The microbiome and bowel cancer

Caroline Anne Young

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Medicine

September 2019

Intellectual Property Statement

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

The author would like to thank Professor Phil Quirke (University of Leeds), Dr Henry Wood (University of Leeds) and Professor Eva Morris (University of Leeds) for their supervision, support and guidance; Professor Quirke's research team for support, advice and assistance with laboratory techniques; and Professor Jenny Barrett (University of Leeds) for statistical advice.

National collaborators include the team at the NHS Bowel Cancer Screening Programme Southern Hub (in particular Dr Sally Benton, Mrs Carole Burtonwood and Mr Martin Brealey), and the Office for Data Release (Public Health England). International collaborators include Professor Ramakrishnan and Dr Mayil Bose (Cancer Institute (WIA), Chennai, India), Dr Pham Van Nang and Dr Mai Van Doi (Can Tho University of Medicine and Pharmacy, Vietnam), Dr Carlos Vaccaro and Dr Tamara Piñero (Italian Hospital, Buenos Aries, Argentina), Dr Luis Contreras Melendez and Mr Camilo Tapia Valladares (Universidad de los Andes, Santiago, Chile), and the CRUK Grand Challenge team 'Optimisticc'.

The author is sincerely grateful to the National Institute for Health Research (NIHR) for sponsoring her Academic Clinical Fellowship; the Wellcome Trust for sponsoring this Clinical Research Training Fellowship; the Academy of Medical Sciences for funding the study 'Global Challenges Research Fund Networking Grant: Large bowel microbiome disease network. Creation of a proof of principle exemplar in colorectal cancer across three continents'; the Cancer Research UK (CRUK) Leeds Centre for funding a 'Future Leader Award'; the University of Leeds for funding an 'International Research Collaboration Award to visit members of the International Network for Cancer Screening Evaluation'; Health Education England Genomics Education Programme for funding to undertake three modules of the University of Manchester's 'Genomic medicine' masters course; the Pathological Society of Great Britain & Ireland, the European Society of Pathology and the German Society of Pathology for sponsorship to attend their respective Junior Pathology Academies; and the Pathological Society of Great Britain & Ireland for funding a 'Visiting Fellowship' to the Meyerson and Huttenhower Laboratories to learn how to perform bioinformatic and statistical analysis of complex microbiome datasets.

The work presented in this thesis is based on the results of 1792 samples. Samples were processed either entirely by the author or jointly with the assistance of laboratory technicians, who were trained by the author in the relevant laboratory techniques. The author also acknowledges the author's supervisor Dr Henry Wood, who handled much of the bioinformatic processing of data, allowing the author to concentrate on planning the data analysis and interpretation of the results.

Abstract

Colorectal cancer (CRC) is the second commonest cause of UK cancer-related deaths and the fourth commonest cause of global cancer-related deaths, with a rising incidence in non-Western countries. Research has demonstrated a CRC-associated microbiome, and the potential for microbiome testing to improve CRC screening accuracy. Currently the majority of studies analyse the microbiome from whole stool, transported and stored refrigerated/frozen; this limits study size and restricts research to Western countries with cold-chain facilities.

This thesis investigated the potential to use guaiac faecal occult blood test (gFOBT) cards or faecal immunochemical test (FIT) samples to collect faeces for 16S ribosomal ribonucleic acid (16SrRNA) analysis. Screening potential was assessed by analysing the microbiome of gFOBT samples collected routinely by the NHS Bowel Cancer Screening Programme (NHSBCSP). The microbiome of non-Western countries was investigated by analysing gFOBT samples collected from healthy volunteers and CRC patients in Argentina, Chile, India and Vietnam.

The microbiome was successfully analysed from processed NHSBCSP gFOBT samples stored for prolonged periods at room temperature and FIT samples for which NHSBCSP conditions were simulated. CRC-associated taxa demonstrated minimal temporal variation, but *Escherichia-Shigella* demonstrated marked variation, negating its potential as a screening biomarker. Microbiome-based models improved screening accuracy; combining gFOBT and microbiome results produced areas under the receiver operating characteristic curve of 0.855 (95% confidence interval (CI): 0.832-0.877) for the detection of CRC and 0.868 (95% CI: 0.848-0.886) for the detection of CRC/adenoma. The microbiome of gFOBT samples stored and transported from abroad at ambient temperature was stable. The combined non-Western CRC-associated microbiome contained CRC-associated bacteria described in Western populations, suggesting that certain taxa may be universally associated with CRC.

The results confirm that gFOBT is suitable for conducting large-scale national and global microbiome research. Clinical application is demonstrated with the

development of novel microbiome-based CRC screening models with improved accuracy.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Full term
AEEC	attaching and effacing *Escherichia coli*
AMER1	APC membrane recruitment protein 1 gene
ASF	Altered Schaedler's Flora
ATM	Ataxia-Telangiectasia mutated gene
AUC	area under the receiver operating characteristic curve
BCSP	Bowel Cancer Screening Programme
BFT	*Bacteroides fragilis* toxin
BIRC3	baculoviral IAP repeat containing 3 gene
BKV	BK virus
BMI	body mass index
bp	base pair
BRAF	BRAF gene
CEA	carcinoembryonic antigen
CHD7/8	chromodomain-helicase-DNA-binding protein 7/8 genes
CI	confidence interval
CIMP	CpG island methylator phenotype
CMV	cytomegalovirus
COX-2	cyclo-oxygenase-2
CRC	colorectal cancer
CRUK	Cancer Research UK
DNA	deoxyribonucleic acid
EBV	Epstein-Barr virus
E. coli	*Escherichia coli*
EDTA	ethylenediaminetetraacetic acid
ELISA	enzyme-linked immunosorbent assay

Abbreviation	Full term
EMP	Earth Microbiome Project
EMT	epithelial-mesenchymal transition
ETBF	enterotoxigenic *Bacteroides fragilis*
FadA	*Fusobacterium* adhesin A
FAP	Familial adenomatous polyposis
FFPE	formalin fixed paraffin embedded
FISH	fluorescent in situ hybridisation
FIT	faecal immunochemical test
FMT	faecal microbiome transplant
F. nucleatum	*Fusobacterium nucleatum*
FOBT	faecal occult blood test
FTA	Flinders Technology Associates
g	gram
g	relative centrifugal force
Gal-Gal-NAc	D-galactose-β(1–3)-N-acetyl-D-galactosamine
gFOBT	guaiac faecal occult blood test
HGD	high grade dysplasia
hMLH1	human mutL homolog 1
HPV	Human Papillomavirus
HSV	herpes simplex virus
IATA	International Air Transport Association
IBD	Inflammatory Bowel Disease
IBS	Irritable Bowel Syndrome
Ig	immunoglobulin
IHC	immunohistochemistry
IRAS	Integrated Research Application System
KRAS	Kirsten rat sarcoma viral oncogene homolog
LEfSe	Linear discriminant analysis Effect Size

Abbreviation	Full term
M	molar
MAPK	mitogen-activated protein kinase
mg	milligram
miR	microRNA gene
ml	millilitre
mM	millimolar
MMR	mismatch repair
MSI	microsatellite instability
MYD88	myeloid differentiation primary response gene 88
NF-kβ	nuclear factor kappa-β subunit
ng	nanogram
NGS	Next Generation Sequencing
NHSBCSP	NHS Bowel Cancer Screening Programme
NIHR	National Institute for Health Research
NK cells	natural killer cells
ODR	Office for Data Release
OOB	out of bag
OTU	operational taxonomic unit
Pa	pascal
PAK1	protein-activated kinase 1
PCA	principal co-ordinate analysis
PCR	polymerase chain reaction
PDXs	patient-derived xenografts
PHE	Public Health England
PICRUSt	Phylogenetic Investigation of Communities by Reconstruction of Unobserved States
pks	polyketide synthases
qPCR	quantitative polymerase chain reaction

Abbreviation	Full term
RCTs	randomised control trials
REC	research ethics committee
ROC	receiver operating characteristic curve
rpm	revolutions per minute
SCFAs	short chain fatty acids
SD	standard deviation
SEED	Secure Electronic Environment for Data
S. gallolyticus	*Streptococcus gallolyticus*
SNP	single nucleotide polymorphism
SPF	Specific-Pathogen-free
TGF-β	transforming growth factor beta
TILs	tumour-infiltrating lymphocytes
TLR	toll-like receptor
TP53	TP53 gene
µl	microlitre
UPEC	uropathogenic *Escherichia coli*
UV	ultraviolet
V	volts
VOCs	volatile organic compounds
V region	hypervariable region of the 16SrRNA gene
16SrRNA	16S ribosomal ribonucleic acid gene
5-FU	5-Fluorouracil

Summary of Chapters

Chapter 1 introduces the microbiome and current understanding of its role in health and disease. Evidence for the CRC-associated microbiome is reviewed and the potential clinical implications of this field of research are discussed. How the microbiome is studied and limitations of existing approaches are outlined. This chapter identifies that there is an urgent need to conduct standardised, large-scale microbiome studies in populations of interest, including non-Western populations.

Chapter 2 proposes harnessing the infrastructure of the NHSBCSP to conduct large-scale microbiome research, by analysing the microbiome directly from processed bowel cancer screening samples. This chapter demonstrates the feasibility of this method using NHSBCSP gFOBT samples and the potential to perform microbiome analysis directly from FIT, which the NHSBCSP is currently adopting.

Chapter 3 applies this method to the analysis of 1283 NHSBCSP gFOBT samples, exploring differences in the microbiome between different clinical groups. Random Forest models which use microbiome data are shown to improve the accuracy of screening.

Chapter 4 investigates whether microbiome analysis performed using gFOBT samples could be applied to non-Western countries (Argentina, Chile, India and Vietnam); this is confirmed. Differences in the microbiome between countries are demonstrated. CRC-associated bacteria traditionally described in Western populations are found to also be enriched in CRC patients from these four non-Western countries.

Chapter 5 discusses the implications of this work and plans for future development.

Chapter 1
Introduction

This chapter introduces the colorectal microbiome, its contribution to health and the evidence for an association between an altered microbiome and CRC. Current methods of investigating the microbiome are described, including their limitations. The chapter concludes by outlining the major challenges facing the field of CRC microbiome research, which will be addressed by the work of this thesis.

1.1 The colorectal microbiome in health and disease

1.1.1 The microbiome as an ecosystem

The point at which the colorectal microbiome is first established is currently unclear; some studies suggest it may be established *in utero*, although critics believe the findings may be secondary to contamination (1). The microbiome of neonates is influenced by the method of delivery (2-4) and feeding (3-8). This initially simplistic microbiome gradually increases in taxonomic and metabolic complexity until a stable diverse 'adult-like' state is reached at approximately three years of age (5, 7). The developed microbiome represents a microbial ecosystem containing bacteria, archaea, viruses, fungi and parasites. Although often studied in isolation, it lies in continuity with the rest of the gastrointestinal tract; associations have been found with both the oral and gastric microbiomes (9-11). The components of the colorectal microbiome will now be discussed in turn.

1.1.1.1 Bacteria

The adult microbiome contains 10^{13} bacteria, comprising approximately 150 different species and 200,000 common bacterial genes which encode a minimum of 6000 different functions (12-14). Approximately half of the bacterial species and genes identified in an individual's microbiome are present in half of all individuals (14).

It had previously been hypothesised that all microbiomes could be categorised into three distinct 'enterotypes', defined by the relative abundance of *Bacteroides*, *Prevotella*, and *Ruminococcus* (15), however this hypothesis is no longer

believed to be correct (16). Instead a diet-dependent continuum of the inversely associated *Prevotella* and *Bacteroides* has been described; microbiomes are either *Prevotella* or *Bacteroides* predominant, with the relative abundance of the remaining taxa demonstrating high inter-individual variability (16). Functional activity demonstrates less inter-individual variation and functional potential the least (10, 17, 18).

Within an individual microbiome, only a small number of species are highly abundant (15). Low-abundant species often make an important functional contribution, they may increase in abundance in response to insults to the microbiome and have been associated with CRC (15, 19).

Opportunistic pathogens have been detected within the microbiomes of healthy individuals but high risk pathogens have not (20).

1.1.1.2 Archaea, viruses and fungi

Non-bacterial members of the microbiome are less well characterised due to their relatively low abundance, the incompleteness of reference databases and the fact that primers and pipelines have been optimised for bacterial detection (21). The majority of faecal DNA is bacterial; 4-17% is viral, 0.8% archaeal, 0.5% other eukaryotic organisms, 0.14% human and 0.01% fungal (15, 22, 23).

Archaea constitute one of the three domains of the tree of life; they have certain features in common with bacteria and certain in common with eukaryotes (24). Archaea have been identified in 50-95% of microbiomes (25, 26); the most frequently detected species is *Methanobrevibacter* (21, 26).

Fungal and viral (including phage) sequences have been detected with 100% prevalence (22, 26, 27).The virome and mycobiome show high inter-individual variability (22, 27). Unlike the mycobiome, the virome is temporally stable (27) and correlates with the bacterial microbiome (22, 23). Importantly the virome can serve as a reservoir for antibiotic resistance genes (23).

1.1.1.3 Parasites

Colorectal parasites (helminths and protozoa) have not been well characterised, as the majority of microbiome research is conducted in Western countries where parasite prevalence is believed to be low. This may be a misconception; one study found a prevalence of faecal parasites in Danish healthy controls of 50% (28). Parasites have co-existed with the microbiome during the course of evolution and it is therefore important to investigate this relationship.

Associations between certain parasites and bacterial diversity or taxonomic composition have been described in both Western and non-Western cohorts (28-33) and investigated mechanistically (34, 35). Confounding factors include changes to the microbiome secondary to parasite-induced gastrointestinal symptoms, past exposure or anti-helmintic treatments rather than the parasites per se and co-existence of more than one type of parasite (36, 37).

1.1.2 The physiological role of the colorectal microbiome

The microbiome makes an important physiological contribution, to the extent that humans are considered 'holobionts', with the microbiome termed a 'forgotten organ' which performs the following metabolic roles (38, 39):

- Synthesis of vitamin K and B group vitamins (40).
- Further metabolism of the material received from the small intestine, increasing host energy availability (41).
- Metabolism of carbohydrates to short chain fatty acids (SCFAs); the most abundant are acetate, propionate and butyrate. SCFAs are a source of energy for colonocytes. SCFAs have been shown to have anti-proliferative and immunomodulatory properties (influencing both innate and adaptive immune responses) (42-45), to play a role in glucose homeostasis and cardiovascular health (40, 46) and to influence microglia development in mouse models (44, 47).
- Metabolism of bile acids to secondary bile acids, some of which are cytotoxic and some of which act as hormones. The microbiome deconjugates bile acids, enabling their return to the enterohepatic circulation (40, 48).

- Metabolism of drugs, affecting toxicity and efficacy (49-51).

- Metabolism of mucus and promotion of mucin secretion (52).

The microbiome influences the development of the immune system and the anatomical development and physiological function of the colon (53-56). It reduces the likelihood of colonisation by potentially pathogenic bacteria through the occupation of niches, competition for nutrients and bactericidal activity (57).

The microbiome is largely confined to the colonic lumen by the mucus barrier, production of antimicrobial peptides, secretion of IgA and activity of mucosal immune cells (54), although bacteria have been identified in colonic crypts of both patients with CRC and healthy volunteers (58, 59).

1.2 Colorectal cancer and the microbiome

There is growing evidence of an association between CRC and an altered (dysbiotic) microbiome. Many of the proposed CRC-associated bacteria are Gram-negative anaerobes typically found in the oral microbiome. They are capable of forming biofilms and produce virulence factors which, in mechanistic studies, have been shown to modulate CRC tumourigenesis. However, whether the CRC-associated microbiome contributes to tumour initiation and/or progression in man or is merely secondary to changes in the colonic environment, has not been determined; prospective longitudinal studies are required to answer this question. Current understanding of the CRC-associated microbiome will now be reviewed.

1.2.1 Epidemiological evidence

Epidemiological research indicates an association between increased CRC risk and microbiome-related factors including: high meat and animal fat consumption (60); obesity (61), alcohol consumption (62), antibiotic exposure (63-66) (confirmed in a mouse model) (67); appendectomy (68); poor dentition (69); certain Toll-like receptor (TLR) single nucleotide polymorphisms (SNPs) (70, 71); and antecedent bacteraemia with certain CRC-associated species (*Bacteroides fragilis, Streptococcus gallolyticus, Fusobacterium nucleatum, Peptostreptococcus species, Clostridium septicum, Clostridium perfringens, or*

Gemella morbillorum) (72, 73). Decreased CRC and adenoma risk has been associated with yogurt consumption in men (74) and whole grain fibre consumption (75).

The age-standardised incidence of small intestinal adenocarcinoma is low (2.8/100,000 in the UK in 2016), compared with CRC (69.3/100,000) (76). The small intestine contains fewer, different and less diverse bacteria and their associated metabolites compared to the colon (77, 78).

1.2.2 Dysbiosis

Dysbiosis denotes a perturbation of the microbiome from the healthy state. Dysbiosis has been described in both patients with colorectal adenomas and patients with CRC. Dysbiosis can be detected in faecal and mucosal (tumour or normal mucosa) samples (79-91), although differences according to sample type have been observed (86, 92, 93). Within patients, the microbiome of tumour-mucosa differs from adjacent normal mucosa (a relative dysbiosis), and these differences become more marked as distance between the two mucosal samples is increased (86, 94-98).

The dysbiotic microbiome contains the same number of bacteria as in a healthy system, but the taxonomic composition and metabolic profile is different (96, 99, 100). There is a relative depletion of SCFA-producing bacteria and an enrichment of bile salt-metabolising and mucin-degrading bacteria (101-110), a lower concentration of butyrate (111, 112) and increased faecal pH (113). Virulence genes are overexpressed (94, 114), bacteria form distinct microbe-microbe and microbe-host co-occurrence networks (94, 108, 115-117) and there is an enrichment of ‘oral pathogen’ and ‘oral biofilm-associated’ bacteria (11, 98, 114, 118-121). The route by which ‘oral’ bacteria reach the colon (intra-luminal or haematogenous) has not been confirmed, but research suggests that these bacteria do not merely transit through the colon but survive and proliferate (10, 11).

Many of these ‘oral’ bacteria have the ability to form biofilms. Biofilms denote polymicrobial consortia enclosed within an extracellular polymer matrix, adherent to a solid surface (122). Biofilms provide a survival advantage by concentrating nutrients and confer protection against host defence mechanisms and antibiotics (both exogenous and endogenous) (123, 124). Biofilms have been identified in

approximately 50% of sporadic CRC, 40-65% of sporadic adenoma (percentage varying with type), 70% of FAP (Familial adenomatous polyposis) colectomy specimens and 10-20% of healthy controls (125-127). Biofilms overlie not only tumours, but also distant normal mucosa and the normal mucosa of FAP post-operative ileal pouches and ano-rectal stumps (125, 127). An association with right-sided lesions was demonstrated in two of four studies (125-128) and one case report which identified biofilms within colonic crypts of a right-sided CRC (129); no association with tumour stage has been found as yet (128). It is hypothesised that biofilms cause E-cadherin disruption of the underlying mucosa, with subsequent increased mucosal permeability, bacterial invasion, inflammation, and epithelial proliferation (125). Biofilms have been associated with an upregulation of tissue polyamine metabolites (130). Transfer of biofilm homogenates from both healthy controls and patients with CRC has been shown to induce tumourigenesis in a mouse model (131).

The evidence for differences in the microbiome between patients with adenoma or CRC and controls is compelling, but the concept of a single adenoma/CRC-associated microbiome is unlikely. Differences in the microbiome have been found within patients with adenoma/CRC according to the following tumour characteristics: location (92, 93, 132-134), type (135), grade (107, 133), stage (120, 136, 137), size (133), mutational and molecular profile (138-142) and associated systemic inflammation (143). Furthermore, both tumourigenesis and the microbiome are dynamic; the dysbiotic signature profiled by a single sample may be merely transient (120).

The aforementioned studies have been unable to distinguish causation from association. Prospective and interventional studies in man are required to investigate a causative role of the microbiome in tumourigenesis; already one has shown an association between microbiome signatures and risk of adenoma development (144). Meanwhile, mechanistic studies allow potential causation to be explored.

1.2.3 Mechanistic studies

Work using CRC cell lines has shown that components of the microbiome signal via TLR4 (expression of which is interestingly reduced by aspirin), to induce epithelial-mesenchymal transition (EMT), cellular migration (145), proliferation (146) and chemokine secretion (147). Non-toxigenic *Bacteroides fragilis*, a

bacterium which is typically depleted in CRC, has been shown to have the opposite effect (148).

In mouse models, tumourigenesis is reduced under germ-free conditions (146), antibiotic administration (149-155), targeted depletion of oncomicrobes (156), high-fibre dietary intervention (157) and knock-out of certain components of the immune signalling cascade (152, 158, 159). Conversely, tumourigenesis is increased with depletion or loss of function of myeloid cells (160, 161) and defective barrier function with subsequent increased bacterial invasion (162). Transfer of the microbiome from tumour-bearing mice or CRC patients increases tumourigenesis in recipient mice (150, 163) in some, but not all, cases (164). Interestingly, dysbiosis has been observed in a genetic CRC mouse model prior to the development of microscopically detectable polyposis; this suggests a host-induced change in the microbiome which occurs extremely early during tumourigenesis, although the cause of this change has not been elucidated and requires confirmation in man (165).

1.2.4 Candidate bacteria of interest

CRC-associated bacteria identified by associative and mechanistic studies show a degree of variability across studies and cohorts, likely due to genuine biological but also technical differences (97, 166). The more consistently identified bacteria, *Fusobacterium nucleatum* (*F. nucleatum*), Enterotoxigenic *Bacteroides fragilis* (ETBF) and pks+ *Escherichia coli* (*E. coli),* became the focus of extensive investigation and have been proposed as putative 'oncomicrobes'. They are discussed below. The association between CRC and *Streptococcus gallolyticus* (formally *Streptococcus bovis*) has been recognised since the 1970s and will also be described.

However the current focus on only a small number of bacteria requires revision; a recent meta-analysis of faecal metagenomes revealed *F. nucleatum, Bacteroides fragilis*, *Porphyromonas asaccharolytica*, *Parvimonas micra*, *Prevotella intermedia*, *Alistipes finegoldii*, and *Thermanaerovibrio acidaminovorans* to be consistently enriched in CRC compared with controls and 62 bacteria to be depleted (166).

1.2.4.1 *Fusobacterium nucleatum*

Fusobacterium nucleatum (*F. nucleatum*) is a Gram-negative anaerobic rod-shaped bacterium found within the oral and colorectal microbiome (167). Within the oral microbiome it exists as an opportunistic pathogen; it facilitates oral biofilm formation and is implicated in periodontitis and gingivitis (168, 169). It has also been associated with infections including appendicitis, osteomyelitis, chorio-amnionitis, pericarditis and brain abscess, Inflammatory Bowel Disease (IBD) and pre-term and still-birth (167) and has been identified in oesophageal and gastric tumours (170).

F. nucleatum is both more prevalent and more abundant in adenoma tissue (particularly adenomas with high grade dysplasia (HGD)) and CRC tissue compared with adjacent normal mucosa, and in the stool and normal rectal mucosa of patients with adenoma or CRC compared with healthy controls (81, 88, 89, 171-184), although amounts of tissue and stool *F. nucleatum* are not correlated (175). Tissue prevalence ranges from 15-50% of CRC (170, 185), 25-45% of adenomas (171, 186), 30% of lesion-adjacent normal mucosa (187) and 20% of normal mucosa of healthy controls (126). The detection of *F. nucleatum* has a sensitivity and specificity for CRC of 0.81 (95% CI: 0.64–0.91) and 0.77 (95% CI: 0.59–0.89) in meta-analysis of faecal and mucosal studies combined and in faecal studies alone a sensitivity and specificity of 0.68 (95% CI: 0.64-0.72) and 0.78 (95% CI: 0.75-0.81) (188, 189).

A high abundance of *F. nucleatum* has been shown to associate consistently with CIMP-high, MSI-high and hMLH1 methylation and in some studies has been found to associate with greater tumour size, poor differentiation, high stage, certain mutations (cytosine to thymine, CHD7/8, AMER1, ATM, KRAS and TGF-β pathway mutations), TP53 wild type and diets with a high propensity to induce inflammation (134, 138, 142, 172, 174, 180, 186, 190-195). Most, but not all (186, 196), studies have found an association with serrated lesions and right-sided lesions (126, 193). No association has been found with BRAF mutations, gender or age (187, 196).

The source of *F. nucleatum* is believed to be the oral microbiome; identical strains can be identified in saliva and CRC tissue (197) although they may exhibit a degree of genetic divergence (117). *F. nucleatum* is more abundant within saliva and CRC tissue compared with faeces (198). Two small pilot studies found no difference in the abundance of *F. nucleatum* in saliva samples from CRC patients

compared with controls(198, 199).and neither did a prospective case-control study (200). *F. nucleatum* has the ability to co-aggregate with bacteria (117) and be bactericidal to others (certain probiotic strains), suggesting that it influences the surrounding microbiome (201). *F. nucleatum* forms biofilms both within the mouth and in the colon, particularly overlying tumours where 'blooms' of *F. nucleatum* have been observed (128).

F. nucleatum binds by the Fap2 surface protein to D-galactose-β(1–3)-N-acetyl-D-galactosamine (Gal-Gal-NAc) which is overexpressed by the epithelial cells of adenomas, CRC and CRC metastases (plus other adenocarcinomas) (202, 203). *F. nucleatum* is able to invade and survive within the mucosa (173, 204), epithelial cells (172, 204, 205) and macrophages (206); interestingly mucosal invasion appears to be independent of the presence of *F. nucleatum* in biofilms (126). Adenoma and CRC have reduced mucin (Muc2) and tight junction proteins which may facilitate bacterial invasion (158).

Given the invasive nature of *F. nucleatum*, it is unsurprising that inflammatory pathways (including COX-2 and NF-kβ) are up-regulated (171). Interestingly *F. nucleatum* has been shown to be positively associated with tumour-infiltrating lymphocytes (TILs) and an intratumoural periglandular lymphocytic reaction in non-MSI-high tumours but negatively associated with these factors in MSI-high tumours (185); another study showed an association with CD68+ macrophages in MSI-high tumours (207). Other studies have found an inverse association between high abundance of *F. nucleatum* and CD3 or CD4 T cell density (190, 208). The potential for immune cell subversion has been shown with both macrophages (206, 209), T cells and NK cells (210, 211). Inflammation or subversion of the anti-tumour immune response may initiate or potentiate tumourigenesis.

F. nucleatum also has the potential to affect tumourigenesis more directly. Mechanistic studies have shown that *F. nucleatum* increases β-catenin signalling through two pathways: the TLR4/p-PAK1 cascade (152) or through binding of the molecule *Fusobacterium* adhesin A (FadA) to E-cadherin/Annexin A1, with subsequent NF-kβ, Myc and Cyclin D1 expression (212, 213). Of note, levels of FadA are increased in CRC tissue compared with adjacent normal tissue and tissue from patients with adenomas or healthy controls (212). *F. nucleatum* has been shown to promote proliferation, invasion and production of inflammatory cytokines by CRC cell lines; pathways such as E-cadherin signalling and miR21-

release of MAPK signalling have been implicated (214, 215). Treatment of *Fusobacterium*-positive (but not negative) patient-derived xenografts (PDXs) with metronidazole (to which *F. nucleatum* is sensitive) reduces the rate of tumour growth (204).

F. nucleatum has been detected in a proportion of CRC metastases (liver and lymph node) in addition to a lower proportion of non-metastatic colorectal lymph nodes (126, 173, 204). *Fusobacterium* is viable within liver metastases and shows >99.9% average nucleotide identity with isolates from the primary CRC. Liver metastases from *F. nucleatum*-negative primary CRC have not been found to contain *F. nucleatum.* The bacteria which co-occur with *F. nucleatum* in CRC are also present within CRC liver metastases (204).

There is conflicting evidence for an association between *F. nucleatum* and CRC recurrence or survival. Some studies have not found an association (185, 186, 190, 193, 204, 216). Others have shown high levels of *F. nucleatum* are associated with poor prognosis (175, 179, 191, 215, 217-221), which may be stage-dependent (174, 181), and recurrence (219). A potential mechanism through which *F. nucleatum* may mediate resistance to 5-Fluorouracil (5-FU) or oxaliplatin has been proposed: *F. nucleatum* reduces apoptosis of CRC cells treated with chemotherapy via TLR4 and MYD88 signalling, which causes a reduction in miR-18a and miR-4802 and an increase in autophagy pathways (219). An alternative mechanism has also been proposed: upregulation of BIRC3 with subsequent inhibition of apoptosis (221).

1.2.4.2 Enterotoxigenic *Bacteroides fragilis*

Like *F. nucleatum*, *Bacteroides fragilis* is a Gram-negative anaerobic rod-shaped bacterium; strains of *Bacteroides fragilis* which are capable of producing the toxin fragilysin (*Bacteroides fragilis* toxin (BFT)) are termed Enterotoxigenic *Bacteroides fragilis* (ETBF) (222, 223).

BFT detection is via quantitative polymerase chain reaction (qPCR) (224) or potentially enzyme-linked immunosorbent assay (ELISA) (225). The BFT toxin is detected in a higher percentage of CRC tissue (usually concordant with adjacent normal mucosa) compared to controls (89% versus 67%) (226) and a higher percentage of stool from CRC patients (27-38%) compared to controls (10-12%) (227, 228). One study has found higher levels of BFT in the normal adjacent

mucosa of adenoma patients compared with controls (and also CRC patients) (229).

BFT increases the permeability of tight junctions which causes epithelial cells to round, become vacuolated and separate from adjacent cells and the basement membrane (230-233). This enables bacterial invasion (234) which triggers an acute (235) and then persistent IL-17-dependent colitis with subsequent epithelial proliferation (233, 236-239). The mucus barrier plays an important role in limiting access of BFT to the epithelium; again, it should be noted that adenoma and CRC have reduced Muc2 (158, 240).

BFT inhibits apoptosis of epithelial cells (241) and induces them to secrete IL8 (via NF-kβ and MAPK signalling) which leads to neutrophil chemotaxis and activation (242-244). BFT also causes nuclear translocation of β-catenin (via E-cadherin cleavage) leading to c-Myc upregulation and increased proliferation (245-248). A mouse model has shown that tumourigenesis is affected by the duration of ETBF colonisation (249).

1.2.4.3 *Escherichia coli*

Escherichia coli (*E. coli*) is a Gram-negative rod-shaped bacterium which is a facultative anaerobe. *E. coli* are capable of producing several toxins which are either genotoxic or cyclomodulating (250). Enzymes required for the synthesis of the genotoxin colibactin are encoded by a mobile genetic element the polyketide synthases (pks) pathogenicity island; *E. coli* which possess the pks island are termed pks+ *E. coli* (251). Other virulence factors include UPEC-pathogenicity islands, bacteriocin production, mucosal attachment, and invasion. *E. coli* with such pathogenic properties are increased in CRC (67%) and IBD (40%) tissue and faeces compared with control (20%) (250, 252-260). Evidence as to whether *E. coli* from patients are more capable of biofilm formation than *E. coli* from controls is conflicting (261, 262).

Colibactin induces DNA alkylation and double strand breaks (263-265). Xenograft studies have shown that this causes cellular senescence, subsequent growth factor release and increased tumour proliferation; senescence markers are increased in human CRC biopsies containing pks+ *E. coli* (266). Attaching and effacing *E. coli* (AEEC) may attach to the epithelium and cause downregulation

of the mismatch repair (MMR) proteins MSH2 and MLH1; they are detected in 20% of CRC tissues (267, 268) and associate with increased Ki67 expression (257). Mice transfected with pks+ *E. coli* or adherent pks+ *E. coli* have been shown to develop significantly more tumours than controls (253, 257).

CRC-associated bacteria may exhibit synergy. Both pks+ *E. coli* and ETBF are enriched in the tissue of FAP patients; a mouse model demonstrates that co-colonisation increases tumourigenesis compared with mono-colonisation, hypothesising that ETBF reduces mucus depth to enable attachment of pks+ *E. coli* (127).

1.2.4.4 *Streptococcus gallolyticus*

Streptococcus gallolyticus (S. gallolyticus) (formally *Streptococcus bovis*) is a Gram-positive coccus which is a facultative anaerobe. An association between *S. gallolyticus* and CRC was first described in 1977 when a high prevalence of *S. gallolyticus*-associated endocarditis was observed in patients with CRC (269). Faecal carriage of *S. gallolyticus* is higher in patients with CRC or large adenomas than controls (269, 270). *S. gallolyticus* has been detected in 30-70% of CRC tissues, significantly more than controls, and at higher abundance than adjacent normal mucosa (271, 272). Serum levels of anti-*S. gallolyticus* antibody are increased in patients with adenoma and CRC compared with controls (273) and can be detected prior to diagnosis (mean 3.4 years) (274). This suggests that *S. gallolyticus* is an early coloniser of tumours. Mice which are predisposed to developing CRC are more likely to become colonised with *S. gallolyticus* than normal mice, suggesting that the tumour microenvironment may offer a selective advantage to certain bacteria such as *S. gallolyticus* (275). Early cell line and mouse model research suggest that *S. gallolyticus* may augment tumourigenesis through inflammation and β-catenin signalling (271, 276).

1.2.5 The possibility of a CRC-associated virome/mycobiome

The association of CRC with an altered virome or mycobiome has been less extensively investigated but is garnering interest; results so far are contradictory. One study found no difference in viral abundance between CRC and adjacent normal tissue (173), whereas other studies have identified viruses including Epstein-Barr virus (EBV) (although with conflicting results), Human Papillomavirus (HPV), BK virus (BKV), Herpes simplex virus (HSV), Cytomegalovirus (CMV) and bacteriophages within CRC tissue and faeces (184,

277-281). One study has shown that the CRC-associated virome associates with oral bacteria and certain viruses associate with reduced survival (282). Professor zur Hausen, who won the Nobel Prize for the discovery of HPV's aetiological role in cervical cancer, cites epidemiological evidence of an association between CRC and dairy/beef consumption and hypothesises that a bovine virus may play a role in CRC development (283-287).

Differences in the mycobiome between patients with adenoma or CRC and controls have been identified in some (288, 289) but not all (290) studies, and potential mechanisms by which fungi may affect tumourigenesis have been proposed (291, 292). The protozoa *Cryptosporidium* has been detected with increased prevalence in CRC patients compared with controls in Western (293) and non-Western cohorts (294), as has *Blastocystis* in non-Western cohorts (295, 296).

1.2.6 Potential clinical implications

If the microbiome plays a role in the initiation or progression of CRC, then it presents an opportunity to improve our understanding of the disease and develop novel therapies. Methods to modify the microbiome include dietary modifications (297), probiotics (298-303), prebiotics, synbiotics (a combination of probiotics and prebiotics) (304), engineered bacteria (305), antibiotics (151), phage (306, 307), predatory bacteria, faecal microbiome transplant (FMT) (308-312) or vaccine (313).

If the microbiome is merely passively associated with CRC, it has the potential to be used as a diagnostic/screening biomarker, as a prognostic marker and or as an indicator of likely response to existing therapies. Research has shown that the microbiome predicts efficacy of anti-PD1 immunotherapy and that this phenotype can be transferred via FMT to mouse models (314-316). The microbiome has also been shown to influence the efficacy of 5-FU (317), to have the potential to inactivate gemcitabine (318) and to influence chemotherapy side-effects (319-321). Research suggests that modifying the microbiome could reduce complications of surgery (322).

Therefore, whether causative or associative, the microbiome offers great promise to improve management of CRC.

1.3 How the colorectal microbiome is studied

Given the increasing appreciation of an association between the microbiome and CRC, it is useful to outline the types of study, samples and techniques on which this evidence is based and to describe their limitations.

1.3.1 Types of study

1.3.1.1 Studies in man

The majority of studies conducted in man have been cross-sectional in design, and have shown a difference in the microbiome of cases compared with controls. Causation cannot be determined by this type of study; prospective, longitudinal studies are required to demonstrate that perturbations of the microbiome precede the development of disease. Randomised controlled trials (RCTs) have and are being conducted to investigate the effect of microbiome-based interventions on certain diseases.

1.3.1.2 Animal models

Causation can be explored mechanistically using animal models. Germ-free, Altered Schaedler's Flora (ASF) and Specific-Pathogen-free (SPF) mice have been used to investigate the microbiome, through the introduction or abrogation of specific bacteria or the transplantation of whole microbiomes (323). However, differences between human and murine anatomy, physiology, life span, diet, behaviour and environment mean that mouse models may not accurately reflect conditions in man (324-328).

1.3.1.3 *In vitro* models

The most simplistic *in vitro* microbiome research assesses the effect of specific bacteria or their associated molecules on cell lines. More complex *in vitro* models include ‘gut simulators’ (329, 330) and ‘gut on a chip’ microfluidic devices (331-334), which model colonic conditions on a large and small-scale respectively, and organoids (335-337).

1.3.2 Types of sample

1.3.2.1 Colonic samples

There are three types of colonic sample: mucosal, luminal and faecal. Mucosal samples are collected by biopsy, swab or balloon device and luminal samples by endoscopic aspirates.

Mucosal samples capture the microbiome which is in direct contact with the outer mucus layer, biofilms, bacteria within crypts and invasive bacteria. The outer mucus layer provides a source of nutrients and attachment sites for bacteria; it sits above a barrier of dense, layered mucus which bacteria do not usually penetrate (338). The advantage of mucosal samples is that the mucosal microbiome is believed to be less transient than the luminal or faecal microbiome, it is in proximity to the mucosa, it is lesion or colorectal location-specific, and it is sampled *in situ*, which avoids exposure to ambient conditions. Disadvantages are that collection is invasive and usually occurs post-bowel preparation, which changes the microbiome, samples are of lower biomass than faecal or luminal samples, and mucosal samples may be contaminated with faecal material during sample collection (339, 340). The corollary applies for the advantages/disadvantages of faecal samples. Luminal samples combine the advantages of *in situ* sample collection and high biomass with the disadvantages of invasive sample collection usually post-bowel preparation.

The mucosal, luminal and faecal microbiomes have been shown to differ in diversity and composition (340-348). The mucosal microbiome has been found to contain more asaccharolytic bacteria which digest mucus, and more aerotolerant bacteria, able to withstand oxygen diffusion from the underlying mucosa (346). However, these findings must be interpreted with caution, as differences in how samples were collected (whether pre/post bowel preparation, the method used to collect and store samples and duration of exposure to ambient oxygen) could underlie some of the reported differences (349, 350).

Differences in the mucosal microbiome along the length of the colon have been described (342, 347, 351, 352), although these findings may be confounded by micro-heterogeneity (significant differences in the microbiome have been described in biopsies taken 1cm apart (353)). The luminal microbiome appears to show less regional variation; the faecal microbiome is most similar to the distal luminal microbiome (347).

1.3.2.2 Extra-colonic samples

The metabolic profile associated with the microbiome can be analysed from breath (354), urine (355, 356) and blood (357, 358).

1.3.3 Methods of analysis

1.3.3.1 16SrRNA sequencing

Bacterial taxonomy was traditionally culture-dependent, based on morphology and nutritional requirements (359). Study of the microbiome was limited; approximately 60% of the faecal microbiome is uncultivable (360, 361). A pivotal point came with the discovery that the bacterial 16S ribosomal ribonucleic acid (16SrRNA) gene could be used to infer bacterial phylogeny and taxonomy as it is universally present amongst bacteria and has a slow rate of evolutionary change (359, 362-364). 16SrRNA contains nine hypervariable (V) regions, the sequences of which differ between taxa. Conserved regions, the sequences of which are mostly conserved between taxa, flank the V regions (365). Primers which are specific to a pair of conserved regions allow amplification of the interim V region. The amplicons are sequenced, usually by Next Generation Sequencing (NGS), and the sequences are then compared to a reference database to infer the bacteria within a sample.

16SrRNA sequencing is a relatively inexpensive and high-throughput method of microbiome analysis. However, it has limited sensitivity: sequencing a single V region has discriminatory power only to genus (not species or strain), with V2 or V4 having the lowest error rates for genus assignment (366). Results can be affected by choice of primer, choice of V region(s), amplicon length and depth of sequencing (367-373). 16SrRNA sequencing does not provide information about bacterial function, although it is possible to crudely infer function from taxonomic information using the software package PICRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) (374).

1.3.3.2 Metagenomic sequencing

Metagenomic sequencing uses a shotgun sequencing approach to sequence all DNA (human, bacterial, viral and fungal) within a sample. Sequences are aligned to reference databases to determine the genes present within a sample. The taxonomic composition and functional potential of the microbiome can be derived.

However, horizontal gene transfer and sub-speciation confound the assignment of taxonomy and the presence of DNA does not necessarily signify the presence of viable organisms.

Fewer samples can be sequenced per lane with metagenomic sequencing compared to 16SrRNA sequencing, making the technique more expensive. However, existing low coverage whole genome sequencing data derived from formalin fixed paraffin embedded (FFPE) material can be mined for bacterial metagenomic data using the software package PathSeq (375). In contrast, few studies have reported successful 16SrRNA sequencing using DNA extracted from FFPE tissue (137, 376, 377). Analysis of metagenomic data is computationally intense, although analysis pipelines are continually being improved (378, 379).

1.3.3.3 Sequencing biases

Both 16SrRNA and metagenomic analysis are subject to biases inherent to sequencing, including sequencing errors and clustering by run (380). Both methods assign reads to reference databases; the stringency of the assignment and choice of database can affect the results (15, 370). The databases are continually expanding as new computational techniques are developed and as new organisms are discovered (381, 382). It is hoped that the emerging technique of long read sequencing will improve sensitivity and accuracy, although this will not be applicable to FFPE samples (383, 384).

1.3.3.4 Analysis of sequencing data

As the microbiome represents an ecosystem, many of the methods of microbiome data analysis derive from ecology. Bacteria phylogenetic taxa comprise phylum, class, order, family, genus and species. Bacterial diversity can be described by two measures: alpha and beta diversity. Alpha diversity describes the diversity *within* a sample. Beta diversity describes differences in diversity *between* samples. Different alpha and beta diversity metrics take account of presence/absence (species richness) and/or relative abundance (species evenness) (385). Some beta diversity metrics also take phylogeny into consideration; the most commonly used is UniFrac (386). UniFrac can be calculated as unweighted UniFrac (based only on the presence/absence of taxa) or weighted UniFrac (which weights taxa according to their relative abundance) (387). Beta diversity can be displayed as a Principal Co-ordinate Analysis (PCA)

plot – the more similar the bacterial community of two samples, the closer together they are.

Network modelling can be used to determine groups of bacteria which co-occur and those which have an inverse relationship (388). Bacteria which are significantly enriched or depleted in cases compared with controls can be determined using the software package LEfSe (Linear discriminant analysis Effect Size) (389).

More sophisticated methods of analysis are continually being developed.

1.3.3.5 Metatranscriptomics, metaproteomics and metabolomics

Although functional potential can be inferred from metagenomic and 16SrRNA data, metatranscriptomic, metaproteomic and metabolomic analysis determine active microbiome function. This is illustrated by a comparison of metagenomic and metatranscriptomic faecal microbiome data which showed that 20% of transcripts had relative abundances an order of magnitude different to that predicted by DNA abundance (10). A further advantage of these techniques is that information about host transcripts, proteins and metabolites is also captured. All three methods can be performed using mucosal, luminal or faecal samples; metabolomics can also be performed on faecal, urine or breath volatile organic compounds (VOCs) (390-392).

1.3.3.6 Targeted analysis of specific bacteria

The aforementioned techniques give insight into the structure and function of the bacterial microbiome in its entirety. As these sequencing and spectrometry-based techniques are relatively expensive and time consuming, they are often used to generate hypotheses as to which bacteria differ between cases and controls. Simpler and cheaper techniques, including qPCR (204), oligonucleotide-based microarrays (393) and fluorescent *in situ* hybridisation (FISH) (127) can then be used to quantify and visualise specific bacteria of interest. There are no reports in the microbiome literature of immunohistochemistry (IHC) being used to detect specific bacteria of interest, although it has been shown to be possible (394).

1.4 Limitations of existing microbiome research

The field of microbiome research is young, so understandably there has so far been limited translation of research findings to clinical practice. The microbiome is also an inherently complex ecosystem complicated by horizontal gene transfer, redundancy and rapid evolutionary timescales. However, there are also a number of limitations to the current methods of conducting microbiome research which need to be urgently addressed.

1.4.1 The need for standardisation

One reason why many microbiome studies are generating conflicting or irreproducible results is that microbiome analysis is extremely sensitive to technical aspects of study design. Factors which can affect microbiome results are outlined in Table 1 (395-401).

Table 1. Factors which can affect microbiome results.

Sample processing	**Laboratory processing**	**Bioinformatic analysis**
Biological factors Colonic location Sample type Homogenisation Number of replicates Collection media Transport conditions Storage temperature Storage duration	Extraction method Choice of 16SrRNA V region Choice of DNA polymerase Choice of PCR primer Amplicon length Number of rounds of PCR Sequencing depth Sequencing run variation	Bioinformatic pipeline used Choice of reference database Choice of alpha and beta diversity metrics Choice of statistical analysis method

There is therefore a need for consistency within and between studies in order to reduce variability in results and permit meaningful meta-analysis. The International Human Microbiome Standards project and the Microbiome Quality Control project aim to investigate which variables might be large enough to

overwhelm biological variability, and to create openly-available protocols, a positive reference standard and potentially an External Quality Assessment scheme (399, 402). The Earth Microbiome Project, which is focused on investigating all microbial habitats not just human, has open-access standardised protocols for sample collection, processing and analysis which have been adopted by many microbiome research groups (403-406).

1.4.2 The need to conduct large-scale studies in representative populations

Many microbiome studies to date have sampled small numbers of participants due to expense and logistical (recruitment and sample collection) constraints. Small studies are inadequately powered to detect subtle differences between cases and controls, especially in light of high inter-individual and temporal variation in the microbiome. Recently published power calculations estimate that for faecal microbiome studies to detect a difference with an odds ratio of 3.5, 100–400 cases would be needed; for an odds ratio of 1.5, 1000–3000 cases would be required (407). In order to increase sample size, studies often pool results from separate cohorts, yet technical and biological differences between cohorts introduce bias. This is true also of microbiome research consortia (408-411).

It is, therefore, important to conduct large-scale microbiome research using a single cohort and consistent methodology. The British Gut Project (412) and the American Gut Project (413) represent two such studies, based on a citizen-science, crowd-funded model whereby interested members of the public pay a donation to submit samples for microbiome analysis. However, limitations include participation bias, variability in sample collection, self-reported clinical metadata and bacterial blooms.

1.4.3 The need to conduct longitudinal studies

In order to confirm a causative role of the microbiome in CRC development, prospective longitudinal case-control studies with repeated sampling or biobanks are required. Existing longitudinal studies such as the Nurses' Health Study are starting to collect microbiome samples and new studies are being established (414). To do so successfully, a stable method of microbiome collection and transport is required that will be acceptable to study participants, in addition to the timely collection of accurate clinical follow-up data.

1.4.4 The need to conduct microbiome research in non-Western countries

The majority of microbiome research to date has been conducted in Western countries, largely due to expense. There is a need to expand microbiome research to non-Western countries, in order to investigate differences between the Western and non-Western microbiome in health and to determine whether the disease-microbiome associations identified in Western cohorts are universal or geography-specific. It is important to investigate the microbiome of people living in non-Western countries now, as the incidence of Western disease (including CRC) is increasing rapidly as these countries adopt a Western lifestyle; there exists a critical window in which to analyse the non-Western microbiome in its native state, before it is potentially changed irreversibly by Westernisation.

1.5 Chapter Summary

- The microbiome is a complex microbial ecosystem with bacterial, archaeal, viral, fungal and parasitic components.

- The microbiome contributes to host development, metabolism, physiology and immunology. Perturbations of the microbiome are associated with disease.

- A dysbiotic microbiome is associated with CRC. Candidate oncomicrobes include *F. nucleatum*, ETBF, pks+ *E. coli* and *S. gallolyticus*.

- Potential clinical implications of the CRC-associated microbiome include it acting as a novel diagnostic/screening marker, as a prognostic marker or as a potentially modifiable predictor of drug response and toxicity. If the CRC-associated microbiome is found to play a role in tumour initiation or progression, this would change current understanding of CRC and its management.

- The microbiome is a relatively new field of research which arose secondary to the advent of NGS.

- The microbiome is studied using *in vitro* techniques, animal models and human association studies. Types of sample include mucosal, luminal, faecal and extra-colonic. Types of analysis include 16SrRNA sequencing,

metagenomic sequencing, metatranscriptomics, metaproteomics metabolomics, FISH and qPCR.

- A lack of standardised, large-scale, longitudinal microbiome research studies limits progress and needs to be addressed. There is also an urgent need to conduct microbiome research in non-Western populations.

1.6 Aims and objectives

The aim of this PhD is to address current limitations to the clinical translation of CRC microbiome research by developing a method to conduct a large-scale, single-methodology microbiome study in the UK; to use this study to assess the utility of a microbiome-based CRC screening model; and to assess whether the same methodology could be used to conduct CRC microbiome research in non-Western populations.

Objectives:

- To develop and assess a method to perform microbiome analysis directly from faeces on processed NHSBCSP samples.

- To use this methodology to conduct a large-scale microbiome study in order to identify CRC/adenoma-associated taxa and to determine whether microbiome analysis improves the accuracy of CRC screening.

- To assess whether the same methodology can be applied to cohorts from Argentina, Chile, India and Vietnam and whether the CRC-associated bacteria identified in Western cohorts are also identified within these cohorts.

Chapter 2
Investigating the potential to use NHSBCSP samples for microbiome analysis

2.1 Introduction

This chapter describes the development of a method to conduct large-scale single-methodology microbiome research through harnessing the existing infrastructure of the NHSBCSP. The method involves analysing the microbiome directly from processed bowel cancer screening samples; traditionally these were gFOBT, however the NHSBCSP is currently transitioning to FIT. The rationale for this choice of methodology will first be outlined, followed by laboratory aspects of the study design.

2.1.1 Collection and storage of faecal samples

Within the microbiome literature there is no consensus as to the optimum method of faecal sample collection and storage. The gold standard is considered to be fresh stool which undergoes either immediate DNA extraction or immediate freezing at -80°C. However this is rarely practical: precluding studies where participants collect faecal samples at home, studies conducted in institutions without -80°C freezing facilities and studies conducted in remote locations where cold-chain transport would be required. Alternative methods of sample collection and storage have therefore been proposed with the aim of limiting both DNA degradation and bacterial overgrowth, relative to the gold standard. When choosing a method to perform microbiome research, the following points require consideration (Table 2):

Table 2. Factors related to method of faecal sample collection.

Item to consider	Possible options
Type of stool specimen	Whole stool sample Stool subsample (+/- prior homogenisation) Replicate subsamples Collection at a single or multiple time points
Media/device for sample collection and storage	Ethanol RNA*later* FIT device Card: gFOBT or Flinders Technology Associates (FTA) OMNIgene.GUT
Storage conditions	Ambient temperature 4°C -20°C -80°C
What will be measured	Bacterial DNA Human DNA RNA Metabolites Proteins Bacterial culture
Participant acceptability	
Cost	

Each of these items will now be discussed, and the implications of using NHSBCSP samples will be outlined.

2.1.1.1 Type of stool specimen

There is no consensus as to the optimum type of stool sample for microbiome research. Many studies ask participants to collect whole stool samples, yet this is surplus to the requirements of most DNA extraction kits, which limit the amount of stool/sample to ~250mg (equivalent to 2ml). A potential advantage of collecting whole stool samples is that they can be homogenised; however this is not performed by all research groups and if performed, is usually done so manually which may not be entirely effective. Some devices (FIT device, card, OMNIgene.GUT) collect subsamples of stool in the absence of prior homogenisation. There has been limited investigation of the impact of these different methodologies. Two studies reported that sampling from stool affords good reproducibility of taxa at abundance greater than 1% but not lower abundance taxa (395, 415), although another study, which used qPCR rather than 16SrRNA, reported marked variability in the relative abundance of taxa between subsamples, including differences between the inner and outer part of the stool (416), and a third study showed that reproducibility of subsamples varied by individual (417).

Most studies perform analysis on a single sample collected at a single timepoint. However the microbiome exhibits dynamic temporal variation in response to changes in environmental conditions such as diet (20, 417-419). Different taxa are affected to differing degrees (for example *Fusobacteria* has been reported to exhibit high temporal variation); calculating an average from multiple timepoint samples is therefore recommended (407).

2.1.1.1.1 Implications of using NHSBCSP samples

NHSBCSP gFOBT instructions ask participants to collect two subsamples from three separate stools and to record the dates of collection. Stool is not homogenised prior to sampling. For the current study, it was decided to combine three of the six subsamples (one from each of the three stools) and to leave the remaining three subsamples available for subsequent analysis or to be used as extraction replicates; the implications of this approach will be evaluated in this chapter. Temporal variability of the CRC-associated microbiome will also be evaluated.

The NHSBCSP is currently transitioning to FIT. FIT instructions ask participants to scrape the tip of the device across the surface of a stool. This subsamples a

single stool; the biomass collected is significantly less than that collected by gFOBT.

2.1.1.2 Media/device for stool collection and storage

A number of different media/devices for stool collection and storage have been proposed for use in microbiome research: ethanol (which in theory would allow metabolomic analysis), ethylenediaminetetraacetic acid (EDTA), RNA*later* (which in theory would allow metatranscriptomic analysis), card (gFOBT or FTA), FIT, and OMNIgene.GUT. Technical studies of these devices evaluate three key metrics: accuracy, reproducibility and stability. Accuracy denotes how similar the microbiome result is to that of the gold standard (stool at -80°C); reproducibility denotes how similar technical replicates are to one another; stability denotes the degree to which the microbiome changes after storage over time.

In most 16SrRNA studies, the collection method has not been shown to significantly affect alpha or beta diversity metrics, with subject being the greatest source of variation (396, 420, 421). Regarding relative abundance of taxa, gFOBT, OMNIgene.GUT and FIT have variously emerged as optimum methods of sample collection/storage. These will be outlined below, after a review of the commonly used alternative, frozen stool samples.

2.1.1.2.1 Frozen stool

The gold standard for comparison in technical microbiome studies is either whole stool stored immediately at -80°C or mock bacterial communities (395). Although gFOBT, OMNIgene.GUT and FIT have been proposed as sample collection/storage devices for microbiome research, the majority of microbiome studies still use frozen stool samples. It is therefore important to evaluate the impact of freezing stool on microbiome results.

The microbiome of stool stored at -80°C is considered to be stable long term (confirmed in one study to be stable at six months) (422) and a mock microbiome community stored at -20°C has been shown to be stable at four weeks (423). However, freezing has also been shown to increase the *Firmicutes*:*Bacteroidetes* ratio (424) and to cause a change in 20% of genes compared to immediate DNA extraction (425). Five cycles of freeze-thawing, storage for longer than three days in domestic frost-free freezers (which may be encountered by studies which ask

participants to freeze their samples at home), and delays greater than one hour between defrosting and DNA extraction have been shown to affect results (416, 426). No benefit has been shown for snap-freezing (427).

Often freezing of stool is not immediate, particularly in the case of studies whereby participants collect samples at home. Several studies have reported that the microbiome of stool stored at room temperature for 24 hours is stable (422, 426, 428), whereas others contest this, detecting differences after as little as 30 minutes storage at room temperature (416) which increase gradually over time (with an average change of bacterial community composition of 3% at 12 hours and 9% at 24 hours) (429). The American Gut Project shipped stool samples at ambient temperature in the absence of preservative; bacteria blooms of the class *gammaproteobacteria* were observed. These were sufficient to obscure some biological effects and had to be computationally subtracted from the analysis (430). Stool samples stored for 14 days at room temperature have markedly altered taxonomic composition and it has been observed that some develop visible fungal growth (426, 431).

2.1.1.2.2 gFOBT

In most technical comparison studies, the gFOBT card has shown high reproducibility, stability and acceptable accuracy (432). Stool collected on gFOBT and stored, undeveloped, at room temperature for four days showed high stability with a marked fold change of only 1-3 Operational Taxonomic Units (OTUs) (which are broadly equivalent to taxa) compared to an equivalent change in 20-37 OTUs for whole stool (396, 421); stability was also shown after gFOBT storage for seven days at room temperature (433). The Leeds group has demonstrated similar microbiome profiles of fresh whole stool and replicate gFOBT samples and stability of the microbiome from stool on developed gFOBT cards stored at room temperature for up to three years (434) and beyond to five years (unpublished data).

Correlation for OTUs between undeveloped and developed gFOBT has been shown to be high, indicating that application of the hydrogen peroxide-based developer solution does not alter the microbiome result (396).

FTA cards, which are similar to gFOBT but contain a nucleic acid stabiliser, have also been shown to afford high accuracy after storage at room temperature for 24 hours (420) and high stability up to eight weeks despite marked fluctuations in temperature (4-40°C) (435). Similarly gFOBT had high stability, reproducibility and acceptable accuracy in a technical microbiome study conducted in Bangladesh, indicating the potential for gFOBT to be used to conduct microbiome research in non-Western populations (436).

2.1.1.2.3 OMNIgene.GUT

OMNIgene.GUT is a commercial microbiome collection device. Stool is scraped into the cap of the device, the lid is sealed and the device (containing media and a ball-bearing) is shaken to homogenise the stool. The manufacturer guarantees sample stability at ambient temperature for up to 60 days.

Technical studies have shown that OMNIgene.GUT yields high quality extracted DNA and RNA, and has high accuracy and stability (437-440). Some differences in taxa relative abundance between OMNIgene.GUT and stool immediately stored at -80°C have been recorded (438, 440). OMNIgene.GUT has been shown to afford high stability for up to eight weeks despite marked fluctuations in temperature (4-40°C) (435).

2.1.1.2.4 FIT

FIT has been shown to afford high reproducibility, stability and acceptable accuracy (432), although had poor stability in a technical study conducted in Bangladesh, suggesting that it may be less appropriate for studies conducted in non-Western populations (436). Importantly it should be noted that different FIT devices are available, and both studies only tested one type of FIT device, although this happened to be the device which will be used by the NHSBCSP (variously referenced as OC-Sensor, OC-Auto or Polymedco).

One study tested the OC-Sensor FIT device, comparing the microbiome of devices immediately stored at different temperatures (-86°C, -20°C, 4°C, 20°C, 30°C) and devices stored at 4°C for two days followed by 20°C for two days (441). It should be noted that these conditions do not reflect the conditions that screening samples would be exposed to. The authors reported difficulty extracting sufficient DNA from FIT, and resorted to lyophilisation of the samples.

The authors reported a decrease in alpha diversity and Gram-negative bacteria over time, but acknowledged that they did not assess this in non-FIT samples and did not collect replicates. One study showed that microbiome analysis could be performed from FIT and gFOBT after 10 years storage at -80°C, however whole stool samples were not available for comparison, which prevents assessment of stability, accuracy and reproducibility (442).

One study has importantly confirmed that cross-sample contamination does not occur during automated processing of FIT and that processed (perforated) FIT samples can be stored frozen without sample evaporation (443).

2.1.1.2.5 Implications of using NHSBCSP samples

The aforementioned studies suggest that gFOBT and FIT may be suitable methods of collecting and storing stool samples at ambient temperature for microbiome analysis. However, the majority of these technical studies were conducted using stool samples from small numbers of healthy volunteers. Stability, reproducibility and accuracy ideally need to be assessed using stool from the population of interest (i.e. patients with adenoma and CRC) as the relative abundance of taxa, metabolite and enzyme profiles will differ from that of healthy volunteers, which may give rise to different results. For example one study demonstrated that stool samples from healthy volunteers were more stable over time than samples from Irritable Bowel Syndrome (IBS) patients (422) and another showed differences in accuracy between meconium and stool stored on gFOBT (433).

There is a need to perform the assessment using the make of device that will be used by the NHSBCSP and conditions that the device will be exposed to. For NHSBCSP samples these include:

- Sample collection at home
- Sample transport to the Screening Hub at ambient temperature
- Sample processing:
    - gFOBT: application of hydrogen peroxide-based developer solution
    - FIT: piercing of the foil cap by machine and aspiration of an aliquot of the buffer
- Storage by the Screening Hub:
    - gFOBT: samples are stored together in large batches at room temperature
    - FIT: the Screening Hub has not finalised storage conditions of FIT
- Transport to a laboratory
- Storage by a laboratory

It is not possible to perform comparison of the microbiome analysed from NHSBCSP samples with immediately frozen stool samples as it is vital that routine screening is not disrupted. Stability of the microbiome on gFOBT stored at room temperature will be assessed by comparing extraction replicates after storage for different lengths of time.

2.1.1.3 What will be measured

Most technical microbiome studies have performed 16SrRNA analysis; few studies have assessed other methods of microbiome analysis. For metabolomic analysis, 95% ethanol has been shown to afford optimal accuracy followed by FTA card and OMNIgene.GUT (420). Another study also demonstrated high accuracy and stability after four days storage at room temperature for 95% ethanol and gFOBT but not OC-Sensor FIT (444).

It is also possible to analyse human DNA from faecal samples; mutation detection could potentially be used to improve screening (it forms part of the Cologuard test (445)). Several studies have assessed this using different storage buffers but to the author's knowledge this has not been assessed using gFOBT, FIT or OMNIgene.GUT (437, 446, 447).

2.1.1.3.1 Implications of using NHSBCSP samples

The current study will perform 16SrRNA analysis, however only three of the six subsamples will be processed, leaving the remaining available for other methods of analysis (such as metabolomics).

The current study only has ethical approval to analyse bacterial DNA. However, as samples are link-anonymised, the option of targeted mutation-specific analysis of human DNA may be acceptable to a research ethics committee (REC); this is an area for future work.

2.1.1.4 Participant acceptability

Participant acceptability is an important consideration as it can affect study uptake and consequently the degree of selection bias. In one study, sample collection at home was shown to be more acceptable to the majority of participants than sample collection in clinic (448). Study participants using an OMNIgene.GUT device have reported good acceptability (449).

2.1.1.4.1 Implications of NHSBCSP samples

NHSBCSP uptake with gFOBT has been reported as 35-60% (450); this may in part be related to the method of sample collection but also reluctance to participate in the screening programme. Pilot work anticipates that uptake will improve with the introduction of FIT (451), which is considered to be “easier to complete” and “less disgusting” (452). Indeed this has proven to be the case in Scotland where screening uptake has improved from 56% to 64% (453).

2.1.1.5 Cost

Costs include those of the kit, transport and storage. OMNIgene.GUT currently cost £11/kit (but it may be possible to reduce this cost if ordering in bulk). Creating bespoke collection kits with instructions and packaging approved by the postal service is expensive.

2.1.1.5.1 Implications of using NHSBCSP samples

Microbiome analysis from processed NHSBCSP samples negates the cost of collection device and the cost of delivery to and collection from participants. The following costs apply: link-anonymisation of samples and data extraction by the Screening Hub and transport of samples from the Screening Hub to the microbiome laboratory. Transport can be performed in batches by courier at ambient temperature, as gFOBT are exempt from the infectious substances category of the World Health Organisation 'Guidance on regulations for the transport of infectious substances 2019– 2020' (454). For the current study, gFOBT were stored at room temperature, therefore storage costs were minimal.

2.1.2 Considerations for laboratory processing

The rationale for assessing whether microbiome analysis can be performed directly from the faeces of processed NHSBCSP samples has been presented and informs the work of this chapter. Several laboratory processing factors also require consideration as follows:

2.1.2.1 Choice of laboratory methodologies

As outlined in the Introduction, choice of laboratory methodologies can influence microbiome results; a standardised method for DNA extraction, PCR and library preparation and sequencing underpinned this work.

Many different DNA extraction kits, optimised for bacterial DNA extraction, are available. Many contain a buffer to prevent DNA degradation by enzymes present in stool and to remove PCR inhibitors. A mechanical lysis (bead-beating) step is required to achieve lysis of all bacteria, particularly Gram-positive bacteria which have a thick cell wall.

Choice of DNA extraction method has been shown to affect the amount and purity of extracted DNA and the relative abundance of taxa (395, 423, 438). A standardised methodology, optimised from the laboratory's existing protocol, will be used throughout the thesis.

Choice of library preparation method (including choice of V region and primers) can affect microbiome results. The Earth Microbiome Project (EMP) 16SrRNA microbiome methodology will be used throughout the thesis (403-406). This is an open-access, straightforward protocol which other groups will be able to follow should they wish to replicate the studies.

Microbiome results have been shown to be affected by sequencing run (380). Up to 1500 samples will be processed on each Illumina HiSeq run to reduce the need to process samples from the same study on different runs.

The work will be conducted under routine laboratory conditions; negative controls will therefore be used to test for possible contamination.

2.1.2.2 The potential for high-throughput sample processing

In order to conduct large-scale microbiome research using NHSBCSP samples, high-throughput processing is an important consideration. The rate limiting step in sample processing is DNA extraction; this process can be automated using a robot. An experiment to compare results generated by manual and automated DNA extraction has been performed, although sequencing results are not available at this point.

2.2 Aims

- To determine whether the microbiome can be successfully analysed from processed NHSBCSP gFOBT samples.

- To assess whether the microbiome is stable if NHSBCSP gFOBT samples are stored at room temperature.

- To assess temporal variation of the microbiome of screening participants.

- To determine whether the microbiome can be successfully analysed from the FIT devices which the NHSBCSP will use, after simulation of the conditions that the FIT samples will be exposed to.

- To determine whether the DNA extraction process can be up-scaled.

2.3 Methods

2.3.1 NHSBCSP gFOBT samples

2.3.1.1 Collaborators

This study was conducted in collaboration with the team at the NHSBCSP Southern Hub, located in Guildford. This is the largest NHSBCSP Hub, serving a population of 14.5 million. The Hub provided link-anonymised processed gFOBT (Immunostics Inc., USA) samples and matched clinical data.

2.3.1.2 Ethical approval

To prevent disruption to screening, ethical approval was sought to collect and analyse link-anonymised samples and matched clinical data without consent (on the proviso that only bacterial DNA, and not human DNA, would be analysed). The following ethical approvals and subsequent amendments were granted (Table 3):

Table 3. Details of ethical approvals.

Committee	Ethical approval reference
North East-Tyne & Wear South Research Ethics Committee (REC)	IRAS project ID: 188007 REC reference: 16/NE/0210
Bowel Cancer Screening Programme Research Committee	BCSPID_160
Office for Data Release (ODR)	ODR1617_126 ODR1617_126/A1
Amendments	
Original protocol: The collection and analysis of 400 gFOBT blood-negative cards, 600 gFOBT blood-positive cards and matched link-anonymised clinical data (limited to age, gender and greatest risk for the current screening episode).	
Amendment 1: The continued collection and analysis of all gFOBT cards with a clear positive result until the time when the NHSBCSP stops issuing gFOBT cards.	
Amendment 2: The collection and analysis of more detailed link-anonymised clinical data (pathological data and the greatest risk for preceding screening episodes).	
Amendment 3: The collection and analysis of 9000 FIT samples.	

2.3.1.3 Usual processing of samples by the Southern Hub

In order not to interfere with the screening process, the samples comprised gFOBT cards which had been processed routinely by the team at the Southern Hub; these cards would otherwise have been disposed of. A brief summary of what happens to gFOBT samples is given below.

The NHSBCSP posts gFOBT cards to adults aged 60-74 biennially. There are three flaps on the front. A participant applies faeces from a single stool to the two squares beneath the first flap, seals the flap and records the date. This is repeated for the second and third flaps. The completed card is posted back to the Screening Hub (Figure 1).

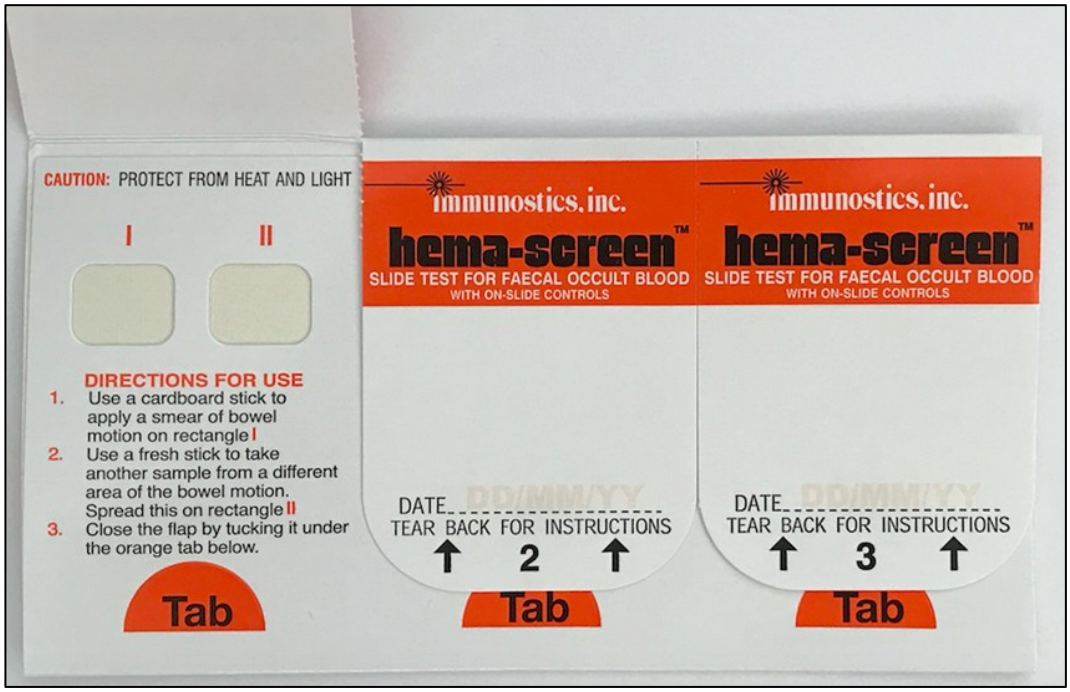


Figure 1. Application of stool to gFOBT cards by screening participants. Photograph illustrating how a screening participant would receive a gFOBT card. One of the three flaps has been opened to reveal two squares beneath.

At the Screening Hub, a strip is removed from the back of the card and a hydrogen peroxide-based developer solution (Immunostics Inc., USA) is applied to the reverse of the six squares. If blood is present, blue discolouration occurs. If no colour change is detected, the result is deemed ‘normal/negative’ and screening is complete. If a blue colour change occurs in five or six squares, the result is deemed ‘abnormal/positive’ and colonoscopy is offered. If a blue colour change occurs in one to four squares, the result is deemed ‘unclear’ and up to two further gFOBT cards are dispatched (Figure 2).

Figure 2. The collection of 'positive' and 'negative' gFOBT samples. This diagram indicates 'normal/negative' and 'abnormal/positive' gFOBT samples (marked with red stars). Diagram adapted from NHS Cancer Screening Programme's "Guidance for public health and commissioners"(455).

2.3.1.4 Sample collection

Sample collection began in October 2016. Samples were link-anonymised and stored collectively at room temperature before being placed in plastic boxes for storage and transport. Prior to June 2018, samples were sealed in individual plastic bags; this was subsequently deemed unnecessary (as cards were stored collectively prior to being placed into individual bags). Boxes of samples were then transported at room temperature to the University of Leeds where they were stored at room temperature prior to processing.

2.3.1.5 Sample preparation

2.3.1.5.1 gFOBT dissection

From each gFOBT three squares of faecally-loaded card, one from beneath each of the three flaps, were dissected and processed as a single combined sample (Figure 3). The rationale was that the three remaining squares could be used for alternative analysis or as technical replicates.

Figure 3. gFOBT dissection. The left-hand image is a gFOBT card viewed from the front. One of the three flaps is lifted to reveal two squares. The right-hand image show the reverse side of a gFOBT card, after the strip has been removed. Six squares, two underlying each of the three flaps, are visible. One square from beneath each of the three flaps was dissected; these were processed as a single combined sample.

Time between stool collection and DNA extraction (during which samples were stored at room temperature) is displayed in Figure 4. The wide range reflects variation in time between stool collection and receipt by the team at the Southern Hub, time until gFOBT cards were anonymised and labelled by the team at the Southern Hub (dependent upon staff availability and the demands of routine screening), time until gFOBT cards were sent to the Leeds laboratory (samples were sent as batches to reduce courier costs), and time until DNA extraction (dependent upon release of clinical metadata so that samples corresponding to a diagnosis could be processed and manual DNA extraction being constrained to 24 gFOBT/day). gFOBT cards were stored at room temperature throughout the different phases of storage. Differences in time until DNA extraction were shown not to affect microbiome results (see section 2.4.4 and 3.5.3).

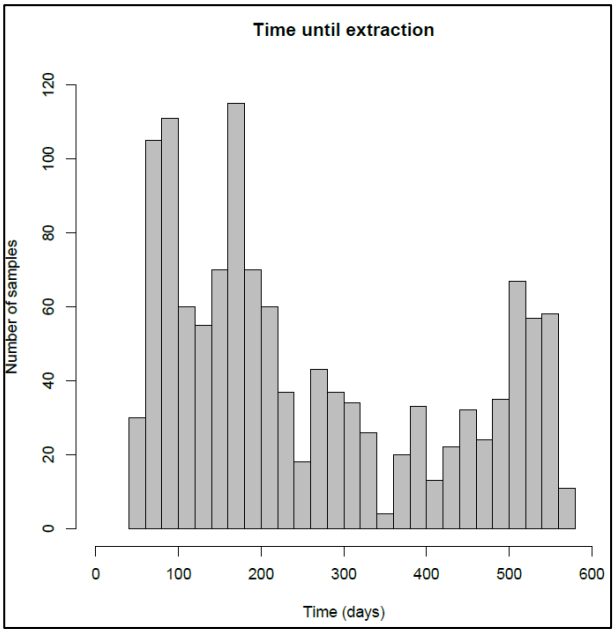


Figure 4. Time between stool collection and DNA extraction for NHSBCSP samples.

2.3.1.5.2 Extraction replicates

Extraction replicates were created to determine whether the choice of which three squares to extract altered the microbiome result. Three squares (one from beneath each window) were dissected and combined to make a single sample and the alternate three squares were dissected and combined to make a replicate sample (n=51).

An additional set of extraction replicates was prepared to determine whether prolonged storage of samples at room temperature altered the microbiome result. Three squares (one from beneath each window) were dissected and combined to make a single sample and, after a period of time (6-23 months) the alternate three squares were dissected and combined to make a replicate sample (n=26) (Figure 5). The replicates extracted after ~200 days reflect the median length of time until DNA extraction for all the NHSBCSP samples (Figure 4); the replicates extracted after ~ 700 days reflect the maximum length of time until DNA extraction for all the NHSBCSP samples (Figure 4).

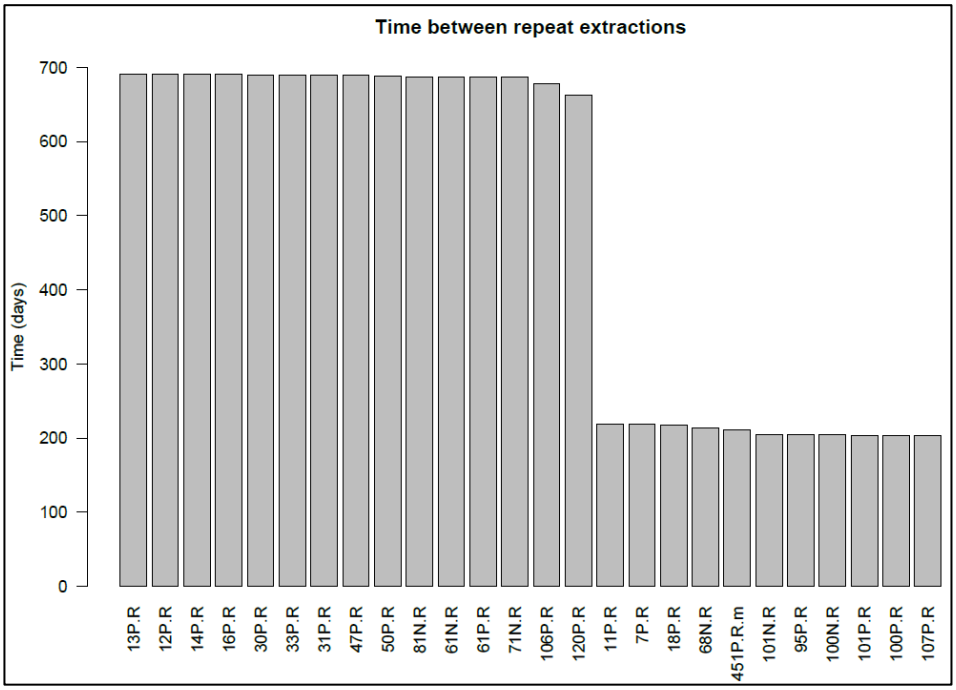


Figure 5. Time between DNA extractions for replicate sample pairs. Each bar represents the second member of a replicate pair. The y axis indicates the time between DNA extractions of each replicate pair. Two time points were chosen to reflect the median and maximum length of time until DNA extraction for all the NHSBCSP samples.

2.3.1.5.3 Assessing temporal variation of the microbiome

A set of samples was used to assess the temporal variation of the microbiome of screening participants, with particular focus on: the temporal variability of CRC-associated bacteria; whether stool samples collected on different days would produce similar results; and whether a single stool sample (as FIT will be) would produce similar results to the existing gFOBT samples derived from a combination of three stool samples. NHSBCSP gFOBT samples with visibly large amounts of faecal material/square were selected and dissected in the following ways (Figure 6):

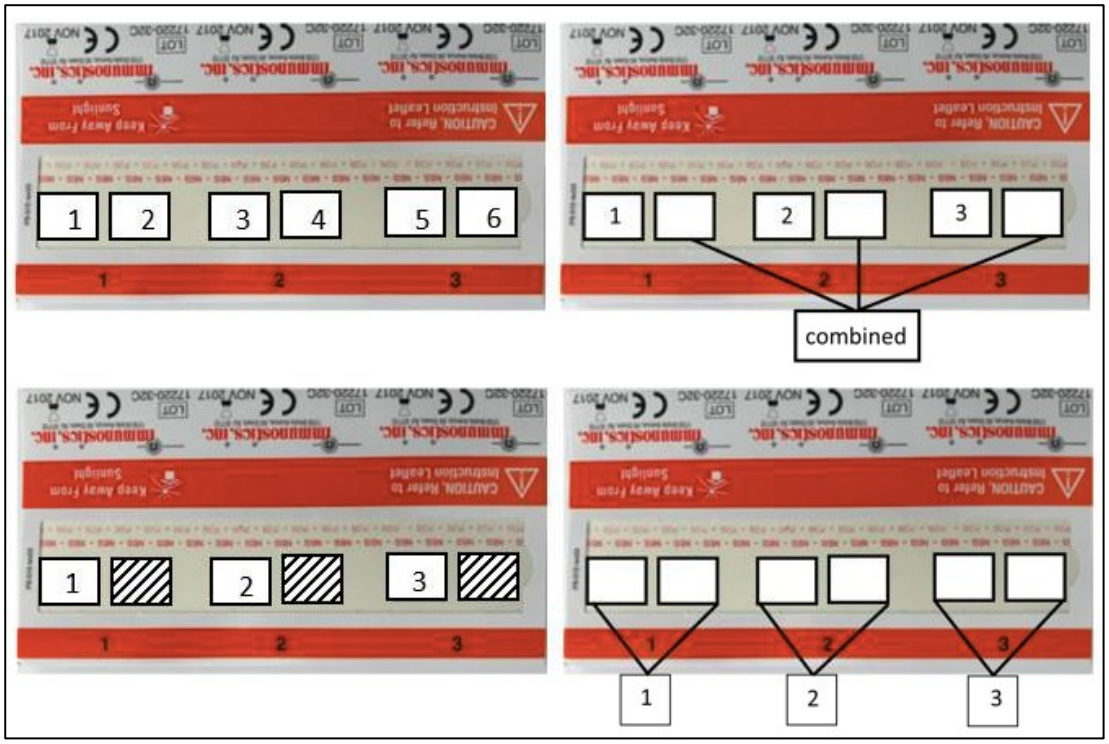


Figure 6. Dissection of Temporal samples. Labelled squares were dissected. *Top left:* Temporal 1-6 samples. *Top right:* Temporal 1.2.3.combined samples. *Bottom left:* Temporal N1-3 samples (3 squares had previously been extracted – indicated by hash). *Bottom right:* Temporal P1-3 samples.

Details of the temporal samples are described in Table 4 and Figure 7.

Table 4. Temporal sample characteristics.

Type of sample	Number of gFOBT cards	Temporal period over which stools were collected
Temporal 1-6	6 blood-positive	3 consecutive days
Temporal 1.2.3.combined	24 blood-positive 2 blood-negative	3 consecutive days
Temporal N1-3	5 blood-negative	3 consecutive days apart from one sample where the stools were collected over days 1, 3 and 4.
Temporal P1-3	8 blood-positive	Variety of days (see Figure 7)

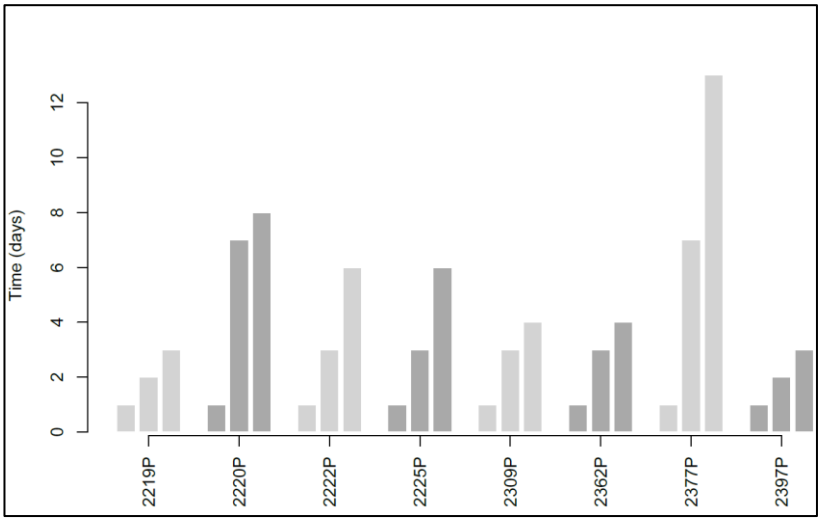


Figure 7. Day of stool collection for Temporal P 1-3 samples. Each gFOBT sample is represented by three bars; each bar represents a subsample (two gFOBT squares) derived from a single stool. The y axis indicates the relative day of each stool collection.

2.3.1.5.4 FIT experiment

In order to assess whether microbiome analysis could be performed on the FIT devices which the NHSBCSP will be using (OC-Sensor FIT (Mast Group Ltd)), the supplier was approached and generously provided blank devices. Two healthy volunteers (A and B) provided two stool samples on two separate days. The healthy volunteers were aged 31 and 63 and had no history of antibiotic use within the previous five years or of colonoscopy. Stool was manually homogenised using a spatula. Triplicate samples were made using FIT stored at different temperatures; gFOBT stored at room temperature; and stool immediately stored at -80°C (Figure 8).

In order to assess whether the FIT solution has an immediate effect on the microbiome, replicates from one sample from each of the volunteers underwent DNA extraction after one day of storage.

In an attempt to recreate the conditions that NHSBCSP FIT samples will be exposed to, a set of FIT replicates from all four samples were stored for three days at room temperature; this simulated postage of FIT devices to the Screening Hub. Processing of the FIT was then simulated by piercing the foil of the devices with a sterile pipette tip and squeezing liquid into the chamber. The FIT were then sealed with parafilm and stored upright for five days at either room temperature,

4°C or -80°C to simulate transfer from the Screening Hub to the laboratory in Leeds. Samples underwent DNA extraction either on day eight (the soonest timepoint that DNA extraction of NHSBCSP FIT samples could be expected to occur) or after eight weeks storage (Figure 8). The results of samples extracted on day one and day eight will be presented; samples extracted after eight weeks have been processed but the results are not yet available.

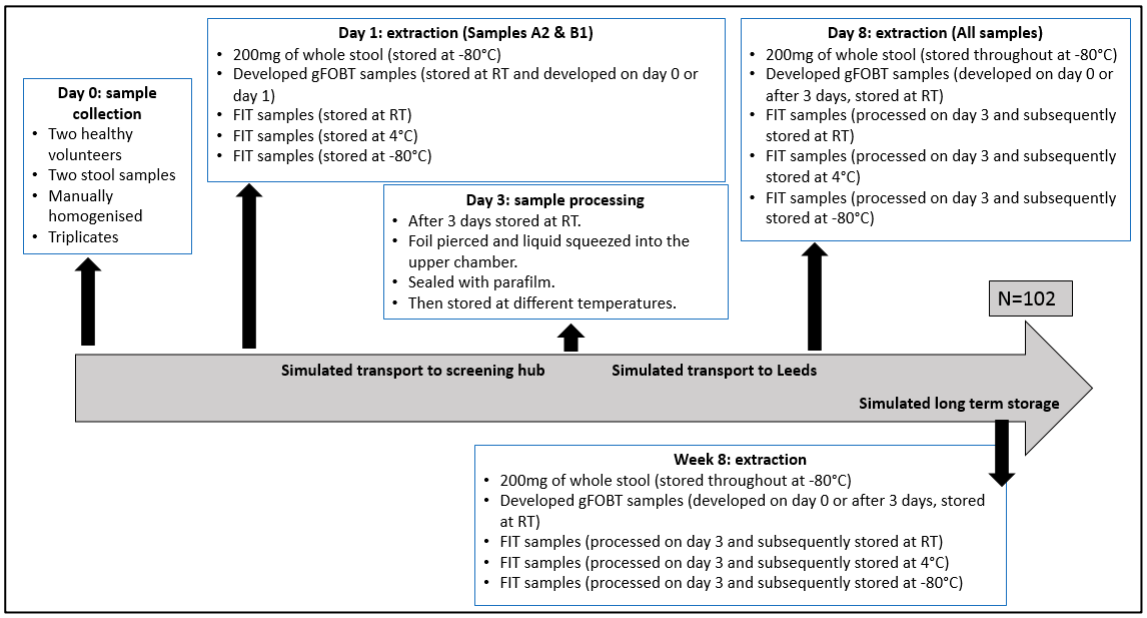


Figure 8. Processing of 'FIT experiment' samples. This diagram illustrates production, storage and extraction of 'FIT experiment' samples. RT = room temperature.

2.3.2 DNA extraction

2.3.2.1 Modifying the laboratory's existing DNA extraction protocol

The laboratory's existing bacterial DNA extraction protocol (434) had been developed from the protocol of Yu and Morrison (456). It was based on the protocol of the QIAamp DNA Stool Mini Kit (Qiagen, Germany) and had been optimised to enable DNA extraction from low biomass stool samples as follows (Table 5):

Table 5. Modifications to the laboratory's existing DNA extraction protocol.

Modification to the QIAamp DNA Stool Mini Kit protocol	**Rationale**
Omission of InhibitEX tablets (designed to remove PCR inhibitors present in stool).	PCR was successful when InhibitEX tablets were omitted; they were deemed unnecessary.
Addition of a bead-beating pathogen lysis step using pathogen lysis tubes (Qiagen, Germany) and mechanical shaking.	To ensure mechanical lysis of bacterial cell walls (including Gram-positive bacteria).
Addition of an ammonium acetate precipitation step.	To precipitate impurities.
Addition of a propan-2-ol precipitation step.	To precipitate the DNA.
Addition of a 70% ethanol wash step.	To remove salts from the DNA pellet.

The laboratory's existing DNA extraction protocol was further modified to optimise it for high-throughput processing of NHSBCSP gFOBT samples as follows:

- The existing protocol had been used to extract DNA from mock gFOBT samples (prepared by volunteers from the laboratory); DNA was extracted from two squares on each card (combined to make a single sample). It was noted that the amount of stool on NHSBCSP samples (prepared by screening participants) was very variable and was often less than the mock samples. The protocol was modified so that DNA was extracted from *three* squares. The microbiome result therefore represents a composite of three separate stool samples and leaves three squares intact to be used as replicates or for future studies.

- Using the existing protocol, twelve gFOBT samples could be extracted per day. The limiting step was that each square was initially processed separately, combined at the pathogen lysis step, then divided into two for the propan-2-ol precipitation step, before recombining the DNA for subsequent purification and elution. The protocol was modified by adjusting the volumes of the reagents at these steps to enable each sample to be processed as a single entity, without division and

recombination. This modification enabled DNA to be extracted from twice as many gFOBT samples per day and reduced the likelihood of error. A system was devised whereby two researchers, working in parallel, could extract DNA from forty-eight gFOBT samples per day. This up-scaling was necessary in order to meet the sample size of the NHSBCSP study (Chapter 3).

2.3.2.2 The modified DNA extraction protocol

The modified DNA extraction protocol is now described. Originally the QIAamp DNA Stool Mini Kit (Qiagen, Germany) was used, however this kit was discontinued in August 2018. Subsequently the QIAamp DNA Mini Kit (Qiagen, Germany) was used; this contains the same buffers apart from Buffer ASL (Qiagen, Germany) which was purchased separately. To avoid confusion, only the QIAamp DNA Mini Kit is referenced.

The first part of the process (until the point when DNA is extracted) was undertaken within a biohazard fume-hood as a health and safety precaution. In order to reduce batch effects, batches of samples contained a mixture of different sample types (i.e. both blood-positive and blood-negative gFOBT samples).

For each gFOBT sample, one square from beneath each of the three flaps on the front of the card was dissected using a sterile scalpel (Swann-Morton, UK). The three dissected squares were placed into a single collective microcentrifuge tube (Eppendorf AG, Germany). This step was either completed the day before or on the day of DNA extraction.

In order to remove faeces from the card, 800µl of Buffer ASL (Qiagen, Germany) was added to each microcentrifuge tube and these were incubated at 23°C on a ‘Thermomixer Comfort’ (Eppendorf UK) at 850 revolutions per minute (rpm) for 1 hour (note, there is no relative centrifugal force (RCF) equivalent as the Thermomixer is not a centrifuge). Samples were briefly centrifuged (Hettich Mikro 200, DJB Labcare) and the supernatant transferred to pathogen lysis tubes (S) (Qiagen, Germany). Bacterial cell lysis was achieved by placing the samples on a shaker (Vibrax VXR, IKA, UK) at a motor setting of 1800-2200 for 10 minutes followed by incubation at 95°C on the Thermomixer at 850rpm for 15 minutes.

To precipitate impurities, samples were centrifuged at 14000rpm (18625g) for one minute and the supernatant transferred to a microcentrifuge tube containing 173µl of 10M ammonium acetate (made in-house). Samples were vortexed (Vortex Genie 2, Scientific Industries, USA) and placed on ice for five minutes before being centrifuged at 14000rpm (18625g) for five minutes. To achieve DNA precipitation, the supernatant was transferred to a microcentrifuge tube containing 725µl of propan-2-ol (Fisher Scientific, UK), vortexed and placed on ice for 30 minutes.

To remove salts, samples were centrifuged at 14000rpm (18625g) for ten minutes, supernatant was discarded and 1ml of 70% ethanol (Sigma-Aldrich, USA) was added to the DNA pellet. Samples were centrifuged at 14000rpm (18625g) for five minutes, the supernatant was discarded and 500µl 70% ethanol was added. Samples were centrifuged at 14000rpm (18625g) for three minutes, the supernatant was discarded and the samples were left with lids open for ten minutes to allow residual ethanol to evaporate.

200µl tris-EDTA (Fisher Scientific, USA) was added to re-suspend the DNA pellet. After ten minutes, samples were vortexed and added to microcentrifuge tubes containing 200µl of Buffer AL from the QIAamp DNA Mini Kit (Qiagen, Germany). 15µl of Proteinase K from the QIAamp DNA Mini Kit was added, the samples were vortexed and incubated at 70°C on the Thermomixer at 650rpm for ten minutes.

The QIAamp DNA Mini Kit protocol was then followed, with centrifuge speeds of 14000rpm (18625g). To elute the DNA, 100µl of ultraviolet (UV) irradiated molecular biology grade water (Fisher Scientific, USA) was added to samples for five minutes before centrifuging at 14000rpm (18625g) for one minute.

2.3.2.3 Modifications for DNA extraction from FIT

The following modifications applied for DNA extraction from FIT:

FIT were vortexed for one minute. The probe was removed from the device and liquid squeezed into a sterile microcentrifuge tube. Samples were centrifuged at 14000rpm (18625g) for ten minutes. Supernatant was discarded and 800µl of Buffer ASL (Qiagen, Germany) was added to the pellet and vortexed to mix.

Samples were then processed according to the protocol; as FIT are lower biomass samples, elution was with 50µl rather than 100µl of water.

2.3.2.4 Modifications for DNA extraction from whole stool samples

The following modifications applied for DNA extraction from 200mg stool samples:

The volume of Buffer ASL (Qiagen, Germany) was increased to 1800µl in accordance with the manufacturer's recommendation (due to increased biomass relative to gFOBT or FIT). At the lysis step, 1000µl of liquid was transferred to the pathogen lysis tube (S) (Qiagen, Germany). Samples were then processed according to the protocol; elution was with 50µl rather than 100 µl of water, in order to be consistent with the 'FIT experiment' FIT samples.

2.3.2.5 Automated DNA extraction

The QIAcube HT instrument (Qiagen, Germany) has the capacity to perform automated DNA extraction of 96 samples, with a hands-on set-up time of approximately two hours and a run time of approximately two hours; in contrast manual extraction is limited to 24 samples and takes approximately seven hours hands-on laboratory work. Automated DNA extraction has the potential to save time and reduce the likelihood of error; elution is into a 96 well plate, which saves time at the subsequent library preparation step.

Initial sample preparation was as previously described (i.e. dissection of gFOBT samples; precipitation of a faecal pellet from FIT samples). Samples were transferred to a 96-well plate. The DNeasy 96 PowerSoil Pro QIAcube HT Kit (Qiagen, Germany) was used according to the manufacturer's instructions. Sample lysis was performed using PowerBead Pro Plates (Qiagen, Germany) on a TissueLyserII (Qiagen, Germany). Elution volume was 80µl.

2.3.2.6 Quantification and storage of extracted DNA

Manually extracted DNA was quantified using a NanoDrop-1000 spectrophotometer (Thermo Fisher Scientific Incorporated, UK). Originally DNA was stored in the cold room at 4°C (following the laboratory's original DNA extraction and storage protocol). This was later reviewed and a decision made to

store extracted DNA at -20°C from July 2018, as there is some evidence that DNA, eluted in water, shows greater long-term stability at -20°C (457).

DNA extracted using the QIAcube HT instrument was quantified using the Quant-iT 'dsDNA Assay Kit, broad range' (Invitrogen, Thermo Fisher Scientific, USA) as this is compatible with samples in a 96 well plate. 198µl master-mix (comprising 197µl Quant-iT Buffer and 1µl Quant-iT dsDNA BR reagent) was added to the wells of an opaque 96 well plate (Costar, Life Sciences, USA). 2µl of extracted DNA was added. Fluorescence was read using a microplate fluorometer (Fluoroskan Ascent, Thermo Fisher Scientific, USA) and a standard curve, constructed using the kit's DNA standards, was used to calculate DNA concentration.

2.3.3 PCR amplification and library preparation

2.3.3.1 Changing from the laboratory's existing PCR amplification and library preparation methodology to the Earth Microbiome Project (EMP) protocol

The laboratory's original protocol for PCR amplification and library preparation for microbiome analysis was devised 'in-house' using a combination of NEBNext kits (New England BioLabs, France), as a standardised protocol did not exist at that time. This protocol was followed until February 2018, when the EMP 16S Illumina Amplicon methodology was explored (403). The EMP protocol is open-access and is being used by an increasing number of microbiome studies in an effort to achieve standardisation (enabling cross-study validation, comparison and meta-analysis). An additional advantage is that the EMP protocol comprises a single PCR amplification (as the Illumina adapter is contained within the 16S V4 primers), whereas the laboratory's original protocol comprised three thermocycler-based reactions and two bead-based purification steps. The EMP protocol is therefore significantly cheaper, quicker and reduces the likelihood of error.

After confirming that microbiome analysis could be successfully performed on NHSBCSP gFOBT samples using the EMP methodology, a decision was made to convert to the EMP protocol and re-process the existing samples (as sample preparation has a considerable effect on microbiome results). Comparison between the two methods was not performed as the profound effect of differences

in methodology on microbiome results is well-documented (as outlined in section 1.4.1) and the 'in-house' methodology was not used by other groups, such that a comparison would be of no benefit.

2.3.3.2 Minor modifications to the EMP PCR amplification and library preparation protocol

The EMP 16S Illumina Amplicon protocol comprises the following steps (403):

- Dilution of extracted DNA
- PCR amplification of the V4 region of the 16SrRNA gene
- Visualisation of the PCR amplicons by gel electrophoresis
- Quantification of the PCR amplicons
- Pooling of 240ng of each PCR amplicon
- Purification of the pools
- Quantification of the pools
- Pooling prior to NGS

This protocol was followed using a 96-well Thermocycler (C1000 Touch, Bio-Rad, UK) with the following minor modifications (Table 6):

Table 6. Modifications to the EMP 16S Illumina Amplicon protocol.

EMP protocol	Modification	Rationale
Starting amount of DNA not specified (only a volume of 1µl).	For each sample, 3µl of extracted DNA was diluted to 20ng/µl with UV irradiated molecular grade water. 1µl (20ng) was used per PCR reaction.	This was the starting amount used in the existing laboratory protocol. It represents an optimal amount; PCR can fail if there is too little DNA or too high a concentration of inhibitors.
Samples are amplified in triplicate (3x25µl reaction volumes) and subsequently pooled.	Samples are amplified singly (1x25µl reaction volume).	Amplicons from a single reaction are of sufficient concentration for subsequent processing: this saves time and money and reduces the likelihood of error.
0.5µl of 10µM Reverse primer is required per PCR reaction.	1µl of 5µM Reverse primer was used per PCR reaction.	This reduces pipetting error.
Amplicons are visualised by gel electrophoresis. The existing laboratory protocol required three separate gels to visualise a 96 well plate.	A 96-well gel electrophoresis tank (Alpha Laboratories, UK) was obtained.	This saves time and money, and reduces the likelihood of error.
Amplicons are quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Thermo Fisher Scientific, USA).	Amplicons are quantified using the Quant-iT ‘dsDNA Assay Kit, broad range’ (Invitrogen, Thermo Fisher Scientific, USA).	Equivalent methodology; the Quant-iT ‘dsDNA Assay Kit, broad range’ was already available in the laboratory.

EMP protocol	Modification	Rationale
Amplicon pools are purified using UltraClean PCR Clean-Up Kit (MoBio, Qiagen, Germany).	Amplicon pools are purified using MinElute PCR Purification kit (Qiagen, Germany) and eluted with 2x10µl Buffer EB (Qiagen, Germany).	Equivalent methodology; the MinElute PCR Purification kit was already available in the laboratory. The double-elution step ensures sufficient volume for subsequent quantification and pooling.
Amplicon pools are quantified using a NanoDrop-1000 spectrophotometer (Thermo Fisher Scientific Incorporated, UK).	Amplicon pools are quantified using a 2200 TapeStation (Agilent Technologies, USA).	The TapeStation quantifies the DNA concentration more accurately and allows visual inspection of the pool (for identification of adaptor peaks).

The EMP 16S Illumina Amplicon protocol provides details of 960 unique V4 16SrRNA Forward primers. An additional 575 unique Forward primers were designed by Dr Wood, using the following criteria (inferred from the existing EMP primers) (Table 7):

Table 7. Criteria used to design additional V4 16SrRNA Forward primers.

EMP primers	Inferred criteria for primer design	Rationale
The maximum number of consecutive matching bases/primer was three.	Maximum of three consecutive matching bases.	Too many consecutive matching bases may cause slippage during PCR amplification or errors during NGS.
The GC content of the EMP primers was normally distributed.	No criteria set.	GC-rich regions have higher melting points than AT-rich regions leading to increased specificity of binding.
The EMP primers were twelve bases long; none of the primers had more than seven identically positioned bases compared to any of the other primers.	Maximum of seven identically positioned bases to be shared by any of the primers in the set.	The fewer the number of identically positioned bases shared by primers, the less likely that an error at PCR or NGS will be interpreted as a different primer.
The EMP primers had a normal distribution of 'offset base' scores per primer.	No criteria set.	The offset base score is determined by calculating the number of identically positioned bases shared between primers if the first base of one primer was omitted or repeated (due to PCR errors).

Primers were re-suspended with molecular grade water to the concentrations specified by the EMP protocol and stored at -20°C.

2.3.4 Controls

Potential contamination was assessed at each stage of laboratory processing using the following controls (Table 8):

Table 8. DNA extraction and PCR amplification controls.

DNA extraction	
Control	
Negative	no template
Negative	blank gFOBT card
Negative	blank FIT
Negative	UV irradiated molecular biology grade water (Fisher Scientific, USA)
PCR amplification	
Control	
Positive	20ng *E. coli* strain B DNA (Sigma-Aldrich, USA)
Negative	Microbial DNA-free water (Qiagen)

2.3.5 Pooling and sequencing of libraries

An equal amount of each library was used to create a single pool for sequencing. The laboratory's original protocol used the Illumina MiSeq. This limited the maximum number of samples per sequencing run to 400 and a run-to-run discrepancy in microbiome results was observed. A decision was therefore made to instead use the Illumina HiSeq; this enables 1500 samples to be sequenced per run, mitigating the risk of run-to-run variation. In order to assess the impact of run-run variability of the Illumina HiSeq on microbiome results, a subset of libraries (n=145) were sequenced on two separate Illumina HiSeq runs. This was to determine whether results from one sequencing run could be compared with results from an alternative sequencing run.

Sequencing was performed by the Sequencing Facility at the University of Leeds. The sequencing team followed the methods described by the EMP protocol and used the HiSeq 3000/4000 paired end cluster kit (Illumina Inc. USA) and the HiSeq 3000/4000 SBS kit (300 cycles) (Illumina Inc. USA) with a ten base pair index, according to the manufacturer's protocol. The initial sequencing run failed

(only 5% of reads passed quality filters) due to inadequate PhiX. PhiX is spiked into the sequencing reaction to increase nucleotide diversity so that the location of clusters can be accurately identified. The run was successfully repeated using 20% PhiX (67% of reads passed quality filters) and a subsequent run was performed using 50% PhiX (76% of reads passed quality filters). A total of 1518 samples were sequenced on the first NGS run and 996 samples were sequenced on the second NGS run.

2.3.6 Bioinformatic and statistical analysis

Reads were stripped of adaptor sequences using cutadapt (458). Quality plots were examined and a decision was made to truncate the final five base pairs of reads owing to low quality; this was performed using the DADA2 (459) package within QIIME2 (version 2019.4) (460). Reads were filtered, de-noised, merged as pairs and representative sequences were chosen using the DADA2 package within QIIME2. Samples with fewer than 10,000 reads were removed from subsequent analysis.

Shannon index (461) alpha diversity was calculated and significance was assessed by the Kruskal-Wallis test (462) in QIIME2. Bray-Curtis beta diversity (463) was performed and plotted as principal co-ordinate analysis (PCA) plots (464) in QIIME2. The significance of differences in beta diversity between groups was assessed by PERMANOVA analysis (465) performed using the Adonis package (466) within QIIME2 with 9999 iterations. Taxa which differed significantly according to metadata groups were identified using MaAsLin2 (Multivariate Association with Linear Models) (467).

Taxa were assigned to the representative sequences by the QIIME2 feature classifier (468), using the BLAST+ algorithm (469), aligning sequences against the SILVA version 132 99% similarity database (470-472). Taxa which differed significantly between groups were identified using LEfSe (Linear discriminant analysis Effect Size)(389). PCA plots and taxonomy bar plots generated in QIIME2 were re-plotted using R (version 3.5.0).

Bland-Altman plots (473, 474) were generated in R (version 3.5.0) using the package blandr (475); histograms of differences were checked visually to confirm a normal distribution.

2.4 Results

2.4.1 Summary of sample processing and sequencing

The majority of samples were successfully sequenced on the first attempt. The following samples were processed on the first sequencing run (Figure 9):

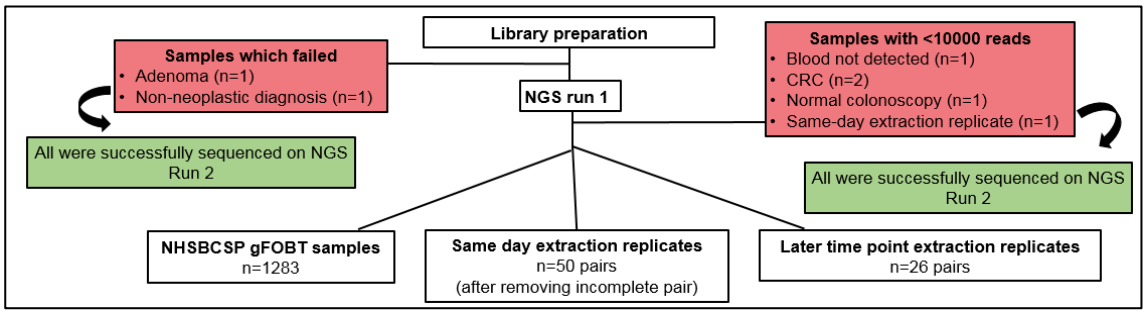


Figure 9. Workflow of samples processed on the first sequencing run.

