.R., *Development of clinical radiotherapy since 1896.* Acta Oncol, 1995. **34**(8): p. 995-1003.

197. Baskar, R., et al., *Biological response of cancer cells to radiation treatment.* Front Mol Biosci, 2014. **1**: p. 24.

198. Dutta, S.W., et al., *Short-course Versus Long-course Neoadjuvant Therapy for Non-metastatic Rectal Cancer: Patterns of Care and Outcomes From the National Cancer Database.* Clin Colorectal Cancer, 2018. **17**(4): p. 297-306.

199. Bousquet, P.A., et al., *Markers of Mitochondrial Metabolism in Tumor Hypoxia, Systemic Inflammation, and Adverse Outcome of Rectal Cancer.* Transl Oncol, 2019. **12**(1): p. 76-83.

200. Sun, Y., et al., *Hypoxia-induced autophagy reduces radiosensitivity by the HIF-1alpha/miR-210/Bcl-2 pathway in colon cancer cells.* Int J Oncol, 2015. **46**(2): p. 750-6.

201. Leszczynska, K.B., et al., *Hypoxia-induced p53 modulates both apoptosis and radiosensitivity via AKT.* J Clin Invest, 2015. **125**(6): p. 2385-98.

202. Kuiper, C., et al., *Increased Tumor Ascorbate is Associated with Extended Disease-Free Survival and Decreased Hypoxia-Inducible Factor-1 Activation in Human Colorectal Cancer.* Front Oncol, 2014. **4**: p. 10.

203. Hsiao, H.T., et al., *Hypoxia-targeted triple suicide gene therapy radiosensitizes human colorectal cancer cells.* Oncol Rep, 2014. **32**(2): p. 723-9.

204. Ali, R. and E.E. Graves, *Targeted therapies and hypoxia imaging.* Q J Nucl Med Mol Imaging, 2013. **57**(3): p. 283-95.

205. Peng, F. and M. Chen, *Antiangiogenic therapy: a novel approach to overcome tumor hypoxia.* Chin J Cancer, 2010. **29**(8): p. 715-20.

206. Mehta, S.R., et al., *Radiotherapy: Basic Concepts and Recent Advances.* Med J Armed Forces India, 2010. **66**(2): p. 158-62.

207. Berkey, F.J., *Managing the adverse effects of radiation therapy.* Am Fam Physician, 2010. **82**(4): p. 381-8, 394.

208. Delibegovic, S., *Introduction to Total Mesorectal Excision.* Med Arch, 2017. **71**(6): p. 434-438.

209. Rociu, E., et al., *Normal anal sphincter anatomy and age- and sex-related variations at high-spatial-resolution endoanal MR imaging.* Radiology, 2000. **217**(2): p. 395-401.

210. Sergio P Regadas, F., et al., *Anal canal anatomy showed by three-dimensional anorectal ultrasonography*. Vol. 21. 2007. 2207-11.

211. Nivatvongs, S., H.S. Stern, and D.S. Fryd, *The length of the anal canal.* Dis Colon Rectum, 1981. **24**(8): p. 600-1.

212. Saunders, B.P., et al., *Why is colonoscopy more difficult in women?* Gastrointestinal Endoscopy, 1996. **43**(2): p. 124-126.

213. La Monica, G., et al., *Incidence of sexual dysfunction in male patients treated surgically for rectal malignancy.* Dis Colon Rectum, 1985. **28**(12): p. 937-40.

214. Brouillette, J.N., E. Pryor, and T.A. Fox, Jr., *Evaluation of sexual dysfunction in the female following rectal resection and intestinal stoma.* Dis Colon Rectum, 1981. **24**(2): p. 96-102.

215. Hartley, A., et al., *Pathological complete response following pre-operative chemoradiotherapy in rectal cancer: analysis of phase II/III trials.* Br J Radiol, 2005. **78**(934): p. 934-8.

216. Erlandsson, J., et al., *Tumour regression after radiotherapy for rectal cancer - Results from the randomised Stockholm III trial.* Radiother Oncol, 2019. **135**: p. 178-186.

217. Bitterman, D.S., et al., *Predictors of Complete Response and Disease Recurrence Following Chemoradiation for Rectal Cancer.* Front Oncol, 2015. **5**: p. 286.

218. Zeng, W.G., et al., *Clinical parameters predicting pathologic complete response following neoadjuvant chemoradiotherapy for rectal cancer.* Chin J Cancer, 2015. **34**(10): p. 468-74.

219. Dayde, D., et al., *Predictive and Prognostic Molecular Biomarkers for Response to Neoadjuvant Chemoradiation in Rectal Cancer.* Int J Mol Sci, 2017. **18**(3): p. 573.

220. Lopez-Crapez, E., et al., *p53 status and response to radiotherapy in rectal cancer: a prospective multilevel analysis.* Br J Cancer, 2005. **92**(12): p. 2114-21.

221. Derbel, O., et al., *Impact of KRAS, BRAF and PI3KCA mutations in rectal carcinomas treated with neoadjuvant radiochemotherapy and surgery.* BMC Cancer, 2013. **13**(1): p. 200.

222. Lee, J.W., et al., *KRAS Mutation Status Is Not a Predictor for Tumor Response and Survival in Rectal Cancer Patients Who Received Preoperative Radiotherapy With 5-Fluoropyrimidine Followed by Curative Surgery.* Medicine (Baltimore), 2015. **94**(31): p. e1284.

223. Wang, Q., et al., *PIK3CA mutations confer resistance to first-line chemotherapy in colorectal cancer.* Cell Death Dis, 2018. **9**(7): p. 739.

224. Chen, M.B., et al., *P53 status as a predictive biomarker for patients receiving neoadjuvant radiation-based treatment: a meta-analysis in rectal cancer.* PLoS One, 2012. **7**(9): p. e45388.

225. Leichsenring, J., A. Koppelle, and A. Reinacher-Schick, *Colorectal Cancer: Personalized Therapy.* Gastrointest Tumors, 2014. **1**(4): p. 209-20.

226. Krishnan, S. and G.J. Chang, *KRAS mutations and rectal cancer response to chemoradiation: are we closer to personalization of therapy?* Ann Surg Oncol, 2013. **20**(11): p. 3359-62.

227. Guinney, J., et al., *The consensus molecular subtypes of colorectal cancer.* Nat Med, 2015. **21**(11): p. 1350-6.

228. Thanki, K., et al., *Consensus Molecular Subtypes of Colorectal Cancer and their Clinical Implications.* Int Biol Biomed J, 2017. **3**(3): p. 105-111.

229. Fontana, E., et al., *Context matters-consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials.* Ann Oncol, 2019. **30**(4): p. 520-527.

230. Lenz, H.J., et al., *Impact of Consensus Molecular Subtype on Survival in Patients With Metastatic Colorectal Cancer: Results From CALGB/SWOG 80405 (Alliance).* J Clin Oncol, 2019. **37**(22): p. 1876-1885.

231. Williamson, J.S., et al., *Review of the development of DNA methylation as a marker of response to neoadjuvant therapy and outcomes in rectal cancer.* Clin Epigenetics, 2015. **7**(1): p. 70.

232. Murcia, O., et al., *Colorectal cancer molecular classification using BRAF, KRAS, microsatellite instability and CIMP status: Prognostic implications and response to chemotherapy.* PLoS One, 2018. **13**(9): p. e0203051.

233. Gaedcke, J., et al., *Identification of a DNA methylation signature to predict disease-free survival in locally advanced rectal cancer.* Oncotarget, 2014. **5**(18): p. 8123-35.