Two samples failed library preparation due to an inadequate concentration of PCR amplicons. The extracted DNA from these samples underwent successful repeat PCR amplification, library preparation and sequencing on NGS run 2, indicating that the fault was technical.

Five samples had fewer than 10,000 reads and were deemed to have failed sequencing. The extracted DNA from these samples underwent successful repeat PCR amplification, library preparation and sequencing on NGS run 2, indicating that the fault was technical.

The following samples were processed on the second sequencing run (Figure 10):

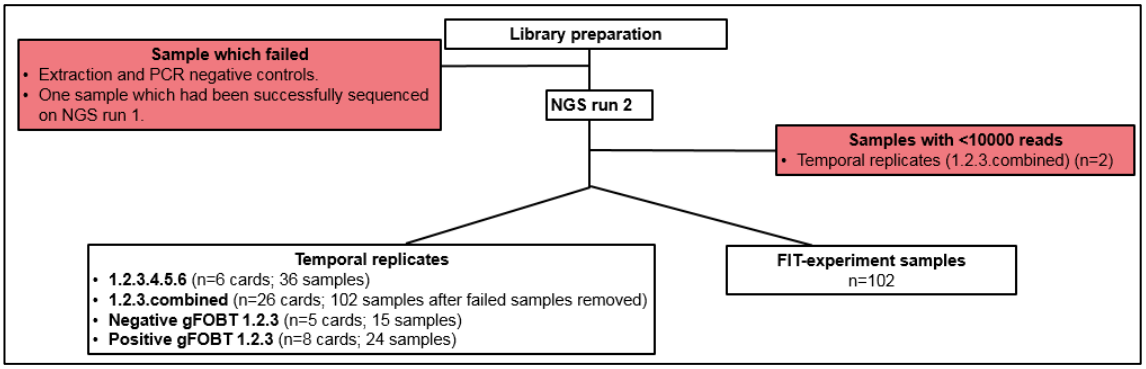


Figure 10. Workflow of samples processed on the second sequencing run.

The extraction-negative controls and PCR-negative controls had insufficient concentration of PCR amplicons to sequence. One sample had a low concentration of PCR amplicon (13ng/µl) prior to the first NGS run. As there was concern that it might fail sequencing, the extracted DNA was re-processed. However the sample was successfully sequenced on NGS run 1; the concentration of PCR amplicon remained ~ 13ng/µl on repeated library preparation, therefore the sample was not re-sequenced. Two samples had fewer than 10,000 reads and were deemed to have failed sequencing. An attempt will be made to sequence these samples on a subsequent NGS run.

2.4.1.1 NHSBCSP gFOBT samples

The number of reads/sample is displayed in Figure 11 Four samples had fewer than 10,000 reads and were removed from subsequent analysis. With these samples removed, the range of read counts/sample was 13,000-223,000 with a median of 80,000.

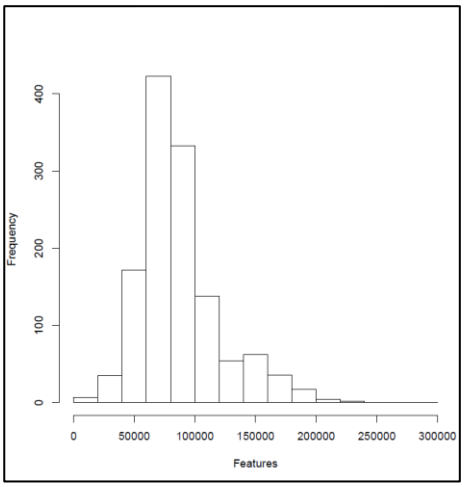


Figure 11. Number of reads/sample for the NHSBCSP samples. The number of reads (features) is plotted on the x axis; the y axis indicates the number of samples.

2.4.1.2 Samples which were sequenced on two sequencing runs

The cumulative number of reads/sample is displayed in Figure 12 and number of reads/sample for each NGS run in Figure 13. The cumulative range of read counts/sample was 26,000-155,000 with a median of 92,000.

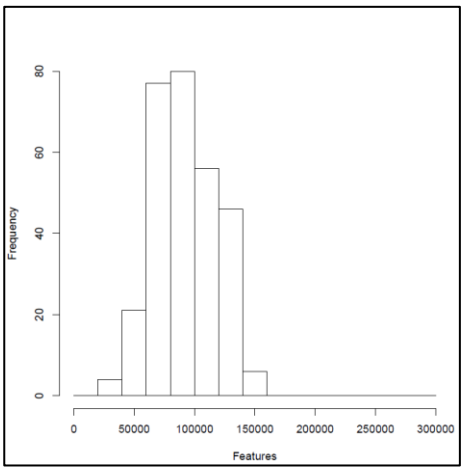


Figure 12. Cumulative number of reads/sample for samples which were sequenced on two sequencing runs. The number of reads (features) is plotted on the x axis; the y axis indicates the number of samples.

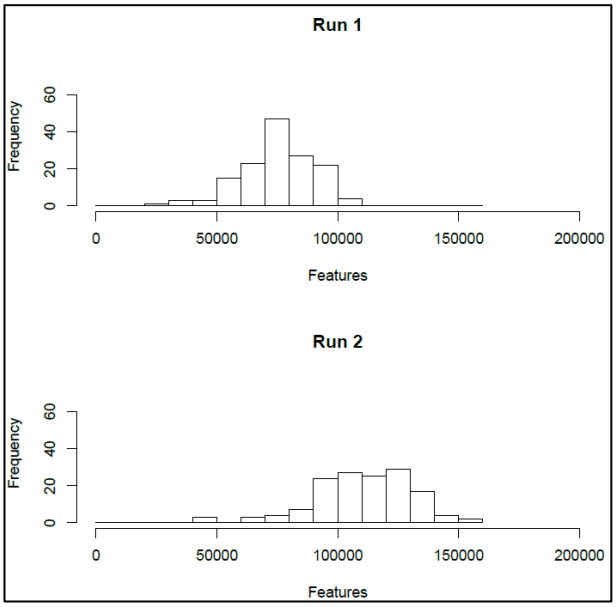


Figure 13. Number of reads/sample/NGS run for samples which were sequenced on two sequencing runs. The number of reads (features) is plotted on the x axis; the y axis indicates the number of samples.

2.4.1.3 Extraction replicate samples

The number of reads/sample is displayed in Figure 14. The range of read counts/sample was 17,000-188,000 with a median of 78,000.

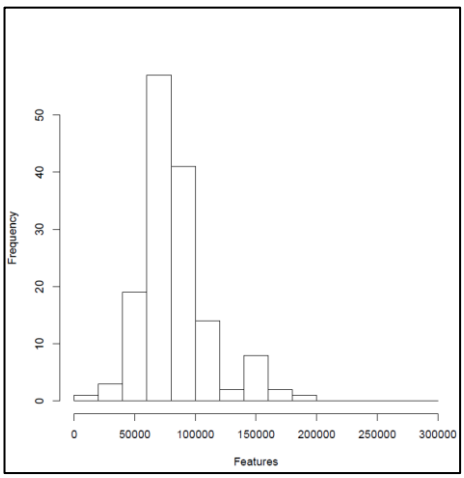


Figure 14. Number of reads/sample for the extraction replicate samples. The number of reads (features) is plotted on the x axis; the y axis indicates the number of samples.

2.4.1.4 Temporal replicates

The number of reads/sample is displayed in Figure 15. Two of the ‘temporal 1.2.3.combined’ samples had fewer than 10,000 reads and were removed from subsequent analysis. With these samples removed, the range of read counts/sample was:

- 69,000-176,000 (median 114,000) for ‘temporal 1.2.3.4.5.6’ samples
- 59,000-192,000 (median 123,000) for ‘temporal 1.2.3.combined’ samples
- 62,000-162,000 (median 129,000) for ‘temporal N 1-3’ samples
- 24,000-153,000 (median 125,000) for ‘temporal P 1-3’ samples'

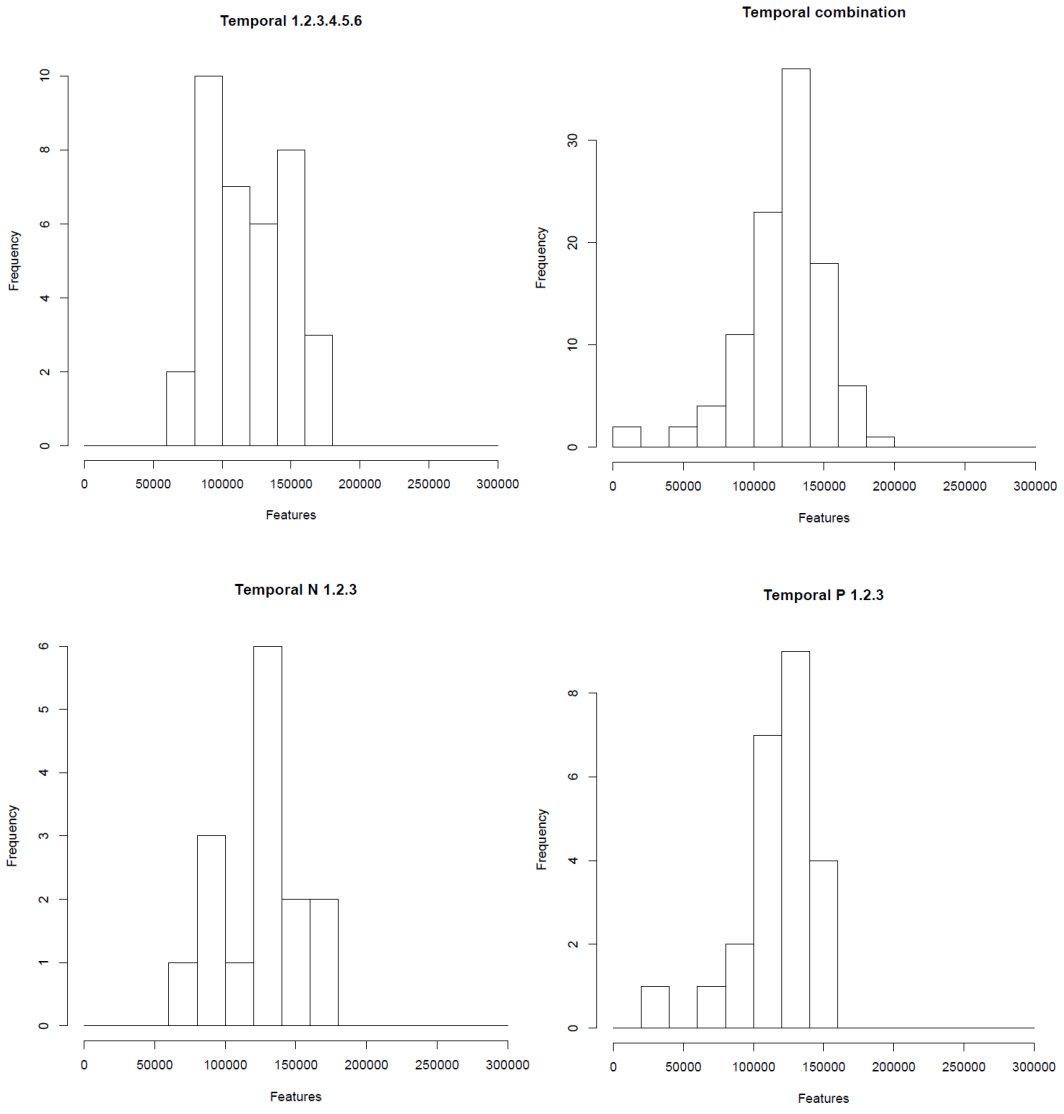


Figure 15. Number of reads/sample for the temporal samples. The number of reads (features) is plotted on the x axis; the y axis indicates the number of samples. ‘Temporal combination’ = ‘Temporal 1.2.3.combined’ samples.

2.4.1.5 'FIT experiment' samples

The number of reads/sample is displayed in Figure 16. The range of read counts/sample was 26,000-284,000 with a median of 147,000.

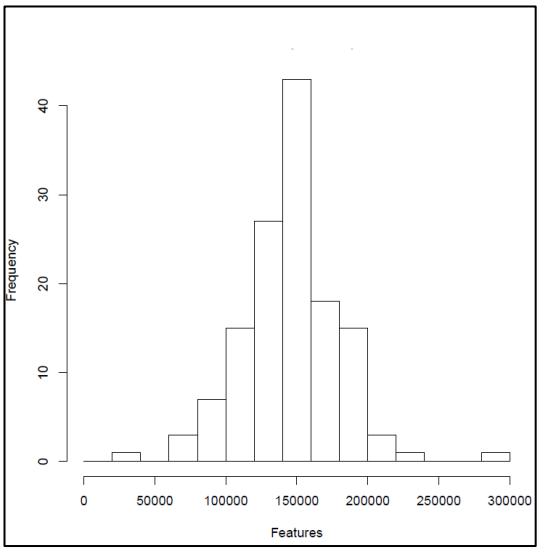


Figure 16. Number of reads/sample for the 'FIT experiment' samples. The number of reads (features) is plotted on the x axis; the y axis indicates the number of samples.

2.4.2 Controls

The concentration of DNA extracted from the extraction-negative controls is detailed in Table 9; it was less than 4ng/µl and of poor quality (Figure 17). For comparison the median extracted DNA concentration of all samples was 45ng/µl; some samples had extracted DNA concentrations as low as the extraction-negative controls, however unlike the extraction-negative controls, these samples were successfully amplified by V4 16SrRNA PCR. This suggests that a proportion of the DNA as measured by the NanoDrop-1000 spectrophotometer (Thermo Fisher Scientific Incorporated, UK) in the extraction-negative controls may be non-bacterial. Blank media from FIT devices has undergone DNA extraction and will be sequenced on the subsequent NGS run.

Table 9. Concentration of DNA extracted from the extraction-negative controls.

Type of sample	Extracted DNA concentration (ng/µl)	
	Minimum	Maximum
No template (n=2)	3.4	3.7
Water (n=2)	0.3	0.4
Blank gFOBT card (n=3)	0	2.3

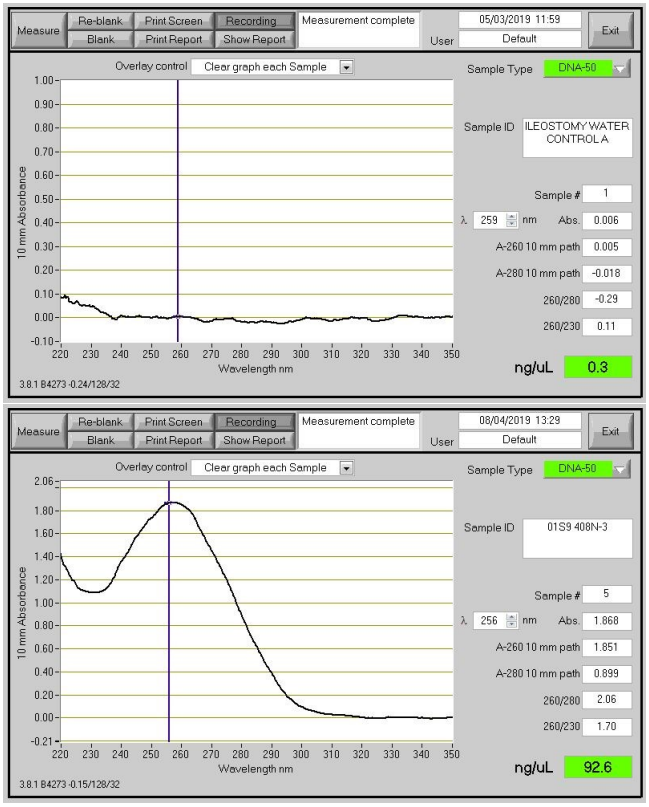


Figure 17. Comparison of typical NanoDrop-1000 spectrophotometer traces of an extraction-negative control and a sample. The upper trace is derived from DNA extracted from an extraction-negative control; the lower trace is derived from DNA extracted from a NHSBCSP gFOBT sample.

The concentration of PCR amplicons of the extraction-negative controls and PCR negative and positive controls is detailed in Table 10 and Figure 18. The maximum PCR amplicon concentration of the negative controls was 5.6 ng/µl. For comparison the minimum concentration of PCR amplicon that was pooled for sequencing was 13.4ng/µl and the median of each plate of PCR amplicons ranged from 29-53ng/µl.

Table 10. Concentration of PCR amplicons from extraction and PCR controls.

Type of sample	PCR amplicon concentration (ng/µl)	
	Minimum	Maximum
No template extraction control (n=2)	3.6	4
Water extraction control (n=2)	3.4	4
Blank gFOBT card extraction control (n=3)	3.2	4.7
Water PCR control (n=24)	1.1	5.6
E. coli PCR control (n=24)	36.5	72.7

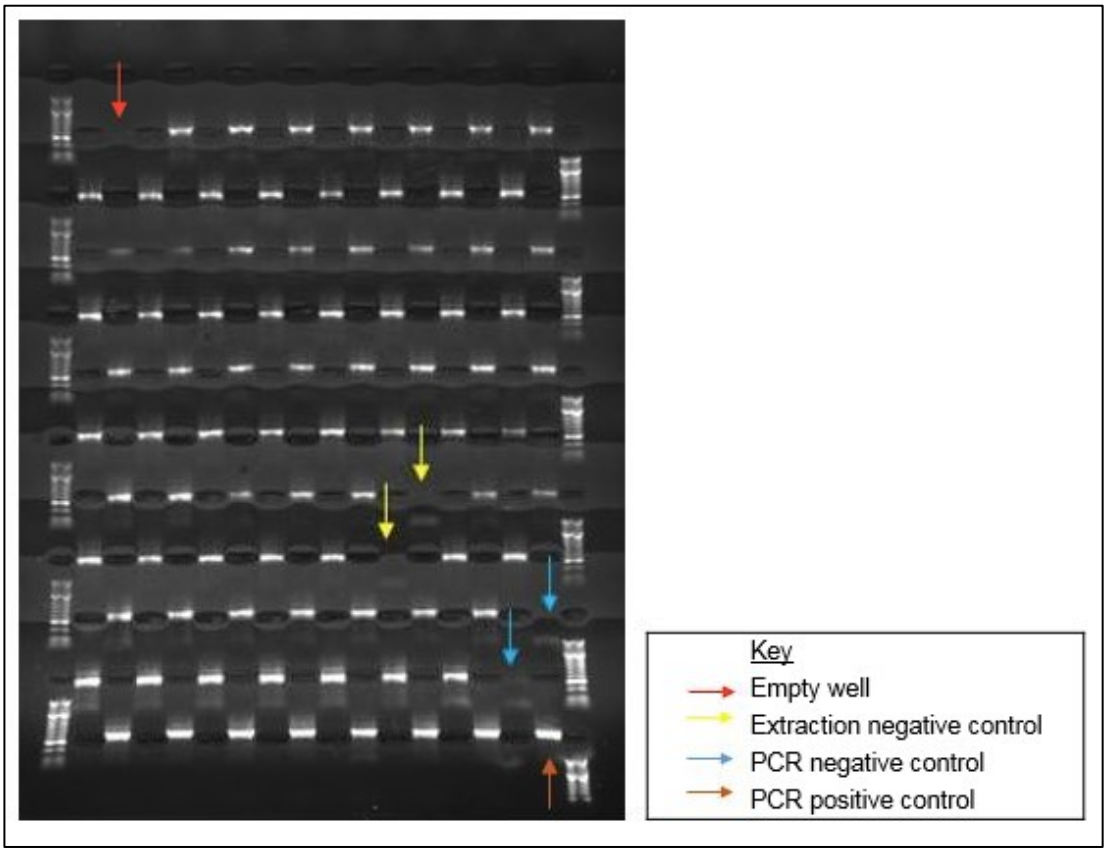


Figure 18. Gel electrophoresis image of PCR amplicons. There is an absence of bands corresponding to wells containing the extraction-negative controls and PCR-negative controls compared with visible bands for samples. (Note: the first two bands of the third row from the top are faint due to pipetting error).

2.4.3 Sequencing run-sequencing run variability

2.4.3.1 Sequencing metrics

Table 11 details the sequencing data for samples (libraries) which were sequenced on both sequencing runs. Samples on NGS run 2 had on average a higher number of reads than their equivalent samples sequenced on NGS run 1. This may have been caused by differences between the total number of samples per run (1518 for NGS run 1 and 996 for NGS run 2) and/or differences between the amount of PhiX per run (20% for NGS run 1 and 50% for NGS run 2). Read numbers per sample demonstrated correlation between sequencing runs, with the majority of samples maintaining their relative position (Figure 19), however as correlation indicates association only, a Bland-Altman plot was generated to assess agreement (Figure 20). This confirms a negative bias in the number of reads/sample measured by NGS run 1 compared with NGS run 2 (mean ~35,000 less) and demonstrates that the range of differences is wide (-15,000 to -55,000) i.e. agreement is poor.

Table 11. Number of reads/sample for samples (libraries) which were sequenced on two sequencing runs. Figures are rounded to the nearest 1000.

Number of reads/sample		
	Run 1	**Run 2**
Minimum	26,000	40,000
Maximum	106,000	155,000
Median	77,000	112,000

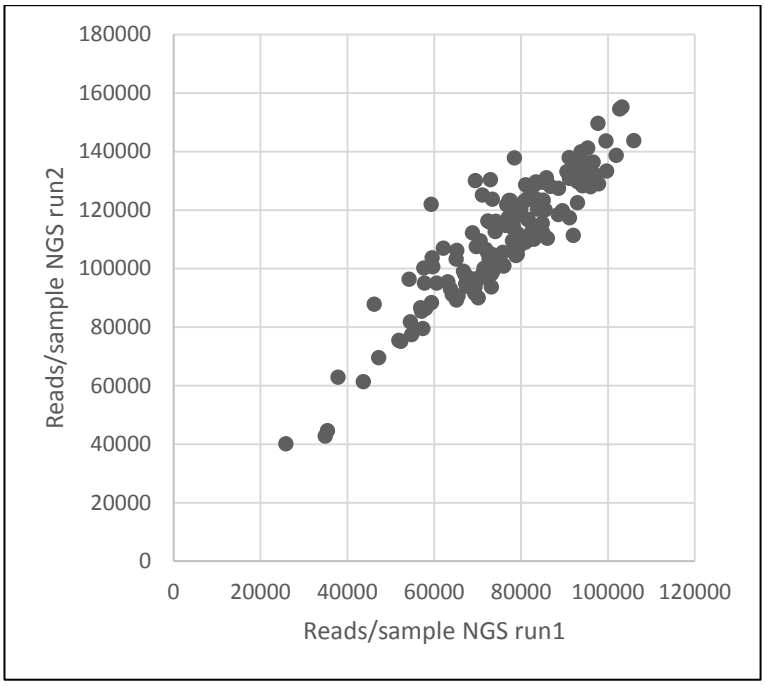


Figure 19. Scatter-plot showing the number of reads/sample for samples (libraries) which were sequenced on two sequencing runs.

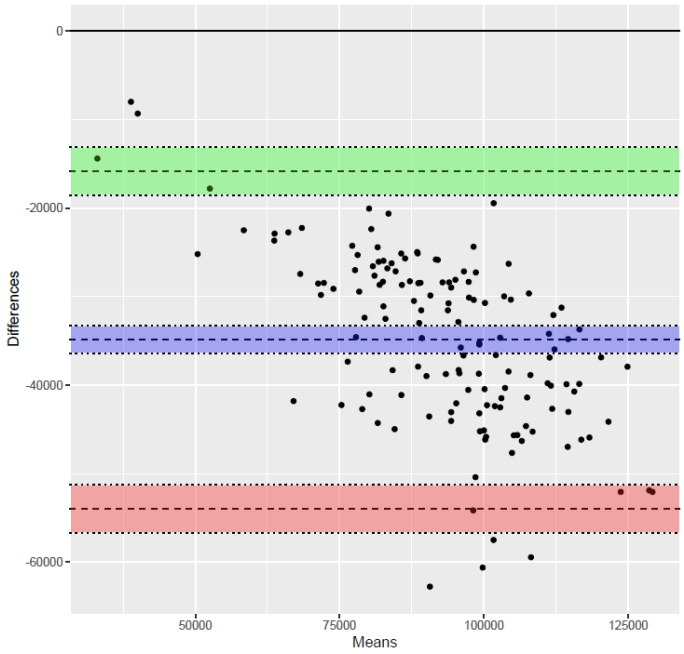


Figure 20. Bland-Altman plot of the number of reads/sample for samples (libraries) which were sequenced on two sequencing runs. The x axis shows the mean of the number of reads/sample across both NGS runs. The y axis shows the difference in the number of reads/sample across both NGS runs. The purple band represents the mean difference (i.e. bias) (plus 95% CI) in the number of reads/sample between NGS run 1 and NGS run 2. The green and red bands represent the upper and lower 95% limits of agreement plus 95% CI. The upper and lower limits of agreement are calculated as mean difference +/-1.96 Standard Deviations (SD); they are expected to contain 95% of the differences measured between both methods.

2.4.3.2 Alpha diversity

No significant difference in Shannon diversity index was detected between runs (Kruskal-Wallis p = 0.24) (Figure 21).

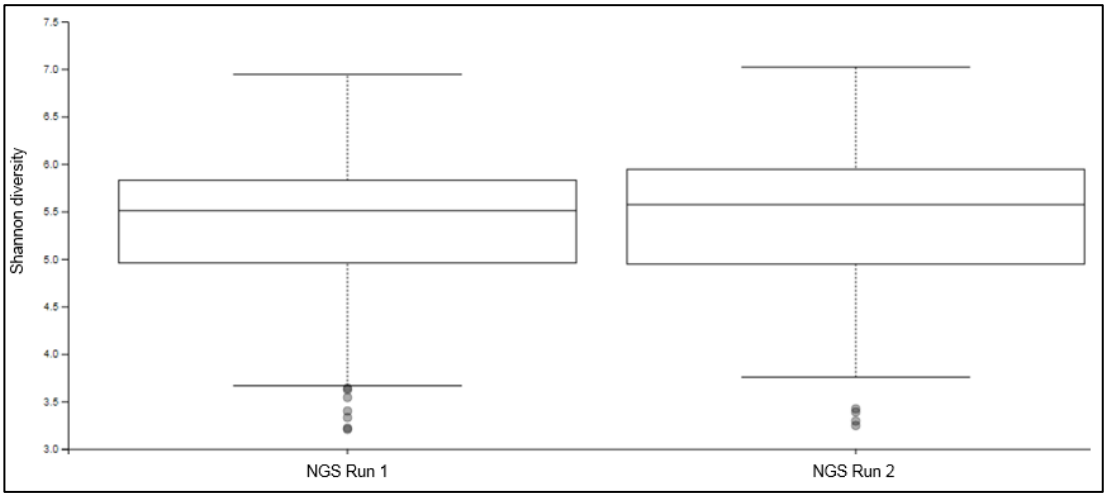


Figure 21. Boxplots of Shannon diversity index for samples (libraries) which were sequenced on two sequencing runs.

2.4.3.3 Beta diversity

Pairs of equivalent samples processed on either NGS run 1 or NGS run 2 clustered together on a PCA plot of Bray-Curtis distances, suggesting that sequencing run causes relatively little change to the microbiome community as measured by Bray-Curtis distance (Figure 22). No significant difference in Bray-Curtis distances was detected between runs (PERMANOVA: $R^2 = 0.002$, $p = 0.91$).

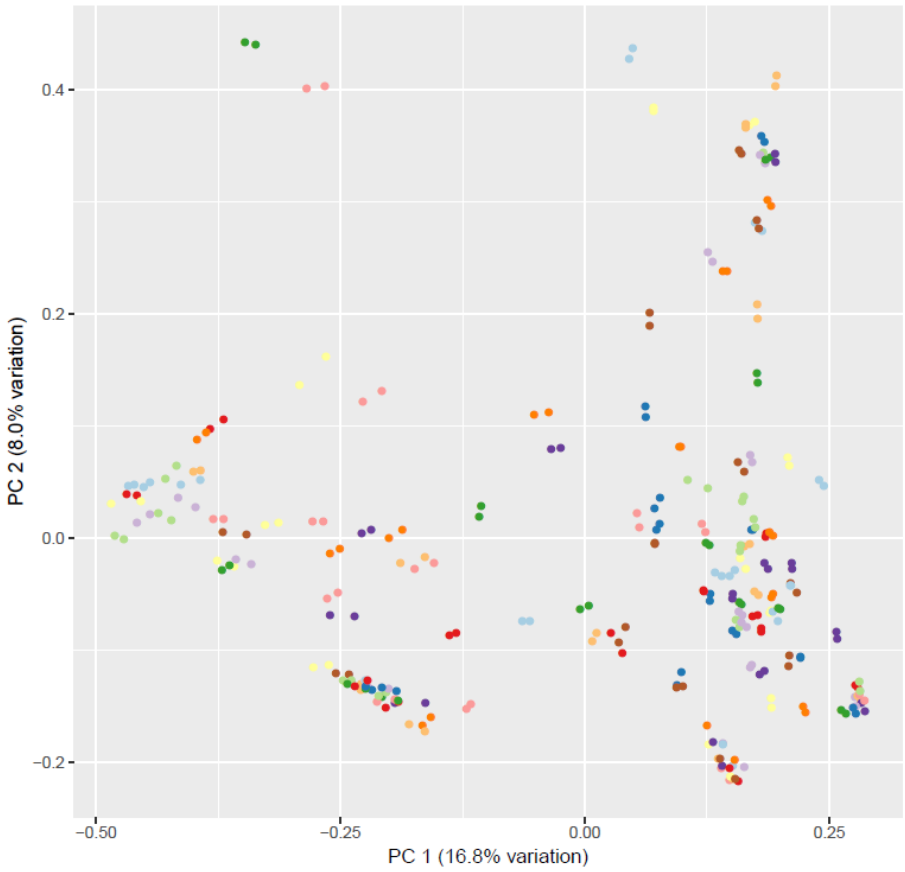


Figure 22. PCA of Bray-Curtis distances for samples (libraries) sequenced on separate NGS runs. Points on the graph are coloured according to gFOBT sample.

2.4.3.4 Taxonomy

The taxonomic composition of sample pairs was very similar but not identical (Figure 23). The relative abundances of CRC-associated bacteria (identified by the Random Forest models described in Chapter 3) were compared for pairs of samples sequenced on NGS run 1 and NGS run 2 (Figure 24). Association (Figure 25) and agreement (Figure 26) between pairs of measurements was high.

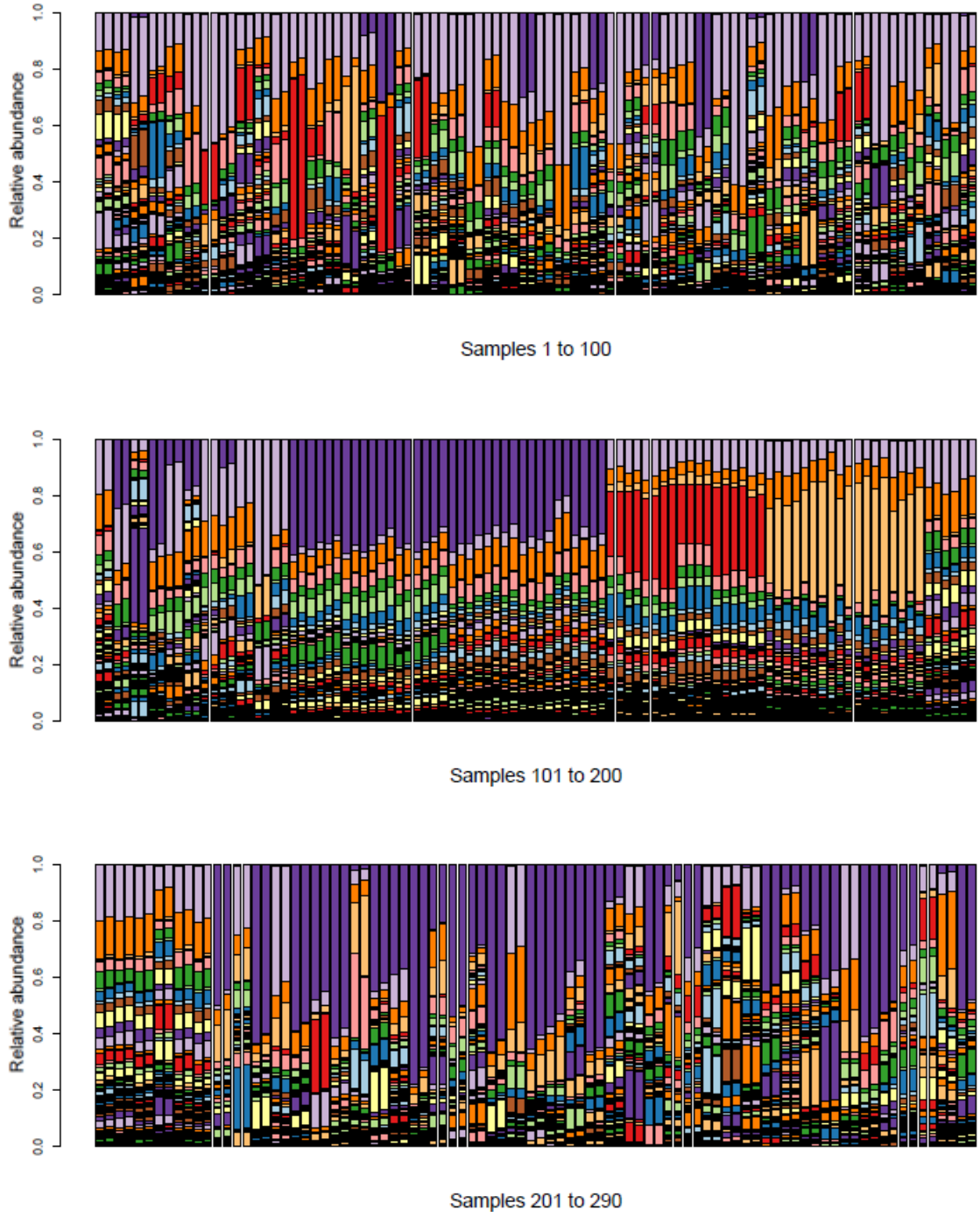
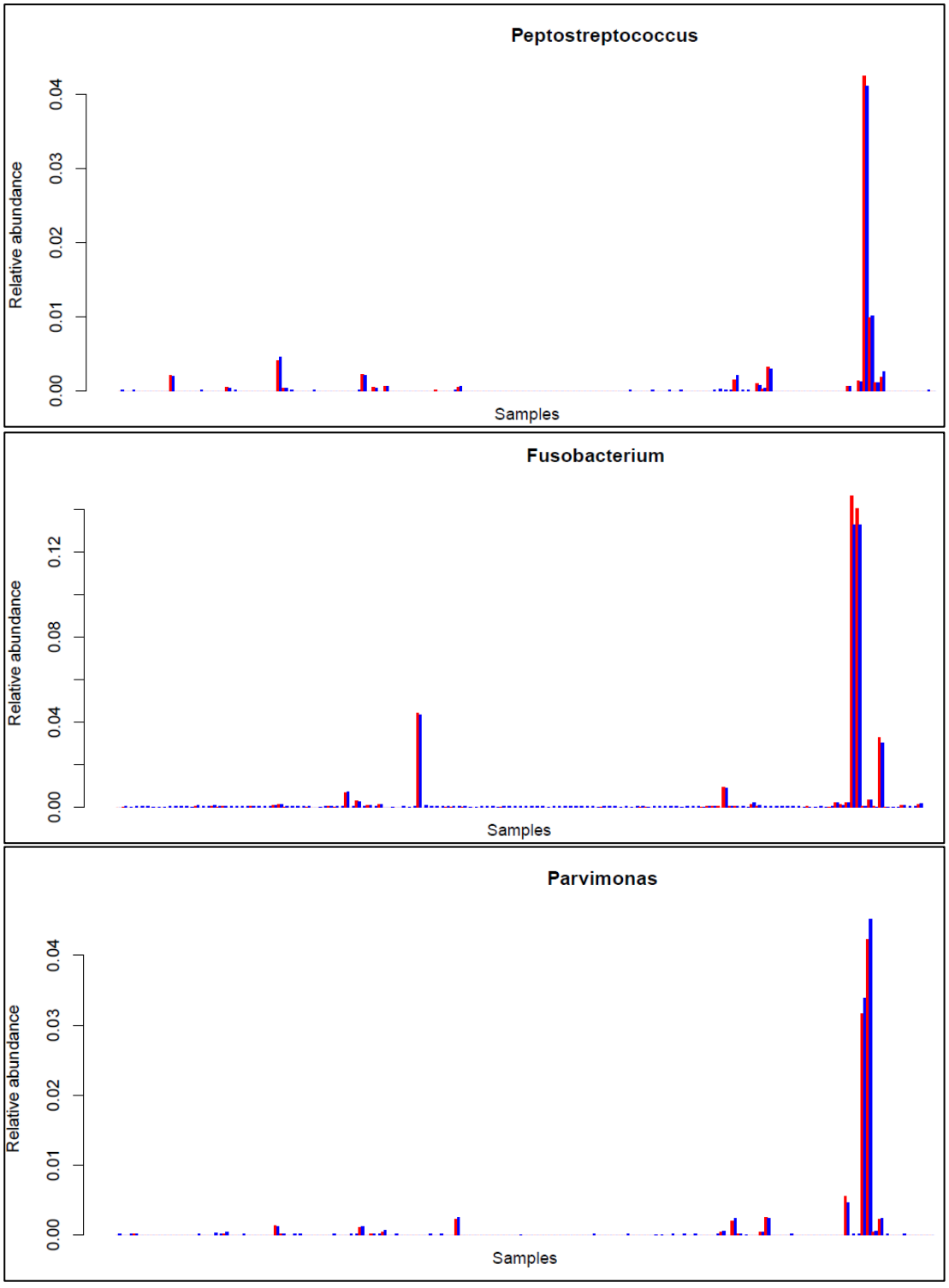


Figure 23. Taxonomy bar charts for samples (libraries) sequenced on separate NGS runs. Each pair of adjacent bars represents a sample sequenced on NGS run 1 (left-hand bar) or NGS run 2 (right-hand bar). Colours denote the relative abundance of taxa (genus level); there are too many to include a legend.


Peptostreptococcus
Relative abundance
0.00
0.01
0.02
0.03
0.04
Samples
Fusobacterium
Relative abundance
0.00
0.04
0.08
0.12
Samples
Parvimonas
Relative abundance
0.00
0.01
0.02
0.03
0.04
Samples

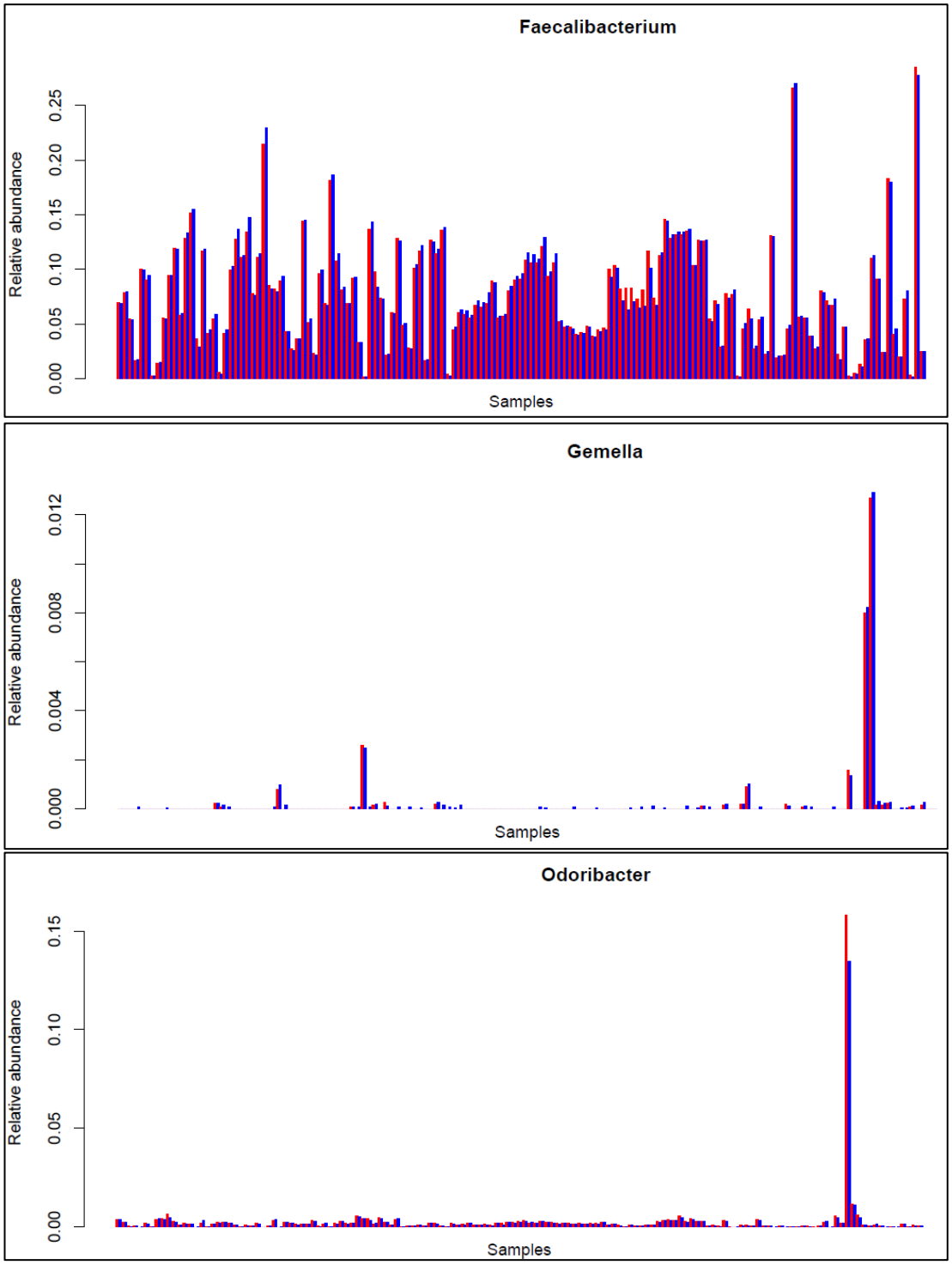


Figure 24. The relative abundance of CRC-associated taxa for samples (libraries) sequenced on separate NGS runs. Each pair of adjacent bars represents a sample sequenced on NGS run 1 (left-hand bar, red) or NGS run 2 (right-hand bar, blue).

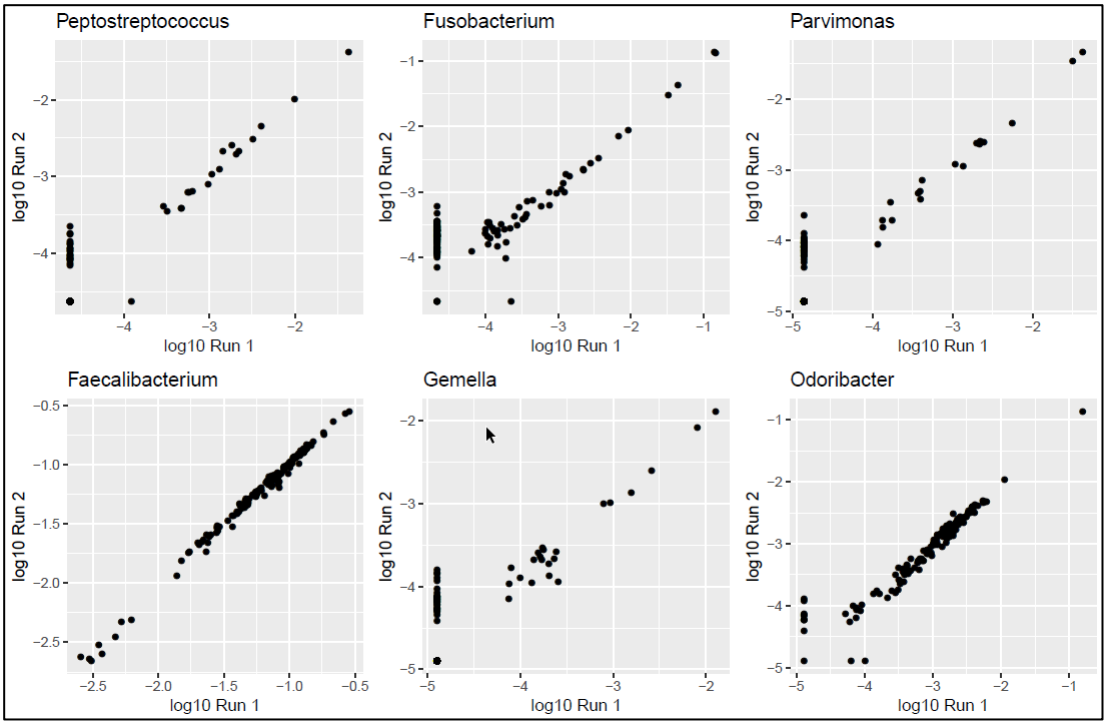


Figure 25. Scatter-plots of CRC-associated taxa for samples (libraries) sequenced on separate NGS runs. Log relative abundance is plotted on the x (NGS run 1) and y (NGS run 2) axes respectively. As it is not possible to log relative abundance values equal to 0, these correspond to the smallest value on each axis (i.e. the bottom left-hand corner of each graph). Points with identical x and y values are overlaid and represented by a single point.

The Bland-Altman plots (Figure 26) indicate an absence of bias for all but *Gemella* (which is small); variation between relative abundances recorded by the two NGS runs is minimal.

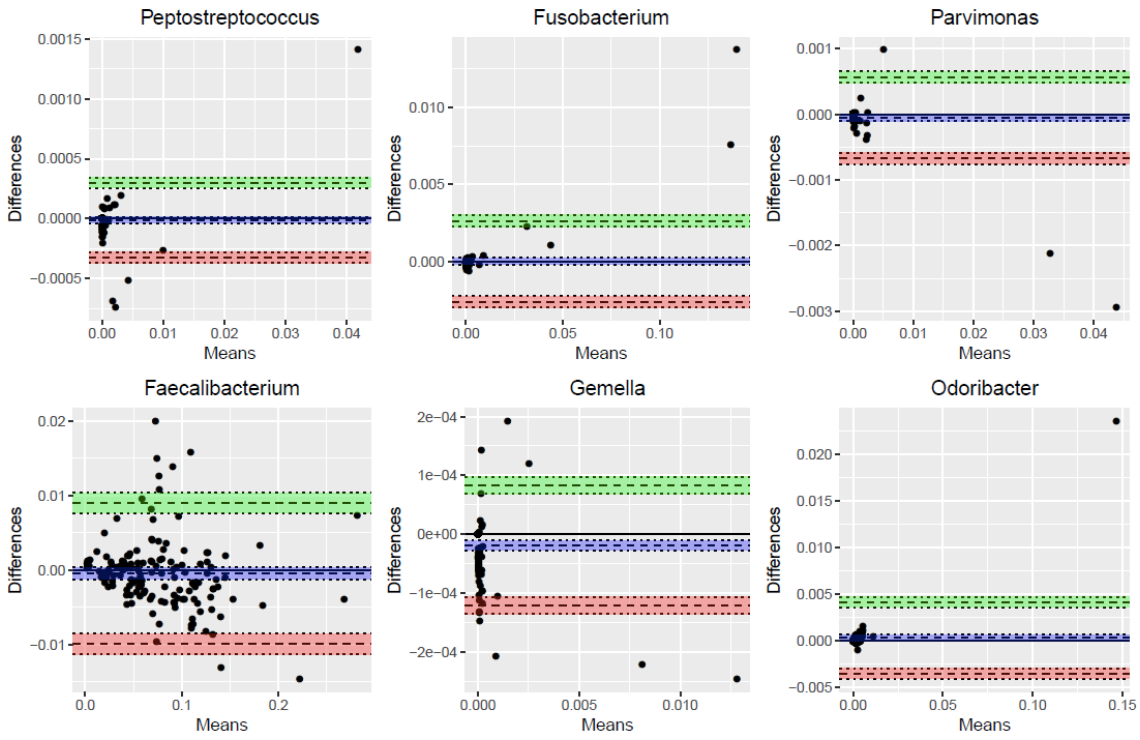


Figure 26. Bland-Altman plots of the relative abundances of CRC-associated taxa between samples (libraries) sequenced on separate NGS runs. The x axis shows the mean of the relative abundance of CRC-associated taxa across both NGS runs. The y axis shows the difference in the relative abundance of CRC-associated taxa across both NGS runs. The purple band represents the mean difference (i.e. bias) (plus 95% CI) in the relative abundance of CRC-associated taxa between NGS run 1 and NGS run 2. The green and red bands represent the upper and lower 95% limits of agreement plus 95% CI. The upper and lower limits of agreement are calculated as mean difference +/- 1.96 SD; they are expected to contain 95% of the differences measured between both methods.

During analysis of the Temporal Replicate samples (described in section 2.4.5), it was observed that the relative abundance of *Escherichia-Shigella* displayed marked variation between certain replicates. Whether the relative abundance of *Escherichia-Shigella* varies between sequencing run was therefore investigated. This was not found to be the case (Figure 27 and Figure 28).

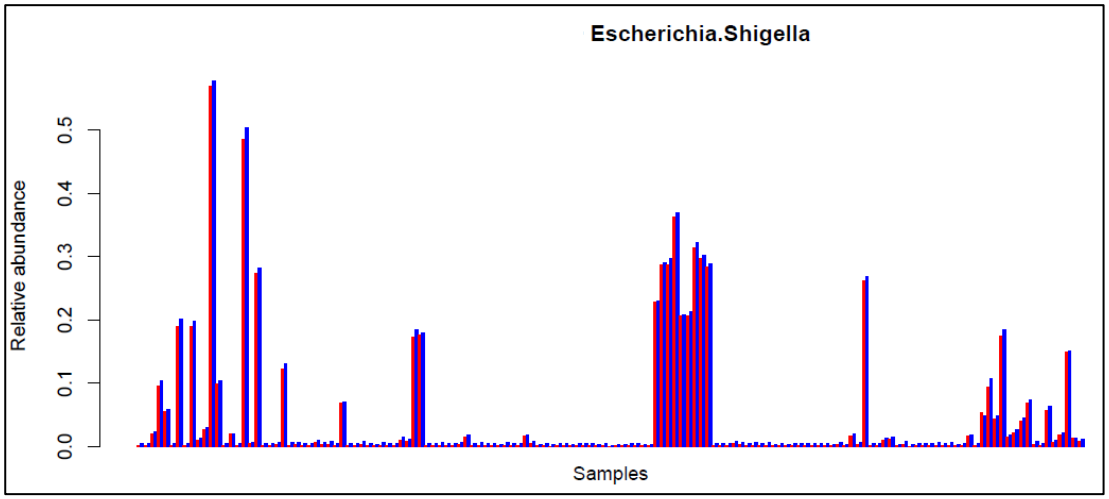


Figure 27. The relative abundance of *Escherichia-Shigella* for samples (libraries) sequenced on separate NGS runs. Each pair of adjacent bars represents a sample sequenced on NGS run 1 (left-hand bar, red) or NGS run 2 (right-hand bar, blue).

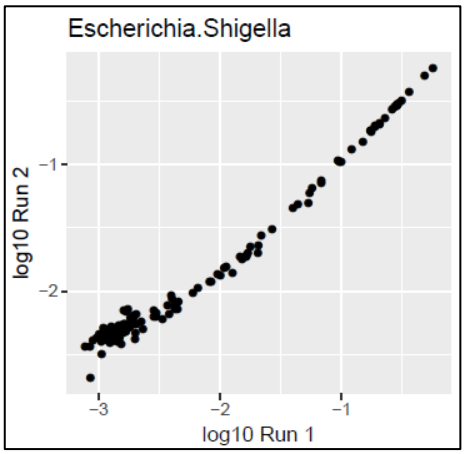


Figure 28. Scatter-plot of the relative abundance of *Escherichia-Shigella* for samples (libraries) sequenced on separate NGS runs. Log relative abundance is plotted on the x (NGS run 1) and y (NGS run 2) axes respectively. As it is not possible to log relative abundance values equal to 0, these correspond to the smallest value on each axis (i.e. the bottom left-hand corner of the graph). Points with identical x and y values are overlaid and represented by a single point.

The Bland-Altman plot (Figure 29) indicates a small bias but minimal variation in relative abundance measured by the two NGS runs.

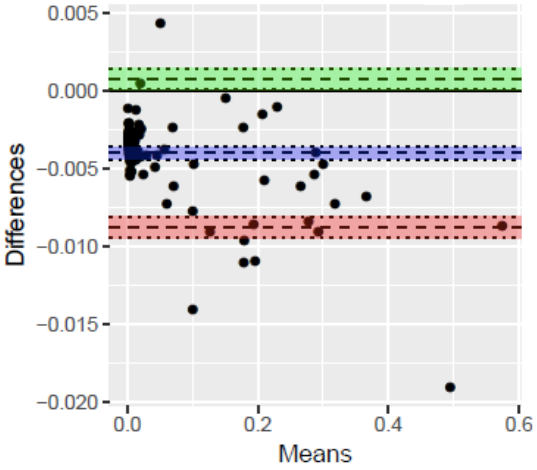


Figure 29. Bland-Altman plot of the relative abundance of *Escherichia-Shigella* for samples (libraries) sequenced on separate NGS runs. The x axis shows the mean of the relative abundance of *Escherichia-Shigella* across both NGS runs. The y axis shows the difference in the relative abundance of *Escherichia-Shigella* across both NGS runs. The purple band represents the mean difference (i.e. bias) (plus 95% CI) in the relative abundance of *Escherichia-Shigella* between NGS run 1 and NGS run 2. The green and red bands represent the upper and lower 95% limits of agreement plus 95% CI. The upper and lower limits of agreement are calculated as mean difference +/- 1.96 SD; they are expected to contain 95% of the differences measured between both methods.

2.4.4 Extraction replicates

Same-day extraction replicates were created to determine whether the choice of which three squares to extract altered the microbiome result. These were denoted with the suffix A or B. An additional set of extraction replicates was prepared to determine whether prolonged storage of samples at room temperature altered the microbiome result. These samples were denoted with the suffix ‘original’ or ‘repeat’.

2.4.4.1 Discovery of outliers

Upon initial data analysis, two outliers were identified on the PCA of Bray-Curtis distances (Figure 30). These outliers were samples 1763P.A/B and 398N.A/B. Figure 30 demonstrates that each member of the two pairs clusters with a member of the alternate pair. Given the high inter-individual variation in the microbiome relative to intra-individual variation, this indicated probable sample labelling error.

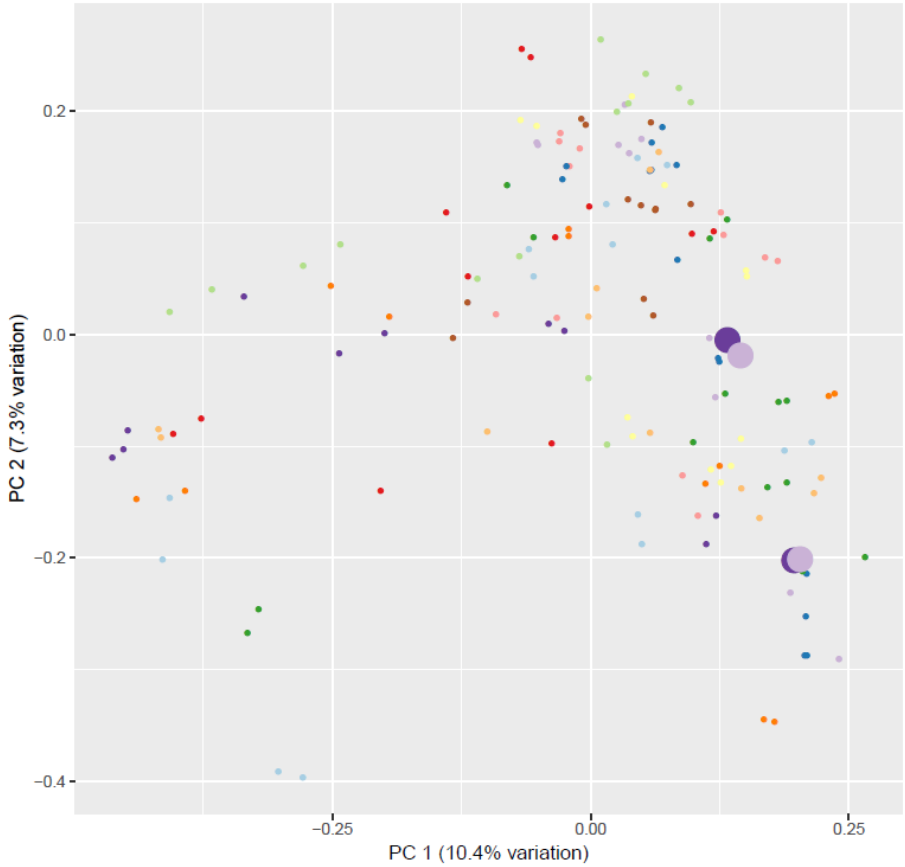


Figure 30. PCA of Bray-Curtis distances for extraction replicates. Points on the graph are coloured according to sample. Points corresponding to replicates of samples 1763P and 398N are enlarged for ease of identification.

In order to determine whether the error occurred at the time of DNA extraction or during subsequent laboratory processing, the original extracted DNA was re-processed and underwent sequencing on NGS run 2. Figure 31 demonstrates the taxonomic composition of the four samples. It indicates that each member of the two pairs is more similar to a member of the alternate pair than its partner, suggesting that the error occurred at the time of DNA extraction. These samples were removed from subsequent analysis. The need to pay particular attention to sample labelling was re-iterated; automated DNA extraction would minimise this risk.

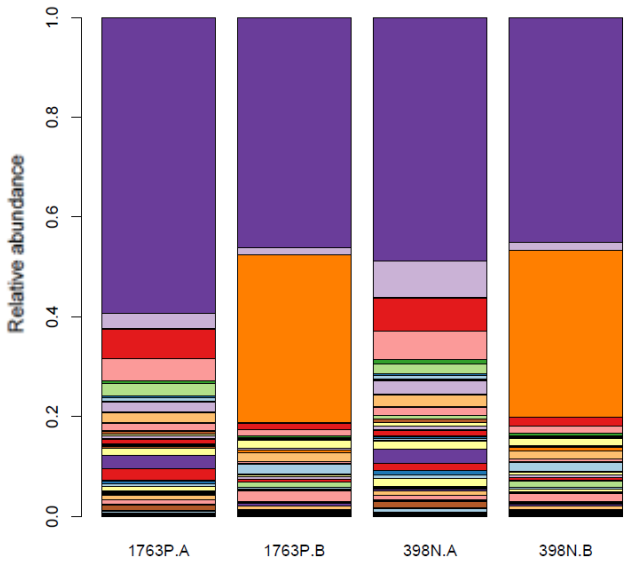
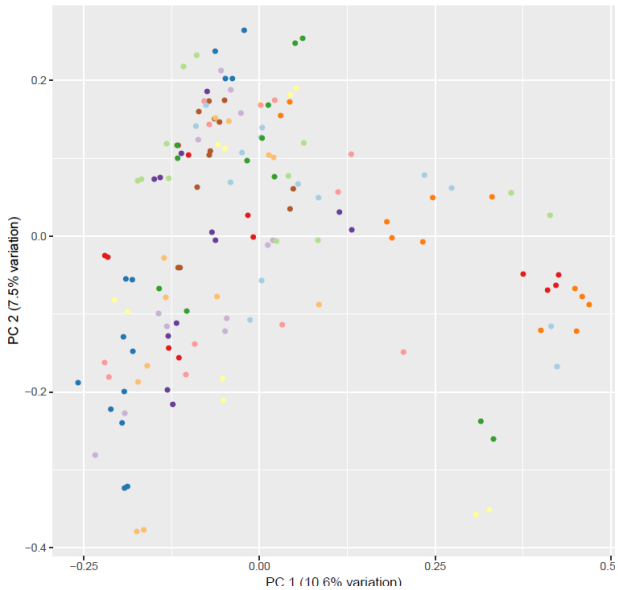
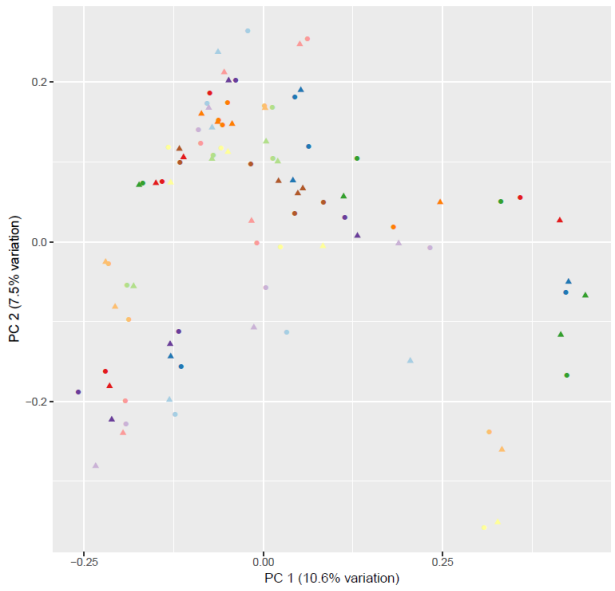


Figure 31. Taxonomy bar chart of samples 1763P.A/B and 398N.A/B. Colours denote the relative abundance of taxa (genus level); there are too many to include a legend.

2.4.4.2 Beta diversity

PCA of Bray-Curtis distances demonstrated that the majority of replicate pairs clustered closely together, indicating high similarity in microbiome community structure (Figure 32). Interestingly, a minority of replicate pairs were separated to a greater degree than most on the PCA plot; however, this did not occur more frequently for replicate pairs whereby replicates were extracted after a period of storage at ambient temperature, compared with replicate pairs whereby replicates were extracted at the same time. This is confirmed by Figure 33, which demonstrates a similar range of Bray-Curtis distances within pairs of samples whereby replicates were extracted after a period of storage at ambient temperature, compared with replicate pairs whereby replicates were extracted at the same time. This suggests that variability in the microbiome between replicate pairs is not due to prolonged storage at ambient temperature, but is due to other sources of variation common to both groups e.g. stool subsampling or other sources of technical variation. For both groups, Bray-Curtis distances within pairs of samples were markedly smaller than Bray-Curtis distances between unrelated samples, indicating that intra-individual microbiome community structure was preserved.


0.2
0.0
−0.2
−0.4
PC 2 (7.5% variation)
−0.25
0.00
0.25
0.5
PC 1 (10.6% variation)


0.2
0.0
−0.2
PC 2 (7.5% variation)
−0.25
0.00
0.25
PC 1 (10.6% variation)

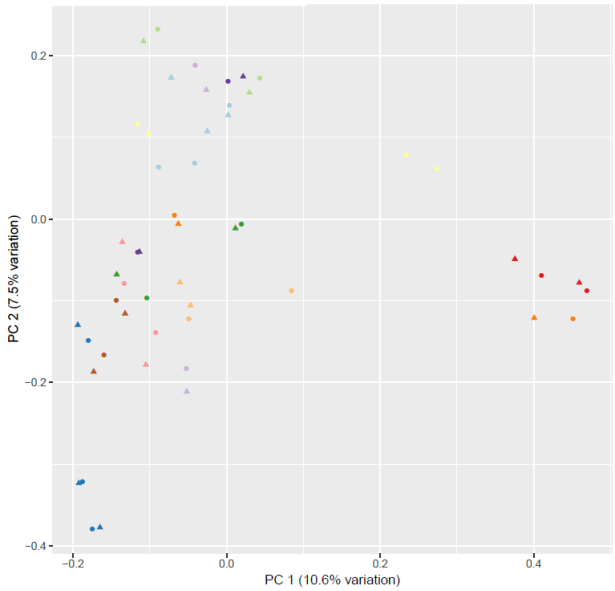

0.2
0.0
−0.2
−0.4
PC 2 (7.5% variation)
−0.2
0.0
0.2
0.4
PC 1 (10.6% variation)

Figure 32. PCA of Bray-Curtis distances for extraction replicates. Points on the graph are coloured according to gFOBT sample. *Upper plot:* all samples. *Middle plot:* samples whereby replicates were extracted at the same time. *Lower plot:* samples whereby replicates were extracted after a period of storage at ambient temperature (triangles represent the replicates which were extracted after a period of storage.)

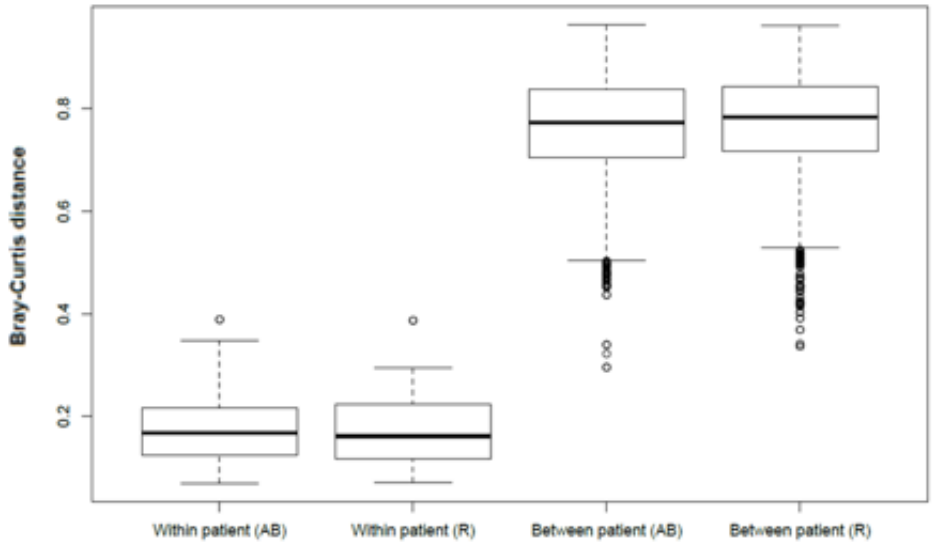


Figure 33. Box plots of Bray-Curtis distances for extraction replicates. The two left-hand boxplots depict the range of Bray-Curtis distances within pairs of replicates whereby samples were extracted at the same time (AB) or after a period of storage at ambient temperature (R). The two right-hand boxplots depict the range of Bray-Curtis distances between all of the samples within each group respectively.

2.4.4.3 Taxonomy

The taxonomic composition of sample pairs was very similar (Figure 34).

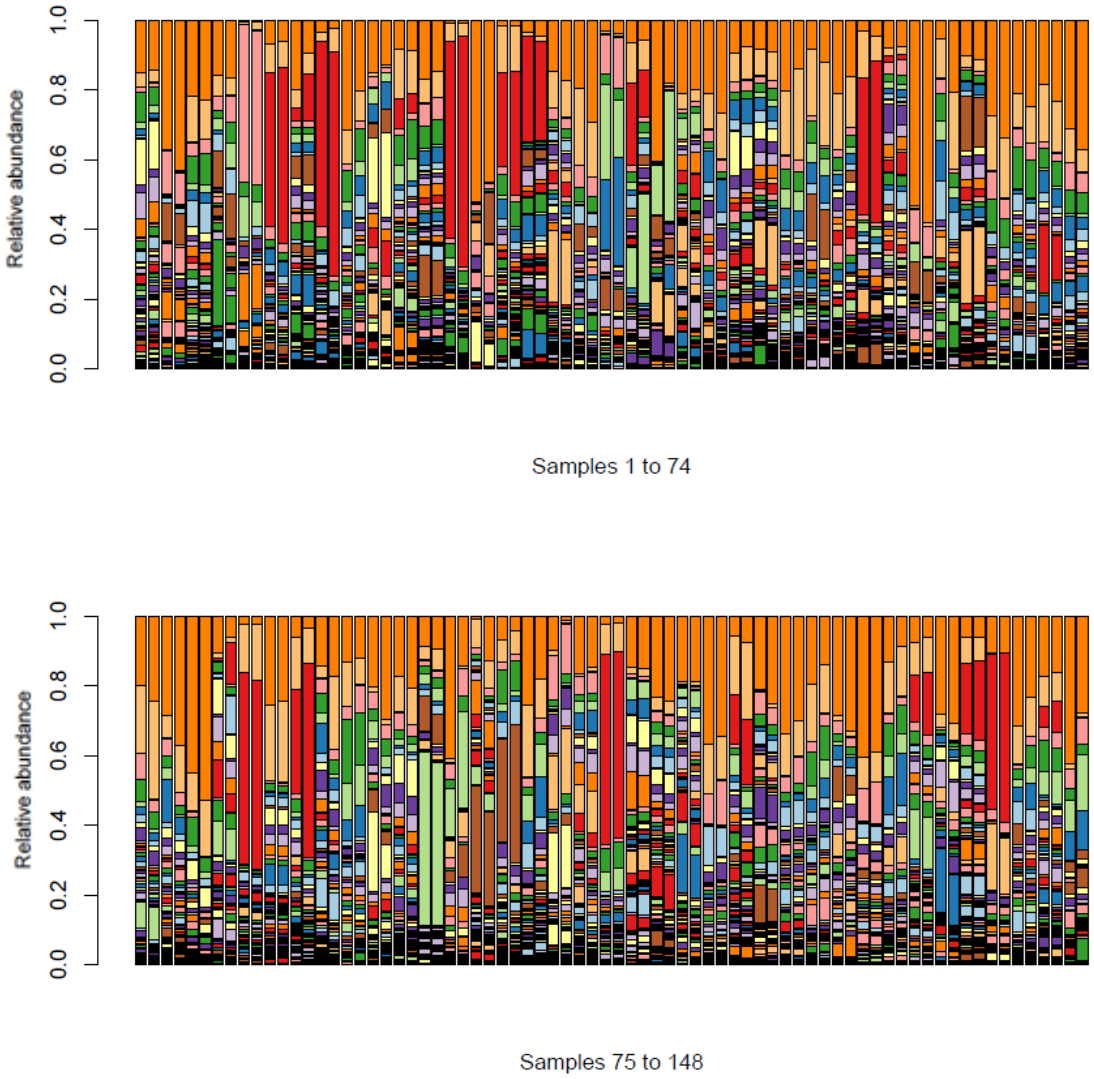


Figure 34. Taxonomy bar charts for extraction replicates. Each pair of adjacent bars represents a pair of extraction replicates. Colours denote the relative abundance of taxa (genus level); there are too many to include a legend.

LEfSe analysis was performed to determine whether the relative abundance of any taxa differed significantly between samples within each group. No taxa were significantly different between the two groups of replicates extracted at the same time. One taxon (*prevotellaceae*NK3B31 group) was found to be significantly enriched in the group of replicates extracted after a period of storage at ambient temperature compared with the group containing the original samples (Figure 35).

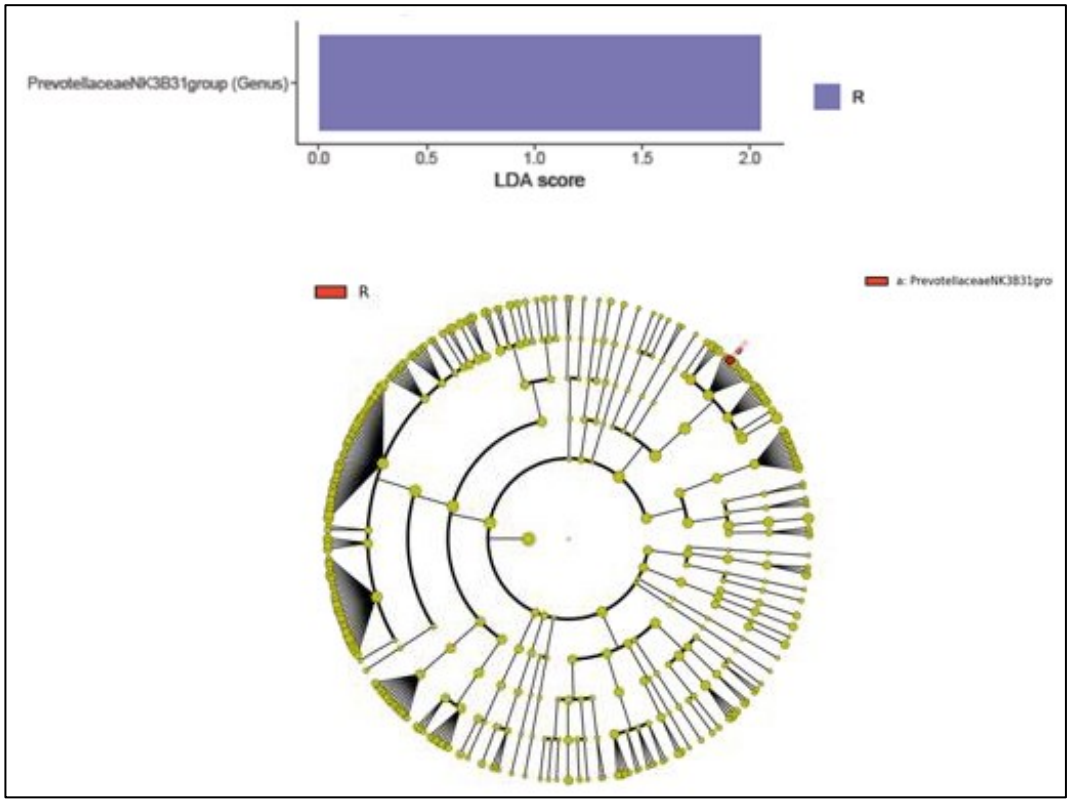


Figure 35. LEfSe plot and cladogram of samples whereby replicates were extracted after a period of storage at ambient temperature. LEfSe plot indicates the taxon which is significantly enriched in the group of replicates extracted after a period of storage at ambient temperature (R) compared with the group containing the original samples. The cladogram indicates the phylogenetic relationship between taxa. Working outwards, the five rings denote phylum, class, order, family, and genus. Taxa are represented by circles. Circle diameter is proportional to abundance. The red circle indicates the taxon which is significantly enriched in the group of replicates extracted after a period of storage at ambient temperature.

Figure 36 demonstrates that the majority of samples in both groups of extraction replicates contain a low relative abundance of *prevotellaceae*NK3B31 group; the relative abundance of *prevotellaceae*NK3B31 group is not consistently higher in the group of replicates extracted after a period of storage at ambient temperature compared with the group of replicates extracted at the same time; and within the group of replicates extracted after a period of storage at ambient temperature, a single replicate pair has a high relative abundance of *prevotellaceae*NK3B31 group.

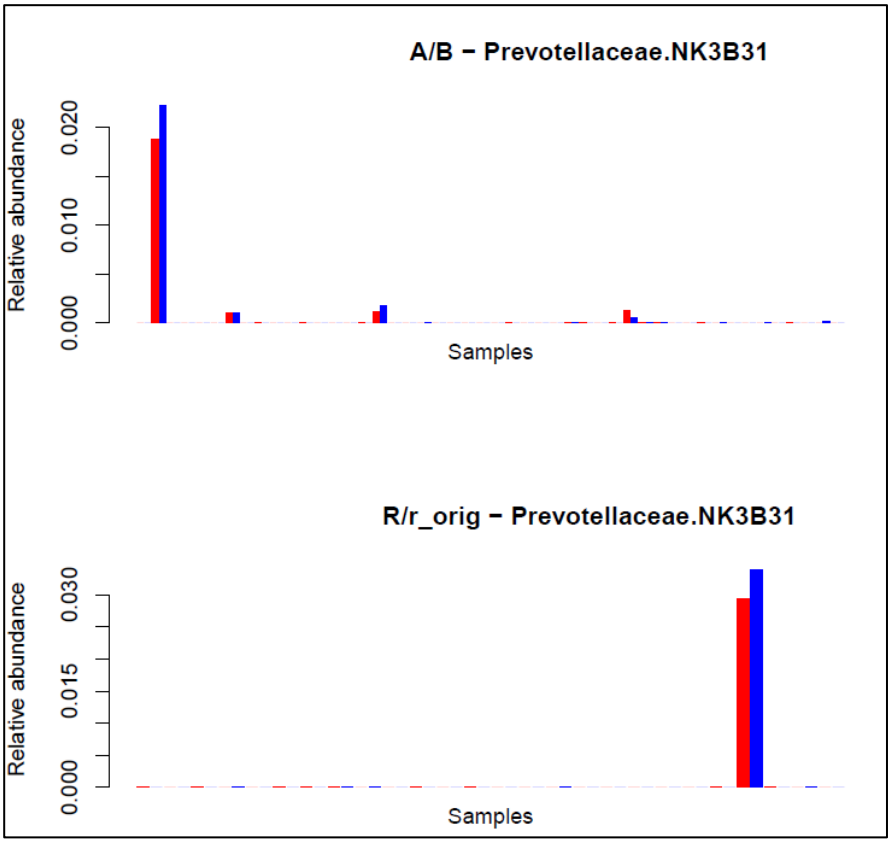


Figure 36. The relative abundance of *Prevotellaceae.*NK3B31 group for extraction replicate samples. Each pair of adjacent bars represents a pair of extraction replicates. The upper plot (A/B) = extraction replicates whereby replicates were extracted at the same time; the lower plot (R/r_orig) = extraction replicates whereby replicates were extracted after a period of storage at ambient temperature. Red = original sample.

There is good association and agreement, with an absence of bias, between the relative abundance of *prevotellaceae*NK3B31 group for extraction replicate pairs (Figure 37 and Figure 38). This, taken together with the fact that only a single taxon was identified as being significantly different, suggests that the LEfSe result is secondary to chance rather than a consistent biological effect of prolonged storage at ambient temperature.

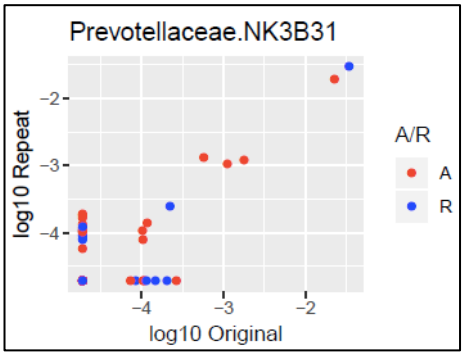


Figure 37. Scatter-plot of *Prevotellaceae.NK3B31* group for extraction replicate samples. Log relative abundance is plotted on the x (original sample) and y (replicate) axes respectively. As it is not possible to log relative abundance values equal to 0, these correspond to the smallest value on each axis (i.e. the bottom left-hand corner of the graph). Points with identical x and y values are overlaid and represented by a single point. Red (A) = extraction replicates whereby replicates were extracted at the same time. Blue (R) = extraction replicates whereby replicates were extracted after a period of storage at ambient temperature.

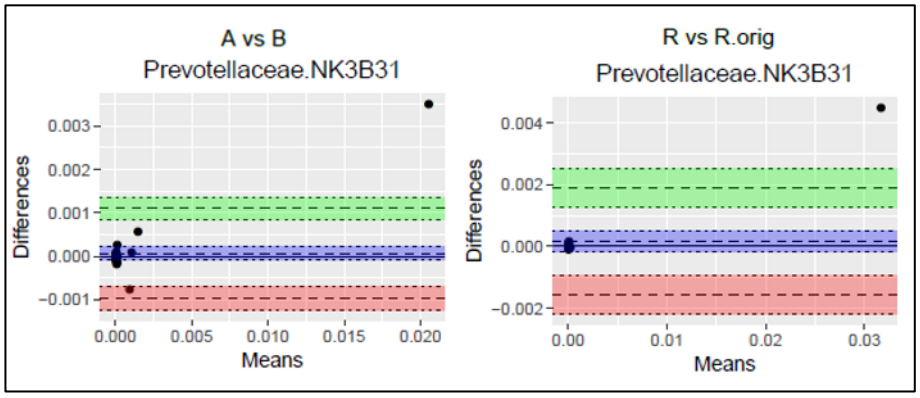
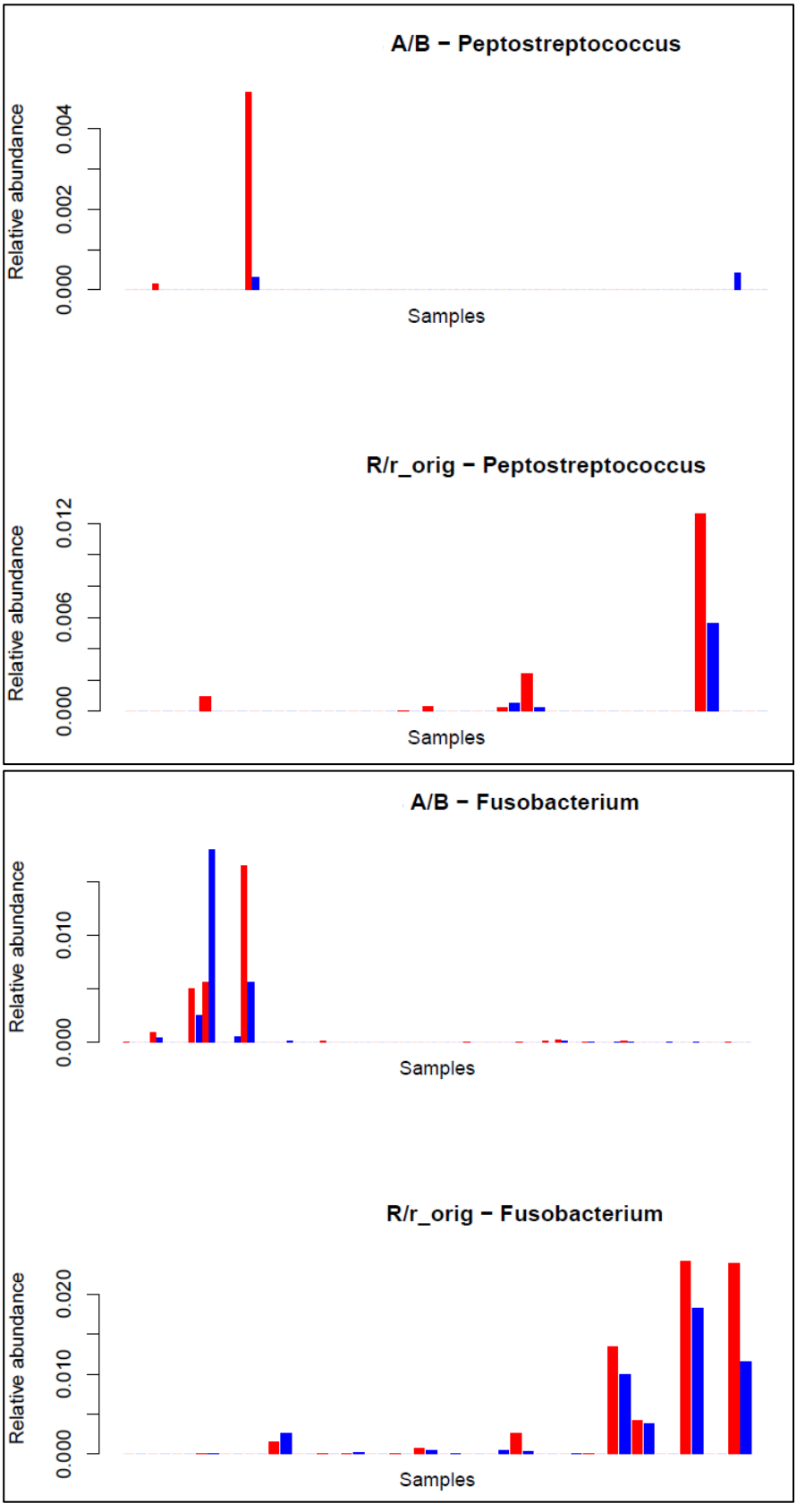
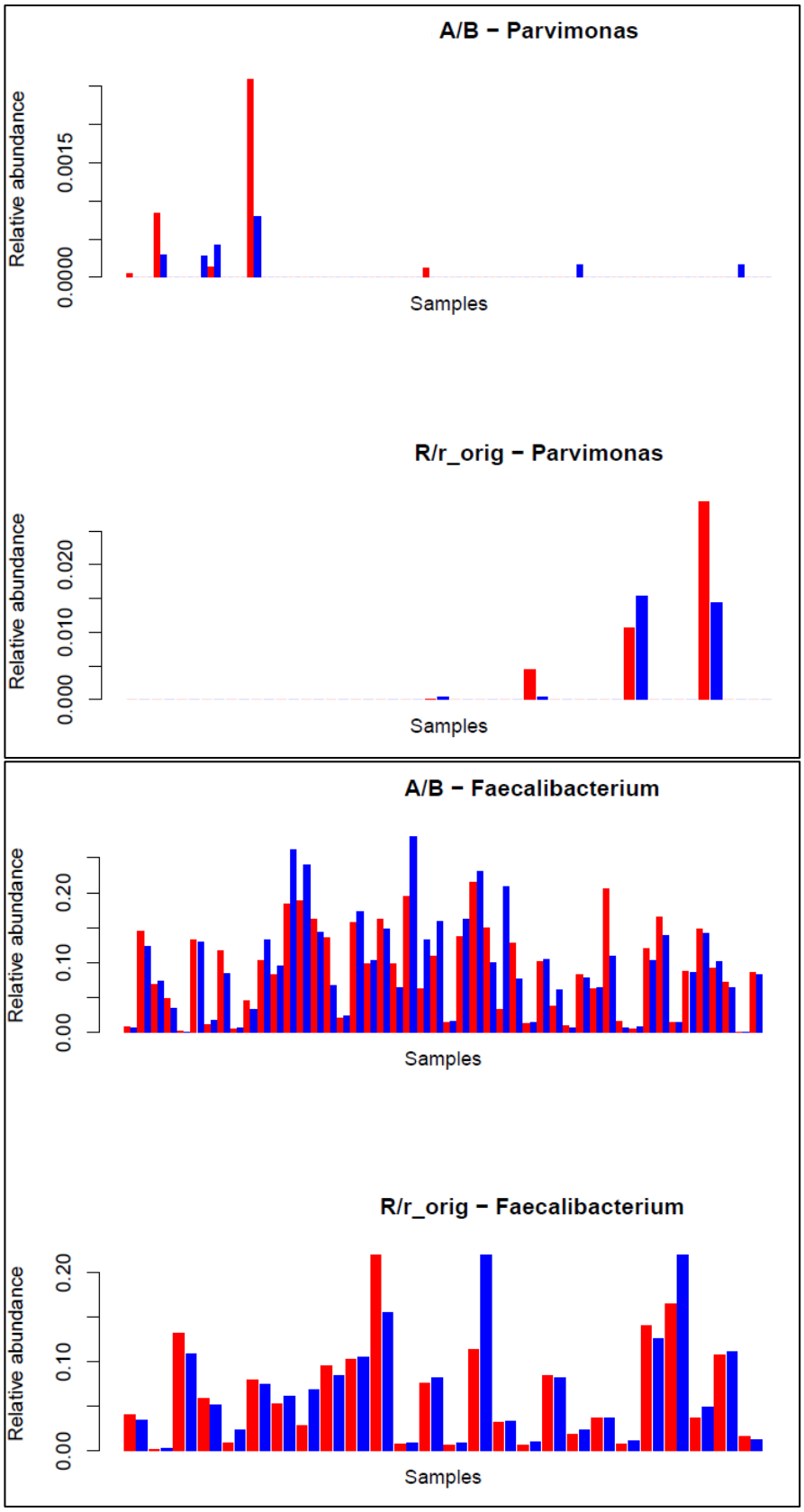


Figure 38. Bland-Altman plots of *Prevotellaceae.NK3B31 group* for extraction replicate samples. The left-hand plot (A vs B) shows extraction replicates whereby replicates were extracted at the same time. The right-hand plot (R vs R.orig) shows extraction replicates whereby replicates were extracted after a period of storage at ambient temperature. The x axis shows the mean relative abundance of *Prevotellaceae.*NK3B31 group across replicates. The y axis shows the difference in relative abundance of *Prevotellaceae.*NK3B31 group across replicates. The purple band represents the mean difference (i.e. bias) (plus 95% CI) in relative abundance of *Prevotellaceae.*NK3B31 group across replicates. The green and red bands represent the upper and lower 95% limits of agreement plus 95% CI. The upper and lower limits of agreement are calculated as mean difference +/- 1.96 SD; they are expected to contain 95% of the differences measured between replicates.

The relative abundances of CRC-associated bacteria were compared between replicate pairs (Figure 39 and Figure 40). There is good association between replicate measurements.




A/B − Parvimonas
Relative abundance
0.0000
0.0015
Samples
R/r_orig − Parvimonas
Relative abundance
0.000
0.010
0.020
Samples
A/B − Faecalibacterium
Relative abundance
0.00
0.10
0.20
Samples
R/r_orig − Faecalibacterium
Relative abundance
0.00
0.10
0.20
Samples

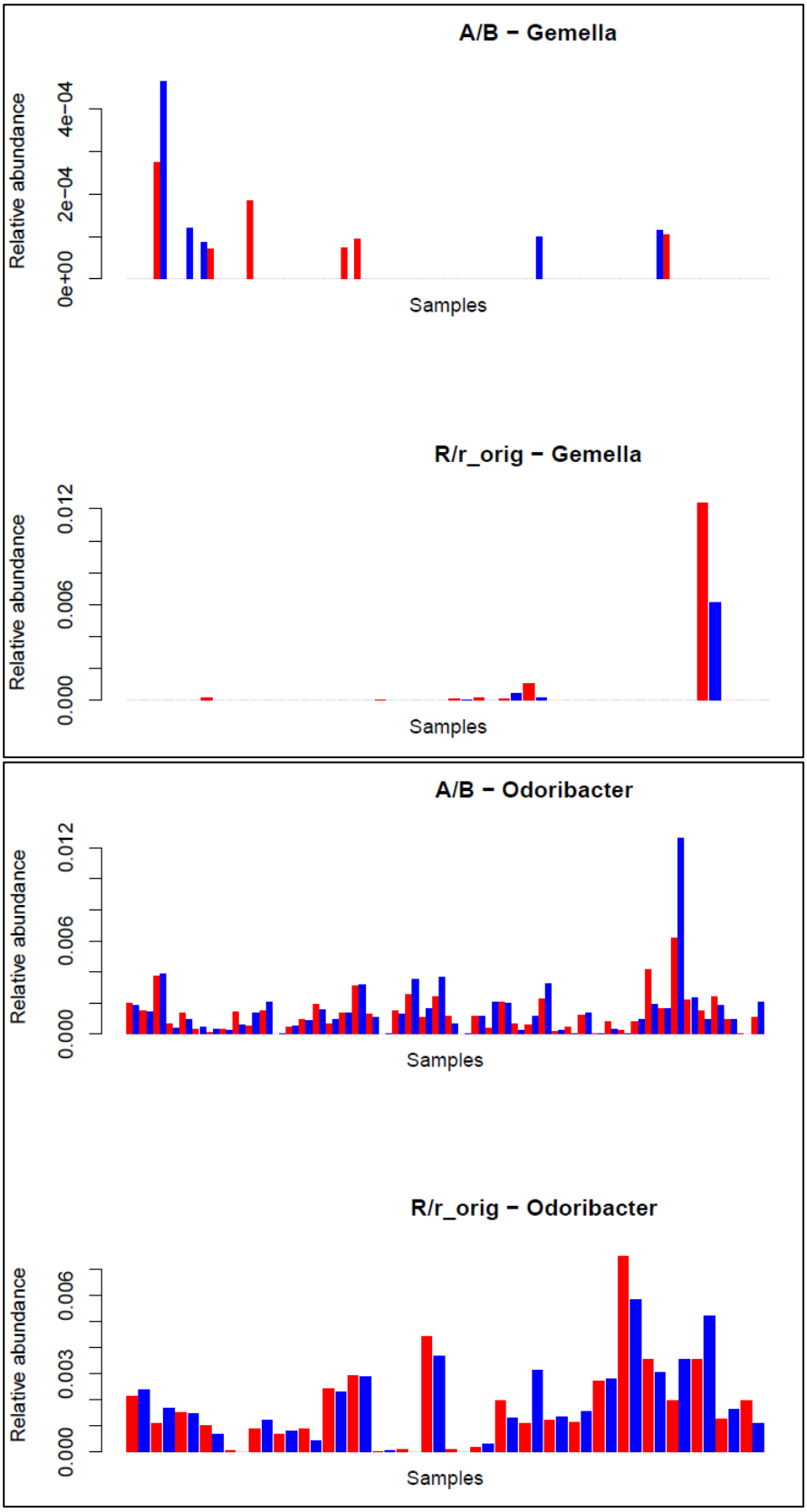

A/B − Gemella
Relative abundance
0e+00
2e−04
4e−04
Samples
R/r_orig − Gemella
Relative abundance
0.000
0.006
0.012
Samples
A/B − Odoribacter
Relative abundance
0.000
0.006
0.012
Samples
R/r_orig − Odoribacter
Relative abundance
0.000
0.003
0.006
Samples

Figure 39. The relative abundances of CRC-associated taxa for extraction replicate samples. Each pair of adjacent bars represents a pair of extraction replicates. For each taxon, the upper plot (A/B) = extraction replicates whereby replicates were extracted at the same time; the lower plot (R/r_orig) = extraction replicates whereby replicates were extracted after a period of storage at ambient temperature. Red = original sample. e = x 10^.

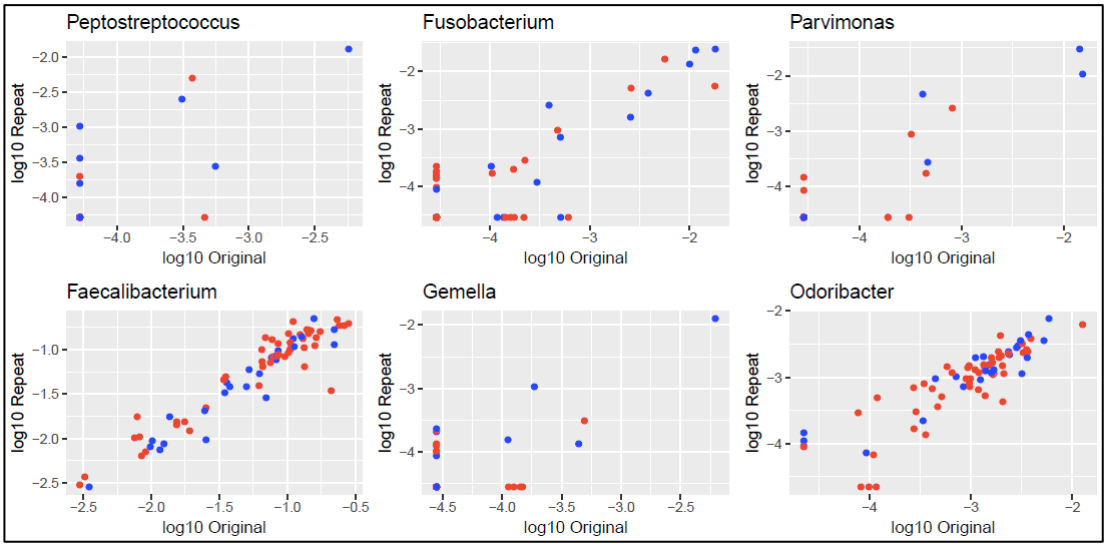


Figure 40. Scatter-plots of CRC-associated taxa for extraction replicate samples. Log relative abundance is plotted on the x (original sample) and y (replicate) axes respectively. As it is not possible to log relative abundance values equal to 0, these correspond to the smallest value on each axis (i.e. the bottom left-hand corner of each graph). Points with identical x and y values are overlaid and represented by a single point. Red = extraction replicates whereby replicates were extracted at the same time. Blue = extraction replicates whereby replicates were extracted after a period of storage at ambient temperature.

The Bland-Altman plots (Figure 41) indicate an absence of bias between replicate measurements; variation in relative abundances between replicates is minimal, and similar between extraction replicates extracted at the same time or after a period of storage at ambient temperature.

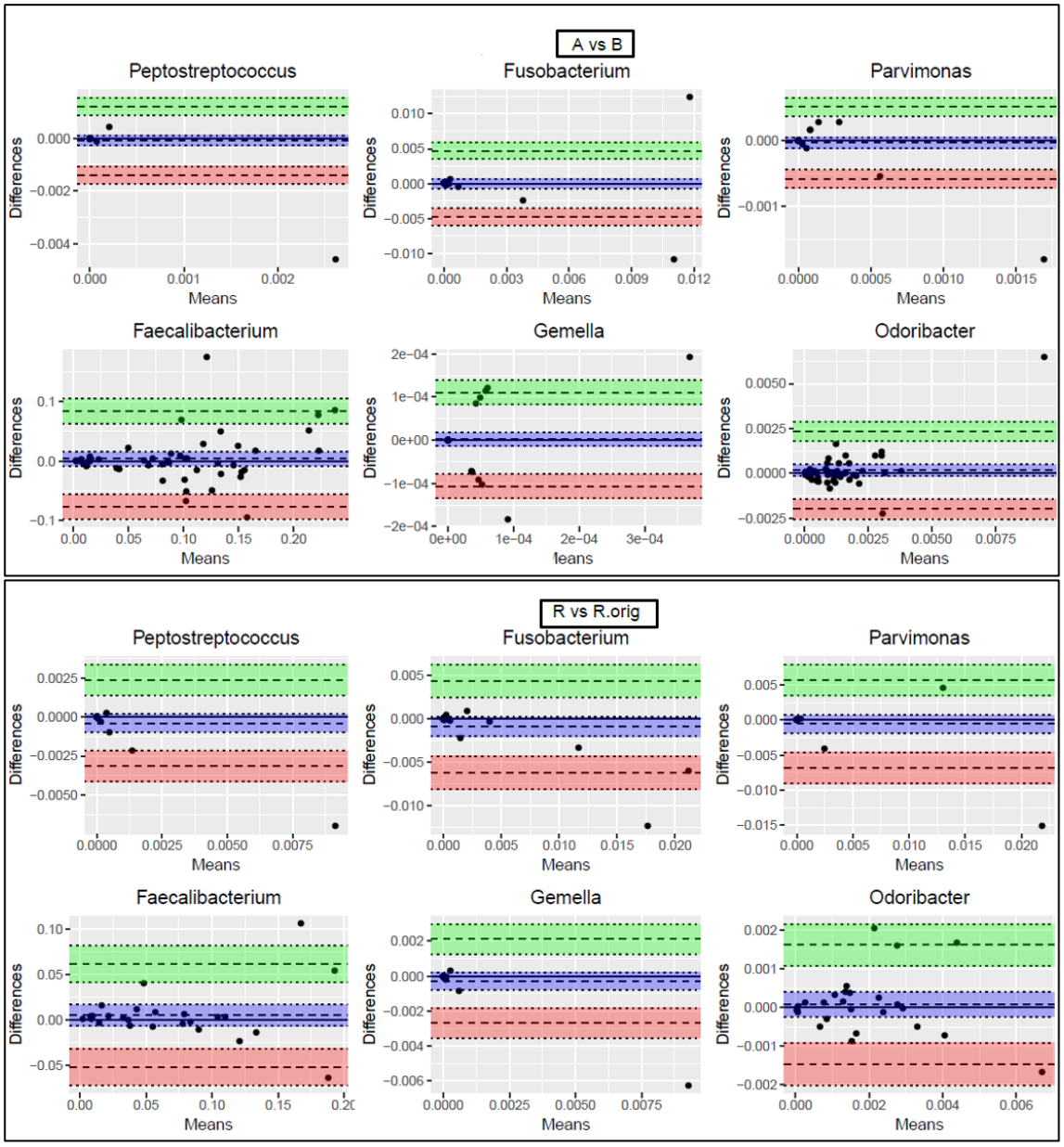


Figure 41. Bland-Altman plots of CRC-associated taxa for extraction replicate samples. The upper plots (A vs B) show extraction replicates whereby replicates were extracted at the same time. The lower plots (R vs R.orig) show extraction replicates whereby replicates were extracted after a period of storage at ambient temperature. The x axis shows the mean relative abundance of CRC-associated taxa across replicates. The y axis shows the difference in relative abundance of CRC-associated taxa across replicates. The purple band represents the mean difference (i.e. bias) (plus 95% CI) in relative abundance of CRC-associated taxa across replicates. The green and red bands represent the upper and lower 95% limits of agreement plus 95% CI. The upper and lower limits of agreement are calculated as mean difference +/- 1.96 SD; they are expected to contain 95% of the differences measured between replicates. e = x 10^.

There is a degree of variation of the relative abundance of *Escherichia-Shigella* between replicates but no bias (Figure 42-Figure 44); variation is not higher in the group of replicates extracted after a period of storage at ambient temperature compared with the group of replicates extracted at the same time, which suggests that variation in the relative abundance of *Escherichia-Shigella* is not related to differences in duration of storage.

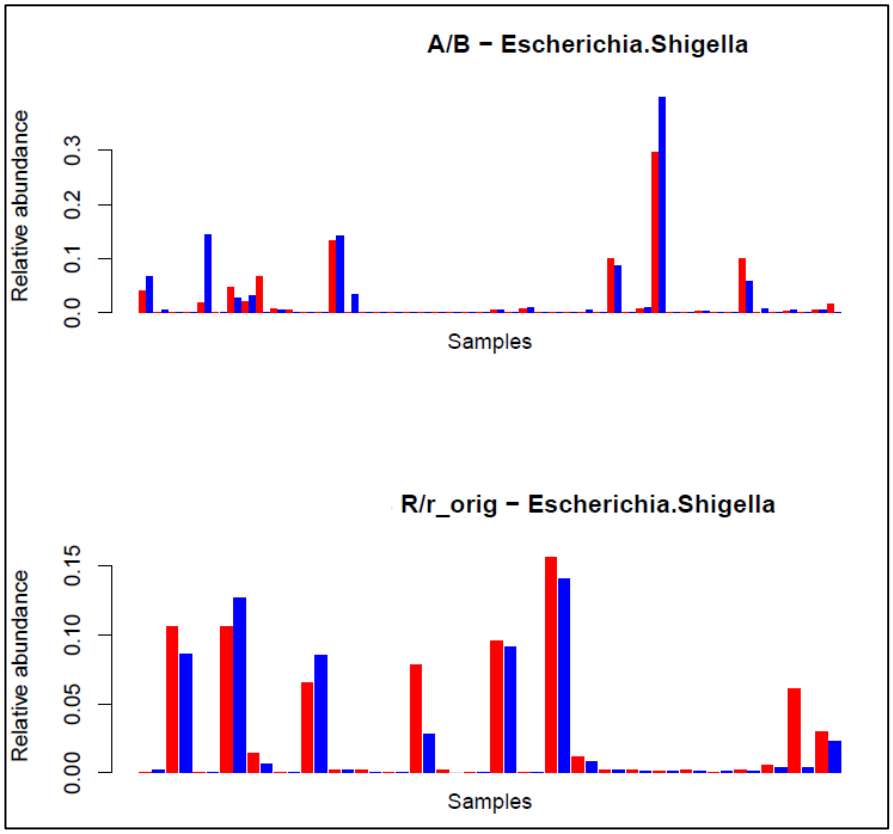


Figure 42. The relative abundance of *Escherichia-Shigella* for extraction replicate samples. Each pair of adjacent bars represents a pair of extraction replicates. The upper plot (A/B) = extraction replicates whereby replicates were extracted at the same time; the lower plot (R/r_orig) = extraction replicates whereby replicates were extracted after a period of storage at ambient temperature. Red = original sample.

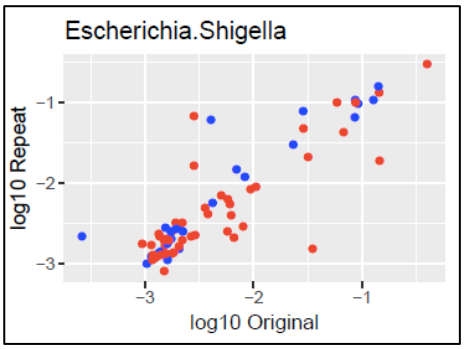


Figure 43. Scatter-plot of *Escherichia-Shigella* for extraction replicate samples. Log relative abundance is plotted on the x (original sample) and y (replicate) axes respectively. As it is not possible to log relative abundance values equal to 0, these correspond to the smallest value on each axis (i.e. the bottom left-hand corner of the graph). Points with identical x and y values are overlaid and represented by a single point. Red = extraction replicates whereby replicates were extracted at the same time. Blue = extraction replicates whereby replicates were extracted after a period of storage at ambient temperature.

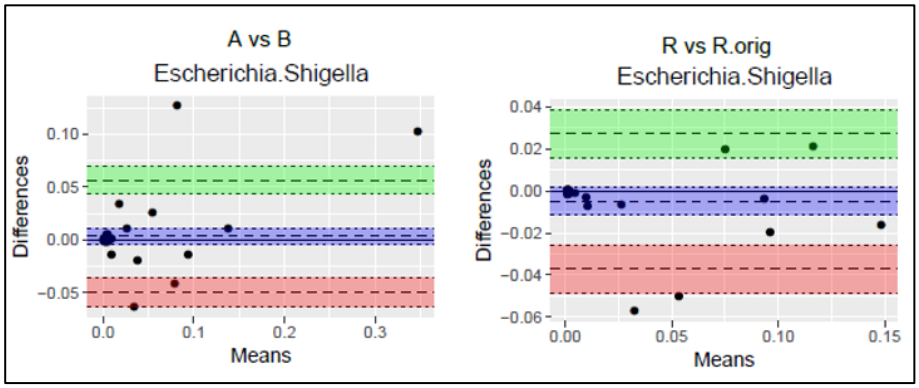


Figure 44. Bland-Altman plots of *Escherichia-Shigella* for extraction replicate samples. The left-hand plot (A vs B) shows extraction replicates whereby replicates were extracted at the same time. The right-hand plot (R vs R.orig) shows extraction replicates whereby replicates were extracted after a period of storage at ambient temperature. The x axis shows the mean relative abundance of *Escherichia-Shigella* across replicates. The y axis shows the difference in relative abundance of *Escherichia-Shigella* across replicates. The purple band represents the mean difference (i.e. bias) (plus 95% CI) in relative abundance of *Escherichia-Shigella* across replicates. The green and red bands represent the upper and lower 95% limits of agreement plus 95% CI. The upper and lower limits of agreement are calculated as mean difference +/- 1.96 SD; they are expected to contain 95% of the differences measured between replicates.