234. Ma, S., et al., *Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis.* Genome Biol, 2018. **19**(1): p. 142.

235. Agostini, M., et al., *A functional biological network centered on XRCC3: a new possible marker of chemoradiotherapy resistance in rectal cancer patients.* Cancer Biol Ther, 2015. **16**(8): p. 1160-71.

236. Cecchin, E., et al., *Tumor response is predicted by patient genetic profile in rectal cancer patients treated with neo-adjuvant chemo-radiotherapy.* Pharmacogenomics J, 2011. **11**(3): p. 214-26.

237. Karagkounis, G., et al., *NPTX2 is associated with neoadjuvant therapy response in rectal cancer.* J Surg Res, 2016. **202**(1): p. 112-7.

238. Agostini, M., et al., *An integrative approach for the identification of prognostic and predictive biomarkers in rectal cancer.* Oncotarget, 2015. **6**(32): p. 32561-74.

239. Fessler, J., V. Matson, and T.F. Gajewski, *Exploring the emerging role of the microbiome in cancer immunotherapy.* J Immunother Cancer, 2019. **7**(1): p. 108.

240. Pouncey, A.L., et al., *Gut microbiota, chemotherapy and the host: the influence of the gut microbiota on cancer treatment.* Ecancermedicalscience, 2018. **12**: p. 868.

241. Kumagai, T., F. Rahman, and A.M. Smith, *The Microbiome and Radiation Induced-Bowel Injury: Evidence for Potential Mechanistic Role in Disease Pathogenesis.* Nutrients, 2018. **10**(10): p. 1405.

242. Zhang, S., et al., *Colorectal cancer, radiotherapy and gut microbiota.* Chin J Cancer Res, 2019. **31**(1): p. 212-222.

243. Ishiguro, L., et al., *Folic Acid Supplementation Adversely Affects Chemosensitivity of Colon Cancer Cells to 5-fluorouracil.* Nutr Cancer, 2016. **68**(5): p. 780-90.

244. Tsukihara, H., K. Tsunekuni, and T. Takechi, *Folic Acid-Metabolizing Enzymes Regulate the Antitumor Effect of 5-Fluoro-2'-Deoxyuridine in Colorectal Cancer Cell Lines.* PLoS One, 2016. **11**(9): p. e0163961.

245. Maynard, C., et al., *A bacterial route for folic acid supplementation.* BMC Biol, 2018. **16**(1): p. 67.

246. Virk, B., et al., *Folate Acts in E. coli to Accelerate C. elegans Aging Independently of Bacterial Biosynthesis.* Cell Rep, 2016. **14**(7): p. 1611-1620.

247. Gao, Y.D., Y. Zhao, and J. Huang, *Metabolic modeling of common Escherichia coli strains in human gut microbiome.* Biomed Res Int, 2014. **2014**: p. 694967.

248. Spanogiannopoulos, P., et al., *The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism.* Nat Rev Microbiol, 2016. **14**(5): p. 273-87.

249. Stein, A., W. Voigt, and K. Jordan, *Chemotherapy-induced diarrhea: pathophysiology, frequency and guideline-based management.* Ther Adv Med Oncol, 2010. **2**(1): p. 51-63.

250. Yilmaz, M.T., A. Elmali, and G. Yazici, *Abscopal Effect, From Myth to Reality: From Radiation Oncologists' Perspective.* Cureus, 2019. **11**(1): p. e3860.

251. Peterson, D.E., et al., *Management of oral and gastrointestinal mucositis: ESMO Clinical Practice Guidelines.* Ann Oncol, 2011. **22 Suppl 6**(Suppl 6): p. vi78-84.

252. van Vliet, M.J., et al., *The role of intestinal microbiota in the development and severity of chemotherapy-induced mucositis.* PLoS Pathog, 2010. **6**(5): p. e1000879.

253. Ong, Z.Y., et al., *Pro-inflammatory cytokines play a key role in the development of radiotherapy-induced gastrointestinal mucositis.* Radiat Oncol, 2010. **5**(1): p. 22.

254. Touchefeu, Y., et al., *Systematic review: the role of the gut microbiota in chemotherapy- or radiation-induced gastrointestinal mucositis - current evidence and potential clinical applications.* Aliment Pharmacol Ther, 2014. **40**(5): p. 409-21.

255. Schmidt, S.V., A.C. Nino-Castro, and J.L. Schultze, *Regulatory dendritic cells: there is more than just immune activation.* Front Immunol, 2012. **3**: p. 274.

256. Merrick, A., et al., *Immunosuppressive effects of radiation on human dendritic cells: reduced IL-12 production on activation and impairment of naive T-cell priming.* Br J Cancer, 2005. **92**(8): p. 1450-8.

257. Wild, J., Z. Hradecna, and W. Szybalski, *Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones.* Genome Res, 2002. **12**(9): p. 1434-44.

258. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA.* Genomics, 2016. **107**(1): p. 1-8.

259. Berglund, E.C., A. Kiialainen, and A.C. Syvanen, *Next-generation sequencing technologies and applications for human genetic history and forensics.* Investig Genet, 2011. **2**: p. 23.

260. Ambardar, S., et al., *High Throughput Sequencing: An Overview of Sequencing Chemistry.* Indian J Microbiol, 2016. **56**(4): p. 394-404.

261. Loman, N.J., J. Quick, and J.T. Simpson, *A complete bacterial genome assembled de novo using only nanopore sequencing data.* Nat Methods, 2015. **12**(8): p. 733-5.

262. Sovic, I., et al., *Fast and sensitive mapping of nanopore sequencing reads with GraphMap.* Nat Commun, 2016. **7**: p. 11307.

263. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.

264. Li, H., *Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.* Bioinformatics, 2016. **32**(14): p. 2103-10.

265. Berlin, K., et al., *Assembling large genomes with single-molecule sequencing and locality-sensitive hashing.* Nat Biotechnol, 2015. **33**(6): p. 623-30.

266. Links, M.G., et al., *The chaperonin-60 universal target is a barcode for bacteria that enables de novo assembly of metagenomic sequence data.* PLoS One, 2012. **7**(11): p. e49755.

267. Fuselli, S., et al., *A new hybrid approach for MHC genotyping: high-throughput NGS and long read MinION nanopore sequencing, with application to the non-model vertebrate Alpine chamois (Rupicapra rupicapra).* Heredity (Edinb), 2018. **121**(4): p. 293-303.

268. Chang, F. and M.M. Li, *Clinical application of amplicon-based next-generation sequencing in cancer.* Cancer Genet, 2013. **206**(12): p. 413-9.

269. Sharpton, T.J., *An introduction to the analysis of shotgun metagenomic data.* Front Plant Sci, 2014. **5**: p. 209.

270. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biol, 2013. **14**(4): p. R36.

271. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

272. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

273. Salzberg, S.L., *Next-generation genome annotation: we still struggle to get it right.* Genome Biol, 2019. **20**(1): p. 92.

274. Martinez-Carranza, E., et al., *Variability of Bacterial Essential Genes Among Closely Related Bacteria: The Case of Escherichia coli.* Front Microbiol, 2018. **9**: p. 1059.

275. Bolotin, E. and R. Hershberg, *Horizontally Acquired Genes Are Often Shared between Closely Related Bacterial Species.* Front Microbiol, 2017. **8**(1536): p. 1536.

276. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

277. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.

278. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

279. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic Acids Res, 2014. **42**(Database issue): p. D459-71.

280. Barry, W.T., A.B. Nobel, and F.A. Wright, *Significance analysis of functional categories in gene expression studies: a structured permutation approach.* Bioinformatics, 2005. **21**(9): p. 1943-9.