2.4.5 Assessing temporal variation of gFOBT samples

2.4.5.1 Temporal 1-6 samples

Each ‘Temporal 1-6’ sample was derived from one of the six squares of a gFOBT card; this represents subsamples of three stools collected over three consecutive days.

The PCA of Bray-Curtis distances (Figure 45) and taxonomy bar chart (Figure 47) demonstrated that the majority of extraction replicates derived from a single gFOBT card were highly similar to one another. For each gFOBT card, the Bray-Curtis distances between the first subsample (arbitrarily taken as a reference) and each of the five remaining subsamples were lower than the mean of the Bray-Curtis distances between the reference subsample and subsamples derived from the remaining ‘Temporal 1-6’ gFOBT cards (Figure 46). Bray-Curtis distances between the first and second subsample (derived from the same stool) were often, but not always, smaller than Bray-Curtis distances between subsamples derived from stool samples collected on consecutive days.

However, a degree of variability for some samples was apparent. The sample which is furthest from its counterparts on the PCA plot is 2146P-6; the taxonomy bar plot demonstrates that it contains a higher relative abundance of *Escherichia-Shigella* than its replicate counterparts. This was shown to be the case for several of the samples (Figure 47 and Figure 48). The difference between the maximum and minimum relative abundance of each taxa (at genus level) was calculated across the six samples derived from each gFOBT card. The greatest difference was in the relative abundance of *Escherichia-Shigella* (Table 12).

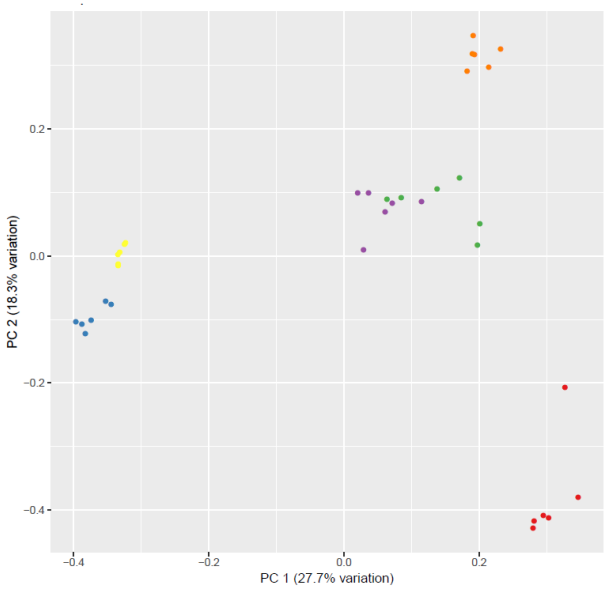


Figure 45. PCA of Bray-Curtis distances for Temporal 1-6 samples. Points on the graph are coloured according to gFOBT sample.

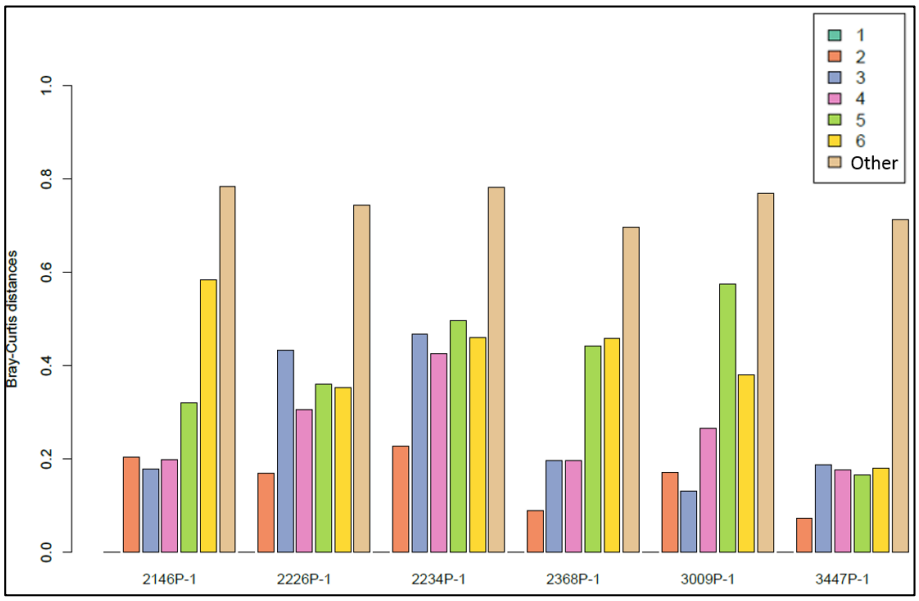


Figure 46. Bar chart of Bray-Curtis distances for Temporal 1-6 samples. Seven bars correspond to each gFOBT card. For each, the first six bars (coloured according to the key) show the Bray-Curtis distances between each of the six subsamples and the first subsample. The seventh (brown) bar labelled 'Other' shows the mean of the Bray-Curtis distances between the first subsample of that gFOBT card and subsamples derived from the remaining 'Temporal 1-6' gFOBT cards.

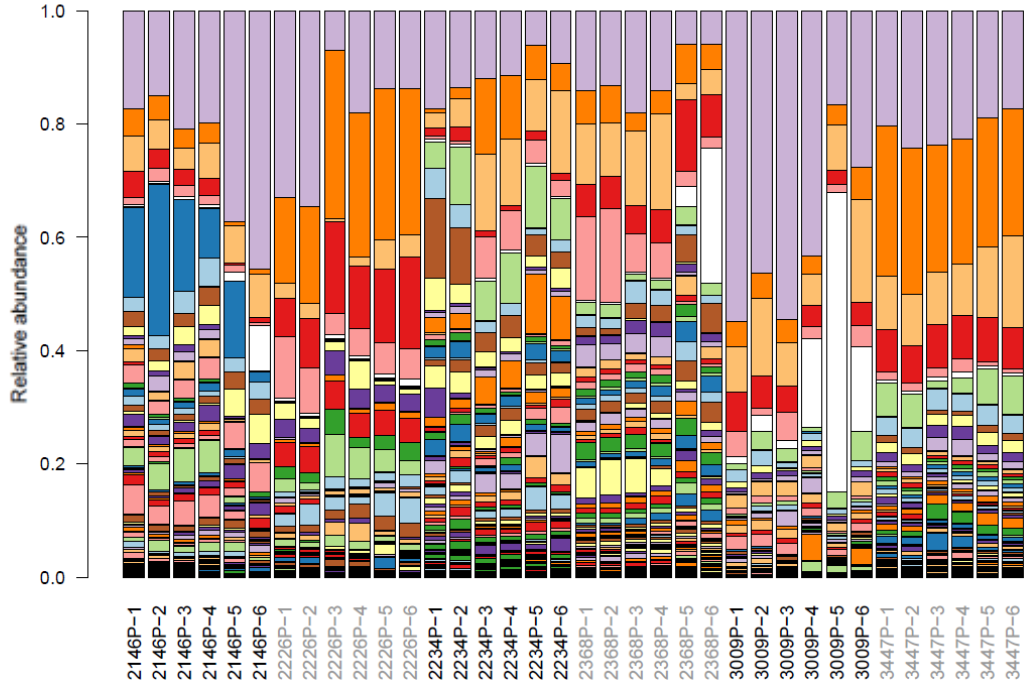


Figure 47. Taxonomy bar chart for Temporal 1-6 samples. Each bar represents a subsample derived from a single gFOBT square. Colours denote the relative abundance of taxa (genus level); there are too many to include a legend. *Escherichia-Shigella* is white for ease of identification.

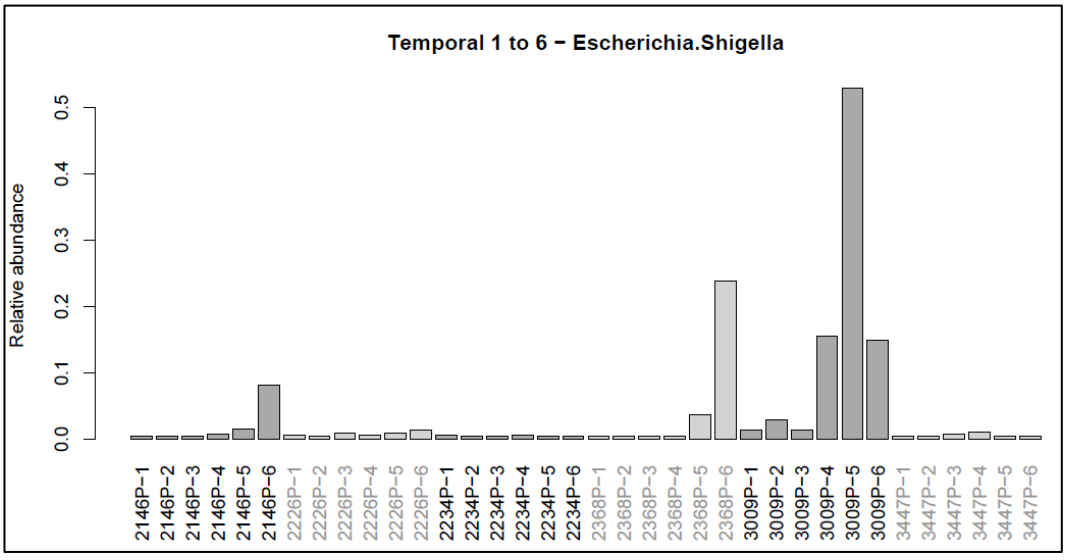
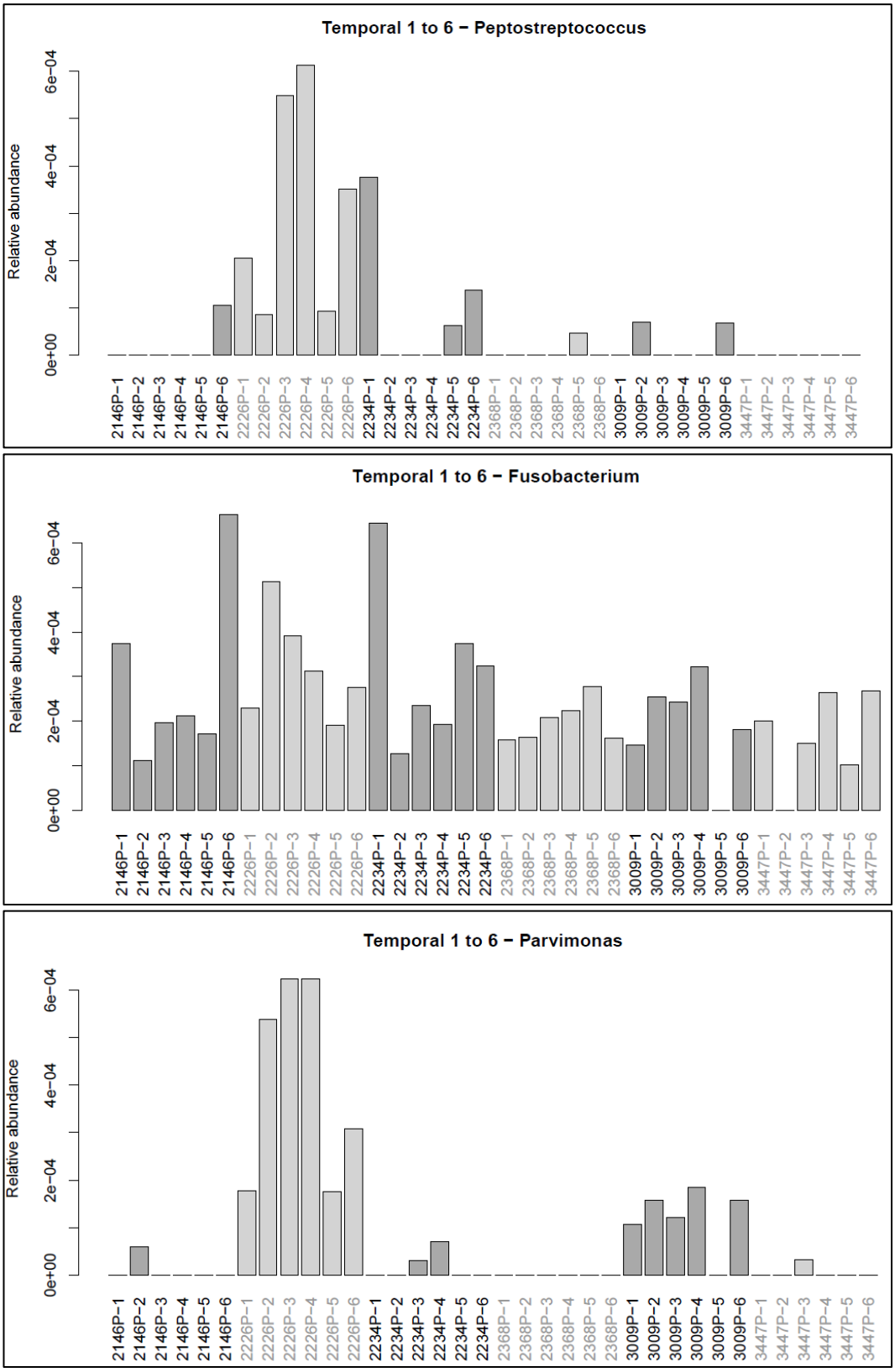


Figure 48. The relative abundance of *Escherichia-Shigella* for Temporal 1-6 samples. Each bar represents a subsample derived from a single gFOBT square.

Table 12. The greatest difference in relative abundance of taxa (at genus level) across the six samples derived from each gFOBT card for Temporal 1-6 samples.

gFOBT card	Greatest difference in relative abundance of taxa (at genus level) across the six subsamples	Taxa
3009P	0.52	*Escherichia.Shigella*
2146P	0.31	*Bacteroides*
2226P	0.28	*Bacteroides*
2368P	0.23	*Escherichia.Shigella*
2234P	0.13	*Lachnospiraceae*
3447P	0.07	*Faecalibacterium*

CRC-associated bacteria showed relatively less variability between replicate samples (Figure 49); especially when the scale of the y axis is taken into consideration. This suggests a degree of subsample and temporal (over three consecutive days) consistency between CRC-associated taxa.


Temporal 1 to 6 − Peptostreptococcus
Relative abundance
0e+00
2e−04
4e−04
6e−04
2146P−1
2146P−2
2146P−3
2146P−4
2146P−5
2146P−6
2226P−1
2226P−2
2226P−3
2226P−4
2226P−5
2226P−6
2234P−1
2234P−2
2234P−3
2234P−4
2234P−5
2234P−6
2368P−1
2368P−2
2368P−3
2368P−4
2368P−5
2368P−6
3009P−1
3009P−2
3009P−3
3009P−4
3009P−5
3009P−6
3447P−1
3447P−2
3447P−3
3447P−4
3447P−5
3447P−6
Temporal 1 to 6 − Fusobacterium
Relative abundance
0e+00
2e−04
4e−04
6e−04
2146P−1
2146P−2
2146P−3
2146P−4
2146P−5
2146P−6
2226P−1
2226P−2
2226P−3
2226P−4
2226P−5
2226P−6
2234P−1
2234P−2
2234P−3
2234P−4
2234P−5
2234P−6
2368P−1
2368P−2
2368P−3
2368P−4
2368P−5
2368P−6
3009P−1
3009P−2
3009P−3
3009P−4
3009P−5
3009P−6
3447P−1
3447P−2
3447P−3
3447P−4
3447P−5
3447P−6
Temporal 1 to 6 − Parvimonas
Relative abundance
0e+00
2e−04
4e−04
6e−04
2146P−1
2146P−2
2146P−3
2146P−4
2146P−5
2146P−6
2226P−1
2226P−2
2226P−3
2226P−4
2226P−5
2226P−6
2234P−1
2234P−2
2234P−3
2234P−4
2234P−5
2234P−6
2368P−1
2368P−2
2368P−3
2368P−4
2368P−5
2368P−6
3009P−1
3009P−2
3009P−3
3009P−4
3009P−5
3009P−6
3447P−1
3447P−2
3447P−3
3447P−4
3447P−5
3447P−6

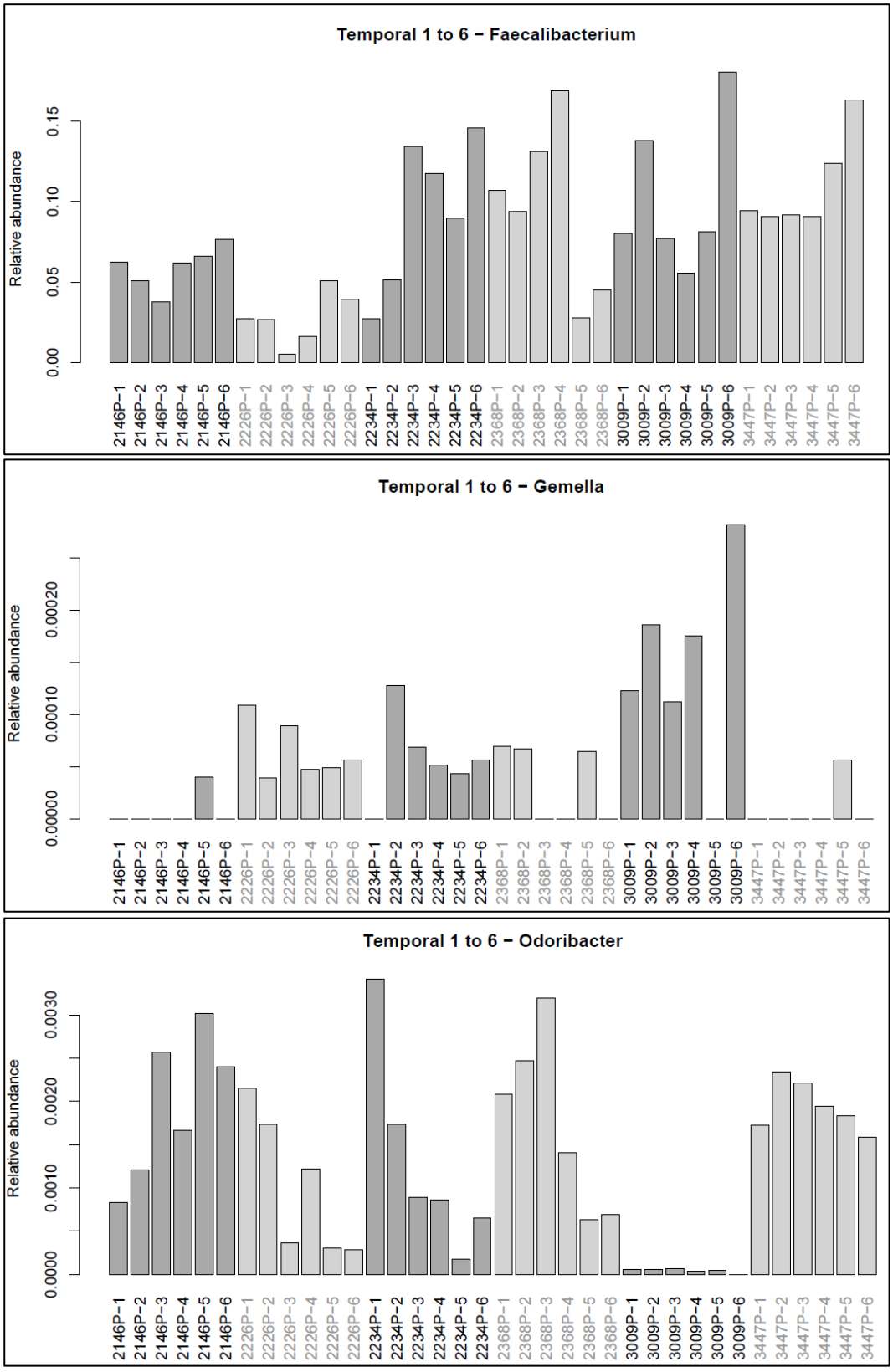


Figure 49. The relative abundance of CRC-associated taxa for Temporal 1-6 samples. Each bar represents a subsample derived from a single gFOBT square. e = x 10^.

2.4.5.2 Temporal 1.2.3 combined samples

Each ‘Temporal 1.2.3.combined’ sample was derived from one of the six squares of a gFOBT card (samples 1.2.3), representing subsamples of three stools collected over three consecutive days, or a combined sample (of the three alternate squares). The two samples with fewer than 10,000 reads were removed from analysis.

The PCA of Bray-Curtis distances (Figure 50) and taxonomy bar chart (Figure 52) demonstrated that most extraction replicates were most similar to one another, but that a few of the samples showed a high degree of separation from the parent cluster. For each gFOBT card, the Bray-Curtis distances between the combined subsample (arbitrarily taken as a reference) and each of the remaining subsamples were lower than the mean of the Bray-Curtis distances between the reference subsample and subsamples derived from the remaining ‘Temporal 1.2.3 combined’ cards (although in three cases the Bray-Curtis distances were similar) (Figure 51). This suggests that in most cases, a single subsample has a similar overall microbiome community to a combined subsample derived from stool collected over three consecutive days.

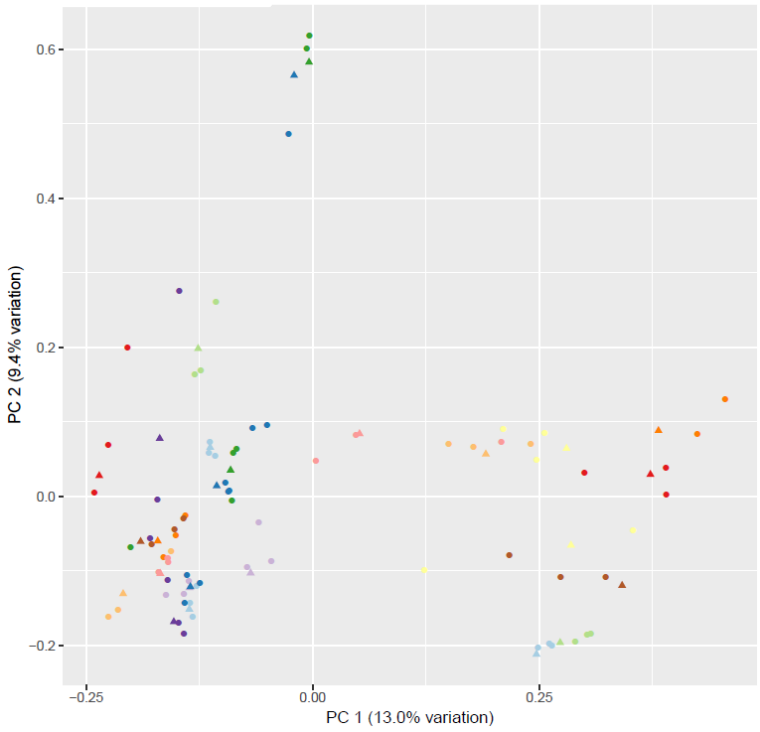


Figure 50. PCA of Bray-Curtis distances for Temporal 1.2.3.combined samples. Points on the graph are coloured according to gFOBT sample. The ‘combined’ subsamples are indicated by a triangle.

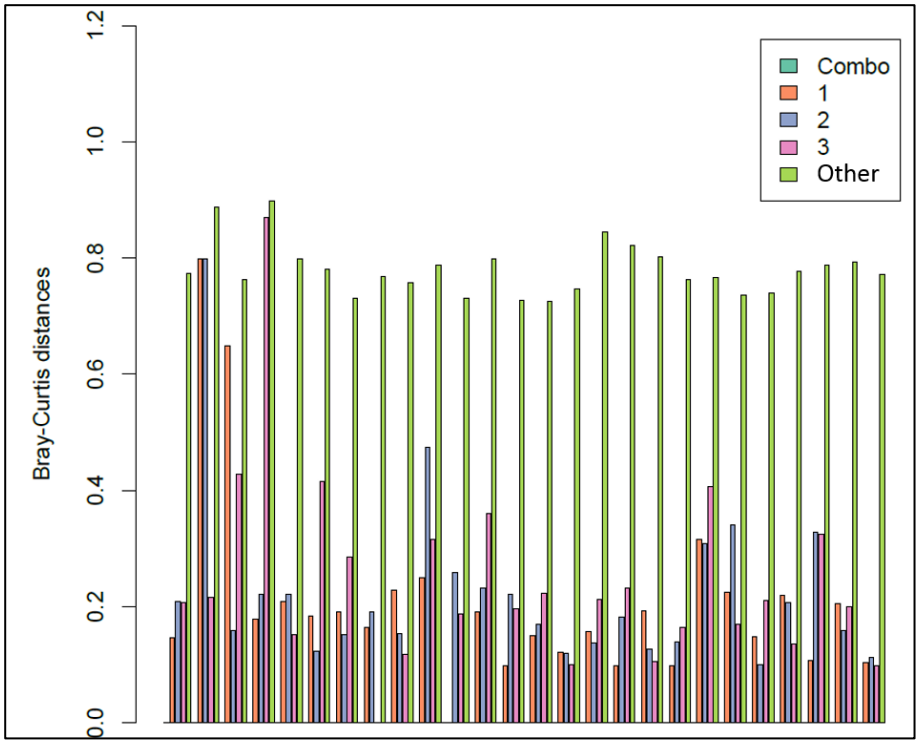


Figure 51. Bar chart of Bray-Curtis distances for Temporal 1.2.3.combined samples. Five bars correspond to each gFOBT card (the names are not included due to limited space). For each, the first four bars (coloured according to the key) show the Bray-Curtis distances between each of the three subsamples and the combined (combo) subsample. The fifth (light green) bar labelled 'Other' shows the mean of the Bray-Curtis distances between the combined subsample of that gFOBT card and subsamples derived from the remaining 'Temporal 1.2.3.combined' gFOBT cards.

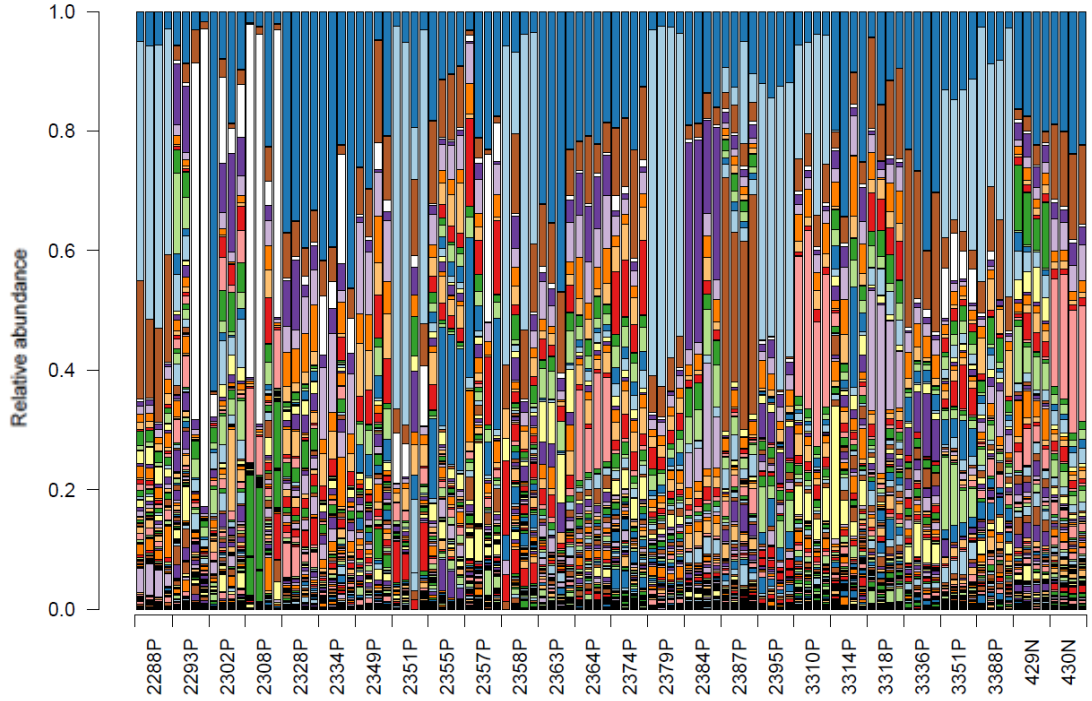


Figure 52. Taxonomy bar chart for Temporal 1.2.3.combined samples. Each gFOBT sample is represented by four bars, corresponding (from left to right) to subsample 1 (from a single gFOBT square), 2 (from a single gFOBT square), 3 (from a single gFOBT square) and 'combined' (a combination of three gFOBT squares). Colours denote the relative abundance of taxa (genus level); there are too many to include a legend. *Escherichia-Shigella* is white for ease of identification.

Figure 52 and Figure 53 demonstrate a high degree of variability of *Escherichia-Shigella* relative abundance for some replicates. The difference between the maximum and minimum relative abundance of each taxa (at genus level) was calculated across the four samples derived from each gFOBT card. The greatest difference was in the relative abundance of *Escherichia-Shigella* (Table 13).

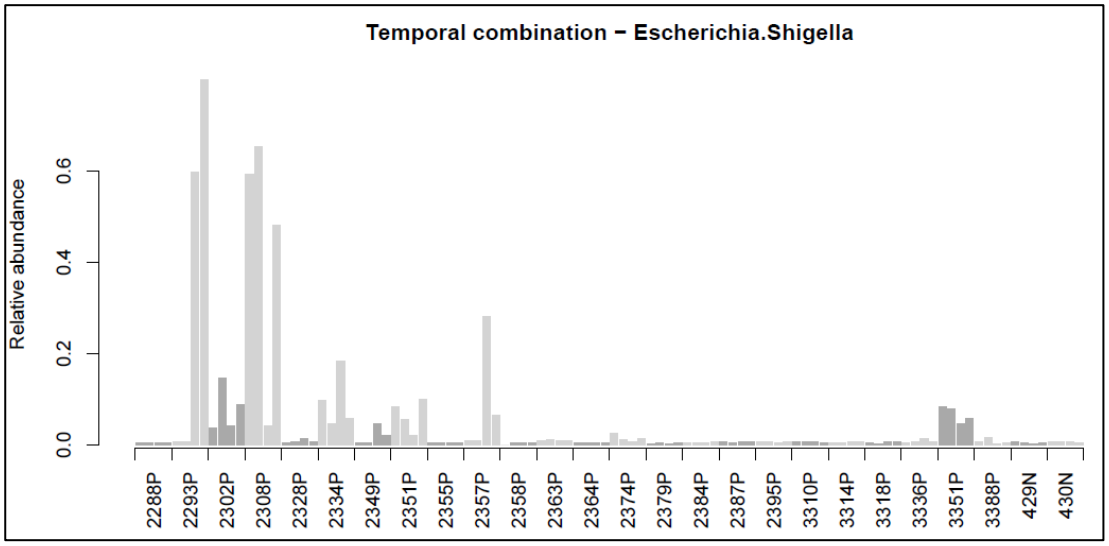


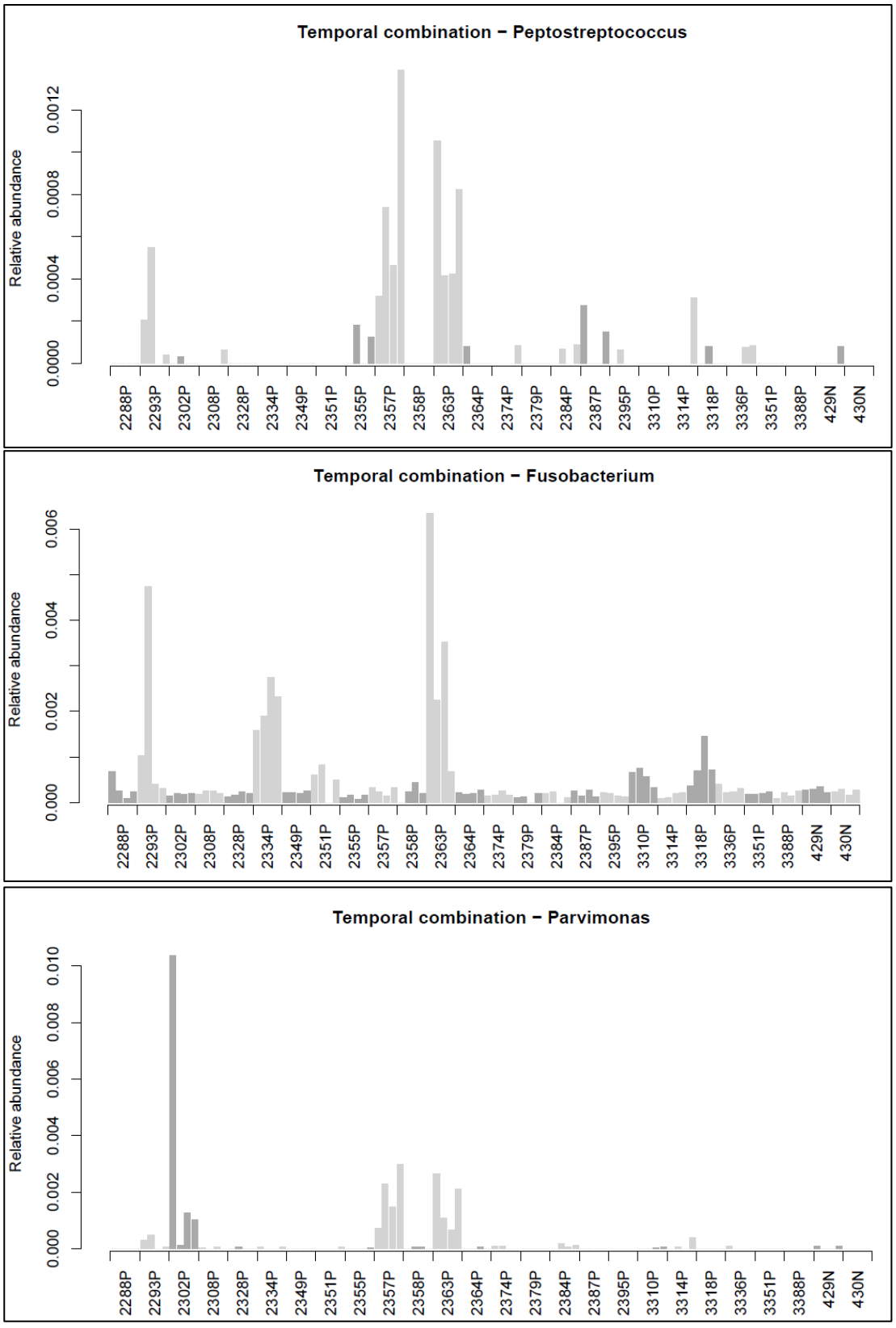
Figure 53. The relative abundance of *Escherichia-Shigella* for Temporal 1.2.3.combined samples. Each gFOBT sample is represented by four bars, corresponding (from left to right) to subsample 1 (from a single gFOBT square), 2 (from a single gFOBT square), 3 (from a single gFOBT square) and 'combined' (a combination of three gFOBT squares).

Table 13. The greatest difference in relative abundance of taxa (at genus level) across the four samples derived from each gFOBT card for Temporal 1.2.3.combined samples.

gFOBT card	Greatest difference in relative abundance of taxa (at genus level) across the four subsamples	Taxa
2293P	0.79	*Escherichia.Shigella*
2308P	0.61	*Escherichia.Shigella*
2302P	0.56	*Bacteroides*
2351P	0.55	*Prevotella.9*
2358P	0.40	*Prevotella.9*
2357P	0.27	*Escherichia.Shigella*
2349P	0.25	*Bacteroides*
3388P	0.24	*Prevotella.9*
3314P	0.24	*Bacteroides*
2334P	0.24	*Bacteroides*
2363P	0.24	*Bacteroides*

gFOBT card	Greatest difference in relative abundance of taxa (at genus level) across the four subsamples	Taxa
3336P	0.20	*Faecalibacterium*
2384P	0.19	*Alistipes*
3318P	0.19	*Rikenellaceae.RC9. gut.group*
2387P	0.17	*Bacteroidales.S24.7. group.D_5* uncultured bacterium
3310P	0.15	*Prevotella.9*
2395P	0.11	*Prevotella.9*
2374P	0.11	*Bacteroides*
2328P	0.10	*Alistipes*
2379P	0.10	*Prevotella.9*
2288P	0.10	*Prevotella.9*
2355P	0.09	*Bacteroides*
2364P	0.09	*Ruminococcaceae.D _5__uncultured*
3351P	0.09	*Prevotella.9*
429N	0.06	*Bacteroides*
430N	0.05	*Bacteroides*

CRC-associated bacteria demonstrated relatively less variability between replicate samples (Figure 54); especially when the scale of the y axis is taken into consideration. This suggests that a single subsample (as will be obtained with FIT) may give similar relative abundances of these CRC-associated taxa compared with a combined sample (derived from three separate stools, as is currently obtained with gFOBT).


Temporal combination − Peptostreptococcus
Relative abundance
0.0000
0.0004
0.0008
0.0012
2288P
2293P
2302P
2308P
2328P
2334P
2349P
2351P
2355P
2357P
2358P
2363P
2364P
2374P
2379P
2384P
2387P
2395P
3310P
3314P
3318P
3336P
3351P
3388P
429N
430N
Temporal combination − Fusobacterium
Relative abundance
0.000
0.002
0.004
0.006
2288P
2293P
2302P
2308P
2328P
2334P
2349P
2351P
2355P
2357P
2358P
2363P
2364P
2374P
2379P
2384P
2387P
2395P
3310P
3314P
3318P
3336P
3351P
3388P
429N
430N
Temporal combination − Parvimonas
Relative abundance
0.000
0.002
0.004
0.006
0.008
0.010
2288P
2293P
2302P
2308P
2328P
2334P
2349P
2351P
2355P
2357P
2358P
2363P
2364P
2374P
2379P
2384P
2387P
2395P
3310P
3314P
3318P
3336P
3351P
3388P
429N
430N

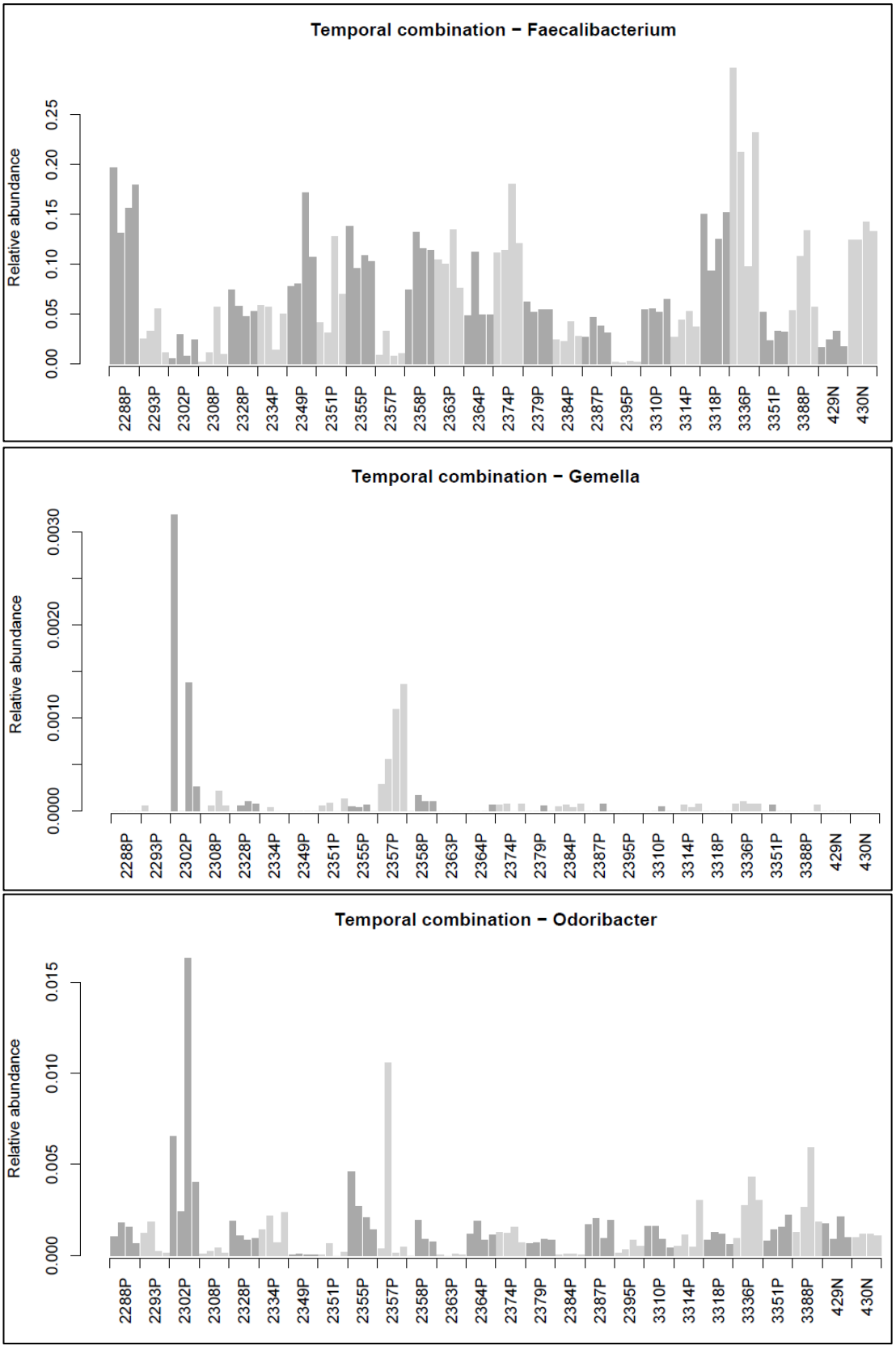


Figure 54. The relative abundance of CRC-associated taxa for Temporal 1.2.3.combined samples. Each gFOBT sample is represented by four bars, corresponding (from left to right) to subsample 1 (from a single gFOBT square), 2 (from a single gFOBT square), 3 (from a single gFOBT square) and 'combined' (a combination of three gFOBT squares).

2.4.5.3 Temporal N 1-3 samples

Each ‘Temporal N1-3’ sample was derived from one of the six squares of a gFOBT card; this represents subsamples of three stools collected over three consecutive days (except for one gFOBT card where the stools were collected on days 1, 3 and 4).

The PCA of Bray-Curtis distances (Figure 55) and taxonomy bar chart (Figure 57) demonstrated that two of the sets of extraction replicates showed a high degree of intra-gFOBT similarity, but three of the sets contained samples that were very different from their counterparts (on the PCA red=408N, purple=427N, orange=428N). This is confirmed by Figure 56 which shows that for sample 408N, the Bray-Curtis distance between the first subsample (arbitrarily taken as a reference) and the third subsample, exceeds the mean of the Bray-Curtis distances between the reference subsample and subsamples derived from the remaining ‘Temporal N1-3’ cards. Such a large difference in Bray-Curtis distances between subsamples was not observed in the ‘Temporal 1-6 samples’ (Figure 46).

These samples demonstrated a high degree of variability in the relative abundance of *Escherichia-Shigella* (Figure 58) The difference between the maximum and minimum relative abundance of each taxa (at genus level) was calculated across the three samples derived from each gFOBT card. The greatest difference was in the relative abundance of *Escherichia-Shigella* (Table 14).

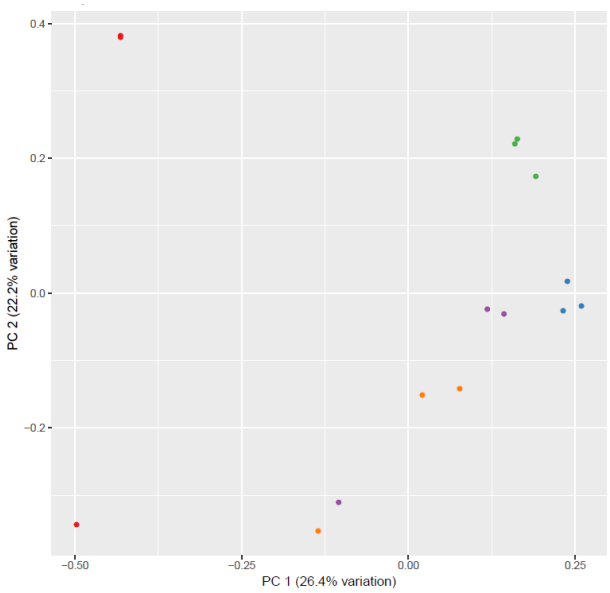


Figure 55. PCA of Bray-Curtis distances for Temporal N 1-3 samples. Points on the graph are coloured according to gFOBT sample.

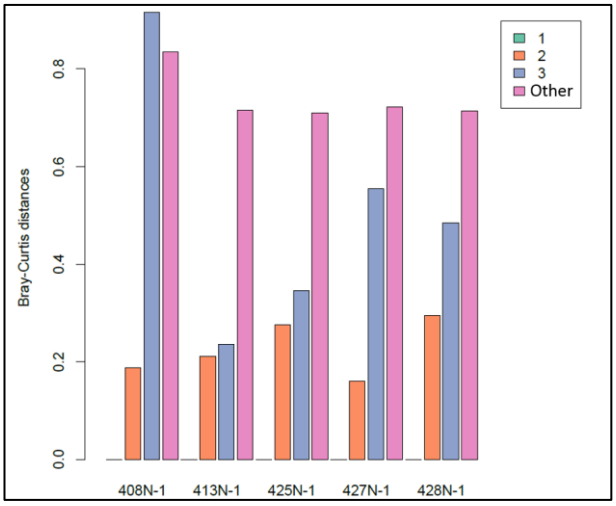


Figure 56. Bar chart of Bray-Curtis distances for Temporal N 1-3 samples. Four bars correspond to each gFOBT card. For each, the first three bars (coloured according to the key) show the Bray-Curtis distances between each of the subsamples and the first subsample. The fourth (pink) bar labelled 'Other' shows the mean of the Bray-Curtis distances between the first subsample of that gFOBT card and subsamples derived from all other 'Temporal N 1-3' gFOBT cards.

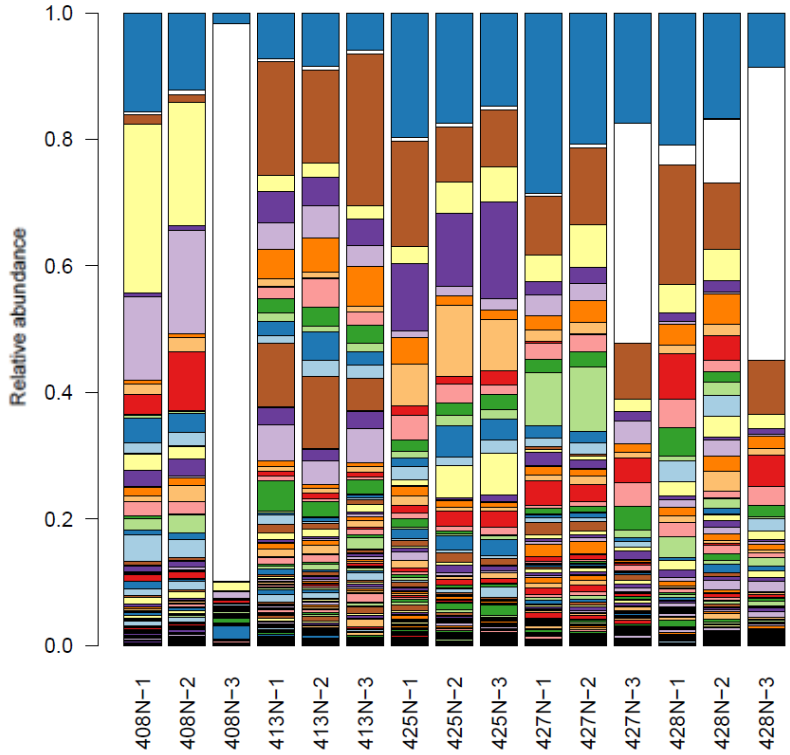


Figure 57. Taxonomy bar chart for Temporal N 1-3 samples. Each gFOBT sample is represented by three bars; each bar represents a subsample derived from a single gFOBT square. Colours denote the relative abundance of taxa (genus level); there are too many to include a legend. *Escherichia-Shigella* is white for ease of identification.

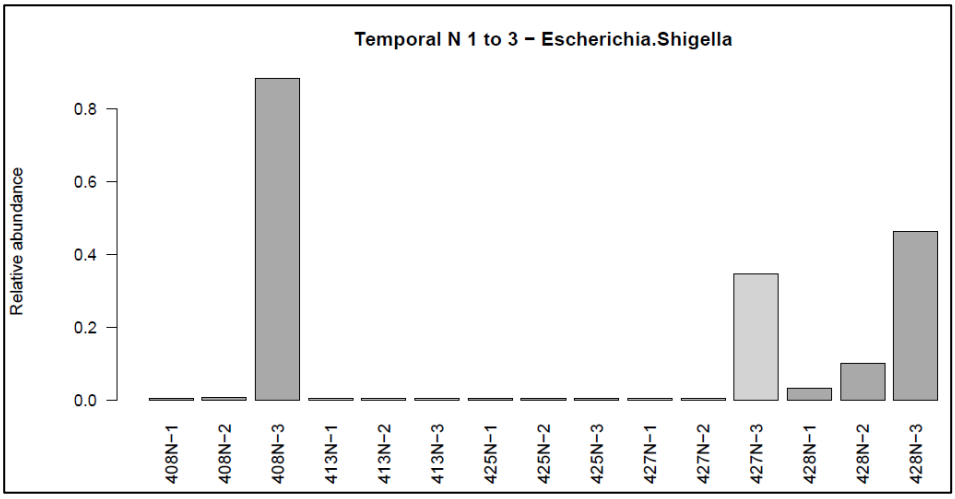
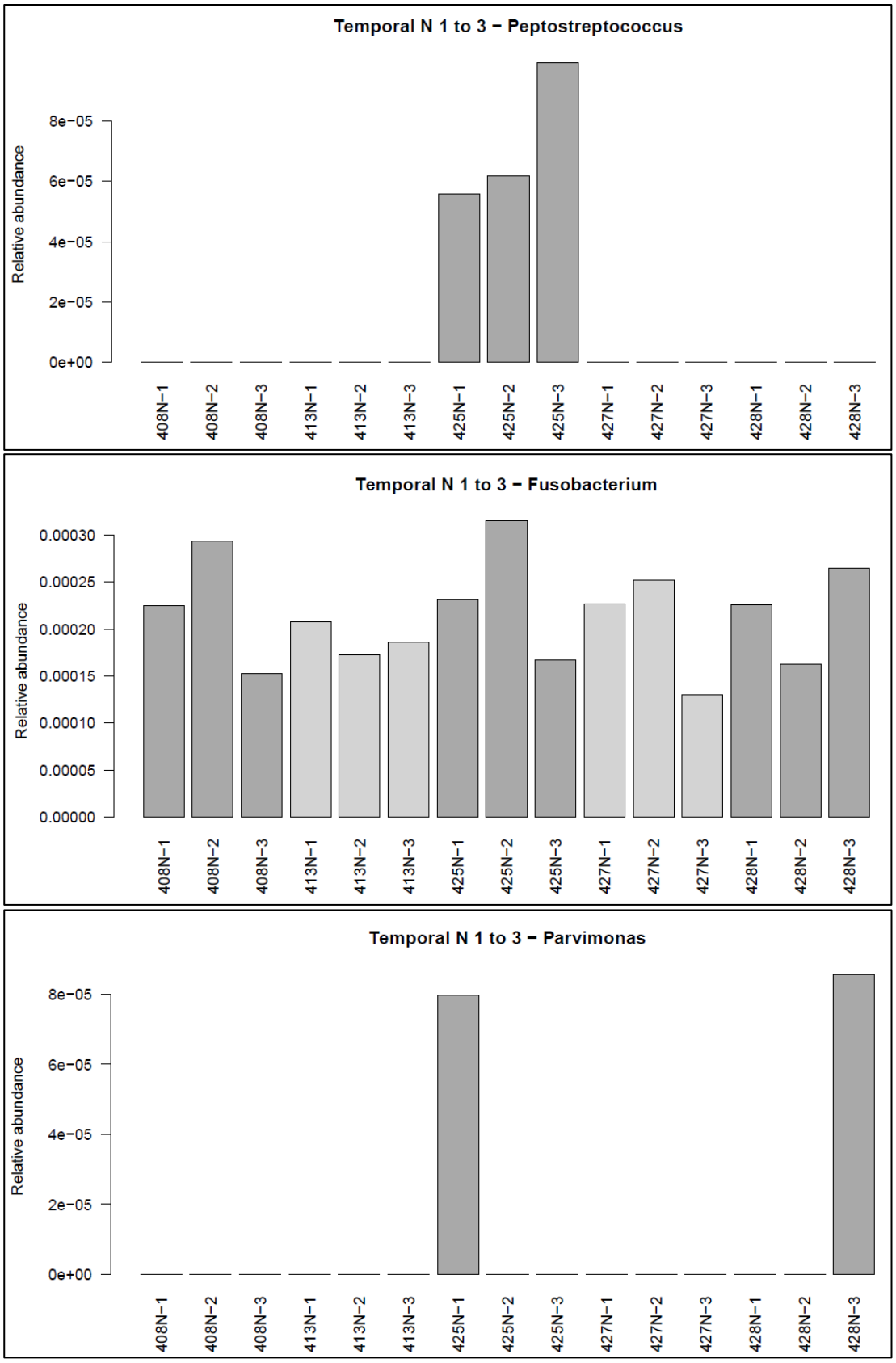


Figure 58. The relative abundance of *Escherichia-Shigella* for Temporal N 1-3 samples. Each gFOBT sample is represented by three bars; each bar represents a subsample derived from a single gFOBT square.

Table 14. The greatest difference in relative abundance of taxa (at genus level) across the three samples derived from each gFOBT card for Temporal N 1-3 samples.

gFOBT card	Greatest difference in relative abundance of taxa (at genus level) across the three subsamples	Taxa
408N	0.88	*Escherichia.Shigella*
428N	0.43	*Escherichia.Shigella*
427N	0.34	*Escherichia.Shigella*
413N	0.09	*Faecalibacterium*
425N	0.08	*Faecalibacterium*

CRC-associated bacteria demonstrated relatively less variability between replicate samples (Figure 59); especially when the scale of the y axis is taken into consideration. This suggests that the relative abundances of these CRC-associated taxa are relatively consistent between subsamples from three stools collected over three consecutive days.


Temporal N 1 to 3 − Peptostreptococcus
Relative abundance
8e−05
6e−05
4e−05
2e−05
0e+00
408N−1
408N−2
408N−3
413N−1
413N−2
413N−3
425N−1
425N−2
425N−3
427N−1
427N−2
427N−3
428N−1
428N−2
428N−3
Temporal N 1 to 3 − Fusobacterium
Relative abundance
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
408N−1
408N−2
408N−3
413N−1
413N−2
413N−3
425N−1
425N−2
425N−3
427N−1
427N−2
427N−3
428N−1
428N−2
428N−3
Temporal N 1 to 3 − Parvimonas
Relative abundance
8e−05
6e−05
4e−05
2e−05
0e+00
408N−1
408N−2
408N−3
413N−1
413N−2
413N−3
425N−1
425N−2
425N−3
427N−1
427N−2
427N−3
428N−1
428N−2
428N−3

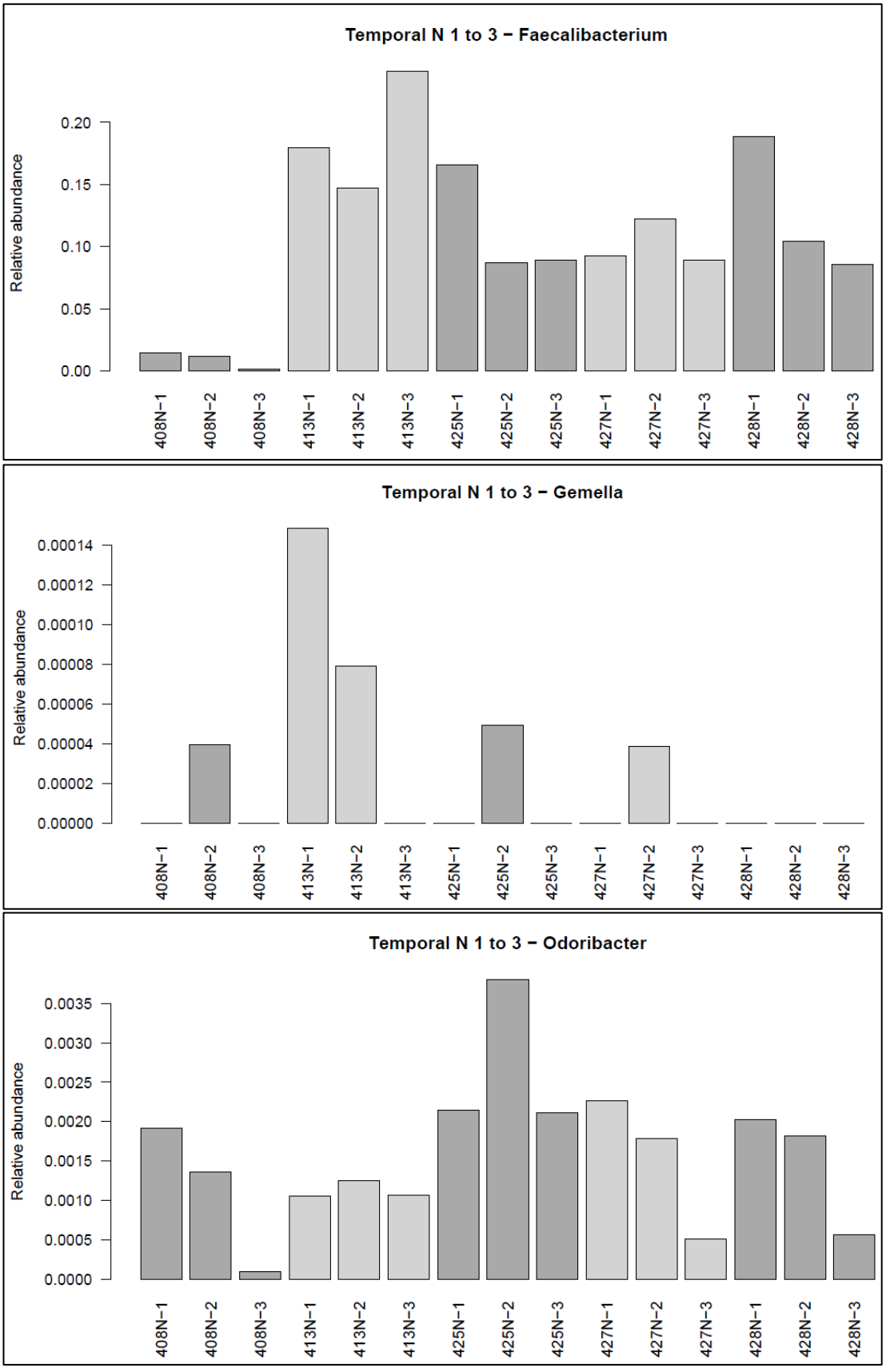


Figure 59. The relative abundance of CRC-associated taxa for Temporal N 1-3 samples. Each gFOBT sample is represented by three bars; each bar represents a subsample derived from a single gFOBT square. e = x 10^.

2.4.5.4 Temporal P 1-3 samples

Each ‘Temporal P1-3’ sample was derived from two squares of a gFOBT card, representing subsamples of three stool samples, collected over a range of days (maximum 12).

The PCA and bar plot of Bray-Curtis distances (Figure 60 and Figure 61) and taxonomy bar chart (Figure 62) demonstrated that the majority of extraction replicates were similar to one another. One sample showed a high degree of variability of *Escherichia-Shigella* relative abundance (2397P) (Figure 62 and Figure 63); the three stools were collected on consecutive days for this sample. The difference between the maximum and minimum relative abundance of each taxa (at genus level) was calculated across the three samples derived from each gFOBT card. The greatest difference was in the relative abundance of *Escherichia-Shigella* (Table 15).

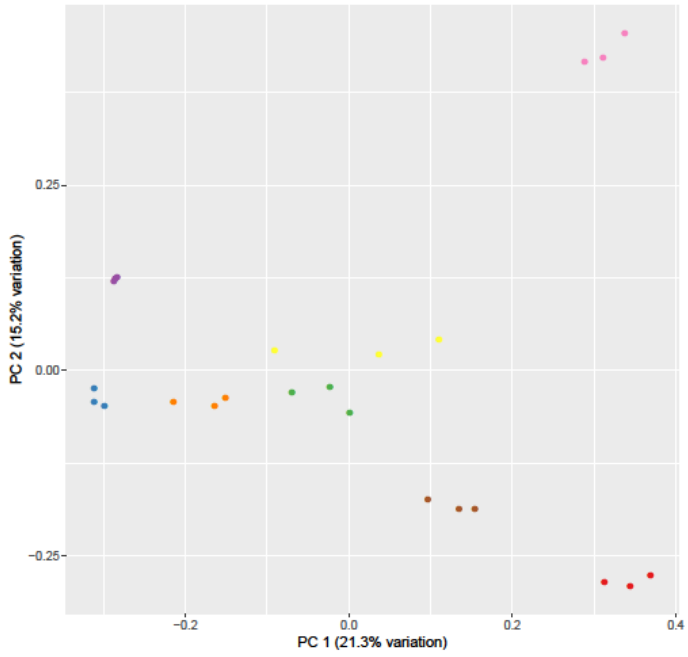


Figure 60. PCA of Bray-Curtis distances for Temporal P 1-3 samples. Points on the graph are coloured according to gFOBT sample.

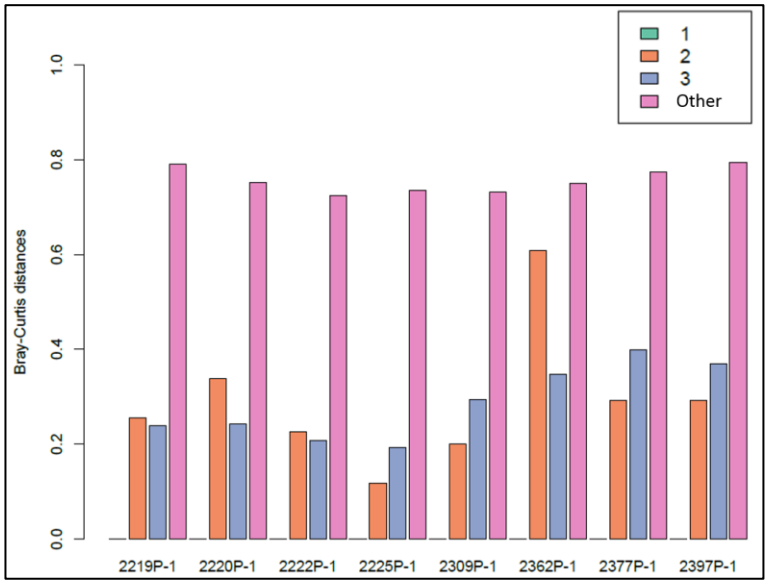


Figure 61. Bar chart of Bray-Curtis distances for Temporal P 1-3 samples. Four bars correspond to each gFOBT card. For each, the first three bars (coloured according to the key) show the Bray-Curtis distances between each of the subsamples and the first subsample. The fourth (pink) bar labelled 'Other' shows the mean of the Bray-Curtis distances between the first subsample of that gFOBT card and subsamples derived from all other 'P 1-3' gFOBT cards.

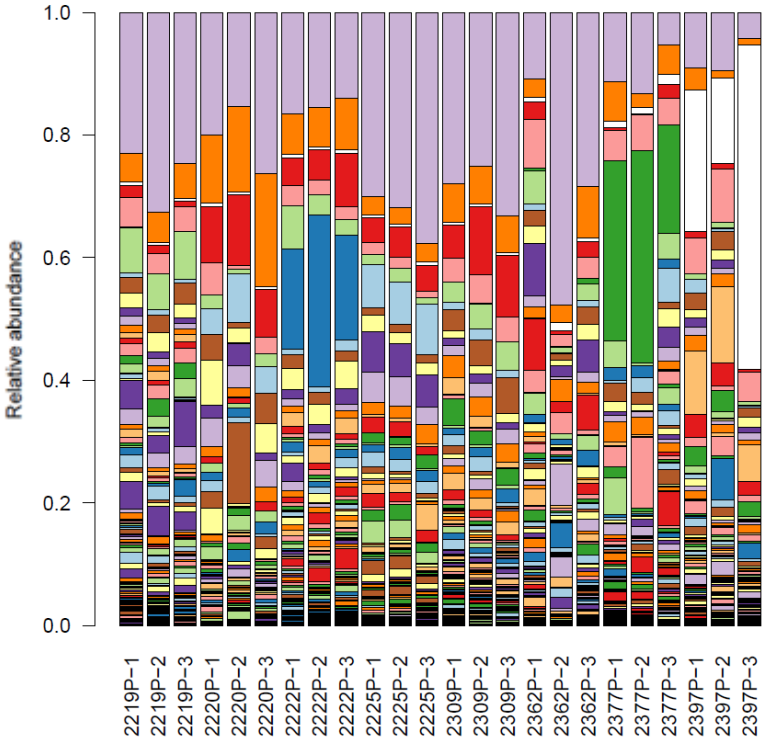


Figure 62. Taxonomy bar chart for Temporal P 1-3 samples. Each gFOBT sample is represented by three bars; each bar represents a subsample derived from two gFOBT squares. Colours denote the relative abundance of taxa (genus level); there are too many to include a legend. *Escherichia-Shigella* is white for ease of identification.

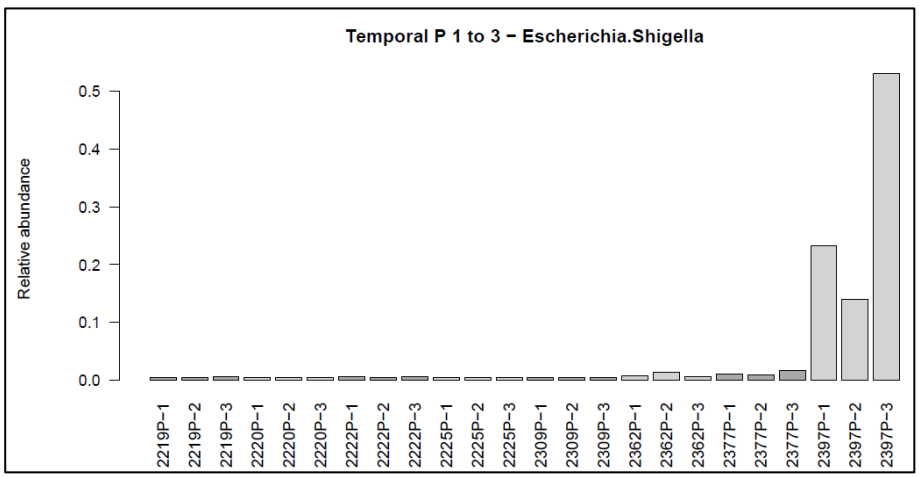
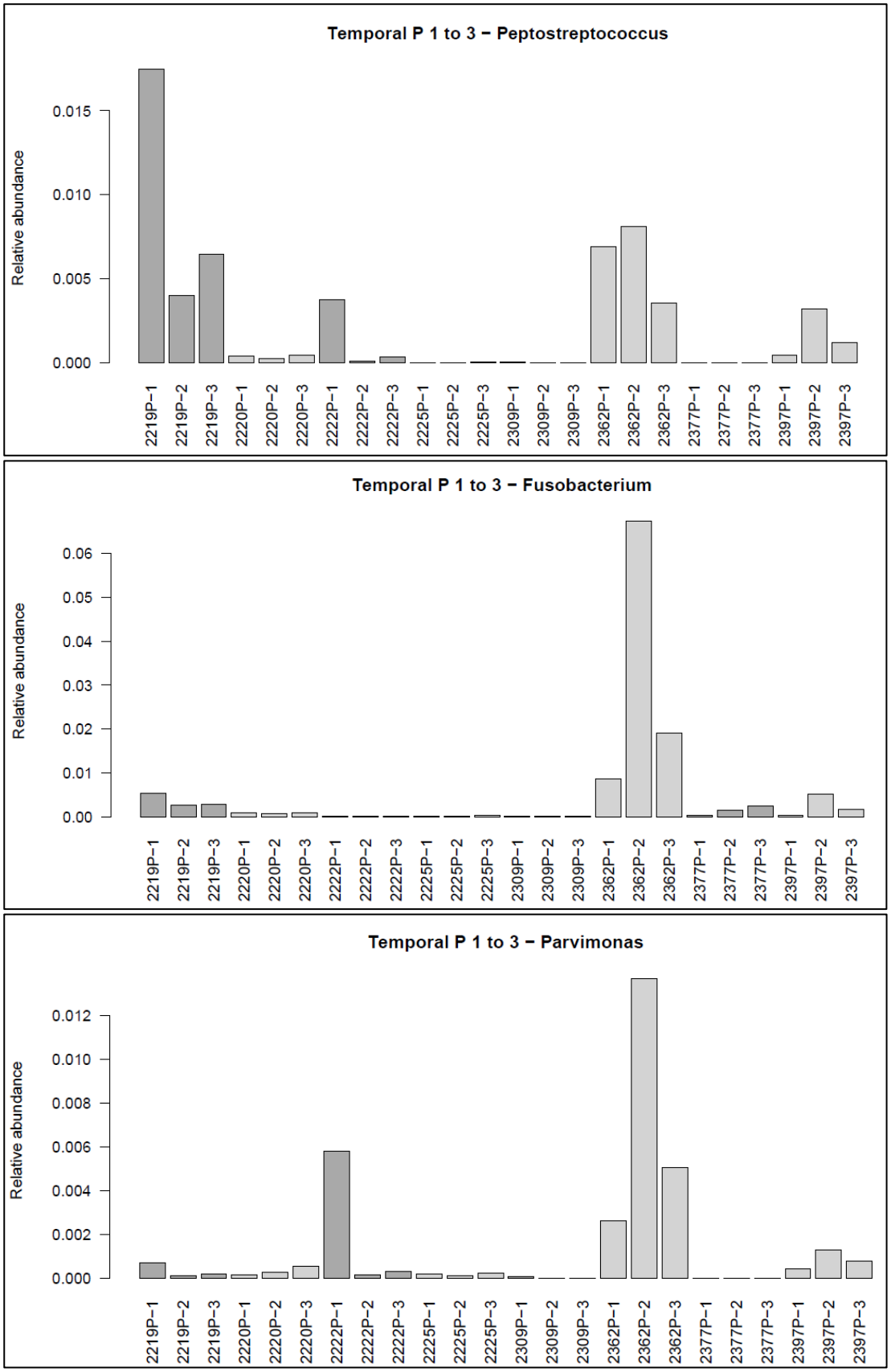


Figure 63. The relative abundance of *Escherichia-Shigella* for Temporal P 1-3 samples. Each gFOBT sample is represented by three bars; each bar represents a subsample derived from two gFOBT squares.

Table 15. The greatest difference in relative abundance of taxa (at genus level) across the three samples derived from each gFOBT card for Temporal P 1-3 samples.

gFOBT card	Greatest difference in relative abundance of taxa (at genus level) across the three subsamples	Taxa
2397P	0.39	*Escherichia.Shigella*
2362P	0.37	*Bacteroides*
2377P	0.17	*Bacteroidales.S24.7.group (genus not specified)*
2222P	0.12	*Rikenellaceae.RC9.gut.group*
2220P	0.11	*Haemophilus*
2219P	0.10	*Bacteroides*
2309P	0.08	*Bacteroides*
2225P	0.08	*Bacteroides*

CRC-associated bacteria demonstrated relatively less variability between replicate samples (Figure 64); especially when the scale of the y axis is taken into consideration. This suggests that the relative abundances of these CRC-associated taxa are relatively consistent between subsamples from three stools collected over a range of days.


Temporal P 1 to 3 − Peptostreptococcus
Relative abundance
0.015
0.010
0.005
0.000
2219P−1
2219P−2
2219P−3
2220P−1
2220P−2
2220P−3
2222P−1
2222P−2
2222P−3
2225P−1
2225P−2
2225P−3
2309P−1
2309P−2
2309P−3
2362P−1
2362P−2
2362P−3
2377P−1
2377P−2
2377P−3
2397P−1
2397P−2
2397P−3
Temporal P 1 to 3 − Fusobacterium
Relative abundance
0.06
0.05
0.04
0.03
0.02
0.01
0.00
2219P−1
2219P−2
2219P−3
2220P−1
2220P−2
2220P−3
2222P−1
2222P−2
2222P−3
2225P−1
2225P−2
2225P−3
2309P−1
2309P−2
2309P−3
2362P−1
2362P−2
2362P−3
2377P−1
2377P−2
2377P−3
2397P−1
2397P−2
2397P−3
Temporal P 1 to 3 − Parvimonas
Relative abundance
0.012
0.010
0.008
0.006
0.004
0.002
0.000
2219P−1
2219P−2
2219P−3
2220P−1
2220P−2
2220P−3
2222P−1
2222P−2
2222P−3
2225P−1
2225P−2
2225P−3
2309P−1
2309P−2
2309P−3
2362P−1
2362P−2
2362P−3
2377P−1
2377P−2
2377P−3
2397P−1
2397P−2
2397P−3

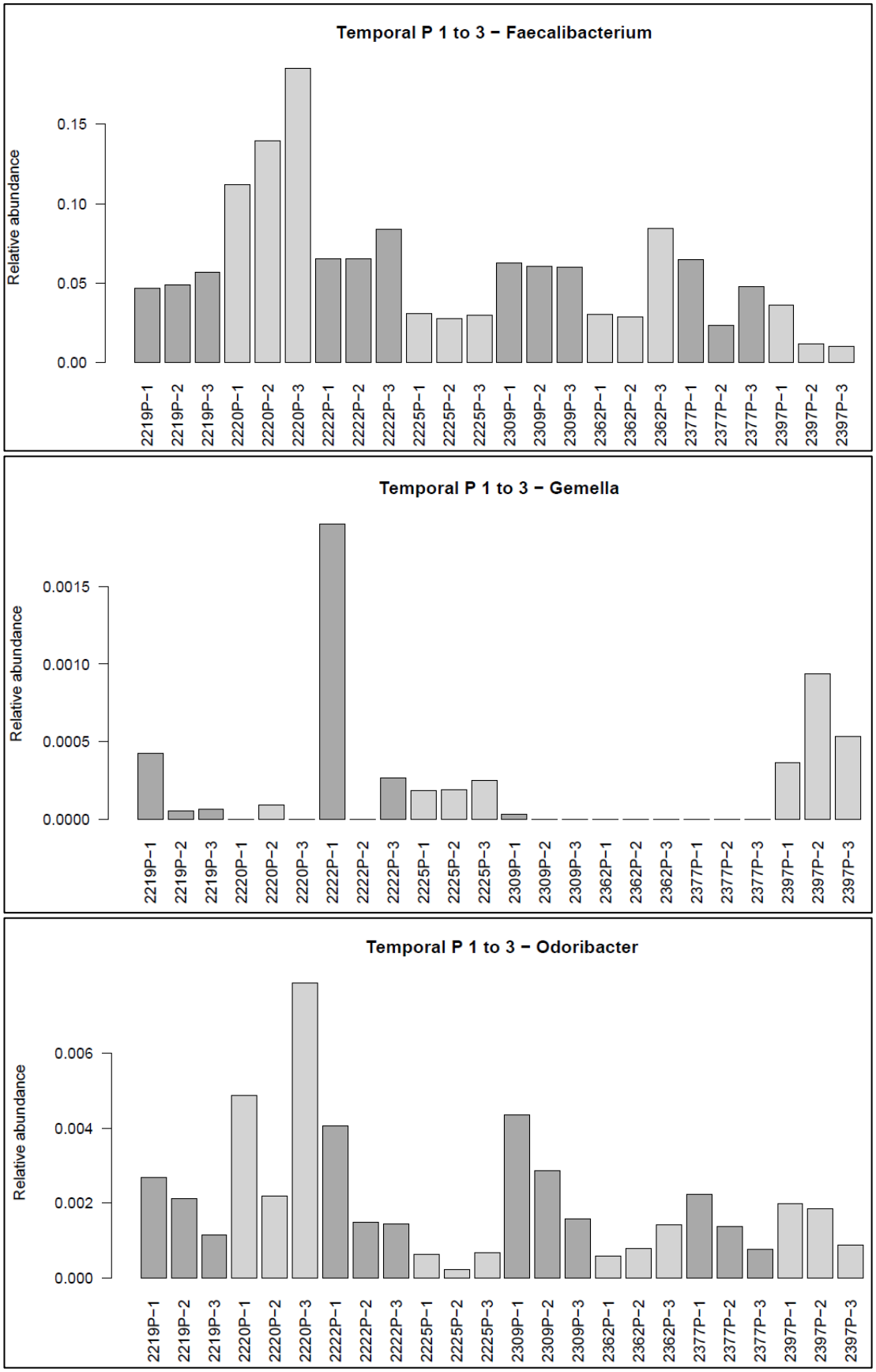


Figure 64. The relative abundance of CRC-associated taxa for Temporal P 1-3 samples. Each gFOBT sample is represented by three bars; each bar represents a subsample derived from two gFOBT squares.

2.4.6 FIT experiment

2.4.6.1 All samples

The PCA of Bray-Curtis distances (Figure 65) demonstrated that samples cluster by participant and stool sample of origin. PERMANOVA analysis of the variables: 'stool sample of origin' (i.e. A1, A2, B1 or B2), 'extraction day' and 'sample type/temperature of storage', confirmed that 'stool sample' contributes to the largest amount of variation in Bray-Curtis distance (Table 16). 'Type of sample' and 'extraction day' contributed significant but small amounts to variation in Bray-Curtis distance; variation due to 'extraction day' is illustrated in Figure 66.

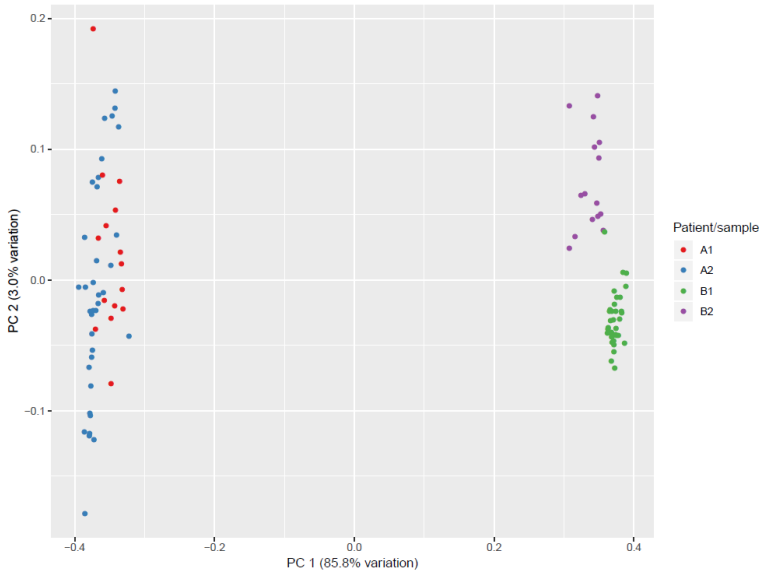


Figure 65. PCA of Bray-Curtis distances of all of the samples processed as part of the FIT experiment. Points on the graph are coloured according to the legend. Stool samples A1 and A2 derive from participant A. Stool samples B1 and B2 derive from participant B.

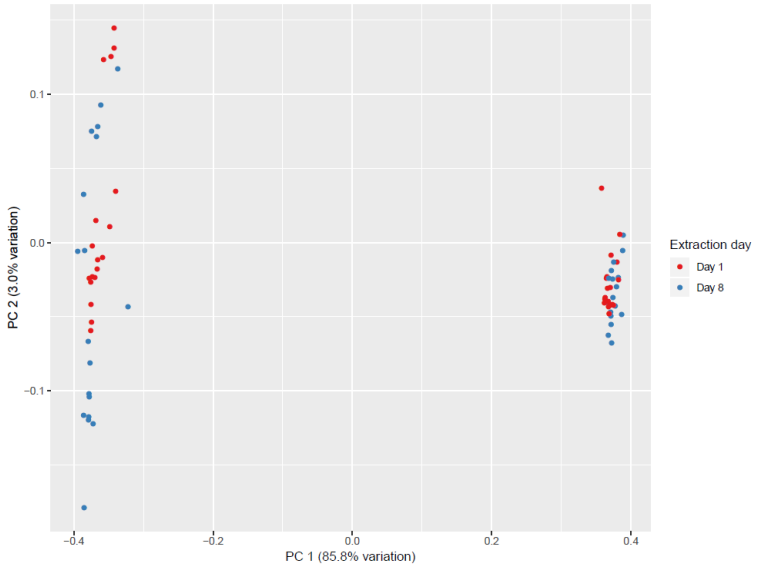


Figure 66. PCA of Bray-Curtis distances for samples extracted on day 1 and day 8 of the FIT experiment, which were derived from the same stool (A2 or B1). Points are coloured according to the day that DNA extraction was performed. Points derived from stool A2 are on the left of the plot; points derived from stool B1 are on the right of the plot.

Table 16. Results of PERMANOVA analysis of 'FIT experiment' samples. Df = degrees of freedom. Type of sample includes: gFOBT, stool, FIT stored at room temperature, FIT stored at 4°C, FIT stored at -80°C. Stool sample of origin denotes stool samples A1, A2, B1 or B2. NA = not applicable. Significant p values are shaded grey. Values are recorded to two decimal places.

	Df	Sums of squares	F.Model	R^2	Pr(>F)
Type of sample	4	0.35	6.47	0.02	2×10^{-4}
DNA extraction day	1	0.05	3.96	3.49×10^{-3}	4.14×10^{-2}
Stool sample of origin	3	13.74	338.02	0.89	1×10^{-4}
Residuals	93	1.26	NA	0.08	NA
Total	101	15.40	NA	1.00	NA

Boxplots indicated that Bray-Curtis distances were smaller for samples derived from a single stool sample or participant than Bray-Curtis distances between samples derived from different stool samples or participants (Figure 67). This is confirmed in Figure 68, which shows the Bray-Curtis distances between one of the stool samples stored at -80°C (arbitrarily taken as a reference) and the remaining samples. Bray-Curtis distances between the reference and replicate stool samples stored at -80°C appear slightly lower than between the reference and FIT or gFOBT samples; however there is no distinction in Bray-Curtis distances between the reference and FIT or the reference and gFOBT samples, nor is there a trend in Bray-Curtis distances between the reference and FIT samples stored under different conditions or extracted on different days.

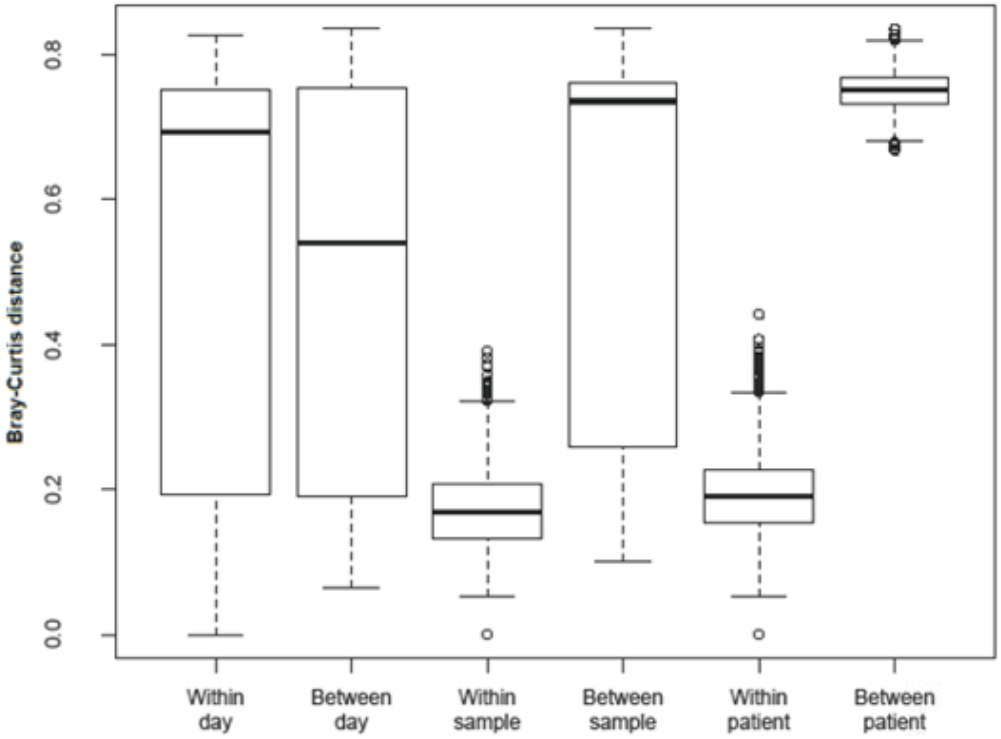
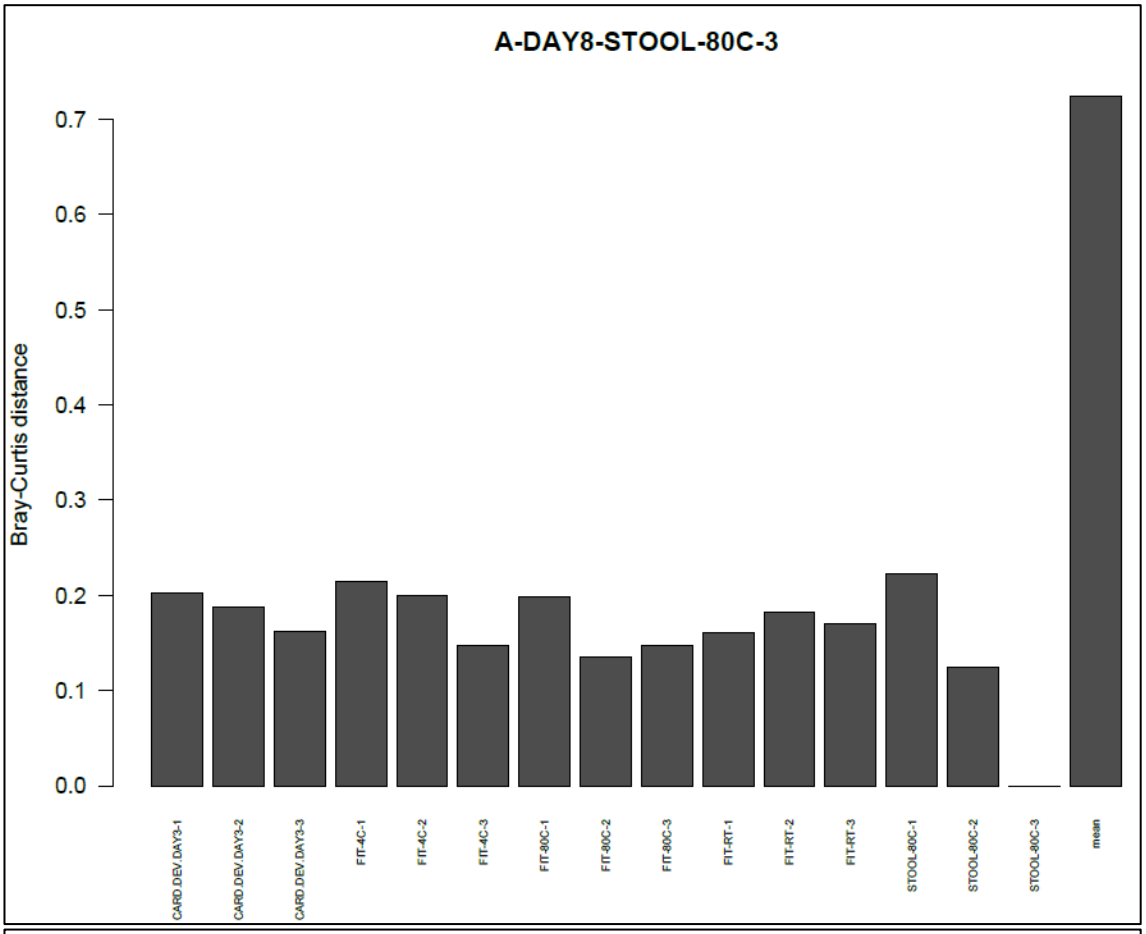
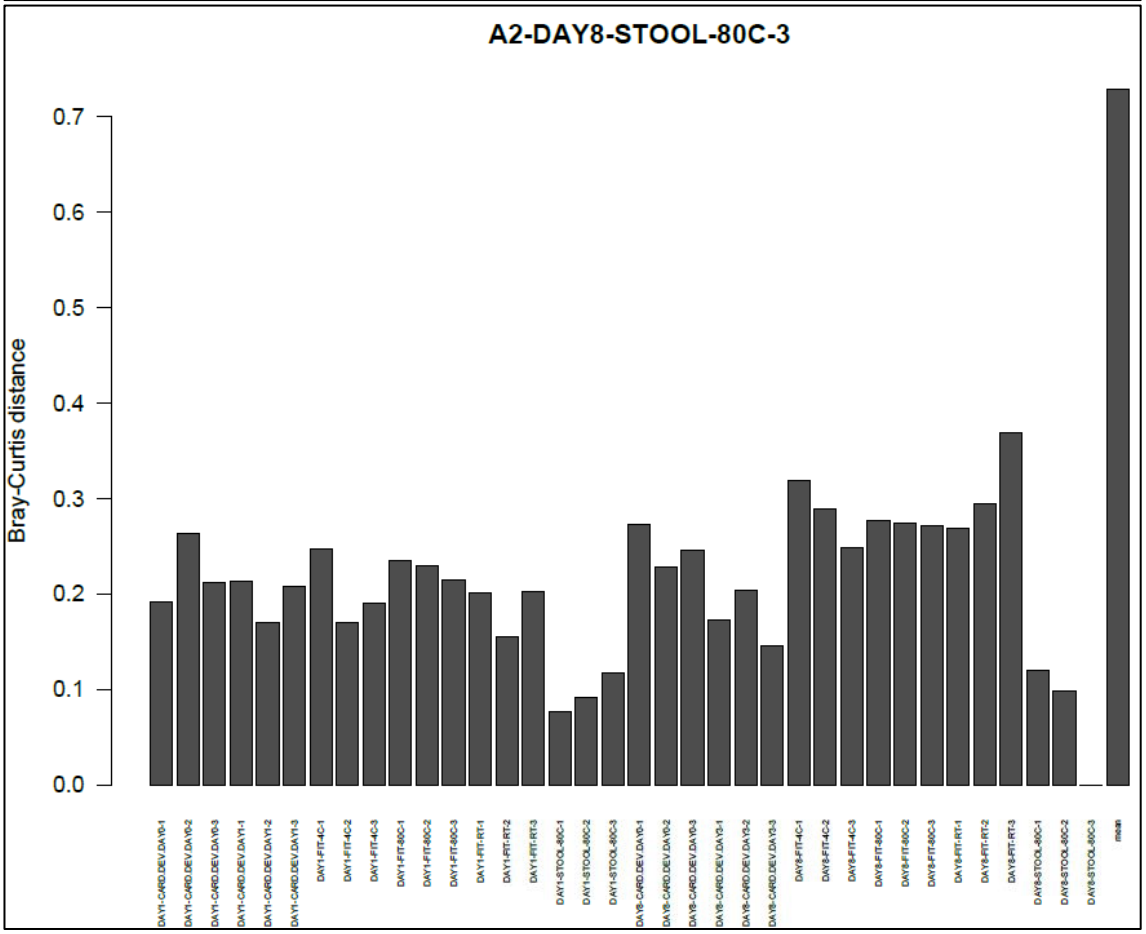
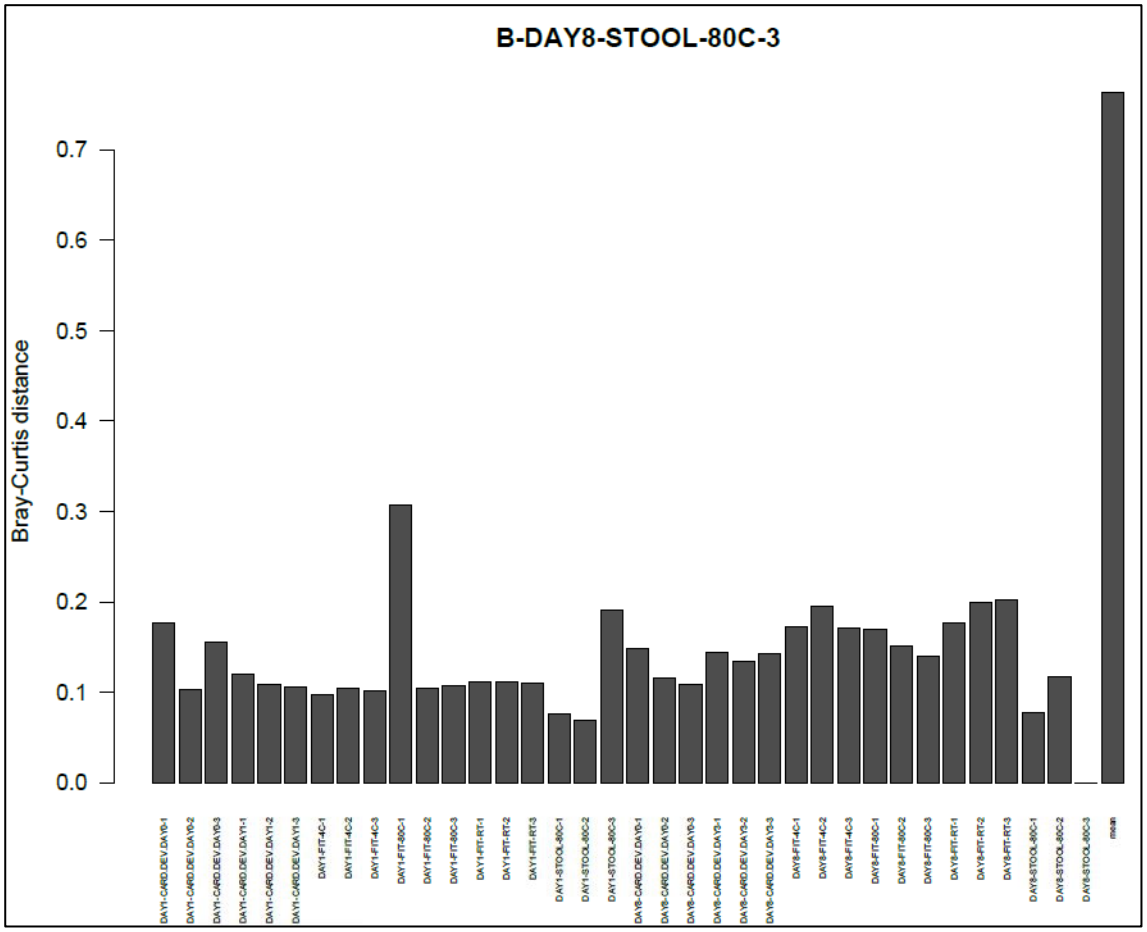


Figure 67. Box plots of Bray-Curtis distances for FIT-experiment samples. The two left-hand boxplots depict the range of Bray-Curtis distances within all samples extracted on the same day and between samples extracted on different days respectively. The third and fourth boxplots depict the range of Bray-Curtis distances within all samples derived from a single stool sample and between all samples derived from different stool samples respectively. The two right-hand boxplots depict the range of Bray-Curtis distances within all samples derived from a single participant and between all samples derived from the two different participants respectively.


A-DAY8-STOOL-80C-3
Bray-Curtis distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
CARD.DEV.DAY3-1
CARD.DEV.DAY3-2
CARD.DEV.DAY3-3
FIT-4C-1
FIT-4C-2
FIT-4C-3
FIT-80C-1
FIT-80C-2
FIT-80C-3
FIT-RT-1
FIT-RT-2
FIT-RT-3
STOOL-80C-1
STOOL-80C-2
STOOL-80C-3
mean


A2-DAY8-STOOL-80C-3
Bray-Curtis distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
DAY1-CARD.DEV.DAY8-1
DAY1-CARD.DEV.DAY8-2
DAY1-CARD.DEV.DAY8-3
DAY1-CARD.DEV.DAY1-1
DAY1-CARD.DEV.DAY1-2
DAY1-CARD.DEV.DAY1-3
DAY1-FIT-4C-1
DAY1-FIT-4C-2
DAY1-FIT-4C-3
DAY1-FIT-80C-1
DAY1-FIT-80C-2
DAY1-FIT-80C-3
DAY1-FIT-RT-1
DAY1-FIT-RT-2
DAY1-FIT-RT-3
DAY1-STOOL-80C-1
DAY1-STOOL-80C-2
DAY1-STOOL-80C-3
DAY8-CARD.DEV.DAY8-1
DAY8-CARD.DEV.DAY8-2
DAY8-CARD.DEV.DAY8-3
DAY8-CARD.DEV.DAY1-1
DAY8-CARD.DEV.DAY1-2
DAY8-CARD.DEV.DAY1-3
DAY8-FIT-4C-1
DAY8-FIT-4C-2
DAY8-FIT-4C-3
DAY8-FIT-80C-1
DAY8-FIT-80C-2
DAY8-FIT-80C-3
DAY8-FIT-RT-1
DAY8-FIT-RT-2
DAY8-FIT-RT-3
DAY8-STOOL-80C-1
DAY8-STOOL-80C-2
DAY8-STOOL-80C-3
mean


B-DAY8-STOOL-80C-3
Bray-Curtis distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
DAY1-CARD.DEV.DAY8-1
DAY1-CARD.DEV.DAY8-2
DAY1-CARD.DEV.DAY8-3
DAY1-CARD.DEV.DAY1-1
DAY1-CARD.DEV.DAY1-2
DAY1-CARD.DEV.DAY1-3
DAY1-FIT-4C-1
DAY1-FIT-4C-2
DAY1-FIT-4C-3
DAY1-FIT-80C-1
DAY1-FIT-80C-2
DAY1-FIT-80C-3
DAY1-FIT-RT-1
DAY1-FIT-RT-2
DAY1-FIT-RT-3
DAY1-STOOL-80C-1
DAY1-STOOL-80C-2
DAY1-STOOL-80C-3
DAY8-CARD.DEV.DAY8-1
DAY8-CARD.DEV.DAY8-2
DAY8-CARD.DEV.DAY8-3
DAY8-CARD.DEV.DAY9-1
DAY8-CARD.DEV.DAY9-2
DAY8-CARD.DEV.DAY9-3
DAY8-FIT-4C-1
DAY8-FIT-4C-2
DAY8-FIT-4C-3
DAY8-FIT-80C-1
DAY8-FIT-80C-2
DAY8-FIT-80C-3
DAY8-FIT-RT-1
DAY8-FIT-RT-2
DAY8-FIT-RT-3
DAY8-STOOL-80C-1
DAY8-STOOL-80C-2
DAY8-STOOL-80C-3
mean

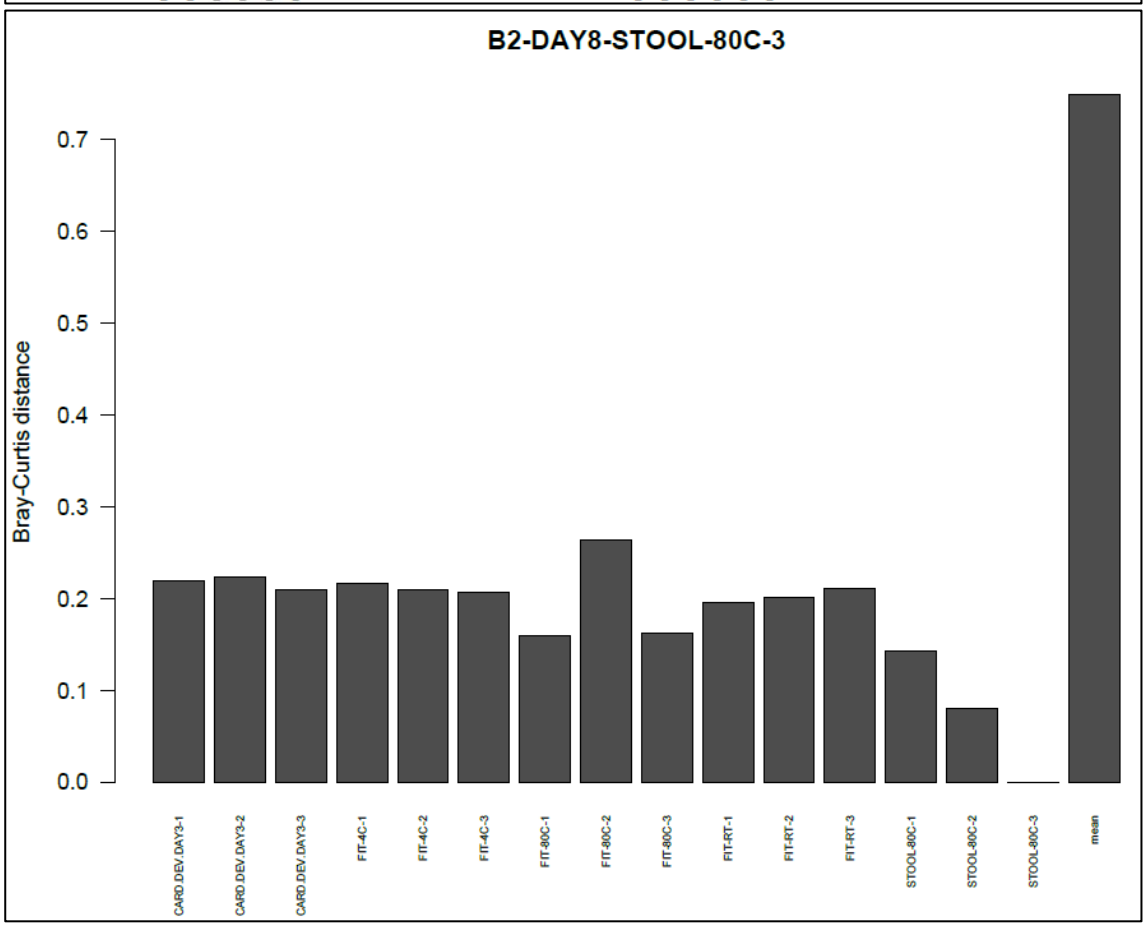

B2-DAY8-STOOL-80C-3
Bray-Curtis distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
CARD.DEV.DAY3-1
CARD.DEV.DAY3-2
CARD.DEV.DAY3-3
FIT-4C-1
FIT-4C-2
FIT-4C-3
FIT-80C-1
FIT-80C-2
FIT-80C-3
FIT-RT-1
FIT-RT-2
FIT-RT-3
STOOL-80C-1
STOOL-80C-2
STOOL-80C-3
mean

Figure 68. Bar chart of Bray-Curtis distances for samples sequenced as part of the FIT experiment. The four plots show data corresponding to stool sample A, A2, B and B2 (from top to bottom, respectively). Within each plot, each bar shows the Bray-Curtis distances between that sample and the reference sample (one of the stool samples stored at -80°C and extracted on day 8, triplicate number 3). The far right-hand bar shows the mean of the Bray-Curtis distances between the reference sample and samples derived from the other participant. Each bar is labelled as follows: extraction day (if not labelled, this was day 8); sample type (where CARD.DEV.DAY0/1/3 = gFOBT developed on day 0/1/3 respectively); storage temperature; triplicate number.

Taxonomy bar charts (Figure 69) indicated a greater difference in taxonomic composition between participants A and B than between stool samples derived from the same participant. There was a degree of variation of taxonomic composition across all of the samples, however a trend associated with a certain sample type or extraction day was not apparent.

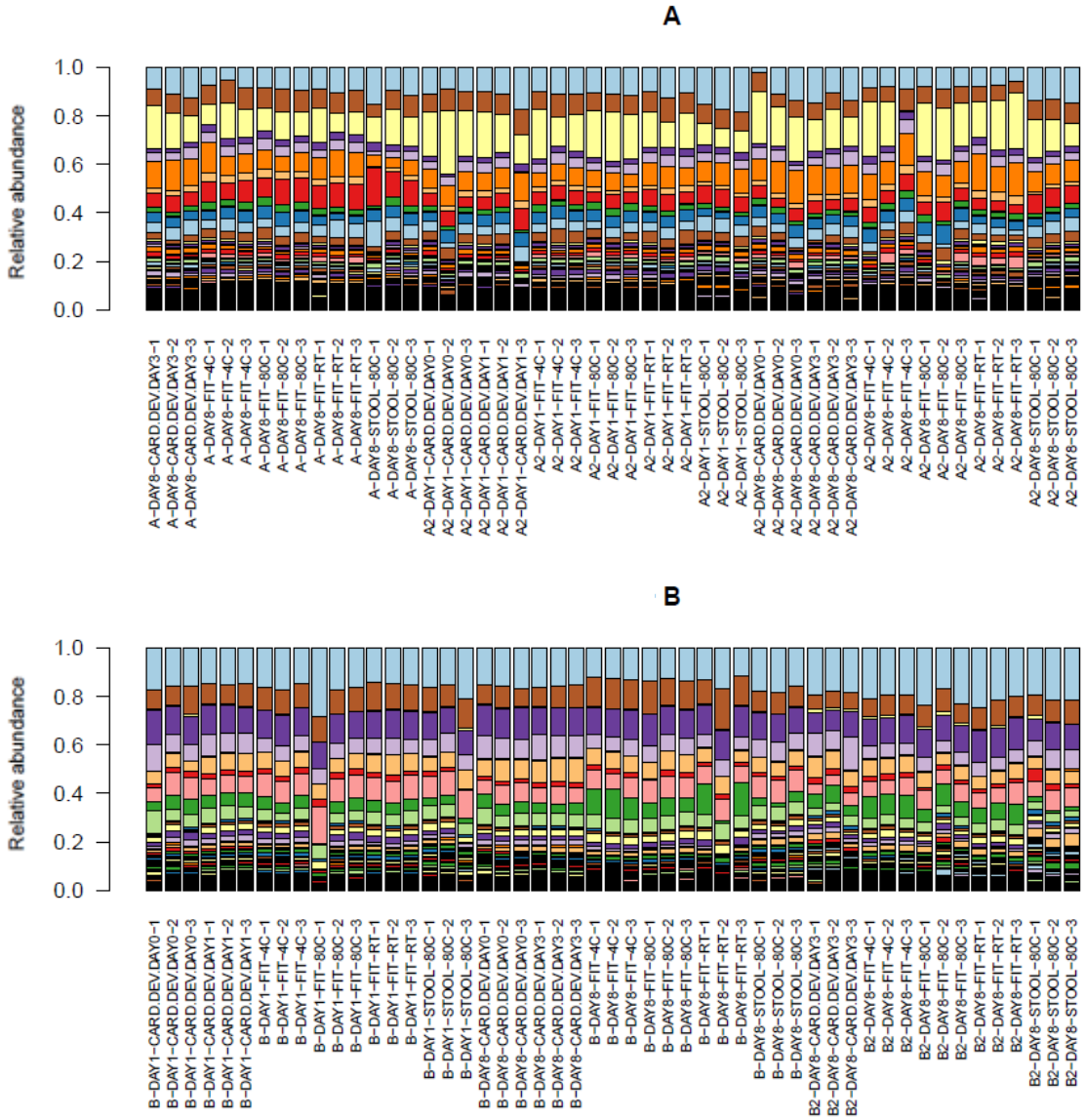


Figure 69. Taxonomy bar charts of all the samples sequenced as part of the FIT experiment. Each bar represents a sample. Colours denote the relative abundance of taxa (genus level); there are too many to include a legend. Samples derived from participant A (stool samples A and A2) are displayed in the upper chart; samples derived from participant B (stool samples B and B2) are displayed in the lower chart. Each bar is labelled as follows: stool from which the sample was derived; extraction day; sample type (where CARD.DEV.DAY0/1/3 = gFOBT developed on day 0/1/3 respectively); storage temperature; triplicate number.

CRC-associated bacteria showed a degree of variability between samples (Figure 70 to Figure 75); however the scale of the y axis should be taken into consideration. Some of the variability occurred between triplicates of a single sample type, suggesting variation secondary to subsampling or chance rather than sample type itself. There did not appear to be a trend associated with sample

type. This suggests that all sample types are appropriate for the analysis of these CRC-associated taxa. However, this should be confirmed using patient samples, where the relative abundances of such taxa are expected to be higher. *Escherichia-Shigella* demonstrated minimal variation in relative abundance between samples (note the y axis) (Figure 76).

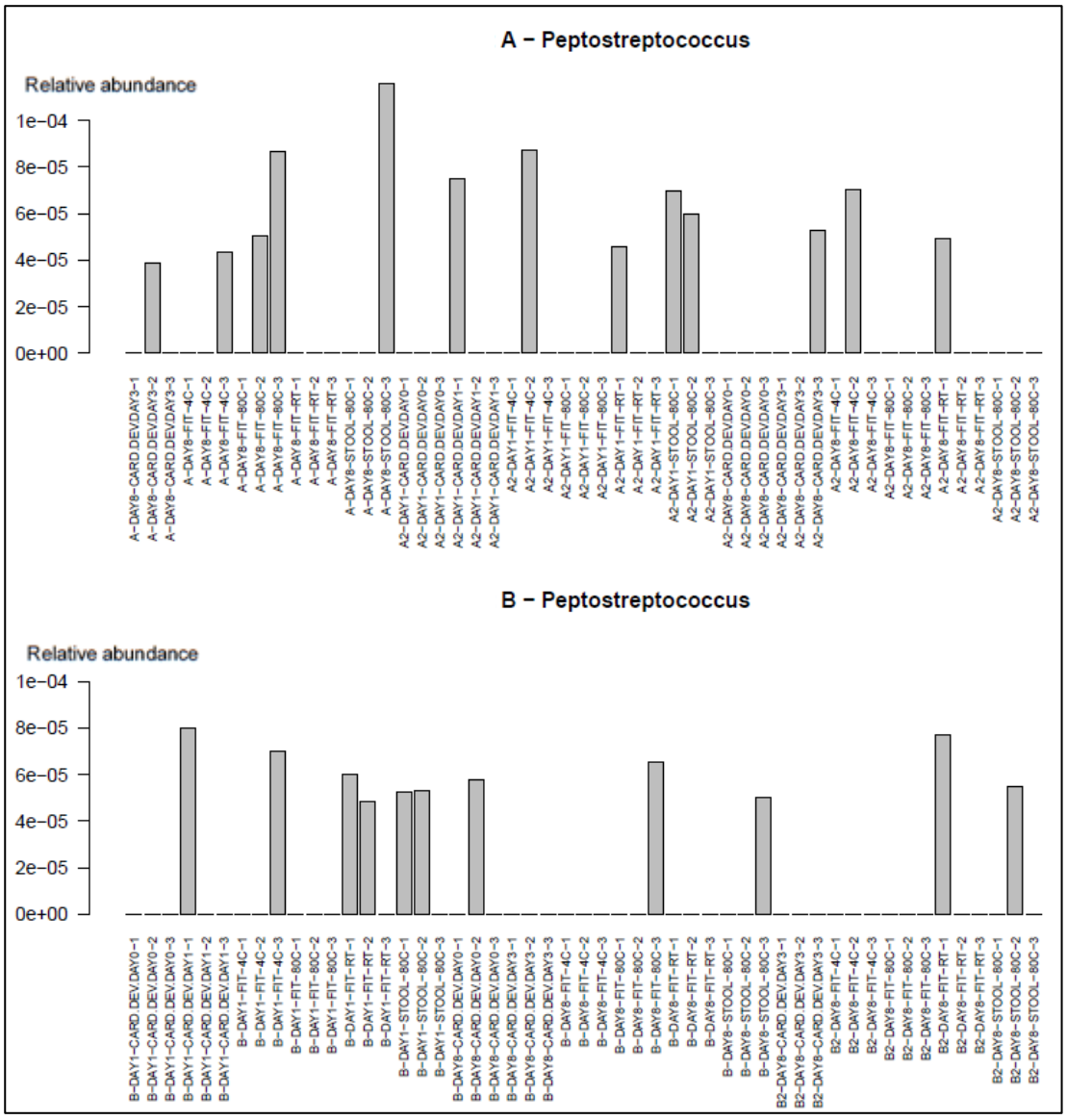


Figure 70. The relative abundance of *Peptostreptococcus* of all the samples sequenced as part of the FIT experiment. Each bar represents a sample. Samples derived from participant A (stool samples A and A2) are displayed in the upper chart; samples derived from participant B (stool samples B and B2) are displayed in the lower chart. Each bar is labelled as follows: stool which the sample was derived from; extraction day; sample type (where CARD.DEV.DAY0/1/3 = gFOBT developed on day 0/1/3 respectively); storage temperature; triplicate number. e = x 10^.

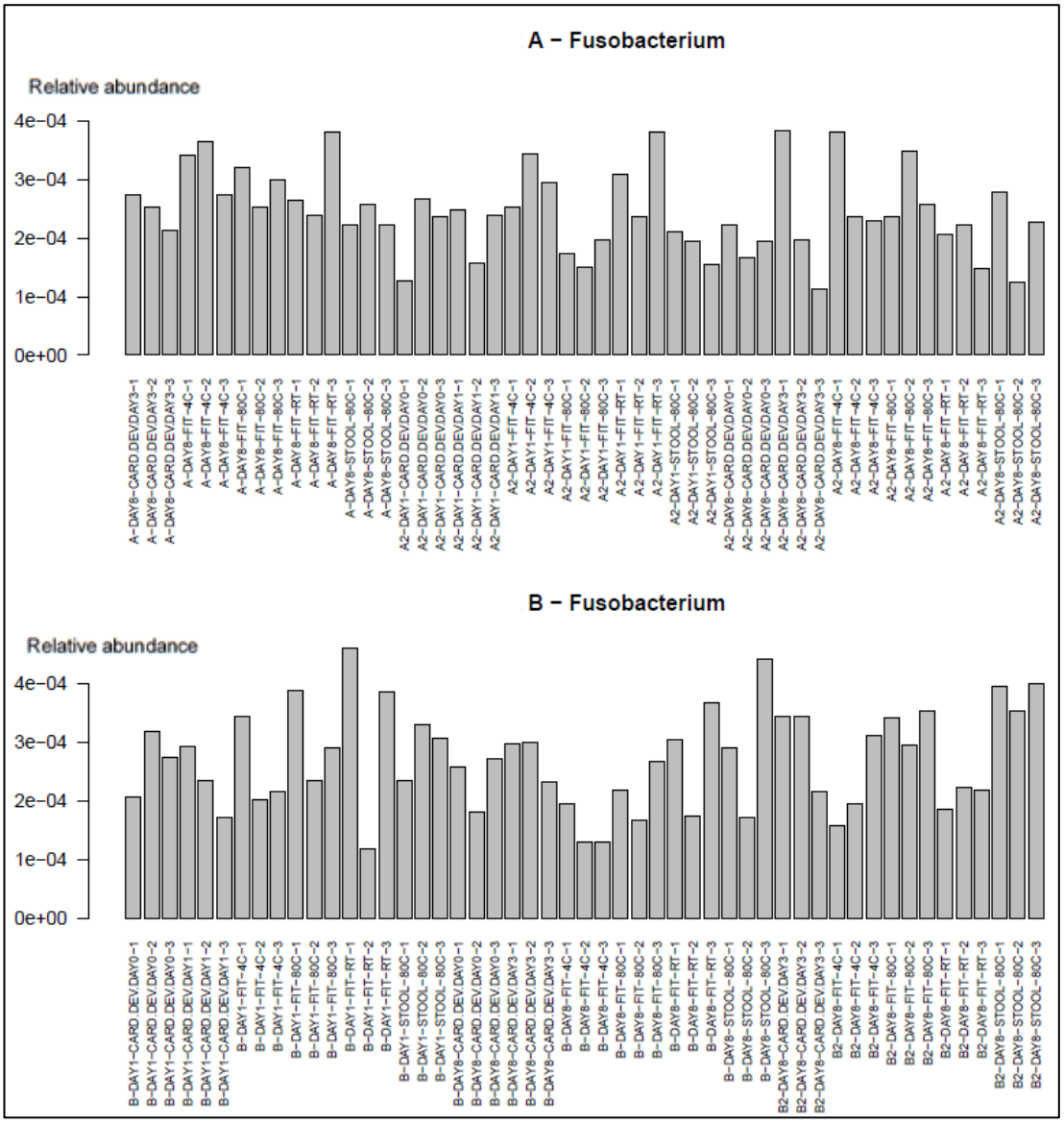


Figure 71. The relative abundance of *Fusobacterium* of all the samples sequenced as part of the FIT experiment. Each bar represents a sample. Samples derived from participant A (stool samples A and A2) are displayed in the upper chart; samples derived from participant B (stool samples B and B2) are displayed in the lower chart. Each bar is labelled as follows: stool which the sample was derived from; extraction day; sample type (where CARD.DEV.DAY0/1/3 = gFOBT developed on day 0/1/3 respectively); storage temperature; triplicate number. e = x 10^.

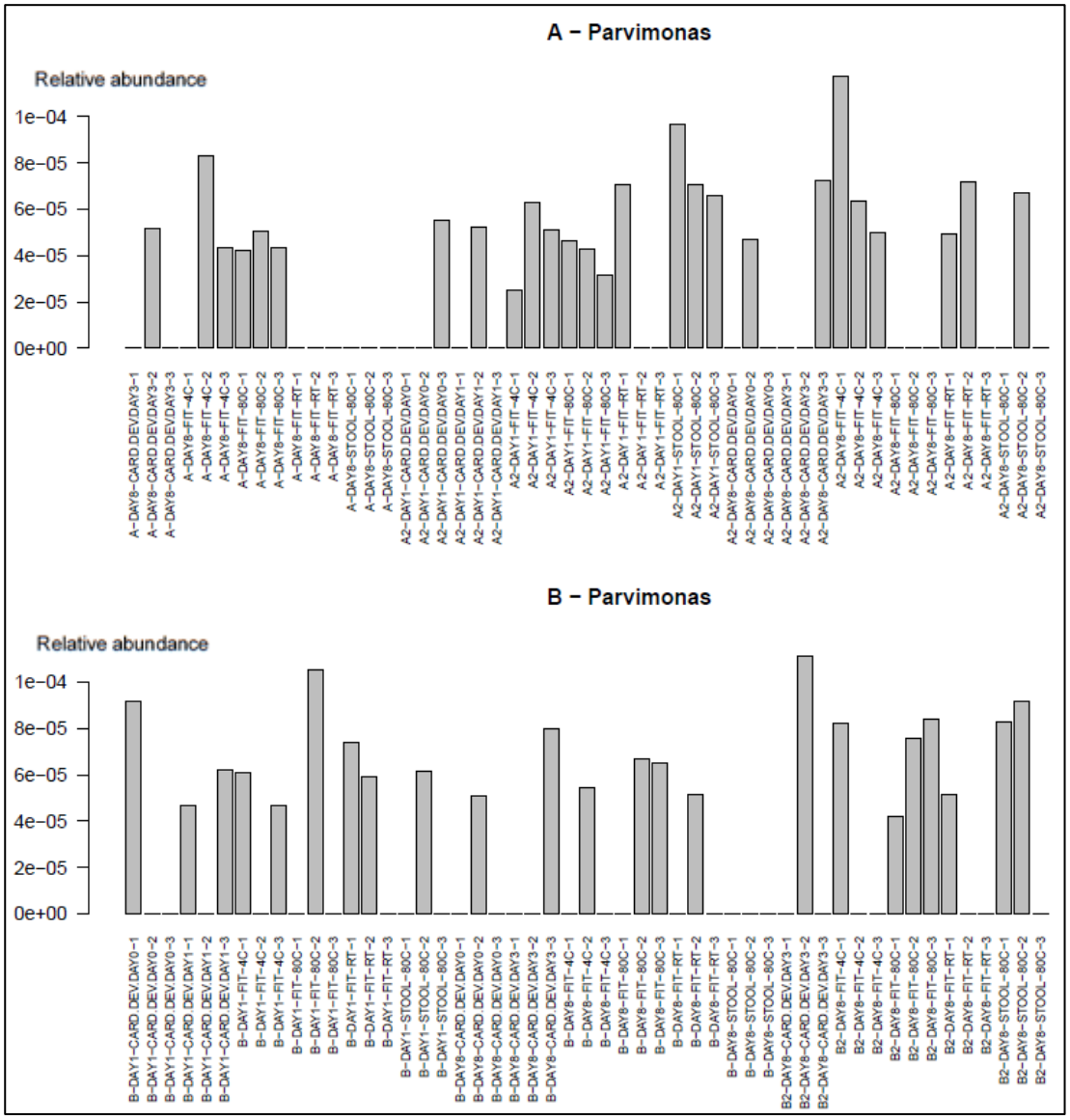


Figure 72. The relative abundance of *Parvimonas* of all the samples sequenced as part of the FIT experiment. Each bar represents a sample. Samples derived from participant A (stool samples A and A2) are displayed in the upper chart; samples derived from participant B (stool samples B and B2) are displayed in the lower chart. Each bar is labelled as follows: stool which the sample was derived from; extraction day; sample type (where CARD.DEV.DAY0/1/3 = gFOBT developed on day 0/1/3 respectively); storage temperature; triplicate number. e = x 10^.

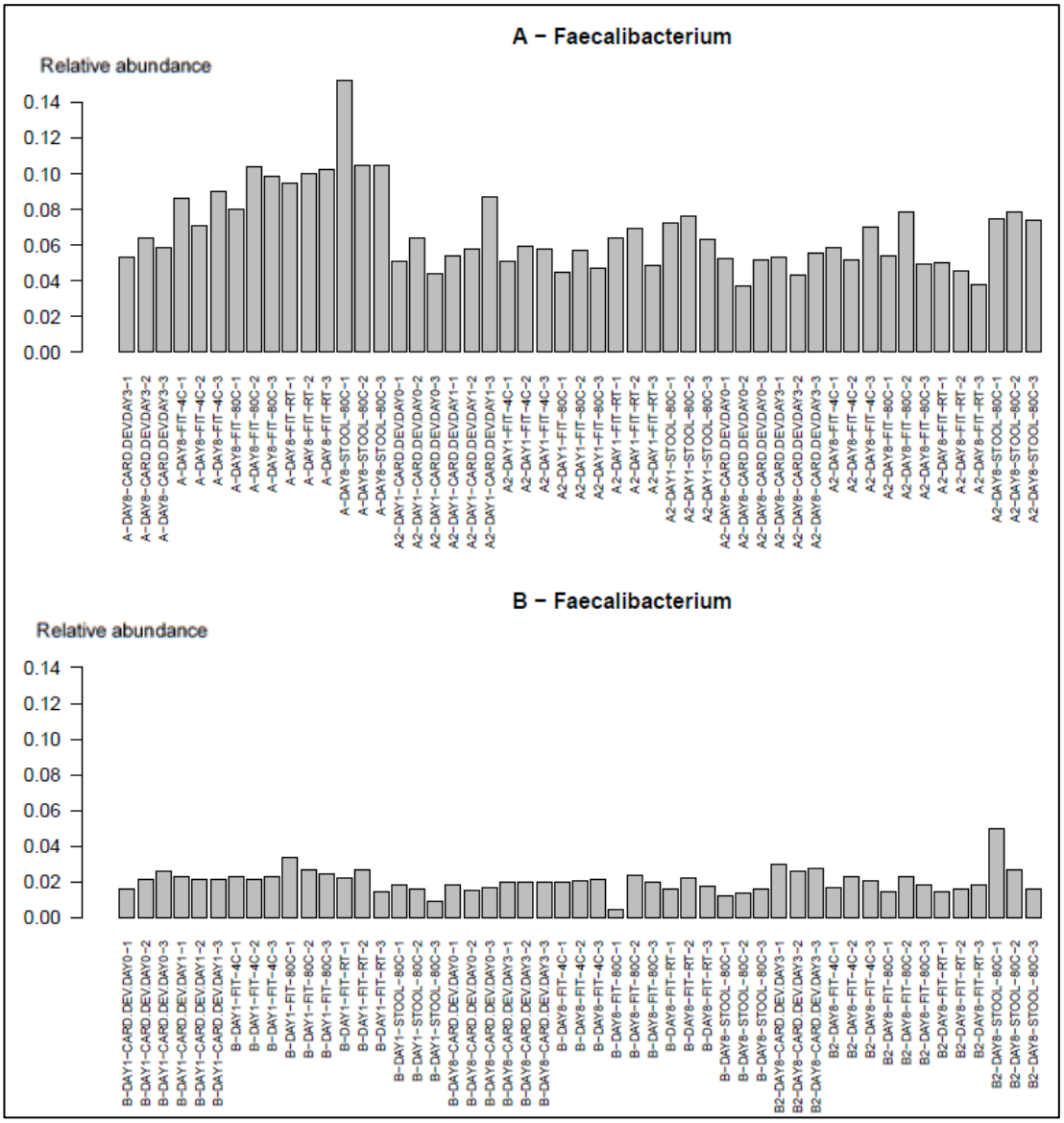


Figure 73. The relative abundance of *Faecalibacterium* of all the samples sequenced as part of the FIT experiment. Each bar represents a sample. Samples derived from participant A (stool samples A and A2) are displayed in the upper chart; samples derived from participant B (stool samples B and B2) are displayed in the lower chart. Each bar is labelled as follows: stool which the sample was derived from; extraction day; sample type (where CARD.DEV.DAY0/1/3 = gFOBT developed on day 0/1/3 respectively); storage temperature; triplicate number.

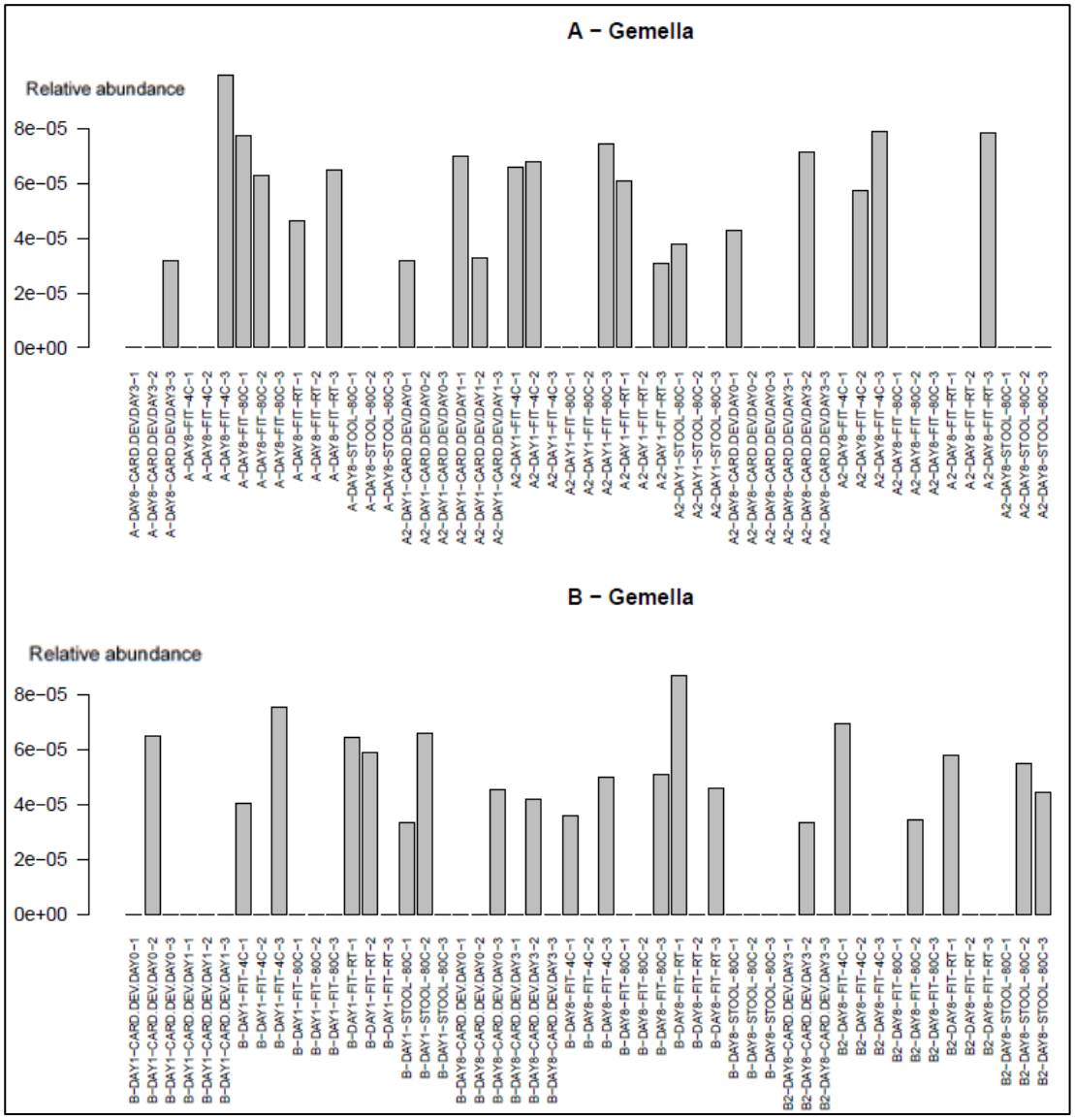


Figure 74. The relative abundance of *Gemella* of all the samples sequenced as part of the FIT-experiment. Each bar represents a sample. Samples derived from participant A (stool samples A and A2) are displayed in the upper chart; samples derived from participant B (stool samples B and B2) are displayed in the lower chart. Each bar is labelled as follows: stool which the sample was derived from; extraction day; sample type (where CARD.DEV.DAY0/1/3 = gFOBT developed on day 0/1/3 respectively); storage temperature; triplicate number. e = x 10^.

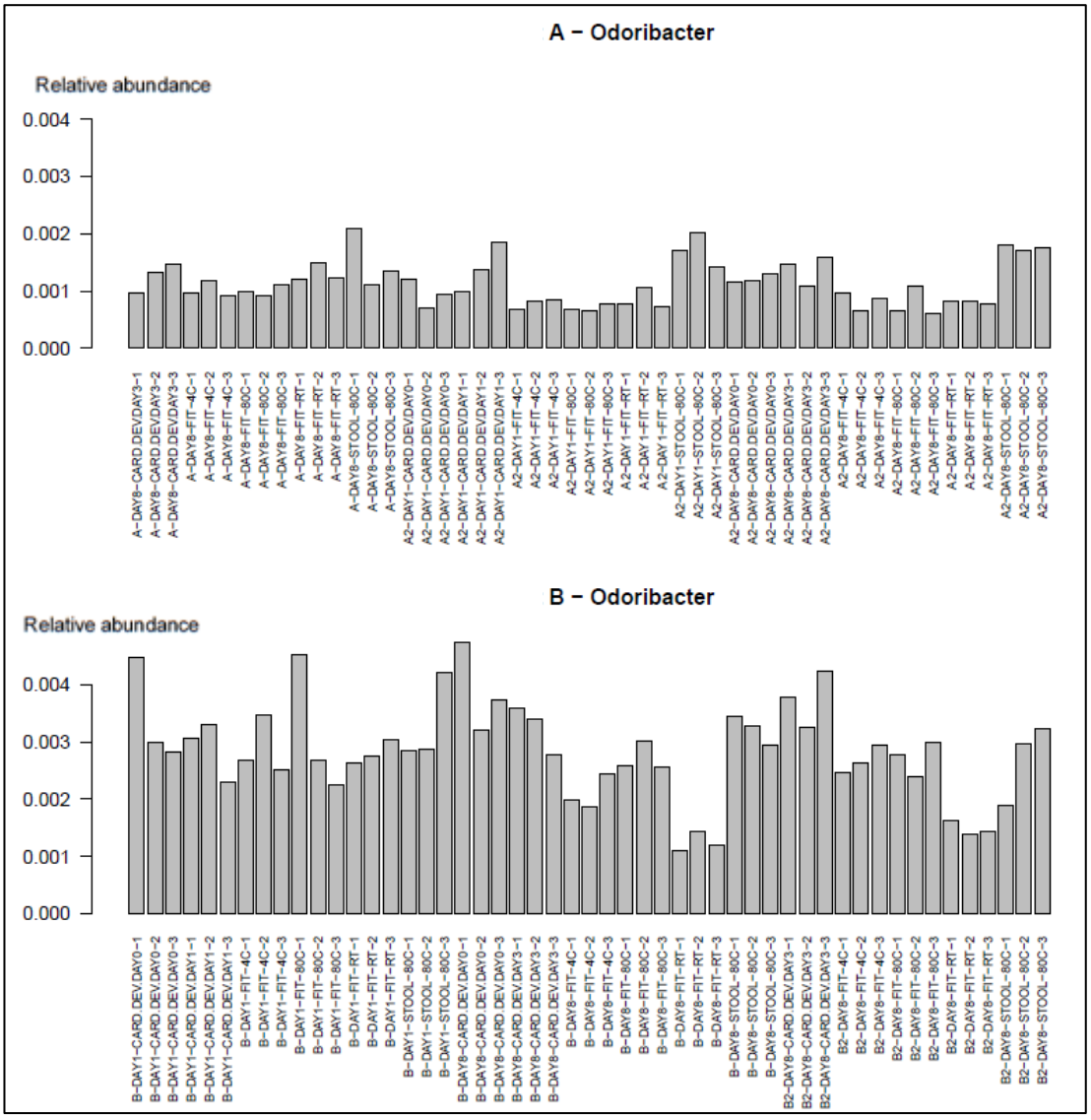


Figure 75. The relative abundance of *Odoribacter* of all the samples sequenced as part of the FIT experiment. Each bar represents a sample. Samples derived from participant A (stool samples A and A2) are displayed in the upper chart; samples derived from participant B (stool samples B and B2) are displayed in the lower chart. Each bar is labelled as follows: stool which the sample was derived from; extraction day; sample type (where CARD.DEV.DAY0/1/3 = gFOBT developed on day 0/1/3 respectively); storage temperature; triplicate number.

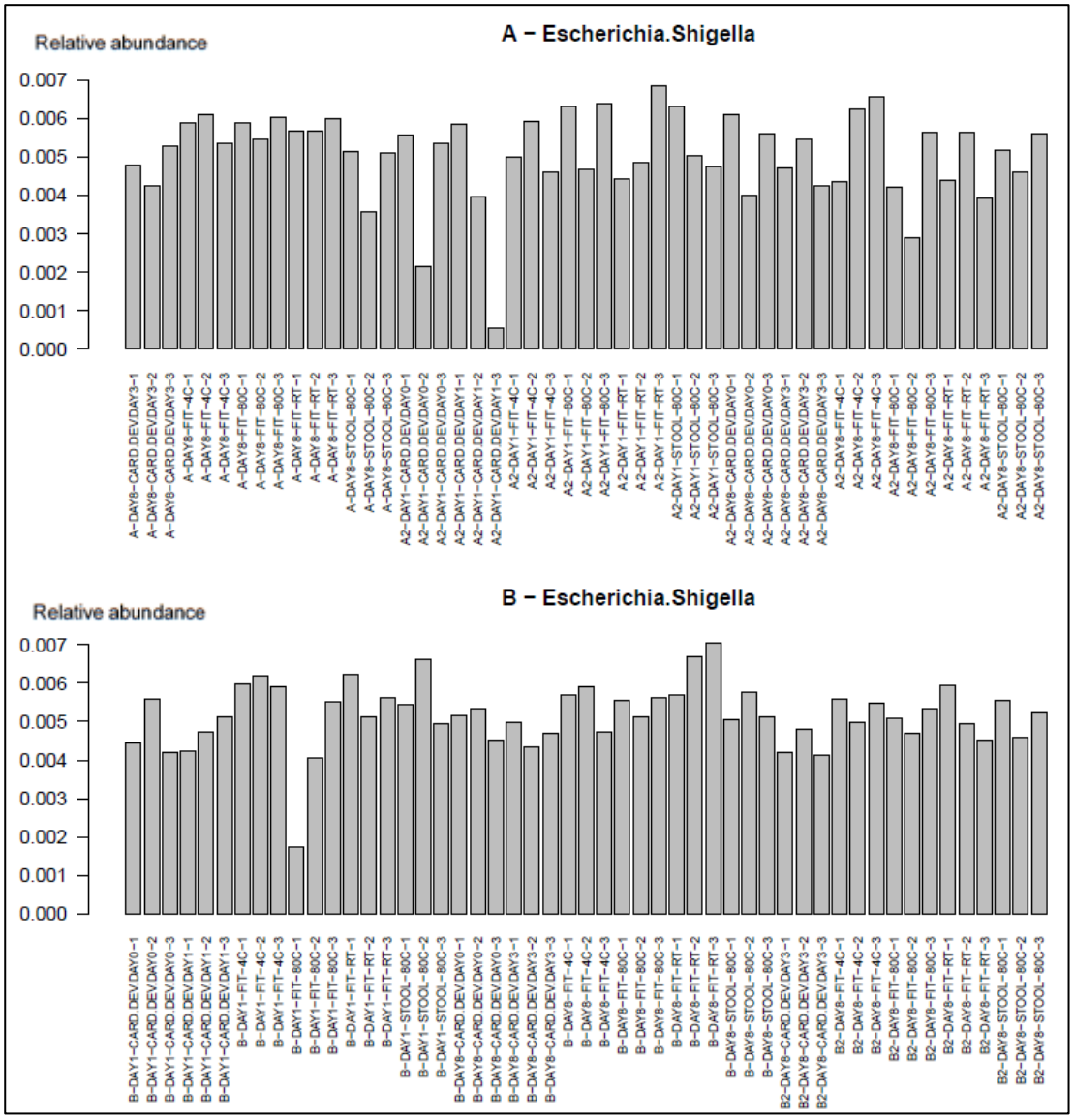


Figure 76. The relative abundance of *Escherichia-Shigella* of all the samples sequenced as part of the FIT experiment. Each bar represents a sample. Samples derived from participant A (stool samples A and A2) are displayed in the upper chart; samples derived from participant B (stool samples B and B2) are displayed in the lower chart. Each bar is labelled as follows: stool which the sample was derived from; extraction day; sample type (where CARD.DEV.DAY0/1/3 = gFOBT developed on day 0/1/3 respectively); storage temperature; triplicate number.

2.4.6.2 Samples extracted on day 1

For samples extracted on day 1, no significant difference in Shannon diversity index was detected between sample types (gFOBT, stool, FIT stored at room temperature, FIT stored at 4°C, FIT stored at -80°C) (Kruskal-Wallis p = 0.8) (Figure 77).

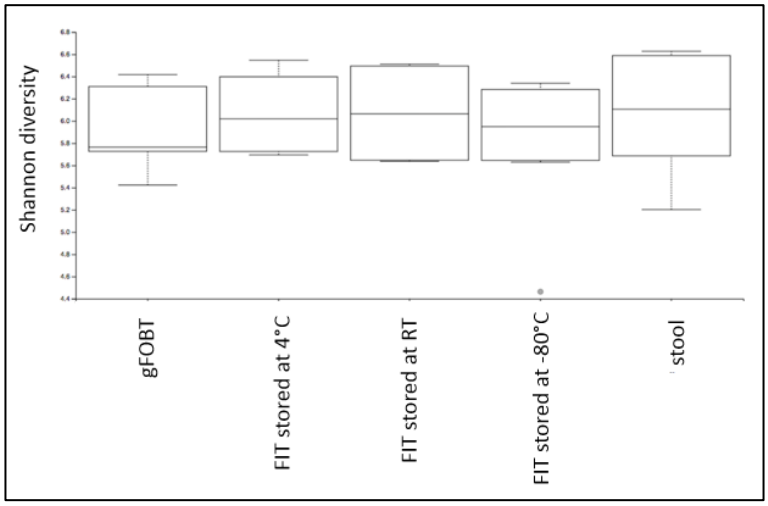


Figure 77. Boxplots of Shannon diversity for 'FIT experiment' samples which were extracted on day 1.

PCA of Bray-Curtis distances (Figure 78) demonstrated that points cluster by participant rather than sample type. Taxonomy bar chart (Figure 79) confirmed that the taxonomic composition of samples derived from the same participant is similar, with no apparent trend due to sample type.

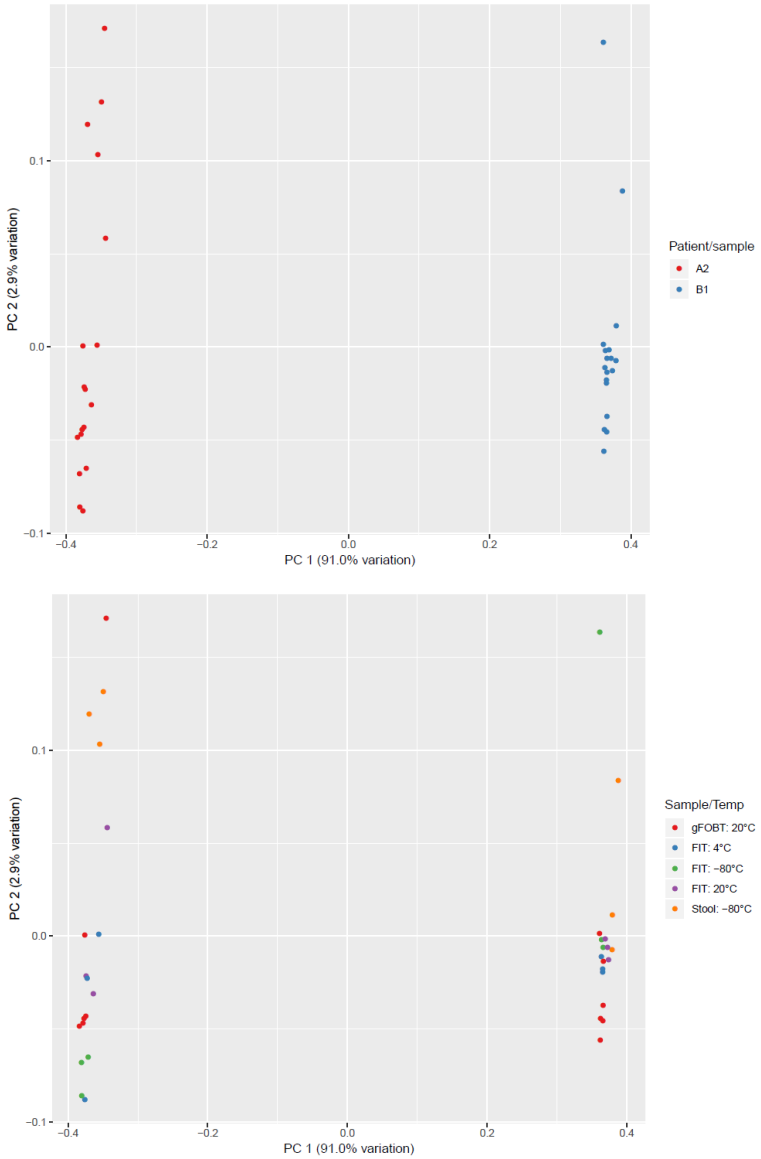


Figure 78. PCA of Bray-Curtis distances for samples extracted on day 1 of the FIT experiment. In the upper graph, points are coloured according to the stool the samples were derived from. In the lower graph, points are coloured according to sample type and storage temperature.

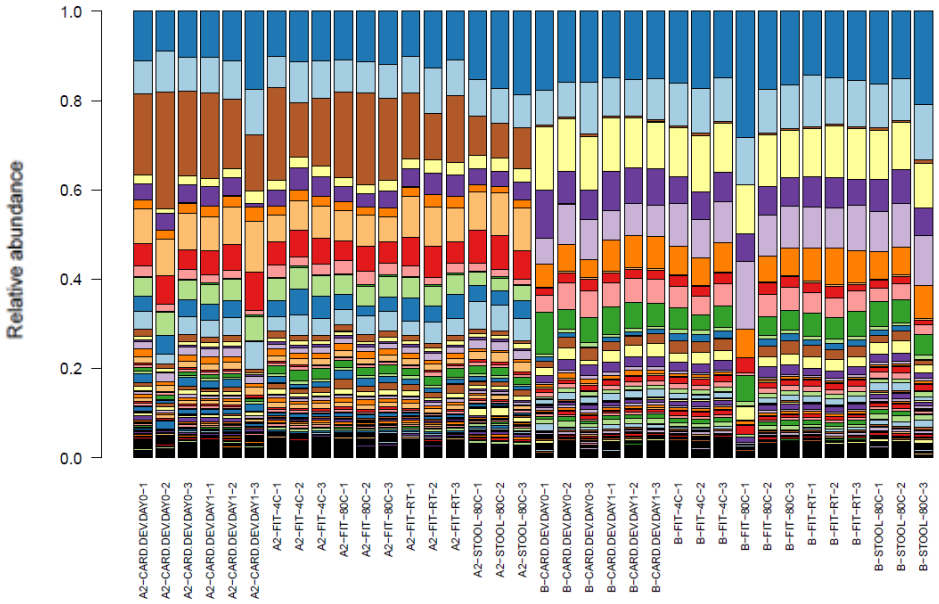


Figure 79. Taxonomy bar chart for samples extracted on Day 1 of the FIT experiment. Each bar represents a sample. Colours denote the relative abundance of taxa (genus level); there are too many to include a legend. Samples derived from stool sample A2 are on the left-hand side of the chart; samples derived from stool sample B are on the right-hand side. Each bar is labelled as follows: stool which the sample was derived from; sample type (where CARD.DEV.DAY0/1 = gFOBT developed on day 0/1 respectively); storage temperature; triplicate number.

2.4.6.3 Samples extracted on day 8

For samples extracted on day 8, no significant difference in Shannon diversity index was detected between sample types (gFOBT, stool, FIT stored at room temperature, FIT stored at 4°C, FIT stored at -80°C) (Kruskal-Wallis p = 0.92) (Figure 80).

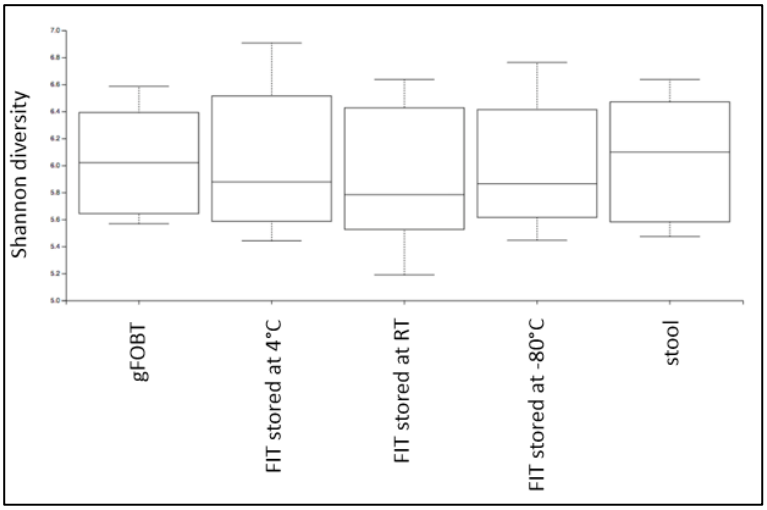
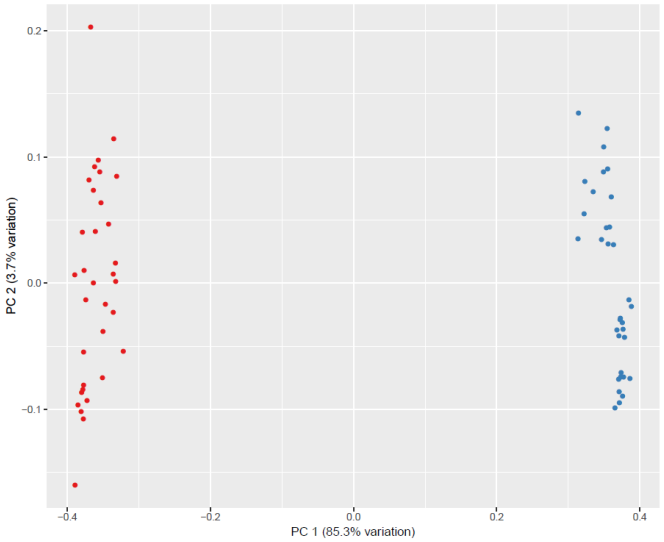
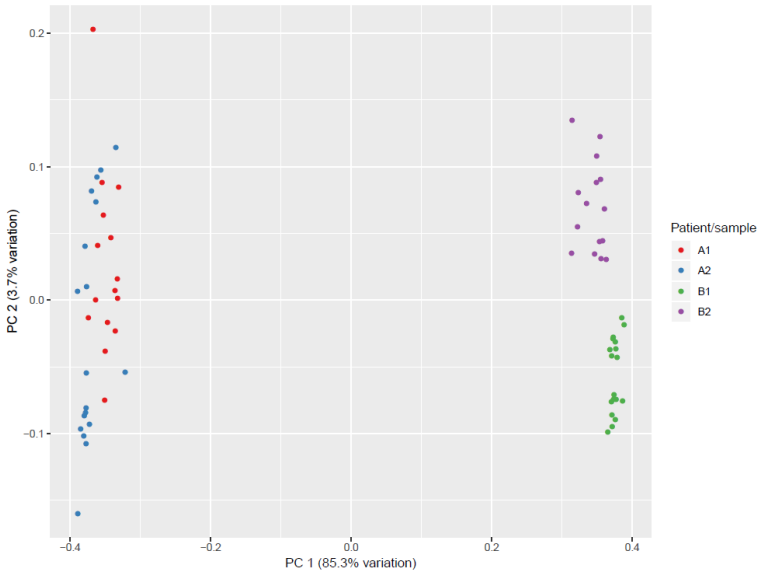


Figure 80. Boxplots of Shannon diversity for 'FIT experiment' samples which were extracted on day 8.

PCA of Bray-Curtis distances (Figure 81) demonstrated that points cluster by participant and for each participant they cluster by stool sample of origin rather than sample type. Taxonomy bar chart (Figure 82) confirmed that the taxonomic composition of samples derived from the same participant is similar, with no apparent trend due to sample type.


PC 2 (3.7% variation)
PC 1 (85.3% variation)
Patient
A
B


PC 2 (3.7% variation)
PC 1 (85.3% variation)
Patient/sample
A1
A2
B1
B2

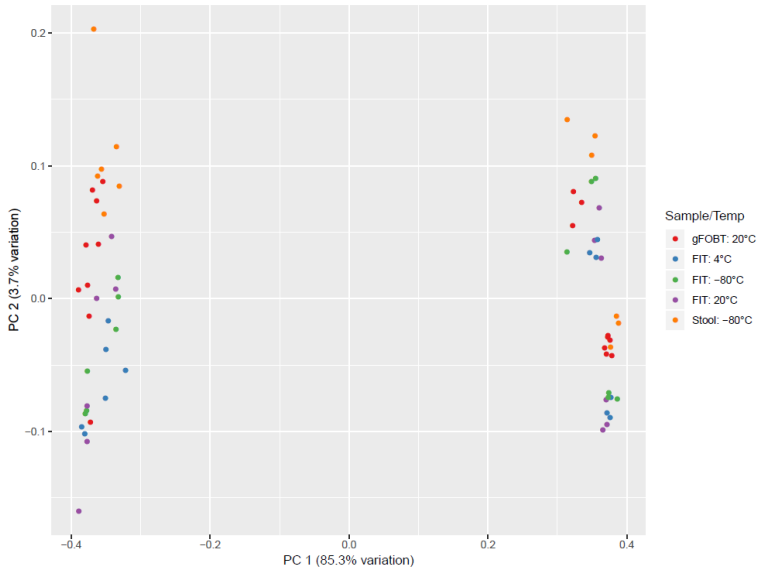

PC 2 (3.7% variation)
PC 1 (85.3% variation)
Sample/Temp
gFOBT: 20°C
FIT: 4°C
FIT: −80°C
FIT: 20°C
Stool: −80°C

Figure 81. PCA of Bray-Curtis distances for samples extracted on day 8 of the FIT experiment. In the upper graph, points are coloured according to the participant the samples were derived from. In the middle graph, points are coloured according to the stool the samples were derived from. In the lower graph, points are coloured according to sample type and storage temperature.

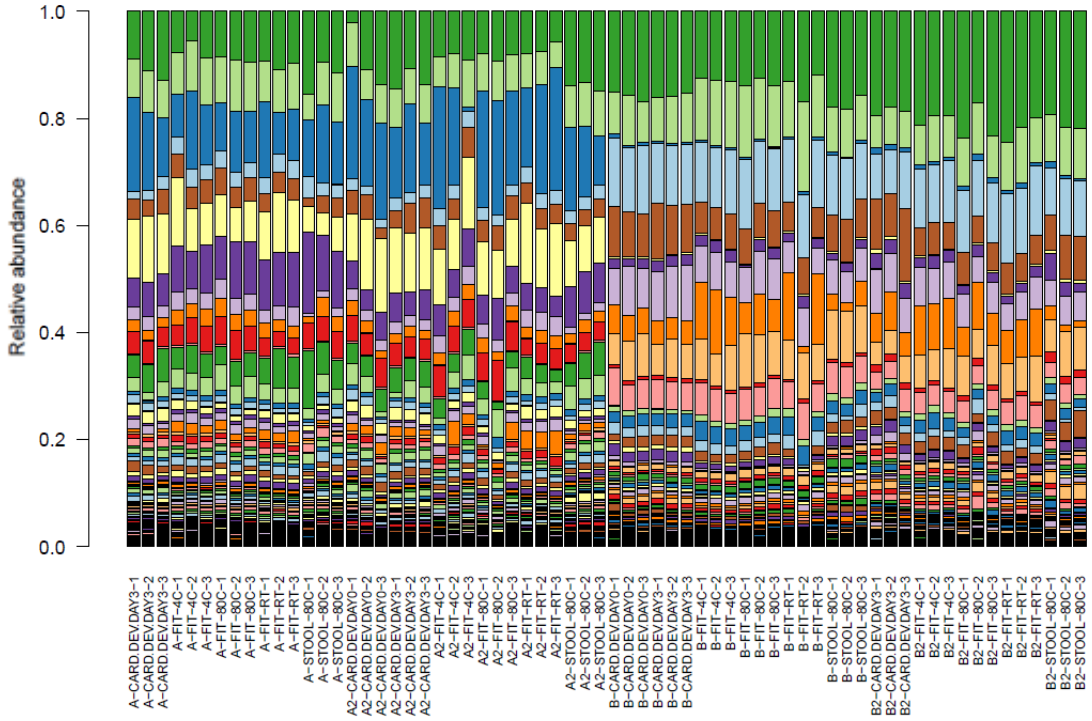


Figure 82. Taxonomy bar chart of samples extracted on day 8 of the FIT experiment. Each bar represents a sample. Colours denote the relative abundance of taxa (genus level); there are too many to include a legend. Samples derived from participant A (stool samples A and A2) are on the left-hand side of the chart; samples derived from participant B (stool samples B and B2) are on the right-hand side. Each bar is labelled as follows: stool which the sample was derived from; sample type (where CARD.DEV.DAY0/3 = gFOBT developed on day 0/3 respectively); storage temperature; triplicate number.

2.5 Discussion

2.5.1 Microbiome analysis of NHSBCSP samples

This is the first study to assess the feasibility of performing microbiome analysis using samples collected routinely by a national CRC screening programme. To date, the majority of microbiome research has been performed using whole stool samples either frozen by study participants at home or expedited to the laboratory for immediate DNA extraction or freezing. This methodology limits sample size, introduces technical variation and potential bias and precludes research in remote locations or locations without access to freezing facilities.

Previous technical microbiome studies demonstrated that the microbiome can be analysed from the faeces of gFOBT samples with high reproducibility, stability and acceptable accuracy (396, 421, 432-434). However, adoption of this method of sample collection by the microbiome research community has been limited; the author is aware of only one study which used gFOBT, however the gFOBT were stored refrigerated/frozen and transported using cold-chain transport, obviating many of the advantages associated with gFOBT as a method of sample collection (85, 476). One study demonstrated that the microbiome can be analysed from the faeces of OC-Sensor FIT with high reproducibility, stability and acceptable accuracy (432).

These technical microbiome studies were performed using samples prepared by laboratory staff using stool from small numbers of healthy volunteers. This is not reflective of the conditions that NHSBCSP samples are exposed to (i.e. collection by participants, storage within their homes and transportation to the Screening Hub at ambient temperature via the post). This study sought to assess whether microbiome analysis could be performed from NHSBCSP samples (processed gFOBT samples and the FIT device that has been adopted by the NHSBCSP, exposed to simulated NHSBCSP conditions).

2.5.1.1 It is possible to perform microbiome analysis from NHSBCSP gFOBT samples

The vast majority of NHSBCSP gFOBT samples underwent successful library preparation and NGS at the first attempt. Of the first batch of NHSBCSP gFOBT samples, two failed library preparation (0.1%), five failed NGS with fewer than

10,000 reads (0.3%) and 1435 were successfully sequenced (99.5%). The samples which failed library preparation or sequencing were successfully sequenced on the second NGS run. This indicates that DNA extracted from processed NHSBCSP gFOBT samples is of sufficient quality and concentration for subsequent library preparation and sequencing, despite the marked variation which was observed in the amount of stool per sample.

A recent meta-analysis of studies performed using 16SrRNA analysis of faecal samples and a meta-analysis of studies which used the EMP methodology both set a sequencing depth of fewer than 5000 reads as a study exclusion criterion; this indicates that the number of reads generated by the NHSBCSP samples is comparable with existing studies and sufficiently high for robust analysis (404, 477).

2.5.1.2 Minimal bacterial contamination is introduced during laboratory processing

Laboratory processing was not performed under sterile conditions. Although this is also the case for the majority of microbiome research, it was important to assess the amount of contamination introduced during laboratory processing. Assessment of contamination has not been routinely performed in microbiome studies; a review of 265 microbiome publications discovered that only 30% reported use of a negative control and 10% use of a positive control (478). The issue of contamination potentially impacting microbiome results has only recently started to gain recognition (479); it is thought to be a particular problem for low biomass samples (such as tissue). Although lists of contaminant taxa have been published (480-482) there is no consensus as to how to account for these findings during analysis (as removal of contaminant taxa or taxa below a minimum relative abundance may inadvertently remove taxa which are genuinely present within the sample, particularly as many contaminant taxa are present within faeces).

In the current study, some of the extraction controls had similar amounts of extracted DNA to some NHSBCSP gFOBT samples as measured by the NanoDrop-1000 spectrophotometer (Thermo Fisher Scientific Incorporated, UK). However, none of the extraction controls generated PCR amplicons of high enough concentration for sequencing. This indicates either that the concentrations recorded by the NanoDrop-1000 spectrophotometer (Thermo Fisher Scientific Incorporated, UK) were inaccurate or the DNA was non-

bacterial. Similarly none of the PCR water controls generated PCR amplicons of sufficient concentration for sequencing.

These results indicate that minimal contamination is introduced during laboratory processing. However, one limitation is that possible contamination which may have occurred prior to laboratory processing could not be assessed. To do this, replicate blank gFOBT cards would have to have been issued to participants to be stored alongside the gFOBT samples during collection in the participants' homes, storage at the Screening Hub and transit to the processing laboratory. This was not possible as it would disrupt routine screening and was not covered by the study's ethics.

In future work, assessment of contamination will be even more rigorously assessed, in light of recent guidance and given that FIT samples are of lower biomass than gFOBT (483). The guidance recommends random allocation of samples to positions on plate layouts to reduce the impact of bias and contamination on results; this was already performed during the current study. The guidance also recommends including extraction controls in every extraction batch (which would become more cost-effective if automated extraction of larger batches is adopted); use of a diverse 'mock community' microbial positive extraction control; sequencing of controls with bioinformatic modelling to identify potential contamination; and validating findings using alternative methodologies (qPCR will be performed on the NHSBCSP samples). Given that laboratory contamination has been shown to exhibit temporal and technician-specific variation, in future work, contamination will be continually assessed at each DNA extraction and PCR/library preparation batch (484). Well-to-well contamination has been reported with automated DNA extraction (485) – this will therefore be formally assessed with the QIAcube HT instrument.

2.5.1.3 The choice of which three squares of a gFOBT sample to process has minimal effect on microbiome results

From each gFOBT three squares of faecally-loaded card, one from beneath each of the three flaps, were dissected and processed as a single combined sample. The rationale was that the three remaining squares could be used for alternative analysis or technical replicates. This methodology differs from the existing microbiome technical studies, the majority of which applied a single homogenised stool to gFOBT and did not specify the number of squares which were extracted.

It was therefore important to assess whether the microbiome result would differ significantly between pairs of composite samples extracted on the same day. The existing literature suggested that variability between subsample replicates may be apparent, particularly for lower abundance taxa (395, 415-417).

In general replicates were separated by low Bray-Curtis distances. LEfSe did not identify taxa which differed significantly between a group comprising one member from each replicate pair and another group containing the second member from each replicate pair. Within pairs, replicates demonstrated good agreement for the relative abundance of CRC-associated taxa and *Escherichia-Shigella*. These findings will be confirmed by running the replicate pairs through the Random Forest model (described in Chapter 3) once it is validated, to determine whether there is any effect on sample classification.

2.5.1.4 The microbiome of processed NHSBCSP gFOBT samples is stable when samples are stored at room temperature for a prolonged time

Processed NHSBCSP gFOBT samples are stored for variable lengths of time at ambient temperature, both at the Screening Hub (prior to link-anonymisation and transport to the processing laboratory) and within the processing laboratory prior to DNA extraction. Due to the constraints of staff availability within a busy screening programme and the logistics of transporting such a large number of samples, it was not possible to standardise the time between sample collection and DNA extraction, reflected in a wide range (55-570 days, median 202 days).

Storage at -20°C or -80°C would require considerable freezer space and potentially introduce technical bias; it was therefore not considered. Our laboratory had previously demonstrated stability of the microbiome extracted from processed gFOBT stored for up to three years at room temperature, although this study was performed using samples prepared by laboratory staff using stool from small numbers of healthy volunteers (434). The majority of existing microbiome technical studies assessed the stability of the microbiome on gFOBT samples after storage for 4-7 days at ambient temperature (396, 421, 433). It was therefore important to assess the effect of prolonged storage at room temperature on the microbiome extracted from processed NHSBCSP gFOBT cards.

Three squares of faecally-loaded card, one from beneath each of the three flaps, were dissected and processed as a single combined sample; the alternate three squares were extracted after prolonged storage reflective of either the median or maximum length of storage experienced by the NHSBCSP samples. A limitation of this approach is that the extraction replicates were not true replicates (made from a single homogenised stool); instead they represent subsample replicates extracted after prolonged storage. This means that any variability in the microbiome between replicates could be due to a combination of biological variation (subsample) and technical variation (storage duration). It was not possible to ask participants to create replicate gFOBT samples as this would interfere with routine screening and was not covered by the study's ethics; furthermore, to comprehensively overcome the risk of subsampling bias, this would require participants to homogenise stool prior to loading the gFOBT cards. It was not possible to divide the squares of faecally-loaded card to create replicates, as for the majority of NHSBCSP samples there would have been inadequate biomass; furthermore, heterogeneity within a square could not be excluded.

It was not possible to compare results with gFOBT cards which were extracted immediately as the minimum time between sample collection and DNA extraction for the NHSBCSP samples was 55 days; a change in microbiome stability which occurred prior to this point could therefore not be excluded. This was however appropriate for the assessment of temporal stability within the design of the current study.

The microbiome results of samples extracted after a prolonged period of storage were no more different from their replicate partners than pairs of samples extracted on the same day; Bray-Curtis distances between replicates were low and agreement was high for the relative abundance of CRC-associated taxa and *Escherichia-Shigella*. LEfSe identified one taxon (*prevotellaceae*NK3B31 group) which was significantly enriched in the group comprising samples extracted after a prolonged period of storage. However, review of the relative abundances of *prevotellaceae*NK3B31 group for individual samples indicated that there was no consistent increase in *prevotellaceae*NK3B31 group in the samples extracted after a prolonged period of storage and that the LEfSe result was likely influenced by a single replicate pair. A literature search revealed only a few microbiome studies, none of which were CRC-specific, which mentioned

*prevotellaceae*NK3B31 group; none described an association with microbiome temporal stability or bacterial overgrowth.

These findings, together with the fact that time until DNA extraction was not shown to affect variation in Bray-Curtis distances (presented in section 3.5.3), suggest that the microbiome is stable when extracted from NHSBCSP gFOBT samples stored for prolonged periods at ambient temperature, compared with a baseline of replicate samples stored for a minimum of 55 days prior to DNA extraction. This will be confirmed by running the replicate samples through the Random Forest model (described in Chapter 3) once it is validated, to determine whether there is any effect on sample classification. Future work will also investigate the underlying mechanism, by attempting to culture gFOBT samples in order to determine whether bacteria on gFOBT are viable or not.

2.5.1.5 The microbiome of NHSBCSP gFOBT samples demonstrates relative temporal stability, although marked intra-participant variability is noted for the taxon *Escherichia-Shigella*

The majority of research of the CRC-associated microbiome has been conducted using samples collected at a single timepoint. The health-associated microbiome has been shown to exhibit temporal variation in response to changes in environmental conditions (20, 417-419); the extent to which the CRC-associated microbiome demonstrates temporal variation had not been investigated. It was important to determine whether temporal variation would affect the predictive accuracy of a microbiome-based screening model.

Four different types of replicates were created to assess subsample and temporal variation of the microbiome of NHSBCSP participants. Overall, the microbiome results of samples derived from a single gFOBT card were more similar to one another than the microbiome results of different screening participants, which confirms the high degree of inter-individual variation of the faecal microbiome described in the literature. However some subsamples did show a greater degree of difference between their microbiome community structure and that of their respective replicates; this was not associated with type of temporal replicate or type of subsample, suggesting differences arose randomly perhaps due to temporal variation, subsampling, contamination or chance. The fact that the microbiome of individual subsample squares was similar to the combined sample derived from three squares (the methodology used in Chapter 3) suggests that

FIT (which collects a low biomass sample from a single stool) may give similar microbiome results to gFOBT samples (which collect higher biomass subsamples from three separate stools); however this will need to be confirmed in future work by comparing the microbiome results of FIT and gFOBT NHSBCSP samples.

There was minimal variation in the relative abundances of CRC-associated bacteria between replicates, suggesting subsample and at least short-term temporal stability of these taxa. To assess the effect on sample classification, these replicates will be run through the Random Forest model (described in Chapter 3) once it is validated.

These NHSBCSP gFOBT samples represent a relatively unique resource to assess short-term stability of the CRC-associated microbiome. It would be interesting to investigate the longer-term stability of the CRC-associated microbiome. Future work could compare the microbiome of NHSBCSP samples collected over consecutive screening rounds, derived from the same individuals. However, this may only allow assessment of blood-negative or colonoscopy-normal samples, as participants with adenoma or CRC would receive treatment; such treatment has been shown to alter the microbiome in the short-term, but long-term effects have not been assessed (486, 487). Prospective longitudinal cohort studies or biobanks with repeated sampling would be an alternative method to investigate the long-term stability of the CRC-associated microbiome, but they are resource-intensive. Assessing the long-term stability in CRC patients that have already received a diagnosis is not possible, due to the limited time between diagnosis and treatment and the effect of bowel-preparation on the microbiome.

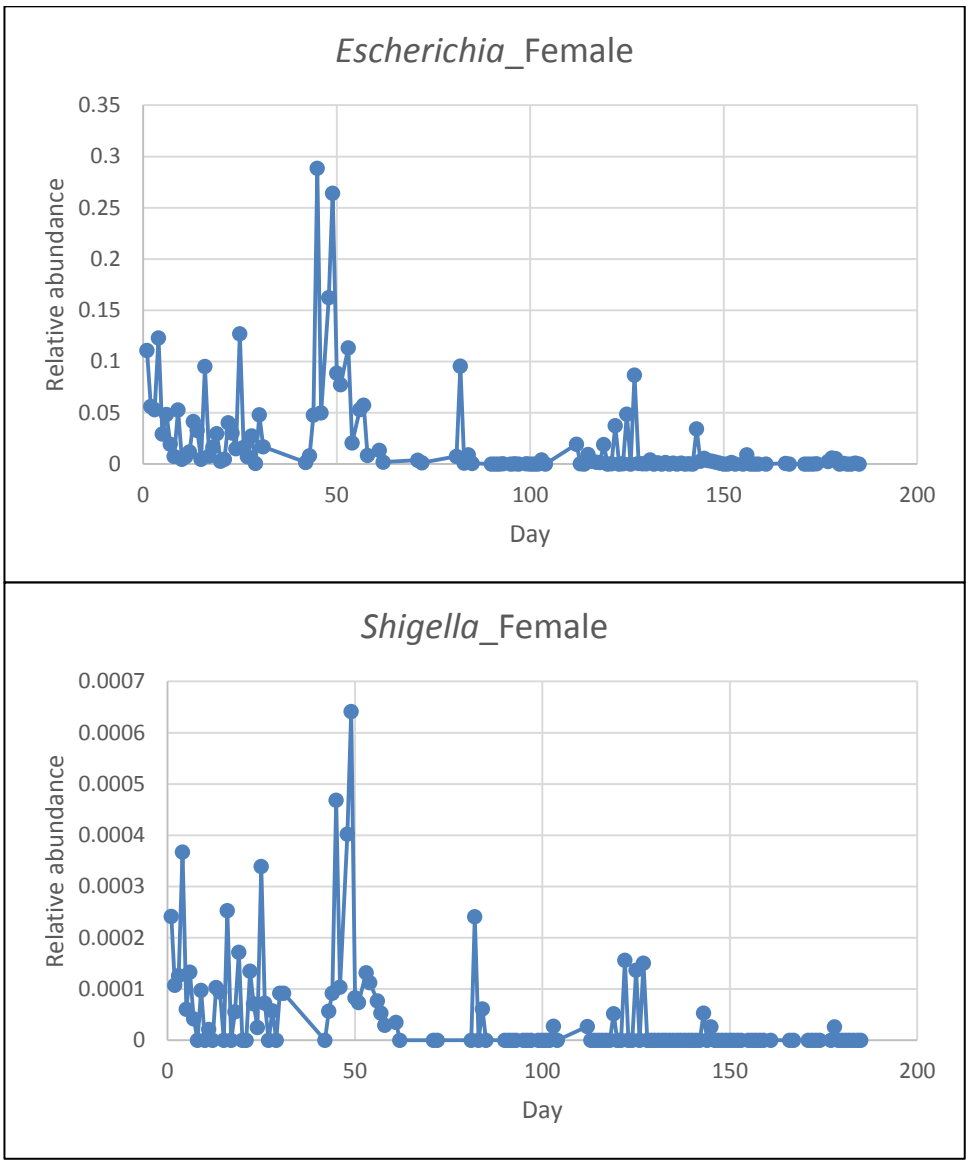
The extremely high intra-participant variability in the relative abundance of *Escherichia-Shigella* was an unexpected finding, and one which could not have been identified using single-timepoint sampling, as employed by the majority of microbiome studies. The relative abundance of *Escherichia-Shigella* varied between subsamples derived from stool collected on the same day, which suggests either heterogeneity of *Escherichia-Shigella* within a faecal sample or a change in relative abundance due to a technical reason (e.g. overgrowth, contamination or sequencing error). None of the aforementioned studies which investigated faecal subsampling specifically mentioned variability of the relative abundance of *Escherichia-Shigella* (395, 415-417). However, a study which performed FISH on faecal cores reported focal areas of concentrated taxa, which

could potentially account for the current study's findings (488). Alternative explanations include contamination; *Escherichia* has been cited in studies as a laboratory contaminant (480); or errors in detection: compared with culture based methods, 16SrRNA NGS has been shown to have a high positive predictive value but low negative predictive value for the detection of *Escherichia-Shigella* (489). To exclude possible contamination post-DNA extraction or detection error, the relative abundance of *Escherichia-Shigella* in these samples will be confirmed by qPCR.

The relative abundance of *Escherichia-Shigella* also varied between subsamples derived from stools collected on separate days. It is impossible to determine whether this reflects true temporal differences in the relative abundance of *Escherichia-Shigella* or differences due to subsampling. Temporal variability in the relative abundance of *E. coli* has been described by a study which performed metagenomic analysis of stool samples collected over a period of 3 years from a patient who had Crohn's disease (490). The relative abundance ranged from 0.1% to 42.6% and it was noted that high abundance of *E. coli* did not necessarily correlate with inflammatory levels. Importantly the paper did not describe how the stool samples were collected and processed (whether homogenised or sub-sampled), which could potentially contribute to or account for the study's findings. The authors compared their results with the mean relative abundance of *E. coli* (0.008) detected in faecal samples of healthy volunteers by the Human Microbiome Project. However, on inspection of the supplementary information for the Human Microbiome Project, the maximum relative abundance recorded for *E. coli* was high (0.96), and was the highest relative abundance recorded of all the faecal taxa (491). The methods paper which describes the Human Microbiome Project protocol indicates that stool subsamples (~2 ml) were processed, with no mention of prior homogenisation (492).

A paper which assessed temporal variability of the faecal microbiome in two individuals over hundreds of daily time points provided the raw data in the supplementary materials, from which it was possible to calculate the relative abundance of taxa (493). The microbiome was analysed from stool which was swabbed from used toilet paper or faecal stabs, indicating subsampling (418). The range of relative abundance of *Escherichia* recorded for each individual was wide (0-0.29 and 0-0.44) whereas that of *Shigella* was narrow (0-0.0006 and 0-0.0009) (493). This is illustrated in Figure 83. The relative abundance of *Escherichia* and *Shigella* mirror one another, likely due to a limitation of the

discriminative ability of the OTUs. Importantly, variability between consecutive days is high: the maximum relative abundance of *Escherichia* of 0.29 in the female participant is preceded and succeeded by relative abundances of 0.05; the maximum relative abundance of *Escherichia* of 0.44 in the male participant is preceded and succeeded by relative abundances of 0.02 and 0.06 respectively.



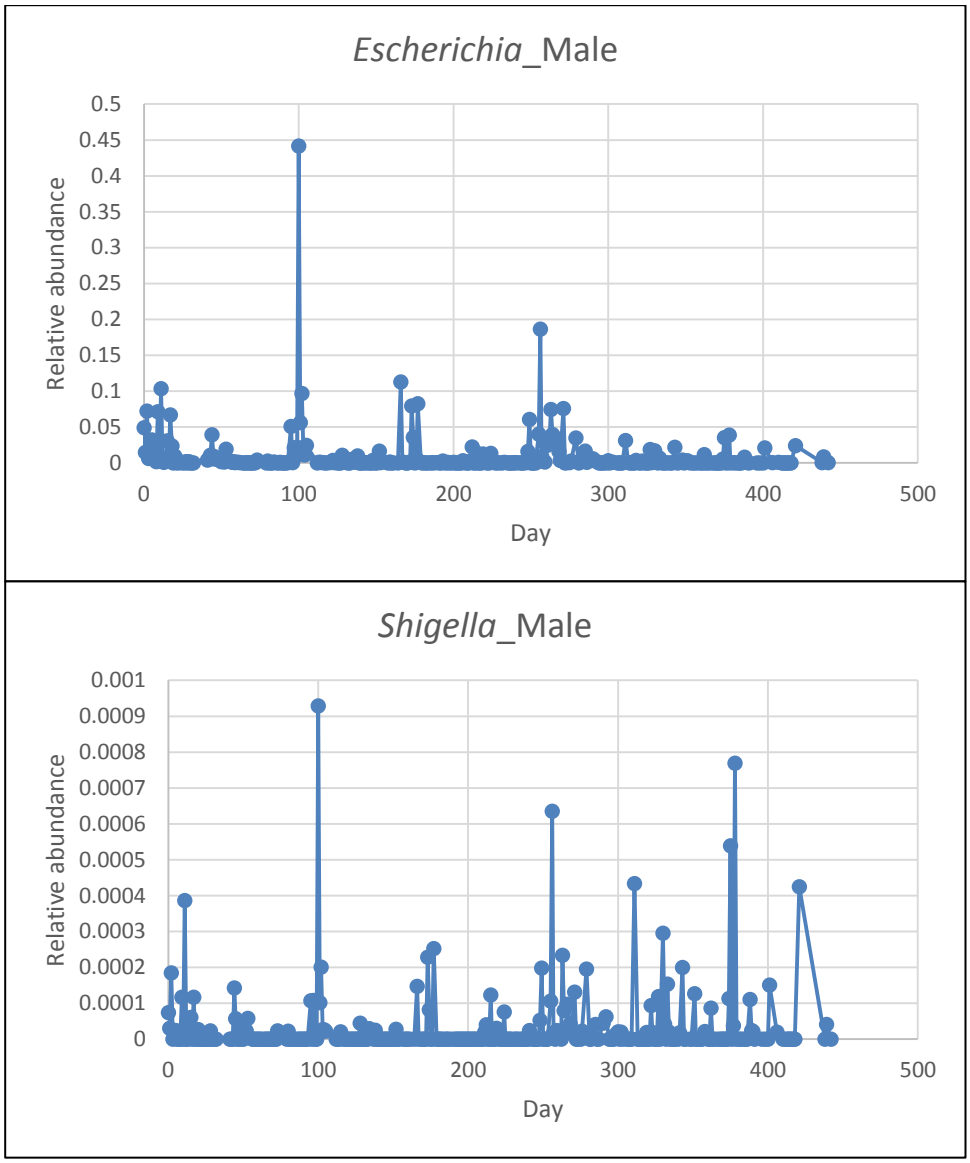


Figure 83. The relative abundance of *Escherichia* and *Shigella* calculated from the raw data provided in the paper 'Moving pictures of the human microbiome' (493). The first two plots show the relative abundance for *Escherichia* and *Shigella* for participant F4. The last two plots show the relative abundance for *Escherichia* and *Shigella* for participant M3.

Whilst the data from these papers indicate marked inter-subject and intra-subject variability of the relative abundance of *Escherichia-Shigella*, it is unclear to what extent this is due to true temporal variability or technical factors such as subsampling. Temporal variability of *E. coli* clones (as opposed to overall relative abundance) has been reported by a study which did perform faecal homogenisation (489) and another which performed culture from anal swabs (494).

As FIT collects a lower biomass sample, scraped from the surface of a stool, it may exhibit subsampling variation in taxa; this will need to be formally assessed.

In the current study and the Human Microbiome Project, the maximum relative abundance of *Escherichia-Shigella* was extremely high (0.88 in the current study). In the NHSBCSP study described in Chapter 3, an unexpectedly high relative abundance (0.40) of *Fusobacterium* was observed (see section 3.5.6). Subsample variation in the relative abundance of *Escherichia-Shigella* described in the current study raises the question as to whether subsample variation might explain the unexpected *Fusobacterium* finding. It also raises the question as to whether samples containing an unexpectedly high relative abundance of a taxon could/should be excluded from analysis.

The high intra-sample variability in the relative abundance of *Escherichia-Shigella* also has implications for the design of screening models. *Escherichia-Shigella* was the taxon which was most enriched in samples from CRC patients compared with healthy volunteers in the study described in Chapter 4. Other studies have also reported enrichment of *Escherichia-Shigella* in CRC cases compared with controls (121, 495) and in a meta-analysis of faecal studies, *E. coli* was the 14th most important variable contributing to a Random Forest model to detect CRC (496). However, in light of the current study's findings caution should be exercised; *Escherichia-Shigella* may not be an appropriate component of a microbiome-based screening tool.

2.5.1.6 The microbiome can be analysed from mock NHSBCSP FIT samples

The NHSBCSP is transitioning from gFOBT to FIT. A few technical studies have assessed the potential to perform microbiome analysis from FIT samples with mixed results, although some of these studies suffered methodological limitations and none of them replicated the conditions that NHSBCSP FIT samples will be exposed to (432, 441, 442). The current study assessed the potential to perform microbiome analysis from the FIT device which has been adopted by the NHSBCSP (OC-Sensor FIT (Mast Group Ltd)) under conditions simulated to reflect those that NHSBCSP samples will be subjected to. FIT-processing was simulated by piercing the FIT device and squeezing liquid into the upper chamber, as the laboratory did not have access to a FIT-processing machine. It has been shown in one study that cross-sample contamination does not occur during

automated FIT-processing (443); ideally this will be confirmed using the FIT-processing machine at the Southern Hub, although appropriate approvals will first need to be sought.

As the NHSBCSP had not started to collect FIT samples at the time of the current study and in order to compare results with reference whole stool and gFOBT samples, samples were provided by healthy volunteers rather than NHSBCSP participants. It will be important to expand the study to NHSBCSP FIT samples collected as part of routine screening from both participants with a blood-negative screening result and colonoscopy-confirmed diagnosis, in order to determine the effect of routine screening conditions on the microbiome, and in particular CRC-associated taxa. As it will not be possible to confirm accuracy by comparison with whole stool samples (as this would disrupt routine screening), this will be assessed using paired FIT and frozen whole stool samples collected as part of a symptomatic FIT trial (this study is currently being planned).

Unlike one study which reported difficulty extracting sufficient DNA from FIT (441), the current study successfully extracted DNA from all FIT samples tested. It should be noted that the FIT samples were prepared by the author; it will be important to assess the yield of DNA from NHSBCSP FIT samples prepared by NHSBCSP participants. All of the FIT samples successfully underwent library preparation and sequencing on the first attempt, indicating that the extracted DNA from FIT is of sufficient quality and concentration for microbiome analysis.

The type of sample (whether FIT, gFOBT or whole stool) contributed a small but significant amount to variability in microbiome community structure (measured by Bray-Curtis distances). The Bray-Curtis distances between a reference stool sample and replicate stool samples appeared slightly reduced compared with the Bray-Curtis distances between the reference stool sample and FIT or gFOBT samples. However, there was no appreciable difference in the relative abundance of CRC-associated taxa between any of the types of sample, storage conditions or storage duration beyond that occurring due to subsampling. These findings will be confirmed by running the samples through the Random Forest model (described in Chapter 3) once it is validated, to determine whether there is any effect on sample classification. There was minimal subsample variation in the relative abundance of *Escherichia-Shigella,* this may be due to the fact that stool was manually homogenised prior to the creation of replicate samples or due to

the fact that replicate samples were made from only four stools originating from only two healthy volunteers.

Sequencing results from the ‘FIT experiment’ samples processed after eight weeks of storage are pending.

2.5.1.7 Microbiome analysis from NHSBCSP samples can be performed at scale

The feasibility of the laboratory processing used in this study ultimately being adopted by the NHSBCSP was investigated. Methods to reduce manual processing and improve throughput were explored. These serve a dual purpose: to reduce staff-time and to reduce the likelihood of error. The possibility of an error occurring when manually processing such large numbers of samples was illustrated by the identification and investigation of a sample labelling error. This emphasises the need for cross-checking and automation. A Random Forest model-based method to identify potentially mislabelled microbiome samples has been proposed (497) and used by at least one study (498), however for metadata categories which do not associate with distinct microbiomes (such as the clinical categories of the NHSBCSP samples), this method is unlikely to be successful and risks falsely excluding samples due to natural biological variation. Random Forest modelling (as performed in Chapter 3) has been shown to be robust to minor degrees of sample mislabelling, although the consequence is a decrease in the reported accuracy of the model (497).

Automated DNA extraction using the QIAcube HT instrument (Qiagen, Germany) was shown to increase DNA extraction throughput and reduce DNA extraction time. The sequencing results from a comparison between manual and automated DNA extraction are pending. The EMP16S Illumina Amplicon methodology was adopted; it is a quick and straightforward protocol which can be performed using multichannel pipettes. A 96-well gel electrophoresis mould was introduced to save time; however, given that all PCR amplicons processed to date were of the correct size, this step is felt to be unnecessary and will be omitted in future work. It was confirmed that it is possible to sequence ~1500 samples/run on an Illumina HiSeq provided adequate concentration of PhiX.

Variability between sequencing runs was investigated by repeated sequencing of amplicon pools. A study had shown that microbiome results can be affected by sequencing run (380), and this was found to be the case for the initial Illumina MiSeq work that was the fore-runner to this study. It was important to determine whether this would also be the case for samples processed using the EMP methodology and sequenced on the Illumina HiSeq, as this would inform the design of future studies: whether the set of samples used to validate the Random Forest model (described in Chapter 3) could be processed on a separate sequencing run; and whether samples collected from abroad which were received in two batches (described in Chapter 4) could be sequenced as two batches or should be sequenced as a single batch.

Sequencing run was shown to affect the number of reads/sample, which has been shown, in another study, to affect the prevalence of sequence detection (404). However, a limitation was that the number of samples and the concentration of PhiX differed between the two sequencing runs; in future, these will be held constant (the sequencing team were titrating the concentration of PhiX across sequencing runs in an attempt to optimise sequencing depth). Sequencing run did not affect sample alpha diversity (measured by the Shannon diversity index) or beta diversity (as measured by the Bray-Curtis distance) and agreement between sequencing runs for the relative abundance of CRC-associated taxa was high (with minimal bias detected for *Gemella* and *Escherichia-Shigella* only). This suggests that future work (which will involve the processing of large numbers of samples) could combine and compare microbiome results from samples sequenced on different sequencing runs. In order to confirm this, once the Random Forest model (described in Chapter 3) has been validated, results from the duplicate pools which were sequenced on two separate sequencing runs will be run through the model to determine whether sequencing run influences sample classification.

2.5.2 Chapter Summary

- The microbiome can be successfully analysed from processed NHSBCSP gFOBT samples.

- The microbiome is stable if NHSBCSP gFOBT samples are stored for prolonged periods at room temperature.

- The relative abundances of CRC-associated taxa demonstrate minimal temporal variation. The relative abundance of *Escherichia-Shigella* demonstrates marked variation potentially secondary to technical factors such as subsampling or temporal variation; this suggests that *Escherichia-Shigella* is not a useful CRC screening biomarker.

- The microbiome can be successfully analysed from the FIT devices which the NHSBCSP will use, after simulation of the conditions that NHSBCSP FIT samples will be exposed to.

- Microbiome analysis of NHSBCSP samples can be performed at scale.

Chapter 3
Investigating the potential of the microbiome to improve the accuracy of CRC screening

3.1 Introduction

Chapter 2 confirmed the feasibility of performing microbiome analysis directly from the faeces of processed NHSBCSP samples. In this chapter, the utility of this approach will be evaluated by:

- determining whether microbiome analysis can be used to improve the accuracy of screening

- characterising the microbiome of a large number of patients with colonoscopy-confirmed diagnoses

The principles of CRC screening, details of the NHSBCSP and the need to improve its accuracy will be presented. Literature suggesting that microbiome analysis has the potential to improve screening accuracy will be reviewed, with limitations of the existing studies outlined. These limitations will be surmounted if microbiome analysis of NHSBCSP samples is found to improve screening. The potential advantages and limitations of this approach will be reviewed.

3.2 CRC screening

3.2.1 The principle of CRC screening

Screening reduces mortality by detecting asymptomatic colorectal adenomas or early stage CRC (499). The mortality rate of CRC increases with increasing stage; five year survival rate for Stage I exceeds 90% whereas it is less than 10% for Stage IV disease (76).

3.2.2 The NHS Bowel Cancer Screening Programme

The NHSBCSP was founded in 2006; it reduces the relative risk of death from CRC by 15% (500). Men and women aged 60-74 are screened biennially for faecal occult blood using gFOBT. 2% of participants have an abnormal test and

are offered colonoscopy (501). CRC is detected at 10% of colonoscopies and adenoma at 40%; 50% of colonoscopies reveal a normal bowel or benign non-neoplastic condition (501), reflecting the limited specificity of gFOBT (502) (Table 17).

Table 17. Factors influencing the sensitivity and specificity of gFOBT.

Factors which may cause a false negative result	Factors which may cause a false positive result
Not all CRC bleed	Dietary haem
CRC may bleed intermittently	Vegetables containing peroxidase
	Non-CRC bowel conditions which bleed
	Anticoagulant use

The NHSBCSP is transitioning to a more sensitive and specific faecal occult blood test, FIT, which has several advantages compared with gFOBT. FIT is expected to improve participant uptake from 35-60% with gFOBT (450, 452), and this has proven to be the case in Scotland (FIT was introduced in November 2017 and uptake has increased from 56% to 64%) (453). FIT requires collection of a single stool sample and minimal participant contact with the stool, compared with gFOBT.

The FIT result is machine-read and quantitative; this means that the cut-off to determine a ‘positive’ test can potentially be adjusted dependent upon screening capacity or participant demographics.

3.2.2.1 There is a need to improve screening accuracy further

In Scotland, the introduction of FIT has been associated with a higher percentage of positive results (3.2% compared with 2.1%); of which 6.1% had CRC and 41.2% adenoma detected (453). This still results in a large number of unnecessary colonoscopies; this represents a significant cost, resource and patient burden. There is a need to improve screening accuracy further.

An improved method of screening, Cologuard, has been approved by the Food and Drug Administration (445). Cologuard combines FIT with stool DNA testing. This test is relatively onerous for participants (requiring the collection of a whole stool sample, subsampling an aliquot, and the addition of a buffer to the remaining sample) and is expensive ($649), meaning that it is unsuitable for routine screening by the NHS.

Research suggests that analysis of the microbiome may augment screening methods which rely on the detection of faecal blood.

3.2.3 The potential for microbiome analysis to improve screening accuracy

3.2.3.1 Non-faecal samples as a potential screening adjunct

The majority of research investigating whether microbiome analysis can improve the accuracy of screening has been conducted using stool samples. However, as mentioned in the Introduction, it is possible to analyse the microbiome using non-faecal samples (blood, saliva, urine and breath). Preliminary studies have shown that it is possible to use non-faecal microbiome samples to screen for CRC. However, they suffer the following limitations: most were conducted using small numbers of study participants; most discriminated between patients with CRC and healthy volunteers and did not attempt to detect adenomas; most collected samples post-bowel preparation, which changes the microbiome; in some studies only the CRC patients underwent bowel preparation; and very few studies performed model validation. Bearing these limitations in mind, the results of these studies will now be outlined.

A model which used serum metabolites was able to distinguish patients with CRC from controls with 100% sensitivity and specificity (503). An ELISA to serum IgG to Fap2 was able to distinguish patients with CRC from controls with a sensitivity of 100% and specificity of 68% (504). Serum antibodies to FliC (a *Salmonella* flagellin) were shown to be elevated in patients with CRC or adenomas compared with healthy controls, although the potential for screening was not formally evaluated (505).

The oral microbiome of patients with CRC has been shown to differ significantly from controls; a model which combined data from the oral and faecal microbiome improved the area under the receiver operating characteristic curve (AUC) over a model using the faecal microbiome alone, achieving AUC for CRC of 0.94 (95% CI: 0.87-0.94) and AUC for adenoma of 0.98 (95% CI: 0.95-0.98) (98).

Metabolomic analysis of urine has been shown to distinguish CRC patients from controls (506) with an AUC of 0.998 (95% CI: 0.992-1.000) (507) and AUC 0.98 (95% CI: 0.93-1) (356). It has also been shown to be capable of discriminating between different stages of CRC (355) and between pre and post-operative states (355, 508).

Metabolomic analysis can also be performed on VOCs. In 2011 it was reported that a dog could correctly identify faecal or breath samples from CRC patients versus controls (509). A model using breath VOCs has been shown to distinguish CRC patients from controls with 91% accuracy (510), and another breath VOC model reported a sensitivity and specificity for CRC detection of 86% and 83% respectively (511). A model using faecal VOCs has been shown to distinguish CRC patients from controls with an AUC of 0.92 (95% CI: 0.89-0.95) (512). Faecal VOC models which tested participants with a positive gFOBT reported a sensitivity for CRC detection of 87.9% (95% CI: 0.87-0.99) and specificity 84.6% (95% CI: 0.65-1.0) (513); and a similar study reported a sensitivity for CRC detection of 72% and specificity of 78% (514). A model using urinary VOCs has been shown to distinguish CRC patients from controls with an AUC of 0.71 (95% CI: 0.62-0.8) but was unable to distinguish CRC patients from their spouses or first degree relatives (515). Another urinary VOC model reported a sensitivity and specificity for CRC detection of 88% and 60% respectively (516); another reported a low sensitivity and specificity in a cohort of symptomatic patients but good sensitivity (0.97, 95% CI: 0.90–1.0) and specificity (0.72, 95% CI: 0.68–0.76) for CRC detection in those with a FIT-negative result (517).

3.2.3.2 The faecal microbiome as a potential screening adjunct

Many studies have investigated the screening potential of the CRC/adenoma-associated faecal microbiome using (in descending order of cost) metabolomics, metagenomics, 16SrRNA or qPCR. Results from these studies will now be summarised, followed by a review of their limitations.

3.2.3.2.1 Metabolite-based models

A faecal metabolite-based test distinguished advanced neoplasia from controls with an AUC of 0.94 (95% CI: 0.84 to 0.99) (518).

3.2.3.2.2 Metagenomic-based models

Combining a metagenomic-based model with gFOBT improved sensitivity of CRC detection by 45% (AUC 0.87, 95% CI not provided) (519) and in another study a metagenomics-based model achieved an AUC for detection of CRC of 0.91 (95% CI not provided) (520). One study compared a metagenomic versus 16SrRNA-based screening model and demonstrated similar performance (519).

3.2.3.2.3 16SrRNA-based models

In one study, a 16SrRNA (V3-V4)-based model distinguished adenoma from controls with an AUC of 0.77 (95% CI not provided) (521). In another study, combining 16SrRNA (V4) data with clinical data (age, race and body mass index (BMI)) distinguished adenoma from controls with an AUC of 0.896 (95% CI: 0.816–0.976), CRC from controls with an AUC of 0.922 (95% CI: 0.858–0.986) and tumour from controls with an AUC of 0.936(95% CI: 0.887–0.985) (522). In a third study, combining16SrRNA (V4) data with FIT results distinguished tumour from controls with an AUC of 0.83 (95% CI not provided), detecting 37% of adenomas and 70% of CRC which would have been missed using FIT alone (523).

3.2.3.2.4 qPCR-based models

The aforementioned studies demonstrate the potential of microbiome data to improve screening accuracy. However the expense of metabolomics or NGS precludes the application of these models to national screening. Instead they generate hypotheses as to which taxa are differentially abundant; these could be detected using the cheaper and quicker methodology of qPCR. The importance of designing qPCR targets based on the results of metagenomics/16SrRNA data is exemplified by the fact that an early study which used qPCR to a limited number of taxa, not based on a prior hypothesis, did not show improvement in screening accuracy (524). In contrast, the following qPCR-based screening studies have shown an improvement.

The following studies have investigated the accuracy of a screening model based on qPCR detection of *F. nucleatum*. qPCR of *F. nucleatum* of faecal samples combined with FIT data distinguished CRC from controls with an AUC of 0.95 (95% CI: 0.92-0.98) representing an improvement over an AUC of 0.86 (95% CI: 0.81-0.90) using FIT data alone; and advanced adenoma from controls with an AUC of 0.65 (95% CI: 0.58-0.73) compared with an AUC of 0.57 (95% CI: 0.53-0.61) using FIT data alone (525). Another qPCR *F. nucleatum*-based model showed an improvement over a model which used age and gender alone (526); a further *F. nucleatum*-based model (using instead droplet digital PCR) distinguished CRC from controls with an AUC of 0.75 (95% CI not provided) (527).

The following studies have investigated the accuracy of a screening model based on qPCR detection of *F. nucleatum* plus additional taxa. A model combining the *F. nucleatum*:*Faecalibacterium prausnitzii* and *F. nucleatum*:*Bifidobacterium* ratios produced a superior AUC of 0.91 (95% CI not provided) for the detection of CRC (including Stage I CRC with an AUC of 0.80) compared to *F. nucleatum* alone (201). A model combining *F. nucleatum*, *Bacteroides clarus*, *Clostridium hathewayi*, and m7 (unidentified taxon) with FIT data, distinguished CRC from controls with a sensitivity of 92.8% and specificity of 81.5%, an improvement over a model using FIT data alone (sensitivity of 70.3%) or *F. nucleatum* alone (528). A model combining *F. nucleatum* and *Parvimonas micra* distinguished CRC (equal or greater than Stage II) from controls with an AUC of 0.84 (95% CI not provided) (529). A model combining clbA1 bacteria (pks+) and *F. nucleatum* detected CRC with a sensitivity of 84% and specificity of 63% and sensitivity was improved with the addition of FIT data (530). A model combining *F. nucleatum, Enterococcus feacalis*, *Streptococcus bovis*, ETBF and *Porphyromonas* was able to distinguish adenoma or Stage 1 CRC from controls with an AUC of 0.97 (95% CI not provided) (531). One study has shown that *Clostridium symbiosum* (detected via qPCR) is superior as a biomarker for the detection of advanced adenoma/early stage CRC compared to *F. nucleatum*. The optimum model for early stage CRC combined *Clostridium symbiosum* and FIT data to produce an AUC of 0.743 (95% CI: 0.612–0.848), and for all stages of CRC a model which combined *Clostridium symbiosum*, *F. nucleatum*, FIT data and carcinoembryonic antigen (CEA) produced an AUC of 0.876 (95% CI: 0.810–0.926) (532).

3.2.3.2.5 Meta-analysis

The aforementioned studies demonstrate the potential of a microbiome-based model to improve screening. However, they encompass a number of different screening models and a range of AUC. Meta-analysis provides a consensus estimate. One meta-analysis of faecal metagenomic studies (conducted in the USA, China and Europe) identified seven species which were consistently enriched in CRC (*Bacteroides fragilis*, *F. nucleatum*, *Porphyromonas asaccharolytica*, *Parvimonas micra*, *Prevotella intermedia*, *Alistipes finegoldii*, and *Thermanaerovibrio acidaminovorans*) which, when combined with age, gender and BMI, produced an AUC of 0.88 (95% CI not provided) for the distinction of CRC from controls (166). A different meta-analysis of ten studies has demonstrated that *F. nucleatum* could distinguish CRC from controls with an AUC of 0.86 (95% CI: 0.83–0.89); however, the result may not be generalisable to screening as some of the studies performed microbiome analysis using tissue samples (188). Another meta-analysis demonstrated that faecal *F. nucleatum* could distinguish CRC from controls with an AUC of 0.80 (95% CI: 0.76-0.83) and adenomas from controls with an AUC of 0.60 (95% CI: 0.56-0.65), although the authors reported high inter-study heterogeneity (533). Two meta-analyses of faecal 16SrRNA studies demonstrated that a combined model based on genera could distinguish CRC from controls with an AUC of 0.835 or AUC of 0.75 but a model for adenoma had poor discriminatory power (121, 477). Another meta-analysis of faecal 16SrRNA studies demonstrated that AUC could be influenced by choice of bioinformatic analysis; a model based on genera could distinguish CRC from controls with an AUC of 0.766 (sensitivity 55.3%, specificity 82.9%) and combining clinical and microbiome markers gave an AUC of 0.824 (534).

3.2.3.3 Limitations of existing studies

The results of the aforementioned studies should be interpreted with caution. The studies had one to several of the following limitations, all of which prevent their direct translation to a national screening programme.

- **The number of participants in some studies was low.** This limits statistical power.

- **Study participants were often not representative of the screening-eligible population.** In some studies participants were symptomatic; had

a history which made them high risk for CRC; or were outside the normal screening age (particularly controls).

- **Samples were collected in a manner which would be incompatible with national screening.** In many studies participants were asked to collect whole stool samples; samples were collected under anaerobic conditions; or samples were frozen within a limited time (hours) of collection, requiring either home freezing, expedited transport of the sample to the laboratory or cold-chain transport.

- **The timing of sample collection and definition of cases and controls often did not reflect the time-course or aim of screening.** Many studies collected samples post-colonoscopy (with bowel preparation +/- biopsy) which changes the microbiome; control groups often included small adenomas (<1cm); models were sometimes designed for the detection of CRC alone (sometimes Stage II and above) and neglected the need to detect adenomas and early stage CRC; and few studies evaluated the specificity of the model in patients with other microbiome associated colorectal diseases such as IBD.

- **Not all studies performed model validation**. In addition, few models were validated using cohorts from other countries; this risks limiting generalisability of the models.

- **All of the studies used single-timepoint colonoscopy diagnosis as the gold standard**. Colonoscopy has a ~2-5% CRC miss-rate (535) and ~25% adenoma miss-rate (536) therefore ideally longitudinal follow-up of controls should also be performed to exclude future adenoma/CRC development.

These limitations were confirmed by the authors of a systematic review of 19 studies assessing the screening potential of faecal samples (537). The authors concluded that statistical power may not be met by studies with small numbers of participants (the largest study contained 490 participants of which 120 were CRC patients); all studies used refrigerated or frozen faecal samples; some of the studies used samples taken post-colonoscopy, not all of the studies recorded antibiotic status; and the majority of models were not validated The authors noted

that there was inter-study heterogeneity regarding laboratory methods, reference databases and statistical analysis.

Of the limitations described, the two that are pervasive in the microbiome literature are the collection of stool post-bowel preparation and the collection of whole stool samples. These will therefore be discussed in more detail.

3.2.3.3.1 The effects of bowel preparation on the microbiome

Bowel preparation profoundly and rapidly changes the microbiome in the short-term (538, 539). The microbiome has been shown to subsequently return to baseline within a fortnight (538, 539) and after six months one study reported no more difference from baseline than due to temporal variation alone (417). However, such short and long term responses have not been consistent among all individuals studied (349, 540). This could be due to the use of different bowel preparation regimens or differences in baseline microbiomes. Given this uncertainty, research designed to investigate the microbiome as a screening tool should be conducted in bowel preparation-naïve individuals (so that results are generalisable to the bowel preparation-naïve screening population).

3.2.3.3.2 Collection of whole stool samples

Chapter 2 described alternatives to the collection of whole stool samples (gFOBT, FIT, OMNIgene.GUT). Only one study has explored whether the microbiome analysed directly from screening samples can improve screening (541). This study used simulated screening samples: FIT devices were spiked with thawed whole stool and frozen prior to DNA extraction. A microbiome-based model was able to distinguish CRC from controls with an AUC of 0.853 (95% CI not provided) and neoplasm from controls with an AUC of 0.686 (95% CI not provided). This result needs to be confirmed using real screening samples.

Two research studies have collected pre-bowel preparation samples in large numbers of individuals (including a German screening cohort), although they did not perform microbiome analysis directly from screening samples (526, 530). Instead one study required participants to freeze samples at home and the other collected samples in RNA*later*; as was discussed in Chapter 2, these methods of sample collection are suboptimal for microbiome analysis.

3.2.4 Investigating the potential to use NHSBCSP samples for microbiome-based screening

Using NHSBCSP samples for microbiome analysis would address the limitations of existing microbiome-based screening studies:

- samples would originate from the screening-eligible population
- sample collection would not affect routine screening
- samples would be collected prior to bowel preparation
- large numbers of samples would be available for model validation
- theoretically, longitudinal follow-up data could be collected

However, there are also some limitations to using NHSBCSP samples:

- Only participants with a positive gFOBT test are referred for colonoscopy and receive a definitive diagnosis. It is not possible to perform colonoscopy on participants with a negative gFOBT test as this would disrupt routine screening. The number of false negatives is expected to be small; research has shown that of those participants with a negative gFOBT test who are subsequently screened two years later, 0.112% (95% CI: 0.100–0.125) are diagnosed with CRC and 0.315% (95% CI: 0.295–0.336) intermediate or high-risk adenoma (542). It should be noted however that these figures represent a conservative estimate; they only consider participants who attended two consecutive screening rounds, they do not consider the miss-rate of the second of the two screening rounds, and they do not include low-risk adenomas.
- Screening colonoscopy does have an associated miss-rate. However, the number of false negatives within this group is expected to be small; research has shown that of those participants who have had a colonoscopy (or other screening diagnostic investigation) negative for CRC or intermediate/high-risk adenoma who are subsequently screened at the next round, 0.172% (95% CI: 0.131–0.221) are diagnosed with CRC and 0.718% (95% CI: 0.632–0.813) with intermediate or high-risk

adenoma (542). Although for the same reasons as noted above, these figures represent a conservative estimate.

- Limited demographic/clinical information is recorded by the NHSBCSP database. Although most causes of microbiome variability are currently unknown, a small number of research studies try to record and account for recognised sources of variation through participant questionnaires. This would not be possible as it would disrupt routine screening. This caused uncertainty as to whether a microbiome-based screening model, in the absence of this information, would prove successful. The hypothesis was that it would, as the majority of the aforementioned microbiome-based screening studies did not include any or only limited participant metadata in their screening models. Potential sources of microbiome variation will now be discussed.

3.2.4.1 Potential sources of microbiome variation

3.2.4.1.1 Temporal variation

The faecal microbiome is broadly temporally stable over a period of months-years (up to two have been studied) with repeat samples being more similar to one another than to samples from another individual (20, 417, 418). However the microbiome is not static; diurnal circadian fluctuations in taxonomic composition and function occur (543, 544) as do minor taxonomic shifts in response to day-to-day variation in aspects such as diet (419). Larger taxonomic shifts occur in response to perturbations (e.g. foreign travel and gastrointestinal infection) (419). Seasonal changes in the microbiome have been recorded for remote tribes (545) and farming communities (546), probably reflective of seasonal changes in diet; to the author's knowledge this has not been investigated in urban cohorts, but could be expected to occur.

The extent to which the CRC-associated microbiome varies temporally is unknown; short-term variation of the established CRC-associated microbiome was considered in Chapter 2 by comparing the microbiome of stool samples collected over several days.

Two other areas of uncertainty are:

- When is the adenoma/CRC-associated microbiome established?
- How does the microbiome change during tumourigenesis?

These questions are very difficult to address in human studies and instead are being addressed by animal work, where it is possible to analyse the microbiome at different stages of tumourigenesis (150). Interestingly one study showed that in a mouse genetic CRC model, dysbiosis occurred prior to the development of microscopically detectable polyposis (165).

3.2.4.1.2 Inter-individual variation

Inter-individual variation accounts for the highest source of taxonomic variation in the microbiome (344). Interestingly inter-individual variation of the microbiome associates with physiological inter-individual variation, including inter-individual variation of drug metabolism, post prandial glucose responses and male/postmenopausal circulating oestrogen levels, and associates with anthropometric phenotypes (e.g. BMI, blood cholesterol concentration) to the same extent as SNPs (297, 547, 548).

Inter-individual microbiome variation is likely a result of the sources of variation which are described below, in addition to prior exposures (for example during childhood) which are not yet understood.

3.2.4.1.3 Gender

Gender has been shown to associate with differences in the microbiome (549, 550). The underlying mechanism is not known; hypotheses include gender differences in hormonal status, genetics or gastrointestinal physiology (551). Many microbiome studies do not consider gender in their analysis, yet there is some evidence that the association between the microbiome and certain variables may be gender-specific (549, 552). Ideally microbiome studies should perform gender matching of cases and controls.

3.2.4.1.4 Age

The microbiome of infants differs to adults; the microbiome of adults at retirement age or above differs to younger adults; the microbiome of centenarians has been shown to differ from adults of retirement age; and the microbiome of people aged greater than 105 has been shown to differ from centenarians (5, 553-558). Ideally microbiome studies should select cohorts of a defined age range.

Clearly any association between the microbiome and age is complex; changes in the microbiome are likely to be influenced by age-related changes in diet, exercise and physiology (559). However, mechanistic animal studies suggest that the association could be bi-directional. In a fish model, the transplant of microbiomes from younger donors to older recipients led to increased survival (560). In a mouse model the microbiome was associated with increased colonic permeability and inflammatory cytokine profiles with age (561) i.e. contributing to the concept of 'inflammaging' (562)

3.2.4.1.5 Genetics

It is estimated that genetics account for 1.9-8.1% of variability in the microbiome; no association has been found with genetic ancestry or individual SNPs (548). As small studies investigating the association between the microbiome and genetics have produced conflicting results; a large-scale meta-analysis is currently being conducted (373).

3.2.4.1.6 Diet

Broad mammalian diet categories (carnivorous, herbivorous, omnivorous) associate with differences in microbiome composition and function (563). The microbiome of individuals consuming a 'Western' diet differs from the microbiome of individuals consuming a low-fat, high-fibre diet, such as that consumed in rural Africa (564). Western diet associates with a predominance of *Bacteroides* and *Clostridiales* whereas high-fibre low-protein diets associate with a predominance of *Prevotella* (16). Within Western cohorts, fibre intake associates with differences in microbiome composition (565).

Changes in diet induce rapid and reversible changes in the taxonomic composition and function of the microbiome; changes occur within one day of food reaching the colon and reverse two days after cessation of the dietary

intervention (566, 567). Diet-induced changes in the microbiome reflect a change in the resident microbiome in addition to the introduction of novel food-borne bacteria, fungi and viruses (566). Depending upon the dietary intervention and the individual response, changes in the microbiome secondary to diet may exceed inter-individual variability (566).

3.2.4.1.7 Medications and antibiotics

The colorectal microbiome is particularly sensitive to antibiotics (more so than the oral microbiome, for example) (568). Antibiotics have a dramatic and immediate effect on diversity (observed within four days), taxa (affecting up to a third of taxa) and function (569, 570), and lead to an increase in the abundance of antibiotic resistance genes (568, 570), microbial virulence factors (569) and low-abundance taxa (569). The majority of taxa return to baseline abundances within a month, but studies suggest that a few taxa and functions remain perturbed at one year post-antibiotic cessation (417, 569, 571, 572). One quarter of non-antibiotic medications have been shown *in vitro* to inhibit the growth of a gut commensal; some were also shown to inhibit the growth of the CRC-associated bacteria ETBF and *F. nucleatum* (573).

Many microbiome studies make antibiotic usage an exclusion criteria; this will not be possible in this study as it would disrupt national screening. There is no universal ‘antibiotic microbiome signature’; there are many different antibiotics, routes and course durations and responses of individual taxa vary by antibiotic and individual (553, 570, 571). Antibiotic resistance genes cannot be used as an indicator of recent antibiotic exposure as they are present in the antibiotic-naïve microbiome (574), partly reflecting environmental sources of antibiotic exposure (575).

3.2.4.1.8 Smoking

Smoking status has been shown to associate with changes in the microbiome, although few microbiome studies account for this (576, 577).