281. Balkwill, F.R., M. Capasso, and T. Hagemann, *The tumor microenvironment at a glance.* J Cell Sci, 2012. **125**(Pt 23): p. 5591-6.

282. Yoshihara, K., et al., *Inferring tumour purity and stromal and immune cell admixture from expression data.* Nat Commun, 2013. **4**(1): p. 2612.

283. Aran, D., Z. Hu, and A.J. Butte, *xCell: digitally portraying the tissue cellular heterogeneity landscape.* Genome Biol, 2017. **18**(1): p. 220.

284. Chen, B., et al., *Profiling Tumor Infiltrating Immune Cells with CIBERSORT.* Methods Mol Biol, 2018. **1711**: p. 243-259.

285. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

286. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

287. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nat Methods, 2017. **14**(4): p. 417-419.

288. Vertessy, B.G. and J. Toth, *Keeping uracil out of DNA: physiological role, structure and catalytic mechanism of dUTPases.* Acc Chem Res, 2009. **42**(1): p. 97-106.

289. Belkaid, Y. and T.W. Hand, *Role of the microbiota in immunity and inflammation.* Cell, 2014. **157**(1): p. 121-41.

290. Shi, N., et al., *Interaction between the gut microbiome and mucosal immune system.* Mil Med Res, 2017. **4**(1): p. 14.

291. Petrova, M.I., et al., *Lactobacillus iners: Friend or Foe?* Trends Microbiol, 2017. **25**(3): p. 182-191.

292. Goodrich, J.K., et al., *Human genetics shape the gut microbiome.* Cell, 2014. **159**(4): p. 789-99.

293. Hall, A.B., A.C. Tolonen, and R.J. Xavier, *Human genetic variation and the gut microbiome in disease.* Nat Rev Genet, 2017. **18**(11): p. 690-699.

294. Antwis, R.E., et al., *Fifty important research questions in microbial ecology.* FEMS Microbiol Ecol, 2017. **93**(5).

295. Wang, X. and L. Liotta, *Clinical bioinformatics: a new emerging science.* J Clin Bioinforma, 2011. **1**(1): p. 1.

296. Gevers, D., et al., *Bioinformatics for the Human Microbiome Project.* PLoS Comput Biol, 2012. **8**(11): p. e1002779.

297. Dubos, R.J. and R.W. Schaedler, *The Effect Of The Intestinal Flora On The Growth Rate Of Mice, And On Their Susceptibility To Experimental Infections.* The Journal of Experimental Medicine, 1960. **111**(3): p. 407-417.

298. Schaedler, R.W., R. Dubos, and R. Costello, *The Development of the Bacterial Flora in the Gastrointestinal Tract of Mice.* J Exp Med, 1965. **122**(1): p. 59-66.

299. Alain, K. and J. Querellou, *Cultivating the uncultured: limits, advances and future challenges.* Extremophiles, 2009. **13**(4): p. 583-94.

300. Arnold, J.W., J. Roach, and M.A. Azcarate-Peril, *Emerging Technologies for Gut Microbiome Research.* Trends Microbiol, 2016. **24**(11): p. 887-901.

301. Liu, J., G. Liu, and Z. Li, *Importance of metabolomics analyses of maternal parameters and their influence on fetal growth.* Exp Ther Med, 2017. **14**(1): p. 467-472.

302. Hiraoka, S., C.C. Yang, and W. Iwasaki, *Metagenomics and Bioinformatics in Microbial Ecology: Current Status and Beyond.* Microbes Environ, 2016. **31**(3): p. 204-12.

303. Scholz, M.B., C.C. Lo, and P.S. Chain, *Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis.* Curr Opin Biotechnol, 2012. **23**(1): p. 9-15.

304. Case, R.J., et al., *Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies.* Appl Environ Microbiol, 2007. **73**(1): p. 278-88.

305. Stahl, D.A., et al., *Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences.* Science, 1984. **224**(4647): p. 409-11.

306. Olsen, G.J., et al., *Microbial ecology and evolution: a ribosomal RNA approach.* Annu Rev Microbiol, 1986. **40**: p. 337-65.

307. Schmidt, T.M., E.F. DeLong, and N.R. Pace, *Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing.* J Bacteriol, 1991. **173**(14): p. 4371-8.

308. Amann, R.I., W. Ludwig, and K.H. Schleifer, *Phylogenetic identification and in situ detection of individual microbial cells without cultivation.* Microbiol Rev, 1995. **59**(1): p. 143-69.

309. Giovannoni, S.J., et al., *Genetic diversity in Sargasso Sea bacterioplankton.* Nature, 1990. **345**(6270): p. 60-3.

310. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Appl Environ Microbiol, 2009. **75**(23): p. 7537-41.

311. Edgar, R.C., *UPARSE: highly accurate OTU sequences from microbial amplicon reads.* Nat Methods, 2013. **10**(10): p. 996-8.

312. Callahan, B.J., et al., *DADA2: High-resolution sample inference from Illumina amplicon data.* Nat Methods, 2016. **13**(7): p. 581-3.

313. Kuczynski, J., et al., *Using QIIME to analyze 16S rRNA gene sequences from microbial communities.* Curr Protoc Bioinformatics, 2011. **Chapter 10**: p. Unit 10 7.

314. Caporaso, J.G., et al., *QIIME allows analysis of high-throughput community sequencing data.* Nature Methods, 2010. **7**: p. 335.

315. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2009, 2016.

316. McMurdie, P.J. and S. Holmes, *phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.* PLoS One, 2013. **8**(4): p. e61217.

317. Oksanen J, B.F., Kindt R, Legendre P, Minchin PR, et al. , *Vegan: ecological diversity - R Package. R package version 2.0-10.* 2013.

318. Lu, J. and S.L. Salzberg, *Ultrafast and accurate 16S microbial community analysis using Kraken 2.* bioRxiv, 2020: p. 2020.03.27.012047.

319. Glassman, S.I. and J.B.H. Martiny, *Broadscale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units.* mSphere, 2018. **3**(4): p. e00148-18.

320. Edgar, R., *Taxonomy annotation and guide tree errors in 16S rRNA databases.* PeerJ, 2018. **6**: p. e5030.

321. Breitwieser, F.P., J. Lu, and S.L. Salzberg, *A review of methods and databases for metagenomic classification and assembly.* Brief Bioinform, 2019. **20**(4): p. 1125-1136.

322. DeSantis, T.Z., et al., *Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.* Appl Environ Microbiol, 2006. **72**(7): p. 5069-72.

323. Balvociute, M. and D.H. Huson, *SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare?* BMC Genomics, 2017. **18**(Suppl 2): p. 114.

324. Bolyen, E., et al., *Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2.* Nat Biotechnol, 2019. **37**(8): p. 852-857.

325. Eren, A.M., et al., *Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences.* ISME J, 2015. **9**(4): p. 968-79.

326. Callahan, B.J., P.J. McMurdie, and S.P. Holmes, *Exact sequence variants should replace operational taxonomic units in marker-gene data analysis.* ISME J, 2017. **11**(12): p. 2639-2643.

327. Needham, D.M., R. Sachdeva, and J.A. Fuhrman, *Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters.* ISME J, 2017. **11**(7): p. 1614-1629.

328. Ranjan, R., et al., *Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing.* Biochem Biophys Res Commun, 2016. **469**(4): p. 967-77.

329. Kreznar, J.H., et al., *Host Genotype and Gut Microbiome Modulate Insulin Secretion and Diet-Induced Metabolic Phenotypes.* Cell Rep, 2017. **18**(7): p. 1739-1750.

330. Macke, E., et al., *Host-genotype dependent gut microbiota drives zooplankton tolerance to toxic cyanobacteria.* Nat Commun, 2017. **8**(1): p. 1608.