3.2.4.1.9 Comorbidities

Disease-microbiome associations are complicated by the fact that individuals (particularly the elderly) often have more than one comorbidity; controls may have a past medical history of disease (with associated long-term changes in the microbiome); controls may have undiagnosed latent disease; and many diseases are accompanied by medication use. Disease-microbiome associations have been described for a large number of diseases:

- gastrointestinal diseases: IBD (578), IBS(579)
- rheumatic diseases (580)
- metabolic diseases (obesity, cardiovascular disease, hypertension and diabetes) (581-585)
- neurological diseases (Parkinson's disease, Alzheimer's disease, depression, schizophrenia and autism) (586)

3.2.4.1.10 Other factors

Other factors which have been associated with changes in the microbiome include cohabitation (548, 587), being members of the same family (18), pregnancy (588) and average sleep duration (589). These factors encompass some of the aforementioned sources of microbiome variation (both past and present) which makes determining direct associations with the microbiome difficult.

Current understanding of factors which influence the microbiome is incomplete. A comprehensive list of 126 factors which included age, gender, disease status, diet, antibiotic and drug use, smoking status, stool frequency and type, and blood cytokine profile, was only able to account for ~19% of inter-individual microbiome variation (590). Associations were identified between 125 bacterial species and 110 different factors (590). Certain CRC-associated bacteria have been shown to associate with certain factors, although *Fusobacterium* did not show any associations, perhaps indicating that *Fusobacterium* may represent a more robust biomarker (591).

3.3 Aims

- To investigate the microbiome of NHSBCSP gFOBT samples.
- To determine whether the microbiome differs according to the presence/absence of faecal blood.
- To determine whether the microbiome differs according to the underlying pathology.
- To determine whether the CRC-associated bacteria described in the literature are present within a bowel-preparation naïve screening population.
- To assess whether microbiome analysis improves the accuracy of screening.

3.4 Methods

3.4.1 Collaborators & ethical approval

Details of the study collaborators and ethical approvals were as described in Chapter 2.

3.4.2 Samples

3.4.2.1 Sample collection and processing

As this was the first study to analyse the microbiome from gFOBT screening samples, a power calculation was not possible. A target sample size of 200 samples per clinical group (blood-negative gFOBT; colonoscopy normal; adenoma; and CRC) was proposed based on the sample sizes of published microbiome studies. These target sizes were exceeded due to the increased capacity of the Illumina HiSeq. An additional clinical group (benign diagnosis at colonoscopy) was subsequently included (with a subsequent smaller sample size), so that a Random Forest model could be trained using all possible colonoscopy outcomes (Figure 84).

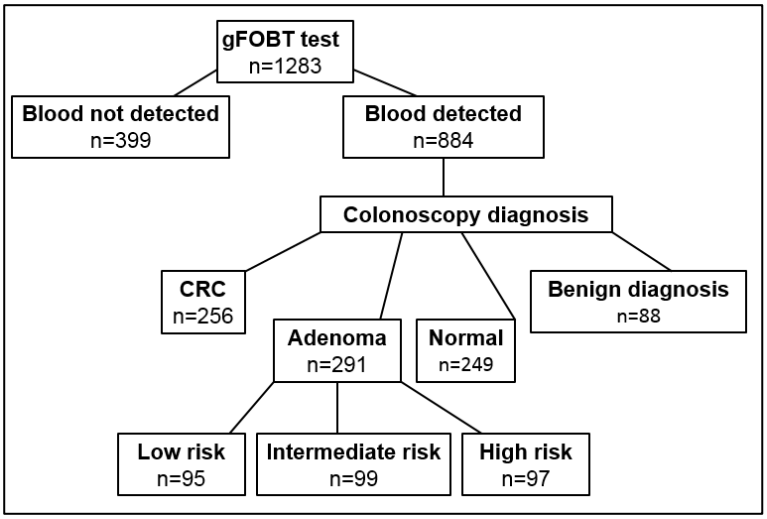


Figure 84. The number of NHSBCSP samples processed.

Details of sample collection, link-anonymisation, transport, storage and laboratory processing were as described in Chapter 2.

3.4.3 Clinical data

3.4.3.1 Extraction

The Southern Hub extracted the link-anonymised data from the OBI-EE national NHSBCSP database (Table 18). Extraction took place outside of office hours to minimise disruption to the NHSBCSP. Data is based on information collected and quality assured by Public Health England (PHE) Population Screening Programmes. Access to the data was facilitated by the PHE Office for Data Release (Table 18).

Table 18. Link-anonymised clinical metadata. NHSBCSP definitions of adenoma risk are described by Logan *et al* (592).

Clinical category	**Possible sub-groups**
Age	
Gender	
Round of screening	
Episode outcome	definitively normal
	definitively abnormal
Diagnostic test result (greatest risk)	normal
	low-risk adenoma: • 1-2 adenomas each of diameter <1 cm
	intermediate-risk adenoma: • 3-4 adenomas each of diameter <1 cm • or one adenoma with a diameter ≥1 cm
	high-risk adenoma: • ≥ 5 adenomas • or ≥3 adenomas of which one has a diameter ≥1 cm
	CRC
	'other': non-neoplastic colonoscopy finding

The following data has been requested but is not yet available:

- outcome of preceding screening episodes
- CRC: number, location, type, grade
- adenoma: number, location
- 'other' non-neoplastic colonoscopy finding: specific diagnosis. Diagnoses recorded as 'other' in the NHSBCSP database include diverticular disease, haemorrhoids, ulcerative colitis, angiodysplasia, Crohn's disease, solitary rectal ulcer syndrome and radiation proctitis (592).

3.4.3.2 Data transfer and storage

Link-anonymised data was transferred to an NHS computer at the University of Leeds via encrypted NHS Secure File Transfer (with passwords conveyed separately via telephone). This data was transferred to a University of Leeds computer via encrypted memory stick and uploaded to the secure University of Leeds Secure Electronic Environment for Data (SEED) system for storage.

3.4.4 Bioinformatic processing

Bioinformatic processing was performed as described in Chapter 2.

3.4.5 Statistical analysis

Statistical analysis was performed as described in Chapter 2. Additionally, Random Forest models and AUC were generated in R (version 3.5.0) using the packages randomForest (593, 594) and pROC (595) respectively. A decision was made to run the models using the relative abundance of bacteria at genus level (as opposed to higher taxa or individual OTUs) as the output would be more biologically meaningful and easier to convert into qPCR primers. Examples of Random Forest commands in R were sourced from online material; relevant commands, including those for partial dependence plots and variable importance analysis, were combined (by the author) into a novel composite script (written by Dr Wood). Each forest was built with 1000 trees which ensured a stable Out of Bag (OOB) error (an example of which is shown in (Figure 85)). The number of variables available for testing at each node (mtry) was determined based on the mtry value corresponding to the lowest OOB error (Figure 85).

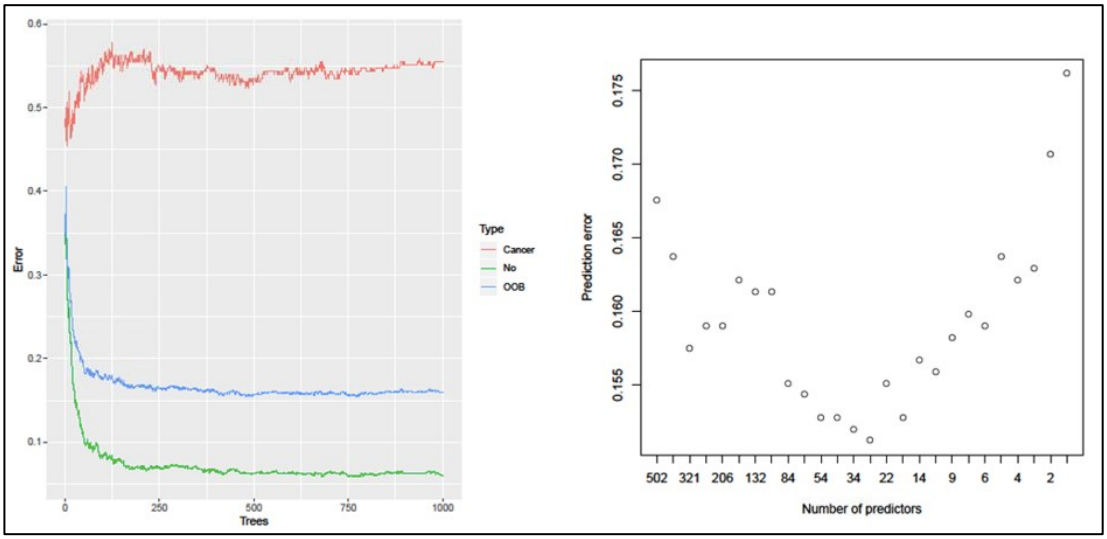


Figure 85. The effect of the number of trees on the Random Forest Out of Bag error rate and of the number of predictors on prediction error.. Left-hand plot: this shows that with a small number of trees (<125) the error for predicting CRC (red), not-CRC (green) and the OOB error (blue) have not stabilised; with 1000 trees the errors are stable. The OOB error is based on data that each tree has not been presented with. Right-hand plot: the number of predictors corresponding to the lowest prediction error (the lowest point on the graph) was selected as the mtry value.

95% CI for the receiver operating characteristic (ROC) curves and AUC were created using the default setting of 2000 stratified bootstrap replicates. AUC were compared using the roc.test command in R, which compares paired ROC using the method of DeLong *et al* (596).

3.5 Results

3.5.1 Table of characteristics

Table 19 demonstrates the characteristics of the clinical groups of NHSBCSP samples. The gender bias in the groups reflects the male predominance of CRC. The mean age per group reflects the NHSBCSP eligible screening age range (60-74 years). The age range of the NHSBCSP samples (60-89) indicates that a minority of participants (28 = 2%) were older than the upper age limit of screening.

Table 19. Table of characteristics for NHSBCSP samples.

	Male (%)	Mean age (SD)
Blood-negative (n=399)	170 (42.6)	67.1 (4.5)
Blood-positive (n=884)	546 (61.8)	67.2 (4.9)
CRC (n=256)	172 (67.2)	68.5 (5.2)
Adenoma (n=291)	196 (67.4)	66.3 (4.8)
Normal colonoscopy (n=249)	130 (52.2)	66.8 (4.3)
Non-neoplastic diagnosis (n=88)	48 (54.5)	67.8 (4.9)

3.5.2 Alpha diversity

A significant difference in Shannon diversity index was detected between the different clinical groups (Kruskal-Wallis $p = 1.24 \times 10^{-16}$) (Figure 86). It should be noted that the difference between group medians was relatively small and the range of Shannon diversity values within each group large.

Pairwise analysis (Table 20) indicated significant differences between all groups apart from: Other vs Adenoma; Cancer vs Negative; Normal colonoscopy vs Other. Interestingly CRC samples had a higher average alpha diversity and colonoscopy-normal samples had a lower average alpha diversity.

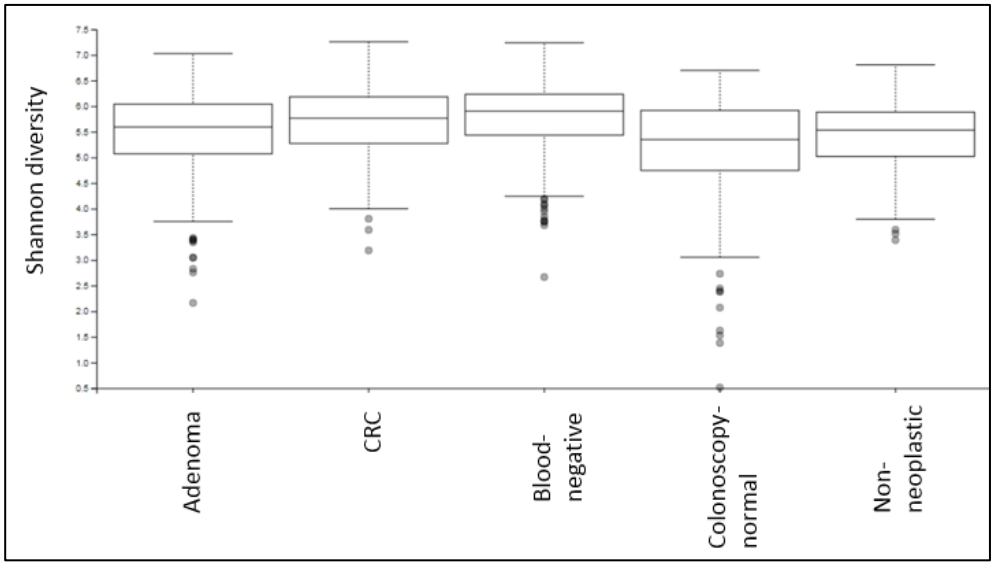


Figure 86. Boxplots of Shannon diversity index for NHSBCSP samples.

Table 20. Pairwise Kruskal-Wallis analysis of Shannon diversity index for NHSBCSP samples. Other = non-neoplastic colonoscopy diagnosis. Significant q values are shaded grey. Values are recorded to two decimal places.

Group 1	Group 2	H	p-value	q-value
Adenoma	Cancer	12.47	4.15×10^{-4}	8.29×10^{-4}
	Negative	28.89	7.67×10^{-8}	2.56×10^{-7}
	Normal colonoscopy	8.98	2.73×10^{-3}	3.90×10^{-3}
	Other	0.84	0.36	0.36
Cancer	Negative	1.86	0.17	0.22
	Normal colonoscopy	35.70	2.30×10^{-9}	1.15×10^{-8}
	Other	11.26	7.94×10^{-4}	1.32×10^{-3}
Negative	Normal colonoscopy	59.75	1.07×10^{-14}	1.07×10^{-13}
	Other	18.96	1.34×10^{-5}	3.34×10^{-5}
Normal colonoscopy	Other	1.650	0.20	0.22

3.5.3 Beta diversity

PCA of Bray-Curtis distances indicated a degree of separation of samples according to clinical status (Figure 87). However, whilst there is visible separation on PCA, there is also a wide range of Bray-Curtis distances within groups. PERMANOVA analysis of the variables ‘clinical group’, gender, age, ‘screening episode’ and ‘time until DNA extraction’ confirmed that ‘clinical group’ contributes to the largest amount of variation in Bray-Curtis distance, although the amount is small with $R^2 = 0.02$ (Table 21).

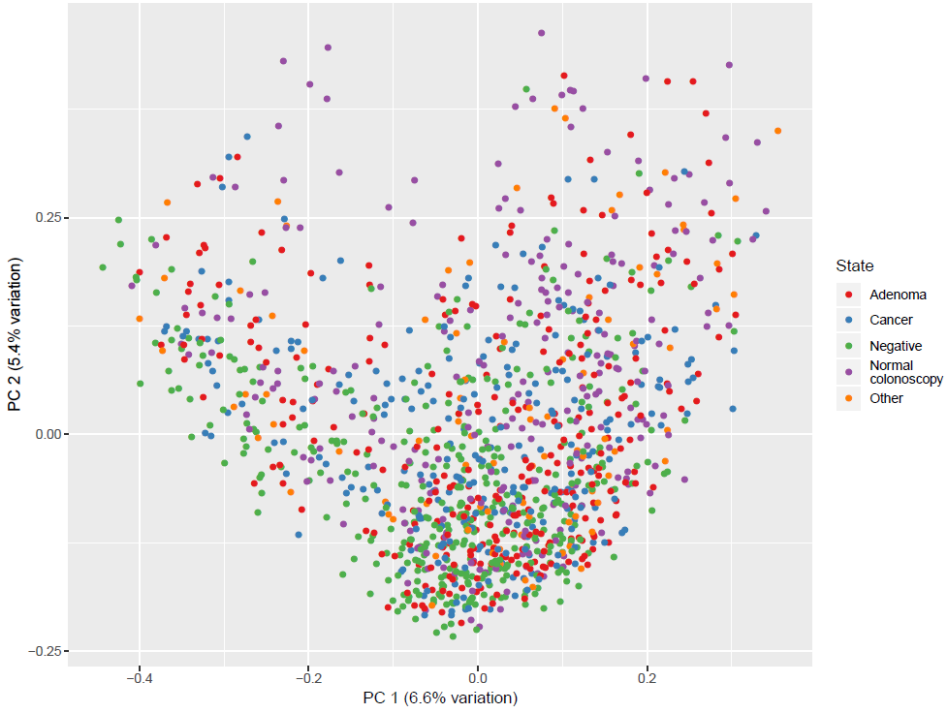


Figure 87. PCA of Bray-Curtis distances for NHSBCSP samples. Points on the graph are coloured according to disease status.

Table 21. Results of PERMANOVA analysis of NHSBCSP samples.. Df = degrees of freedom. Clinical group includes adenoma, CRC, blood-negative, colonoscopy-normal and non-neoplastic colonoscopy diagnosis. NA = not applicable. Significant p values are shaded grey. Values are recorded to two decimal places.

	Df	**Sums of Squares**	**F.Model**	**R^2**	**Pr(>F)**
Clinical group	4	5.72	5.04	0.02	1×10^{-4}
Gender	1	1.23	4.35	3.33×10^{-3}	1×10^{-4}
Age	1	1.06	3.73	2.86×10^{-3}	1×10^{-4}
Screening episode	5	1.26	0.89	3.41×10^{-3}	0.89
Time until DNA extraction	1	0.37	1.32	1.01×10^{-3}	0.08
Residuals	1270	360.04	NA	0.97	NA
Total	1282	369.69	NA	1.00	NA

3.5.4 LEfSe analysis

LEfSe analysis did not identify taxa uniquely enriched in one of the clinical groups, when samples from all the clinical groups were analysed together. However, LEfSe analysis of pairs of clinical groups did identify taxa that were significantly enriched/depleted.

3.5.4.1 Comparison of CRC/neoplasm samples with all other samples

Figure 88 and Figure 89 show taxa significantly enriched in CRC and neoplasm samples compared with all other samples respectively. Phylogenetic differences between the two pairs of groups can be appreciated by comparing the cladograms. The following CRC-associated bacteria described in meta-analyses of faecal studies as being enriched in CRC compared to controls were identified: *Alistipes, Porphyromonas, Bacteroides, Parabacteroides, Ruminococcus_torques group* and *Prevotella7* (121, 166, 477, 496, 534, 597). Interestingly *Fusobacterium* is enriched in non-neoplasm samples compared with neoplasm samples. This may be due to differences between blood-negative and colonoscopy-normal samples, described in the next section.

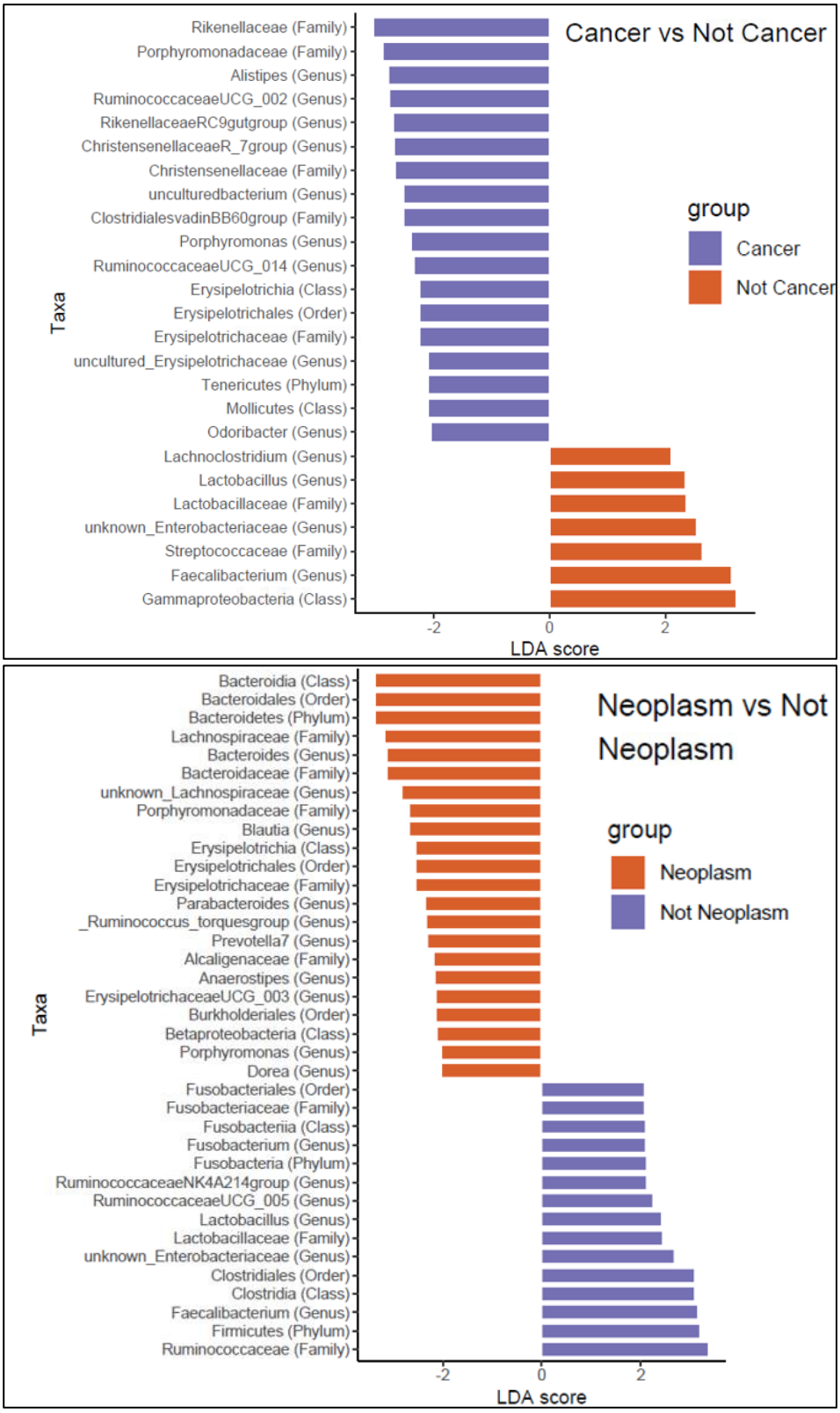


Figure 88. LEfSe plot of NHSBCSP samples (CRC or neoplasm compared with not). Upper plot: taxa which are significantly enriched in CRC samples (purple) compared with not-CRC samples (orange). Lower plot: taxa which are significantly enriched in neoplasm samples (orange) compared with not-neoplasm samples (purple). Taxa are ranked according to effect size.

The genera from Figure 88 are displayed in Table 22 for clarity:

Table 22. Genera enriched/depleted in CRC compared with 'not CRC' and neoplasm compared with 'not neoplasm'.

Genera enriched in CRC compared with 'not CRC'	**Genera depleted in CRC compared with 'not CRC'**
Alistipes	*Faecalibacterium*
RuminococcaceaeUCG_002	Unknown *Enterobacteriaceae*
RikenellaceaeRC9 gut group	*Lactobacillus*
ChristensenellaceaeR_7group	*Lachnoclostridium*
Uncultured bacterium	
Porphyromonas	
RuminococcaceaeUCG_014	
Uncultured *Erysipelotrichaceae*	
Odoribacter	
Genera enriched in neoplasm compared with 'not neoplasm'	**Genera depleted in neoplasm compared with 'not neoplasm'**
Bacteroides	*Faecalibacterium*
Unknown *Lachnospiraceae*	Unknown *Enterobacteriaceae*
Blautia	*Lactobacillus*
Parabacteroides	*RuminococcaceaeUCG_005*
Ruminococcus_torques group	*RuminococcaceaeNK4A214 group*
Prevotella7	*Fusobacterium*
Anaerostipes	
ErysipelotrichaceaeUCG_003	
Porphyromonas	
Dorea	

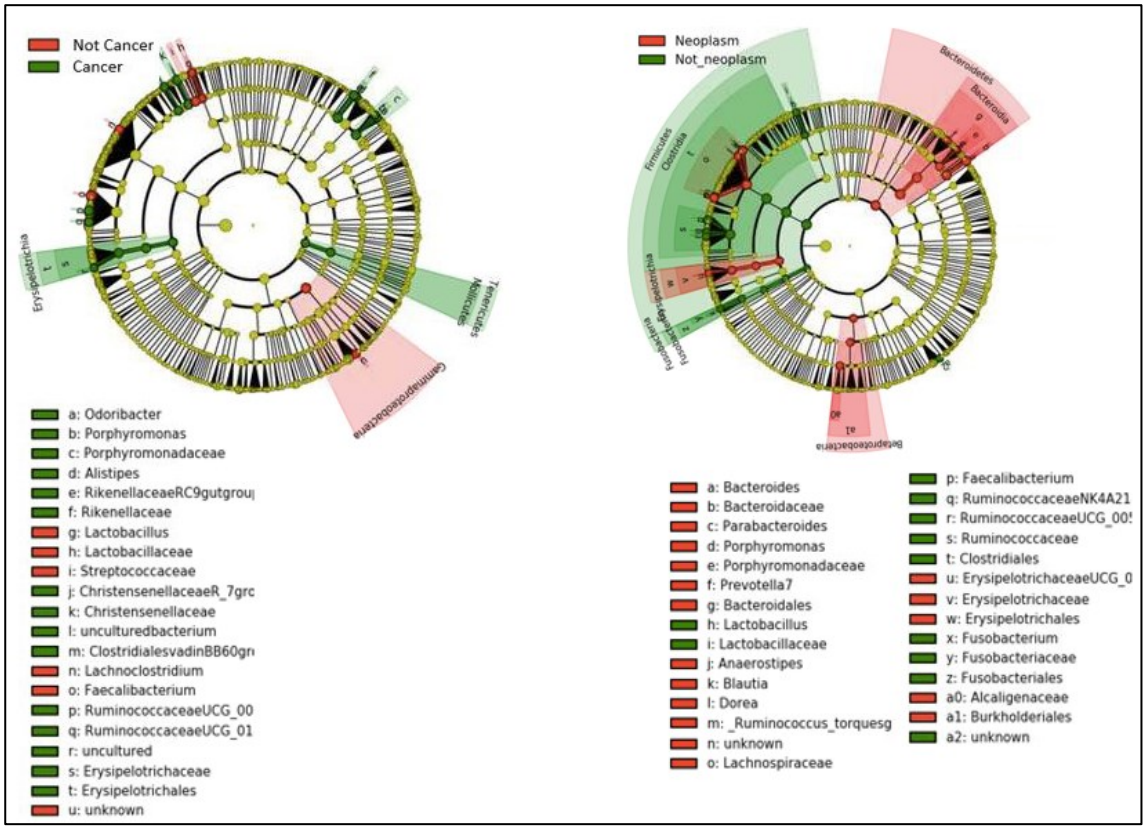


Figure 89. Cladograms of NHSBCSP samples (CRC or neoplasm compared with not). Cladograms indicate the phylogenetic relationship between taxa. Working outwards, the five rings denote phylum, class, order, family, and genus. Taxa are represented by circles. Circle diameter is proportional to abundance. Left-hand plot: circle colour indicates taxa which are significantly enriched in CRC samples (green) and non-CRC samples (red). Right-hand plot: circle colour indicates taxa which are significantly enriched in neoplasm samples (red) and non-neoplasm samples (green).

3.5.4.2 CRC compared with adenoma samples

The following taxa described as CRC or adenoma-associated in meta-analyses of faecal studies were identified (Figure 90): *Alistipes, Porphyromonas, Fusobacterium, Barnesiella* and *Ruminococcus_torques group* (121, 166, 477, 496, 534, 597). It should be noted that the meta-analyses determined differentially abundant taxa between CRC compared with controls or adenoma compared with controls, rather than between CRC and adenoma as per the current analysis.

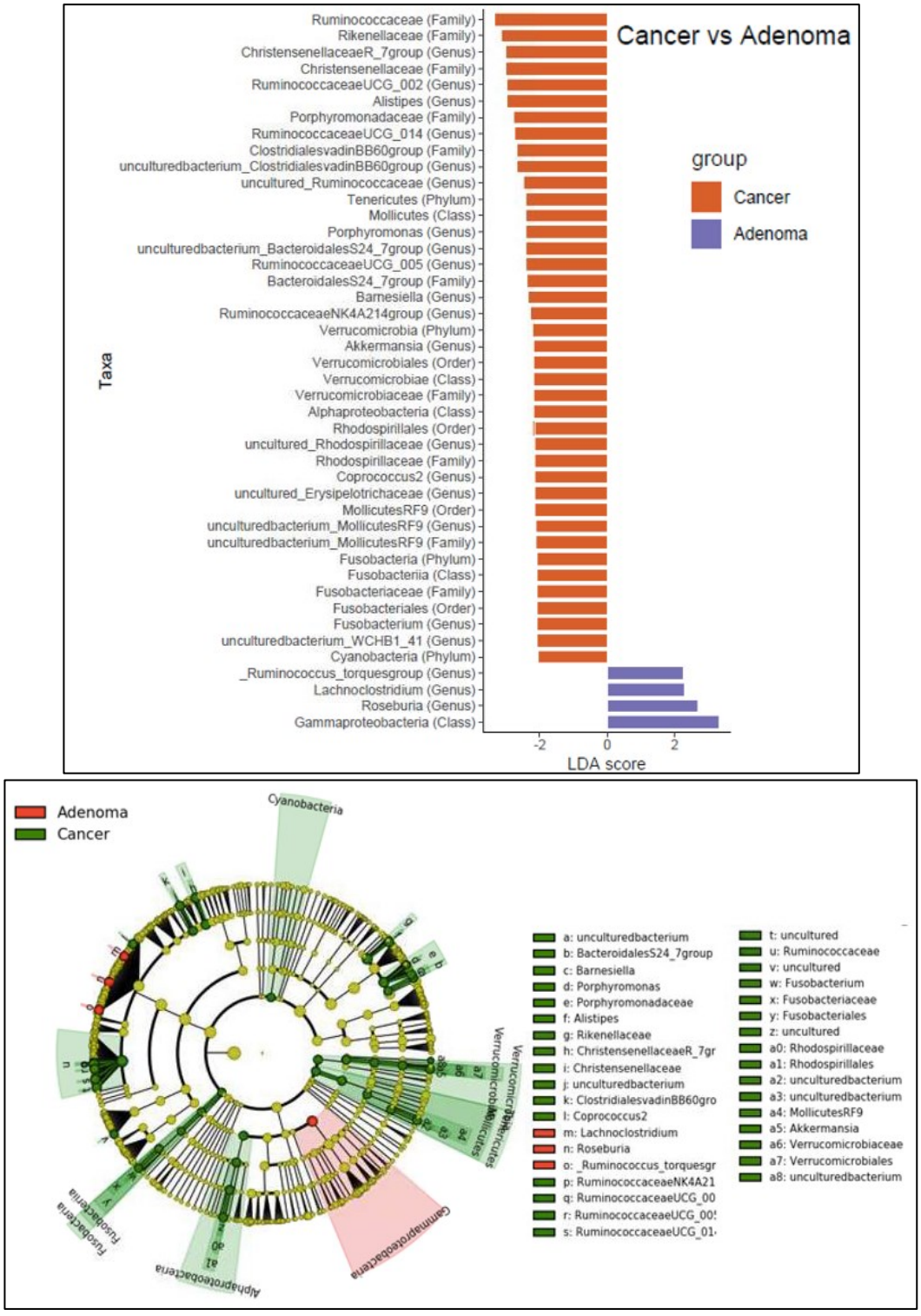


Figure 90. LEfSe plot and cladogram of NHSBCSP samples (CRC compared with adenoma). LEfSe plot indicates taxa which are significantly enriched in CRC samples (orange) and adenoma samples (purple), ranked according to effect size. The cladogram indicates the phylogenetic relationship between taxa. Working outwards, the five rings denote phylum, class, order, family, and genus. Taxa are represented by circles. Circle diameter is proportional to abundance. Circle colour indicates taxa which are significantly enriched in adenoma samples (red) and CRC samples (green).

The genera from Figure 90 are displayed in Table 23 for clarity:

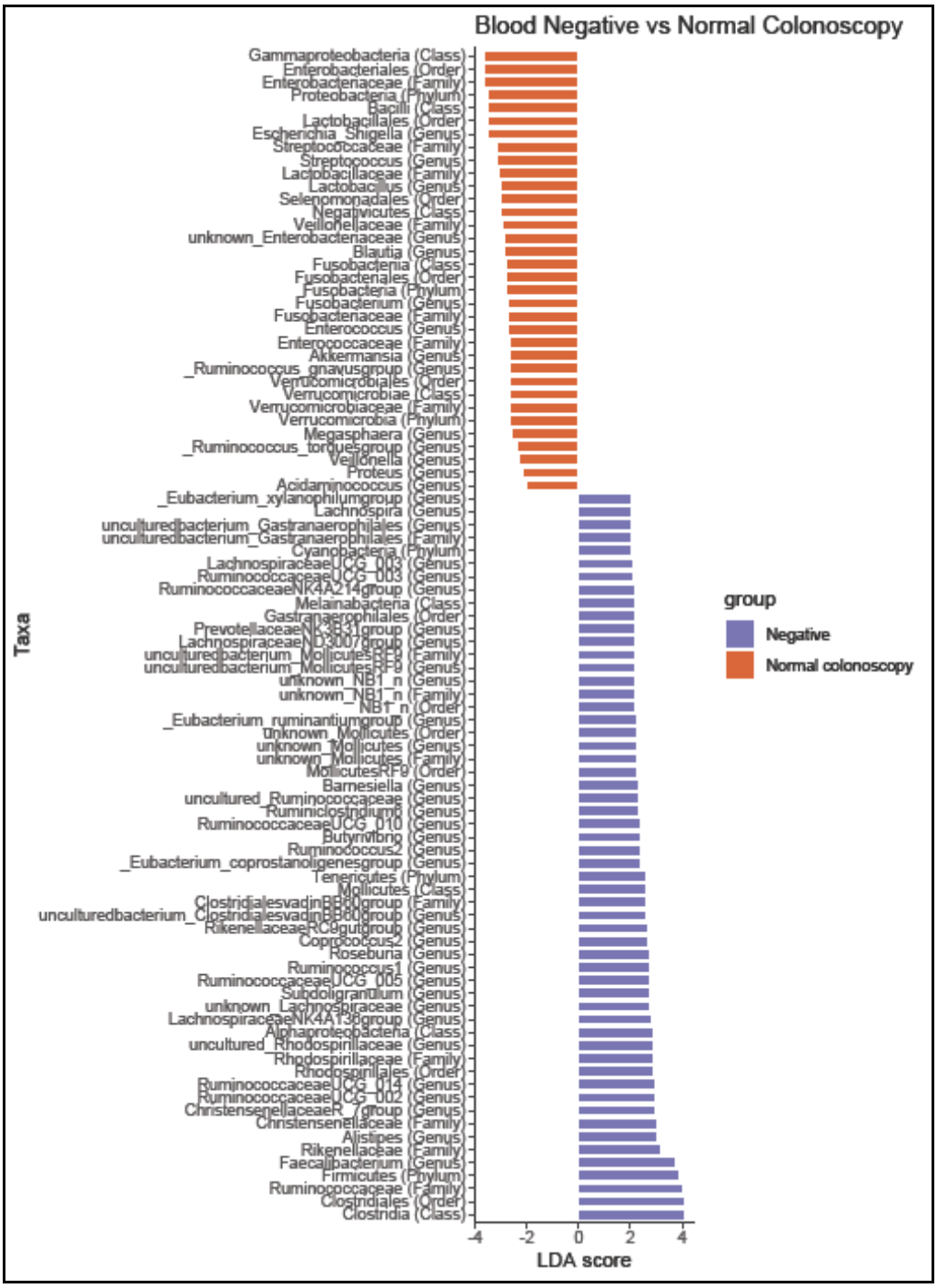
Table 23. Genera enriched/depleted in CRC compared with adenoma.

Genera enriched in CRC compared with adenoma	Genera enriched in adenoma compared with CRC
ChristensenellaceaeR_7 group	*Roseburia*
RuminococcaceaeUCG_002	*Lachnoclostridium*
Alistipes	*Ruminococcus_torques group*
RuminococcaceaeUCG_014	
Uncultured bacterium *ClostridialesvadinBB60 group*	
Uncultured *Ruminococcaceae*	
Porphyromonas	
Uncultured bacterium *BacteroidalesS24_7 group*	
RuminococcaceaeUCG_005	
Barnesiella	
RuminococcaceaeNK4A214 group	
Akkermansia	
Uncultured *Rhodospirillaceae*	
Coprococcus2	
Uncultured *Erysipelotrichaceae*	
Uncultured bacterium *MollicutesRF9*	
Fusobacterium	
Uncultured bacterium *WCHB1_41*	

3.5.4.3 Blood-negative compared with colonoscopy-normal samples

Most microbiome research studies compare CRC samples to either healthy volunteer samples (the equivalent of blood-negative samples) or colonoscopy-normal samples; it is uncommon for both control groups to be included within a single study. Figure 91 indicates taxa which were significantly enriched in colonoscopy-normal compared with blood-negative samples. This includes taxa which have been identified in existing studies of dietary haem or iron supplementation. Interestingly, the list of taxa which are differentially abundant between the two groups includes CRC or adenoma-associated taxa which have been described in meta-analyses of faecal studies as follows (121, 166, 477, 496, 534, 597):

- Enriched in colonoscopy-normal compared with blood-negative:
    - Taxa described in meta-analyses as enriched in CRC compared with controls: *Escherichia-Shigella, Streptococcus, Fusobacterium, Ruminococcus_torques group*
    - Taxon described in meta-analyses as enriched in adenoma compared with controls: *Veillonella*
- Enriched in blood-negative compared with colonoscopy-normal:
    - Taxa described in meta-analyses as enriched in CRC compared with controls: *Alistipes, Subdoligranulum*
    - Taxon described in meta-analyses as enriched in adenoma compared with controls: *Barnesiella*
    - Taxa described in meta-analyses as depleted in CRC compared with controls: *Roseburia, Coprococcus2*
    - Taxon described in meta-analyses as depleted in adenoma compared with controls: *Lachnospira*


Blood Negative vs Normal Colonoscopy
Gammaproteobacteria (Class)
Enterobacteriales (Order)
Enterobacteriaceae (Family)
Proteobacteria (Phylum)
Bacilli (Class)
Lactobacillales (Order)
Escherichia_Shigella (Genus)
Streptococcaceae (Family)
Streptococcus (Genus)
Lactobacillaceae (Family)
Lactobacillus (Genus)
Selenomonadales (Order)
Negativicutes (Class)
Veillonellaceae (Family)
unknown_Enterobacteriaceae (Genus)
Blautia (Genus)
Fusobacteriia (Class)
Fusobacteriales (Order)
Fusobacteriia (Phylum)
Fusobacterium (Genus)
Fusobacteriaceae (Family)
Enterococcus (Genus)
Enterococcaceae (Family)
Akkermansia (Genus)
_Ruminococcus_gnavusgroup (Genus)
Verrucomicrobiales (Order)
Verrucomicrobiae (Class)
Verrucomicrobiaceae (Family)
Verrucomicrobia (Phylum)
Megasphaera (Genus)
_Ruminococcus_torquesgroup (Genus)
Veillonella (Genus)
Proteus (Genus)
Acidaminococcus (Genus)
_Eubacterium_xylanophilumgroup (Genus)
Lachnospira (Genus)
unculturedbacterium_Gastranaerophilales (Genus)
unculturedbacterium_Gastranaerophilales (Family)
Cyanobacteria (Phylum)
LachnospiraceaeUCG_003 (Genus)
RuminococcaceaeUCG_003 (Genus)
RuminococcaceaeNK4A214group (Genus)
Melainabacteria (Class)
Gastranaerophilales (Order)
PrevotellaceaeNK3B31group (Genus)
LachnospiraceaeND3007group (Genus)
unculturedbacterium_MollicutesRF9 (Family)
unculturedbacterium_MollicutesRF9 (Genus)
unknown_NB1_n (Genus)
unknown_NB1_n (Family)
NB1_n (Order)
_Eubacterium_ruminantiumgroup (Genus)
unknown_Mollicutes (Order)
unknown_Mollicutes (Genus)
unknown_Mollicutes (Family)
MollicutesRF9 (Order)
Barnesiella (Genus)
uncultured_Ruminococcaceae (Genus)
Ruminiclostridium6 (Genus)
RuminococcaceaeUCG_010 (Genus)
Butyrivibrio (Genus)
Ruminococcus2 (Genus)
_Eubacterium_coprostanoligenesgroup (Genus)
Tenericutes (Phylum)
Mollicutes (Class)
ClostridialesvadinBB60group (Family)
unculturedbacterium_ClostridialesvadinBB60group (Genus)
RikenellaceaeRC9gutgroup (Genus)
Coprococcus2 (Genus)
Roseburia (Genus)
Ruminococcus1 (Genus)
RuminococcaceaeUCG_005 (Genus)
Subdoligranulum (Genus)
unknown_Lachnospiraceae (Genus)
LachnospiraceaeNK4A136group (Genus)
Alphaproteobacteria (Class)
uncultured_Rhodospirillaceae (Genus)
Rhodospirillaceae (Family)
Rhodospirillales (Order)
RuminococcaceaeUCG_014 (Genus)
RuminococcaceaeUCG_002 (Genus)
ChristensenellaceaeR_7group (Genus)
Christensenellaceae (Family)
Alistipes (Genus)
Rikenellaceae (Family)
Faecalibacterium (Genus)
Firmicutes (Phylum)
Ruminococcaceae (Family)
Clostridiales (Order)
Clostridia (Class)
Taxa
-4
-2
0
2
4
LDA score
group
Negative
Normal colonoscopy

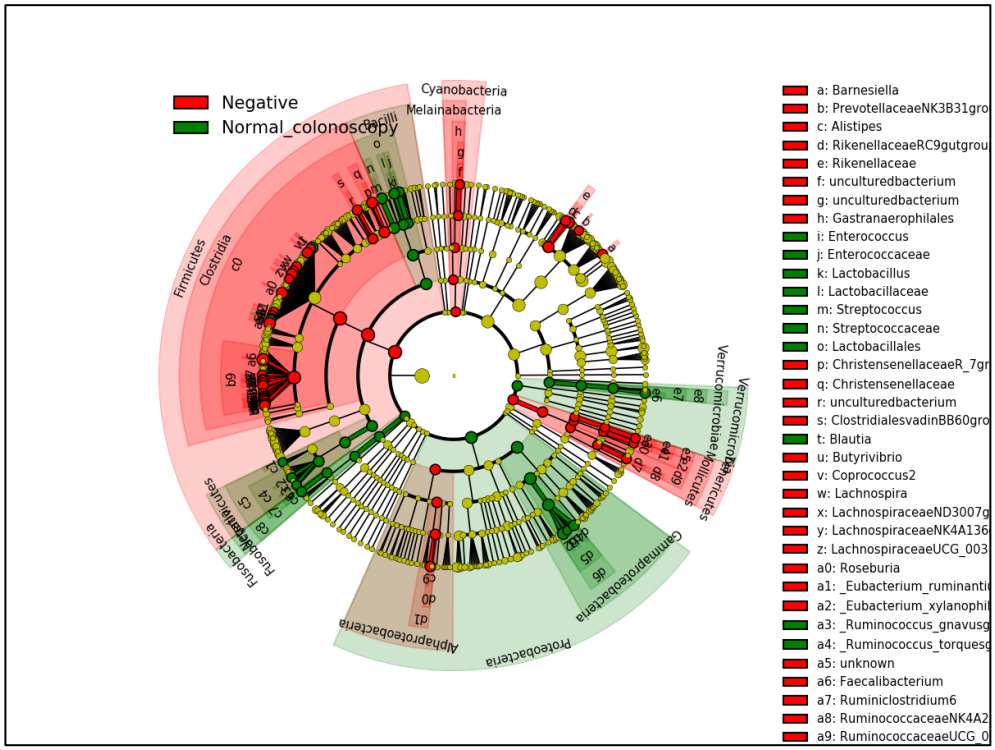


Figure 91. LEfSe plot and cladogram of NHSBCSP samples (blood-negative compared with colonoscopy-normal). LEfSe plot indicates taxa which are significantly enriched in colonoscopy-normal samples (orange) and blood-negative samples (purple), ranked according to effect size. The cladogram indicates the phylogenetic relationship between taxa. Working outwards, the five rings denote phylum, class, order, family, and genus. Taxa are represented by circles. Circle diameter is proportional to abundance. Circle colour indicates taxa which are significantly enriched in blood-negative samples (red) and colonoscopy-normal samples (green).

The genera from Figure 91 are displayed in Table 24 for clarity.

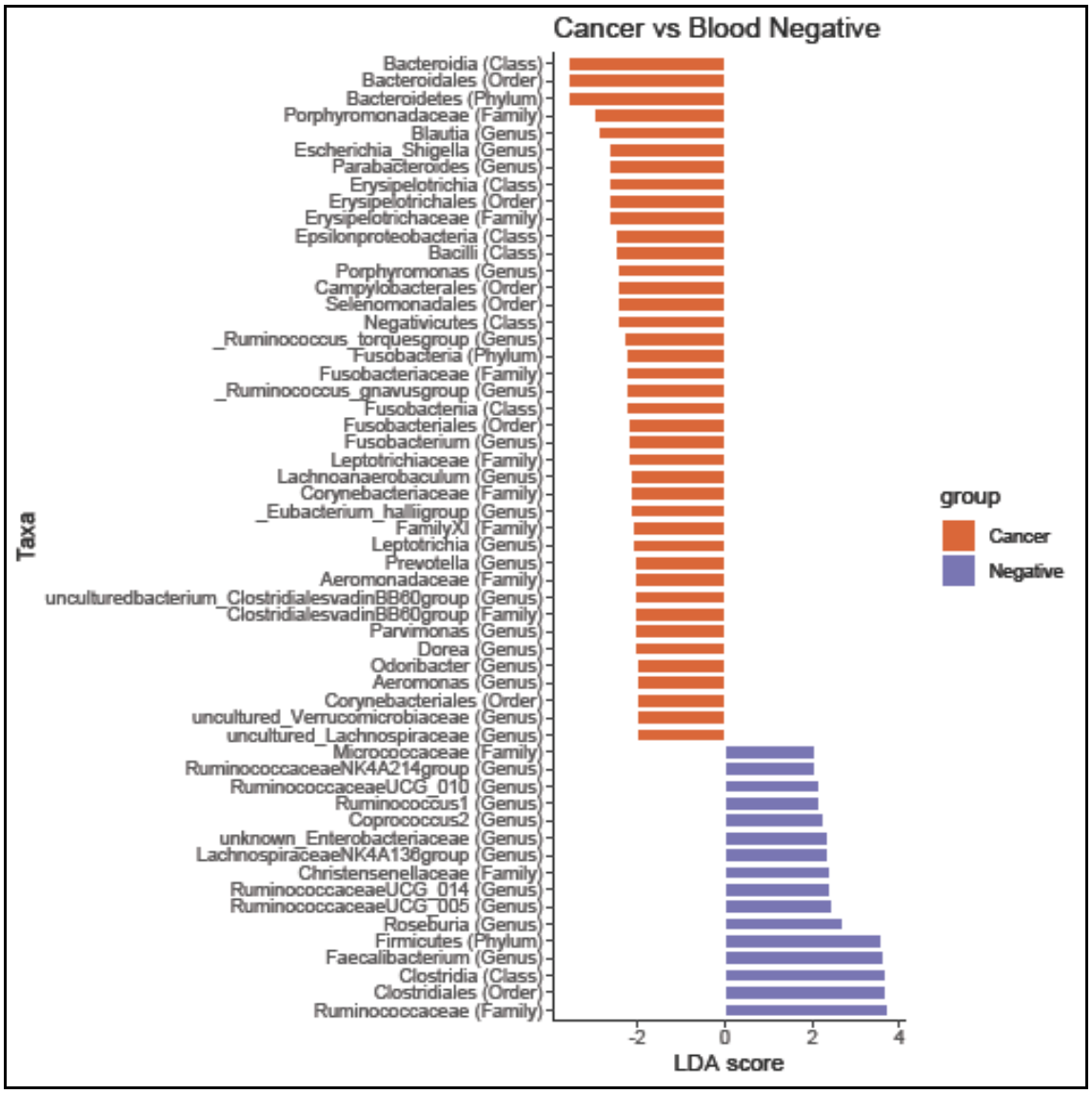
Table 24. Genera enriched/depleted in 'colonoscopy-normal' compared with 'blood-negative' samples. Taxa which have been identified in existing studies of dietary haem or iron supplementation are shaded grey. Taxa which are consistent with the results of these studies are marked (+); those that conflict with the results of these studies are marked (-); those for which the literature is inconsistent are marked (+/-) (598-602).

Genera enriched in colonoscopy-normal compared with blood-negative	Genera enriched in blood-negative compared with colonoscopy-normal
Escherichia-Shigella (+)	*Faecalibacterium*
Streptococcus (-)	*Alistipes*
Lactobacillus (-)	*ChristensenellaceaeR_7 group*
Unknown *Enterobacteriaceae* (+)	*RuminococcaceaeUCG_002*
Blautia (-)	*RuminococcaceaeUCG_014*
Fusobacterium	Uncultured *Rhodospirillaceae*
Enterococcus (+)	*LachnospiraceaeNK4A136 group* (+)
Akkermansia	Unknown *Lachnospiraceae* (+)
Ruminococcus_gnavus group	*Subdoligranulum*
Megasphaera	*RuminococcaceaeUCG_005*
Ruminococcus_torques group	*Ruminococcus1*
Veillonella	*Roseburia* (+/-)
Proteus	*Coprococcus2*
Acidaminococcus	*RikenellaceaeRC9gut group*
	Uncultured bacterium *ClostridialesvadinBB60 group*
	Eubacterium_coprostanoligenes group
	Ruminococcus2
	Butyrivibrio
	RuminococcaceaeUCG_010 (+)
	Ruminiclostridium6 (+)
	Uncultured *Ruminococcaceae* (+)

Genera enriched in colonoscopy-normal compared with blood-negative	Genera enriched in blood-negative compared with colonoscopy-normal
	Barnesiella
	Unknown *Mollicutes*
	Eubacterium_ruminantium group
	Unknown *NB1_n*
	Uncultured bacterium *MollicutesRF9*
	LachnospiraceaeND3007 group
	PrevotellaceaeNK3B31 group
	RuminococcaceaeNK4A214 group (+)
	RuminococcaceaeUCG_003 (+)
	LachnospiraceaeUCG_003
	Uncultured bacterium *Gastranaerophilales*
	Lachnospira
	Eubacterium_xylanophilum group

3.5.4.4 CRC compared with blood-negative and colonoscopy-normal samples

The aforementioned taxonomic differences between blood-negative and colonoscopy-normal groups may account for different taxa being enriched in CRC compared with blood-negative or colonoscopy-normal samples (Figure 92 and Figure 93). Phylogenetic differences between the two pairs of groups can be appreciated by comparing the cladograms. Of the taxa enriched in CRC only two feature in both comparisons (*Odoribacter* and *Porphyromonas*). Of the taxa depleted in CRC, only one features in both comparisons (unknown *Enterobacteriaceae*). Nine taxa show an inverse association with CRC between the two comparisons: *Escherichia-Shigella, Ruminococcus_gnavus group, Fusobacterium, Faecalibacterium, RuminococcaceaeUCG_014, LachnospiraceaeNK4A136 group, Ruminococcus1, Coprococcus2, and RuminococcaceaeUCG_005.*


Cancer vs Blood Negative
Bacteroidia (Class)
Bacteroidales (Order)
Bacteroidetes (Phylum)
Porphyromonadaceae (Family)
Blautia (Genus)
Escherichia_Shigella (Genus)
Parabacteroides (Genus)
Erysipelotrichia (Class)
Erysipelotrichales (Order)
Erysipelotrichaceae (Family)
Epsilonproteobacteria (Class)
Bacilli (Class)
Porphyromonas (Genus)
Campylobacterales (Order)
Selenomonadales (Order)
Negativicutes (Class)
_Ruminococcus_torquesgroup (Genus)
Fusobacteria (Phylum)
Fusobacteriaceae (Family)
_Ruminococcus_gnavusgroup (Genus)
Fusobacteriia (Class)
Fusobacteriales (Order)
Fusobacterium (Genus)
Leptotrichiaceae (Family)
Lachnoanaerobaculum (Genus)
Corynebacteriaceae (Family)
_Eubacterium_halliigroup (Genus)
FamilyXI (Family)
Leptotrichia (Genus)
Prevotella (Genus)
Aeromonadaceae (Family)
unculturedbacterium_ClostridialesvadinBB60group (Genus)
ClostridialesvadinBB60group (Family)
Parvimonas (Genus)
Dorea (Genus)
Odoribacter (Genus)
Aeromonas (Genus)
Corynebacteriales (Order)
uncultured_Verrucomicrobiaceae (Genus)
uncultured_Lachnospiraceae (Genus)
Micrococcaceae (Family)
RuminococcaceaeNK4A214group (Genus)
RuminococcaceaeUCG_010 (Genus)
Ruminococcus1 (Genus)
Coprococcus2 (Genus)
unknown_Enterobacteriaceae (Genus)
LachnospiraceaeNK4A136group (Genus)
Christensenellaceae (Family)
RuminococcaceaeUCG_014 (Genus)
RuminococcaceaeUCG_005 (Genus)
Roseburia (Genus)
Firmicutes (Phylum)
Faecalibacterium (Genus)
Clostridia (Class)
Clostridiales (Order)
Ruminococcaceae (Family)
Taxa
-2
0
2
4
LDA score
group
Cancer
Negative

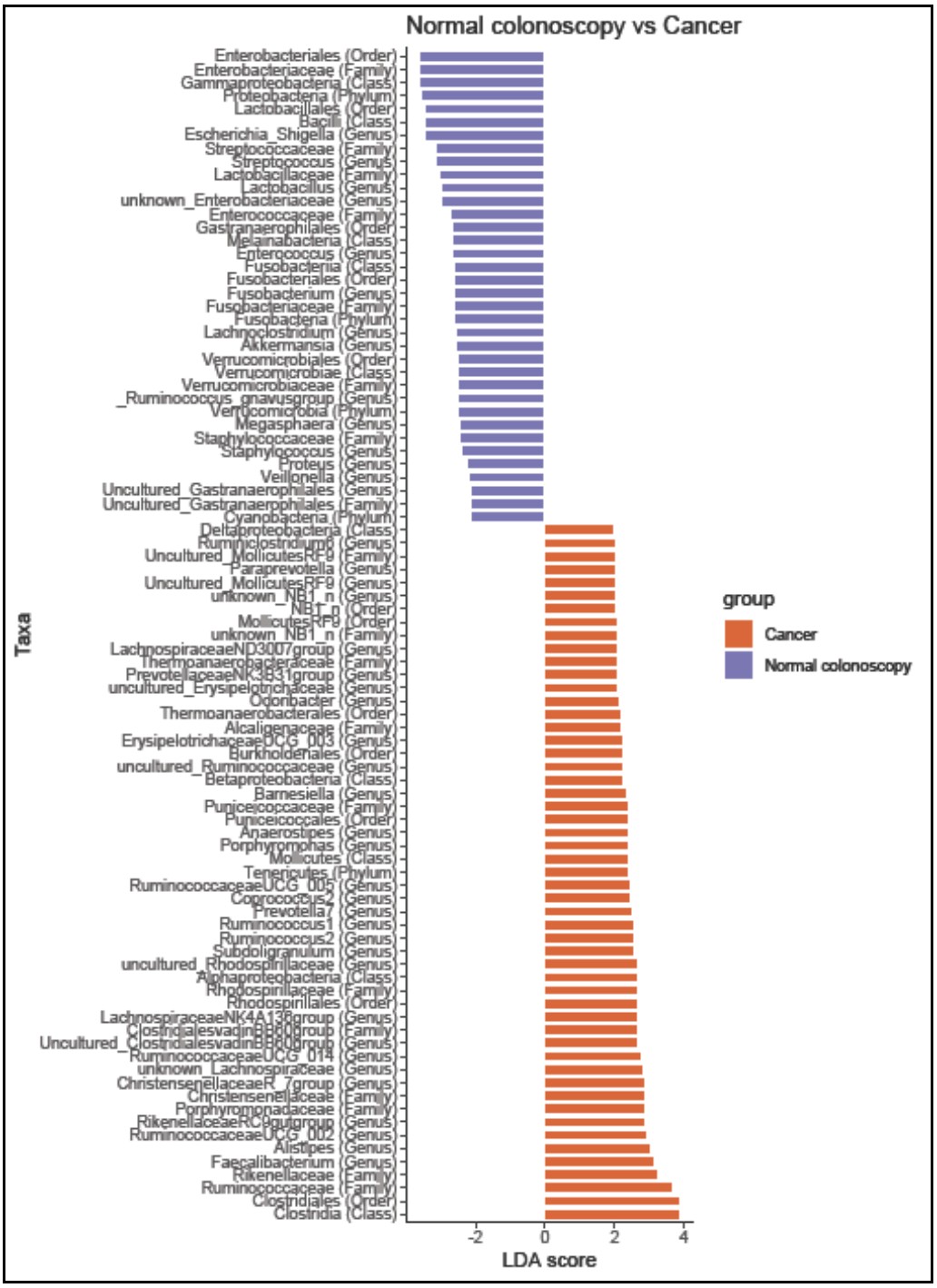


Figure 92. LEfSe plots of NHSBCSP samples (CRC compared with blood-negative and colonoscopy-normal samples). Upper plot: indicates taxa which are significantly enriched in CRC samples (orange) and blood-negative samples (purple), ranked according to effect size. Lower plot: indicates taxa which are significantly enriched in CRC samples (orange) and colonoscopy-normal samples (purple), ranked according to effect size.

The genera from Figure 92 are displayed in Table 25 for clarity:

Table 25. Genera enriched/depleted in CRC compared with 'blood-negative' or 'colonoscopy normal' samples. Taxa which have been identified as being differentially abundant between CRC and controls by meta-analyses of faecal studies are shaded grey (121, 166, 477, 496, 534, 597). Taxa which are consistent with the results of these studies are marked (+); those that conflict with the results of these studies are marked (-).

Genera enriched in CRC compared with blood-negative	Genera depleted in CRC compared with blood-negative
Blautia	*Faecalibacterium*
Escherichia-Shigella (+)	*Roseburia* (+)
Parabacteroides (+)	*RuminococcaceaeUCG_005*
Porphyromonas (+)	*RuminococcaceaeUCG_014*
Ruminococcus_torques group (+)	*LachnospiraceaeNK4A136 group*
Ruminococcus_gnavus group	Unknown *Enterobacteriaceae*
Fusobacterium (+)	*Coprococcus2* (+)
Lachnoanaerobaculum	*Ruminococcus1* (+)
Eubacterium_hallii group	*RuminococcaceaeUCG_010*
Leptotrichia	*RuminococcaceaeNK4A214 group*
Prevotella	
Uncultured bacterium *ClostridialesvadinBB60 group*	
Parvimonas (+)	
Dorea	
Odoribacter	
Aeromonas	
Uncultured *Verrucomicrobiaceae*	
Uncultured *Lachnospiraceae*	

Genera enriched in CRC compared with colonoscopy normal	Genera depleted in CRC compared with colonoscopy normal
Faecalibacterium	*Escherichia-Shigella* (-)
Alistipes (+)	*Streptococcus* (-)
RuminococcaceaeUCG_002	*Lactobacillus*
RikenellaceaeRC9gut group	Unknown *Enterobacteriaceae*
ChristensenellaceaeR_7 group	*Enterococcus*
Unknown *Lachnospiraceae*	*Fusobacterium* (-)
RuminococcaceaeUCG_014	*Lachnoclostridium*
Uncultured *ClostridialesvadinBB60 group*	*Akkermansia*
LachnospiraceaeNK4A136 group	*Ruminococcus_gnavus group*
Uncultured *Rhodospirillaceae*	*Megasphaera*
Subdoligranulum (+)	*Staphylococcus*
Ruminococcus2	*Proteus*
Ruminococcus1	*Veillonella*
Prevotella7 (+)	Uncultured *Gastranaerophilales*
Coprococcus2 (-)	
RuminococcaceaeUCG_005	
Porphyromonas (+)	
Anaerostipes	
Barnesiella	
Uncultured *Ruminococcaceae*	
ErysipelotrichaceaeUCG_003	
Odoribacter	
Uncultured *Erysipelotrichaceae*	
PrevotellaceaeNK3B31 group	
LachnospiraceaeND3007 group	
Unknown *NB1_n*	

Genera enriched in CRC compared with colonoscopy normal	**Genera depleted in CRC compared with colonoscopy normal**
Uncultured *MollicutesRF9*	
Paraprevotella	
Ruminiclostridium6	

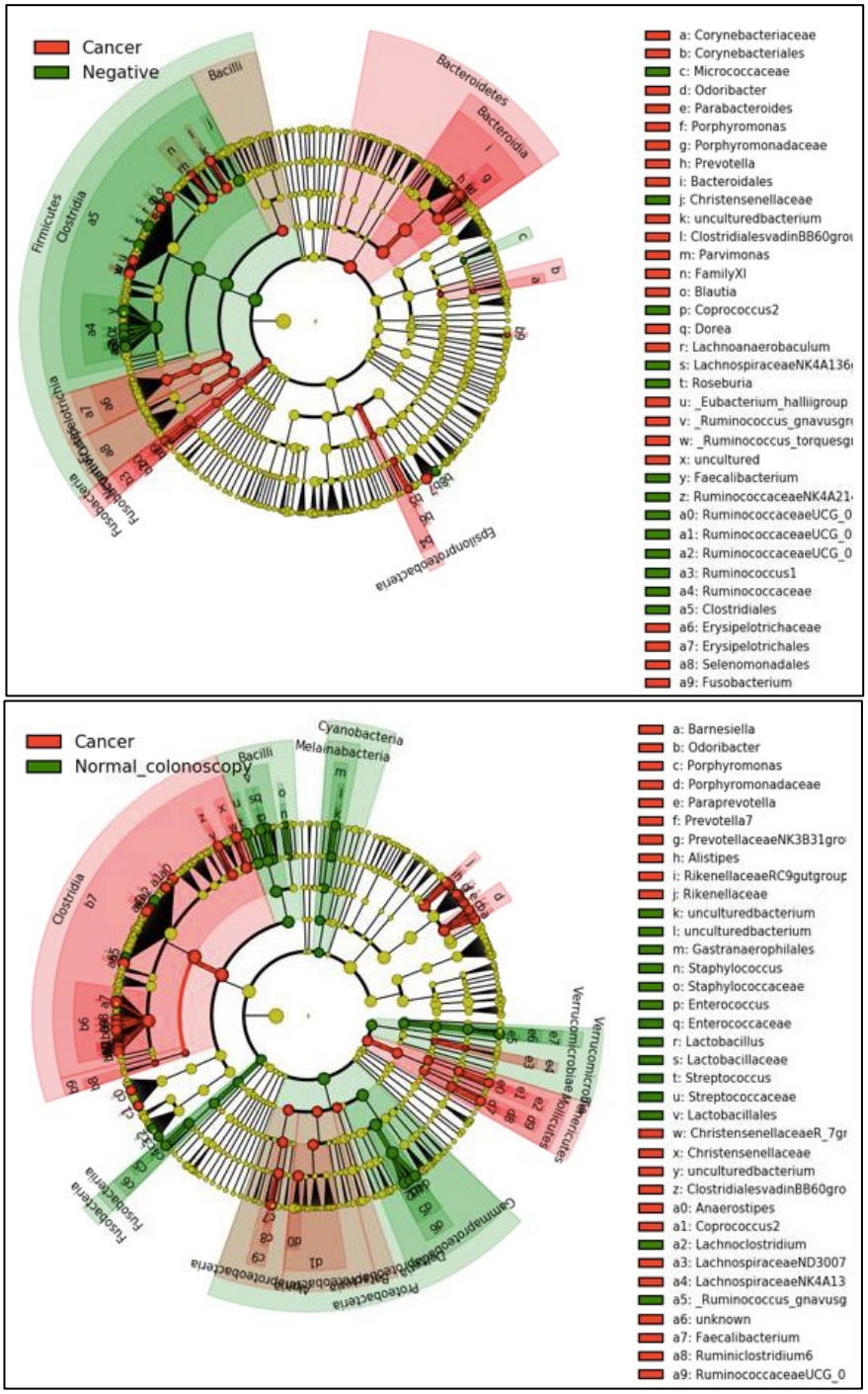
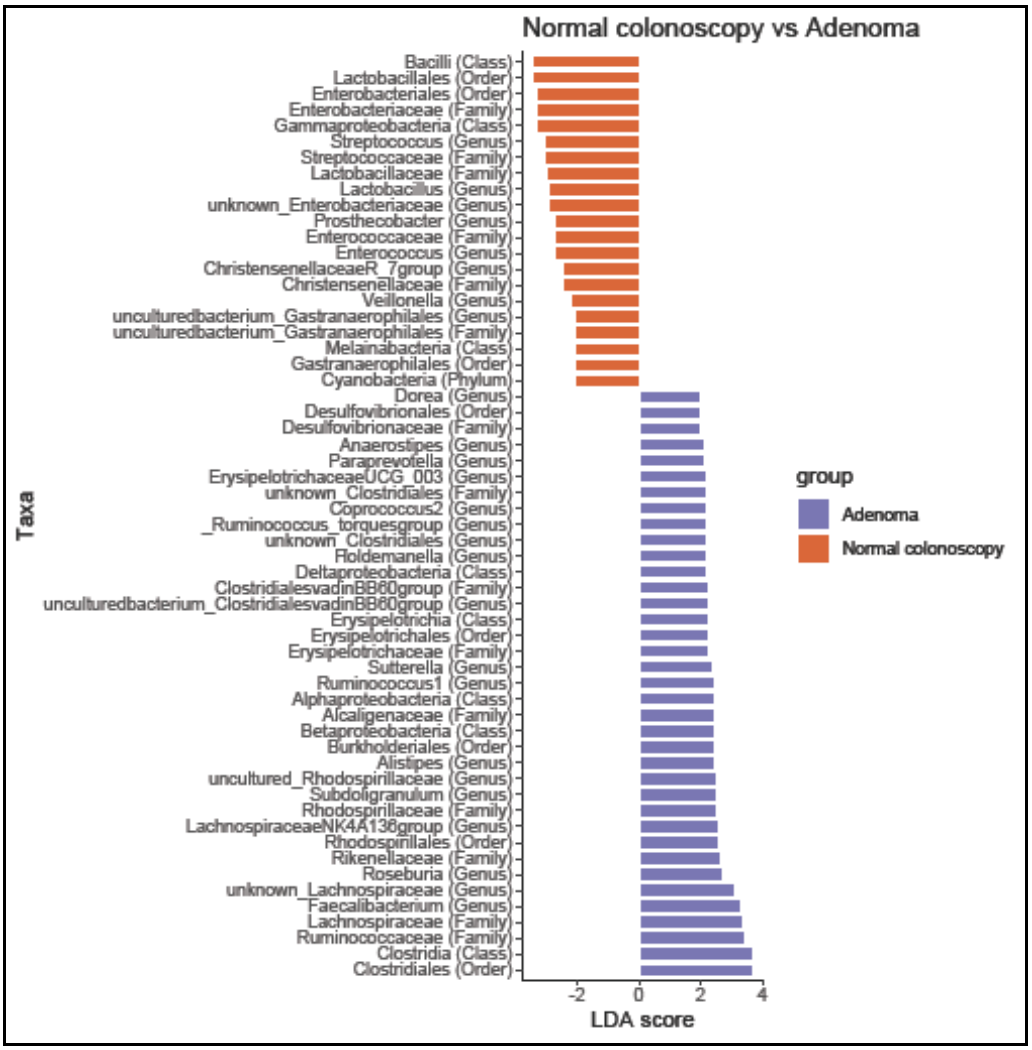


Figure 93. Cladograms of NHSBCSP samples (CRC compared with blood-negative and colonoscopy-normal samples). Cladograms indicate the phylogenetic relationship between taxa. Working outwards, the five rings denote phylum, class, order, family, and genus. Taxa are represented by circles. Circle diameter is proportional to abundance. Upper cladogram: circle colour indicates taxa which are significantly enriched in CRC samples (red) and blood-negative samples (green). Lower cladogram: circle colour indicates taxa which are significantly enriched in CRC samples (red) and colonoscopy-normal samples (green).

3.5.4.5 Adenoma compared with blood-negative and colonoscopy-normal samples

Figure 94 and Figure 95 show taxa which are significantly enriched in adenoma compared with blood-negative and colonoscopy-normal samples. Phylogenetic differences between the two pairs of groups can be appreciated by comparing the cladograms. Of the taxa enriched in adenoma, three feature in both comparisons (unknown *Lachnospiraceae, Ruminococcus_torques group,* and *Dorea*). Of the taxa depleted in adenoma, three feature in both comparisons (unknown *Enterobacteriaceae, ChristensenellaceaeR_7 group, and* uncultured bacterium *Gastranaerophilales*). Eight taxa show an inverse association with adenoma between the two comparisons *(Streptococcus, Faecalibacterium, LachnospiraceaeNK4A136 group,* uncultured *Rhodospirillaceae, Alistipes, Ruminococcus1,* uncultured bacterium *ClostridialesvadinBB60 group,* and *Coprococcus2).*


Normal colonoscopy vs Adenoma
Taxa
Bacilli (Class)
Lactobacillales (Order)
Enterobacteriales (Order)
Enterobacteriaceae (Family)
Gammaproteobacteria (Class)
Streptococcus (Genus)
Streptococcaceae (Family)
Lactobacillaceae (Family)
Lactobacillus (Genus)
unknown_Enterobacteriaceae (Genus)
Prosthecobacter (Genus)
Enterococcaceae (Family)
Enterococcus (Genus)
ChristensenellaceaeR_7group (Genus)
Christensenellaceae (Family)
Veillonella (Genus)
unculturedbacterium_Gastranaerophilales (Genus)
unculturedbacterium_Gastranaerophilales (Family)
Melainabacteria (Class)
Gastranaerophilales (Order)
Cyanobacteria (Phylum)
Dorea (Genus)
Desulfovibrionales (Order)
Desulfovibrionaceae (Family)
Anaerostipes (Genus)
Paraprevotella (Genus)
ErysipelotrichaceaeUCG_003 (Genus)
unknown_Clostridiales (Family)
Coprococcus2 (Genus)
_Ruminococcus_torquesgroup (Genus)
unknown_Clostridiales (Genus)
Roldemanella (Genus)
Deltaproteobacteria (Class)
ClostridialesvadinBB60group (Family)
unculturedbacterium_ClostridialesvadinBB60group (Genus)
Erysipelotrichia (Class)
Erysipelotrichales (Order)
Erysipelotrichaceae (Family)
Sutterella (Genus)
Ruminococcus1 (Genus)
Alphaproteobacteria (Class)
Alcaligenaceae (Family)
Betaproteobacteria (Class)
Burkholderiales (Order)
Alistipes (Genus)
uncultured_Rhodospirillaceae (Genus)
Subdoligranulum (Genus)
Rhodospirillaceae (Family)
LachnospiraceaeNK4A136group (Genus)
Rhodospirillales (Order)
Rikenellaceae (Family)
Roseburia (Genus)
unknown_Lachnospiraceae (Genus)
Faecalibacterium (Genus)
Lachnospiraceae (Family)
Ruminococcaceae (Family)
Clostridia (Class)
Clostridiales (Order)
-2
0
2
4
LDA score
group
Adenoma
Normal colonoscopy

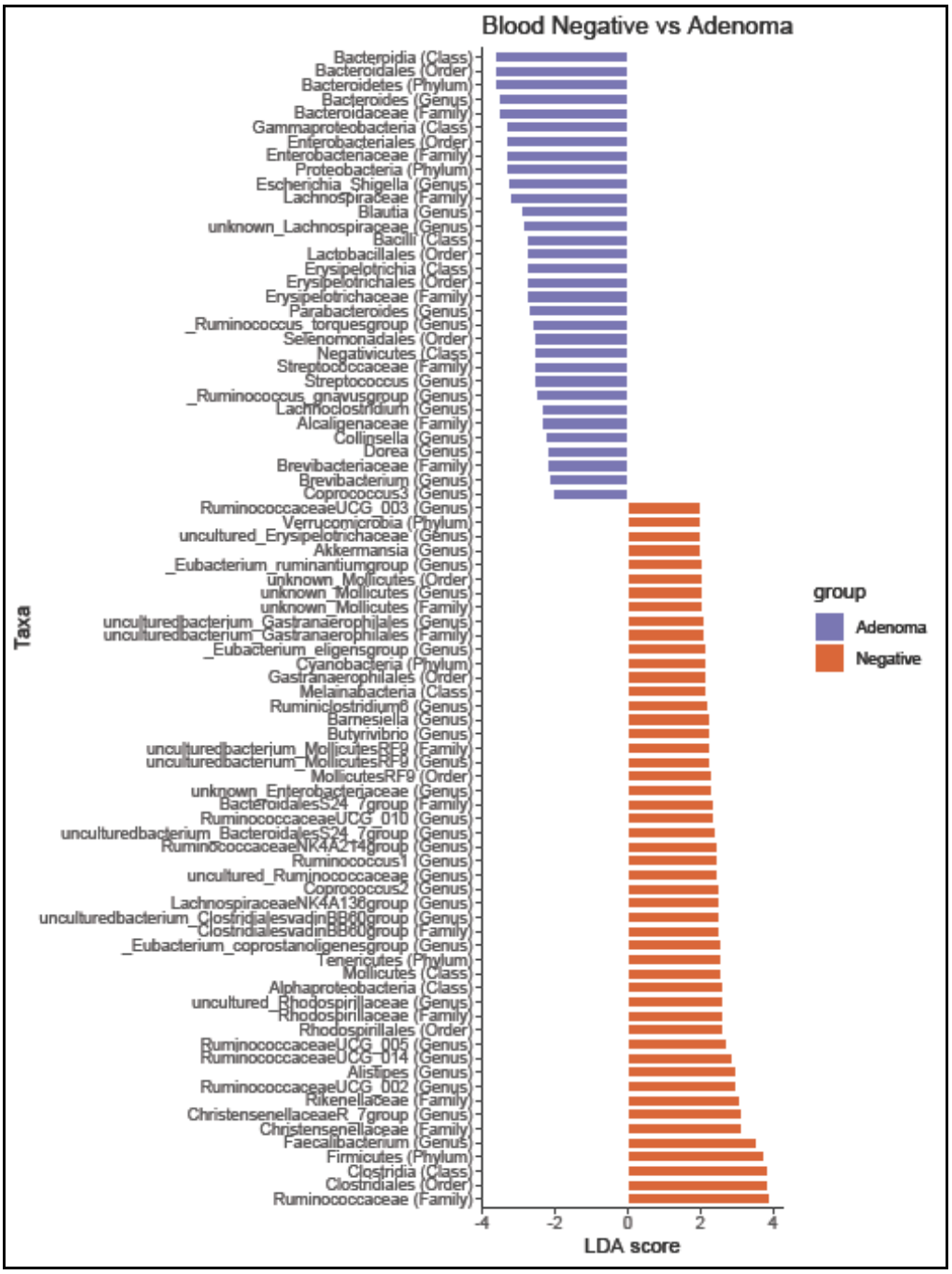


Figure 94. LEfSe plots of NHSBCSP samples (adenoma compared with colonoscopy-normal and blood-negative samples). Upper plot: indicates taxa which are significantly enriched in colonoscopy-normal samples (orange) and adenoma samples (purple), ranked according to effect size. Lower plot: indicates taxa which are significantly enriched in blood-negative samples (orange) and adenoma samples (purple), ranked according to effect size.

The genera from Figure 94 are displayed in Table 26 for clarity:

Table 26. Genera enriched/depleted in adenoma compared with 'colonoscopy-normal' or 'blood-negative' samples. Taxa which have been identified as being differentially abundant between adenoma and controls by meta-analyses of faecal studies are shaded grey (121, 166, 477, 496, 534, 597). Taxa which are consistent with the results of these studies are marked (+); those that conflict with the results of these studies are marked (-).

Genera enriched in adenoma compared with colonoscopy-normal	Genera depleted in adenoma compared with colonoscopy-normal
Faecalibacterium	*Streptococcus*
Unknown *Lachnospiraceae*	*Lactobacillus*
Roseburia	Unknown *Enterobacteriaceae*
LachnospiraceaeNK4A136 group	*Prosthecobacter*
Subdoligranulum	*Enterococcus*
Uncultured *Rhodospirillaceae*	*ChristensenellaceaeR_7 group*
Alistipes	*Veillonella* (-)
Ruminococcus1	Uncultured bacterium *Gastranaerophilales*
Sutterella	
Uncultured bacterium *ClostridialesvadinBB60 group*	
Holdemanella	
Unknown *Clostridiales*	
Ruminococcus_torques group (+)	
Coprococcus2	
ErysipelotrichaceaeUCG_003	
Paraprevotella	
Anaerostipes	
Dorea	

Genera enriched in adenoma compared with blood-negative	Genera depleted in adenoma compared with blood-negative
Bacteroides	*Faecalibacterium*
Escherichia-Shigella	*ChristensenellaceaeR_7 group*
Blautia	*RuminococcaceaeUCG_002*
Unknown *Lachnospiraceae*	*Alistipes*
Parabacteroides	*RuminococcaceaeUCG_014*
Ruminococcus_torques group (+)	*RuminococcaceaeUCG_005*
Streptococcus	Uncultured *Rhodospirillaceae*
Ruminococcus_gnavus group	*Eubacterium_coprostanoligenes group*
Lachnoclostridium	Uncultured bacterium *ClostridialesvadinBB60 group*
Collinsella	*LachnospiraceaeNK4A136 group*
Dorea	*Coprococcus2*
Brevibacterium	Uncultured *Ruminococcaceae*
Coprococcus3 (+)	*Ruminococcus1*
	RuminococcaceaeNK4A214 group
	Uncultured bacterium *BacteroidalesS24_7 group*
	RuminococcaceaeUCG_010
	Unknown *Enterobacteriaceae*
	Uncultured bacterium *MollicutesRF9*
	Butyrivibrio
	Barnesiella (-)
	Ruminiclostridium6
	Eubacterium_eligens group (+)
	Uncultured bacterium *Gastranaerophilales*
	Unknown *Mollicutes*

Genera enriched in adenoma compared with blood-negative	Genera depleted in adenoma compared with blood-negative
	Eubacterium_ruminantium group
	Akkermansia
	Uncultured *Erysipelotrichaceae*
	RuminococcaceaeUCG_003

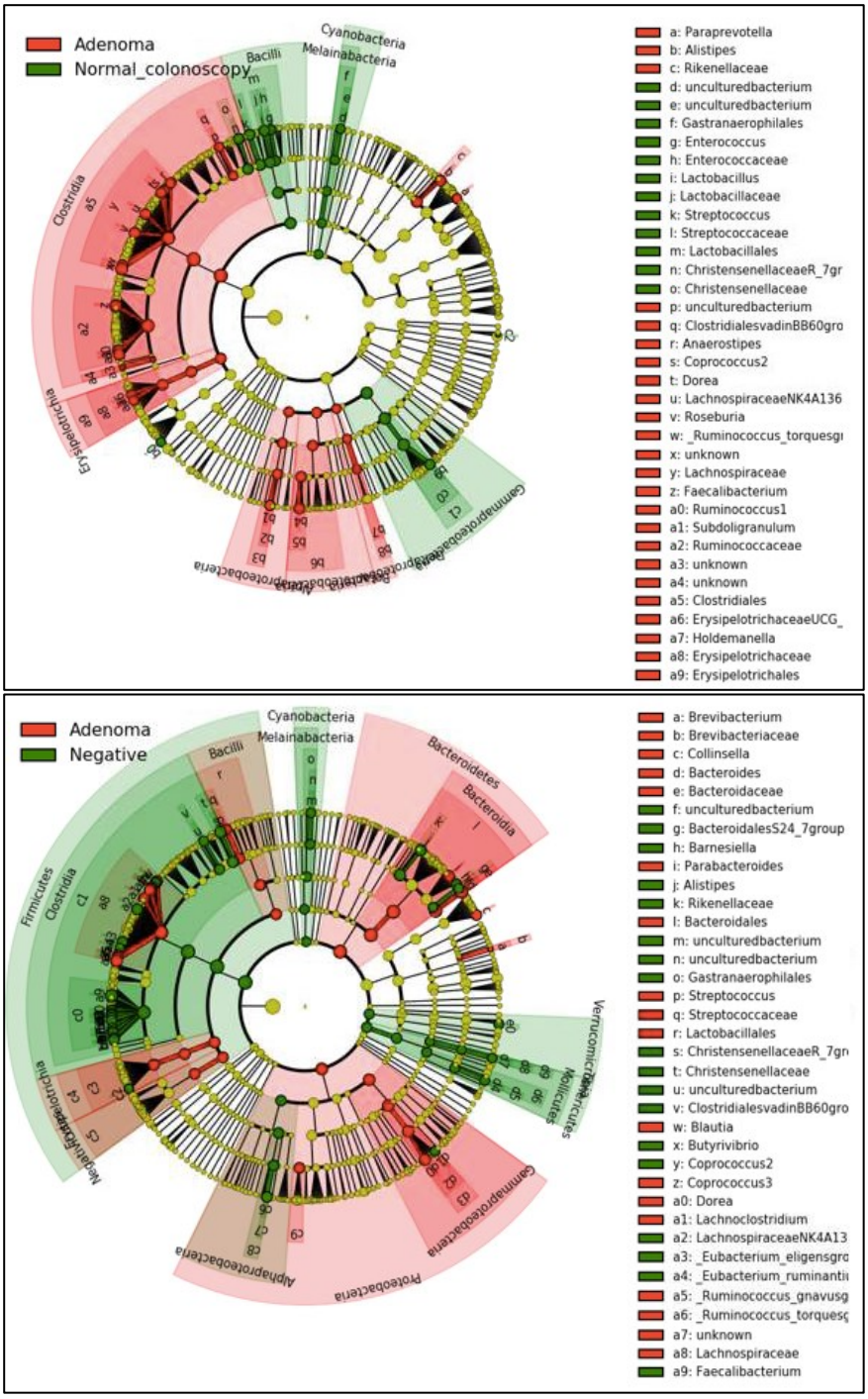
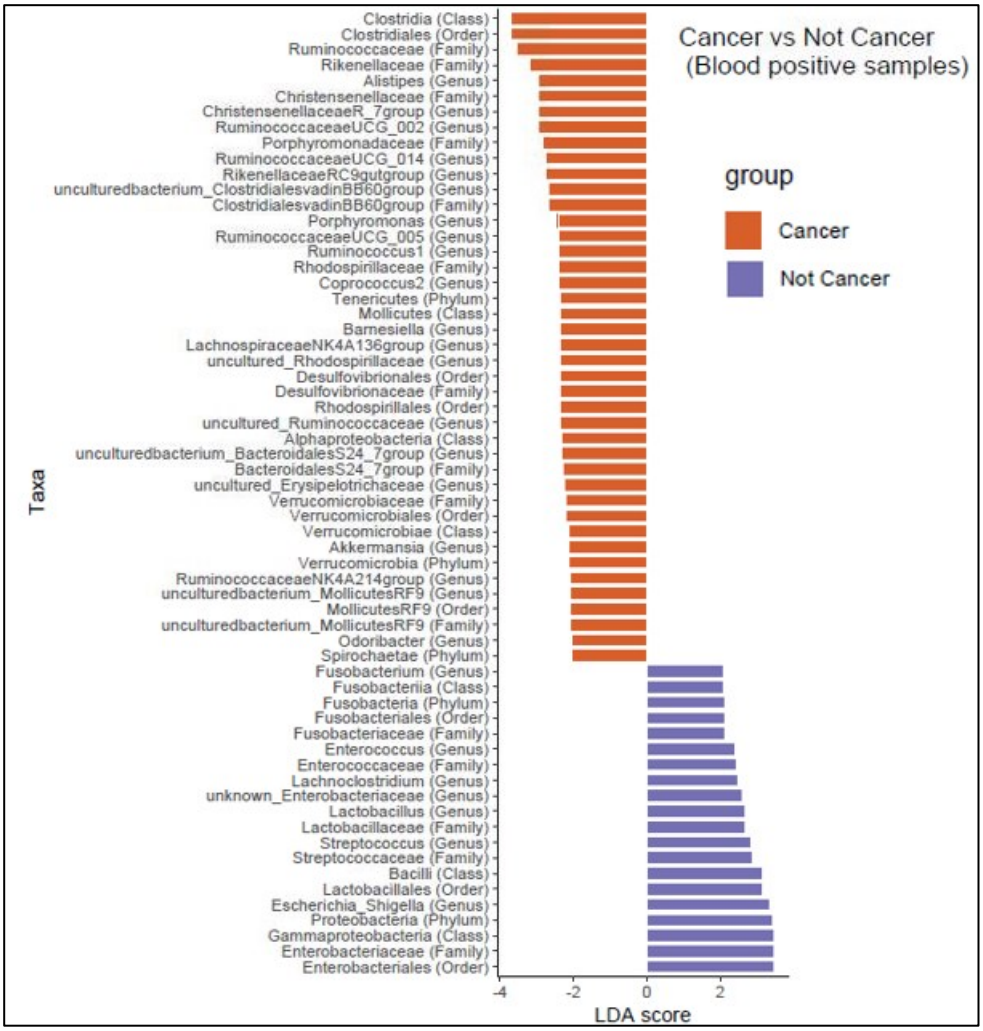


Figure 95. Cladograms of NHSBCSP samples (adenoma compared with colonoscopy-normal and blood-negative samples). Cladograms indicate the phylogenetic relationship between taxa. Working outwards, the five rings denote phylum, class, order, family, and genus. Taxa are represented by circles. Circle diameter is proportional to abundance. Upper cladogram: circle colour indicates taxa which are significantly enriched in adenoma samples (red) and colonoscopy-normal samples (green). Lower cladogram: circle colour indicates taxa which are significantly enriched in adenoma samples (red) and blood-negative samples (green).

3.5.4.6 Comparison of groups within blood-positive samples

LEfSe analysis was performed on pairs of groups within the blood-positive samples only, so that comparisons could be made with the Random Forest models generated for each group (Figure 96 to Figure 98).


Cancer vs Not Cancer
(Blood positive samples)
group
Cancer
Not Cancer
Taxa
Clostridia (Class)
Clostridiales (Order)
Ruminococcaceae (Family)
Rikenellaceae (Family)
Alistipes (Genus)
Christensenellaceae (Family)
ChristensenellaceaeR_7group (Genus)
RuminococcaceaeUCG_002 (Genus)
Porphyromonadaceae (Family)
RuminococcaceaeUCG_014 (Genus)
RikenellaceaeRC9gutgroup (Genus)
unculturedbacterium_ClostridialesvadinBB60group (Genus)
ClostridialesvadinBB60group (Family)
Porphyromonas (Genus)
RuminococcaceaeUCG_005 (Genus)
Ruminococcus1 (Genus)
Rhodospirillaceae (Family)
Coprococcus2 (Genus)
Tenericutes (Phylum)
Mollicutes (Class)
Barnesiella (Genus)
LachnospiraceaeNK4A136group (Genus)
uncultured_Rhodospirillaceae (Genus)
Desulfovibrionales (Order)
Desulfovibrionaceae (Family)
Rhodospirillales (Order)
uncultured_Ruminococcaceae (Genus)
Alphaproteobacteria (Class)
unculturedbacterium_BacteroidalesS24_7group (Genus)
BacteroidalesS24_7group (Family)
uncultured_Erysipelotrichaceae (Genus)
Verrucomicrobiaceae (Family)
Verrucomicrobiales (Order)
Verrucomicrobiae (Class)
Akkermansia (Genus)
Verrucomicrobia (Phylum)
RuminococcaceaeNK4A214group (Genus)
unculturedbacterium_MollicutesRF9 (Genus)
MollicutesRF9 (Order)
unculturedbacterium_MollicutesRF9 (Family)
Odoribacter (Genus)
Spirochaetae (Phylum)
Fusobacterium (Genus)
Fusobacteriia (Class)
Fusobacteria (Phylum)
Fusobacteriales (Order)
Fusobacteriaceae (Family)
Enterococcus (Genus)
Enterococcaceae (Family)
Lachnoclostridium (Genus)
unknown_Enterobacteriaceae (Genus)
Lactobacillus (Genus)
Lactobacillaceae (Family)
Streptococcus (Genus)
Streptococcaceae (Family)
Bacilli (Class)
Lactobacillales (Order)
Escherichia_Shigella (Genus)
Proteobacteria (Phylum)
Gammaproteobacteria (Class)
Enterobacteriaceae (Family)
Enterobacteriales (Order)
-4
-2
0
2
LDA score

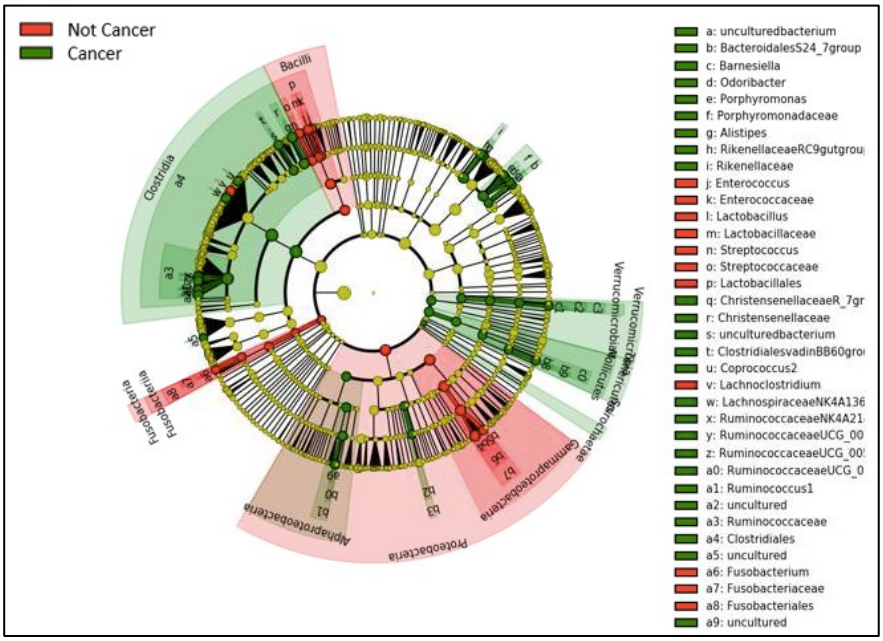
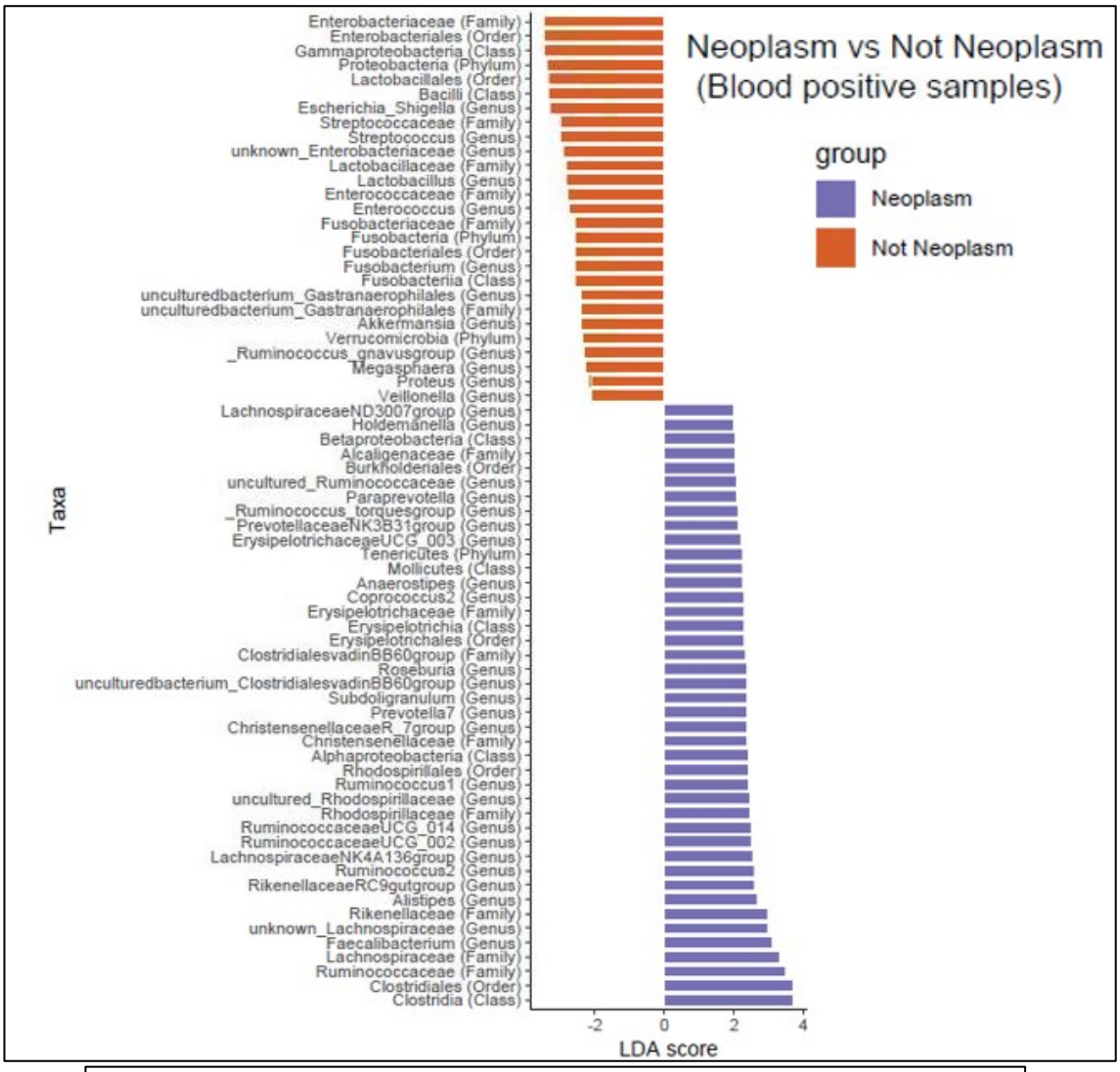

Not Cancer
Cancer
a: unculturedbacterium
b: BacteroidalesS24_7group
c: Barnesiella
d: Odoribacter
e: Porphyromonas
f: Porphyromonadaceae
g: Alistipes
h: RikenellaceaeRC9gutgrou
i: Rikenellaceae
j: Enterococcus
k: Enterococcaceae
l: Lactobacillus
m: Lactobacillaceae
n: Streptococcus
o: Streptococcaceae
p: Lactobacillales
q: ChristensenellaceaeR_7gr
r: Christensenellaceae
s: unculturedbacterium
t: ClostridialesvadinBB60gro
u: Coprococcus2
v: Lachnoclostridium
w: LachnospiraceaeNK4A136
x: RuminococcaceaeNK4A21
y: RuminococcaceaeUCG_00
z: RuminococcaceaeUCG_00
a0: RuminococcaceaeUCG_0
a1: Ruminococcus1
a2: uncultured
a3: Ruminococcaceae
a4: Clostridiales
a5: uncultured
a6: Fusobacterium
a7: Fusobacteriaceae
a8: Fusobacteriales
a9: uncultured
Bacilli
Clostridia
Verrucomicrobiae
Gammaproteobacteria
Proteobacteria
Alphaproteobacteria
Fusobacteriia

Figure 96. LEfSe plot and cladogram of blood-positive NHSBCSP samples (CRC compared with non-CRC). LEfSe plot indicates taxa which are significantly enriched in CRC samples (orange) and non-CRC samples (purple), ranked according to effect size. The cladogram indicates the phylogenetic relationship between taxa. Working outwards, the five rings denote phylum, class, order, family, and genus. Taxa are represented by circles. Circle diameter is proportional to abundance. Circle colour indicates taxa which are significantly enriched in non-CRC samples (red) and CRC samples (green).

The genera from Figure 96 are displayed in Table 27 for clarity.

Table 27. Genera enriched/depleted in CRC compared with 'not CRC' blood-positive samples.

Genera enriched in CRC compared with not CRC (blood-positive samples)	Genera depleted in CRC compared with not CRC (blood-positive samples)
Alistipes	*Escherichia-Shigella*
ChristensenellaceaeR_7 group	*Streptococcus*
RuminococcaceaeUCG_002	*Lactobacillus*
RuminococcaceaeUCG_014	Unknown *Enterobacteriaceae*
RikenellaceaeRC9gut group	*Lachnoclostridium*
Uncultured bacterium *ClostridialesvadinBB60 group*	*Enterococcus*
Porphyromonas	*Fusobacterium*
RuminococcaceaeUCG_005	
Ruminococcus1	
Coprococcus2	
Barnesiella	
LachnospiraceaeNK4A136 group	
Uncultured *Rhodospirillaceae*	
Uncultured *Ruminococcaceae*	
Uncultured bacterium *BacteroidalesS24_7 group*	
Uncultured *Erysipelotrichaceae*	
Akkermansia	
RuminococcaceaeNK4A214 group	
Uncultured bacterium *MollicutesRF9*	
Odoribacter	


Neoplasm vs Not Neoplasm
(Blood positive samples)
group
Neoplasm
Not Neoplasm
Taxa
Enterobacteriaceae (Family)
Enterobacteriales (Order)
Gammaproteobacteria (Class)
Proteobacteria (Phylum)
Lactobacillales (Order)
Bacilli (Class)
Escherichia_Shigella (Genus)
Streptococcaceae (Family)
Streptococcus (Genus)
unknown_Enterobacteriaceae (Genus)
Lactobacillaceae (Family)
Lactobacillus (Genus)
Enterococcaceae (Family)
Enterococcus (Genus)
Fusobacteriaceae (Family)
Fusobacteria (Phylum)
Fusobacteriales (Order)
Fusobacterium (Genus)
Fusobacteriia (Class)
unculturedbacterium_Gastranaerophilales (Genus)
unculturedbacterium_Gastranaerophilales (Family)
Akkermansia (Genus)
Verrucomicrobia (Phylum)
_Ruminococcus_gnavusgroup (Genus)
Megasphaera (Genus)
Proteus (Genus)
Veillonella (Genus)
LachnospiraceaeND3007group (Genus)
Holdemanella (Genus)
Betaproteobacteria (Class)
Alcaligenaceae (Family)
Burkholderiales (Order)
uncultured_Ruminococcaceae (Genus)
Paraprevotella (Genus)
_Ruminococcus_torquesgroup (Genus)
PrevotellaceaeNK3B31group (Genus)
ErysipelotrichaceaeUCG_003 (Genus)
Tenericutes (Phylum)
Mollicutes (Class)
Anaerostipes (Genus)
Coprococcus2 (Genus)
Erysipelotrichaceae (Family)
Erysipelotrichia (Class)
Erysipelotrichales (Order)
ClostridialesvadinBB60group (Family)
Roseburia (Genus)
unculturedbacterium_ClostridialesvadinBB60group (Genus)
Subdoligranulum (Genus)
Prevotella7 (Genus)
ChristensenellaceaeR_7group (Genus)
Christensenellaceae (Family)
Alphaproteobacteria (Class)
Rhodospirillales (Order)
Ruminococcus1 (Genus)
uncultured_Rhodospirillaceae (Genus)
Rhodospirillaceae (Family)
RuminococcaceaeUCG_014 (Genus)
RuminococcaceaeUCG_002 (Genus)
LachnospiraceaeNK4A136group (Genus)
Ruminococcus2 (Genus)
RikenellaceaeRC9gutgroup (Genus)
Alistipes (Genus)
Rikenellaceae (Family)
unknown_Lachnospiraceae (Genus)
Faecalibacterium (Genus)
Lachnospiraceae (Family)
Ruminococcaceae (Family)
Clostridiales (Order)
Clostridia (Class)
-2
0
2
4
LDA score

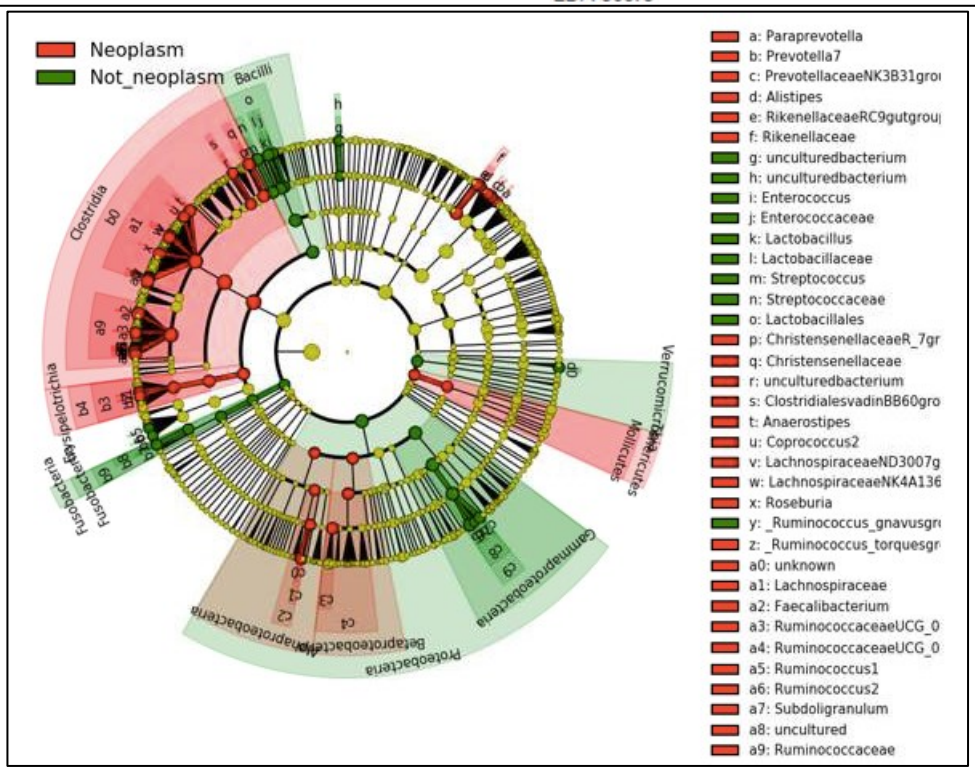
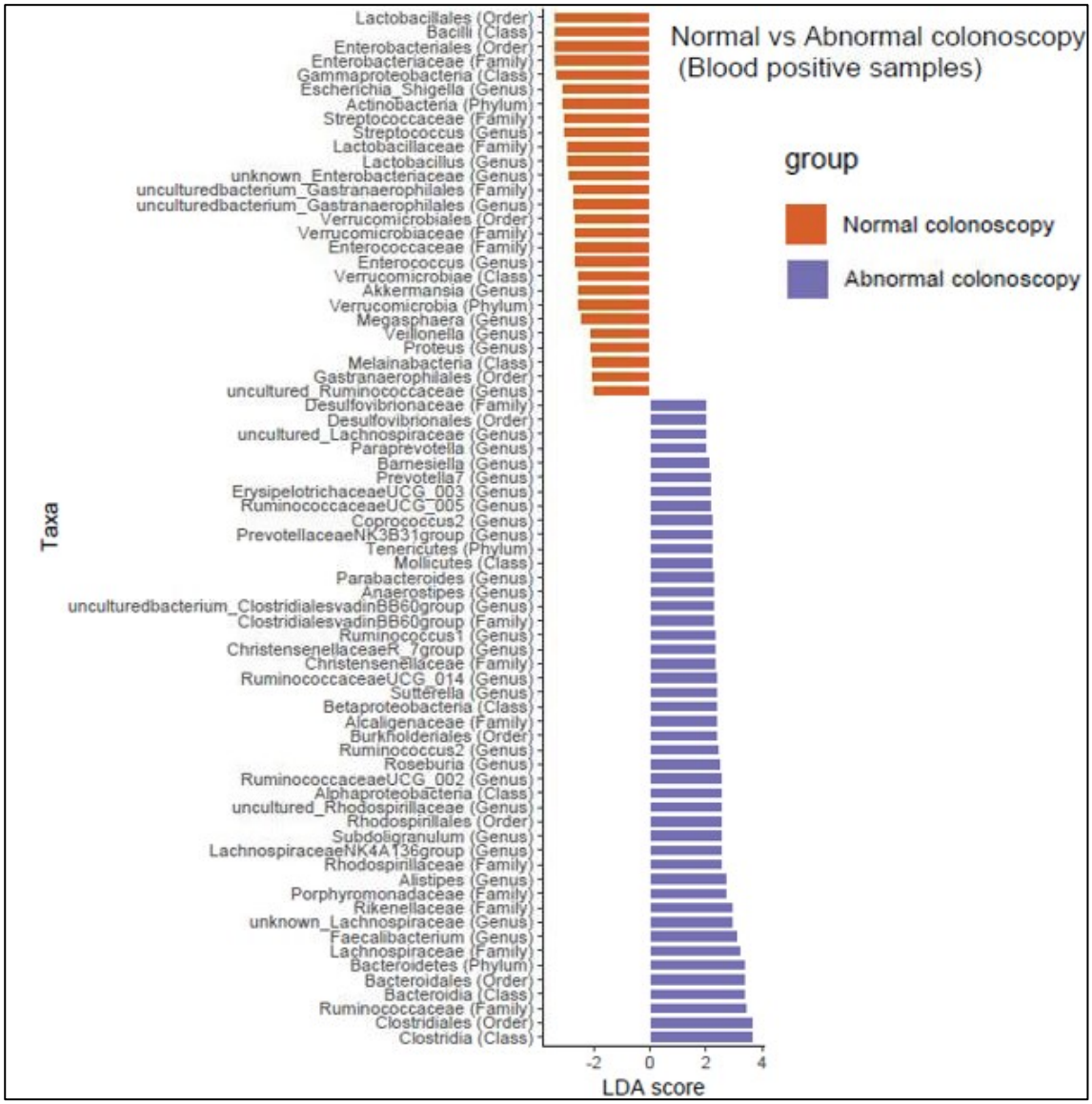

Neoplasm
Not_neoplasm
Bacilli
Clostridia
Verrucomicrobiae
Tenericutes
Mollicutes
Fusobacteria
Gammaproteobacteria
Betaproteobacteria
Proteobacteria
a: Paraprevotella
b: Prevotella7
c: PrevotellaceaeNK3B31gro
d: Alistipes
e: RikenellaceaeRC9gutgrou
f: Rikenellaceae
g: unculturedbacterium
h: unculturedbacterium
i: Enterococcus
j: Enterococcaceae
k: Lactobacillus
l: Lactobacillaceae
m: Streptococcus
n: Streptococcaceae
o: Lactobacillales
p: ChristensenellaceaeR_7gr
q: Christensenellaceae
r: unculturedbacterium
s: ClostridialesvadinBB60gro
t: Anaerostipes
u: Coprococcus2
v: LachnospiraceaeND3007g
w: LachnospiraceaeNK4A136
x: Roseburia
y: _Ruminococcus_gnavusgr
z: _Ruminococcus_torquesgr
a0: unknown
a1: Lachnospiraceae
a2: Faecalibacterium
a3: RuminococcaceaeUCG_0
a4: RuminococcaceaeUCG_0
a5: Ruminococcus1
a6: Ruminococcus2
a7: Subdoligranulum
a8: uncultured
a9: Ruminococcaceae

Figure 97. LEfSe plot and cladogram of blood-positive NHSBCSP samples (neoplasm compared with non-neoplasm). LEfSe plot indicates taxa which are significantly enriched in non-neoplasm samples (orange) and neoplasm samples (purple), ranked according to effect size. The cladogram indicates the phylogenetic relationship between taxa. Working outwards, the five rings denote phylum, class, order, family, and genus. Taxa are represented by circles. Circle diameter is proportional to abundance. Circle colour indicates taxa which are significantly enriched in neoplasm samples (red) and non-neoplasm samples (green).

The genera from Figure 97 are displayed in Table 28 for clarity.

Table 28. **Genera enriched/depleted in neoplasm compared with 'not neoplasm' blood-positive samples.**

Genera enriched in neoplasm compared with not neoplasm (blood-positive samples)	Genera depleted in neoplasm compared with not neoplasm (blood-positive samples)
Faecalibacterium	*Escherichia-Shigella*
Unknown *Lachnospiraceae*	*Streptococcus*
Alistipes	Unknown *Enterobacteriaceae*
RikenellaceaeRC9gut group	*Lactobacillus*
Ruminococcus2	*Enterococcus*
LachnospiraceaeNK4A136 group	*Fusobacterium*
RuminococcaceaeUCG_002	Uncultured bacterium *Gastranaerophilales*
RuminococcaceaeUCG_014	*Akkermansia*
Uncultured *Rhodospirillaceae*	*Ruminococcus_gnavus group*
Ruminococcus1	*Megasphaera*
ChristensenellaceaeR_7 group	*Proteus*
Prevotella7	*Veillonella*
Subdoligranulum	
Uncultured bacterium *ClostridialesvadinBB60 group*	
Roseburia	
Coprococcus2	
Anaerostipes	
ErysipelotrichaceaeUCG_003	
PrevotellaceaeNK3B31 group	
Ruminococcus_torques group	
Paraprevotella	
Uncultured *Ruminococcaceae*	
Holdemanella	
LachnospiraceaeND3007 group	


Normal vs Abnormal colonoscopy
(Blood positive samples)
group
Normal colonoscopy
Abnormal colonoscopy
Taxa
Lactobacillales (Order)
Bacilli (Class)
Enterobacteriales (Order)
Enterobacteriaceae (Family)
Gammaproteobacteria (Class)
Escherichia_Shigella (Genus)
Actinobacteria (Phylum)
Streptococcaceae (Family)
Streptococcus (Genus)
Lactobacillaceae (Family)
Lactobacillus (Genus)
unknown_Enterobacteriaceae (Genus)
unculturedbacterium_Gastranaerophilales (Family)
unculturedbacterium_Gastranaerophilales (Genus)
Verrucomicrobiales (Order)
Verrucomicrobiaceae (Family)
Enterococcaceae (Family)
Enterococcus (Genus)
Verrucomicrobiae (Class)
Akkermansia (Genus)
Verrucomicrobia (Phylum)
Megasphaera (Genus)
Veillonella (Genus)
Proteus (Genus)
Melainabacteria (Class)
Gastranaerophilales (Order)
uncultured_Ruminococcaceae (Genus)
Desulfovibrionaceae (Family)
Desulfovibrionales (Order)
uncultured_Lachnospiraceae (Genus)
Paraprevotella (Genus)
Barnesiella (Genus)
Prevotella7 (Genus)
ErysipelotrichaceaeUCG_003 (Genus)
RuminococcaceaeUCG_005 (Genus)
Coprococcus2 (Genus)
PrevotellaceaeNK3B31group (Genus)
Tenericutes (Phylum)
Mollicutes (Class)
Parabacteroides (Genus)
Anaerostipes (Genus)
unculturedbacterium_ClostridialesvadinBB60group (Genus)
ClostridialesvadinBB60group (Family)
Ruminococcus1 (Genus)
ChristensenellaceaeR_7group (Genus)
Christensenellaceae (Family)
RuminococcaceaeUCG_014 (Genus)
Sutterella (Genus)
Betaproteobacteria (Class)
Alcaligenaceae (Family)
Burkholderiales (Order)
Ruminococcus2 (Genus)
Roseburia (Genus)
RuminococcaceaeUCG_002 (Genus)
Alphaproteobacteria (Class)
uncultured_Rhodospirillaceae (Genus)
Rhodospirillales (Order)
Subdoligranulum (Genus)
LachnospiraceaeNK4A136group (Genus)
Rhodospirillaceae (Family)
Alistipes (Genus)
Porphyromonadaceae (Family)
Rikenellaceae (Family)
unknown_Lachnospiraceae (Genus)
Faecalibacterium (Genus)
Lachnospiraceae (Family)
Bacteroidetes (Phylum)
Bacteroidales (Order)
Bacteroidia (Class)
Ruminococcaceae (Family)
Clostridiales (Order)
Clostridia (Class)
-2
0
2
4
LDA score

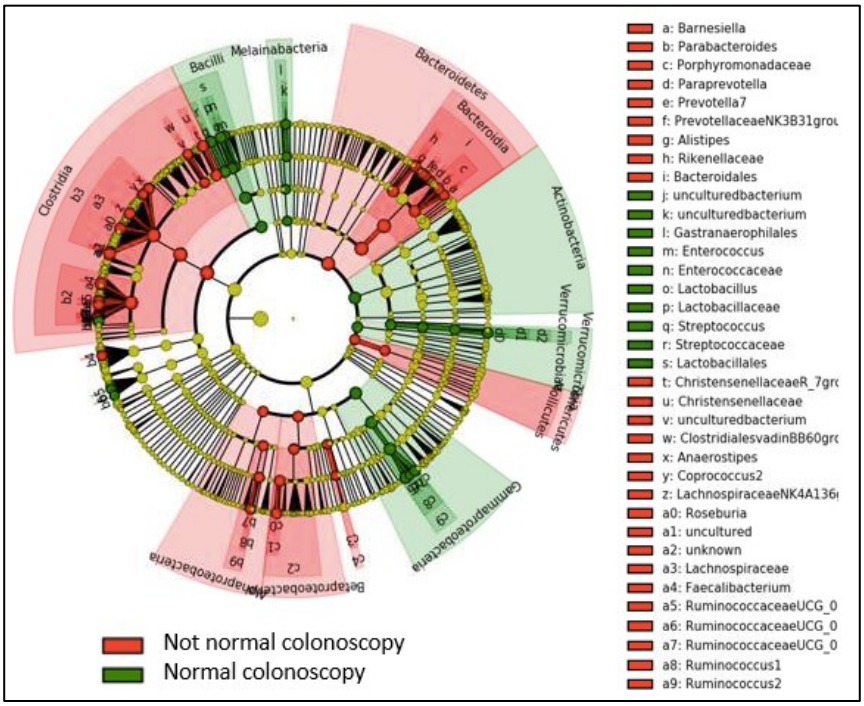

Bacilli
Melainabacteria
Bacteroidetes
Bacteroidia
Clostridia
Actinobacteria
Verrucomicrobiae
Mollicutes
Gammaproteobacteria
Betaproteobacteria
Not normal colonoscopy
Normal colonoscopy
a: Barnesiella
b: Parabacteroides
c: Porphyromonadaceae
d: Paraprevotella
e: Prevotella7
f: PrevotellaceaeNK3B31gro
g: Alistipes
h: Rikenellaceae
i: Bacteroidales
j: unculturedbacterium
k: unculturedbacterium
l: Gastranaerophilales
m: Enterococcus
n: Enterococcaceae
o: Lactobacillus
p: Lactobacillaceae
q: Streptococcus
r: Streptococcaceae
s: Lactobacillales
t: ChristensenellaceaeR_7gr
u: Christensenellaceae
v: unculturedbacterium
w: ClostridialesvadinBB60gr
x: Anaerostipes
y: Coprococcus2
z: LachnospiraceaeNK4A136
a0: Roseburia
a1: uncultured
a2: unknown
a3: Lachnospiraceae
a4: Faecalibacterium
a5: RuminococcaceaeUCG_0
a6: RuminococcaceaeUCG_0
a7: RuminococcaceaeUCG_0
a8: Ruminococcus1
a9: Ruminococcus2

Figure 98. LEfSe plot and cladogram of blood-positive NHSBCSP samples (colonoscopy-normal compared with colonoscopy-abnormal). LEfSe plot indicates taxa which are significantly enriched in colonoscopy-normal samples (orange) and colonoscopy-abnormal samples (purple), ranked according to effect size. The cladogram indicates the phylogenetic relationship between taxa. Working outwards, the five rings denote phylum, class, order, family, and genus. Taxa are represented by circles. Circle diameter is proportional to abundance. Circle colour indicates taxa which are significantly enriched in colonoscopy-abnormal samples (red) and colonoscopy-normal samples (green).

The genera from Figure 98 are displayed in Table 29 for clarity.

Table 29. Genera enriched/depleted in 'colonoscopy-normal' compared with 'colonoscopy-abnormal' blood-positive samples.

Genera enriched in colonoscopy-normal compared with colonoscopy-abnormal (blood-positive samples)	Genera depleted in colonoscopy-normal compared with colonoscopy-abnormal (blood-positive samples)
Escherichia-Shigella	*Faecalibacterium*
Streptococcus	Unknown *Lachnospiraceae*
Lactobacillus	*Alistipes*
Unknown *Enterobacteriaceae*	*LachnospiraceaeNK4A136 group*
Uncultured bacterium *Gastranaerophilales*	*Subdoligranulum*
Enterococcus	Uncultured *Rhodospirillaceae*
Akkermansia	*RuminococcaceaeUCG_002*
Megasphaera	*Roseburia*
Veillonella	*Ruminococcus2*
Proteus	*Sutterella*
Uncultured *Ruminococcaceae*	*RuminococcaceaeUCG_014*
	ChristensenellaceaeR_7 group

Genera enriched in colonoscopy-normal compared with colonoscopy-abnormal (blood-positive samples)	Genera depleted in colonoscopy-normal compared with colonoscopy-abnormal (blood-positive samples)
	Ruminococcus1
	Uncultured bacterium *ClostridialesvadinBB60 group*
	Anaerostipes
	Parabacteroides
	PrevotellaceaeNK3B31 group
	Coprococcus2
	RuminococcaceaeUCG_005
	ErysipelotrichaceaeUCG_003
	Prevotella7
	Barnesiella
	Paraprevotella
	Uncultured *Lachnospiraceae*

3.5.5 Random Forest models

Random Forest models were designed to determine whether microbiome data could improve the accuracy of screening. The results from each Random Forest will now be presented.

3.5.5.1 All samples: distinction between CRC and all other sample types

Two taxa-based Random Forest models were designed to distinguish between CRC and all other sample types; the first used genus-level relative abundance alone and the second combined genus-level relative abundance with gFOBT blood-status. These models were compared with a baseline model which used gFOBT blood status, age and gender. The performance of each model is outlined in Table 30, Figure 99 and Figure 100. The optimum model used both genus-level relative abundance and gFOBT blood status with AUC 0.855 (95% CI: 0.832-0.877). Figure 101 and Figure 102 show the top 15 bacteria which contributed to each model. The most important bacteria were *Fusobacterium,*

Peptostreptococcus, Parvimonas, Porphyromonas, Gemella, Odoribacter and Faecalibacterium.

Table 30. Performance of Random Forest models designed to distinguish CRC samples from all other sample types. The performance of three Random Forest models is tabulated: (1) 'Clinical data only' = a model which used gFOBT blood status, age and gender. (2) 'Bacteria only' = a model which used genus-level relative abundance. (3) 'Bacteria and blood' = a model which used gFOBT blood status and genus-level relative abundance.

Clinical data only (blood status, age, gender)			
	Predicted value		
True value	**Cancer (680)**	**No (603)**	**Error**
Cancer (n=256)	203	53	21%
No (n=1027)	477	550	46%
Sensitivity	203/256=79%		
Specificity	550/1027=54%		
Bacteria only			
	Predicted value		
True value	**Cancer (176)**	**No (1107)**	**Error**
Cancer (n=256)	114	142	55%
No (n=1027)	62	965	6%
Sensitivity	114/256=45%		
Specificity	965/1027=94%		
Bacteria and blood			
	Predicted value		
True value	**Cancer (234)**	**No (1049)**	**Error**
Cancer (n=256)	140	116	45%
No (n=1027)	94	933	9%
Sensitivity	140/256=55%		
Specificity	933/1027=91%		

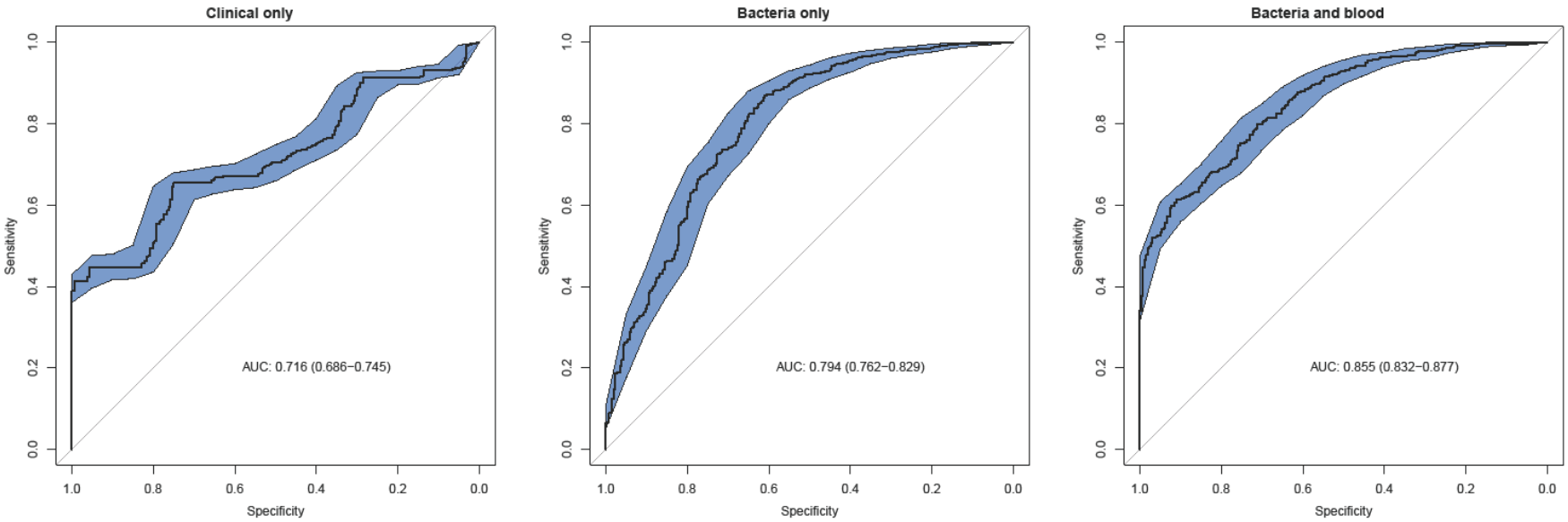


Figure 99. ROC curves of Random Forest models designed to distinguish CRC samples from all other sample types. ROC curves and AUC (with 95% CI) are displayed for three Random Forest models: (left) 'Clinical only' = a model which used gFOBT blood status, age and gender; (middle) 'Bacteria only' = a model which used genus-level relative abundance; (right) 'Bacteria and blood' = a model which used gFOBT blood status and genus-level relative abundance.

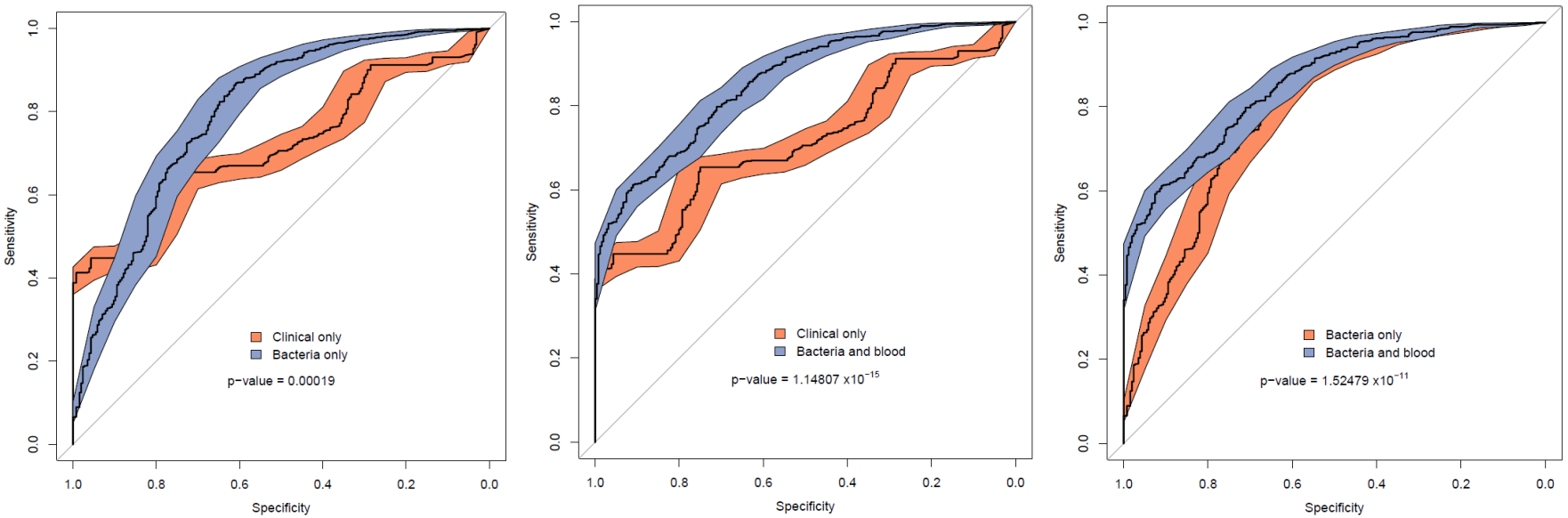


Figure 100. Comparison of ROC curves of Random Forest models designed to distinguish CRC samples from all other sample types. 'Clinical only' = a model which used gFOBT blood status, age and gender; 'Bacteria only' = a model which used genus-level relative abundance; 'Bacteria and blood' = a model which used gFOBT blood status and genus-level relative abundance.

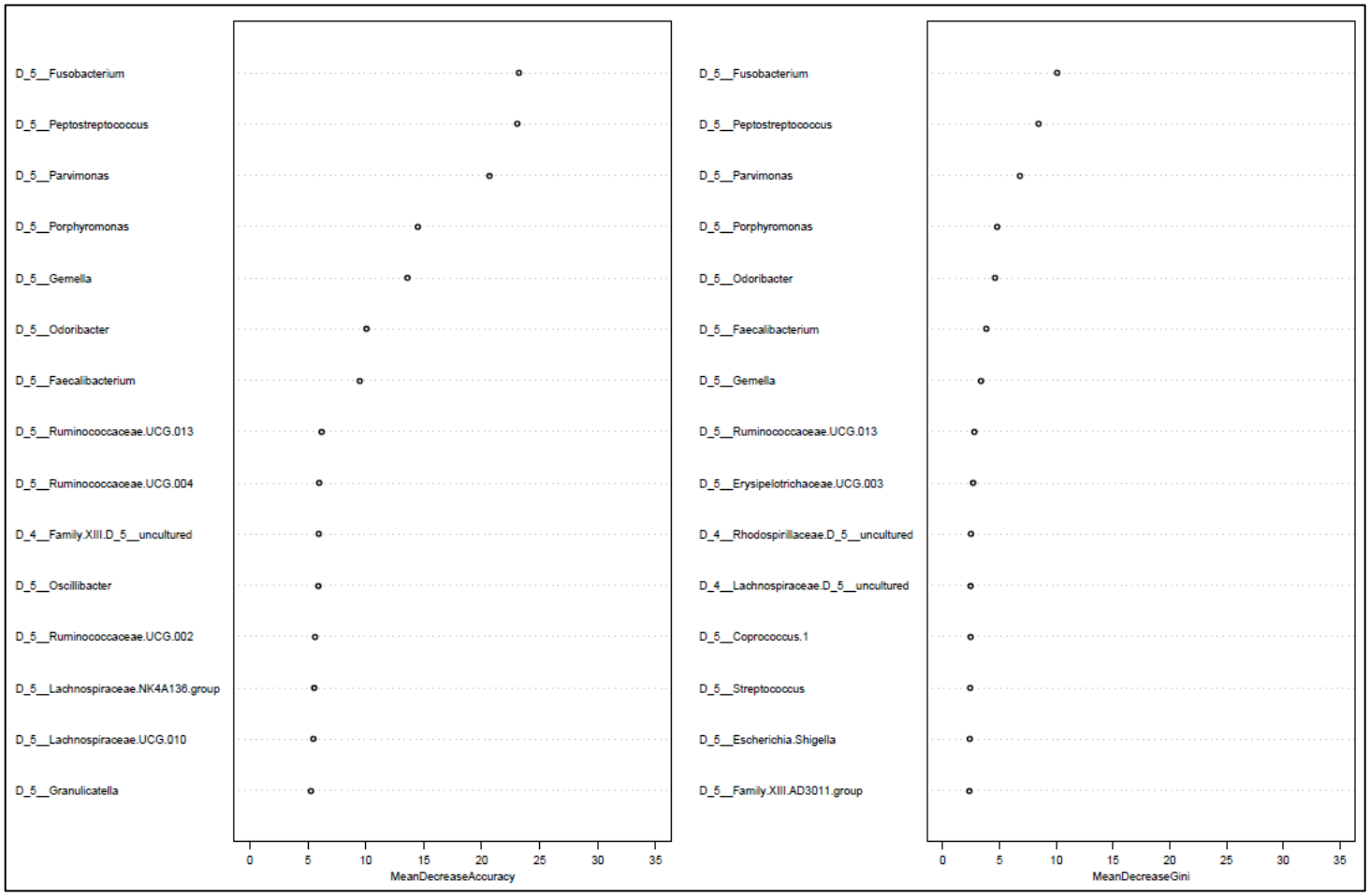


Figure 101. The 15 most important variables in a 'Bacteria only' Random Forest model designed to distinguish CRC samples from all other sample types. 'Bacteria only' = a model which used genus-level relative abundance.

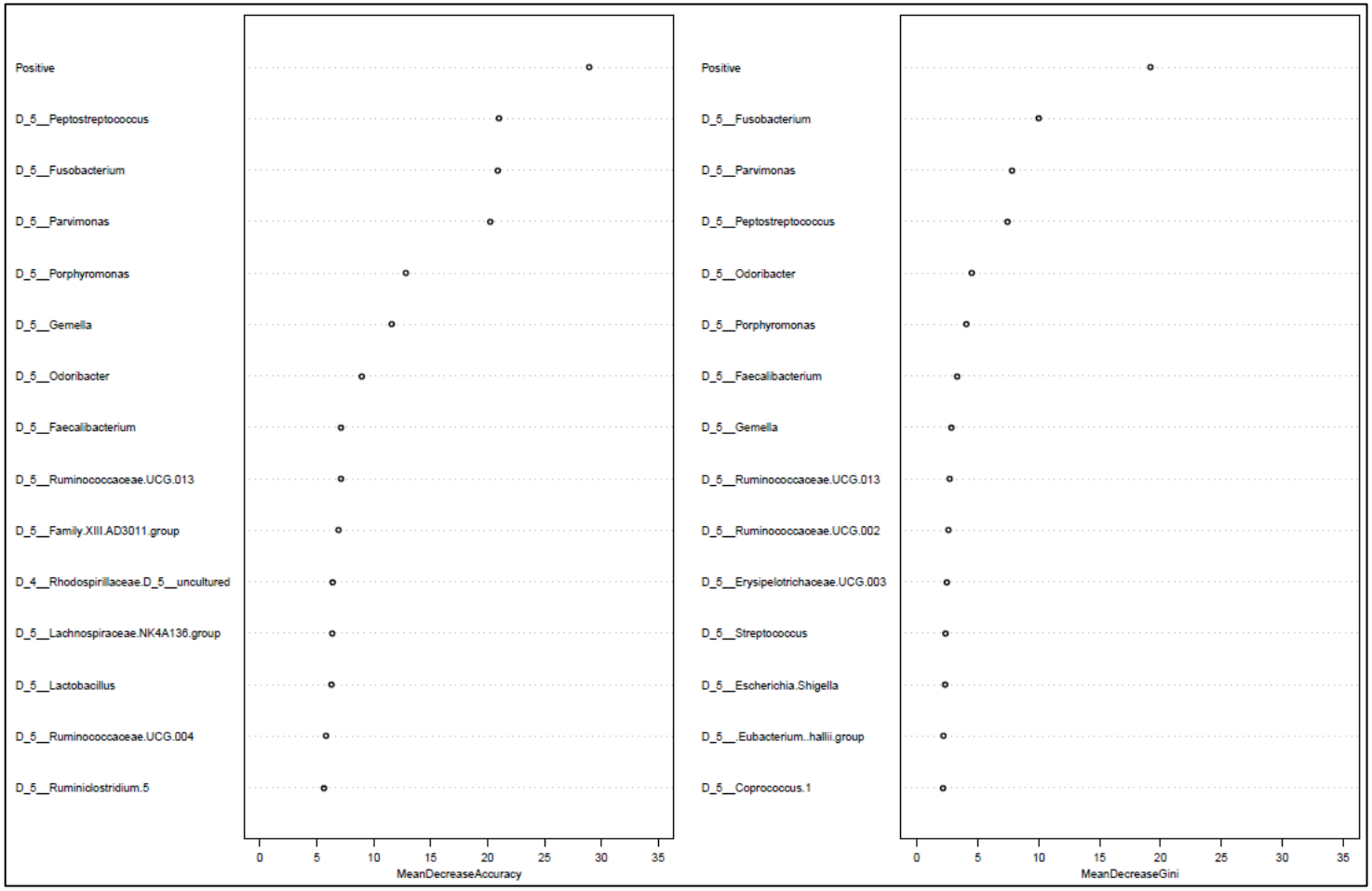


Figure 102. The 15 most important variables in a 'Bacteria and blood' Random Forest model designed to distinguish CRC samples from all other sample types. 'Bacteria and blood' = a model which used gFOBT blood status and genus-level relative abundance. Positive denotes gFOBT blood positivity status.

Partial dependence plots (603) give a crude indication as to how the probability of a class changes as the values of an input variable change. However, they do not account for correlation between variables and the effect on probability is an average only; they are therefore an approximation at best. In the current study, partial dependence plots have been created to give an estimate as to whether the taxa identified by the Random Forest models as important are enriched or depleted in cases compared with controls and the relative abundance at which class probability changes. The results will be confirmed by qPCR in future work.

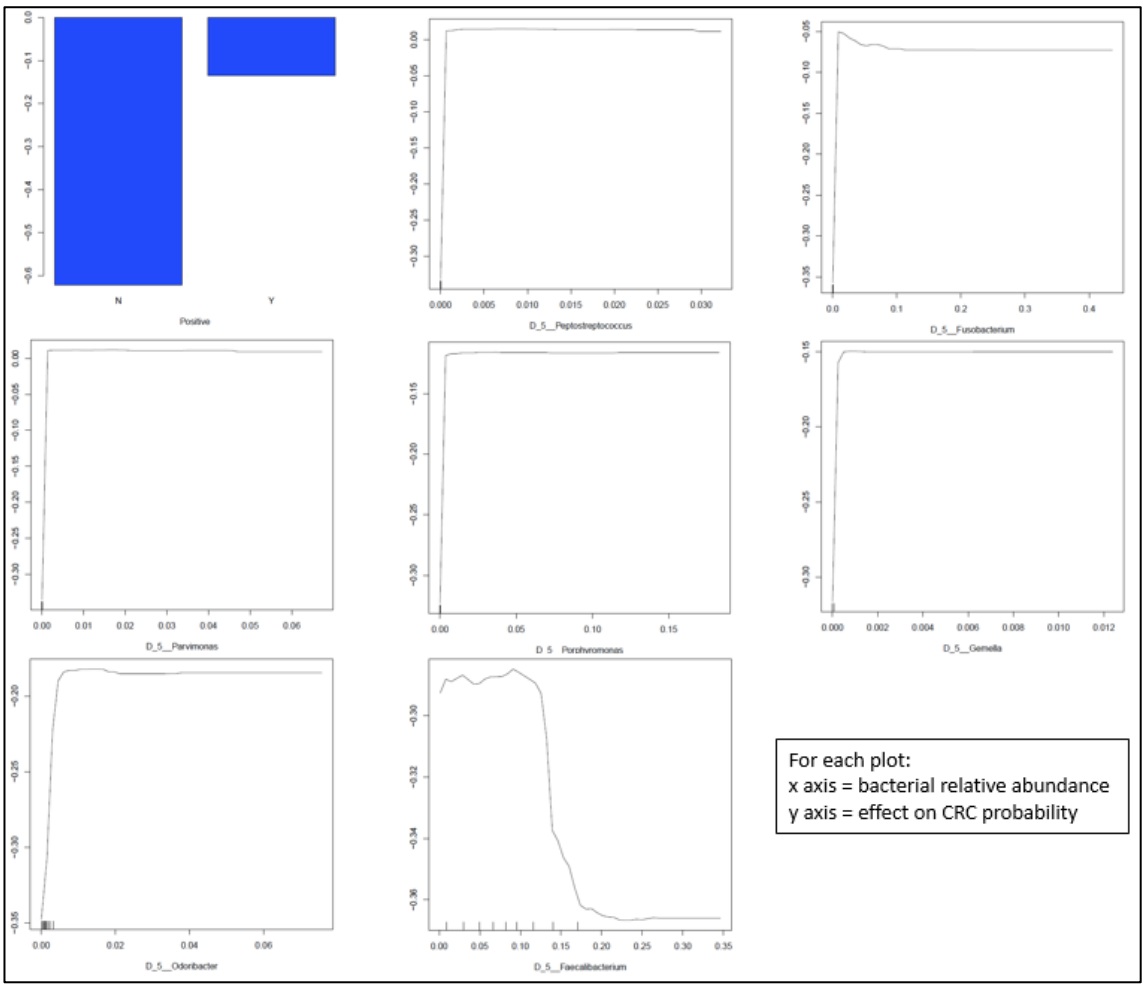


Figure 103. Partial dependence plots of some of the most important variables in a 'Bacteria and blood' Random Forest model designed to distinguish CRC samples from all other sample types. 'Bacteria and blood' = a model which used gFOBT blood status and genus-level relative abundance. The first plot shows the effect of gFOBT blood positivity status on CRC probability (N = blood-negative, Y = blood-positive); CRC probability is higher with gFOBT blood-positive status. The remaining plots show the effect of varying the relative abundances of taxa; for all except *Faecalibacterium*, CRC probability is higher at higher relative abundances.

3.5.5.2 All samples: distinction between neoplasm and all other sample types

Two taxa-based Random Forest models were designed to distinguish between neoplasm and all other sample types; the first used genus-level relative abundance alone and the second combined genus-level relative abundance with gFOBT blood-status. These models were compared with a baseline model which used gFOBT blood status, age and gender. The performance of each model is outlined in Table 31, Figure 104 and Figure 105. Interestingly there was no significant difference between the ROC curves of the model which used genus-level relative abundance alone and the baseline model; these ROC curves can be seen to cross (Figure 105).

The optimum model used both genus-level relative abundance and gFOBT blood status with AUC 0.868 (95% CI: 0.848-0.886). Figure 106 and Figure 107 show the top 15 bacteria which contributed to each model. Like the aforementioned CRC model, the most important bacteria included *Peptostreptococcus, Fusobacterium, Parvimonas, Gemella* and *Faecalibacterium.*

Table 31. Performance of Random Forest models designed to distinguish neoplasm samples from all other sample types. The performance of three Random Forest models is tabulated: (1) 'Clinical data' = a model which used gFOBT blood status, age and gender. (2) 'Bacteria only' = a model which used genus-level relative abundance. (3) 'Bacteria and blood' = a model which used gFOBT blood status and genus-level relative abundance.

Clinical data (blood status, age, gender)			
True value	**Predicted value**		**Error**
	Neoplasm (844)	**No (439)**	
Neoplasm (n=547)	515	32	6%
No (n=736)	329	407	45%
Sensitivity	515/547=94%		
Specificity	407/736=55%		
Bacteria only			
True value	**Predicted value**		**Error**
	Neoplasm (493)	**No (790)**	
Neoplasm (n=547)	334	213	39%
No (n=736)	159	577	22%
Sensitivity	334/547=61%		
Specificity	577/736=78%		
Bacteria and blood			
True value	**Predicted value**		**Error**
	Neoplasm (710)	**No (573)**	
Neoplasm (n=547)	480	67	12%
No (n=736)	230	506	31%
Sensitivity	480/547=88%		
Specificity	506/736=69%		

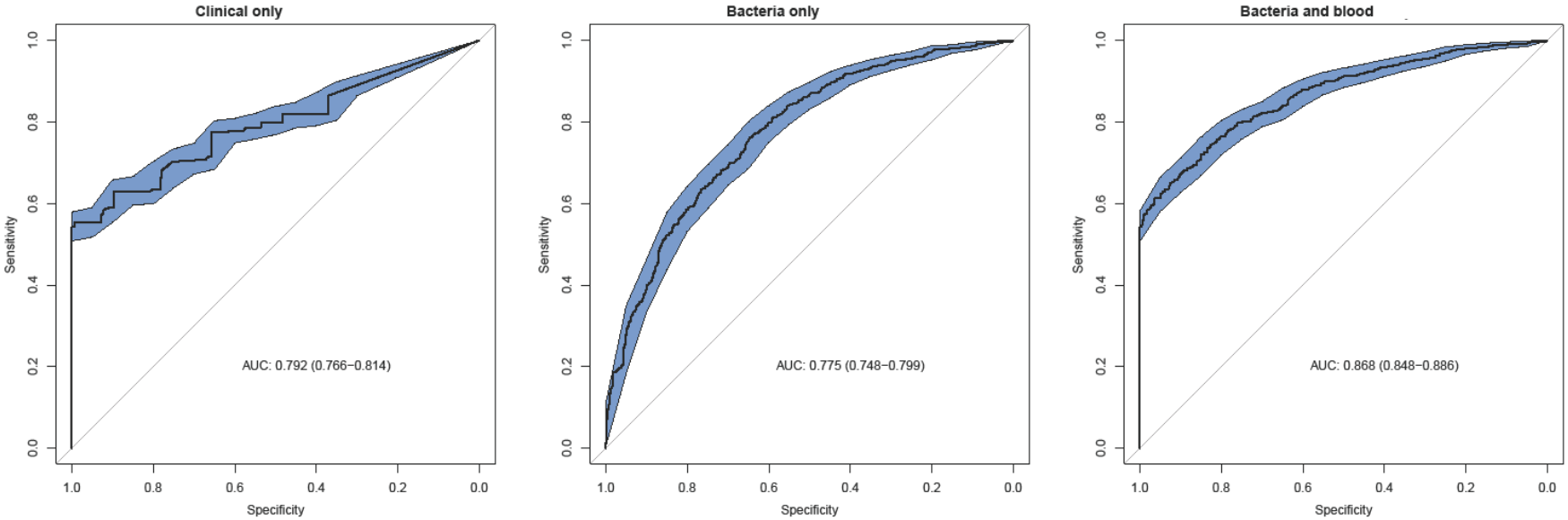


Figure 104. ROC curves of Random Forest models designed to distinguish neoplasm samples from all other sample types. ROC curves and AUC (with 95% CI) are displayed for three Random Forest models: (left) 'Clinical only' = a model which used gFOBT blood status, age and gender; (middle) 'Bacteria only' = a model which used genus-level relative abundance; (right) 'Bacteria and blood' = a model which used gFOBT blood status and genus-level relative abundance.

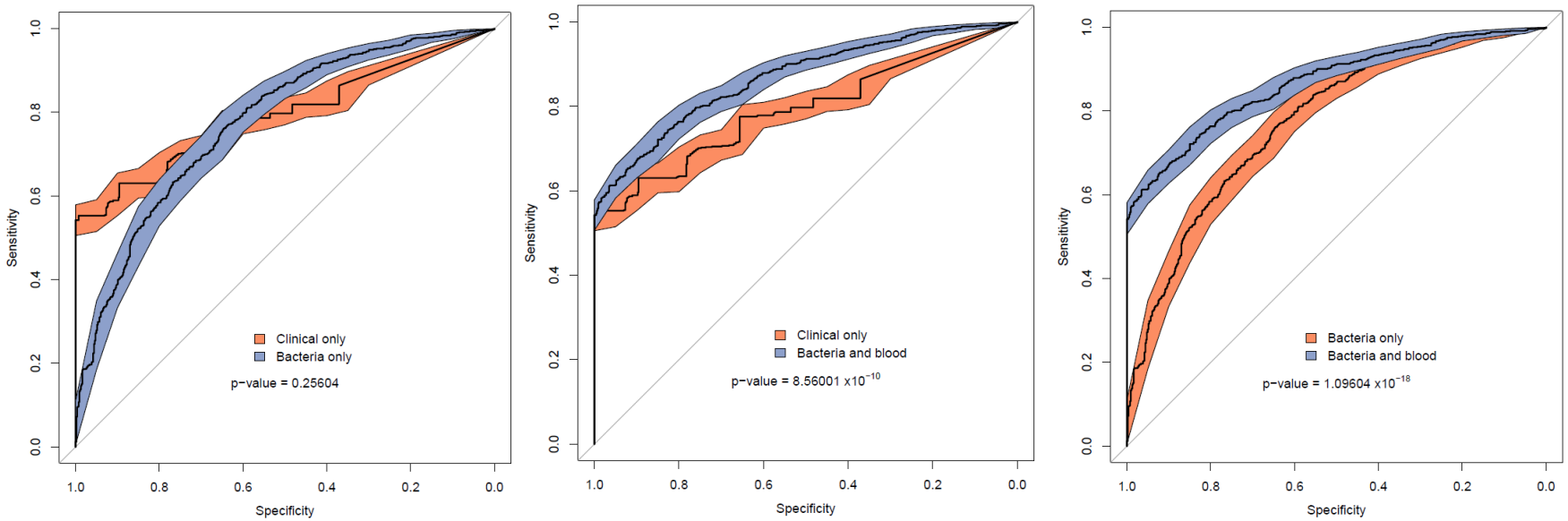


Figure 105. Comparison of ROC curves of Random Forest models designed to distinguish neoplasm samples from all other sample types. 'Clinical only' = a model which used gFOBT blood status, age and gender; 'Bacteria only' = a model which used genus-level relative abundance; 'Bacteria and blood' = a model which used gFOBT blood status and genus-level relative abundance.

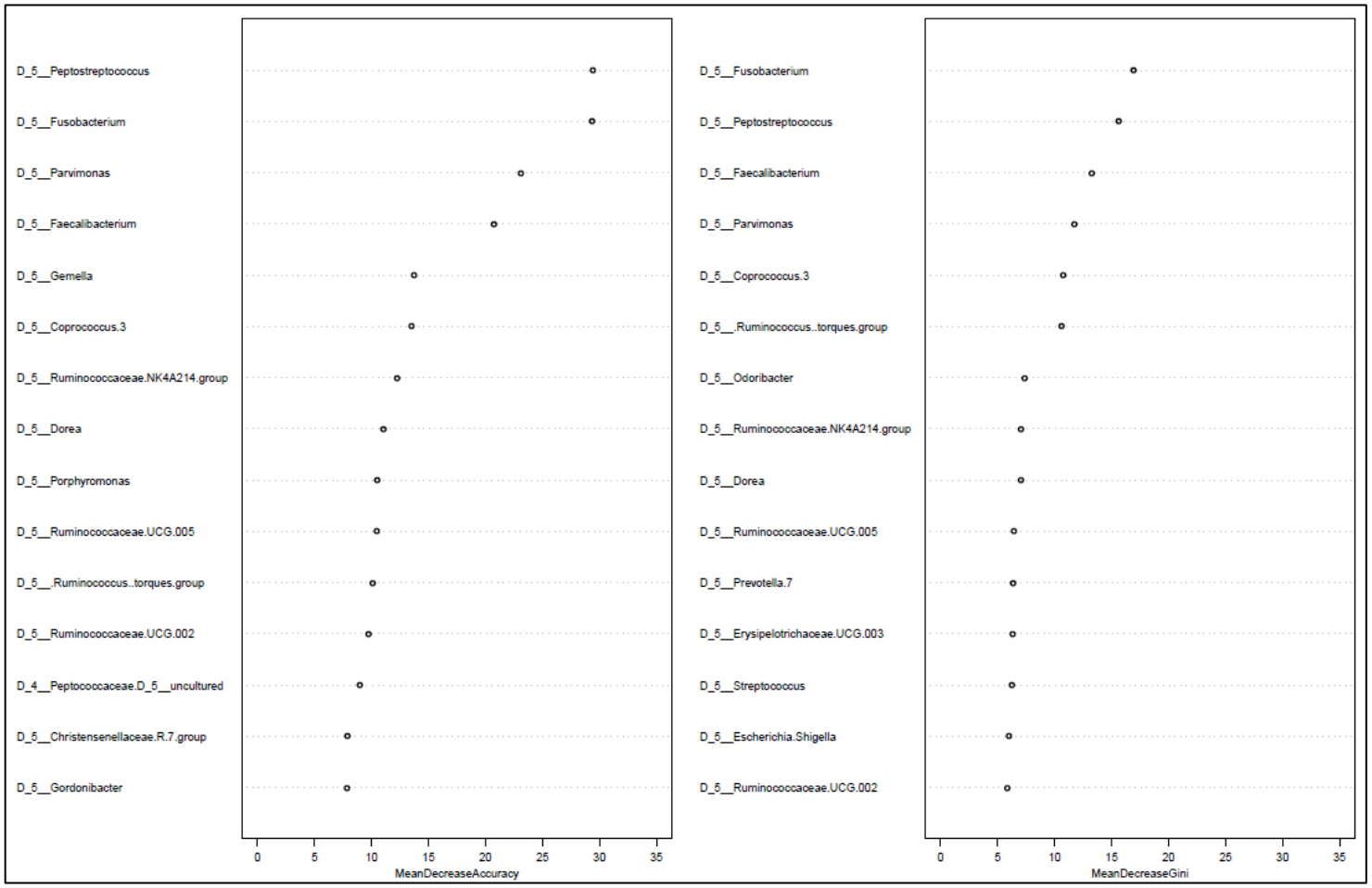


Figure 106. The 15 most important variables in a 'Bacteria only' Random Forest model designed to distinguish neoplasm samples from all other sample types. 'Bacteria only' = a model which used genus-level relative abundance.

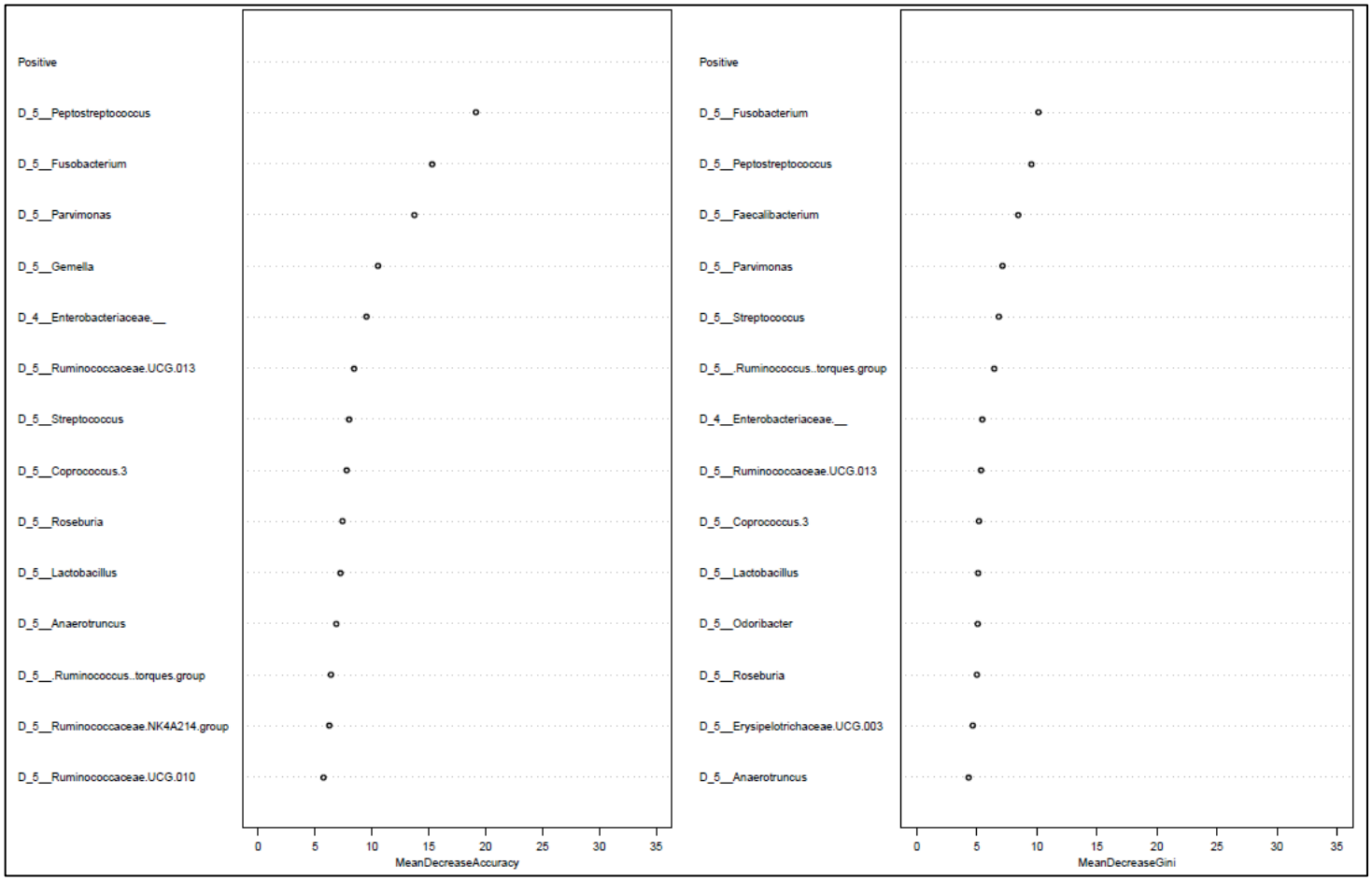


Figure 107. The 15 most important variables in a 'Bacteria and blood' Random Forest model designed to distinguish neoplasm samples from all other sample types. 'Bacteria and blood' = a model which used gFOBT blood status and genus-level relative abundance. Positive denotes gFOBT blood positivity status.

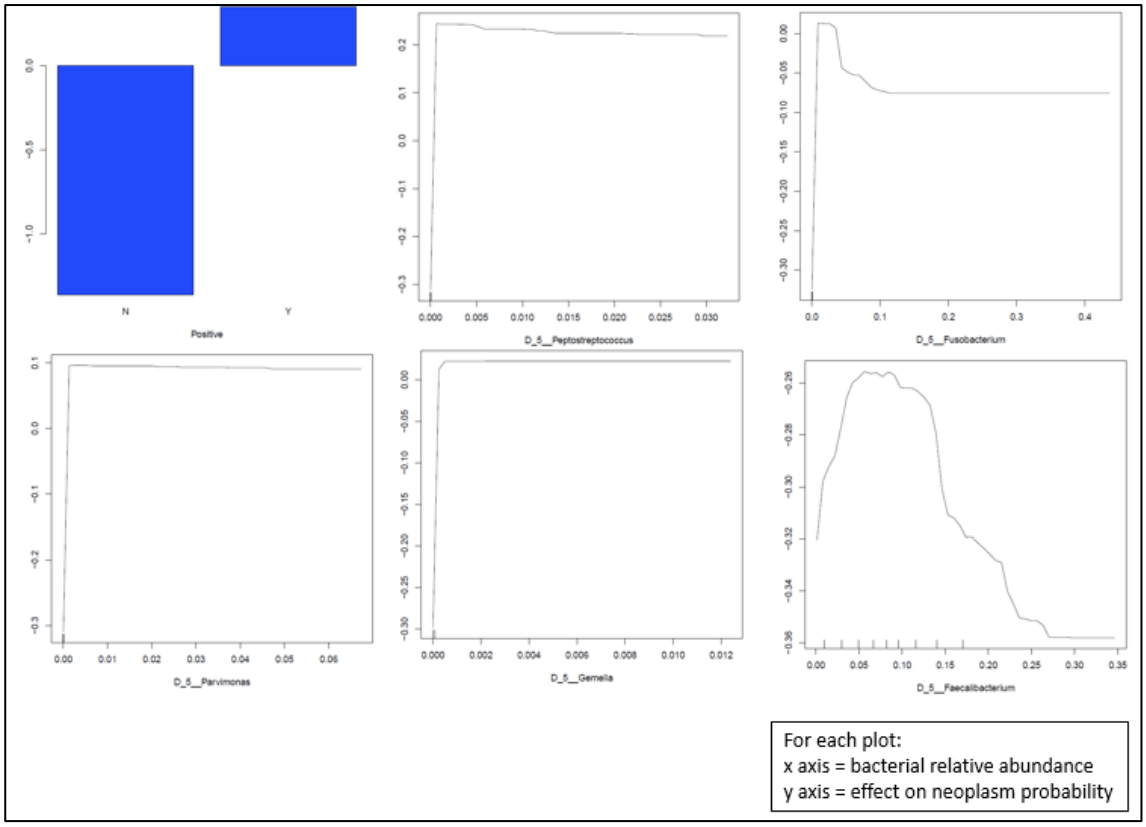


Figure 108. Partial dependence plots of some of the most important variables in a 'Bacteria and blood' Random Forest model designed to distinguish neoplasm samples from all other sample types. 'Bacteria and blood' = a model which used gFOBT blood status and genus-level relative abundance. The first plot shows the effect of gFOBT blood positivity status on neoplasm probability (N = blood-negative, Y = blood-positive); neoplasm probability is higher with gFOBT blood-positive status. The remaining plots show the effect of varying the relative abundances of taxa; for all except *Faecalibacterium*, neoplasm probability is higher at higher relative abundances.

3.5.5.3 Blood-positive samples: distinction between CRC and all other sample types

One taxa-based Random Forest model was designed to distinguish between CRC and all other sample types from within the blood-positive samples only. This model was compared with a baseline model which used age and gender alone. The performance of each model is outlined in Table 32 and Figure 109. The model which used genus-level relative abundance produced a significantly improved AUC relative to baseline with AUC 0.752 (95% CI: 0.714-0.787). Figure 110 shows the top 15 bacteria which contributed to the model. Like the aforementioned models, the most important bacteria included *Parvimonas, Peptostreptococcus, Fusobacterium, Porphyromonas, Gemella* and *Odoribacter.*

Table 32. Performance of Random Forest models designed to distinguish from within the blood-positive samples, CRC samples from all other sample types. The performance of two Random Forest models is tabulated: (1) 'Clinical data' = a model which used age and gender. (2) 'Bacteria only' = a model which used genus-level relative abundance.

Clinical data (age, gender)			
	Predicted value		
True value	**Cancer (387)**	**No (497)**	**Error**
Cancer (n=256)	127	129	50%
No (n=628)	260	368	41%
Sensitivity	127/256=50%		
Specificity	368/628=59%		
Bacteria only			
	Predicted value		
True value	**Cancer (212)**	**No (672)**	**Error**
Cancer (n=256)	124	132	52%
No (n=628)	88	540	14%
Sensitivity	124/256=48%		
Specificity	540/628=86%		

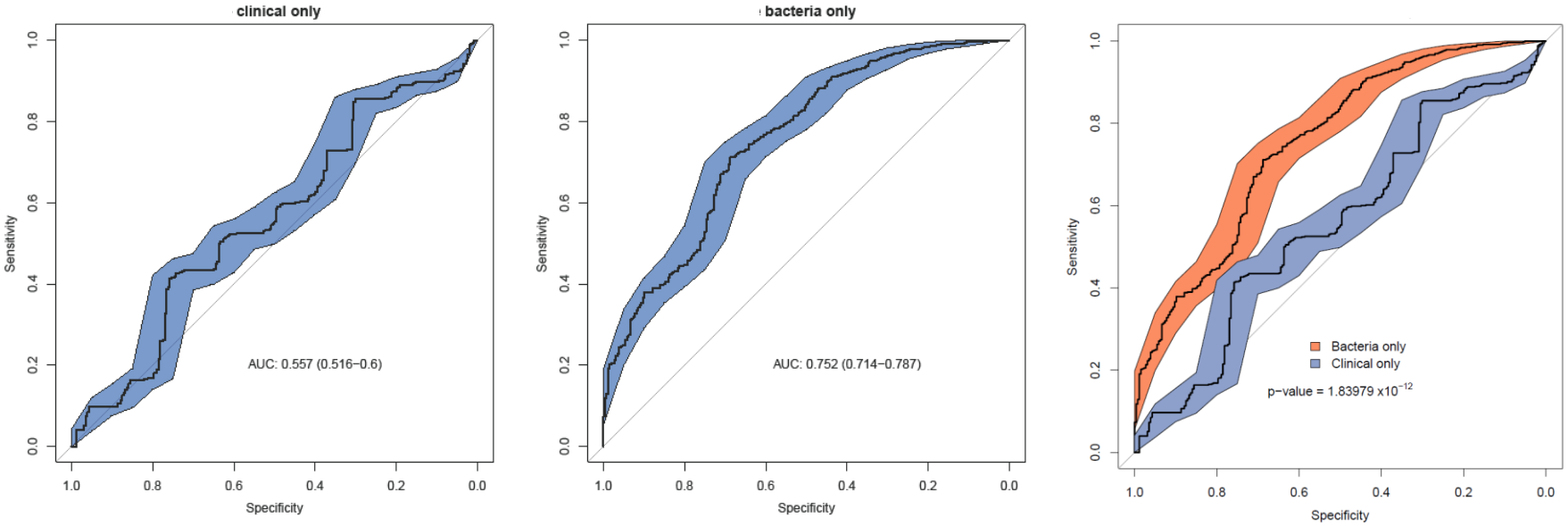


Figure 109. ROC curves of Random Forest models designed to distinguish from within the blood-positive samples, CRC samples from all other sample types. ROC curves and AUC (with 95% CI) are displayed for two Random Forest models: (left) 'Clinical only' = a model which used age and gender; (middle) 'Bacteria only' = a model which used genus-level relative abundance; (right) comparison of the two ROC curves.

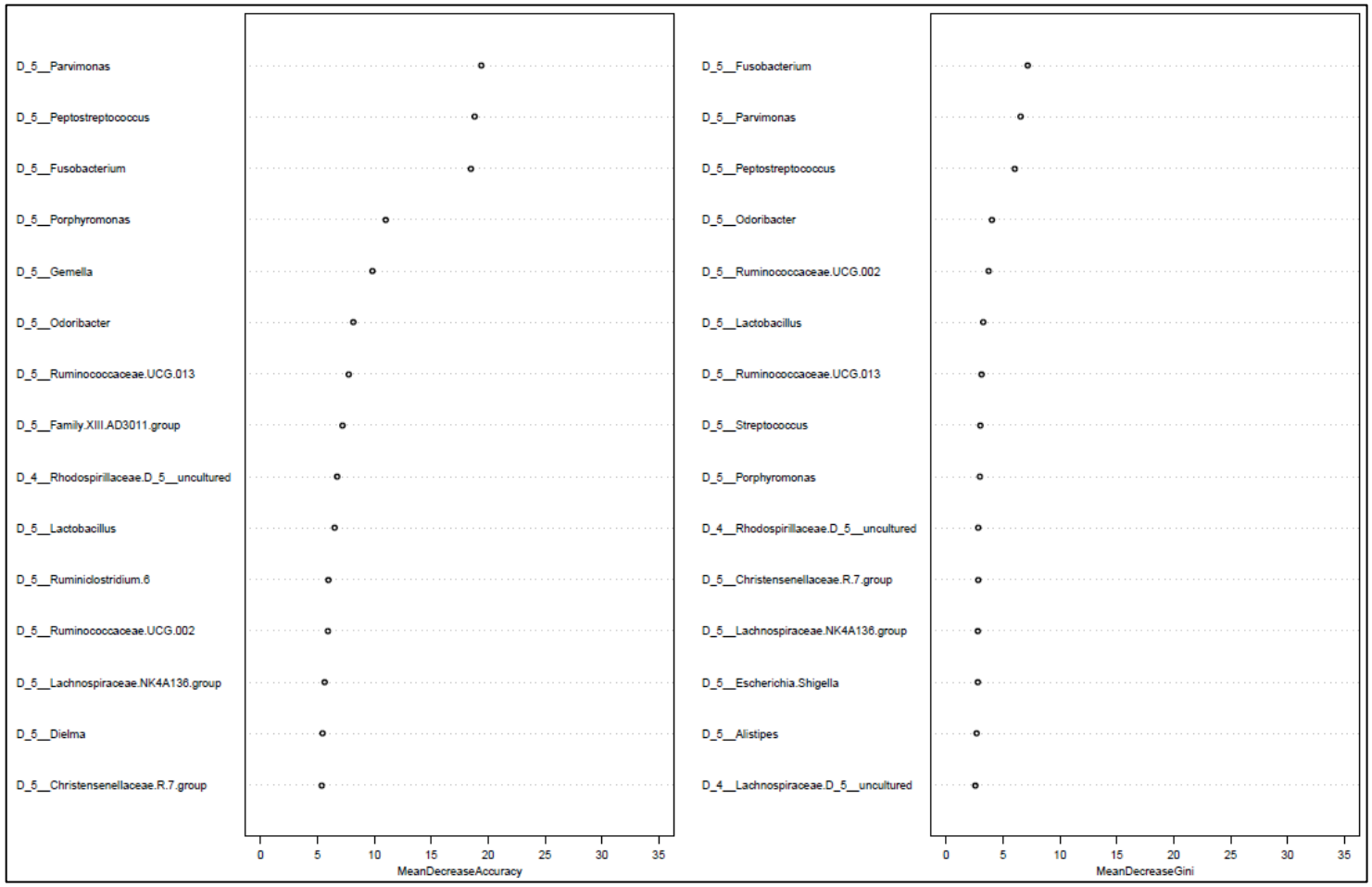


Figure 110. The 15 most important variables in a 'Bacteria only' Random Forest model designed to distinguish from within the blood-positive samples, CRC samples from all other sample types. 'Bacteria only' = a model which used genus-level relative abundance.

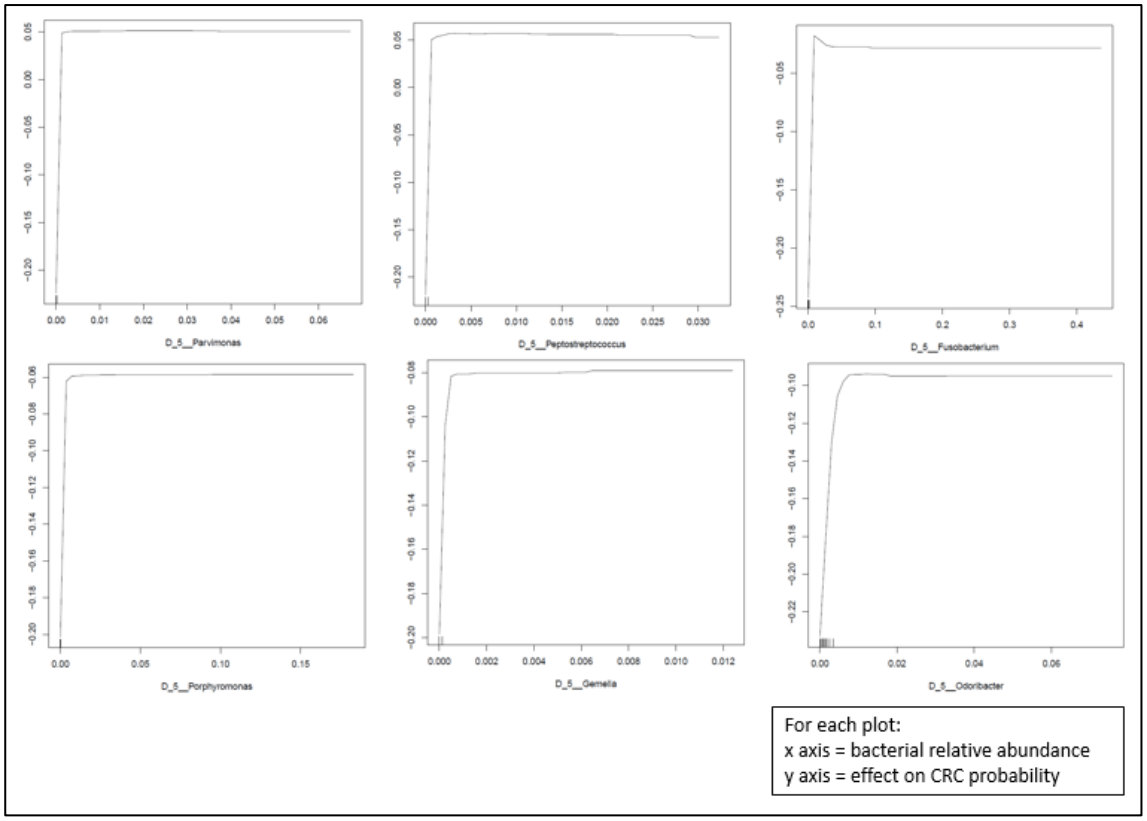


Figure 111. Partial dependence plots of some of the most important variables in a 'Bacteria only' Random Forest model designed to distinguish from within the blood-positive samples, CRC samples from all other sample types. 'Bacteria only' = a model which used genus-level relative abundance. The plots show the effect on CRC probability of varying the relative abundances of taxa; for all taxa CRC probability is higher at higher relative abundances.

3.5.5.4 Blood-positive samples: distinction between neoplasm and all other sample types

One taxa-based Random Forest model was designed to distinguish between neoplasm and all other sample types from within the blood-positive samples only. This model was compared with a baseline model which used age and gender alone. The performance of each model is outlined in Table 33 and Figure 112. The model which used genus-level relative abundance produced a significantly improved AUC relative to baseline with AUC 0.704 (95% CI: 0.669-0.738). Figure 113 shows the top 15 bacteria which contributed to the model. Like the aforementioned models, the most important bacteria included *Odoribacter, Gemella* and *Parvimonas,* however other taxa including *Streptococcus*, *Lactobacillus* and *Prevotella.7* contributed more to the model.

Table 33. Performance of Random Forest models designed to distinguish from within the blood-positive samples, neoplasm samples from all other sample types. The performance of two Random Forest models is tabulated: (1) 'Clinical data' = a model which used age and gender. (2) 'Bacteria only' = a model which used genus-level relative abundance.

Clinical data (age, gender)			
True value	**Predicted value**		**Error**
	Neoplasm (472)	**No (412)**	
Neoplasm (n=547)	326	221	40%
No (n=337)	146	191	43%
Sensitivity	326/547=60%		
Specificity	191/337=57%		
Bacteria only			
True value	**Predicted value**		**Error**
	Neoplasm (569)	**No (315)**	
Neoplasm (n=547)	411	136	25%
No (n=337)	158	179	47%
Sensitivity	411/547=75%		
Specificity	179/337=53%		

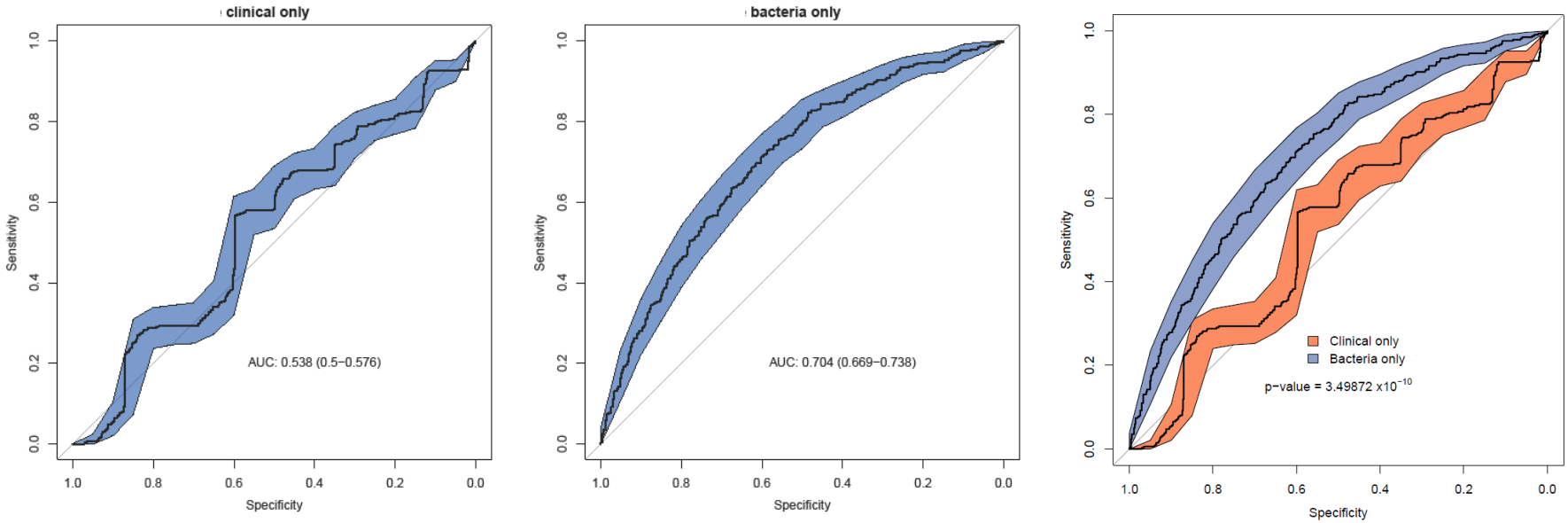


Figure 112. ROC curves of Random Forest models designed to distinguish from within the blood-positive samples, neoplasm samples from all other sample types. ROC curves and AUC (with 95% CI) are displayed for two Random Forest models: (left) 'Clinical only' = a model which used age and gender; (middle) 'Bacteria only' = a model which used genus-level relative abundance; (right) comparison of the two ROC curves.

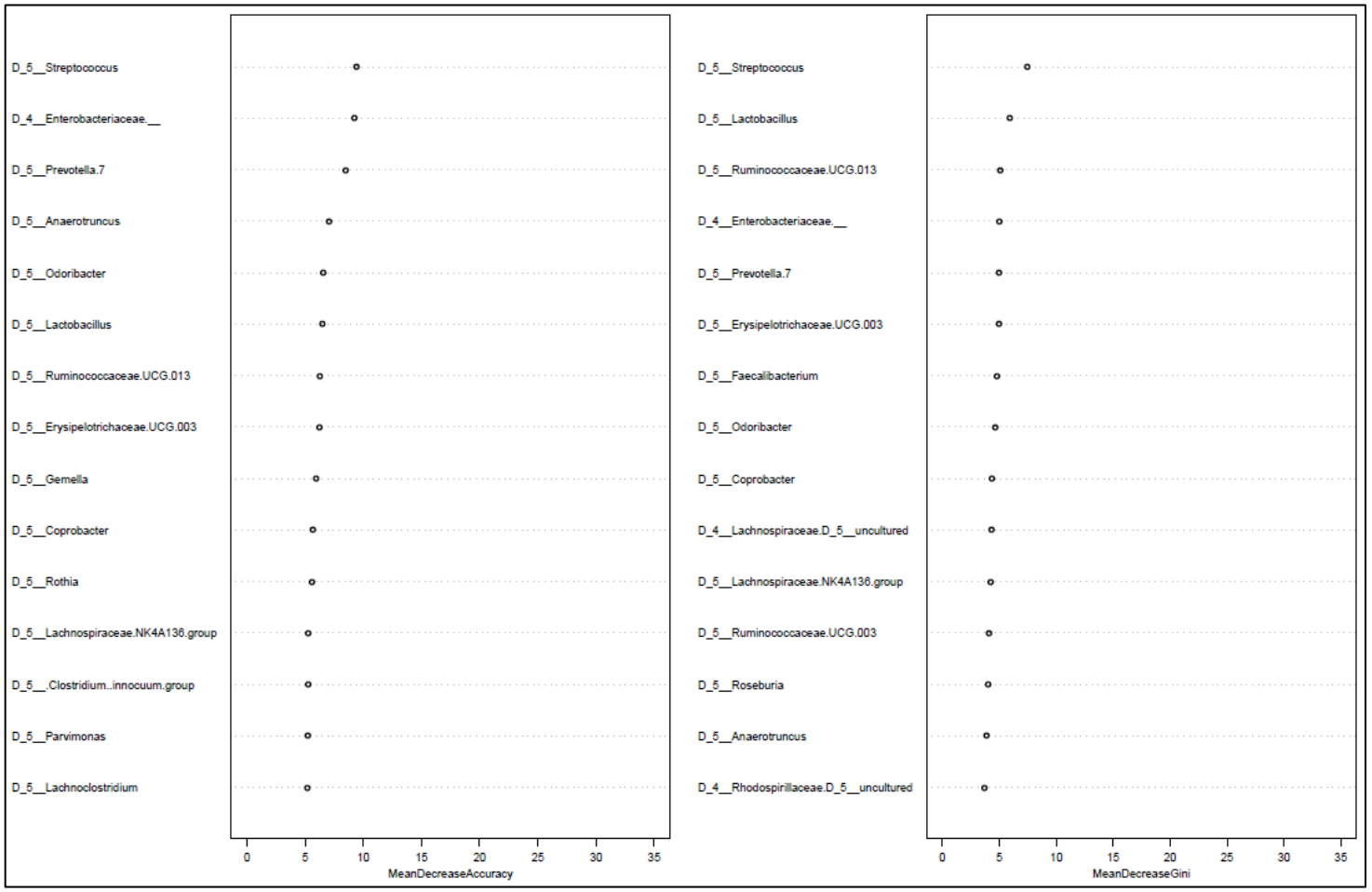


Figure 113. The 15 most important variables in a 'Bacteria only' Random Forest model designed to distinguish from within the blood-positive samples, neoplasm samples from all other sample types. 'Bacteria only' = a model which used genus-level relative abundance.

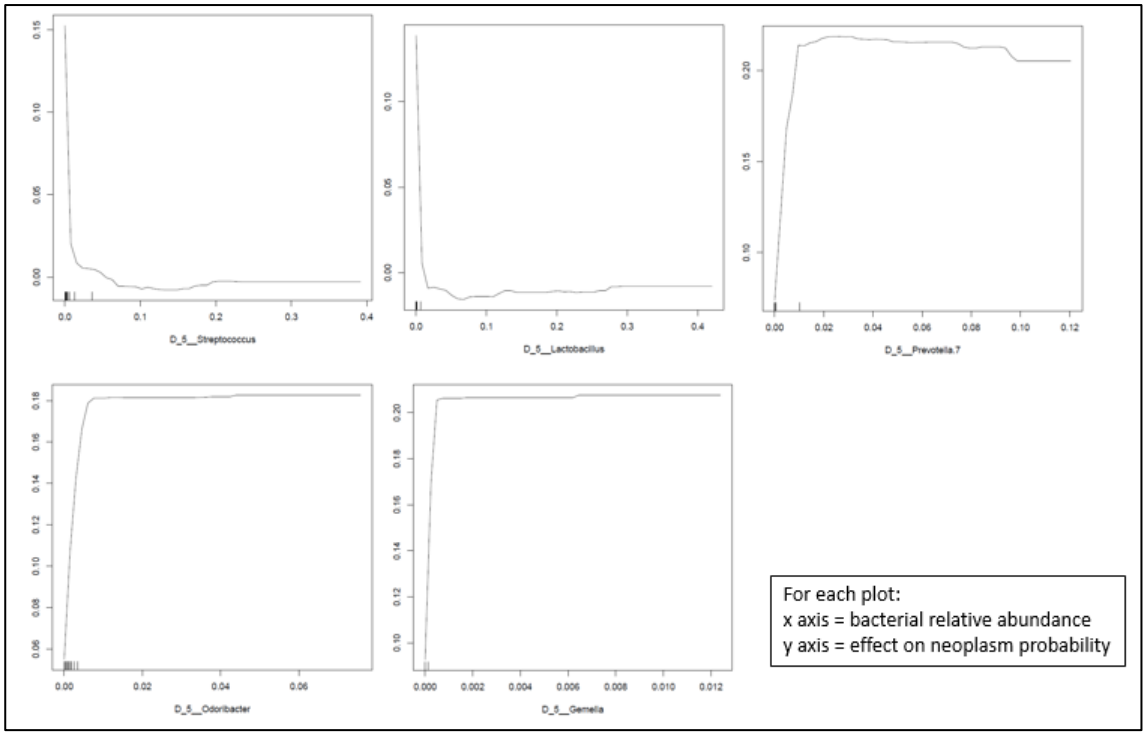


Figure 114. Partial dependence plots of some of the most important variables in a 'Bacteria only' Random Forest model designed to distinguish from within the blood-positive samples, neoplasm samples from all other sample types. 'Bacteria only' = a model which used genus-level relative abundance. The plots show the effect on neoplasm probability of varying the relative abundances of taxa; for all taxa except *Streptococcus* and *Lactobacillus*, neoplasm probability is higher at higher relative abundances.

3.5.5.5 Blood-positive samples: distinction between colonoscopy-normal samples and all other sample types

One taxa-based Random Forest model was designed to distinguish between colonoscopy-normal and all other sample types from within the blood-positive samples only. This model was compared with a baseline model which used age and gender alone. The performance of each model is outlined in Table 34 and Figure 115. The model which used genus-level relative abundance produced a significantly improved AUC relative to baseline with AUC 0.729 (95% CI: 0.692-0.767). Figure 116 shows the top 15 bacteria which contributed to the model. The most important bacteria were different from the aforementioned models apart from *Odoribacter and Faecalibacterium;* the most important were *Lactobacillus, Enterobacteriaceae, Streptococcus* and *Enterococcus*.

Table 34. Performance of Random Forest models designed to distinguish from within the blood-positive samples, colonoscopy-normal samples from all other sample types. The performance of two Random Forest models is tabulated: (1) 'Clinical data' = a model which used age and gender. (2) 'Bacteria only' = a model which used genus-level relative abundance.

Clinical data (age and gender)			
	Predicted value		
True value	**Normal (310)**	**Abnormal (574)**	**Error**
Normal (n=249)	110	139	56%
Abnormal (n=635)	200	435	31%
Sensitivity	110/249=44%		
Specificity	435/635=69%		
Bacteria only			
	Predicted value		
True value	**Normal (257)**	**Abnormal (627)**	**Error**
Normal (n=249)	125	124	50%
Abnormal (n=635)	132	503	21%
Sensitivity	125/249=50%		
Specificity	503/635=79%		

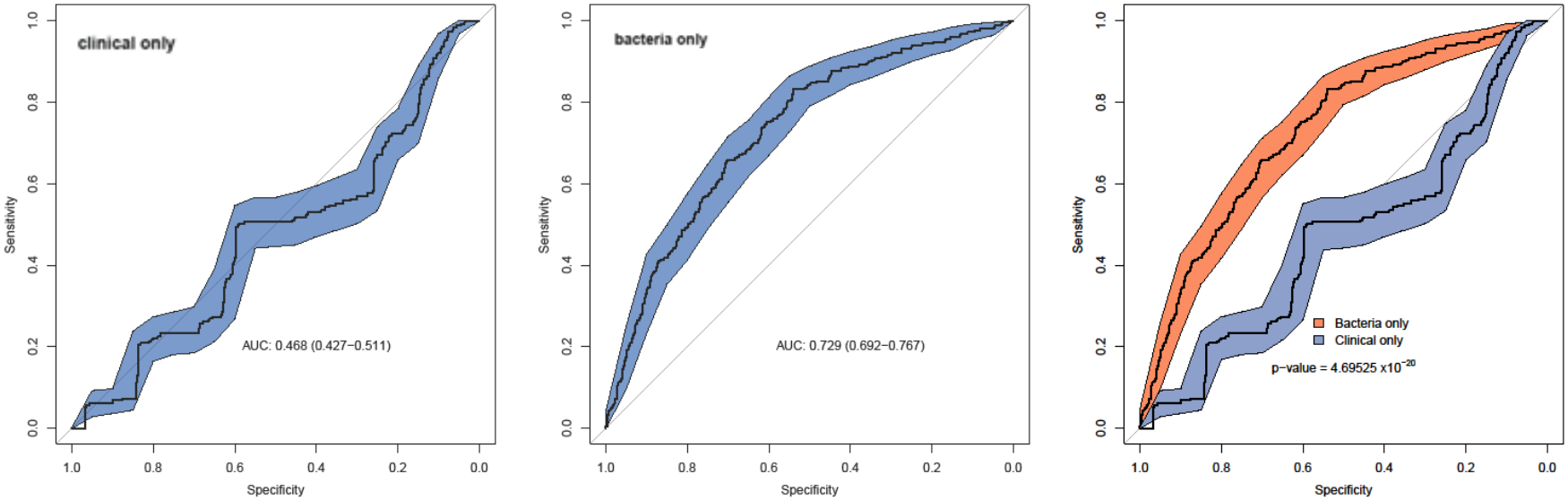


Figure 115. ROC curves of Random Forest models designed to distinguish from within the blood-positive samples, colonoscopy-normal samples from all other sample types. ROC curves and AUC (with 95% CI) are displayed for two Random Forest models: (left) 'Clinical only' = a model which used age and gender; (middle) 'Bacteria only' = a model which used genus-level relative abundance; (right) comparison of the two ROC curves.

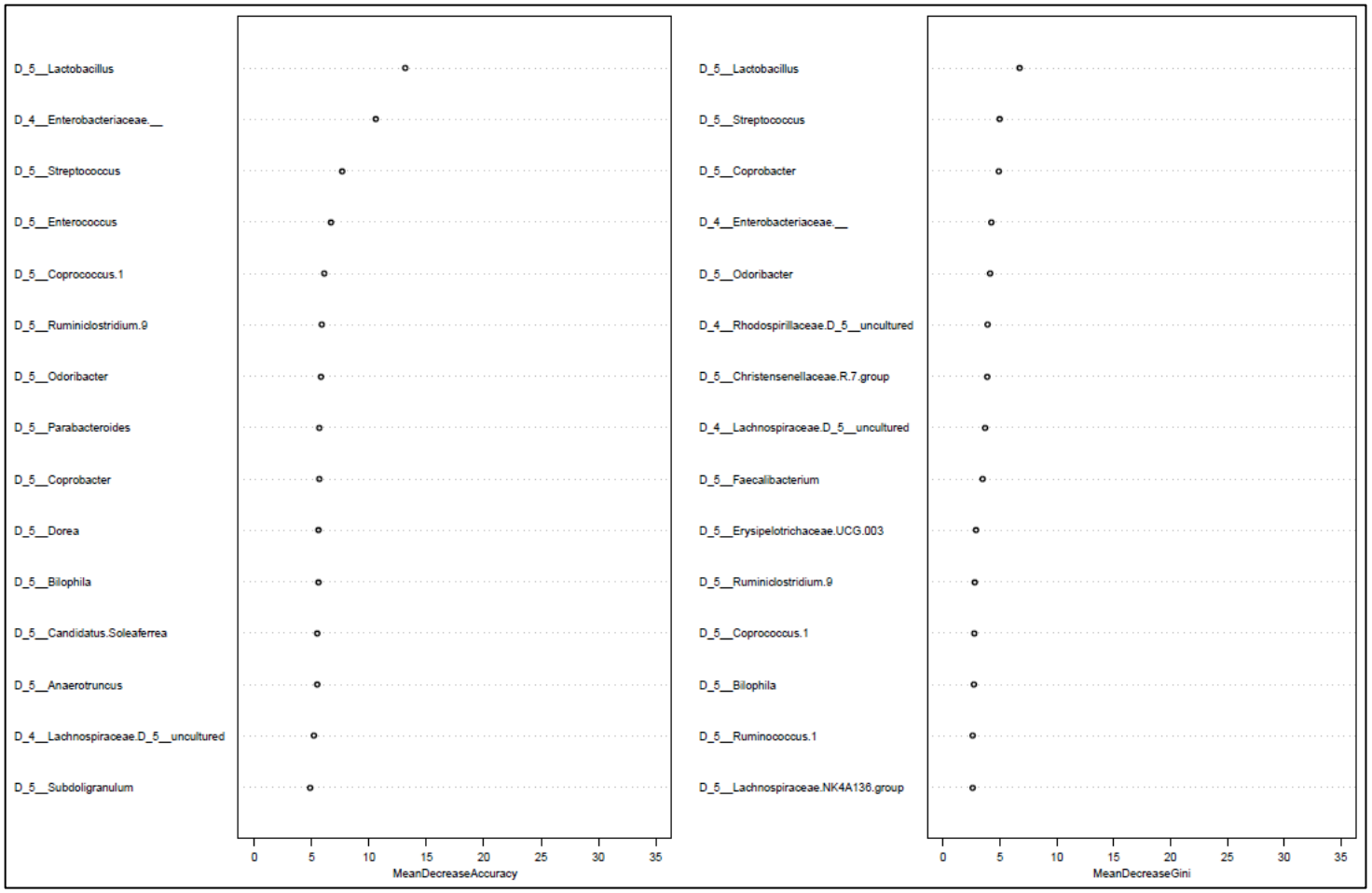


Figure 116. The 15 most important variables in a 'Bacteria only' Random Forest model designed to distinguish from within the blood-positive samples, colonoscopy-normal samples from all other sample types. 'Bacteria only' = a model which used genus-level relative abundance.

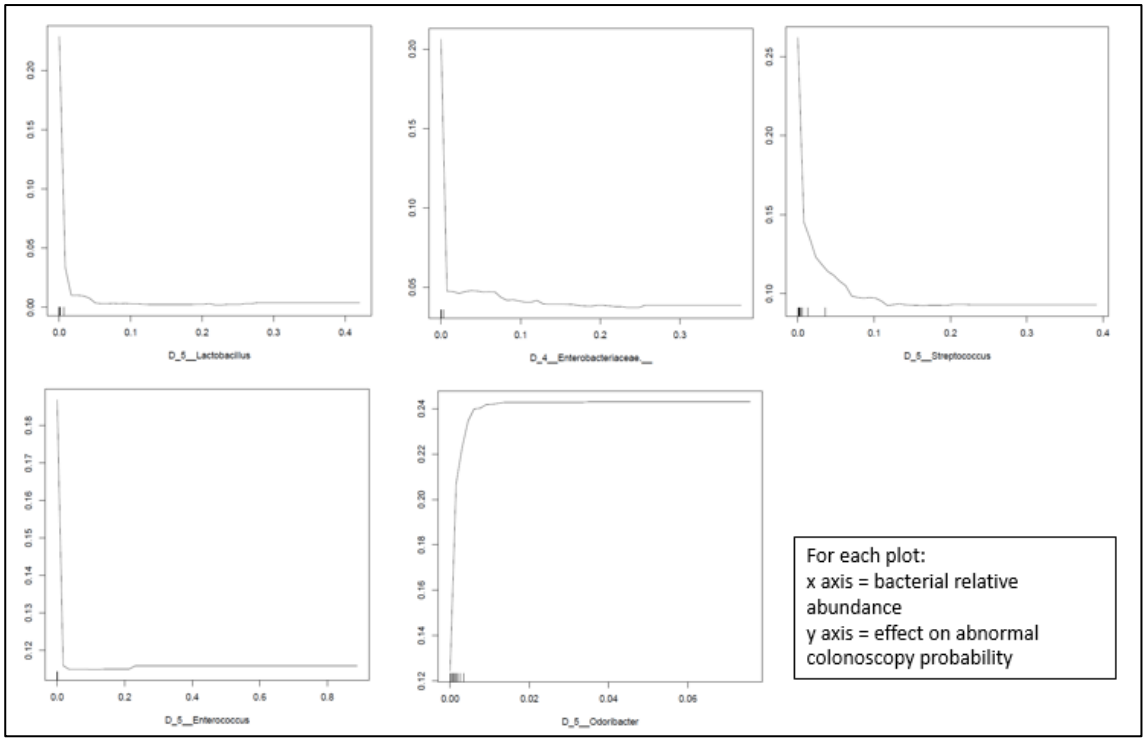
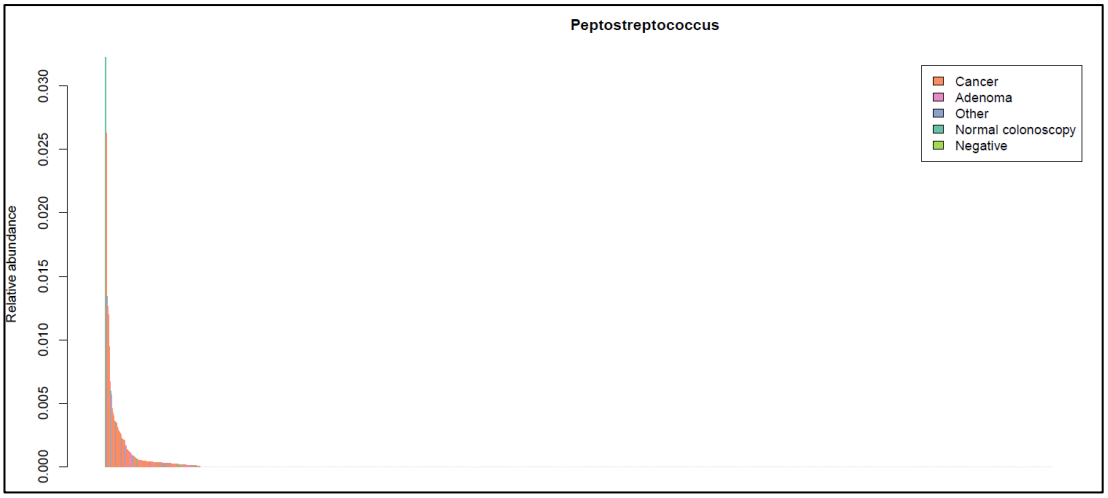
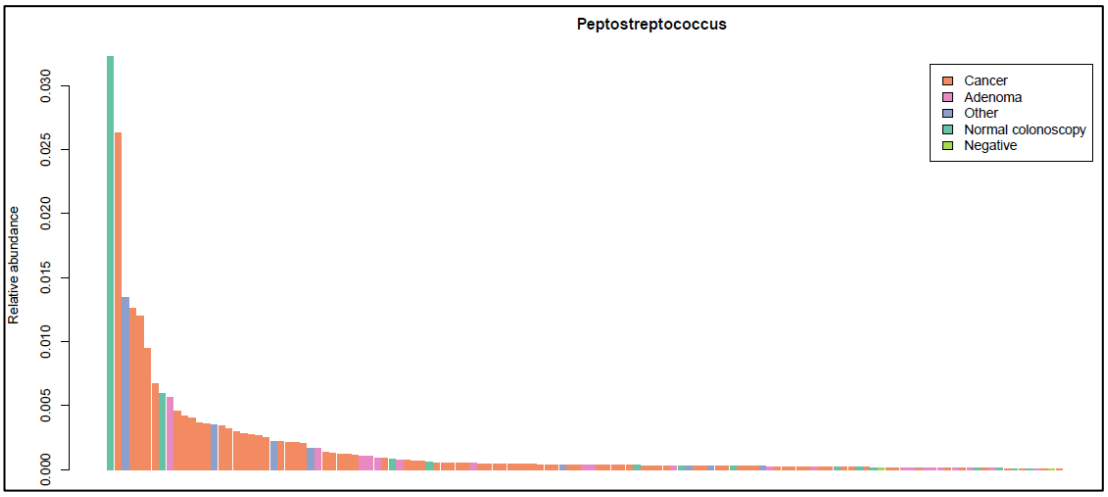


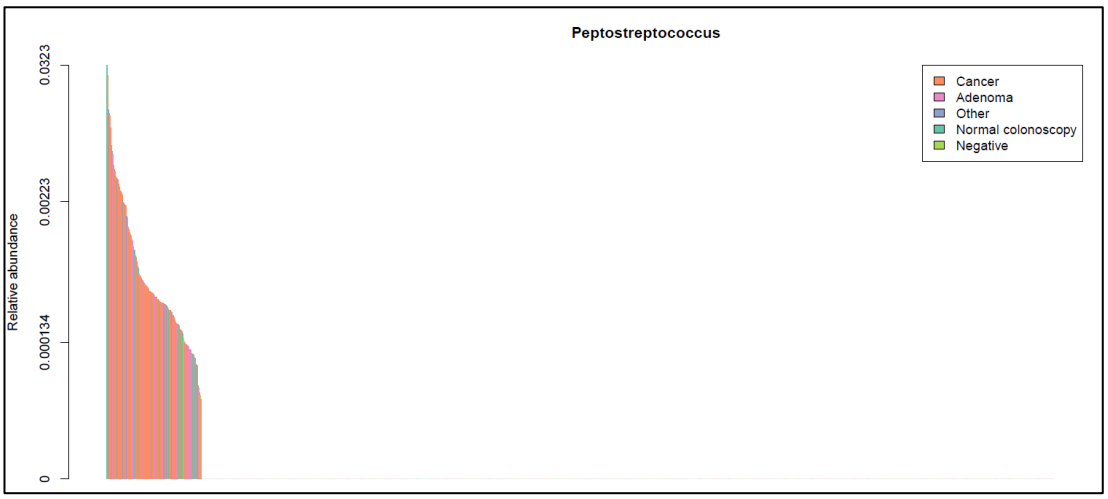
Figure 117. Partial dependence plots of some of the most important variables in a 'Bacteria only' Random Forest model designed to distinguish from within the blood-positive samples, colonoscopy-abnormal samples from all other sample types. 'Bacteria only' = a model which used genus-level relative abundance. The plots show the effect on 'abnormal colonoscopy' probability of varying the relative abundances of taxa; for all taxa except *Odoribacter*, 'abnormal colonoscopy' probability is lower at higher relative abundances.

3.5.6 Specific bacteria of interest

The distribution of relative abundances of the most important taxa identified by the Random Forest models was compared across NHSBCSP samples (Figure 118 to Figure 123).


Peptostreptococcus
Relative abundance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Cancer
Adenoma
Other
Normal colonoscopy
Negative


Peptostreptococcus
Relative abundance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Cancer
Adenoma
Other
Normal colonoscopy
Negative


Peptostreptococcus
Relative abundance
0
0.000134
0.00223
0.0323
Cancer
Adenoma
Other
Normal colonoscopy
Negative

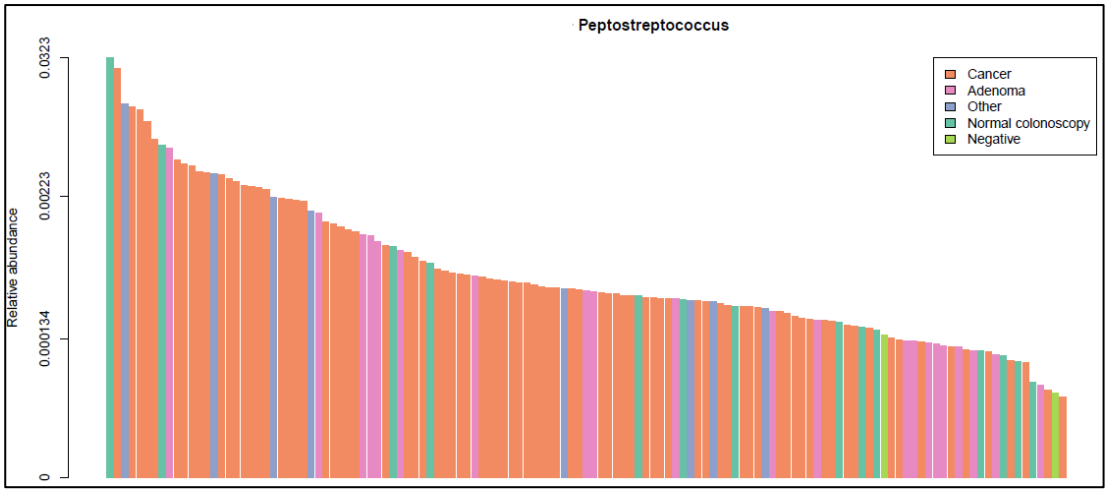
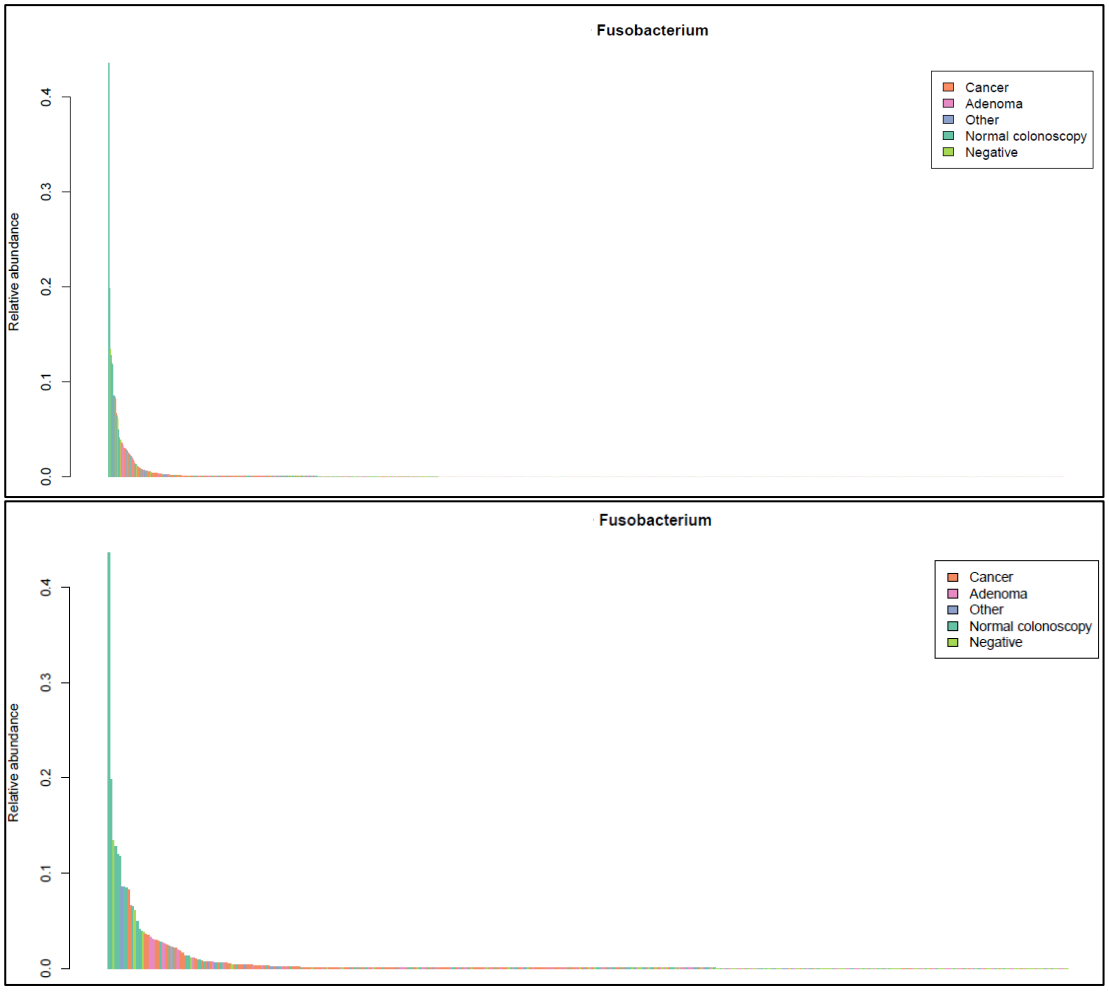


Figure 118. Waterfall plots of the relative abundance of *Peptostreptococcus* for NHSBCSP samples. The upper two plots have a normal y axis; the first shows all samples along the x axis and the second shows only samples with a relative abundance greater than 0, to enable better visualisation. The lower two plots have a logarithmic y axis to enable visualisation of low-abundance samples; the first shows all samples along the x axis and the second shows only samples with a relative abundance greater than 0, to enable better visualisation.



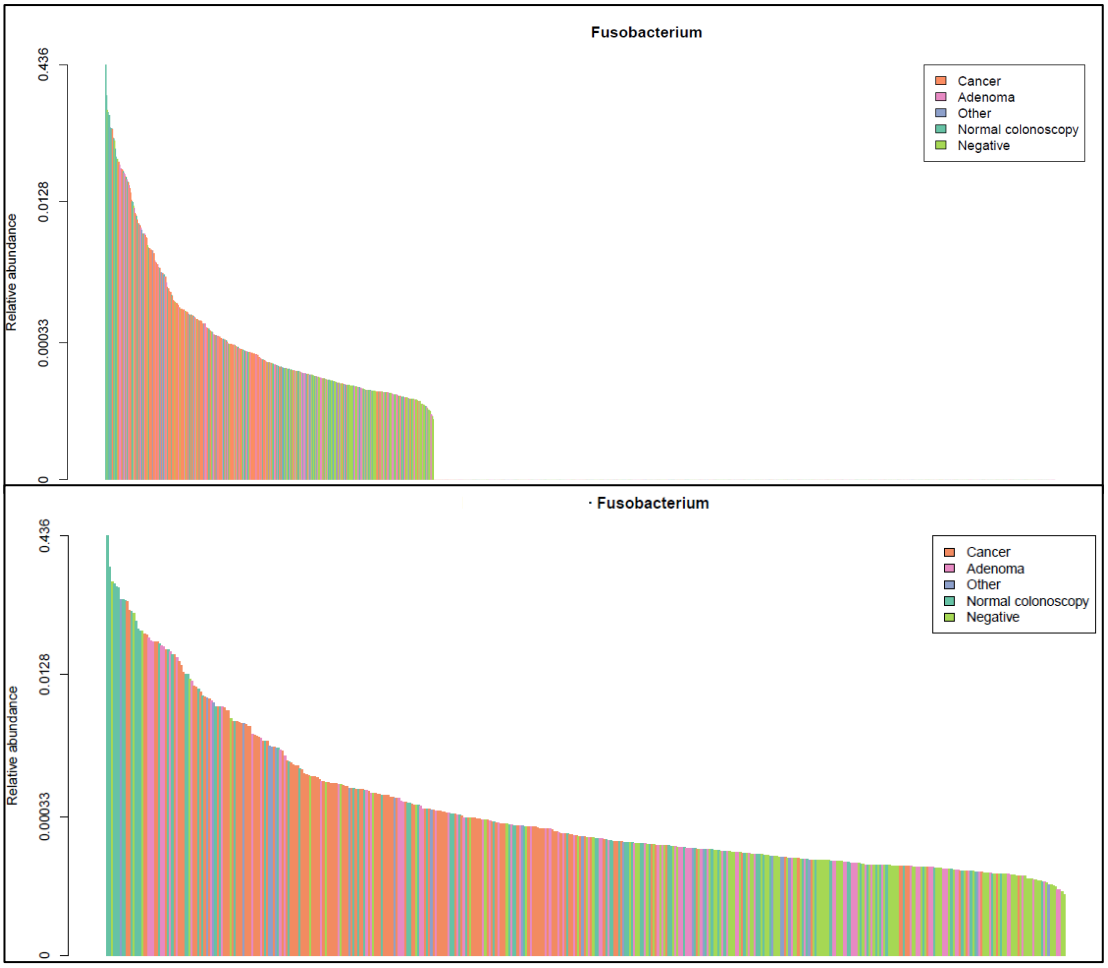
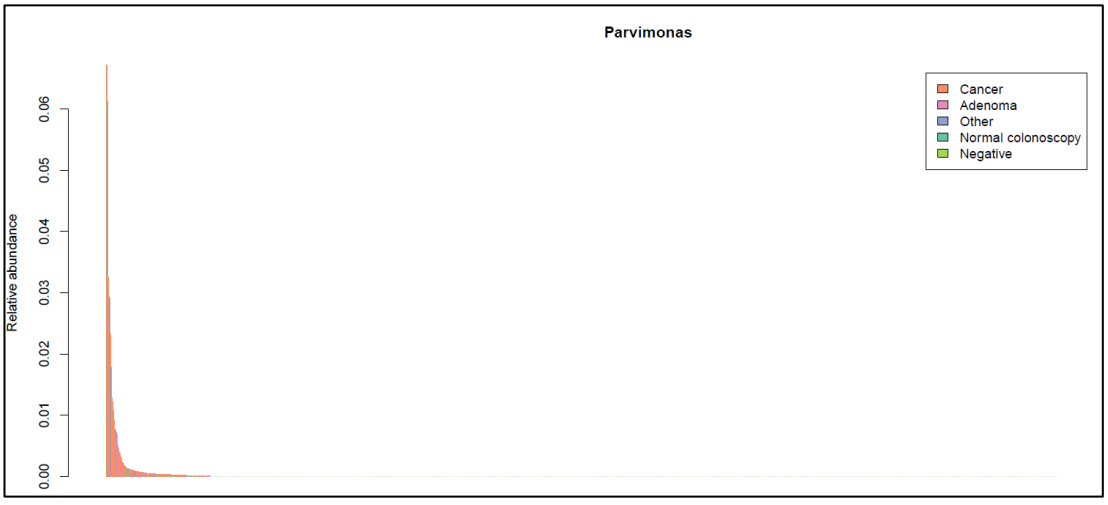


Figure 119. Waterfall plots of the relative abundance of *Fusobacterium* for NHSBCSP samples. The upper two plots have a normal y axis; the first shows all samples along the x axis and the second shows only samples with a relative abundance greater than 0, to enable better visualisation. The lower two plots have a logarithmic y axis to enable visualisation of low-abundance samples; the first shows all samples along the x axis and the second shows only samples with a relative abundance greater than 0, to enable better visualisation.



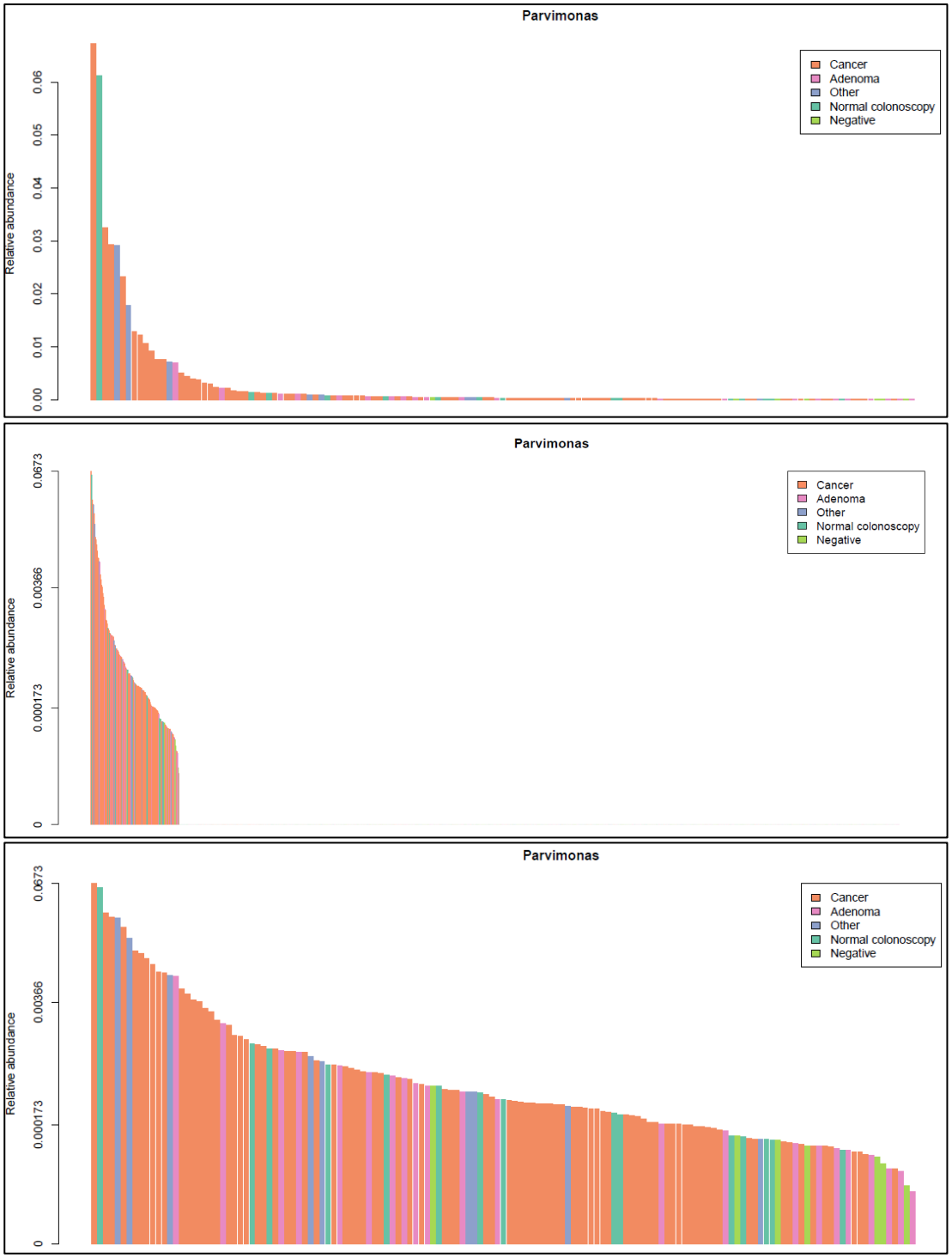


Figure 120. Waterfall plots of the relative abundance of *Parvimonas* for NHSBCSP samples. The upper two plots have a normal y axis; the first shows all samples along the x axis and the second shows only samples with a relative abundance greater than 0, to enable better visualisation. The lower two plots have a logarithmic y axis to enable visualisation of low-abundance samples; the first shows all samples along the x axis and the second shows only samples with a relative abundance greater than 0, to enable better visualisation.