331. Moran-Ramos, S., B.E. Lopez-Contreras, and S. Canizales-Quinteros, *Gut Microbiota in Obesity and Metabolic Abnormalities: A Matter of Composition or Functionality?* Arch Med Res, 2017. **48**(8): p. 735-753.

332. Sanz, Y., *Microbiome and Gluten.* Ann Nutr Metab, 2015. **67 Suppl 2**(Suppl. 2): p. 28-41.

333. Zhang, P., et al., *Commensal Homeostasis of Gut Microbiota-Host for the Impact of Obesity.* Front Physiol, 2017. **8**(JAN): p. 1122.

334. Vinje, H., et al., *Comparing K-mer based methods for improved classification of 16S sequences.* BMC Bioinformatics, 2015. **16**: p. 205.

335. Ounit, R., et al., *CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers.* BMC Genomics, 2015. **16**(1): p. 236.

336. Wood, D.E. and S.L. Salzberg, *Kraken: ultrafast metagenomic sequence classification using exact alignments.* Genome Biol, 2014. **15**(3): p. R46.

337. Kim, D., et al., *Centrifuge: rapid and sensitive classification of metagenomic sequences.* Genome Res, 2016. **26**(12): p. 1721-1729.

338. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2.* Genome Biol, 2019. **20**(1): p. 257.

339. Mangul, S., et al., *Challenges and recommendations to improve the installability and archival stability of omics computational tools.* PLoS Biol, 2019. **17**(6): p. e3000333.

340. Cottier, F., et al., *Advantages of meta-total RNA sequencing (MeTRS) over shotgun metagenomics and amplicon-based sequencing in the profiling of complex microbial communities.* NPJ Biofilms Microbiomes, 2018. **4**(1): p. 2.

341. Lobb, B., et al., *An assessment of genome annotation coverage across the bacterial tree of life.* Microb Genom, 2020. **6**(3).

342. Seemann, T., *Prokka: rapid prokaryotic genome annotation.* Bioinformatics, 2014. **30**(14): p. 2068-9.

343. Warren, A.S., et al., *Missing genes in the annotation of prokaryotic genomes.* BMC Bioinformatics, 2010. **11**(1): p. 131.

344. Visnovska, T., et al., *Metagenomics and transcriptomics data from human colorectal cancer.* Sci Data, 2019. **6**(1): p. 116.

345. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. 2010; Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

346. Bushnell, B. *BBMap*. 2014; Available from: https://sourceforge.net/projects/bbmap/.

347. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018: p. bty191-bty191.

348. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

349. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

350. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2.* Genome Biology, 2019. **20**(1): p. 257.

351. Breitwieser, F.P. and S.L. Salzberg, *Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification.* bioRxiv, 2016: p. 084715.

352. Dixon, P., *VEGAN, a package of R functions for community ecology.* Journal of Vegetation Science, 2003. **14**(6): p. 927-930.

353. Paradis, E., J. Claude, and K. Strimmer, *APE: Analyses of Phylogenetics and Evolution in R language.* Bioinformatics, 2004. **20**(2): p. 289-90.

354. Srivastava, A., et al., *Alignment and mapping methodology influence transcript abundance estimation.* bioRxiv, 2019: p. 657874.

355. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

356. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

357. Rizopoulos, D., *ltm: An R Package for Latent Variable Modeling and Item Response Analysis.* Journal of Statistical Software; Vol 1, Issue 5 (2007), 2006.

358. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.

359. Singh, A., et al., *DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays.* Bioinformatics, 2019. **35**(17): p. 3055-3062.

360. Stevens, J.R., et al., *Power in pairs: assessing the statistical value of paired samples in tests for differential expression.* BMC Genomics, 2018. **19**(1): p. 953.

361. Schubert, A.M., et al., *Microbiome data distinguish patients with Clostridium difficile infection and non-C. difficile-associated diarrhea from healthy controls.* mBio, 2014. **5**(3): p. e01021-14.

362. Slattery, J., D.F. MacFabe, and R.E. Frye, *The Significance of the Enteric Microbiome on the Development of Childhood Disease: A Review of Prebiotic and Probiotic Therapies in Disorders of Childhood.* Clin Med Insights Pediatr, 2016. **10**: p. 91-107.

363. Lewis, J.D., et al., *Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease.* Cell Host Microbe, 2015. **18**(4): p. 489-500.

364. Malan-Muller, S., et al., *The Gut Microbiome and Mental Health: Implications for Anxiety- and Trauma-Related Disorders.* OMICS, 2018. **22**(2): p. 90-107.

365. Ringel, Y., *The Gut Microbiome in Irritable Bowel Syndrome and Other Functional Bowel Disorders.* Gastroenterol Clin North Am, 2017. **46**(1): p. 91-101.

366. Goodman, B. and H. Gardner, *The microbiome and cancer.* J Pathol, 2018. **244**(5): p. 667-676.

367. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2018.* CA Cancer J Clin, 2018. **68**(1): p. 7-30.

368. Ferlay, J., et al., *Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012.* International Journal of Cancer, 2014. **136**(5): p. E359-E386.

369. de Lecea, M.G.M. and M. Rossbach, *Translational genomics in personalized medicine – scientific challenges en route to clinical practice.* The HUGO Journal, 2012. **6**(1): p. 2.

370. Quail, M.A., et al., *A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.* BMC Genomics, 2012. **13**(1): p. 341.

371. Taylor, W.S., et al., *MinION Sequencing of colorectal cancer tumour microbiomes-A comparison with amplicon-based and RNA-Sequencing.* PLoS One, 2020. **15**(5): p. e0233170.

372. Team, R.C., *R: A Language and Environment for Statistical Computing*. 2020.

373. Oliveros, J.C. *Venny. An interactive tool for comparing lists with Venn's diagrams.* 2015; Available from: https://bioinfogp.cnb.csic.es/tools/venny/index.html

374. Kim, K., et al., *Differences Regarding the Molecular Features and Gut Microbiota Between Right and Left Colon Cancer.* Ann Coloproctol, 2018. **34**(6): p. 280-285.

375. Rinninella, E., et al., *What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases.* Microorganisms, 2019. **7**(1): p. 14.

376. Eribe, E.R.K. and I. Olsen, *Leptotrichia species in human infections II.* J Oral Microbiol, 2017. **9**(1): p. 1368848.

377. Hedberg, M.E., et al., *Lachnoanaerobaculum gen. nov., a new genus in the Lachnospiraceae: characterization of Lachnoanaerobaculum umeaense gen. nov., sp. nov., isolated from the human small intestine, and Lachnoanaerobaculum orale sp. nov., isolated from saliva, and reclassification of Eubacterium saburreum (Prevot 1966) Holdeman and Moore 1970 as Lachnoanaerobaculum saburreum comb. nov.* Int J Syst Evol Microbiol, 2012. **62**(Pt 11): p. 2685-2690.

378. Facciola, A., et al., *Campylobacter: from microbiology to prevention.* J Prev Med Hyg, 2017. **58**(2): p. E79-E92.

379. Li, H., et al., *Alterations of gut microbiota contribute to the progression of unruptured intracranial aneurysms.* Nat Commun, 2020. **11**(1): p. 3218.

380. Engels, C., et al., *The Common Gut Microbe Eubacterium hallii also Contributes to Intestinal Propionate Formation.* Front Microbiol, 2016. **7**(713): p. 713.

381. Parker, B.J., et al., *The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health.* Front Immunol, 2020. **11**: p. 906.