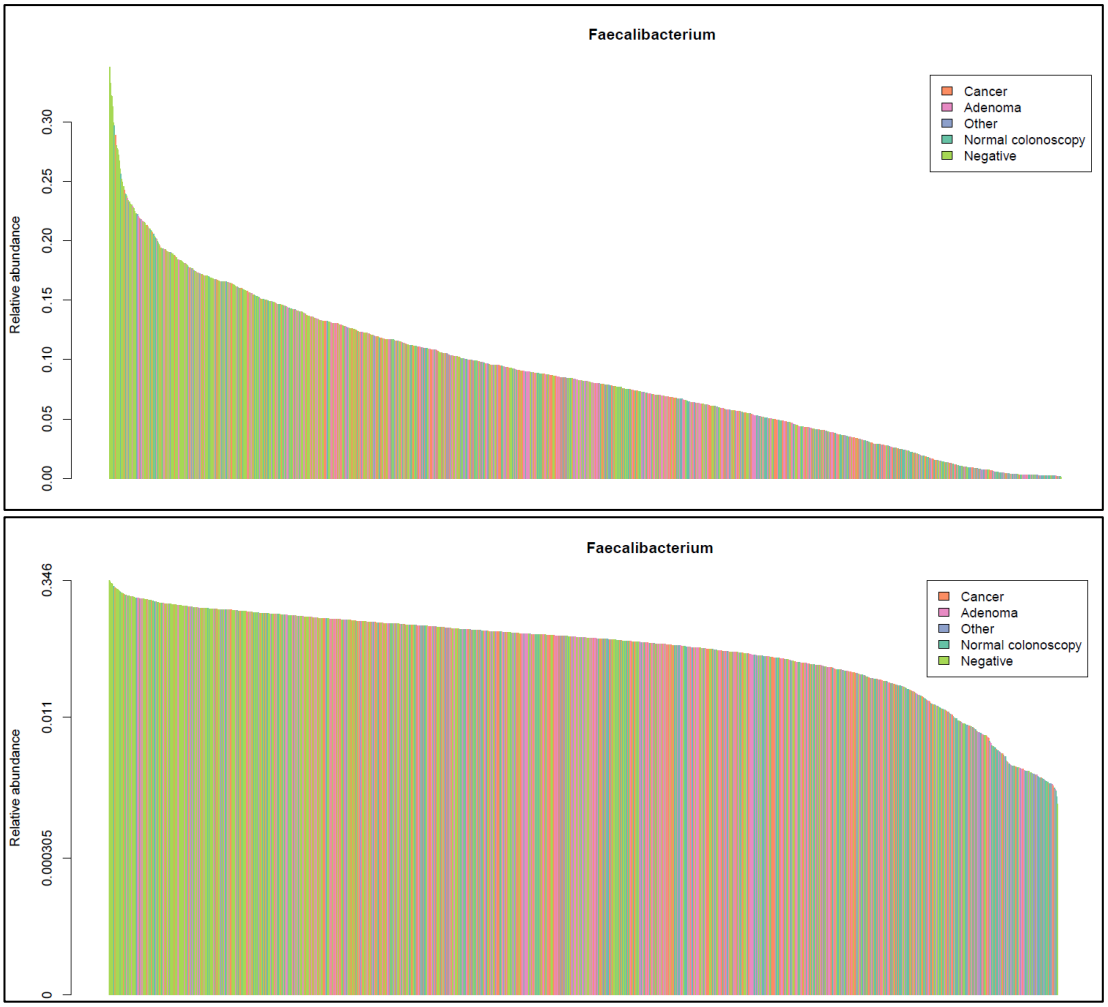
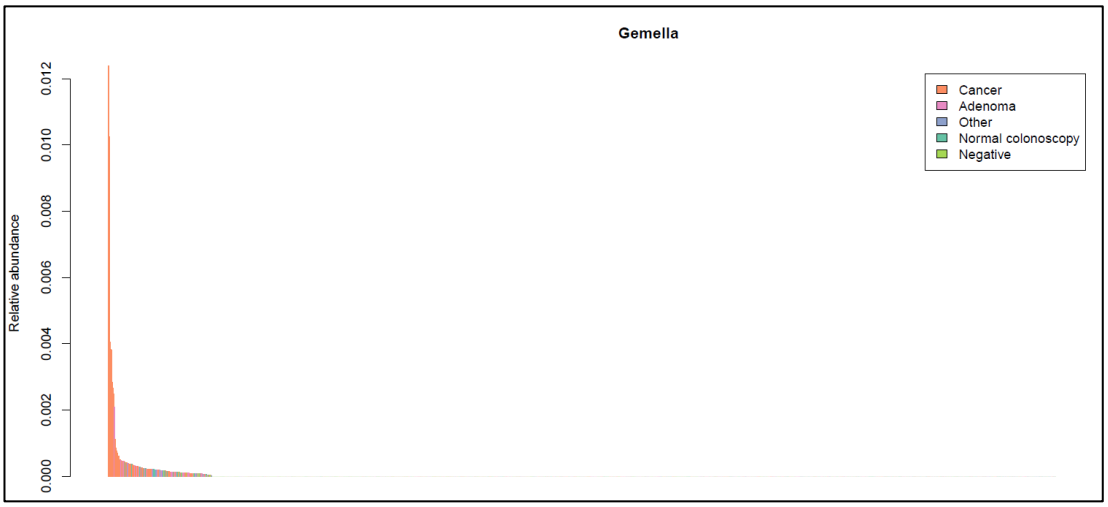


Figure 121. Waterfall plots of the relative abundance of *Faecalibacterium* for NHSBCSP samples. The upper plot has a normal axis; the lower plot has a logarithmic axis to enable visualisation of low-abundance samples.



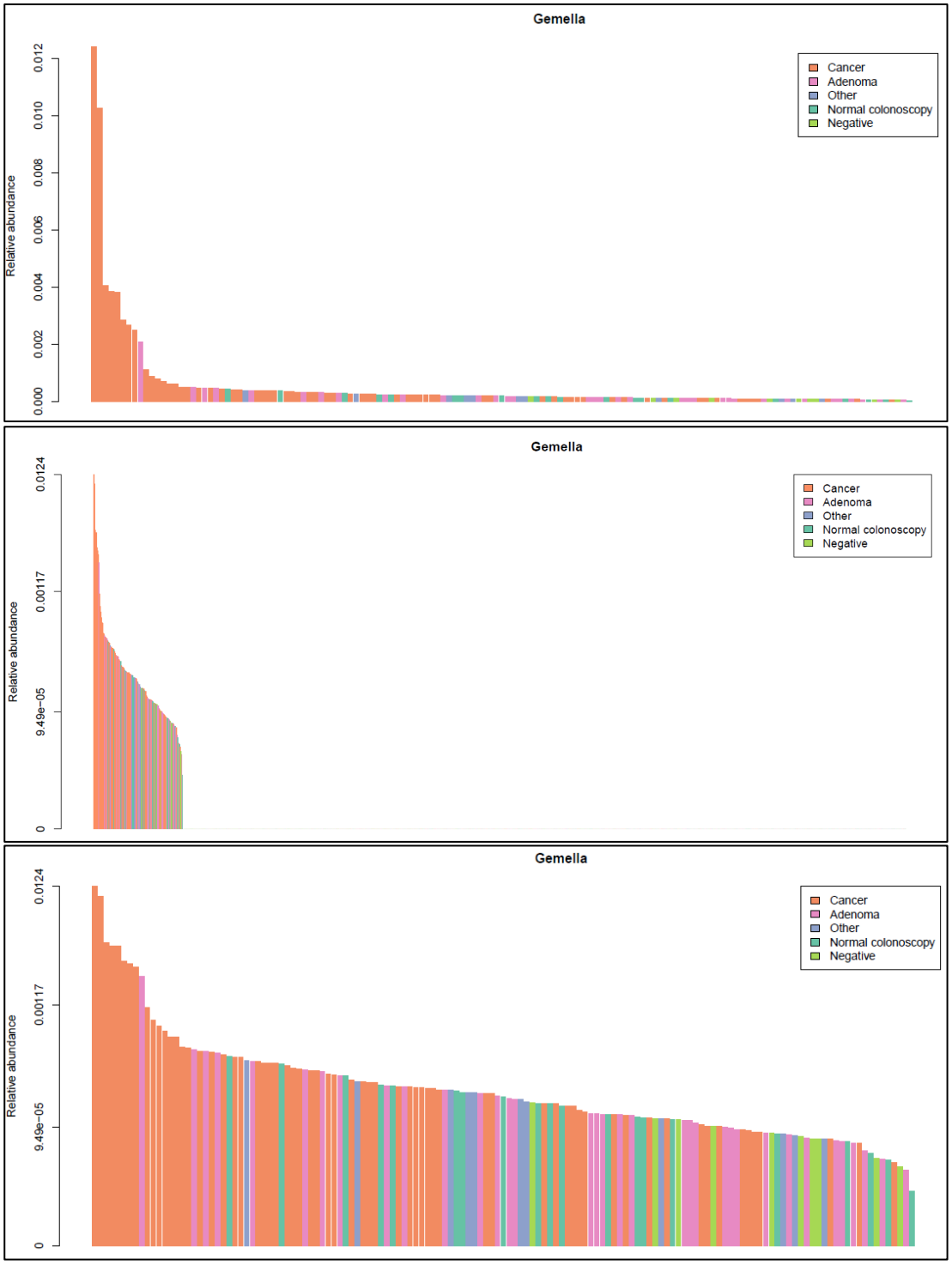


Figure 122. Waterfall plots of the relative abundance of *Gemella* for NHSBCSP samples. The upper two plots have a normal y axis; the first shows all samples along the x axis and the second shows only samples with a relative abundance greater than 0, to enable better visualisation. The lower two plots have a logarithmic y axis to enable visualisation of low-abundance samples; the first shows all samples along the x axis and the second shows only samples with a relative abundance greater than 0, to enable better visualisation.

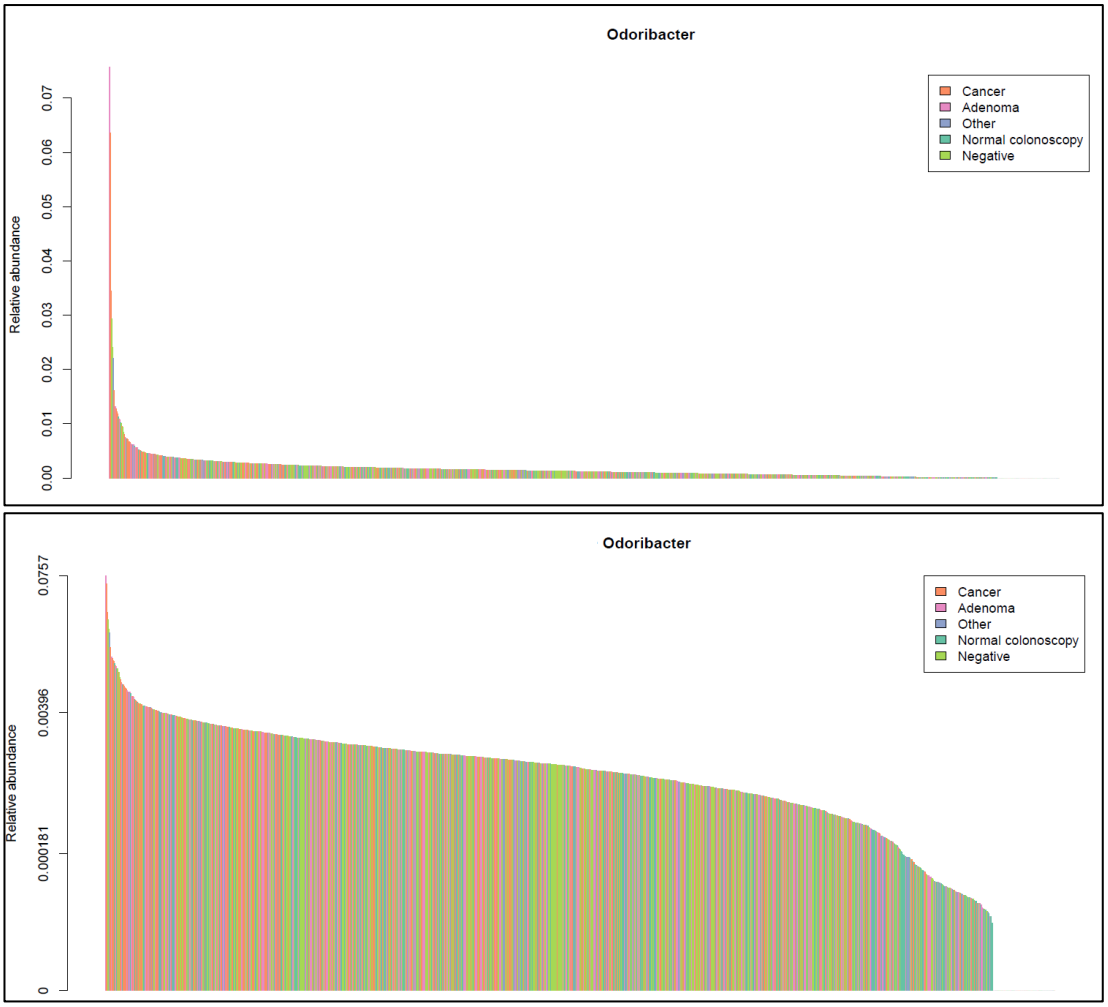


Figure 123. Waterfall plots of the relative abundance of *Odoribacter* for NHSBCSP samples. The upper plot has a normal axis; the lower plot has a logarithmic axis to enable visualisation of low-abundance samples.

Given the wide range of relative abundances of *Escherichia-Shigella* described in Chapter 2, this was investigated in the NHSBCSP cohort. Figure 124 demonstrates a wide range of *Escherichia-Shigella* relative abundance within NHSBCSP samples (0-0.95).

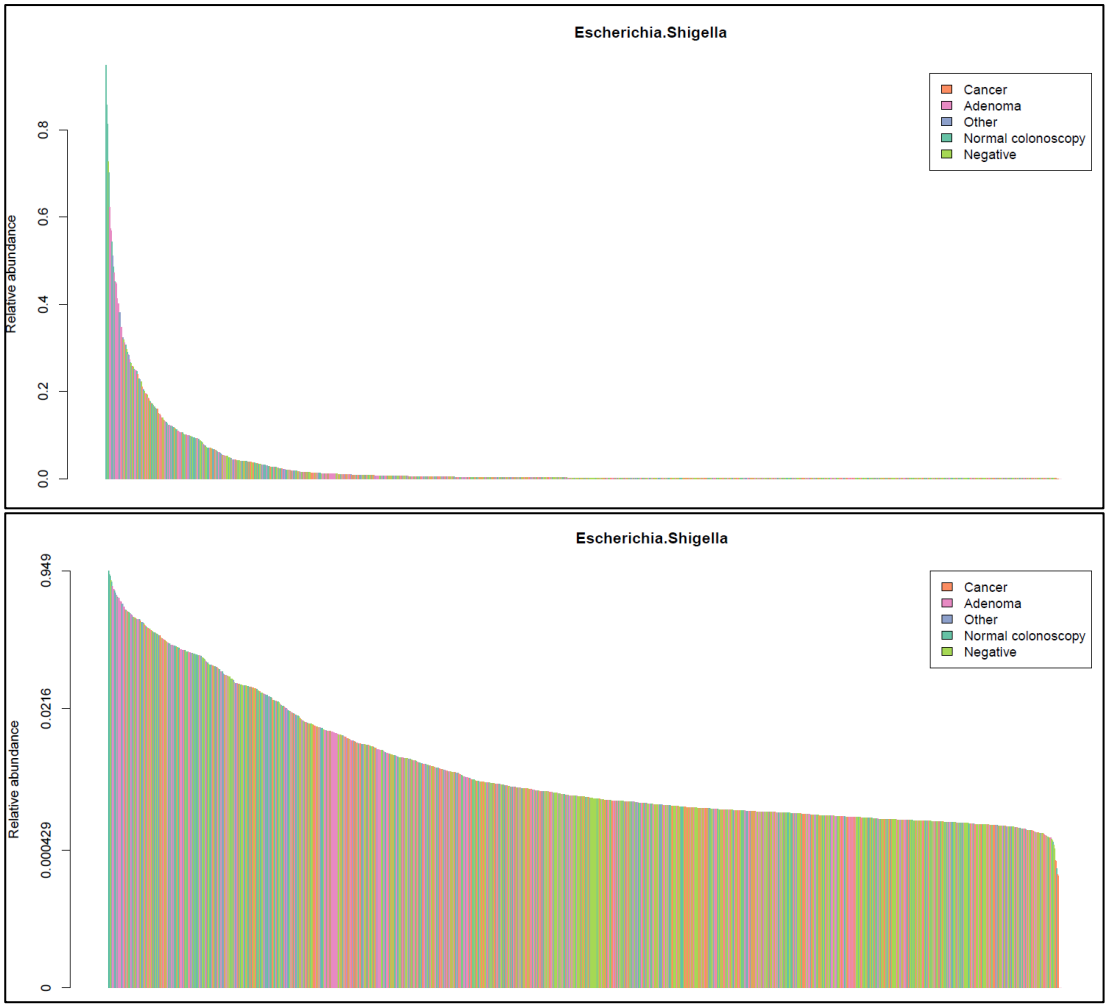


Figure 124. Waterfall plots of the relative abundance of *Escherichia-Shigella* for NHSBCSP samples. The upper plot has a normal axis; the lower plot has a logarithmic axis to enable visualisation of low-abundance samples.

Interestingly the range of relative abundance of *Fusobacterium* was also noted to be wide, with one colonoscopy-normal sample containing 44%. In order to investigate the consistency of *Fusobacterium* relative abundance, DNA from samples with both high and low *Fusobacterium* relative abundances was re-processed and sequenced on a second NGS run (Figure 125). A Bland-Altman plot confirmed good agreement (a lack of bias and minimal variation) between measurements (Figure 126).

For the sample with a relative abundance of *Fusobacterium* of 40%, re-processing and sequencing of the extracted DNA produced a relative abundance of *Fusobacterium* of 37%. Extraction and sequencing of DNA from the alternate three squares of the gFOBT card produced a relative abundance of *Fusobacterium* of 33%. This confirms that the relative abundance of *Fusobacterium* for that gFOBT sample is reproducible; such a high relative

abundance could either be biological or it could be due to a technical factor (such as overgrowth).

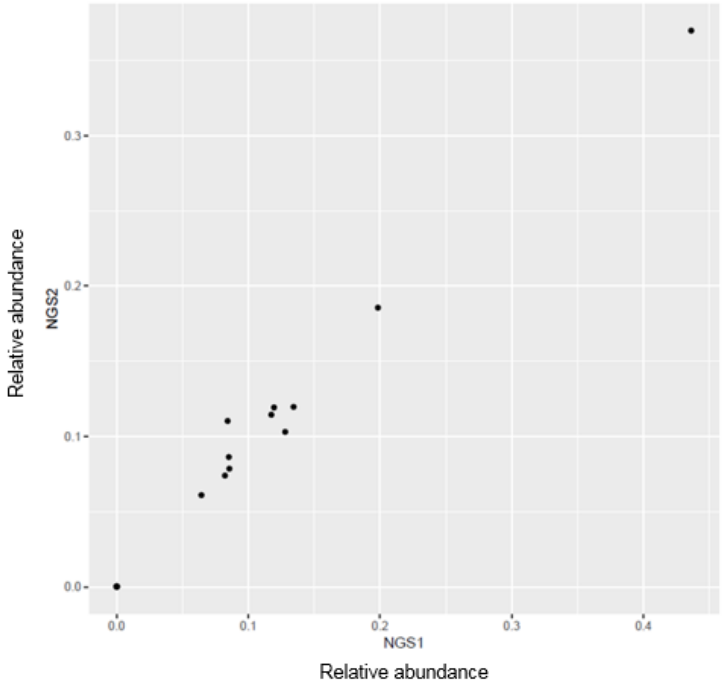


Figure 125. Scatter-plot showing the relative abundance of *Fusobacterium* for samples which were re-processed and sequenced on two sequencing runs.

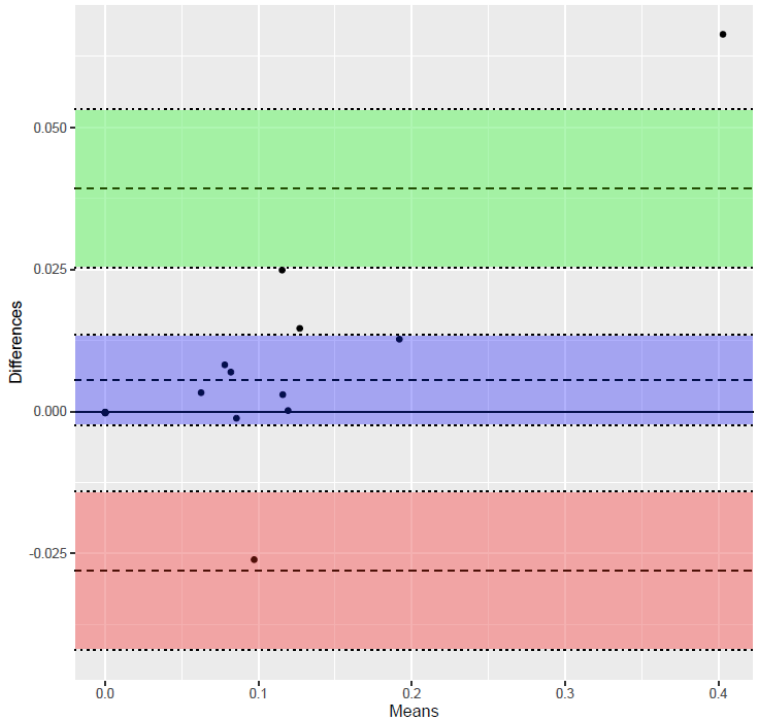


Figure 126. Bland-Altman plot of the relative abundance of *Fusobacterium* for samples which were re-processed and sequenced on two sequencing runs. The x axis shows the mean of *Fusobacterium* relative abundance across both NGS runs. The y axis shows the difference in *Fusobacterium* relative abundance across both NGS runs. The purple band represents the mean difference (i.e. bias) (plus 95% CI) in *Fusobacterium* relative abundance between NGS run 1 and NGS run 2. The green and red bands represent the upper and lower 95% limits of agreement plus 95% CI. The upper and lower limits of agreement are calculated as mean difference +/- 1.96 SD; they are expected to contain 95% of the differences measured between both methods.

3.6 Discussion

3.6.1 Investigating the microbiome of NHSBCSP samples

This is the first study to investigate the microbiome of NHSBCSP samples. To date many studies investigating the CRC-associated microbiome have been limited to small numbers of participants (which limits power); or pooling results from separate cohorts (which introduces technical bias); and collecting samples post-bowel preparation (which changes the microbiome). This study overcame these limitations by investigating the microbiome of NHSBCSP samples, representing a large bowel-preparation naïve cohort with colonoscopy-confirmed diagnosis.

3.6.1.1 Small differences in alpha diversity are identified between clinical groups

Low alpha diversity (as measured by the Shannon diversity index) of faecal samples has been associated with a number of diseases (604), however an association with adenoma or CRC has not been clearly identified in the literature, with meta-analyses of 16SrRNA and metagenomic faecal studies revealing marked inter-study heterogeneity (121, 166, 477, 496).

In the current study, alpha diversity (as measured by the Shannon diversity index) was assessed for the different clinical groups of NHSBCSP samples. Significant pairwise differences in alpha diversity were detected between all groups apart from: the non-neoplastic ‘other’ group compared with the adenoma group; the non-neoplastic ‘other’ group compared with the colonoscopy-normal group; and the CRC group compared with the blood-negative group. CRC samples and blood-negative samples had a higher average alpha diversity compared with the other groups; colonoscopy-normal samples had a lower average alpha diversity. However, the range of alpha diversities within each group was wide and the differences between the group medians small. These findings, together with the existing literature, suggest that alpha diversity (as measured by the Shannon diversity index) has no value as a CRC-discriminatory marker within the NHSBCSP cohort. Alternative measures of alpha diversity exist but were not performed in an effort to mitigate multiple testing.

3.6.1.2 Clinical status contributes minimally to beta diversity

Inter-individual variation accounted for the majority of variation in beta diversity (as measured by Bray-Curtis distance). This finding is in keeping with the existing literature. The PCA of Bray-Curtis distances indicated a degree of separation of samples according to clinical status, which was quantified by PERMANOVA analysis as $R^2 = 0.02$. This result is in keeping with the existing literature. The aforementioned meta-analyses of 16SrRNA and metagenomic faecal-based studies identified inconsistency as to whether studies identified significant associations between Bray-Curtis distance and clinical group or not (121, 166, 477). For comparison with the current study's results, the significant R^2 values from these meta-analyses were retrieved from the supplementary materials; they ranged from 0.006 to 0.07 (although the studies conducted PERMANOVA analysis using *individual* clinical groups, rather than combined clinical status as was performed in the current study).

Gender and age were also found to contribute significantly to variation in Bray-Curtis distances (each with $R^2 = 0.003$). As discussed in the Chapter introduction, an association between gender and the microbiome is being increasingly recognised (549, 550). However, of these two studies only one (549) reported a significant association between gender and beta diversity; the R^2 value was not available for comparison. Whilst age has been associated with differences in the microbiome, a significant association with age in the current study was unexpected, given the narrow age range of NHSBCSP screening (60-74 years). It should be noted that 2% of the NHSBCSP samples originated from participants older than the upper age limit of screening, with a maximum age of 89. Most studies which have investigated the association between age and the microbiome have compared extremes of age or studied differences *between* decades rather than *within* decades as per the current study (5, 553-558).

Importantly, neither screening episode nor time until DNA extraction significantly affected variation in microbiome community structure. Unfortunately it was not possible to assess for the effect of the season of sample collection, as this was confounded by the fact that the majority of blood-negative samples were collected as a single batch at the beginning of the study (whereas collection of blood-positive samples occurred throughout the study, in order to acquire sufficient numbers of samples with an associated diagnosis of CRC or a normal colonoscopy result). Future work will attempt to better match the timing of collection across all sample types. Analysis of the effect of season within the

individual blood-positive gFOBT groups is possible; this will be incorporated into a more comprehensive PERMANOVA analysis which will be performed once additional clinical metadata (lesion location and past screening history) becomes available.

Unexpectedly, the PCA of Bray-Curtis distances appeared to show greatest separation between the colonoscopy-normal and blood-negative groups, although each group contained a wide range of Bray-Curtis distances. Potential explanations for this finding will be discussed in the following section, which reviews the taxonomic differences between these two groups.

3.6.1.3 Differences in the relative abundance of taxa are detected between clinical groups

LEfSe analysis identified enrichment of CRC/adenoma-associated taxa described in the existing literature in CRC, adenoma and neoplasm samples. This provides reassurance that performing microbiome analysis from the faeces of processed NHSBCSP gFOBT samples stored at room temperature generates results comparable with studies which have performed microbiome analysis on whole stool samples collected and stored under controlled conditions. It also validates the findings from the existing literature in a large, bowel-preparation naïve cohort. There is in fact a degree of inconsistency between studies regarding which taxa are differentially abundant between controls and CRC/adenoma. The pooled meta-analyses results from four faecal (16SrRNA and metagenomic) meta-analyses are listed:

- Meta-analyses have indicated enrichment in CRC compared with controls of the following taxa: *Fusobacterium*, *Lachnospiraceae_UCG-010, Mogibacterium, Oscillibacter, Prevotella_7* (477); *Fusobacterium, Parvimonas, Porphyromonas, Peptostreptococcus, Enterobacteriaceae, Escherichia-Shigella* (121); *Fusobacterium, Parabacetroides distasonis, Parvimonas micra, Proteobacteria, Streptococcus anginosus, Porphyromonas* (534); *Fusobacterium nucleatum, Bacteroides fragilis, Porphyromonas asaccharolytica, Parvimonas micra, Prevotella intermedia, Alistipes finegoldii, Thermanaerovibrio acidaminovorans* (166); *Subdoligranulum, Clostridium boltae/clostridioforme, Clostridiales, Porphyromonas uenonis, Anaerotruncus, Anaerococcus obesiensis/vaginalis, Peptostreptococcaceae, Fusobacterium sp. oral taxon 370, Fusobacterium nucleatum s. vincentii, Ruminococcus torques,*

Prevotella nigrescens, Parvimonas, Fusobacterium nucleatum s. nucleatum, Porphyromonas asaccharolytica, Porphyromonas somerae, Prevotella intermedia, Hungatella hathewayi, Clostridium symbiosum, Solobacterium moorei, Dialister, Fusobacterium nucleatum s. animalis, Peptostreptococcus stomatis, Gemella morbillorum, Parvimonas micra (597); *F. nucleatum, Parvimonas Micra, Gemella morbillorum, Peptostreptococcus stomatis, Solobacterium Moorei, Clostridium symbiosum, Anaerococcus vaginalis, Porphyromonas Asaccharolytica, Prevotella intermedia, Bacteroides fagilis, Porphyromonas somerae, Anaerococcus obesiensis, Porphyromonas uenonis, Peptostreptococcus anaerobius, Streptococcus constellatus, Granulicatella adiacens* (496).

- Meta-analyses have indicated depletion in CRC compared with controls of the following taxa: *Anaerostipes, Butyricicoccus, Lachnospiraceae_UCG-001, Romboutsia, Roseburia, Ruminococcaceae_UCG-013, Ruminococcus_1, Coprococcus_2* (477); *Ruminococcus, ClostridiumXI* (121); *Ruminococcus unclassified* (534); 62 species (too many to list)(166); *Roseburia intestinalis, Gordonibacter pamelae, Bifidobacterium catenulatum* (496).

- Meta-analyses have indicated enrichment in adenoma compared with controls of the following taxa: for adenoma <1cm: *Allisonella, Barnesiella, Coprobacter, Odoribacter, Oscillibacter, Phascolarctobacterium, Veillonella* (477).; for adenoma >1cm: *Ruminococcus_torques group, Coprococcus_3* (477).

- Meta-analyses have indicated depletion in adenoma compared with controls of the following taxa: for adenoma <1cm: *Eubacterium_hallii group, Lachnospiraceae* (477).; for adenoma >1cm: *Eubacterium_eligens group, Eubacterium_xylanophilum group, Lachnospira, Lachnospiraceae_UCG-001*(477). It should be borne in mind that adenomas were not separated according to size in the current study as this information was not available.

There was an unexpected result from the current study: the enrichment of *Fusobacterium* within the ‘not neoplasm’ compared with the ‘neoplasm’ group; the ‘colonoscopy-normal’ compared with the ‘blood-negative’ group; the ‘colonoscopy-normal’ compared with the CRC group; and within the blood-

positive samples the ‘not CRC’ compared with the CRC group and the ‘not neoplasm’ compared with the neoplasm group. However, although *Fusobacterium* is widely reported in the literature as being enriched in CRC, the aforementioned meta-analyses of faecal studies demonstrate inconsistency of results between studies, with some studies reporting *Fusobacterium* enrichment in controls relative to cases (477, 534). Inconsistency of *Fusobacterium* enrichment in CRC tissue has also been reported (605, 606).

Inspection of the waterfall plot of *Fusobacterium* relative abundance revealed that one colonoscopy-normal sample had a high *Fusobacterium* relative abundance of 40%. Re-processing and sequencing of the extracted DNA from that sample and extraction of DNA from the alternate three squares of the gFOBT card produced similar results (relative abundances of 37% and 33% respectively). If confirmed by qPCR (a topic for future work), this finding could be biologically genuine or technical e.g. due to contamination, overgrowth affecting the whole of the gFOBT card or subsampling (as has been proposed to account for the extreme variability in the relative abundance of *Escherichia-Shigella* described in Chapter 2). One study has reported an extremely high relative abundance (92%) of a rare taxon *(Borkfalki ceftriaxensis*) following antibiotic use (571). Another study has described a ‘bloom’ of *F. nucleatum* following antibiotic therapy, although relative abundance data was not provided (569).

The differences between the blood-negative and colonoscopy-normal groups identified by beta diversity and LEfSe analysis present an interesting finding and one which has not been described in the literature, as the majority of microbiome studies use only a single control group which has not undergone gFOBT testing (either colonoscopy-normal controls *or* healthy volunteers). A number of possible explanations could account for the difference:

- The first is that methodological differences between the groups could account for differences in the microbiome. As already mentioned, the majority of the blood-negative cards were collected within a short period of time at the beginning of the study, whereas the colonoscopy-normal cards were collected throughout the study (as this group is less prevalent and therefore it took longer to amass adequate numbers of samples). Future work will address this limitation by matching the timing of collection of different sample types.

- The second is that as the blood-negative group had not undergone definitive colonoscopy, it could include participants with undiagnosed lesions. As described in the Chapter's introduction, the likelihood of missed CRC and intermediate/high-risk adenomas within this group is low (542); however it is difficult to estimate the likelihood of low-risk adenomas and non-neoplastic 'other' lesions.

- The third is that the colonoscopy-normal group may not be entirely 'normal'/healthy. As described in the Chapter's introduction, this group could contain lesions that were missed by colonoscopy, although the likelihood of missed CRC or intermediate/high-risk adenoma within this group is low (542). Furthermore, despite a macroscopically normal colon, this group of participants has faecal blood from some source, which may be associated with ill health; research has shown that a blood-positive gFOBT result associates with increased mortality from non-CRC causes (although the authors did not separate blood-positive samples according to their colonoscopy diagnosis) (607).

- The fourth is that the presence of faecal blood per se (i.e. in the absence of pathology) may associate with changes in the microbiome. Animal models of dietary haem and iron fortification studies have demonstrated enrichment of several of the taxa which were found to be enriched in the colonoscopy-normal group (598-602). However, these findings should be interpreted with the caveat that they represent the effect of haem or iron on the microbiome, rather than the effect of faecal blood per se.

Whatever the cause of the difference, the result is that differentially abundant taxa between CRC/adenoma and control are affected by the choice of control group (blood-negative or colonoscopy-normal). In both comparisons *Odoribacter* and *Porphyromonas* were found to be enriched in CRC compared with control and 'unknown *Enterobacteriaceae'* depleted; nine taxa showed an inverse association with CRC between the two comparisons. Future work will explore this further, by creating a microbiome-based Random Forest model to distinguish blood-negative from colonoscopy-normal samples, in order to determine which bacteria contribute most to the model. An advantage of the current study is that the Random Forest models generated included: a comparison between CRC/neoplasm and a 'not CRC'/'not neoplasm' control group which contained *both* blood-negative and colonoscopy-normal samples, reflective of the full range

of NHSBCSP non-neoplastic samples; and within the blood-positive samples a comparison between CRC/neoplasm and a ‘not CRC’/’not neoplasm’ control group, reflective of the types of sample which a second-tier screening test would encounter.

3.6.2 Microbiome-based screening models improve the accuracy of screening

The NHSBCSP is replacing the gFOBT test with the FIT test in order to improve screening accuracy. Research indicates that microbiome-based models could potentially improve the accuracy of screening further; however, these studies suffered a number of limitations. The current study tested the hypothesis that microbiome-based models could improve screening accuracy, using a cohort that was representative of the NHSBCSP screening population and a methodology that could be translated to the NHSBCSP.

Random Forest models were built using the relative abundance of bacteria at genus level (as opposed to individual OTUs) so that the output would be more biologically meaningful, generalisable to independently processed samples and more readily translated to the design of qPCR primers. Two meta-analyses of faecal studies have demonstrated that OTU and genera-based Random Forest models for CRC detection perform equivalently (121, 477). The relative abundances of *all* taxa were made available to the Random Forest models as opposed to only those taxa identified as differentially abundant by LEfSe; a meta-analysis of 16S faecal studies demonstrated that for the majority of studies, the accuracy of models built using relative abundances of all taxa was superior to those built using relative abundances of CRC-associated taxa alone (121).

The first two Random Forest models were designed to discriminate CRC from non-CRC and neoplasm from non-neoplasm. As previously mentioned, a limitation is that participants with blood-negative gFOBT did not undergo definitive colonoscopy (as this is not part of routine screening); this group was treated by the model as non-CRC and non-neoplastic, however it may have contained undiagnosed lesions, although the likelihood of undiagnosed intermediate/high-risk adenomas or CRC is low (542). A second potential limitation is that, as previously mentioned, the majority of the blood-negative gFOBT were collected within a short period of time at the beginning of the study which may have affected the microbiome. However, given that the taxa which

were of greatest importance to the models were in concordance with those described in the literature, any effect is likely to have been minimal.

A Random Forest model which used gFOBT blood status, age and gender was used for comparison; given the gender differences between clinical groups, this model is likely to be an optimistic estimate. Although a true comparison should be made with gFOBT status alone, it is not possible to generate Random Forest models using a single variable.

The optimum models for the detection of CRC or neoplasm combined bacteria and gFOBT blood status; for the detection of CRC the AUC was 0.855 (95% CI: 0.832-0.877), sensitivity 55% and specificity 91%; and for the detection of neoplasm the AUC was 0.868 (95% CI: 0.848-0.886), sensitivity 88% and specificity 69%. These results compare favourably with microbiome-based Random Forest models described in the literature; meta-analyses of faecal studies have generated Random Forest models to discriminate CRC from controls with the following accuracies: AUC 0.75 (121); AUC 0.85 (477); AUC 0.77 (AUC 0.82 with the incorporation of clinical data)(534); AUC 0.73 (AUC 0.88 with the incorporation of clinical data)(166); AUC 0.81 (597); AUC 0.81-0.90 (496). A systematic review of faecal studies revealed AUC for predicting CRC versus control ranged from 0.68-0.95, AUC for predicting neoplasm versus control ranged from 0.59-0.94 and AUC for predicting CRC versus ‘non-CRC’ (adenoma/control) was 0.96 (537).

The Random Forest models generated by the current study still require validation; sequencing results from an independent cohort of NHSBCSP samples are currently pending. It should be noted that the current study used a Random Forest model to discriminate between CRC and ‘non-CRC’, whereas the aforementioned meta-analyses of faecal studies generated Random Forest models to discriminate between CRC and controls; the design of the current study reflects the full range of NHSBCSP samples that a screening test would be presented with. Furthermore, the current study created a Random Forest model to discriminate between neoplasm and ‘not neoplasm’ which performed with similar accuracy to the CRC model, albeit with higher sensitivity and lower specificity. The Random Forest model designed to detect neoplasm is more relevant to the aims of national screening than one which detects CRC alone; its higher sensitivity is more appropriate as a first line screening test. Unfortunately none of

the aforementioned meta-analyses of faecal studies generated neoplasm-specific Random Forest models for comparison.

Comparison with the accuracy of FIT is difficult, as the sensitivity and specificity of FIT was not determined by the NHSBCSP FIT pilot study or the Scottish Bowel Cancer Screening Programme (BCSP) FIT pilot, as only participants with a positive FIT or gFOBT result underwent colonoscopy (451, 608). Positive predictive values were reported for the studies, but cannot be calculated in the current study as they depend upon disease prevalence, which was not reflected in the study cohort used to train the models (609). Meta-analyses can be used to infer the sensitivity and specificity of FIT, although it should be noted that values are influenced by the device, reference standard, number of stool samples and haemoglobin cut-off used (610). Meta-analyses have reported the following results for FIT for the detection of CRC: sensitivity 71% (95% CI: 58-81) and specificity 94% (95% CI: 91-96) (611); sensitivity 87% (95% CI: 73-95) and specificity 93% (95% CI: 84-96) (612); sensitivity 88% (95% CI: 55-99) and specificity 91% (95% CI: 89-92) (613). Meta-analyses have reported the following results for OC-Sensor FIT for the detection of neoplasm: sensitivity 32% (95% CI: 26-38) and specificity 93% (95% CI: 89-96) (614). The following results were reported for the Cologuard test for the detection of: CRC a sensitivity of 92.3% (95% CI: 83.0-97.5); CRC and HGD a sensitivity of 83.7% (95% CI: 75.1-90.2); for non-advanced lesions a specificity of 86.6% (95% CI: 85.9-87.2); for colonoscopy-normal a specificity of 89.8% (95% CI: 88.9-90.7); for CRC an AUC 0.94; for advanced colorectal neoplasia (CRC and advanced pre-cancerous lesions) an AUC 0.73 (445). In summary, these findings suggest that the Random Forest model for the detection of CRC generated by the current study performs with reduced sensitivity but similar specificity to FIT and Cologuard; and the Random Forest model for the detection of neoplasm generated by the current study performs with an improved AUC and sensitivity but reduced specificity compared with the OC-Sensor FIT and Cologuard.

It is anticipated that the models generated by the current study may be improved further. Existing studies have demonstrated that incorporating clinical data such as gender, age, BMI or gFOBT result into microbiome-based Random Forest models improves accuracy (166, 496, 615). In the current study, a decision was made not to incorporate age or gender into the microbiome-based screening models, in order to reduce the likelihood of overfitting the models to the data, and as gender bias in the cohort reflects not only disease prevalence but also

colonoscopy uptake. The addition of age, gender, FIT haemoglobin concentration and potentially faecal-mutation, bacterial virulence-factor or toxin testing (597, 616) is something that will be explored in future work. In order to achieve adequate numbers of samples from each clinical group, samples were not stratified according to which screening episode they originated from. Once information on past screening history becomes available, the number of participants with a screening history incongruent with the current screening outcome will be determined. It is unknown to what extent the microbiome may be altered in participants with a past screening history of adenoma or CRC (one study suggests that post-treatment the microbiome may return to 'normal' but results were inconsistent for adenomas and CRC (487)); this will be assessed by performing Random Forest modelling using only samples originating from a first screening episode. Once information pertaining to lesion location and severity becomes available, the accuracy of the models for these different parameters will be assessed. Finally, the current study created a single Random Forest model to distinguish neoplasm from non-neoplasm; future work will investigate whether two separate Random Forest models (one to identify CRC and one to identify adenoma) perform better; this is unlikely as most (but not all (617)) faecal studies have indicated poor performance of models designed to distinguish adenoma from controls (121, 477, 496).

Three Random Forest models were trained using only the blood-positive gFOBT samples. Under the current NHSBCSP all participants with such a result would be referred for colonoscopy; CRC is detected at 10% of colonoscopies, adenoma at 40%, and 50% of colonoscopies reveal a normal bowel or non-neoplastic condition (501). The high number of unnecessary colonoscopies carries associated risks and strains endoscopy capacity. It was therefore hypothesised that a microbiome-based test could be used at this stage as a second-tier of screening; second-tiers of screening were used by the NHSBCSP and Scottish BCSP to triage participants with a weak blood-positive gFOBT result (618).

In all three scenarios (the prediction of CRC, neoplasm or normal-colonoscopy result), the microbiome-based model performed significantly better than a baseline Random Forest model which used age and gender. For the detection of CRC the AUC was 0.752 (95% CI: 0.714-0.787) with a sensitivity of 48% and a specificity of 86%; for the detection of neoplasm the AUC was 0.704 (95% CI: 0.669-0.738) with a sensitivity of 75% and a specificity of 53%; and for the detection of a normal-colonoscopy result the AUC was 0.729 (95% CI: 0.692-

0.767) with a sensitivity of 50% and a specificity of 79%. These models require validation, but the initial results are promising, particularly if the models could be improved further by the additional work described above.

The next phase of the project will benefit from liaising with the NHSBCSP Research Advisory Committee to determine whether the current design of the models is appropriate (for example, the adenoma group detected by the current models includes low, intermediate and high risk lesions; if endoscopy capacity is limited, restricting detection to higher risk lesions may be required) and to confirm which of the models is likely to be beneficial and should be further developed. Further development will involve the following steps: validation of the model using an independent set of samples (for which sequencing results are pending); development of qPCR primer-probes to identify the taxa of greatest importance to the model (discussions with a manufacturer have taken place); qPCR analysis of samples; Random Forest modelling using the qPCR results and the additional clinical/faecal-testing data described above; validation of the Random Forest model using an independent set of samples; analysis of the costs and logistics of implementation into the NHSBCSP. Prior to undertaking these steps with FIT NHSBCSP samples, 16SrRNA analysis will need to be repeated, as the taxa identified as important by Random Forest models trained using gFOBT samples may differ from those identified as important by Random Forest models trained using FIT samples. The advantage of having performed the current study is that should Random Forest modelling using NHSBCSP FIT samples not produce adequate accuracy, a gFOBT-based microbiome screening test could still be developed and used as an adjunct to the NHSBCSP.

It was important to perform the current study, rather than to simply use taxa identified as being discriminatory in the existing literature, as meta-analyses of faecal studies have demonstrated inter-study heterogeneity as to which taxa are of greatest importance to screening models, presumably reflective of biological and/or technical differences between studies (121, 166, 477, 496, 534, 597). It was important to perform Random Forest modelling rather than LEfSe alone, as Random Forests take into consideration the relationship between taxa; the aforementioned meta-analyses indicated that taxa which contributed highly to Random Forest models were not necessarily identified as being significantly enriched in CRC compared with controls. Reassuringly, the taxa identified as important to the Random Forest models generated by the current study included many taxa identified as CRC/adenoma-associated and as important to Random

Forest models by the aforementioned meta-analyses. Discussions with a manufacturer regarding the design of multiplex qPCR have indicated that it would be possible to design primer-probes to five taxa (plus one house-keeping taxa) and for plates pre-loaded with primer-probes to be produced in order to reduce manual workload; the number of taxa could be increased by performing more than one multiplex qPCR per sample. Meta-analyses of faecal studies have indicated that it is possible to achieve AUC of 0.8 with a small number of taxa (7, 8, 9 or 16 according to the studies) (121, 166, 496, 619).

Specificity of the qPCR-based screening tests should be confirmed by testing FIT samples from patients with other diseases. Meta-analyses have indicated that certain taxa associate with disease/health per se whereas others are more specific to a certain disease (including CRC-associated taxa) (620-622). Meta-analyses of faecal studies demonstrated that microbiome-based models to detect CRC maintained good accuracy when samples from patients with non-CRC diseases (IBD, type 2 diabetes and Parkinson's disease) were included (496, 597). Theoretically it might be possible to use the NHSBCSP samples to detect/screen for diseases other than CRC, including neoplasia elsewhere in the gastrointestinal tract and non-gastrointestinal diseases (623, 624). However questions include:

- Which diseases would it be useful to screen for within the age group 60-74?
- What would be the ethical, cost and risk-benefit considerations?
- What would be the likely performance of such a test?

As knowledge of the CRC-associated virome and mycobiome improves, it may become appropriate to investigate virus or fungi-based screening models (289, 625). However, one faecal study has demonstrated that viral and bacterial Random Forest models to predict CRC performed equivalently, that a combined model showed no improvement over a bacterial model alone and that each virus contributed less to the virus-based model than each bacterium did to the bacteria-based model (with implications for the number of viruses which would need to be screened) (278). Metabolomic analysis could also be explored using the NHSBCSP samples; however one study has shown that SCFA-profiling did not improve a genus-based Random Forest model for the detection of CRC (626).

3.6.3 Chapter Summary

- The microbiome was successfully analysed from NHSBCSP gFOBT samples; this represents a large cohort of bowel preparation-naïve individuals with confirmed colonoscopy diagnosis.

- Small but significant differences of alpha and beta diversity were identified between different clinical groups.

- CRC and adenoma-associated bacteria described in the existing literature were identified.

- Choice of control group (blood-negative or colonoscopy-normal) was found to affect the taxa identified as enriched/depleted in CRC or adenoma.

- A high relative abundance of *Fusobacterium* was identified in a small number of samples.

- Microbiome-based screening models generated using NHSBCSP samples compared favourably with those described in the existing literature and were shown to improve the accuracy of screening.

Chapter 4
Investigating the CRC-associated microbiome of non-Western countries

4.1 Introduction

The CRC-associated microbiome and its potential to improve screening accuracy was investigated in Chapter 3 using a UK cohort. In this chapter, the work is expanded to four non-Western countries (Argentina, Chile, India and Vietnam), which have varying but increasing (for all but India) CRC incidences. This chapter will confirm that gFOBT samples can be used to investigate the microbiome in these countries and will explore the microbiome of patients with CRC and healthy volunteers from these countries.

The rationale for conducting microbiome research in non-Western countries is manifold:

- The incidence of CRC is increasing in these countries as they adopt a Western lifestyle. It is important to track potential concurrent changes in the microbiome.

- It is not known whether the CRC-associated microbiome is universal or geography-specific.

- The microbiomes of Western healthy individuals may not be the ideal comparator for investigating the CRC-associated microbiome; the microbiome of non-Western healthy individuals may be a truer representation of ‘normal’.

- CRC which arises in low incidence countries represents an extreme phenotype that may yield insight into the mechanism of CRC development.

Global differences in CRC incidence and in the microbiome (including the CRC-associated microbiome) will be outlined. Global inequity of microbiome

research will be described and the establishment of a global microbiome research network to address this will be presented.

4.1.1 Global differences in the incidence of CRC

There has traditionally been a disparity in global incidence rates of CRC, being highest in Western countries (627, 628). Worryingly, incidence rates are rising in non-Western countries as they transition to higher socioeconomic status, with a disproportionate rise in young adults (629, 630). This is of major concern as these countries lack the resources and infrastructure to adequately respond. This is reflected in the fact that CRC mortality rates relative to incidence rates are proportionally higher in these countries (627, 628).

The rising incidence of CRC has been attributed to 'Westernisation' (adoption of a 'Western' lifestyle) which accompanies socioeconomic development. Epidemiological studies have demonstrated that migrant populations to Western countries have higher CRC incidence and mortality rates than their native population and that these rates increase with time since immigration (631-633). This is compelling evidence, although it should be interpreted with the following caveats borne in mind: migrants may not be representative of their native population (usually being of a higher socioeconomic status); migrant groups are culturally and ethnically diverse; and environmental factors change over time.

Changes to the microbiome have been shown to occur within the first nine months post-immigration of people from Thailand to the USA (634). The microbiomes became more similar to a Western microbiome the greater the length of time spent in the USA and in the next generation (634). It is unclear which aspects of Westernisation cause changes in the microbiome. Potential factors include (635):

- improved hygiene, sanitation and clean water supplies
- reduced family size and urban accommodation
- increased antibiotics, vaccinations, medications and dental treatment

- changes in diet
- raised BMI and lower levels of exercise
- changes in air pollution and other environmental exposures
- greater usage of plastics
- an increase in the use of hormonal contraceptives, reduction in breastfeeding and increase in caesarean sections

The rise of non-infectious (allergic, metabolic and neoplastic) diseases in Western populations has led to the hypothesis that the Western microbiome may have lost key protective/health-promoting taxa (635) relative to our ancestors and people living in non-Western countries. These temporal and global differences in the microbiome will now be discussed.

4.1.2 Temporal and global differences in the microbiome

4.1.2.1 Changes to the microbiome across evolution

Mammalian microbiomes cluster according to species and diet; humans are most similar to omnivore primates (636). The microbiome of wild apes is significantly more diverse than the human microbiome and contains certain bacterial lineages which are absent in the human microbiome (637, 638); this suggests that diversity and taxa have been lost over the course of evolution. The human microbiome contains a higher relative abundance of taxa associated with meat consumption and a lower relative abundance of taxa associated with plant digestion (638).

Interestingly these trends continue within the human species; one study demonstrated a gradient of bacterial diversity which was highest for the microbiome of wild apes, intermediate for the microbiome of people living in non-Western countries and lowest in an American cohort (637, 638). The microbiomes of the American cohort were taxonomically more different from a Malawian cohort than the Malawian cohort was from the microbiomes of Bonobo apes; and certain bacterial lineages present in the Malawian cohort were absent in the American cohort (637, 638).

4.1.2.2 The ancestral microbiome and the microbiome of hunter-gatherers

The microbiome has been analysed from 1000-year old fossilised faeces and permafrost mummies. Although the results are limited by sample degradation and contamination, the microbiome appears to resemble the modern microbiome of people living a rural lifestyle in non-Western countries (639, 640). An alternative method of attempting to study our ancestral microbiome is to study remote hunter-gatherer tribes. These tribes have a diet and lifestyle similar to Paleolithic humans. The Hazda of Tanzania (545, 641-643), the Matses of Peru (644), the Pygmy of Southwest Cameroon (31) and the Central African Republic (645), and the Amerindians of Venezuela (646) have had their microbiomes characterised. Results from the different studies are similar. Compared with people living in Western countries, the microbiomes of hunter-gatherers have increased richness and diversity (545, 641, 644-646), including strains which are different to those identified in Western microbiomes (646) and a higher percentage of uncharacterised sequences (641).

The microbiomes of the hunter-gatherers reflect their diet. The Hazda's microbiome changes seasonally, following an annual cyclic pattern (545), and differs between men and women, reflecting the difference in diet and lifestyle between the two genders (641). The hunter-gatherers have an enrichment of fibre-digesting bacteria such as *Prevotella* and a reduction in *Bacteroides* and *Bifidobacterium* (associated with meat and dairy consumption) compared with Western microbiomes (545, 641, 645, 646). The metagenome of the hunter-gatherers encodes more enzymes for plant-based carbohydrate digestion (some of the pathways differ from the Western carbohydrate-digesting pathways) (545, 642) and fewer enzymes for xenobiotic and sugar metabolism (642, 645). Differences in SCFA profiles have also been recorded (641).

The microbiome of the hunter-gatherers contains taxa which are typically regarded as pathogens (e.g. *Treponema)* (641, 644, 645), as well as a high burden of pathogenicity related genes (645) and parasites (647). Antibiotic resistance genes are present, but differ from the antibiotic resistance genes described in Western microbiomes, being less diverse, containing fewer mobile elements and resembling the antibiotic resistance genes of soil (545, 642, 646).

The microbiomes of the hunter-gatherers differ to the microbiomes of nearby rural farming or fishing communities (31, 641, 644, 648), with some studies showing that the rural communities have microbiome traits which are more similar to those of Western countries (645, 648). However, the hunter-gatherer and nearby rural groups are overall more similar to one another than either group is to Western countries (642, 645, 646).

4.1.2.3 The microbiome of people living in non-Western countries

Within non-Western countries, there exist differences in levels of urbanisation. The microbiome of people living in urban locations is more similar to the Western microbiome than people living in more rural locations; this includes lower diversity, Western-associated changes in taxonomic abundances, an increase in antibiotic resistance genes and virulence genes; and a decrease in the relative abundance of viruses and archaea (649-653). These changes have been shown to occur even on a small scale, within single villages, and at very early stages of economic development (651).

4.1.2.4 The microbiome of people living in Western countries

The microbiome of Western countries is positioned at the extreme end of the temporal and global microbiome gradient (413, 545). It has been shown to be less diverse (lower alpha diversity), yet more individualised (higher beta diversity) than the microbiome of people living in non-Western countries (30, 654-656).

The Western microbiome has different bacterial abundance profiles (654), with a higher *Firmicutes*:*Bacteroidetes* ratio (shown to correlate with latitude) (656, 657), a higher abundance of *Bacteroides, Shigella* and *Escherichia* and lower abundance of *Prevotella, Xylanibacter* and *Treponema* (5, 37, 564, 646, 655, 656, 658). Certain bacterial lineages present in non-Western microbiomes are absent (654, 655, 658). It also contains fewer novel (newly discovered) species (381, 564, 659).

The Western microbiome differs functionally from the non-Western microbiome (in some cases having different pathways for the same function) (381). It contains more bacterial genes related to secondary bile acid

production, more secondary bile acids and branched-chain fatty acids (indicative of higher levels of dietary fat and protein), fewer SCFAs and fewer bacterial genes involved in methanogenesis and hydrogen sulphide production (5, 564, 660).

Interestingly 'Western microbiome' changes can be rapidly induced. The introduction of a 'Western' diet to native Africans for two weeks has been shown to change the faecal bacterial composition, faecal metabolites (a reduction in the anti-proliferative SCFA butyrate and increase in carcinogenic secondary bile acids) and cause an increase in colonic mucosal proliferation and inflammation (661).

4.1.3 Investigating global differences in the CRC-associated microbiome

The CRC-associated microbiome has been minimally investigated in non-Western countries. This is highlighted by the fact that meta-analyses of faecal metagenomic studies which attempted to define a 'universal CRC-associated microbiome' variously pooled the data of studies conducted in the USA, Canada, Ireland, Austria, Germany, France, Spain, Italy, China and Japan (121, 166, 477, 496, 534, 597).

Another meta-analysis investigating the sensitivity and specificity of faecal *F. nucleatum* as a biomarker of CRC pooled the data of studies conducted in the USA, Spain, China, Japan and Brazil (note the Brazilian sample size was 17) (188). Of the limited number of studies of the CRC-associated microbiome conducted in non-Western countries, most have been small pilot studies and many have had methodological flaws (662-666).

Research has found differences in the CRC-associated microbiome between cohorts from different Western countries, which may suggest that the CRC-associated microbiome is not universal (97, 166). However, it is difficult to discern true biological differences from technical differences between cohorts. This emphasises the need for a global CRC-associated microbiome study conducted using a single standardised methodology.

4.1.4 Global inequity of microbiome research

Limited microbiome research is currently being conducted in non-Western countries primarily due to:

- difficulties of sample collection, storage and transport: as participants may live remotely; ambient temperature is often higher than temperate Western countries; and study participants may lack refrigeration facilities

- the expense of sample processing: including reagents and limited availability of sequencing facilities

This inequity of microbiome research needs to be addressed, as it cannot be assumed that the results of microbiome research conducted in Western cohorts will be generalisable to non-Western populations. Some evidence that this may be true comes from a study which found that the effect of geographical location within a single Chinese province on the microbiome was greater than some microbiome-based signatures of metabolic diseases (667). Associations between the microbiome and certain factors has also been shown to vary geographically; one study demonstrated that the microbiome of vegans living in Western countries had different metabolic potential to the microbiome of vegans living in non-Western countries (668). It is therefore important that the CRC-associated microbiome is investigated in both Western and non-Western populations.

A decision was made to attempt to address the global inequity of microbiome research through the establishment of a global microbiome research network.

4.1.5 The establishment of a global microbiome research network

An application was successfully made to the Academy of Medical Sciences Global Challenge Research Fund Network pilot grant scheme (GCRFNG). The aim of the project was to establish a microbiome network of interdisciplinary research teams from non-Western countries, to deliver training and resources related to microbiome research and to conduct pilot work investigating the microbiomes of patients with CRC and healthy volunteers from countries in the network. As none of the participating

Institutions were conducting microbiome research and had limited resources, a decision was made to process the stool samples and perform microbiome analysis centrally at the University of Leeds.

The countries which were selected were Argentina, Chile, India and Vietnam, owing to an existing relationship with the research teams in these countries. A range of CRC incidence rates is represented by these countries (Table 35); in all but India, incidence rates are increasing (669-673).

Table 35. A comparison of the CRC incidence and mortality rates for the network member countries. Data is taken from the 'Global Cancer Observatory: Cancer Today' report (673).

Country	CRC Incidence	CRC Mortality
	Age Standardised World Rate in 2018 (per 100 000 person-years)	
	(Annual Crude rate in 2018) (per 100 000 individuals at risk)	
India	4.4	3.4
	4.2	3.2
Vietnam	13.4	7.0
	15.3	8.4
Chile	20.7	10.2
	32.5	17.3
Argentina	25.0	12.6
	35.1	19.5
UK	32.1	11.1
	71.9	31.5

Limited microbiome research has been conducted in these countries to date (665, 674-684).

To the author's knowledge, only one study has investigated the CRC-associated microbiome in one of these countries (India), however it suffered methodological limitations including culture-based analysis and 16SrRNA analysis limited to a single patient with CRC (665).

In order to overcome the issues associated with collection of frozen whole stool samples in non-Western countries, it was decided that gFOBT samples would be assessed. As discussed in Chapter 2, only one previous study had assessed the use of gFOBT samples in a non-Western cohort (Bangladesh) and had shown good stability after four days of storage (436).

4.2 Aims

- To assess whether the microbiome is stable if gFOBT samples are stored and transported from abroad at ambient temperature.

- To assess whether the microbiome is stable if gFOBT samples, received from abroad, are stored in the UK at room temperature for different lengths of time prior to DNA extraction.

- To investigate the CRC-associated microbiome of patients from India, Chile, Argentina and Vietnam; to assess whether this differs by country and whether there are universal CRC-associated bacteria.

4.3 Methods

4.3.1 Collaborators

The research network comprised the following researchers from the following institutions (Table 36):

Table 36. Research network collaborators.

Collaborators	Institution
Professor Ramakrishnan Dr Bose	Cancer Institute, Chennai, India
Dr Nang Dr Doi	Can Tho University of Medicine and Pharmacy, Vietnam
Dr Vaccaro Dr Piñero	Italian Hospital, Buenos Aries, Argentina
Dr Melendez Mr Valladares	Universidad de los Andes, Santiago, Chile

4.3.2 Network workshop

Members of the network attended a five-day workshop at the University of Leeds, organised by the author. The workshop provided seminars and group discussions on microbiome research, barriers to conducting microbiome research in the member countries, and practicalities of the research project. Delegates undertook laboratory and bioinformatic training. Protocols, reagent lists and bioinformatic pipelines were shared. Feedback was positive; a summary report is included in the Appendix.

4.3.3 Ethical approval

Members of the research network were granted local ethical approval for the collection of samples and clinical data. The University of Leeds REC granted approval for the receipt, storage and analysis of these samples in addition to the collection, storage and analysis of UK control samples. Tissue transfer agreements were completed. Ethical approval references are as follows (Table 37):

Table 37. Ethical approval references.

Country	Ethical approval references
India	IEC/2018/01 Indian Council of Medical Research: 2018-0337
Vietnam	QD.0604
Argentina	#3507
Chile	CEC201828
UK	University of Leeds REC: MREC17-077

4.3.4 Regulations regarding gFOBT sample and developer solution transport

The regulations regarding the international transport of gFOBT samples were consulted; gFOBT samples are exempt from the infectious substances category of the World Health Organisation 'Guidance on regulations for the transport of infectious substances 2019-2020' (454). Samples were packaged according to this guidance in a leak-proof triplicate packaging system that met International Air Transport Association (IATA) shipping requirements: gFOBT cards were placed in individual sealed bags; these bags were sealed within an IATA 95kPa pressure bag (Alpha Laboratories, UK) containing two absorbent sheets (Alpha Laboratories, UK) in case of leakage; this bag was placed within a rigid outer box (Alpha Laboratories, UK).

Developer solution (Hema Screen, Immunostics, Inc.) contains ethanol solution which is classed as a Dangerous Goods (Hazard classification 3 - Flammable liquid). The University's IATA-advisor was consulted and it was confirmed that limited quantities of developer solution could be sent under the excepted quantity allowance, provided appropriate packaging, labelling and courier agreement.

4.3.5 UK control samples

4.3.5.1 Production

UK control samples were created in order to investigate the effect on microbiome results of storage abroad and international transport of samples.

Five UK healthy volunteers donated a single stool sample. Volunteers were aged between 28 and 66 and had no history of colonoscopy or antibiotic use within the preceding six months.

Each stool sample was used to make ten separate gFOBT samples. These gFOBT samples were anonymised, stored at room temperature and developed after 24 hours. Cards were developed by applying three drops of developer solution (Hema Screen, Immunostics, Inc.) to each of the six windows on the reverse of the card. Cards were left for ten minutes to dry before being stored in individual sealed bags at room temperature.

4.3.5.2 Transport and storage

This image illustrates the transport and storage of the UK control samples, details of which are described below (Figure 127):

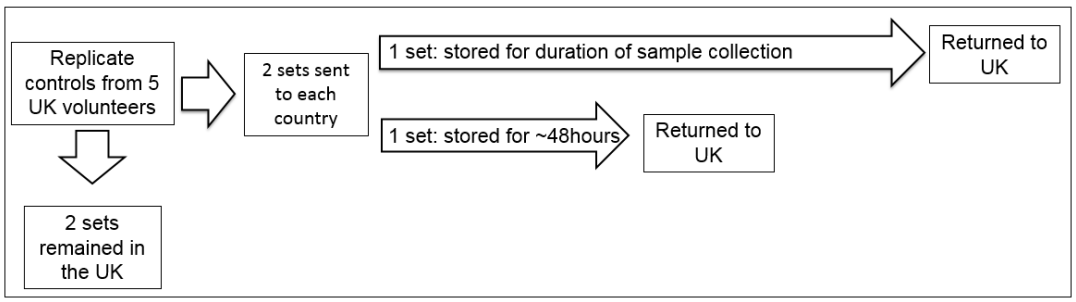


Figure 127. Transport and storage of UK control samples.

Two sets of the UK control samples (each set containing one gFOBT sample from each of the five volunteers) remained within the UK. Two sets of the UK control samples were couriered at room temperature to the abroad Institutions, where they were stored at room temperature as follows:

- One set was returned after a short period of storage in order to assess the effect on the microbiome of short-term storage and transportation.

This was intended to be 48 hours, however the samples sent to Chile and Vietnam were stored for longer (four days and eleven days respectively) due to issues arranging a courier. Samples were returned to the UK at ambient temperature (except for the samples received from Argentina, which were received with ice packs due to the fault of the courier).

- The other set of samples was stored for the duration of healthy volunteer/CRC sample collection and returned at the same time as these samples. This was in order to assess the effect on the microbiome of the storage conditions which the healthy volunteer/CRC samples had been subject to. Samples were returned to the UK at ambient temperature.

Temperature was recorded by the groups (Figure 128). Of note, the laboratory in Vietnam did not keep a temperature record but reported a maximum temperature of 25°C. The daily temperature of the Leeds laboratory was not recorded, but temperature records from the preceding six months (March-September 2019) confirm a mean temperature of 20-23°C and a maximum temperature of 27°C.

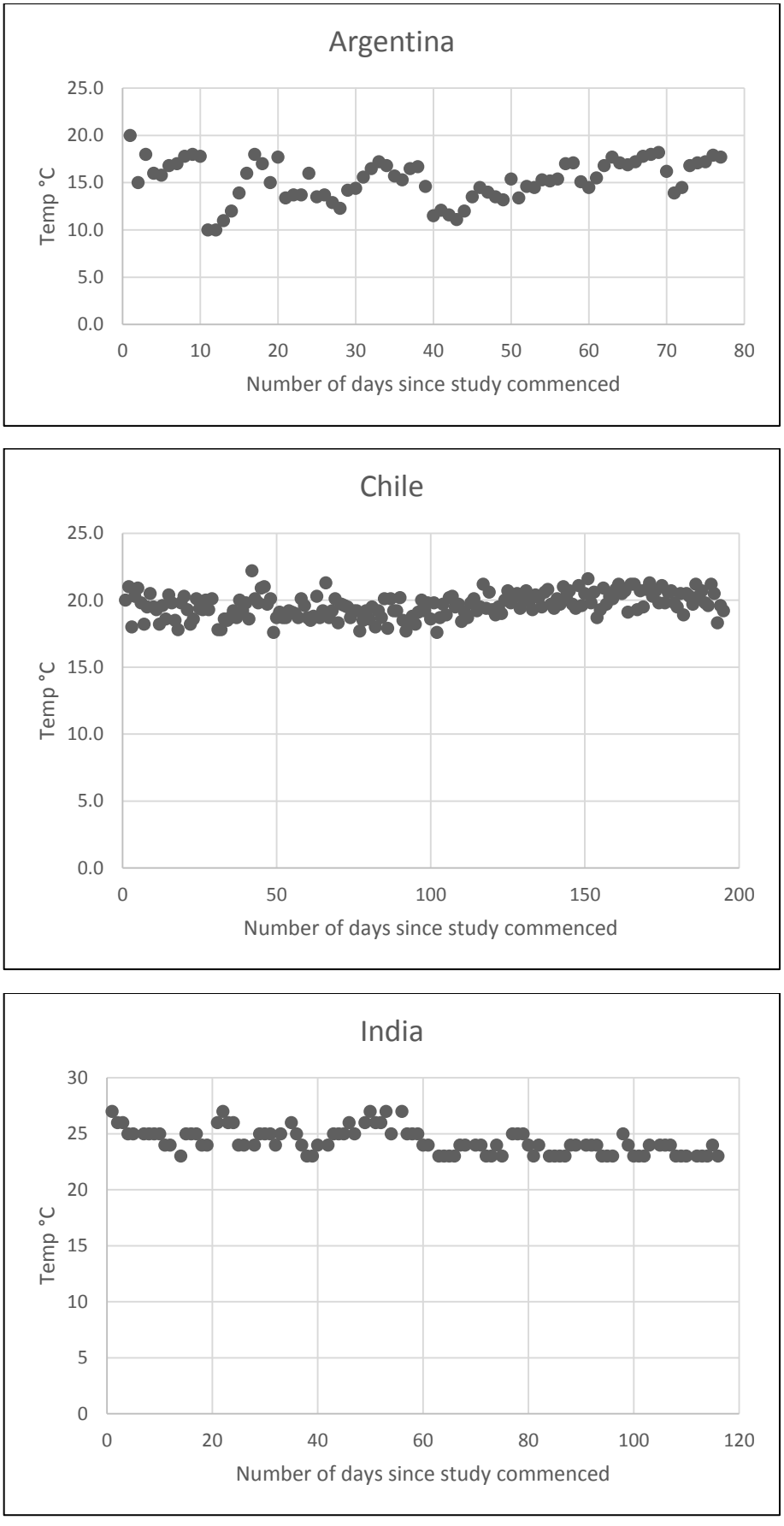


Figure 128. Laboratory temperature records.

4.3.5.3 DNA extraction

Time until DNA extraction of the control samples was as follows (Table 38):

Table 38. Time between sample creation and DNA extraction for UK control samples. The final two columns contain two values, as there were too many samples to perform DNA extraction on a single day.

Time (days):						
	Until transit	**In transit**	**Abroad**	**Return transit**	**Until DNA extraction**	**Total**
Argentina						
quick return	115	6	2	4	63 66	190 193
stored with samples			76	6	1 3	204 206
Chile						
quick return	171	8	4	3	4 7	190 193
stored with samples			196	10	2 3	387 388
India						
quick return	143	5	2	4	36 39	190 193
stored with samples			115	10	1 2	274 275

Time (days):						
	Until transit	**In transit**	**Abroad**	**Return transit**	**Until DNA extraction**	**Total**
Vietnam						
quick return	123	4	11	6	46 49	190 193
stored with samples			29	5	1 2	162 163
UK						
remained in UK	-	-	-	-	-	190 193

The time until transit (outbound) reflects the time until each Institute confirmed that they were ready to start sample collection.

All of the samples which had been stored for a short time abroad had DNA extraction performed at the same time (over two days, with samples received from all four countries being extracted on each day) to limit batch effects. Samples which had remained in the UK were also extracted at this point.

Control samples which had been stored abroad for the duration of healthy volunteer/CRC sample collection had DNA extraction performed at the same time as the healthy volunteer/CRC samples from the respective countries.

4.3.6 Healthy volunteer/CRC samples from abroad

4.3.6.1 Sample size

As this was a pilot study to assess feasibility, a power calculation was not performed. A sample size of ten healthy volunteers and ten CRC patients from each country was considered feasible within the financial constraints and timeframe of the grant.

4.3.6.2 Inclusion and exclusion criteria

Cases and healthy volunteer controls were not age/sex matched due to the preliminary nature of the study and because an extensive number of factors (besides age and gender) are postulated to affect the microbiome. Inclusion and exclusion criteria are detailed in Table 39.

Table 39. Inclusion and exclusion criteria.

Inclusion criteria
Patients with CRC
• Diagnosed with adenocarcinoma of the colon or rectum; treatment-naïve.
Healthy volunteers
• Healthy volunteers from the collaborators' research group/University. OR • People with a normal bowel at colonoscopy.
Both groups
• Aged over 18. • Capacity to give informed consent.
Exclusion criteria
• Antibiotic usage within the preceding six months. • Foreign travel within the preceding two weeks. • Colonoscopy within the preceding 14 days (potential participants could be told about the study at this time, but were asked to collect the stool sample at least 14 days after colonoscopy). • Related to another study participant. • Colostomy. • History of previous CRC or adenoma; history of previous colorectal surgery; history of pelvic radiation or chemotherapy. • Known inherited CRC syndrome (such as Lynch syndrome (according to the Amsterdam criteria) or FAP) or a family history of hereditary CRC. • Coexistent IBD or infectious bowel disease.

4.3.6.3 Sample collection and gFOBT card development

Samples from healthy volunteers and patients with CRC were collected approximately alternately to limit batch effects. Participants were asked to collect a stool sample (or gFOBT sample, depending upon participant preference) and to return it the same day to the research team. In cases where a stool sample was returned, the research team used this to prepare the gFOBT sample. gFOBT samples were developed within 24 hours of receipt by applying three drops of developer solution (Hema Screen, Immunostics, Inc.) to each of the six windows on the reverse of the card. Cards were left for ten minutes to dry before being stored in individual sealed bags at room temperature. Samples were link-anonymised.

Once sample collection was complete, samples were returned to the UK, transported at room temperature. DNA extraction occurred within three days of receipt across two days, with samples from both healthy volunteers and CRC patients processed on each day to limit batch effects.

4.3.6.4 Time between sample collection and DNA extraction

Time between sample collection and DNA extraction is depicted in Figure 129. As the time in transit and time stored in the UK prior to DNA extraction was constant for healthy volunteer and CRC samples originating from a single country, and as time in transit from each of the four countries was similar (Vietnam: five days, Argentina: six days, India and Chile: ten days), differences predominantly reflect differences in abroad storage time. Samples from each Institute were transported to the UK as a single batch in order to standardise the conditions that the samples were exposed to and to minimise courier costs.

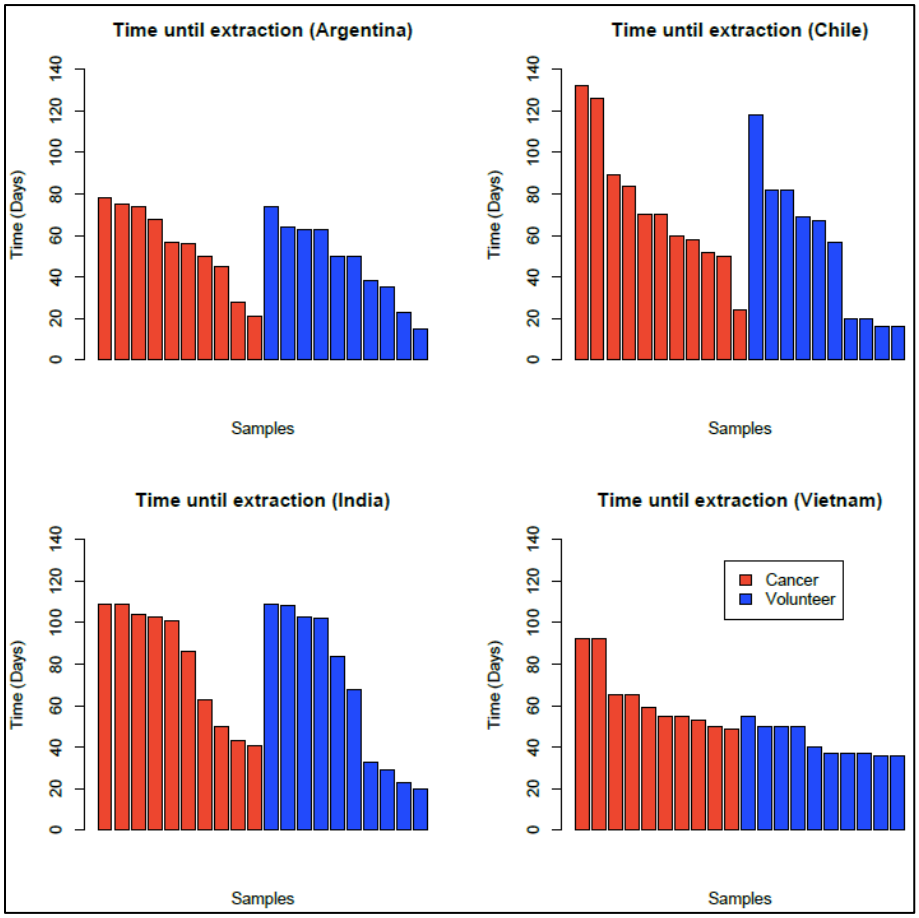


Figure 129. Time between sample collection and DNA extraction.

4.3.6.5 Clinical data

Researchers completed a questionnaire with each participant. The following information was recorded (Table 40):

Table 40. Clinical data collected by questionnaire.

Information
Age
Gender
Has the participant ever had a colonoscopy? • Date of colonoscopy • Type of bowel preparation used • Colonoscopy findings
Date of most recent antibiotic usage • Name of antibiotic • Indication • Duration of antibiotic usage
Medication use
Medical history
Smoking history
Alcohol use
Meat consumption
Family history of CRC

Clinical information was recorded for patients with CRC (Table 41). For patients who received pre-operative neo-adjuvant therapy (after gFOBT sample collection) with a significant response, information was recorded from the biopsy pathology report. Otherwise information was recorded from the final resection pathology report.

Table 41. Information recorded for CRC cases.

Information
Number of tumours
Site of tumour
Size of tumour
Histological type
Histological grade
Stage (TNM8)
Other findings at colonoscopy
Whether the patient had obstructive symptoms
Neutrophil:Lymphocyte ratio (recorded from the pre-operative Full Blood Count if available)
MMR status (if known)

4.3.6.6 Replicate samples

For a subset of samples, a set of extraction replicates were prepared to determine whether, upon receipt of the samples in the UK, prolonged storage at room temperature altered the microbiome result (equivalent to the NHSBCSP replicate experiment described in Chapter 2). Three squares (one from beneath each window) were dissected and combined to make a single sample and, after a period of storage in the UK (27 days for the samples from Chile; 140 days for the samples from India; 211 days for the samples from Argentina; 252 days for the samples from Vietnam), the alternate three squares were dissected and combined to make a replicate sample (n=23 pairs of which n=21 pairs were successfully sequenced) (Figure 130). The difference in time until DNA extraction for the replicates from the different countries reflects when samples from each country were received in the UK (Chilean sample collection started relatively late due to issues with study set-up).

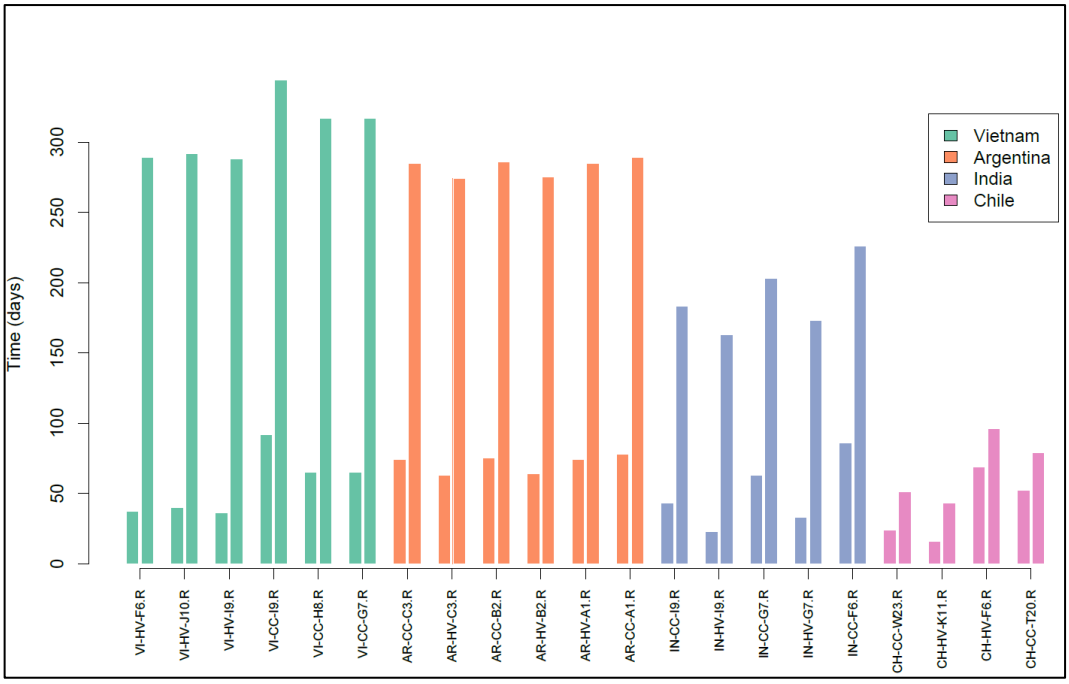


Figure 130. Time from gFOBT collection until DNA extraction for extraction replicate samples. Each pair of adjacent bars represents a pair of extraction replicates. Each bar is labelled as follows: country of origin (AR = Argentina; CH = Chile; IN = India; VI = Vietnam); disease status (CC = CRC; HV = healthy volunteer).

4.3.7 Sample processing

Details of DNA extraction, library preparation and sequencing were as described in Chapter 2. As Chilean sample collection was delayed, non-Chilean samples were sequenced first; subsequently these pools were re-sequenced together with the Chilean samples to avoid sequencing batch effects.

4.3.8 Data transfer and storage

Link-anonymised data was transferred via email and stored on secure University of Leeds servers.

4.3.9 Bioinformatic processing and statistical analysis

Bioinformatic processing and statistical analysis were performed as described in Chapter 2.

4.4 Results

4.4.1 Summary of sample processing and sequencing

The following samples were processed on the first sequencing run (Figure 131). All samples successfully underwent library preparation. All samples were successfully sequenced with more than 10,000 reads/sample, apart from one of the healthy volunteer samples from Vietnam (sample VI-HV-H8).

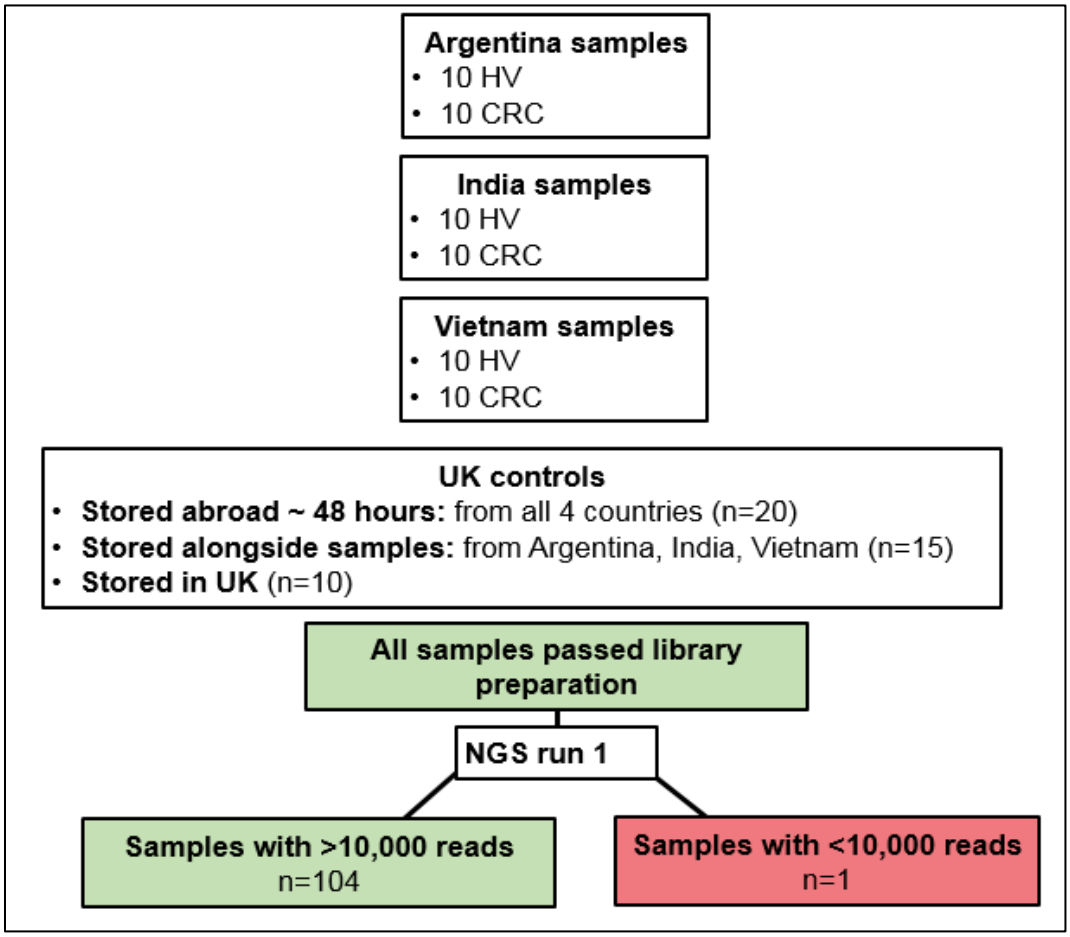


Figure 131. Workflow of samples processed on the first sequencing run.

The following samples were processed on the second sequencing run (Figure 132). Five samples failed library preparation due to an inadequate concentration of PCR amplicons. These samples have been re-processed and will be sequenced on the next available sequencing run.

The pooled amplicons from sample VI-HV-H8 which had failed the first sequencing run, also failed the second sequencing run. However, re-

processed original DNA from this sample and the extraction replicate of this sample were successfully sequenced, indicating the fault lay with the amplicon.

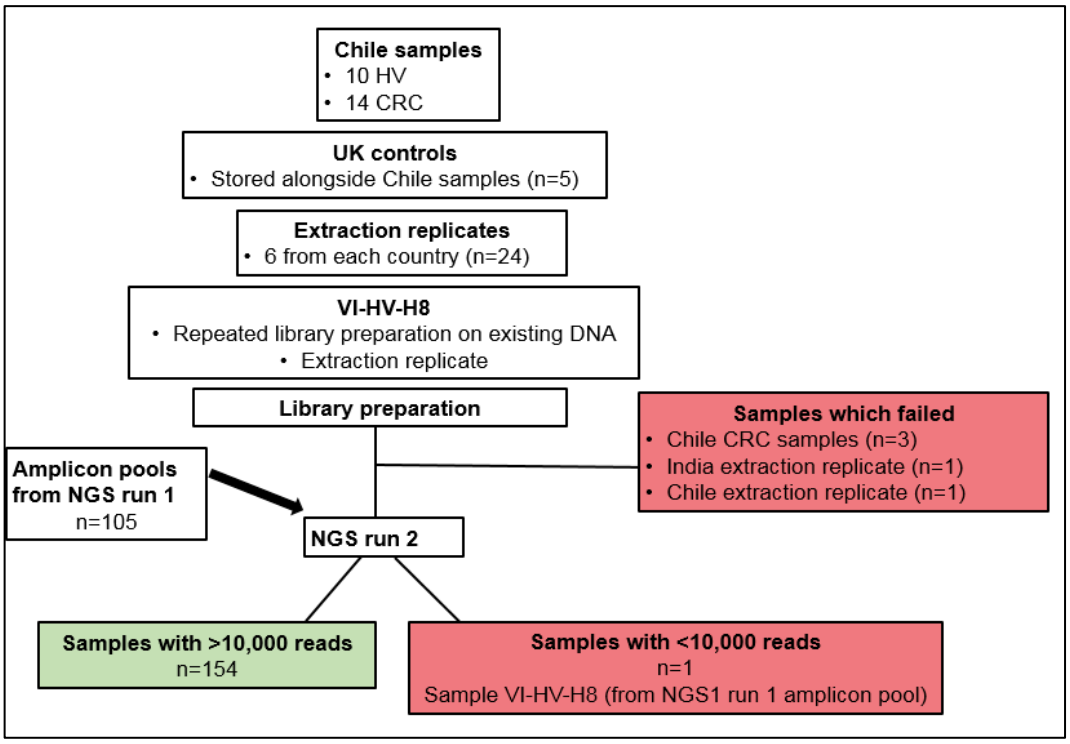


Figure 132. Workflow of samples processed on the second sequencing run.

These results indicate that V4 16SrRNA sequencing can be successfully and robustly performed on gFOBT samples received from abroad. For the remainder of this chapter, only data from the second sequencing run (which contained all of the samples) will be presented (libraries sequenced on both runs were included in the 'sequencing run-sequencing run variability' study detailed in Chapter 2).

4.4.2 Summary of sequencing data

The number of reads/sample is displayed in Figure 133. Sample VI-HV-H8 represents a clear outlier with fewer than 10,000 reads and was removed from subsequent analysis. With VI-HV-H8 removed, the range of read counts/sample was 51,000-167,000 with a median of 117,000.

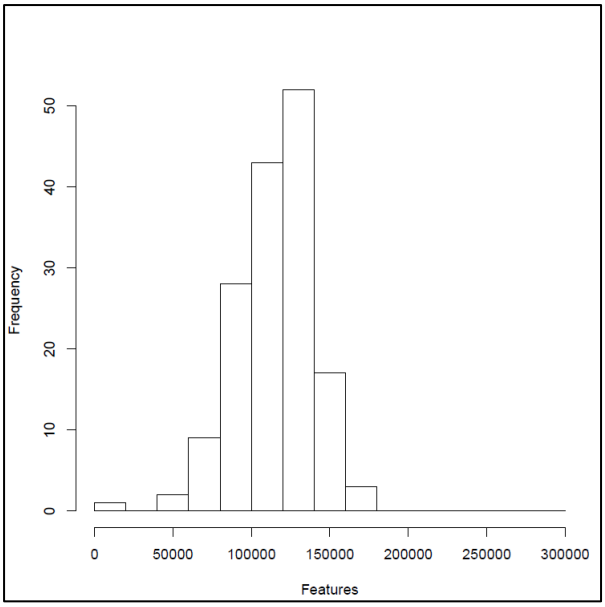
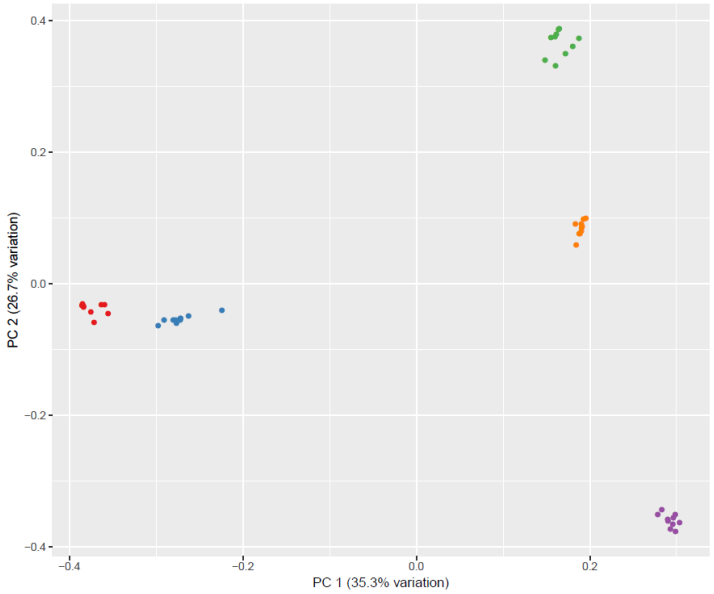


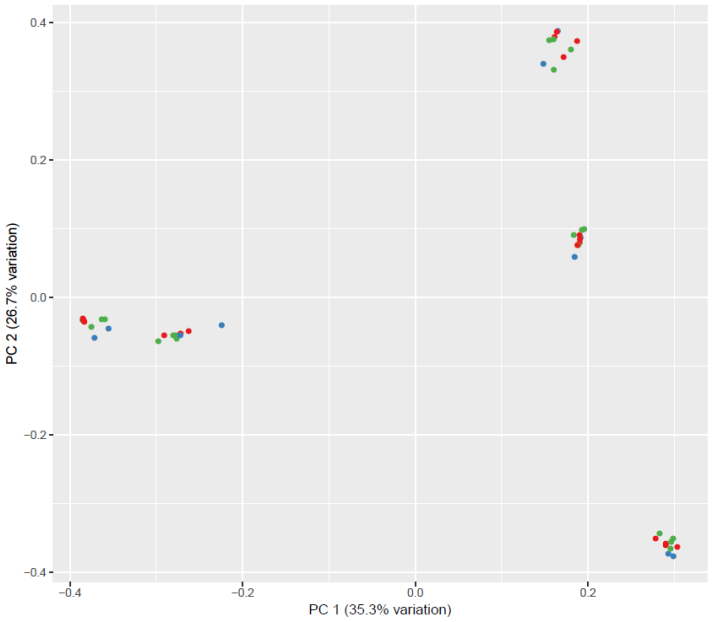
Figure 133. Number of reads/sample for the control, healthy volunteer and CRC samples sequenced on NGS run 2. The number of reads (features) is plotted on the x axis; the y axis indicates the number of samples.

4.4.3 Effects of transport and storage on microbiome results

4.4.3.1 UK control samples

The PCA of Bray-Curtis distances demonstrated that the UK control samples cluster by volunteer, despite differences in transport and storage conditions (Figure 134).


0.4
0.2
0.0
−0.2
−0.4
PC 2 (26.7% variation)
−0.4
−0.2
0.0
0.2
PC 1 (35.3% variation)
Volunteer
A1
B2
C3
D4
E5


0.4
0.2
0.0
−0.2
−0.4
PC 2 (26.7% variation)
−0.4
−0.2
0.0
0.2
PC 1 (35.3% variation)
Transport
Quick return
Remain
With

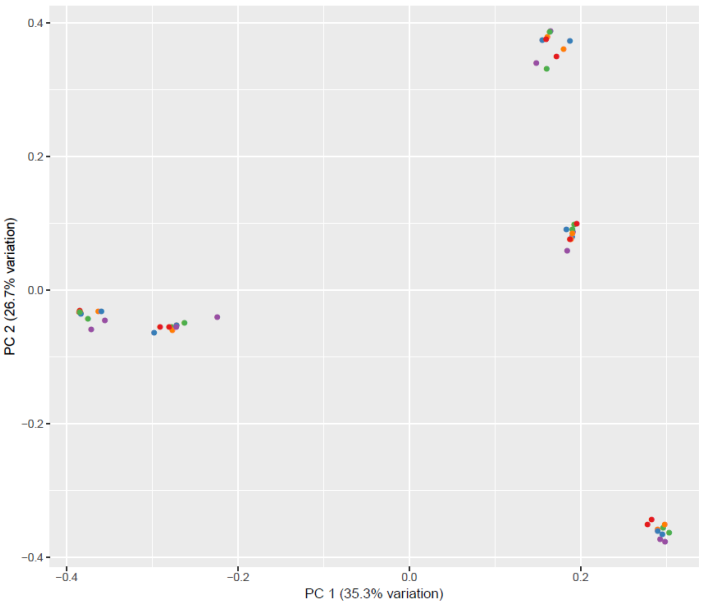

0.4
0.2
0.0
−0.2
−0.4
PC 2 (26.7% variation)
−0.4
−0.2
0.0
0.2
PC 1 (35.3% variation)
Storage
Argentina
Chile
India
UK
Vietnam

Figure 134. PCA of Bray-Curtis distances for UK control samples. Points are coloured by volunteer (upper plot), duration of storage (where 'Quick return' = samples returned after ~48 hours; 'with' = samples stored along with healthy volunteer/CRC samples; 'remain' = samples which remained in the UK) (middle plot) or by country where the sample was stored (lower plot).

The within-group Bray-Curtis distances of samples from each the UK volunteers were smaller than the within-group distances of the combined UK volunteers or healthy volunteer/CRC samples from Argentina, Chile, India or Vietnam (Figure 135). This is confirmed in Figure 136, which shows for each UK volunteer the Bray-Curtis distances between one of the samples which was stored in the UK (arbitrarily taken as a reference) and the remaining samples. Bray-Curtis distances between the reference and the alternate sample which also remained in the UK are similar to Bray-Curtis distances between the reference and the other samples. There is no apparent trend according to sample storage location or duration.

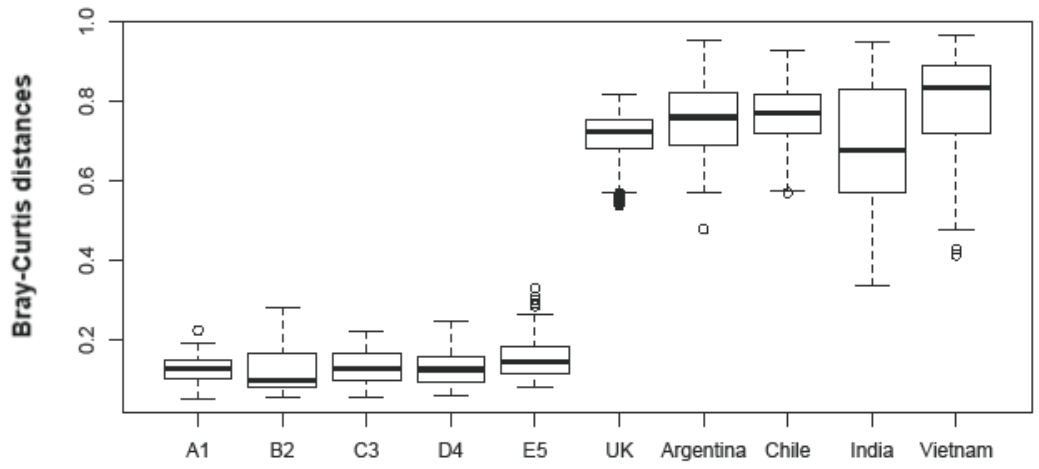
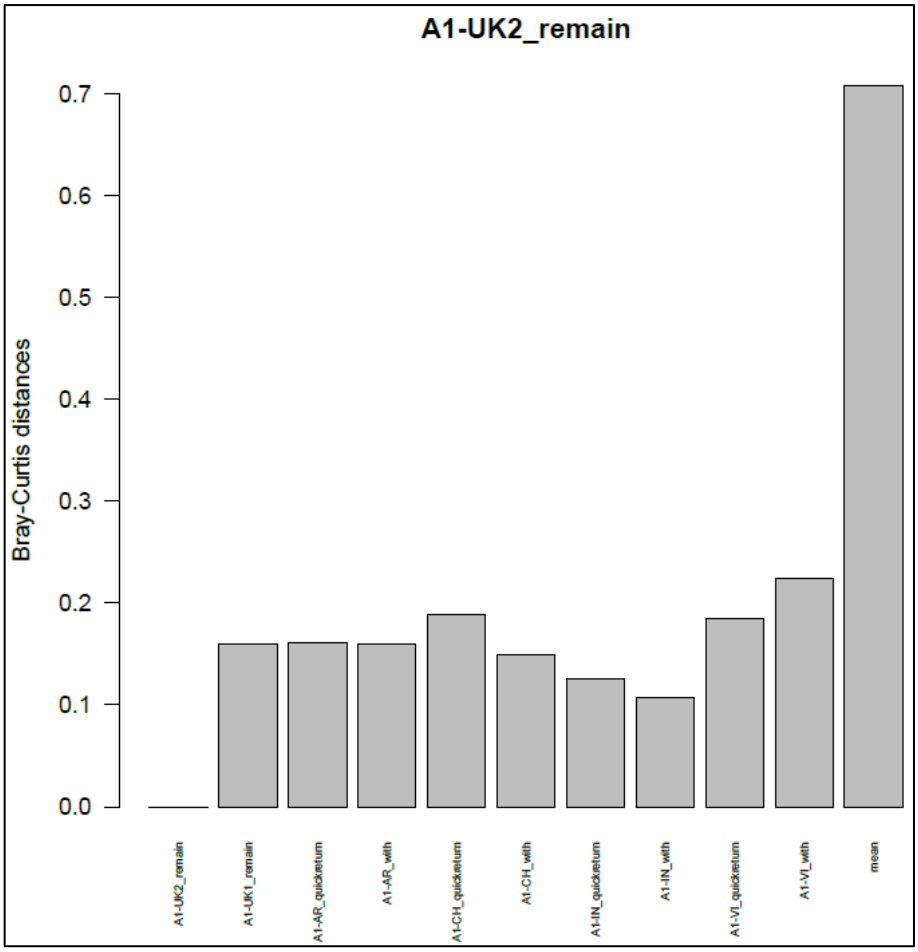
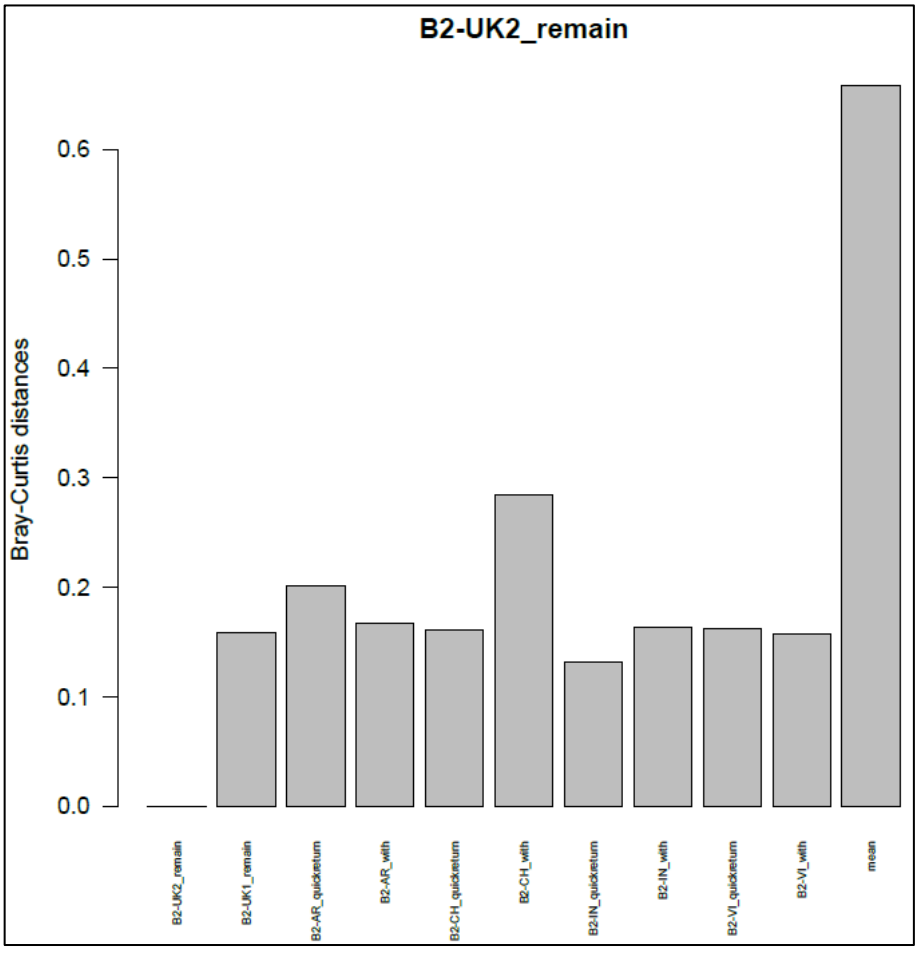
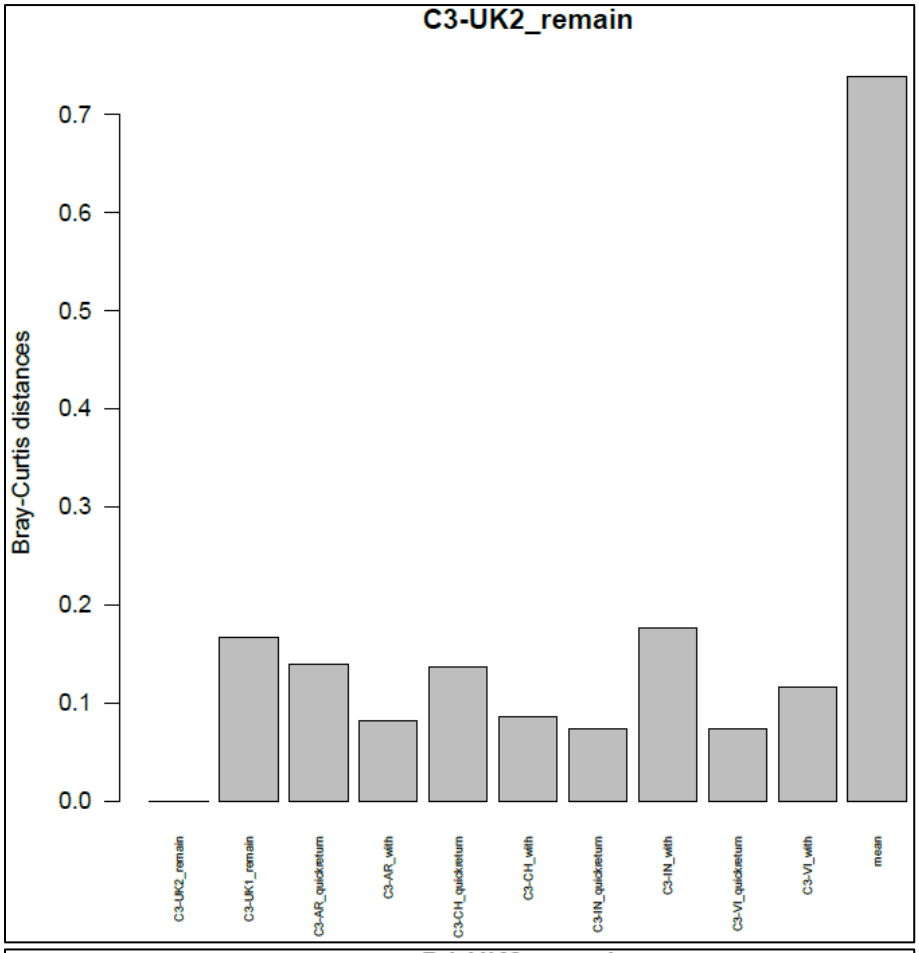
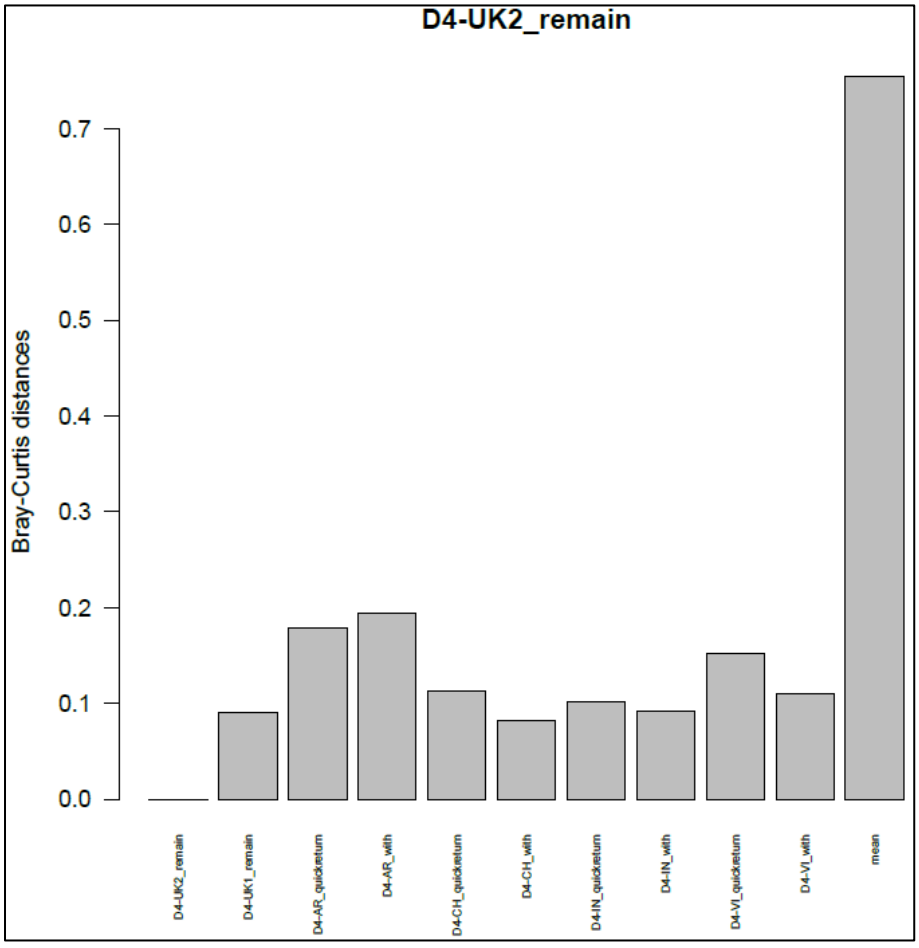


Figure 135. Boxplots of Bray-Curtis distances of UK control samples. The within-participant Bray-Curtis distances of UK control samples A1, B2, C3, D4 and E5 are shown on the left-hand side of the plot. Within-group Bray-Curtis distances of all samples from the UK, Argentina, Chile, India and Vietnam are shown on the right-hand side of the plot.