382. Wang, W., et al., *Metagenomic analysis of microbiome in colon tissue from subjects with inflammatory bowel diseases reveals interplay of viruses and bacteria.* Inflamm Bowel Dis, 2015. **21**(6): p. 1419-27.

383. Flemer, B., et al., *Tumour-associated and non-tumour-associated microbiota: Addendum.* Gut Microbes, 2018. **9**(4): p. 369-373.

384. Wommack, K.E., J. Bhavsar, and J. Ravel, *Metagenomics: read length matters.* Appl Environ Microbiol, 2008. **74**(5): p. 1453-63.

385. Pearman, W.S., N.E. Freed, and O.K. Silander, *The advantages and disadvantages of short- and long-read metagenomics to infer bacterial and eukaryotic community composition.* bioRxiv, 2019: p. 650788.

386. Pearman, W.S., N.E. Freed, and O.K. Silander, *Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads.* BMC Bioinformatics, 2020. **21**(1): p. 220.

387. Lu, J. and S.L. Salzberg, *Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2.* Microbiome, 2020. **8**(1): p. 124.

388. Zou, S., L. Fang, and M.H. Lee, *Dysbiosis of gut microbiota in promoting the development of colorectal cancer.* Gastroenterol Rep (Oxf), 2018. **6**(1): p. 1-12.

389. Zhang, S., et al., *Fusobacterium nucleatum promotes chemoresistance to 5-fluorouracil by upregulation of BIRC3 expression in colorectal cancer.* J Exp Clin Cancer Res, 2019. **38**(1): p. 14.

390. Villeger, R., et al., *Microbial markers in colorectal cancer detection and/or prognosis.* World J Gastroenterol, 2018. **24**(22): p. 2327-2347.

391. Yu, T., et al., *Fusobacterium nucleatum Promotes Chemoresistance to Colorectal Cancer by Modulating Autophagy.* Cell, 2017. **170**(3): p. 548-563 e16.

392. Yuan, L., et al., *The influence of gut microbiota dysbiosis to the efficacy of 5-Fluorouracil treatment on colorectal cancer.* Biomed Pharmacother, 2018. **108**: p. 184-193.

393. Manichanh, C., et al., *The gut microbiota predispose to the pathophysiology of acute postradiotherapy diarrhea.* Am J Gastroenterol, 2008. **103**(7): p. 1754-61.

394. Vesty, A., et al., *Oral microbial influences on oral mucositis during radiotherapy treatment of head and neck cancer.* Support Care Cancer, 2020. **28**(6): p. 2683-2691.

395. Wang, A., et al., *Gut microbial dysbiosis may predict diarrhea and fatigue in patients undergoing pelvic cancer radiotherapy: a pilot study.* PLoS One, 2015. **10**(5): p. e0126312.

396. Reis Ferreira, M., et al., *Microbiota- and Radiotherapy-Induced Gastrointestinal Side-Effects (MARS) Study: A Large Pilot Study of the Microbiome in Acute and Late-Radiation Enteropathy.* Clin Cancer Res, 2019. **25**(21): p. 6487-6500.

397. Nam, Y.D., et al., *Impact of pelvic radiotherapy on gut microbiota of gynecological cancer patients revealed by massive pyrosequencing.* PLoS One, 2013. **8**(12): p. e82659.

398. Kim, Y.S., J. Kim, and S.J. Park, *High-throughput 16S rRNA gene sequencing reveals alterations of mouse intestinal microbiota after radiotherapy.* Anaerobe, 2015. **33**: p. 1-7.

399. Sahly, N., et al., *Effect of radiotherapy on the gut microbiome in pediatric cancer patients: a pilot study.* PeerJ, 2019. **7**(9): p. e7683.

400. Uribe-Herranz, M., et al., *Gut microbiota modulate dendritic cell antigen presentation and radiotherapy-induced antitumor immune response.* J Clin Invest, 2020. **130**(1): p. 466-479.

401. Huang, R., J. Xiang, and P. Zhou, *Vitamin D, gut microbiota, and radiation-related resistance: a love-hate triangle.* J Exp Clin Cancer Res, 2019. **38**(1): p. 493.

402. Goodrich, J.K., et al., *Conducting a microbiome study.* Cell, 2014. **158**(2): p. 250-262.

403. Breitwieser, F.P. and S.L. Salzberg, *Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification.* Bioinformatics, 2020. **36**(4): p. 1303-1304.

404. Pollock, J., et al., *The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies.* Appl Environ Microbiol, 2018. **84**(7): p. e02627-17.

405. Baccarella, A., et al., *Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance.* BMC Bioinformatics, 2018. **19**(1): p. 423.

406. Allaband, C., et al., *Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians.* Clin Gastroenterol Hepatol, 2019. **17**(2): p. 218-230.

407. Shin, N.R., T.W. Whon, and J.W. Bae, *Proteobacteria: microbial signature of dysbiosis in gut microbiota.* Trends Biotechnol, 2015. **33**(9): p. 496-503.

408. Bradley, P.H. and K.S. Pollard, *Proteobacteria explain significant functional variability in the human gut microbiome.* Microbiome, 2017. **5**(1): p. 36.

409. Kubickova, B., et al., *Effects of cyanobacterial toxins on the human gastrointestinal tract and the mucosal innate immune system.* Environmental Sciences Europe, 2019. **31**(1): p. 31.

410. Badri, H., et al., *Molecular investigation of the radiation resistance of edible cyanobacterium Arthrospira sp. PCC 8005.* Microbiologyopen, 2015. **4**(2): p. 187-207.

411. Schnee, A.E. and W.A. Petri, Jr., *Campylobacter jejuni and associated immune mechanisms: short-term effects and long-term implications for infants in low-income countries.* Curr Opin Infect Dis, 2017. **30**(3): p. 322-328.

412. Xia, X., et al., *Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer.* Microbiome, 2020. **8**(1): p. 108.

413. Delday, M., et al., *Bacteroides thetaiotaomicron Ameliorates Colon Inflammation in Preclinical Models of Crohn's Disease.* Inflamm Bowel Dis, 2019. **25**(1): p. 85-96.

414. Ulger Toprak, N., et al., *Butyricimonas virosa: the first clinical case of bacteraemia.* New Microbes New Infect, 2015. **4**: p. 7-8.

415. Litvak, Y., M.X. Byndloss, and A.J. Baumler, *Colonocyte metabolism shapes the gut microbiota.* Science, 2018. **362**(6418): p. eaat9076.

416. Ai, D., et al., *Identifying Gut Microbiota Associated With Colorectal Cancer Using a Zero-Inflated Lognormal Model.* Front Microbiol, 2019. **10**(826): p. 826.

417. Anand, S., H. Kaur, and S.S. Mande, *Comparative In silico Analysis of Butyrate Production Pathways in Gut Commensals and Pathogens.* Front Microbiol, 2016. **7**(1945): p. 1945.

418. Ogita, T., et al., *Oral Administration of Flavonifractor plautii Strongly Suppresses Th2 Immune Responses in Mice.* Front Immunol, 2020. **11**(379): p. 379.

419. Parker, B.J., et al., *The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health.* Front Immunol, 2020. **11**(906): p. 906.

420. Chambers, E.S., et al., *Effects of targeted delivery of propionate to the human colon on appetite regulation, body weight maintenance and adiposity in overweight adults.* Gut, 2015. **64**(11): p. 1744-54.

421. Radka, C.D., et al., *Fatty acid activation and utilization by Alistipes finegoldii, a representative Bacteroidetes resident of the human gut microbiome.* Mol Microbiol, 2020. **113**(4): p. 807-825.

422. Lam, Y.Y., et al., *Increased gut permeability and microbiota change associate with mesenteric fat inflammation and metabolic dysfunction in diet-induced obese mice.* PLoS One, 2012. **7**(3): p. e34233.

423. Klement, R.J., G. Schafer, and R.A. Sweeney, *A ketogenic diet exerts beneficial effects on body composition of cancer patients during radiotherapy: An interim analysis of the KETOCOMP study.* J Tradit Complement Med, 2020. **10**(3): p. 180-187.

424. Breitkreutz, R., et al., *Effects of a high-fat diet on body composition in cancer patients receiving chemotherapy: a randomized controlled study.* Wien Klin Wochenschr, 2005. **117**(19-20): p. 685-92.

425. Tan-Shalaby, J., *Ketogenic Diets and Cancer: Emerging Evidence.* Fed Pract, 2017. **34**(Suppl 1): p. 37S-42S.

426. Weber, D.D., et al., *Ketogenic diet in the treatment of cancer - Where do we stand?* Mol Metab, 2020. **33**: p. 102-121.

427. Yoshida, N., et al., *Bacteroides vulgatus and Bacteroides dorei Reduce Gut Microbial Lipopolysaccharide Production and Inhibit Atherosclerosis.* Circulation, 2018. **138**(22): p. 2486-2498.

428. Goldstein, E.J. and D.R. Snydman, *Intra-abdominal infections: review of the bacteriology, antimicrobial susceptibility and the role of ertapenem in their therapy.* J Antimicrob Chemother, 2004. **53 Suppl 2**: p. ii29-36.

429. Wexler, H.M., *Bacteroides: the good, the bad, and the nitty-gritty.* Clin Microbiol Rev, 2007. **20**(4): p. 593-621.

430. Purcell, R.V., et al., *Colonization with enterotoxigenic Bacteroides fragilis is associated with early-stage colorectal neoplasia.* PLoS One, 2017. **12**(2): p. e0171602.

431. Deng, H., et al., *A novel strain of Bacteroides fragilis enhances phagocytosis and polarises M1 macrophages.* Sci Rep, 2016. **6**(1): p. 29401.

432. Andino, A. and I. Hanning, *Salmonella enterica: survival, colonization, and virulence differences among serovars.* ScientificWorldJournal, 2015. **2015**: p. 520179.

433. Mi, Z., et al., *Salmonella-Mediated Cancer Therapy: An Innovative Therapeutic Strategy.* J Cancer, 2019. **10**(20): p. 4765-4776.

434. Jazeela, K., et al., *Nontyphoidal Salmonella: a potential anticancer agent.* J Appl Microbiol, 2020. **128**(1): p. 2-14.

435. Gibson, B., et al., *The distribution of bacterial doubling times in the wild.* Proc Biol Sci, 2018. **285**(1880): p. 20180789.

436. Cerra, J., et al., *Complete Genome Sequence of Pseudomonas sp. Strain NC02, Isolated from Soil.* Genome Announc, 2018. **6**(7): p. e00033-18.

437. Goker, M., et al., *Complete genome sequence of Odoribacter splanchnicus type strain (1651/6).* Stand Genomic Sci, 2011. **4**(2): p. 200-9.

438. Gomez-Arango, L.F., et al., *Increased Systolic and Diastolic Blood Pressure Is Associated With Altered Gut Microbiota Composition and Butyrate Production in Early Pregnancy.* Hypertension, 2016. **68**(4): p. 974-81.

439. Carvalho, H.A. and R.C. Villar, *Radiotherapy and immune response: the systemic effects of a local treatment.* Clinics (Sao Paulo), 2018. **73**(suppl 1): p. e557s.

440. Liu, M. and F. Guo, *Recent updates on cancer immunotherapy.* Precis Clin Med, 2018. **1**(2): p. 65-74.

441. Hernandez, C., P. Huebener, and R.F. Schwabe, *Damage-associated molecular patterns in cancer: a double-edged sword.* Oncogene, 2016. **35**(46): p. 5931-5941.

442. Genard, G., S. Lucas, and C. Michiels, *Reprogramming of Tumor-Associated Macrophages with Anticancer Therapies: Radiotherapy versus Chemo- and Immunotherapies.* Front Immunol, 2017. **8**: p. 828.

443. Nurieva, R., J. Wang, and A. Sahoo, *T-cell tolerance in cancer.* Immunotherapy, 2013. **5**(5): p. 513-531.

444. Brown, J.M., L. Recht, and S. Strober, *The Promise of Targeting Macrophages in Cancer Therapy.* Clin Cancer Res, 2017. **23**(13): p. 3241-3250.

445. Sinha, P., V.K. Clements, and S. Ostrand-Rosenberg, *Reduction of myeloid-derived suppressor cells and induction of M1 macrophages facilitate the rejection of established metastatic disease.* J Immunol, 2005. **174**(2): p. 636-45.

446. Ley, K., *M1 Means Kill; M2 Means Heal.* J Immunol, 2017. **199**(7): p. 2191-2193.

447. Udagawa, N., et al., *Origin of osteoclasts: mature monocytes and macrophages are capable of differentiating into osteoclasts under a suitable microenvironment prepared by bone marrow-derived stromal cells.* Proc Natl Acad Sci U S A, 1990. **87**(18): p. 7260-4.

448. Kapinas, K., et al., *Bone matrix osteonectin limits prostate cancer cell growth and survival.* Matrix Biol, 2012. **31**(5): p. 299-307.

449. Huang, R., et al., *RANKL-induced M1 macrophages are involved in bone formation.* Bone Res, 2017. **5**(1): p. 17019.

450. Meziani, L., E. Deutsch, and M. Mondini, *Macrophages in radiation injury: a new therapeutic target.* Oncoimmunology, 2018. **7**(10): p. e1494488.

451. Lo Presti, A., et al., *Fecal and Mucosal Microbiota Profiling in Irritable Bowel Syndrome and Inflammatory Bowel Disease.* Front Microbiol, 2019. **10**: p. 1655.

452. Kiernan, M.G., et al., *Systemic Molecular Mediators of Inflammation Differentiate Between Crohn's Disease and Ulcerative Colitis, Implicating Threshold Levels of IL-10 and Relative Ratios of Pro-inflammatory Cytokines in Therapy.* J Crohns Colitis, 2020. **14**(1): p. 118-129.

453. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing.* J R Stat Soc, 1995. **57**(1): p. 289-300.

454. Hameed, A., *Human Immunity Against Campylobacter Infection.* Immune Netw, 2019. **19**(6): p. e38.

455. Moosova, Z., et al., *Immunomodulatory effects of cyanobacterial toxin cylindrospermopsin on innate immune cells.* Chemosphere, 2019. **226**: p. 439-446.

456. Figueiredo, M.C., et al., *Hybrid cluster proteins and flavodiiron proteins afford protection to Desulfovibrio vulgaris upon macrophage infection.* J Bacteriol, 2013. **195**(11): p. 2684-90.

457. L. Weglarz, B.P., A. Mertas, Z. Kondera-Anasz, M. Jaworska-Kik, Z. Dzierzewicz, L. Swiatkowska, *Effect of Endotoxins Isolated from <i>Desulfovibrio desulfuricans</i> Soil and Intestinal Strain on the Secretion of TNF-α by Human Mononuclear Cells.* Polish Journal of Environmental Studies, 2006. **15**(4): p. 615-622.

458. Cavalcante-Silva, L.H., et al., *Obesity-Driven Gut Microbiota Inflammatory Pathways to Metabolic Syndrome.* Front Physiol, 2015. **6**(341): p. 341.

459. Yu, J.C., et al., *Whole Body Vibration-Induced Omental Macrophage Polarization and Fecal Microbiome Modification in a Murine Model.* Int J Mol Sci, 2019. **20**(13): p. 3125.

460. Bobryshev, Y.V., et al., *Macrophages and Their Role in Atherosclerosis: Pathophysiology and Transcriptome Analysis.* Biomed Res Int, 2016. **2016**: p. 9582430.

461. Lam, R.S., et al., *Unprimed, M1 and M2 Macrophages Differentially Interact with Porphyromonas gingivalis.* PLoS One, 2016. **11**(7): p. e0158629.

462. Yu, S., et al., *Porphyromonas gingivalis inhibits M2 activation of macrophages by suppressing alpha-ketoglutarate production in mice.* Mol Oral Microbiol, 2018. **33**(5): p. 388-395.

463. Holden, J.A., et al., *Porphyromonas gingivalis lipopolysaccharide weakly activates M1 and M2 polarized mouse macrophages but induces inflammatory cytokines.* Infect Immun, 2014. **82**(10): p. 4190-203.

464. Chakraborty, S., et al., *Pasteurella multocida Toxin Triggers RANKL-Independent Osteoclastogenesis.* Front Immunol, 2017. **8**(185): p. 185.

465. Krystel-Whittemore, M., K.N. Dileepan, and J.G. Wood, *Mast Cell: A Multi-Functional Master Cell.* Front Immunol, 2015. **6**(620): p. 620.

466. Blatner, N.R., et al., *In colorectal cancer mast cells contribute to systemic regulatory T-cell dysfunction.* Proc Natl Acad Sci U S A, 2010. **107**(14): p. 6430-5.

467. Kasakura, K., et al., *Commensal bacteria directly suppress in vitro degranulation of mast cells in a MyD88-independent manner.* Biosci Biotechnol Biochem, 2014. **78**(10): p. 1669-76.

468. Brzezinska-Blaszczyk, E. and A.K. Olejnik, *Intestinal mucosa-associated bacteria modulate rat mast cell reactivity.* Int J Immunopathol Pharmacol, 1999. **12**(1): p. 31-6.

469. Davis-Richardson, A.G., et al., *Bacteroides dorei dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes.* Front Microbiol, 2014. **5**: p. 678.

470. Pedersen, R.M., E.S. Marmolin, and U.S. Justesen, *Species differentiation of Bacteroides dorei from Bacteroides vulgatus and Bacteroides ovatus from Bacteroides xylanisolvens - back to basics.* Anaerobe, 2013. **24**: p. 1-3.

471. Kim, J.H., et al., *Extracellular vesicle-derived protein from Bifidobacterium longum alleviates food allergy through mast cell suppression.* J Allergy Clin Immunol, 2016. **137**(2): p. 507-516 e8.

472. Troy, E.B. and D.L. Kasper, *Beneficial effects of Bacteroides fragilis polysaccharides on the immune system.* Front Biosci (Landmark Ed), 2010. **15**(1): p. 25-34.

473. Lee, Y., et al., *Therapeutic effects of ablative radiation on local tumor require CD8+ T cells: changing strategies for cancer treatment.* Blood, 2009. **114**(3): p. 589-95.

474. Ramakrishna, C., et al., *Bacteroides fragilis polysaccharide A induces IL-10 secreting B and T cells that prevent viral encephalitis.* Nat Commun, 2019. **10**(1): p. 2153.

475. Guo, F.F. and J.W. Cui, *The Role of Tumor-Infiltrating B Cells in Tumor Immunity.* J Oncol, 2019. **2019**: p. 2592419.

476. Waidhauser, J., et al., *Chemotherapy markedly reduces B cells but not T cells and NK cells in patients with cancer.* Cancer Immunol Immunother, 2020. **69**(1): p. 147-157.

477. de Charette, M., A. Marabelle, and R. Houot, *Turning tumour cells into antigen presenting cells: The next step to improve cancer immunotherapy?* Eur J Cancer, 2016. **68**: p. 134-147.

478. Karcher, N., et al., *Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations.* Genome Biol, 2020. **21**(1): p. 138.

479. Scanvion, Q., et al., *Moderate-to-severe eosinophilia induced by treatment with immune checkpoint inhibitors: 37 cases from a national reference center for hypereosinophilic syndromes and the French pharmacovigilance database.* Oncoimmunology, 2020. **9**(1): p. 1722022.

480. Varricchi, G., et al., *Eosinophils: The unsung heroes in cancer?* Oncoimmunology, 2018. **7**(2): p. e1393134.

481. Vaios, E.J., et al., *Eosinophil and lymphocyte counts predict bevacizumab response and survival in recurrent glioblastoma.* Neurooncol Adv, 2020. **2**(1): p. vdaa031.

482. Grisaru-Tal, S., et al., *A new dawn for eosinophils in the tumour microenvironment.* Nat Rev Cancer, 2020. **20**(10): p. 594-607.

483. Takemura, N., et al., *Eosinophil depletion suppresses radiation-induced small intestinal fibrosis.* Sci Transl Med, 2018. **10**(429): p. eaan0333.

484. Chen, J., et al., *Immunomodulation of NK Cells by Ionizing Radiation.* Front Oncol, 2020. **10**(874): p. 874.

485. Gori, A., et al., *Specific prebiotics modulate gut microbiota and immune activation in HAART-naive HIV-infected adults: results of the "COPA" pilot randomized trial.* Mucosal Immunol, 2011. **4**(5): p. 554-63.

486. Crotty, S., *T follicular helper cell differentiation, function, and roles in disease.* Immunity, 2014. **41**(4): p. 529-42.

487. Singh, D., et al., *CD4+ follicular helper-like T cells are key players in anti-tumor immunity*. 2020, bioRxiv.

488. Nurieva, R.I., et al., *Function of T follicular helper cells in anti-tumor immunity.* The Journal of Immunology, 2019. **202**(1 Supplement): p. 138.18.

489. Perruzza, L., et al., *T Follicular Helper Cells Promote a Beneficial Gut Ecosystem for Host Metabolic Homeostasis by Sensing Microbiota-Derived Extracellular ATP.* Cell Rep, 2017. **18**(11): p. 2566-2575.

490. Nurieva, R.I. and Y. Chung, *Understanding the development and function of T follicular helper cells.* Cell Mol Immunol, 2010. **7**(3): p. 190-7.

491. Wu, H., et al., *Stat3 Is Important for Follicular Regulatory T Cell Differentiation.* PLoS One, 2016. **11**(5): p. e0155040.

492. Zou, S., et al., *Targeting STAT3 in Cancer Immunotherapy.* Mol Cancer, 2020. **19**(1): p. 145.

493. Huynh, J., et al., *Therapeutically exploiting STAT3 activity in cancer - using tissue repair as a road map.* Nat Rev Cancer, 2019. **19**(2): p. 82-96.

494. Musteanu, M., et al., *Stat3 is a negative regulator of intestinal tumor progression in Apc(Min) mice.* Gastroenterology, 2010. **138**(3): p. 1003-11 e1-5.

495. Li, Y., et al., *Gut microbiota accelerate tumor growth via c-jun and STAT3 phosphorylation in APCMin/+ mice.* Carcinogenesis, 2012. **33**(6): p. 1231-8.

496. Radford, K.J., et al., *Recombinant E. coli efficiently delivers antigen and maturation signals to human dendritic cells: presentation of MART1 to CD8+ T cells.* Int J Cancer, 2003. **105**(6): p. 811-9.

497. Radford, K.J., et al., *A recombinant E. coli vaccine to promote MHC class I-dependent antigen presentation: application to cancer immunotherapy.* Gene Ther, 2002. **9**(21): p. 1455-63.

498. Maerz, J.K., et al., *Bacterial Immunogenicity Is Critical for the Induction of Regulatory B Cells in Suppressing Inflammatory Immune Responses.* Front Immunol, 2019. **10**(3093): p. 3093.

499. Wylie, B., et al., *Dendritic Cells and Cancer: From Biology to Therapeutic Intervention.* Cancers (Basel), 2019. **11**(4): p. 521.

500. Wisdom, A.J., et al., *Neutrophils promote tumor resistance to radiation therapy.* Proc Natl Acad Sci U S A, 2019. **116**(37): p. 18584-18589.

501. Schernberg, A., et al., *Neutrophils, a candidate biomarker and target for radiation therapy?* Acta Oncol, 2017. **56**(11): p. 1522-1530.

502. Jaillon, S., et al., *Neutrophil diversity and plasticity in tumour progression and therapy.* Nat Rev Cancer, 2020. **20**(9): p. 485-503.

503. de Oliveira, S., E.E. Rosowski, and A. Huttenlocher, *Neutrophil migration in infection and wound repair: going forward in reverse.* Nat Rev Immunol, 2016. **16**(6): p. 378-91.

504. Wilgus, T.A., S. Roy, and J.C. McDaniel, *Neutrophils and Wound Repair: Positive Actions and Negative Reactions.* Adv Wound Care (New Rochelle), 2013. **2**(7): p. 379-388.

505. Watson, R.W., et al., *Neutrophils undergo apoptosis following ingestion of Escherichia coli.* J Immunol, 1996. **156**(10): p. 3986-92.

506. Russo, T.A., et al., *E. coli virulence factor hemolysin induces neutrophil apoptosis and necrosis/lysis in vitro and necrosis/lysis and lung injury in a rat pneumonia model.* Am J Physiol Lung Cell Mol Physiol, 2005. **289**(2): p. L207-16.

507. Becker, K.A., et al., *Neutrophils Kill Reactive Oxygen Species-Resistant Pseudomonas aeruginosa by Sphingosine.* Cell Physiol Biochem, 2017. **43**(4): p. 1603-1616.

508. Parks, Q.M., et al., *Neutrophil enhancement of Pseudomonas aeruginosa biofilm development: human F-actin and DNA as targets for therapy.* J Med Microbiol, 2009. **58**(Pt 4): p. 492-502.

509. Westerman, T.L., et al., *The Salmonella type-3 secretion system-1 and flagellar motility influence the neutrophil respiratory burst.* PLoS One, 2018. **13**(9): p. e0203698.

510. Takehara, M., et al., *Clostridium perfringens alpha-toxin impairs granulocyte colony-stimulating factor receptor-mediated granulocyte production while triggering septic shock.* Commun Biol, 2019. **2**(1): p. 45.

511. Moschen, A.R., et al., *Lipocalin 2 Protects from Inflammation and Tumorigenesis Associated with Gut Microbiota Alterations.* Cell Host Microbe, 2016. **19**(4): p. 455-69.

512. Tacchini-Cottier, F., et al., *An immunomodulatory function for neutrophils during the induction of a CD4+ Th2 response in BALB/c mice infected with Leishmania major.* J Immunol, 2000. **165**(5): p. 2628-36.

513. Fischer, R., et al., *Th1 and Th2 cells are required for both eosinophil- and neutrophil-associated airway inflammatory responses in mice.* Biochem Biophys Res Commun, 2007. **357**(1): p. 44-9.

514. Hiippala, K., et al., *Novel Odoribacter splanchnicus Strain and Its Outer Membrane Vesicles Exert Immunoregulatory Effects in vitro.* Front Microbiol, 2020. **11**(2906): p. 575455.

515. Richards, H., et al., *Novel role of regulatory T cells in limiting early neutrophil responses in skin.* Immunology, 2010. **131**(4): p. 583-92.

516. Finotello, F. and Z. Trajanoski, *Quantifying tumor-infiltrating immune cells from transcriptomics data.* Cancer Immunol Immunother, 2018. **67**(7): p. 1031-1040.

517. Hafner, M.F. and J. Debus, *Radiotherapy for Colorectal Cancer: Current Standards and Future Perspectives.* Visc Med, 2016. **32**(3): p. 172-7.

518. Kurtul, N., et al., *SPARC: As a prognostic biomarker in rectal cancer patients treated with chemo-radiotherapy.* Cancer Biomark, 2017. **18**(4): p. 459-466.

519. Zhu, Y., et al., *Identification of biomarker microRNAs for predicting the response of colorectal cancer to neoadjuvant chemoradiotherapy based on microRNA regulatory network.* Oncotarget, 2017. **8**(2): p. 2233-2248.

520. Chang, H., et al., *CCR6 Is a Predicting Biomarker of Radiosensitivity and Potential Target of Radiosensitization in Rectal Cancer.* Cancer Res Treat, 2018. **50**(4): p. 1203-1213.

521. Ree, A.H., et al., *Biomarkers of histone deacetylase inhibitor activity in a phase 1 combined-modality study with radiotherapy.* PLoS One, 2014. **9**(2): p. e89750.

522. Wu, H.-J., et al., *Predictive biomarkers for the efficacy of concurrent chemoradiotherapy for patients with colorectal cancer.* Biomarkers and Genomic Medicine, 2014. **6**(4): p. 163-166.

523. Lim, S.H., et al., *Predictive and prognostic biomarkers for neoadjuvant chemoradiotherapy in locally advanced rectal cancer.* Crit Rev Oncol Hematol, 2015. **96**(1): p. 67-80.

524. Kim, I.J., et al., *Microarray gene expression profiling for predicting complete response to preoperative chemoradiotherapy in patients with advanced rectal cancer.* Dis Colon Rectum, 2007. **50**(9): p. 1342-53.

525. Rohart, F., et al., *mixOmics: An R package for 'omics feature selection and multiple data integration.* PLoS Comput Biol, 2017. **13**(11): p. e1005752.

526. Egozcue, J.J., *Isometric Logratio Transformations for Compositional Data Analysis.* Mathematical Geology, 2003. **35**(3): p. 279-300.

527. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

528. Le Cao, K.A., et al., *A sparse PLS for variable selection when integrating omics data.* Stat Appl Genet Mol Biol, 2008. **7**(1): p. Article 35.

529. Aindelis, G. and K. Chlichlia, *Modulation of Anti-Tumour Immune Responses by Probiotic Bacteria.* Vaccines (Basel), 2020. **8**(2): p. 329.

530. Shi, T., G. Gao, and Y. Cao, *Long Noncoding RNAs as Novel Biomarkers Have a Promising Future in Cancer Diagnostics.* Dis Markers, 2016. **2016**: p. 9085195.

531. Furness, J.B., *The enteric nervous system and neurogastroenterology.* Nat Rev Gastroenterol Hepatol, 2012. **9**(5): p. 286-94.

532. Delanian, S., J.L. Lefaix, and P.F. Pradat, *Radiation-induced neuropathy in cancer survivors.* Radiother Oncol, 2012. **105**(3): p. 273-82.

533. Payne, S.C., J.B. Furness, and M.J. Stebbing, *Bioelectric neuromodulation for gastrointestinal disorders: effectiveness and mechanisms.* Nat Rev Gastroenterol Hepatol, 2019. **16**(2): p. 89-105.

534. Krisko, A. and M. Radman, *Biology of extreme radiation resistance: the way of Deinococcus radiodurans.* Cold Spring Harb Perspect Biol, 2013. **5**(7): p. a012765-a012765.

535. Guo, Y., et al., *Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms.* BMC Bioinformatics, 2010. **11**(1): p. 447.