A1-UK2_remain
Bray-Curtis distances
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
A1-UK2_remain
A1-UK1_remain
A1-AR_quickstorm
A1-AR_with
A1-CH_quickstorm
A1-CH_with
A1-IN_quickstorm
A1-IN_with
A1-VI_quickstorm
A1-VI_with
mean


B2-UK2_remain
Bray-Curtis distances
0.0
0.1
0.2
0.3
0.4
0.5
0.6
B2-UK2_remain
B2-UK1_remain
B2-AR_quickstorm
B2-AR_with
B2-CH_quickstorm
B2-CH_with
B2-IN_quickstorm
B2-IN_with
B2-VI_quickstorm
B2-VI_with
mean


C3-UK2_remain
Bray-Curtis distances
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
C3-UK2_remain
C3-UK1_remain
C3-AR_quickreturn
C3-AR_with
C3-CH_quickreturn
C3-CH_with
C3-IN_quickreturn
C3-IN_with
C3-VI_quickreturn
C3-VI_with
mean


D4-UK2_remain
Bray-Curtis distances
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
D4-UK2_remain
D4-UK1_remain
D4-AR_quickreturn
D4-AR_with
D4-CH_quickreturn
D4-CH_with
D4-IN_quickreturn
D4-IN_with
D4-VI_quickreturn
D4-VI_with
mean

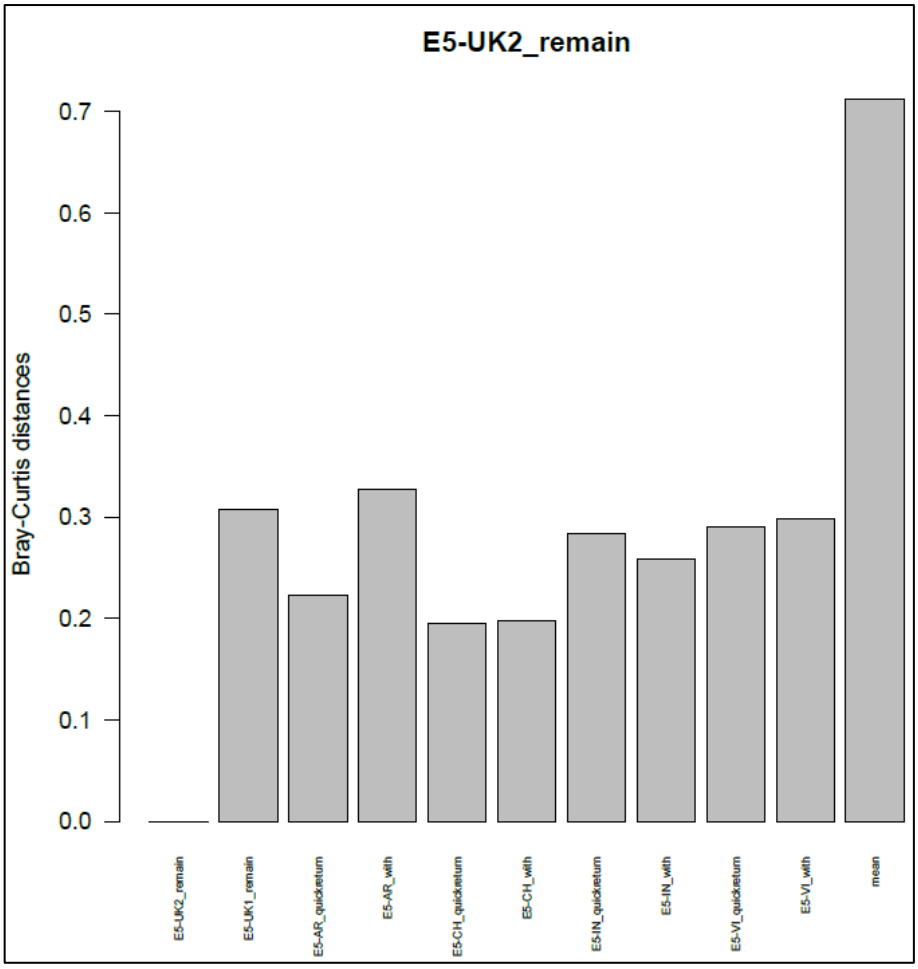


Figure 136. Bar charts of Bray-Curtis distances for UK control samples. The five plots show data corresponding to samples derived from participant A1, B2, C3, D4 and E5 (from top to bottom, respectively). Within each plot, each bar shows the Bray-Curtis distances between that sample and the reference sample (one of the samples which was stored in the UK, replicate number 2). The far right-hand bar shows the mean of the Bray-Curtis distances between the reference sample and samples derived from the other participants. Each bar is labelled as follows: participant; country where the sample was stored (AR = Argentina; CH = Chile; IN = India; VI = Vietnam; UK1/2 = UK, of which there were two sets); duration of storage (where 'quick return' = samples returned after ~48 hours; 'with' = samples stored along with healthy volunteer/CRC samples; 'remain' = samples which remained in the UK).

The taxonomic compositional similarity of replicate samples is demonstrated in Figure 137. Of interest, the microbiome composition of stools from volunteer A1 and B2 were similar, the stool of volunteer C3 contained a high relative abundance of *Escherichia-Shigella* (denoted by pale green bars in Figure 137

and shown separately in Figure 138), and the stool of volunteer D4 contained a high relative abundance of *Bifidobacterium* (denoted by red bars in Figure 137). Of note, volunteers A1 and B2 were co-habiting; volunteers C3 and D4 were co-habiting; and volunteers B2 and E5 were siblings and the offspring of volunteers C3 and D4.

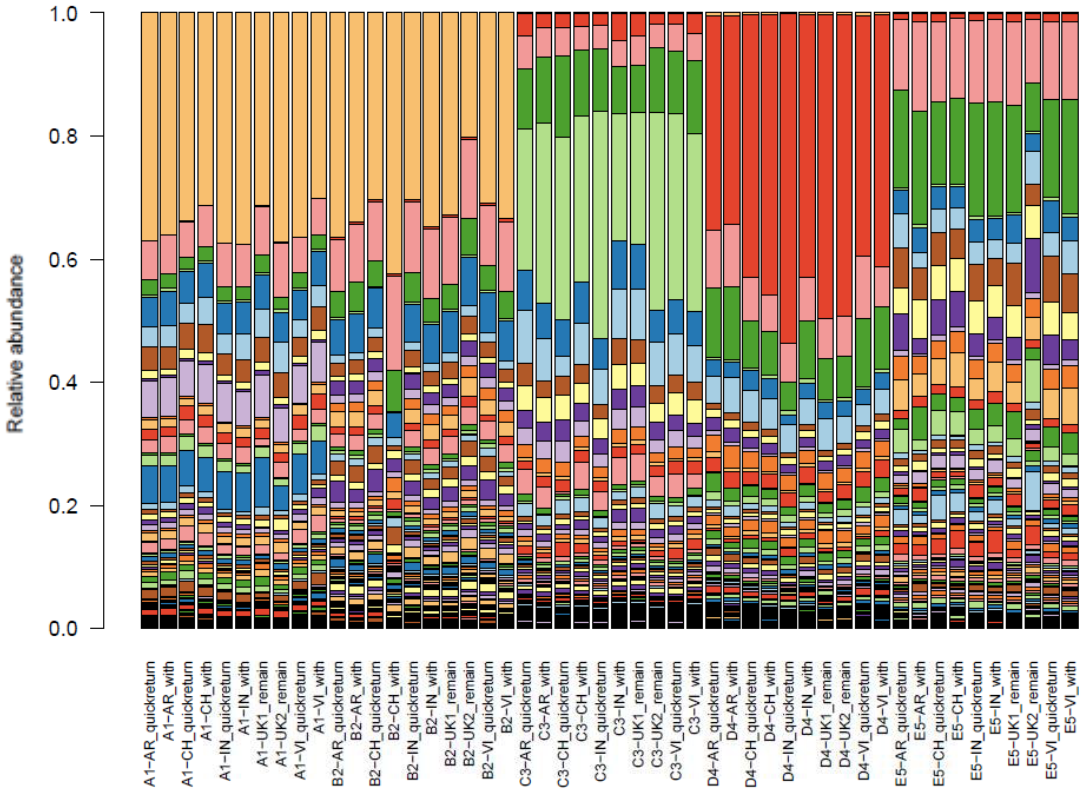


Figure 137. Taxonomy bar chart of UK control samples. Each bar represents a sample. Colours denote the relative abundance of taxa (genus level); there are too many to include a legend. Samples derived from participant A1, B2, C3, D4 and E5 are arranged from left to right respectively. Each bar is labelled as follows: participant; country where the sample was stored (AR = Argentina; CH = Chile; IN = India; VI = Vietnam; UK1/2 = UK, of which there were two sets); duration of storage (where ‘quick return’ = samples returned after ~48 hours; ‘with’ = samples stored along with healthy volunteer/CRC samples; ‘remain’ = samples which remained in the UK).

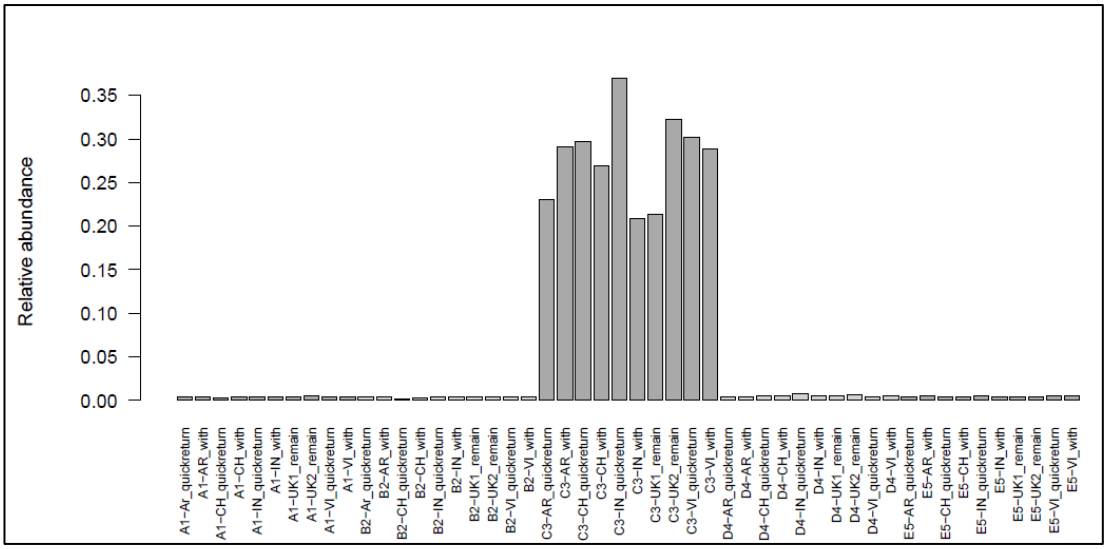
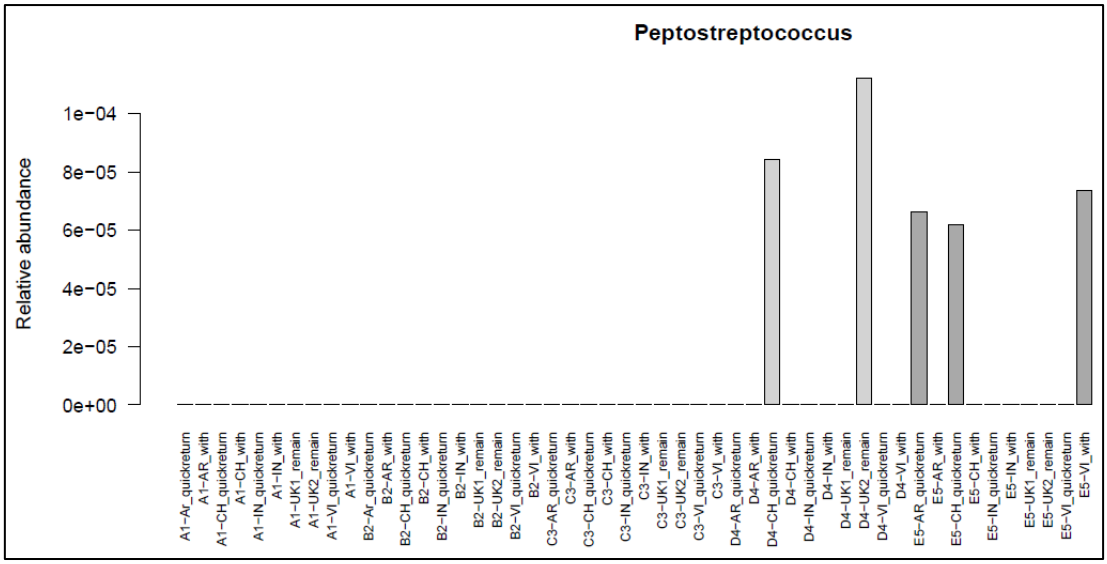
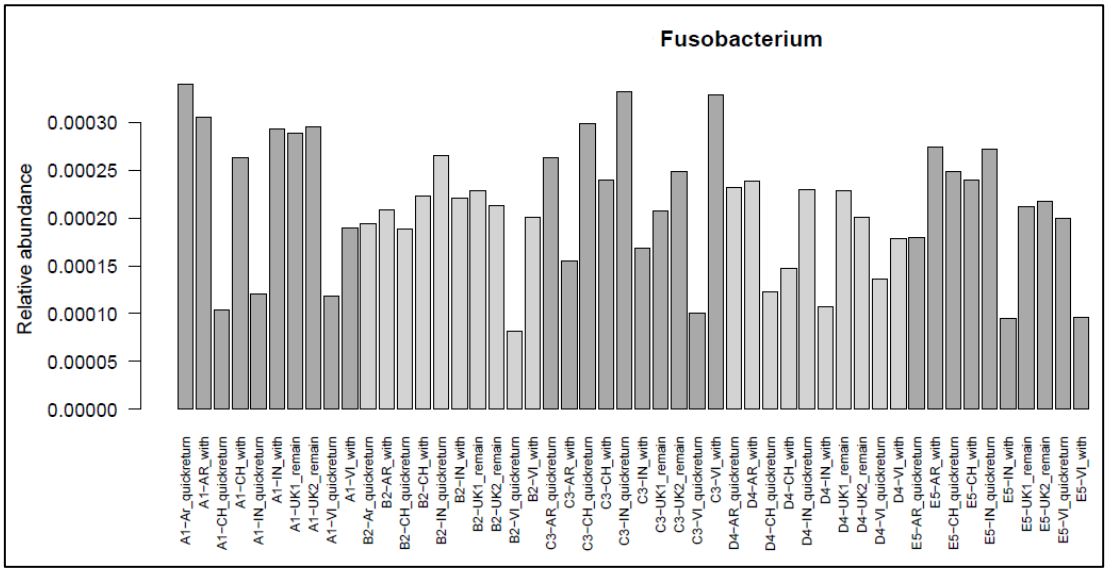
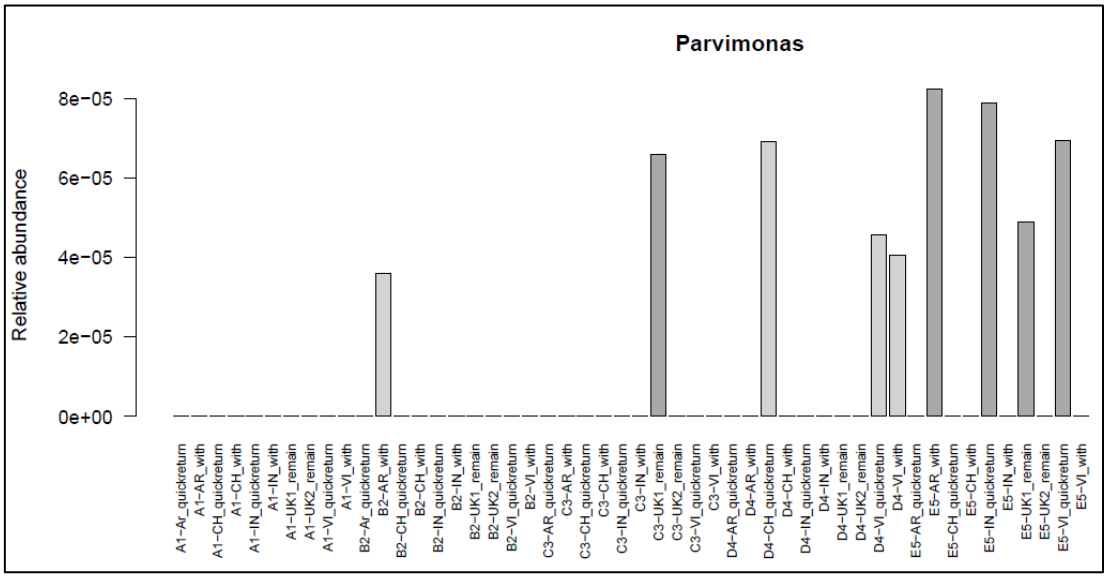


Figure 138. The relative abundance of *Escherichia-Shigella* for UK control samples. Each bar represents a sample. Samples derived from participant A1, B2, C3, D4 and E5 are arranged from left to right respectively. Each bar is labelled as follows: participant; country where the sample was stored (AR = Argentina; CH = Chile; IN = India; VI = Vietnam; UK1/2 = UK, of which there were two sets); duration of storage (where 'quick return' = samples returned after ~48 hours; 'with' = samples stored along with healthy volunteer/CRC samples; 'remain' = samples which remained in the UK).

Variability of the relative abundance of CRC-associated bacteria between replicates is demonstrated in Figure 139. The scale of the y axis of each graph should be borne in mind. Of note, there is no trend in variability with a certain country or duration of storage and some of the variability is shown by samples which remained in the UK, indicating that some of the variability may be due to technical factors such as subsampling rather than transport and storage abroad.


Peptostreptococcus
Relative abundance
1e−04
8e−05
6e−05
4e−05
2e−05
0e+00
A1−Ar_quickreturn
A1−AR_with
A1−CH_quickreturn
A1−CH_with
A1−IN_quickreturn
A1−IN_with
A1−UK1_remain
A1−UK2_remain
A1−VI_quickreturn
A1−VI_with
B2−Ar_quickreturn
B2−AR_with
B2−CH_quickreturn
B2−CH_with
B2−IN_quickreturn
B2−IN_with
B2−UK1_remain
B2−UK2_remain
B2−VI_quickreturn
B2−VI_with
C3−AR_quickreturn
C3−AR_with
C3−CH_quickreturn
C3−CH_with
C3−IN_quickreturn
C3−IN_with
C3−UK1_remain
C3−UK2_remain
C3−VI_quickreturn
C3−VI_with
D4−AR_quickreturn
D4−AR_with
D4−CH_quickreturn
D4−CH_with
D4−IN_quickreturn
D4−IN_with
D4−UK1_remain
D4−UK2_remain
D4−VI_quickreturn
D4−VI_with
E5−AR_quickreturn
E5−AR_with
E5−CH_quickreturn
E5−CH_with
E5−IN_quickreturn
E5−IN_with
E5−UK1_remain
E5−UK2_remain
E5−VI_quickreturn
E5−VI_with


Fusobacterium
Relative abundance
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
A1−Ar_quickreturn
A1−AR_with
A1−CH_quickreturn
A1−CH_with
A1−IN_quickreturn
A1−IN_with
A1−UK1_remain
A1−UK2_remain
A1−VI_quickreturn
A1−VI_with
B2−Ar_quickreturn
B2−AR_with
B2−CH_quickreturn
B2−CH_with
B2−IN_quickreturn
B2−IN_with
B2−UK1_remain
B2−UK2_remain
B2−VI_quickreturn
B2−VI_with
C3−AR_quickreturn
C3−AR_with
C3−CH_quickreturn
C3−CH_with
C3−IN_quickreturn
C3−IN_with
C3−UK1_remain
C3−UK2_remain
C3−VI_quickreturn
C3−VI_with
D4−AR_quickreturn
D4−AR_with
D4−CH_quickreturn
D4−CH_with
D4−IN_quickreturn
D4−IN_with
D4−UK1_remain
D4−UK2_remain
D4−VI_quickreturn
D4−VI_with
E5−AR_quickreturn
E5−AR_with
E5−CH_quickreturn
E5−CH_with
E5−IN_quickreturn
E5−IN_with
E5−UK1_remain
E5−UK2_remain
E5−VI_quickreturn
E5−VI_with


Parvimonas
Relative abundance
8e−05
6e−05
4e−05
2e−05
0e+00
A1−Ar_quickreturn
A1−AR_with
A1−CH_quickreturn
A1−CH_with
A1−IN_quickreturn
A1−IN_with
A1−UK1_remain
A1−UK2_remain
A1−VI_quickreturn
A1−VI_with
B2−Ar_quickreturn
B2−AR_with
B2−CH_quickreturn
B2−CH_with
B2−IN_quickreturn
B2−IN_with
B2−UK1_remain
B2−UK2_remain
B2−VI_quickreturn
B2−VI_with
C3−AR_quickreturn
C3−AR_with
C3−CH_quickreturn
C3−CH_with
C3−IN_quickreturn
C3−IN_with
C3−UK1_remain
C3−UK2_remain
C3−VI_quickreturn
C3−VI_with
D4−AR_quickreturn
D4−AR_with
D4−CH_quickreturn
D4−CH_with
D4−IN_quickreturn
D4−IN_with
D4−UK1_remain
D4−UK2_remain
D4−VI_quickreturn
D4−VI_with
E5−AR_quickreturn
E5−AR_with
E5−CH_quickreturn
E5−CH_with
E5−IN_quickreturn
E5−IN_with
E5−UK1_remain
E5−UK2_remain
E5−VI_quickreturn
E5−VI_with

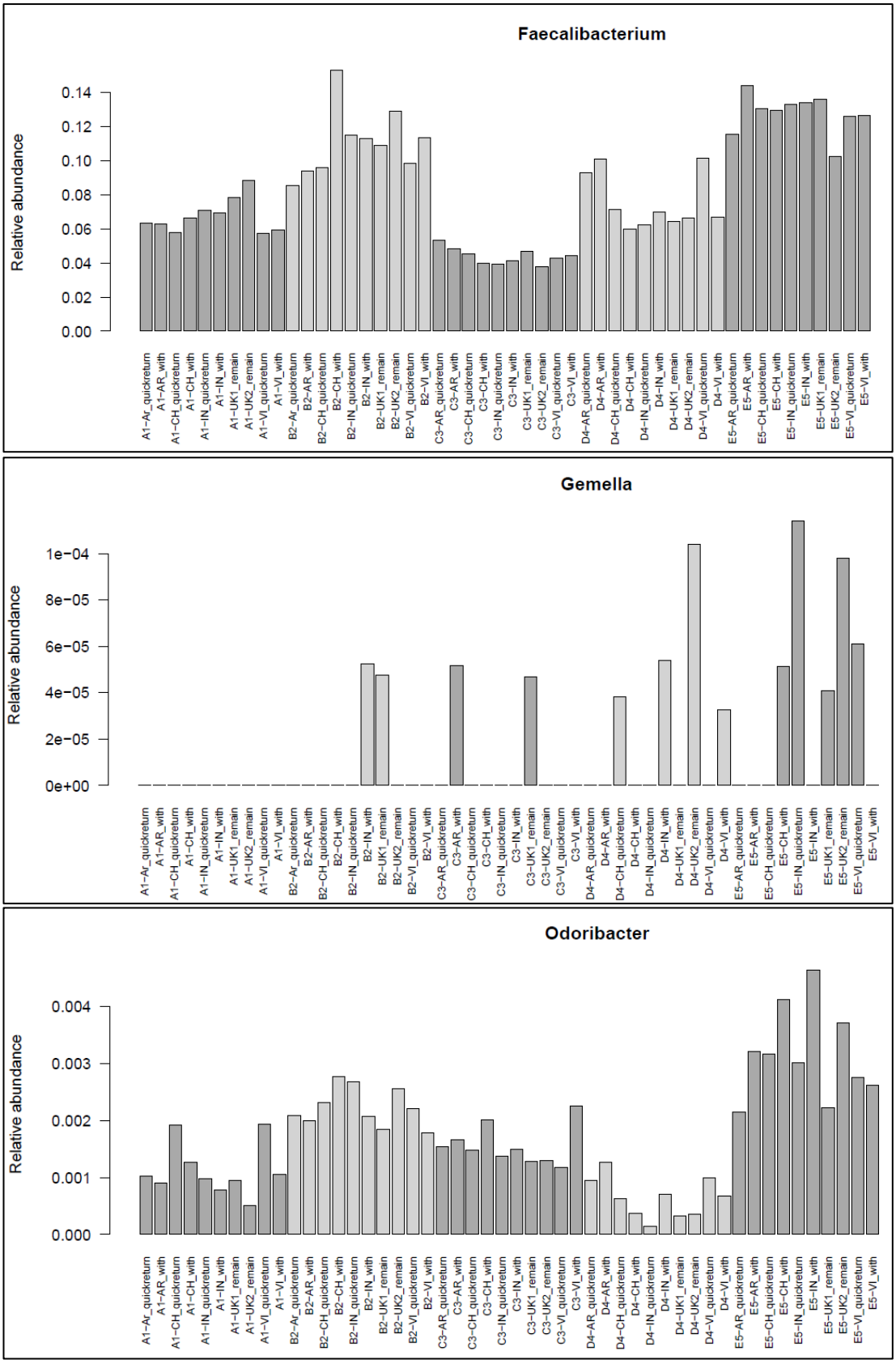

Faecalibacterium
Relative abundance
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
A1−Ar_quickreturn
A1−AR_with
A1−CH_quickreturn
A1−CH_with
A1−IN_quickreturn
A1−IN_with
A1−UK1_remain
A1−UK2_remain
A1−VI_quickreturn
A1−VI_with
B2−Ar_quickreturn
B2−AR_with
B2−CH_quickreturn
B2−CH_with
B2−IN_quickreturn
B2−IN_with
B2−UK1_remain
B2−UK2_remain
B2−VI_quickreturn
B2−VI_with
C3−AR_quickreturn
C3−AR_with
C3−CH_quickreturn
C3−CH_with
C3−IN_quickreturn
C3−IN_with
C3−UK1_remain
C3−UK2_remain
C3−VI_quickreturn
C3−VI_with
D4−AR_quickreturn
D4−AR_with
D4−CH_quickreturn
D4−CH_with
D4−IN_quickreturn
D4−IN_with
D4−UK1_remain
D4−UK2_remain
D4−VI_quickreturn
D4−VI_with
E5−AR_quickreturn
E5−AR_with
E5−CH_quickreturn
E5−CH_with
E5−IN_quickreturn
E5−IN_with
E5−UK1_remain
E5−UK2_remain
E5−VI_quickreturn
E5−VI_with
Gemella
Relative abundance
1e−04
8e−05
6e−05
4e−05
2e−05
0e+00
A1−Ar_quickreturn
A1−AR_with
A1−CH_quickreturn
A1−CH_with
A1−IN_quickreturn
A1−IN_with
A1−UK1_remain
A1−UK2_remain
A1−VI_quickreturn
A1−VI_with
B2−Ar_quickreturn
B2−AR_with
B2−CH_quickreturn
B2−CH_with
B2−IN_quickreturn
B2−IN_with
B2−UK1_remain
B2−UK2_remain
B2−VI_quickreturn
B2−VI_with
C3−AR_quickreturn
C3−AR_with
C3−CH_quickreturn
C3−CH_with
C3−IN_quickreturn
C3−IN_with
C3−UK1_remain
C3−UK2_remain
C3−VI_quickreturn
C3−VI_with
D4−AR_quickreturn
D4−AR_with
D4−CH_quickreturn
D4−CH_with
D4−IN_quickreturn
D4−IN_with
D4−UK1_remain
D4−UK2_remain
D4−VI_quickreturn
D4−VI_with
E5−AR_quickreturn
E5−AR_with
E5−CH_quickreturn
E5−CH_with
E5−IN_quickreturn
E5−IN_with
E5−UK1_remain
E5−UK2_remain
E5−VI_quickreturn
E5−VI_with
Odoribacter
Relative abundance
0.004
0.003
0.002
0.001
0.000
A1−Ar_quickreturn
A1−AR_with
A1−CH_quickreturn
A1−CH_with
A1−IN_quickreturn
A1−IN_with
A1−UK1_remain
A1−UK2_remain
A1−VI_quickreturn
A1−VI_with
B2−Ar_quickreturn
B2−AR_with
B2−CH_quickreturn
B2−CH_with
B2−IN_quickreturn
B2−IN_with
B2−UK1_remain
B2−UK2_remain
B2−VI_quickreturn
B2−VI_with
C3−AR_quickreturn
C3−AR_with
C3−CH_quickreturn
C3−CH_with
C3−IN_quickreturn
C3−IN_with
C3−UK1_remain
C3−UK2_remain
C3−VI_quickreturn
C3−VI_with
D4−AR_quickreturn
D4−AR_with
D4−CH_quickreturn
D4−CH_with
D4−IN_quickreturn
D4−IN_with
D4−UK1_remain
D4−UK2_remain
D4−VI_quickreturn
D4−VI_with
E5−AR_quickreturn
E5−AR_with
E5−CH_quickreturn
E5−CH_with
E5−IN_quickreturn
E5−IN_with
E5−UK1_remain
E5−UK2_remain
E5−VI_quickreturn
E5−VI_with

Figure 139. The relative abundance of CRC-associated taxa for UK control samples. Each bar represents a sample. Samples derived from participant A1, B2, C3, D4 and E5 are arranged from left to right respectively. Each bar is labelled as follows: participant; country where the sample was stored (AR = Argentina; CH = Chile; IN = India; VI = Vietnam; UK1/2 = UK, of which there were two sets); duration of storage (where 'quick return' = samples returned after ~48 hours; 'with' = samples stored along with healthy volunteer/CRC samples; 'remain' = samples which remained in the UK). e = x 10^.

4.4.3.2 Extraction replicates

Two extraction replicates failed library preparation and therefore these pairs were removed from analysis; this resulted in 21 pairs. Samples clustered as their replicate pairs on the PCA of Bray-Curtis distances (Figure 140). The within-pair Bray-Curtis distances were smaller than the distances between unpaired samples (Figure 141), although there was one outlier pair corresponding to sample AR-HV-C3 (Argentina, Healthy Volunteer participant C3). The majority of pairs had similar taxonomic compositions (Figure 142), although the relative abundances of taxa visibly differed between sample AR-HV-C3 and its replicate pair. As this pair of samples was stored and processed in the same manner as the other replicate pairs, it is unclear why it shows a higher degree of dissimilarity. Interestingly, Figure 144 shows very little difference between the relative abundances of CRC-associated taxa for sample AR-HV-C3 and its replicate pair.

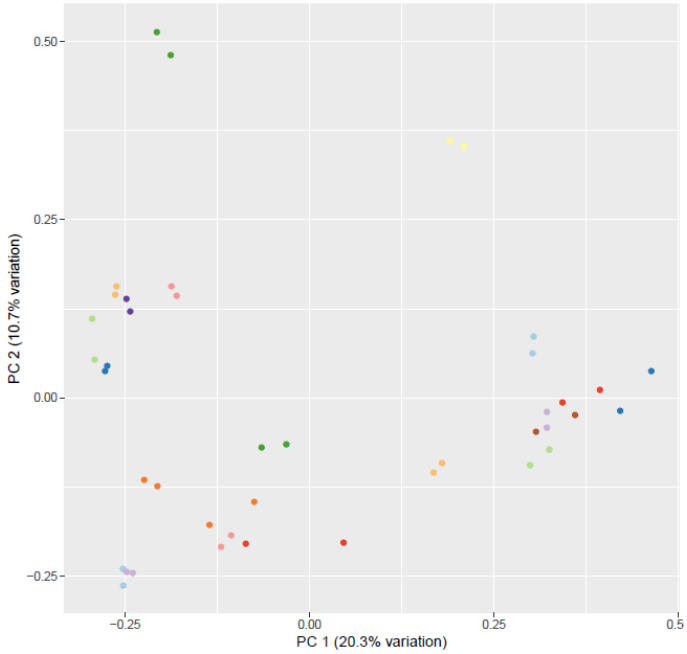


Figure 140. PCA of Bray-Curtis distances for extraction replicates. Points are coloured by gFOBT sample of origin.

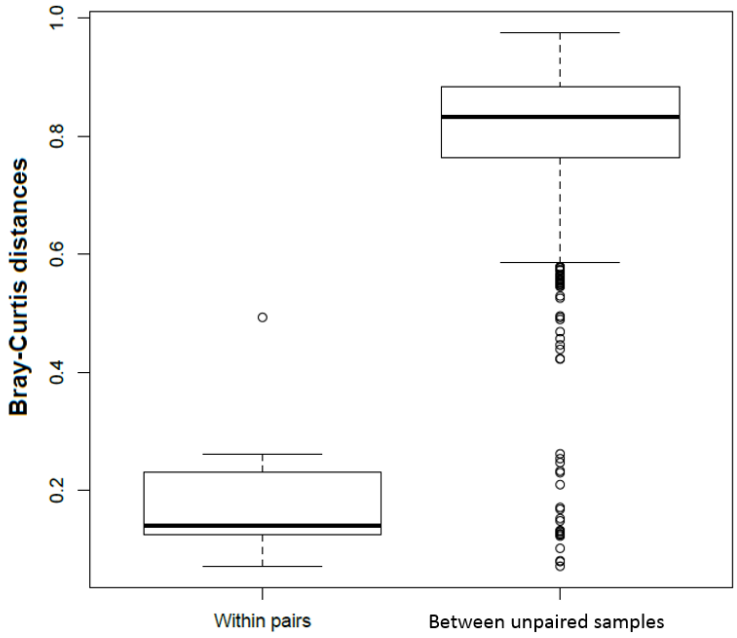


Figure 141. Boxplots of Bray-Curtis distances for extraction replicates. The within-pair Bray-Curtis distances denote Bray-Curtis distances between each sample and its extraction replicate. The between-unpaired samples Bray-Curtis distances denote Bray-Curtis distances between unrelated samples.

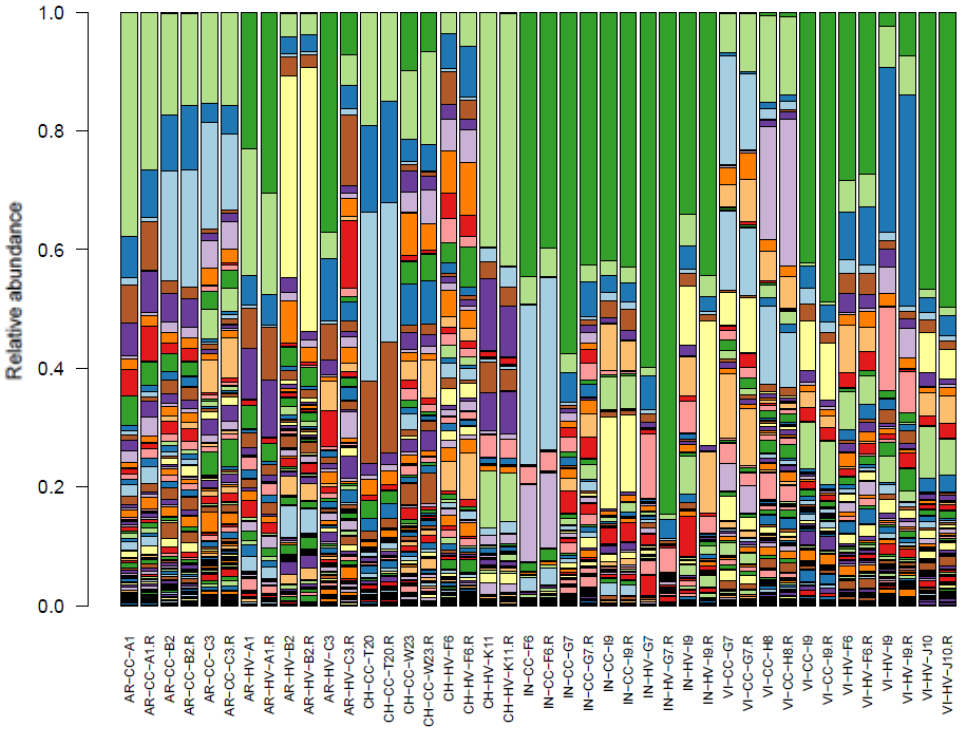


Figure 142. Taxonomy bar chart of extraction replicates. Each pair of adjacent bars represents a pair of extraction replicates. Colours denote the relative abundance of taxa (genus level); there are too many to include a legend. Samples derived from Argentina, Chile, India and Vietnam are arranged from left to right respectively. Each bar is labelled as follows: country of origin (AR = Argentina; CH = Chile; IN = India; VI = Vietnam); disease status (CC = CRC; HV = healthy volunteer); sample name; whether the sample was an extraction replicate (indicated by R).

There was minimal variability in the relative abundance of *Escherichia-Shigella* (Figure 143) and CRC-associated taxa (Figure 144) between members of each extraction replicate pair, similar to the NHSBCSP extraction replicates (Chapter 2). LEfSe analysis did not detect any significant difference in taxa between the original samples and samples whereupon DNA was extracted after a period of storage. These results indicate that storage at room temperature up to 252 days after sample receipt does not markedly alter microbiome results compared with DNA extraction within three days of sample receipt.

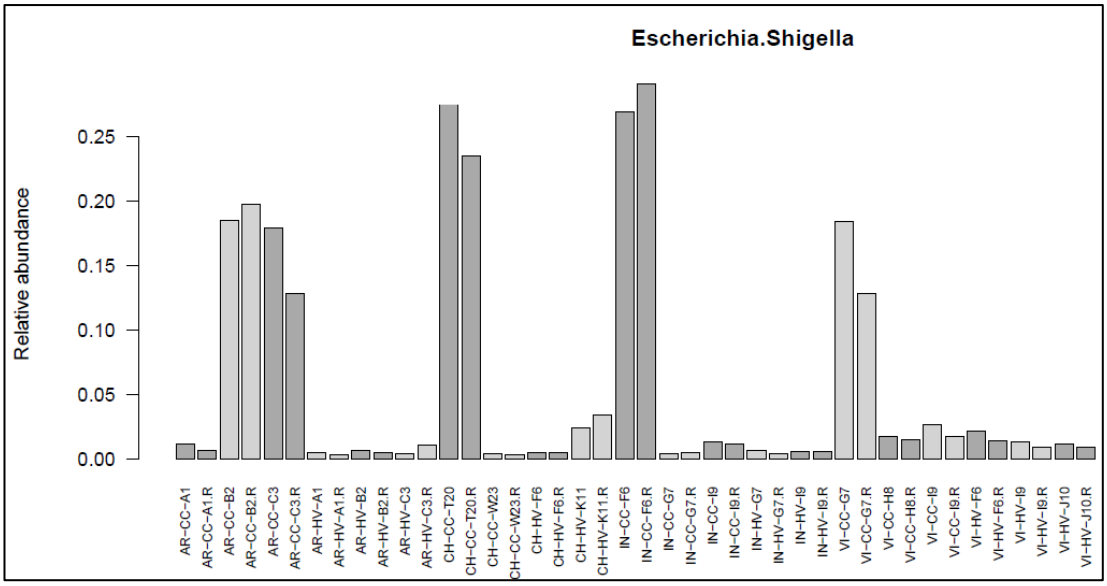
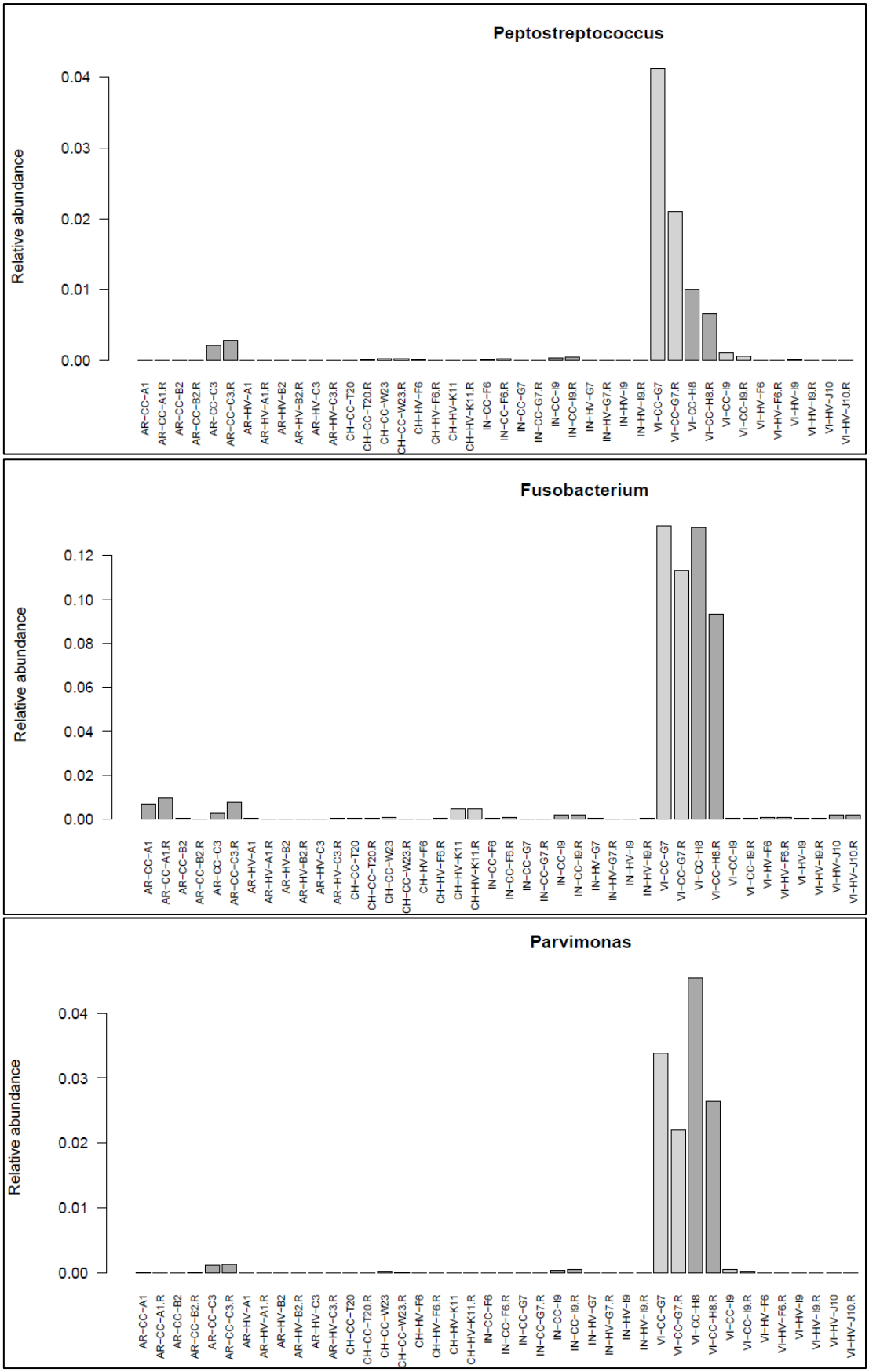


Figure 143. The relative abundance of *Escherichia-Shigella* for extraction replicates. Each pair of adjacent bars represents a pair of extraction replicates. Samples derived from Argentina, Chile, India and Vietnam are arranged from left to right respectively. Each bar is labelled as follows: country of origin (AR = Argentina; CH = Chile; IN = India; VI = Vietnam); disease status (CC = CRC; HV = healthy volunteer); sample name; whether the sample was an extraction replicate (indicated by R).


Peptostreptococcus
Relative abundance
0.04
0.03
0.02
0.01
0.00
AR−CC−A1
AR−CC−A1.R
AR−CC−B2
AR−CC−B2.R
AR−CC−C3
AR−CC−C3.R
AR−HV−A1
AR−HV−A1.R
AR−HV−B2
AR−HV−B2.R
AR−HV−C3
AR−HV−C3.R
CH−CC−T20
CH−CC−T20.R
CH−CC−W23
CH−CC−W23.R
CH−HV−F6
CH−HV−F6.R
CH−HV−K11
CH−HV−K11.R
IN−CC−F6
IN−CC−F6.R
IN−CC−G7
IN−CC−G7.R
IN−CC−I9
IN−CC−I9.R
IN−HV−G7
IN−HV−G7.R
IN−HV−I9
IN−HV−I9.R
VI−CC−G7
VI−CC−G7.R
VI−CC−H8
VI−CC−H8.R
VI−CC−I9
VI−CC−I9.R
VI−HV−F6
VI−HV−F6.R
VI−HV−I9
VI−HV−I9.R
VI−HV−J10
VI−HV−J10.R
Fusobacterium
Relative abundance
0.12
0.10
0.08
0.06
0.04
0.02
0.00
AR−CC−A1
AR−CC−A1.R
AR−CC−B2
AR−CC−B2.R
AR−CC−C3
AR−CC−C3.R
AR−HV−A1
AR−HV−A1.R
AR−HV−B2
AR−HV−B2.R
AR−HV−C3
AR−HV−C3.R
CH−CC−T20
CH−CC−T20.R
CH−CC−W23
CH−CC−W23.R
CH−HV−F6
CH−HV−F6.R
CH−HV−K11
CH−HV−K11.R
IN−CC−F6
IN−CC−F6.R
IN−CC−G7
IN−CC−G7.R
IN−CC−I9
IN−CC−I9.R
IN−HV−G7
IN−HV−G7.R
IN−HV−I9
IN−HV−I9.R
VI−CC−G7
VI−CC−G7.R
VI−CC−H8
VI−CC−H8.R
VI−CC−I9
VI−CC−I9.R
VI−HV−F6
VI−HV−F6.R
VI−HV−I9
VI−HV−I9.R
VI−HV−J10
VI−HV−J10.R
Parvimonas
Relative abundance
0.04
0.03
0.02
0.01
0.00
AR−CC−A1
AR−CC−A1.R
AR−CC−B2
AR−CC−B2.R
AR−CC−C3
AR−CC−C3.R
AR−HV−A1
AR−HV−A1.R
AR−HV−B2
AR−HV−B2.R
AR−HV−C3
AR−HV−C3.R
CH−CC−T20
CH−CC−T20.R
CH−CC−W23
CH−CC−W23.R
CH−HV−F6
CH−HV−F6.R
CH−HV−K11
CH−HV−K11.R
IN−CC−F6
IN−CC−F6.R
IN−CC−G7
IN−CC−G7.R
IN−CC−I9
IN−CC−I9.R
IN−HV−G7
IN−HV−G7.R
IN−HV−I9
IN−HV−I9.R
VI−CC−G7
VI−CC−G7.R
VI−CC−H8
VI−CC−H8.R
VI−CC−I9
VI−CC−I9.R
VI−HV−F6
VI−HV−F6.R
VI−HV−I9
VI−HV−I9.R
VI−HV−J10
VI−HV−J10.R

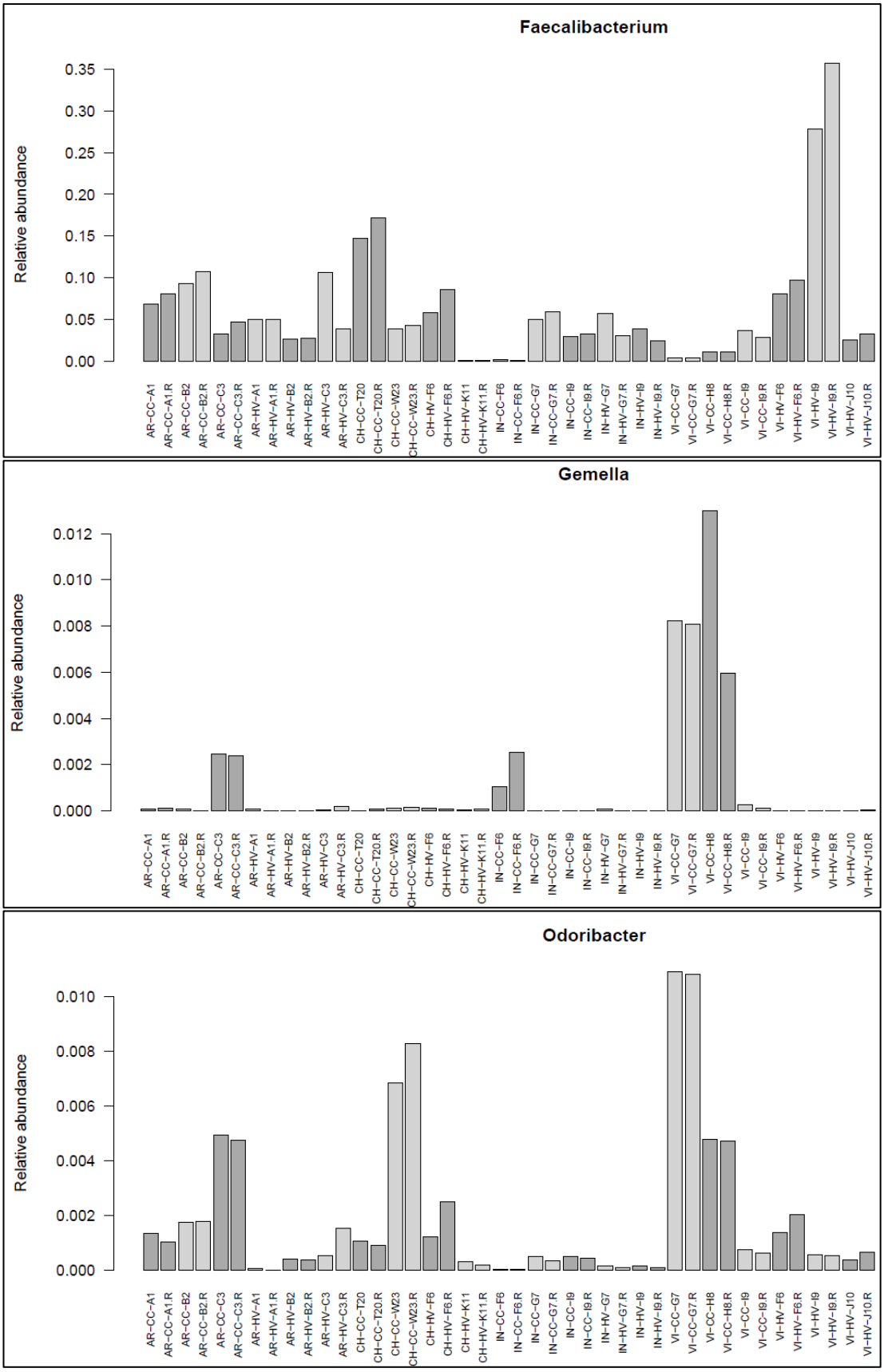

Faecalibacterium
Relative abundance
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
AR−CC−A1
AR−CC−A1.R
AR−CC−B2
AR−CC−B2.R
AR−CC−C3
AR−CC−C3.R
AR−HV−A1
AR−HV−A1.R
AR−HV−B2
AR−HV−B2.R
AR−HV−C3
AR−HV−C3.R
CH−CC−T20
CH−CC−T20.R
CH−CC−W23
CH−CC−W23.R
CH−HV−F6
CH−HV−F6.R
CH−HV−K11
CH−HV−K11.R
IN−CC−F6
IN−CC−F6.R
IN−CC−G7
IN−CC−G7.R
IN−CC−I9
IN−CC−I9.R
IN−HV−G7
IN−HV−G7.R
IN−HV−I9
IN−HV−I9.R
VI−CC−G7
VI−CC−G7.R
VI−CC−H8
VI−CC−H8.R
VI−CC−I9
VI−CC−I9.R
VI−HV−F6
VI−HV−F6.R
VI−HV−I9
VI−HV−I9.R
VI−HV−J10
VI−HV−J10.R
Gemella
Relative abundance
0.012
0.010
0.008
0.006
0.004
0.002
0.000
AR−CC−A1
AR−CC−A1.R
AR−CC−B2
AR−CC−B2.R
AR−CC−C3
AR−CC−C3.R
AR−HV−A1
AR−HV−A1.R
AR−HV−B2
AR−HV−B2.R
AR−HV−C3
AR−HV−C3.R
CH−CC−T20
CH−CC−T20.R
CH−CC−W23
CH−CC−W23.R
CH−HV−F6
CH−HV−F6.R
CH−HV−K11
CH−HV−K11.R
IN−CC−F6
IN−CC−F6.R
IN−CC−G7
IN−CC−G7.R
IN−CC−I9
IN−CC−I9.R
IN−HV−G7
IN−HV−G7.R
IN−HV−I9
IN−HV−I9.R
VI−CC−G7
VI−CC−G7.R
VI−CC−H8
VI−CC−H8.R
VI−CC−I9
VI−CC−I9.R
VI−HV−F6
VI−HV−F6.R
VI−HV−I9
VI−HV−I9.R
VI−HV−J10
VI−HV−J10.R
Odoribacter
Relative abundance
0.010
0.008
0.006
0.004
0.002
0.000
AR−CC−A1
AR−CC−A1.R
AR−CC−B2
AR−CC−B2.R
AR−CC−C3
AR−CC−C3.R
AR−HV−A1
AR−HV−A1.R
AR−HV−B2
AR−HV−B2.R
AR−HV−C3
AR−HV−C3.R
CH−CC−T20
CH−CC−T20.R
CH−CC−W23
CH−CC−W23.R
CH−HV−F6
CH−HV−F6.R
CH−HV−K11
CH−HV−K11.R
IN−CC−F6
IN−CC−F6.R
IN−CC−G7
IN−CC−G7.R
IN−CC−I9
IN−CC−I9.R
IN−HV−G7
IN−HV−G7.R
IN−HV−I9
IN−HV−I9.R
VI−CC−G7
VI−CC−G7.R
VI−CC−H8
VI−CC−H8.R
VI−CC−I9
VI−CC−I9.R
VI−HV−F6
VI−HV−F6.R
VI−HV−I9
VI−HV−I9.R
VI−HV−J10
VI−HV−J10.R

Figure 144. The relative abundance of CRC-associated taxa for extraction replicates. Each pair of adjacent bars represents a pair of extraction replicates. Samples derived from Argentina, Chile, India and Vietnam are arranged from left to right respectively. Each bar is labelled as follows: country of origin (AR = Argentina; CH = Chile; IN = India; VI = Vietnam); disease status (CC = CRC; HV = healthy volunteer); sample name; whether the sample was an extraction replicate (indicated by R).

4.4.4 Analysis of healthy volunteer and CRC samples

4.4.4.1 Tables of characteristics

Table 42 outlines the characteristics of participants. The healthy volunteers from Argentina, Chile and India were on average younger than the corresponding patients with CRC. The average age of the CRC patients from India and Vietnam was younger than the average age of CRC patients from Chile and Argentina. A colonoscopy diagnosis was available for the healthy volunteers from Vietnam and Argentina (but not those from Chile or India); the colonoscopy result was 'normal' for the majority and diverticulosis or haemorrhoids for a few. All of the CRC patients had undergone colonoscopy apart from the CRC patients from Vietnam. In the majority of cases, colonoscopy took place prior to sample collection. The bowel preparation agents used differed by country, apart from Chile and India which used the same regimen. Comorbidities included: hypertension, gastric ulcer, gastro-oesophageal reflux disease, insulin resistance or diabetes, thyroid disease, obesity, and hypercholesterolaemia. There were very few vegetarians, of which the majority were healthy volunteers from India. One CRC patient from Argentina (participant D4) had a history of antibiotic use 1-2 months prior to sample collection, for all other participants no antibiotic use within the previous six months was confirmed. Given that the majority of taxa return to baseline abundances within a month following antibiotic exposure (as described in Chapter 3), the sample from this participant was included in the current study.

Table 43 outlines the characteristics of the tumours. Incompleteness of some of the fields reflects difficulty obtaining some of the information or misunderstanding as to what information was required. The majority of tumours were located in the caecum/ascending colon or sigmoid/rectum. As tumour area was incompletely recorded, the maximum tumour size in one

direction was compared: the maximum size of tumours from Chile and Argentina was greater than the maximum size of tumours from India or Vietnam, although the median sizes were similar. The majority of tumours were grade 2, T3 or T4.

Table 42. Table of characteristics for healthy volunteer and CRC samples. HV = healthy volunteer. SD = standard deviation. History of colonoscopy indicates whether participants had ever had a colonoscopy with bowel preparation prior to sample collection. Current smoker includes participants who recently stopped smoking within the preceding month. If responses were not recorded for all participants within a group, this is denoted by a fraction for which the denominator represents the total number of responses recorded.

	Argentina		**Chile**		**India**		**Vietnam**	
	CRC	**HV**	**CRC**	**HV**	**CRC**	**HV**	**CRC**	**HV**
Number of samples	10	10	11	10	10	10	10	10
Male	4	6	4	4	7	6	5	3
Mean age	78	53.9	69.9	42.1	54.8	34	59.1	55.6
SD age	10.1	10.4	10.2	18.4	11.3	6.6	10.8	10.4
History of colonoscopy	9	10	11	0	8	0	0	8
Medication use	9	6	9	3	6	1	10	10
Comorbidities	8	4	9	5	7	0	3	3
Current smoker	0/8	1	0	3	1	0	3	1
Drinks alcohol	3/8	4	4	9	1	3	3	2
Eats meat	7/7	10	11	10	8	4	10	9

Table 43. Table of tumour characteristics. Size denotes the maximum size in one direction (i.e. length, width or height).

	Argentina n=10	**Chile** n=11	**India** n=10	**Vietnam** n=10
Tumour location				
Caecum or ascending colon	5	5	2	0
Transverse colon	0	0	0	1
Descending colon	0	0	0	1
Sigmoid colon or rectum	4	6	8	8
Not recorded	1	0	0	0
Tumour size				
Number of samples for which size was recorded	8	11	3	10
Range of sizes (cm)	0.7-7	2.5-11	3-5.5	3-5
Median size (cm)	4	6	5	4
Tumour type				
Mucinous	Not recorded	2	Not recorded	Not recorded
Tumour grade				
Grade 2	9	9	9	9
Grade 3	0	2	1	1
Not recorded	1	0	0	0
Tumour stage (TNM8)				
T1	0	2	Clinical stage recorded	0
T2	1	1		0
T3	8	8		0
T4	0	0		10
Not recorded	1	0		0

4.4.4.2 Alpha diversity

A significant difference in Shannon alpha diversity was detected between countries (Kruskal-Wallis p = 0.00004) (Figure 145). Pairwise analysis (Table 44) indicated that the alpha diversities of the Vietnamese samples and Indian samples were significantly lower than the alpha diversities of the Argentinian and Chilean samples; and the alpha diversity of the Indian samples was significantly lower than the Vietnamese samples. Results from the NHSBCSP study described in Chapter 3 are included for comparison; however, as methodological differences between the two studies exist, the NHSBCSP results have not been included in the formal statistical analysis.

Of note, due to small sample numbers results from each country represent a combined set of healthy volunteer and CRC samples; it will be more useful to assess this using larger numbers and separating the healthy volunteers from CRC within and between countries.

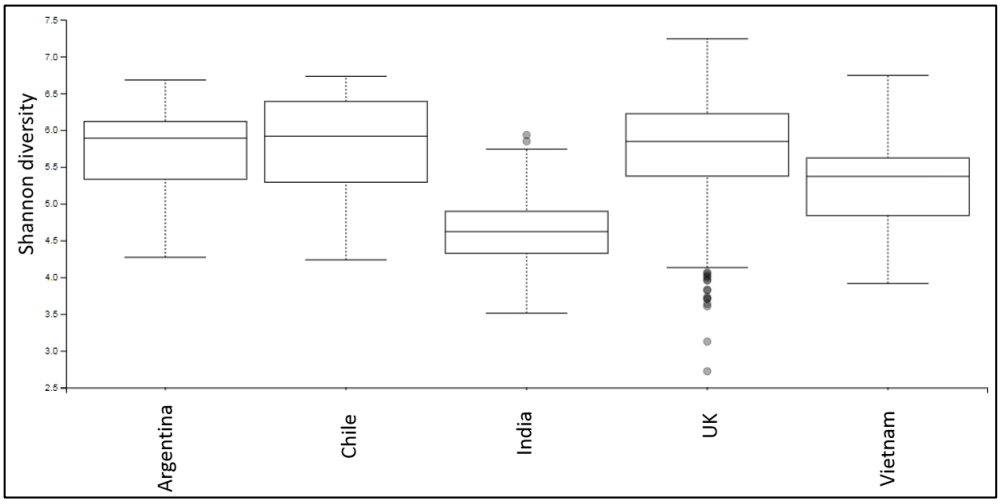


Figure 145. Boxplots of Shannon diversity index for samples from Argentina, Chile, India and Vietnam. Results from the NHSBCSP study described in Chapter 3 (CRC samples and blood-negative samples) are included for comparison.

Table 44. Pairwise Kruskal-Wallis analysis of Shannon diversity index for samples from different countries. AR = Argentina, CH = Chile; IN = India; VI = Vietnam. Significant q values are shaded grey. Values are recorded to two decimal places.

Group 1	Group 2	H	p-value	q-value
AR (n=20)	CH (n=21)	0.39	0.53	0.53
	IN (n=20)	15.60	7.84×10^{-5}	4.70×10^{-4}
	VI (n=20)	5.29	2.15×10^{-2}	2.58×10^{-2}
CH (n=21)	IN (n=20)	13.91	1.92×10^{-4}	5.75×10^{-4}
	VI (n=20)	5.88	1.53×10^{-2}	2.58×10^{-2}
IN (n=20)	VI (n=20)	5.41	2.0×10^{-2}	2.58×10^{-2}

No significant difference in Shannon diversity index was detected between overall CRC and healthy volunteer samples (Kruskal-Wallis p = 0.28) (Figure 146). It will be important to assess this using larger numbers of samples and separating the healthy volunteers from CRC within countries.

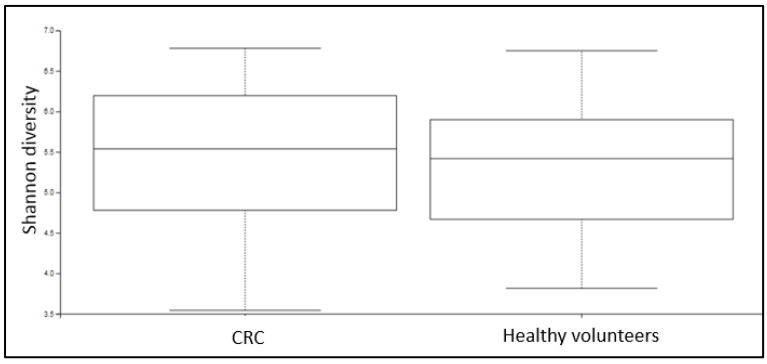


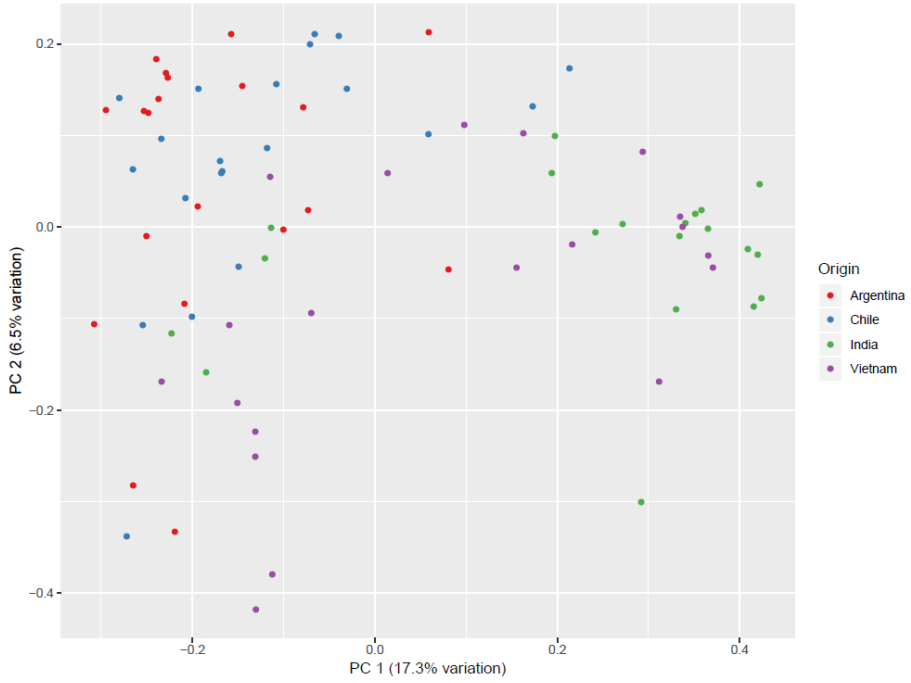
Figure 146. Boxplots of Shannon diversity index for CRC and healthy volunteer samples from Argentina, Chile, India and Vietnam.

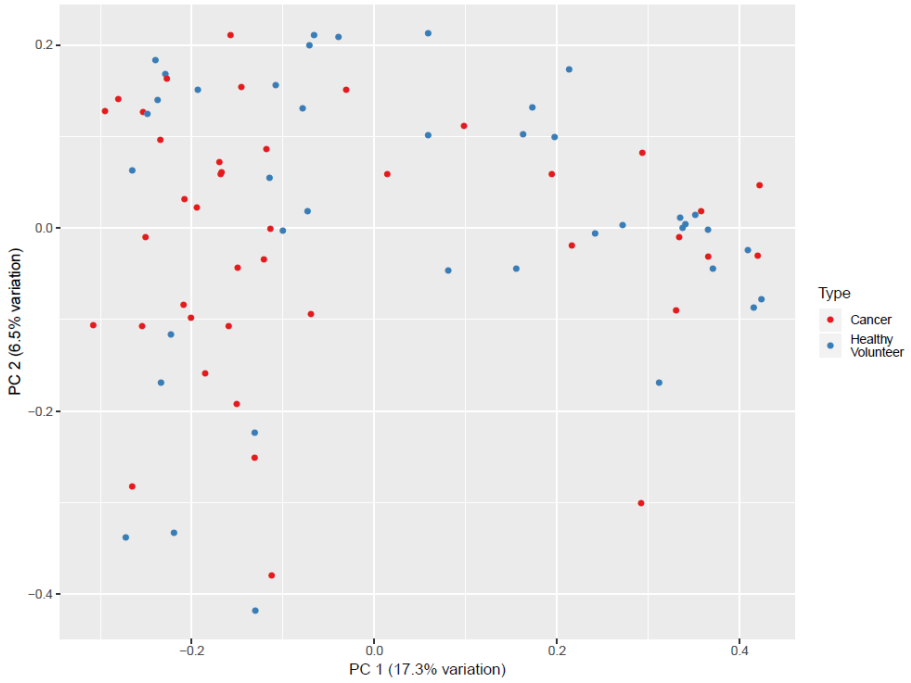
4.4.4.3 Beta diversity

PCA of Bray-Curtis distances demonstrated a degree of sample clustering according to both country of origin and disease status (Figure 147). The microbiomes of Asian samples (India and Vietnam) and South American samples (Chile and Argentina) appeared closer to one another. However, whilst there was visible separation on PCA, there was also a wide range of Bray-Curtis distances within groups. Results from the NHSBCSP study

described in Chapter 3 were included for comparison; the UK samples demonstrated a greater degree of similarity with the samples from Chile and Argentina than the samples from India or Vietnam. However, as methodological differences between the two studies exist, the NHSBCSP results have not been included in the formal statistical analysis.

PERMANOVA analysis of the variables ‘country of origin’, ‘disease status’, age and gender confirmed that ‘country of origin’ contributes to the largest amount of variation in Bray-Curtis distance and that ‘disease status’ and gender contribute small but significant amounts (Table 45).


0.2
0.0
−0.2
−0.4
PC 2 (6.5% variation)
−0.2
0.0
0.2
0.4
PC 1 (17.3% variation)
Origin
Argentina
Chile
India
Vietnam


0.2
0.0
−0.2
−0.4
PC 2 (6.5% variation)
−0.2
0.0
0.2
0.4
PC 1 (17.3% variation)
Type
Cancer
Healthy Volunteer

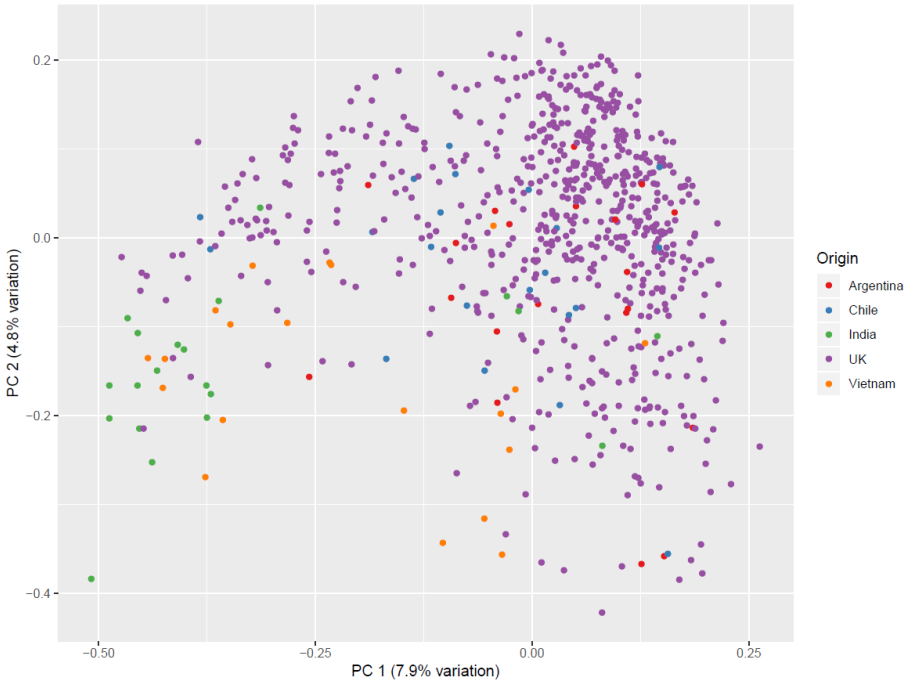


Figure 147. PCA of Bray-Curtis distances of samples derived from healthy volunteers and CRC patients from the network. Upper plot: points are coloured according to country of origin. Middle plot: points are coloured according to disease status. Lower plot: points are coloured according to country of origin and results from the NHSBCSP study described in Chapter 3 (CRC samples and blood-negative samples) are included for comparison.

Table 45. Results of PERMANOVA analysis of samples derived from healthy volunteers and CRC patients from the network. Df = degrees of freedom. Disease status denotes Healthy volunteer or CRC. NA = not applicable. Significant p values are shaded grey. Values are recorded to two decimal places.

	Df	**Sums of squares**	**F.Model**	**R^2**	**Pr(>F)**
Country	3	3.59	4.23	0.14	1×10^{-4}
Disease status	1	0.56	1.99	0.02	0.01
Age	1	0.32	1.11	0.01	0.26
Gender	1	0.55	1.93	0.02	0.01
Residuals	74	20.97	NA	0.81	NA
Total	80	25.99	NA	1.00	NA

4.4.4.4 Taxonomy

The taxonomy bar chart (Figure 148) demonstrates that on average, samples from Argentina and Chile had a high relative abundance of *Bacteroides* (dark grey) compared with samples from India and Vietnam which had a high relative abundance of *Prevotella* (light grey).

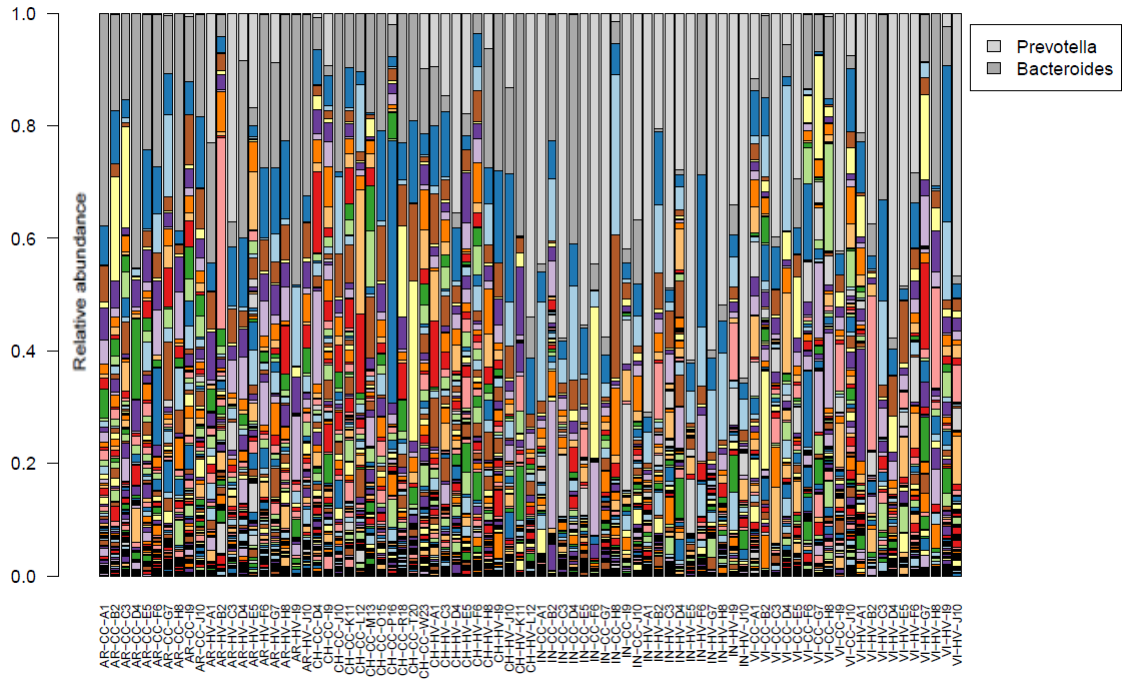


Figure 148. Taxonomy bar chart of samples derived from healthy volunteers and CRC patients from the network. Each bar represents a sample. Colours denote the relative abundance of taxa (genus level); there are too many to include a legend. *Prevotella* and *Bacteroides* are coloured light and dark grey respectively for ease of identification. Samples derived from Argentina, Chile, India and Vietnam are arranged from left to right respectively. Each bar is labelled as follows: country of origin (AR = Argentina; CH = Chile; IN = India; VI = Vietnam); disease status (CC = CRC; HV = healthy volunteer); sample name.

An average taxonomic composition for each country is compared with the NHSBCSP UK samples in Figure 149 (it should be noted that NHSBCSP samples were collected in a different manner to the abroad samples). The mean relative abundance of *Prevotella* and *Bacteroides* appears higher and lower respectively in the non-Western healthy volunteer cohorts compared with the CRC cohorts, whereas this difference is not apparent in the NHSBCSP samples.

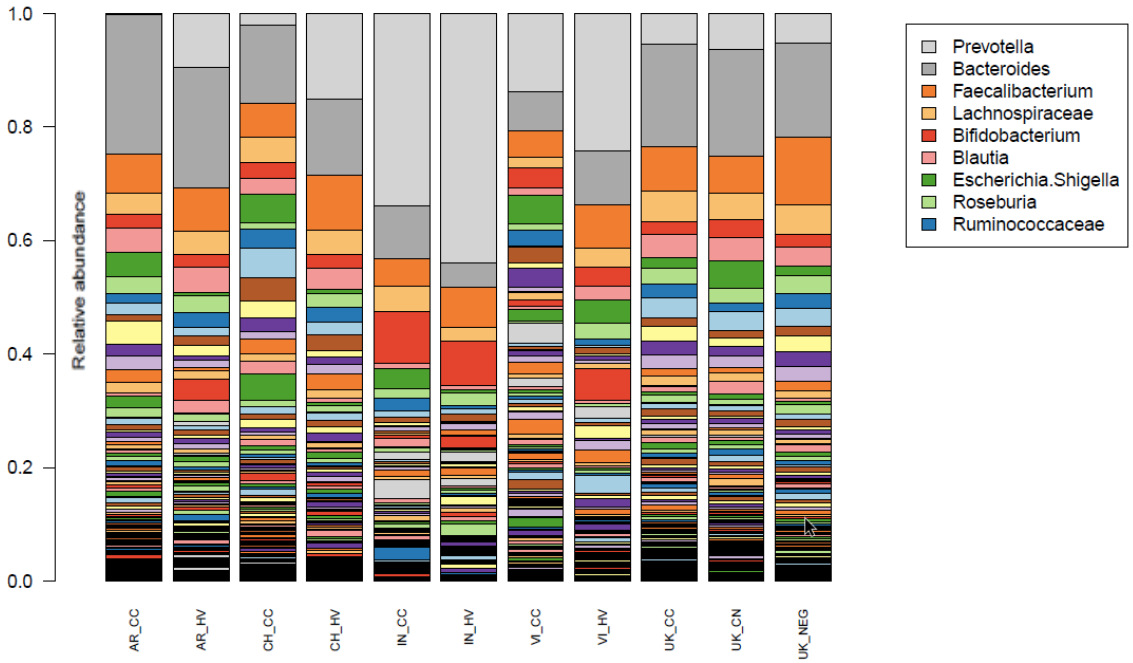
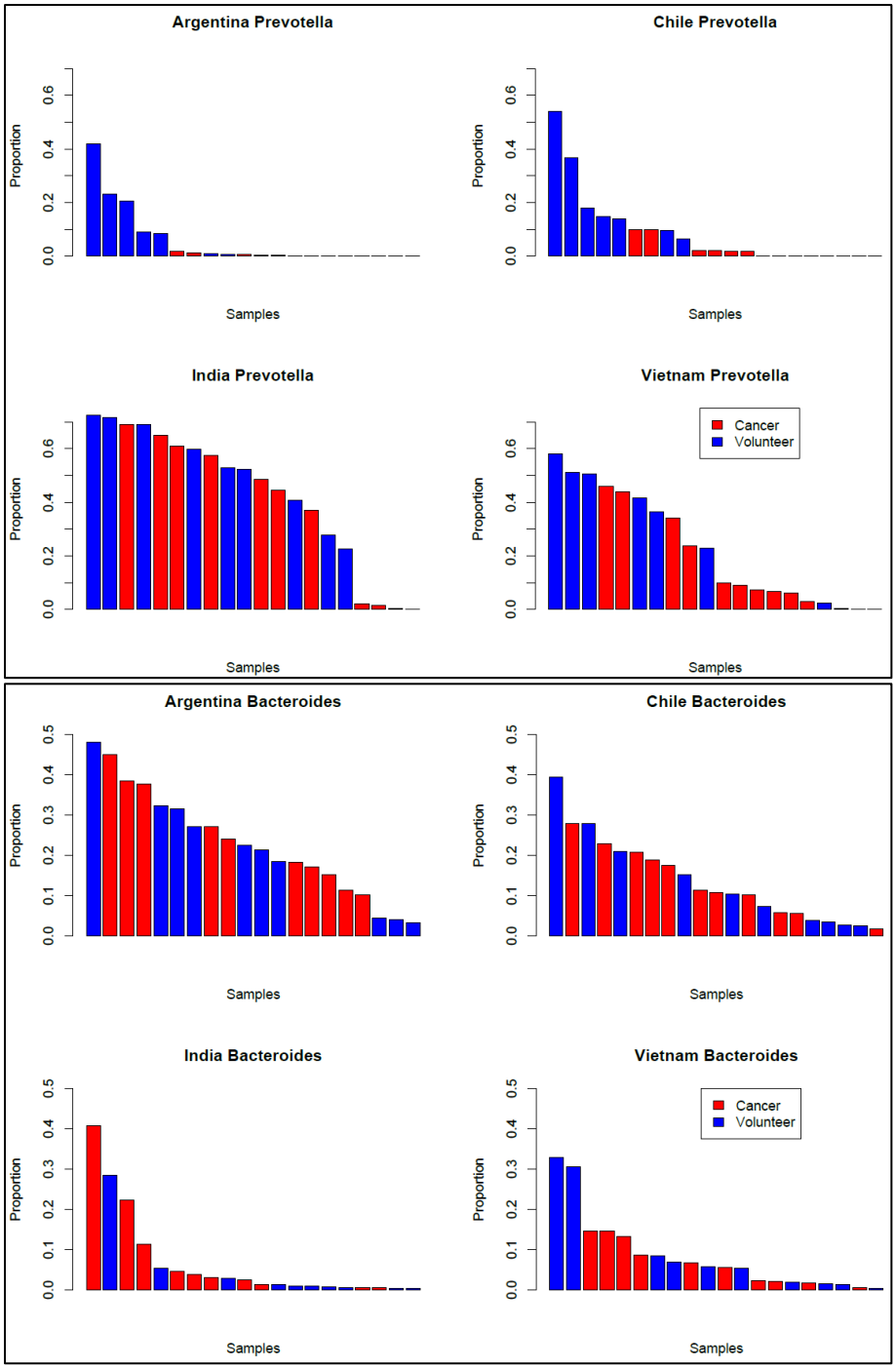


Figure 149. Taxonomy bar chart of mean taxonomic composition of samples derived from healthy volunteers and CRC patients from the network, and NHSBCSP samples. Each bar represents the mean taxonomic composition of samples within a group. Colours denote the relative abundance of taxa (genus level); there are too many to include a comprehensive legend (only the most abundant taxa are labelled). *Prevotella* and *Bacteroides* are coloured light and dark grey respectively for ease of identification. Samples derived from Argentina, Chile, India, Vietnam and the NHSBCSP study are arranged from left to right respectively. Each bar is labelled as follows: country of origin (AR = Argentina; CH = Chile; IN = India; VI = Vietnam; UK = NHSBCSP samples); disease status (CC = CRC; HV = healthy volunteer; CN = colonoscopy-normal; NEG = blood-negative gFOBT).

In order to determine whether this result is consistent among individual samples, waterfall plots depicting the relative abundance of *Bacteroides*, *Prevotella* and the *Bacteroides:Prevotella* and *Prevotella:Bacteroides* ratios were plotted (Figure 150). These plots suggest a trend for a higher relative abundance of *Prevotella* and *Prevotella:Bacteroides* ratio in healthy volunteers compared with CRC patients, although this is not true of all samples and needs confirming in larger cohorts. It should be noted that the *Prevotella* group was a composite of the following taxa: *Prevotella, Prevotella.2, Prevotella.6, Prevotella.7, and Prevotella.9.* The *Bacteroides* group was a composite of *Bacteroides* and *Bacteroides pectinophilus.*


Argentina Prevotella
Chile Prevotella
India Prevotella
Vietnam Prevotella
Cancer
Volunteer
Argentina Bacteroides
Chile Bacteroides
India Bacteroides
Vietnam Bacteroides
Cancer
Volunteer
Proportion
Samples

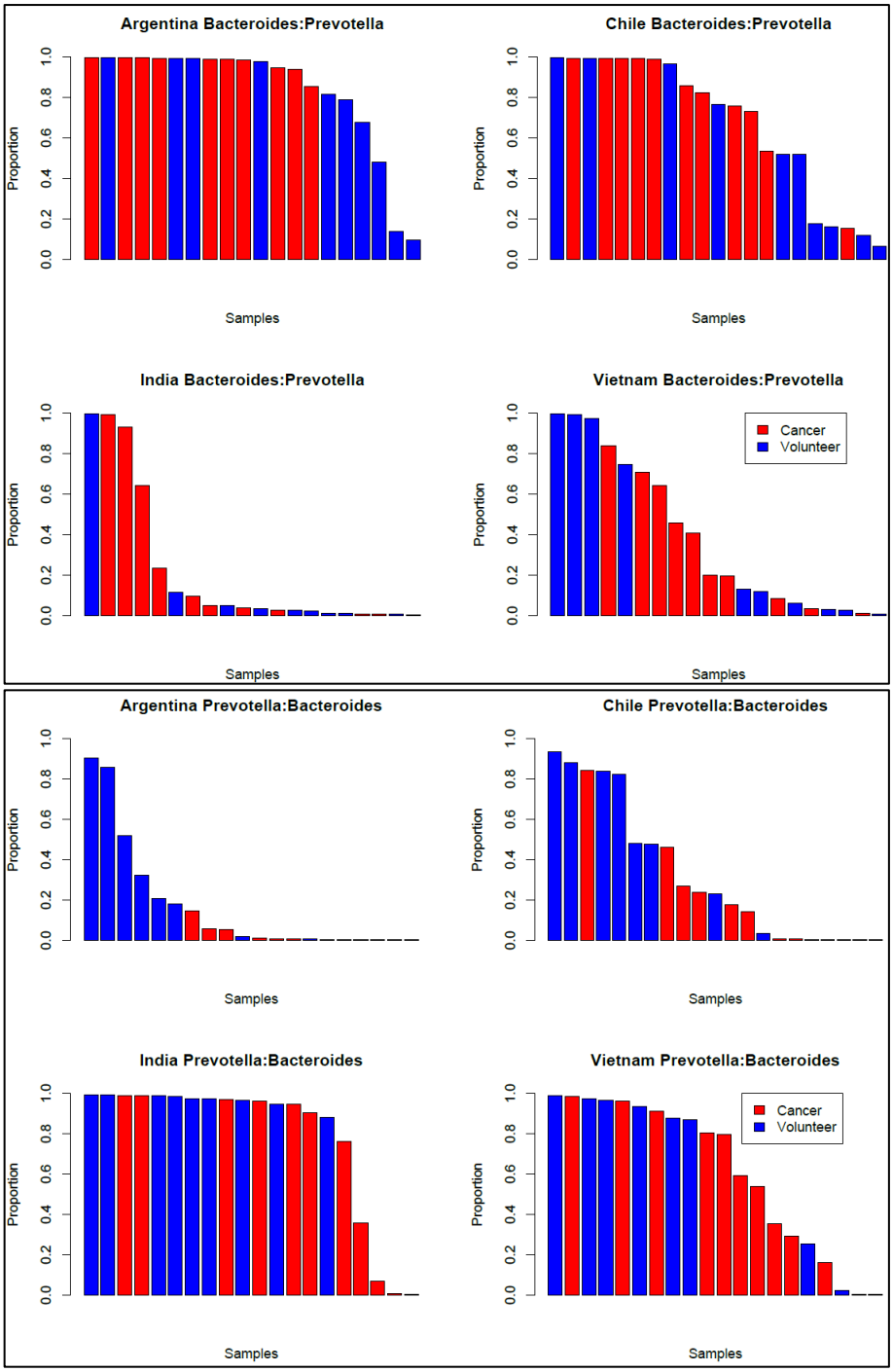
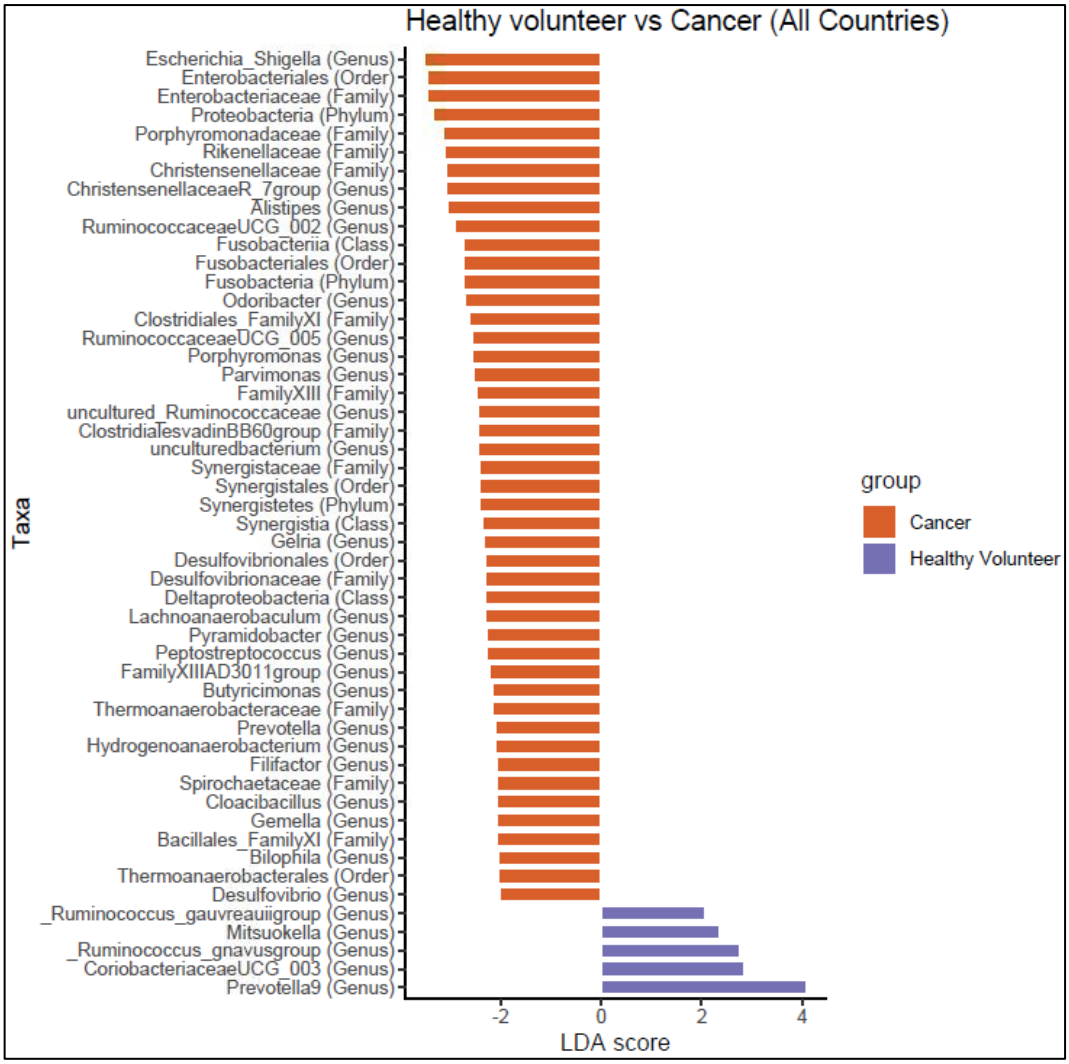


Figure 150. Waterfall plots of the relative abundance of *Prevotella*, *Bacteroides* and their ratios for samples derived from healthy volunteers and CRC patients from the network. The upper plot shows the relative abundance of *Prevotella*, the second plot shows the relative abundance of *Bacteroides*, the third plot shows the *Bacteroides:Prevotella* ratio, the fourth plot shows the *Prevotella:Bacteroides* ratio. Red = samples from CRC patients. Blue = samples from healthy volunteers. The *Bacteroides:Prevotella* ratio was calculated by dividing the relative abundance of *Bacteroides* by the combined relative abundance of *Bacteroides* and *Prevotella*. The

Prevotella:Bacteroides ratio was calculated by dividing the relative abundance of *Prevotella* by the combined relative abundance of *Bacteroides* and *Prevotella*.

4.4.4.5 LEfSe analysis

LEfSe analysis identified taxa which were significantly enriched in healthy volunteer samples compared with CRC samples (Figure 151). Interestingly the genus which was the most significantly enriched in CRC was *Escherichia-Shigella*. The inter-subject variability of the relative abundance of *Escherichia-Shigella* described in Chapter 2, suggests that this may not be appropriate for use as a CRC screening marker. CRC-associated bacteria described in meta-analyses of faecal studies as being enriched in CRC compared to controls were identified (121, 166, 477, 496, 534, 597). The cladogram (Figure 151) showed some similarities but also differences compared with the cladograms of NHSBCSP samples depicted in Chapter 3.


Healthy volunteer vs Cancer (All Countries)
Escherichia_Shigella (Genus)
Enterobacteriales (Order)
Enterobacteriaceae (Family)
Proteobacteria (Phylum)
Porphyromonadaceae (Family)
Rikenellaceae (Family)
Christensenellaceae (Family)
ChristensenellaceaeR_7group (Genus)
Alistipes (Genus)
RuminococcaceaeUCG_002 (Genus)
Fusobacteriia (Class)
Fusobacteriales (Order)
Fusobacteria (Phylum)
Odoribacter (Genus)
Clostridiales_FamilyXI (Family)
RuminococcaceaeUCG_005 (Genus)
Porphyromonas (Genus)
Parvimonas (Genus)
FamilyXIII (Family)
uncultured_Ruminococcaceae (Genus)
ClostridialesvadinBB60group (Family)
unculturedbacterium (Genus)
Synergistaceae (Family)
Synergistales (Order)
Synergistetes (Phylum)
Synergistia (Class)
Gelria (Genus)
Desulfovibrionales (Order)
Desulfovibrionaceae (Family)
Deltaproteobacteria (Class)
Lachnoanaerobaculum (Genus)
Pyramidobacter (Genus)
Peptostreptococcus (Genus)
FamilyXIIIAD3011group (Genus)
Butyricimonas (Genus)
Thermoanaerobacteraceae (Family)
Prevotella (Genus)
Hydrogenoanaerobacterium (Genus)
Filifactor (Genus)
Spirochaetaceae (Family)
Cloacibacillus (Genus)
Gemella (Genus)
Bacillales_FamilyXI (Family)
Bilophila (Genus)
Thermoanaerobacterales (Order)
Desulfovibrio (Genus)
_Ruminococcus_gauvreauiigroup (Genus)
Mitsuokella (Genus)
_Ruminococcus_gnavusgroup (Genus)
CoriobacteriaceaeUCG_003 (Genus)
Prevotella9 (Genus)
Taxa
-2
0
2
4
LDA score
group
Cancer
Healthy Volunteer

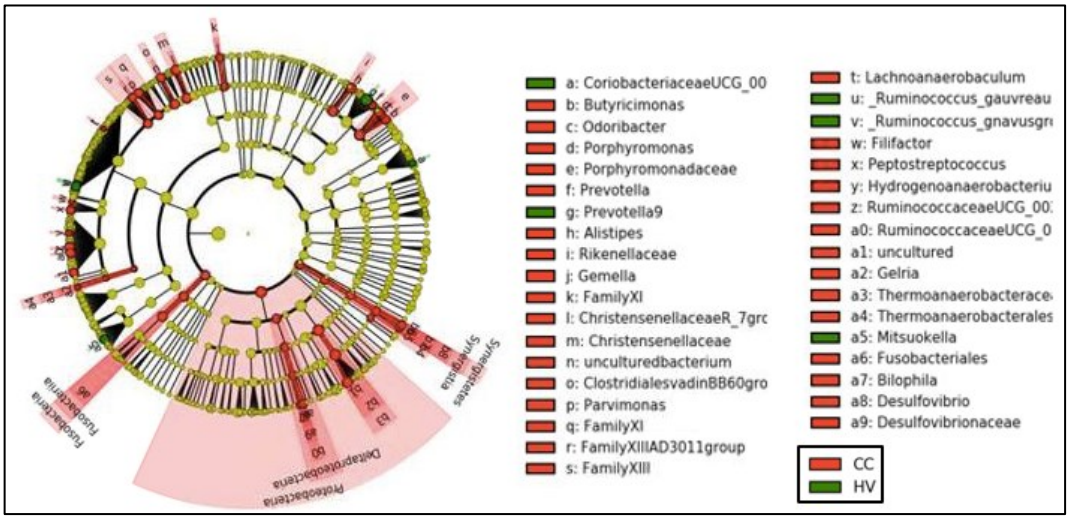

a: CoriobacteriaceaeUCG_00
b: Butyricimonas
c: Odoribacter
d: Porphyromonas
e: Porphyromonadaceae
f: Prevotella
g: Prevotella9
h: Alistipes
i: Rikenellaceae
j: Gemella
k: FamilyXI
l: ChristensenellaceaeR_7grc
m: Christensenellaceae
n: unculturedbacterium
o: ClostridialesvadinBB60gro
p: Parvimonas
q: FamilyXI
r: FamilyXIIIAD3011group
s: FamilyXIII
t: Lachnoanaerobaculum
u: _Ruminococcus_gauvreau
v: _Ruminococcus_gnavusgr
w: Filifactor
x: Peptostreptococcus
y: Hydrogenoanaerobacteriu
z: RuminococcaceaeUCG_00
a0: RuminococcaceaeUCG_0
a1: uncultured
a2: Gelria
a3: Thermoanaerobacterace
a4: Thermoanaerobacterales
a5: Mitsuokella
a6: Fusobacteriales
a7: Bilophila
a8: Desulfovibrio
a9: Desulfovibrionaceae
CC
HV

Figure 151. LEfSe plot and cladogram of samples derived from healthy volunteers and CRC patients from the network. LEfSe plot indicates taxa which are significantly enriched in CRC samples (orange) and healthy volunteer samples (purple), ranked according to effect size. The cladogram indicates the phylogenetic relationship between taxa. Working outwards, the five rings denote phylum, class, order, family, and genus. Taxa are represented by circles. Circle diameter is proportional to abundance. Circle colour indicates taxa which are significantly enriched in CRC samples (red) and healthy volunteer samples (green).

The genera from Figure 151 are displayed in Table 46 for clarity.

Table 46. Genera enriched/depleted in CRC compared with healthy volunteers (all countries). Taxa which have been identified as being differentially abundant between CRC and controls by meta-analyses of faecal studies are shaded grey (121, 166, 477, 496, 534, 597): Taxa which are consistent with the results of these studies are marked (+); those that conflict with the results of these studies are marked (-).

Genera enriched in CRC compared with healthy volunteers (all countries)	Genera depleted in CRC compared with healthy volunteers (all countries)
Escherichia-Shigella (+)	*Prevotella9*
ChristensenellaceaeR_7 group	*CoriobacteriaceaeUCG_003*
Alistipes (+)	*Ruminococcus_gnavus group*
RuminococcaceaeUCG_002	*Mitsuokella*
Odoribacter	*Ruminococcus_gauvreauii group*
RuminococcaceaeUCG_005	
Porphyromonas (+)	
Parvimonas (+)	
Uncultured *Ruminococcaceae*	
Uncultured bacterium	
Gelria	
Lachnoanaerobaculum	
Pyramidobacter	

Genera enriched in CRC compared with healthy volunteers (all countries)	Genera depleted in CRC compared with healthy volunteers (all countries)
Peptostreptococcus (+)	
FamilyXIIIAD3011group	
Butyricimonas	
Prevotella	
Hydrogenoanaerobacterium	
Filifactor	
Cloacibacillus	
Gemella (+)	
Bilophila	
Desulfovibrio	

Within each country, LEfSe analysis was performed (Figure 152 to Figure 155). CRC-associated bacteria described in meta-analyses of faecal studies as being enriched in CRC compared to controls were identified (121, 166, 477, 496, 534, 597). Interestingly the CRC-associated taxa differed by country, although *Peptostreptococcus* and *Parvimonas* were enriched in CRC from all of the countries except Argentina. Comparison of the cladograms (Figure 156) suggests a common phylogenetic position of CRC-associated taxa. These results require confirmation using a larger cohort.

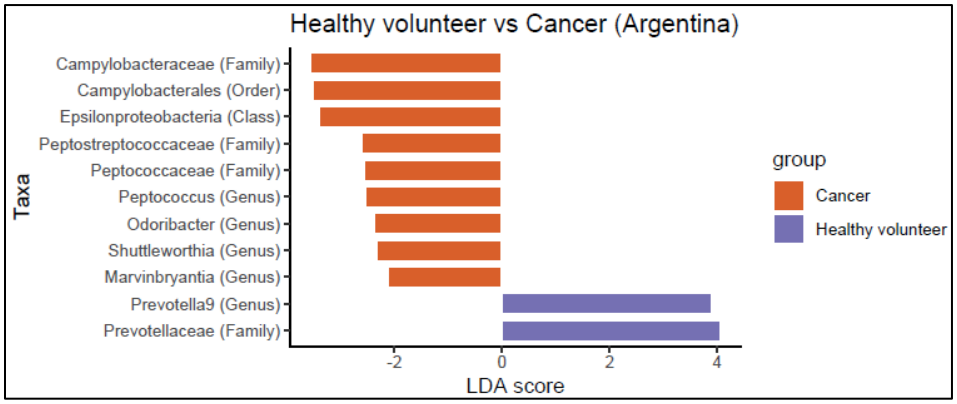


Figure 152. LEfSe plot of Argentina samples (healthy volunteer compared with CRC). LEfSe plot indicates taxa which are significantly enriched in CRC samples (orange) and healthy volunteer samples (purple), ranked according to effect size.

The genera from Figure 152 are displayed in Table 47 for clarity. None of the taxa which have been described as being differentially abundant between CRC and controls by meta-analyses of faecal studies were identified (121, 166, 477, 496, 534, 597).

Table 47. Genera enriched/depleted in CRC compared with healthy volunteers (Argentina).

Genera enriched in CRC compared with healthy volunteers (Argentina)	Genera depleted in CRC compared with healthy volunteers (Argentina)
Peptococcus	*Prevotella9*
Odoribacter	
Shuttleworthia	
Marvinbryantia	

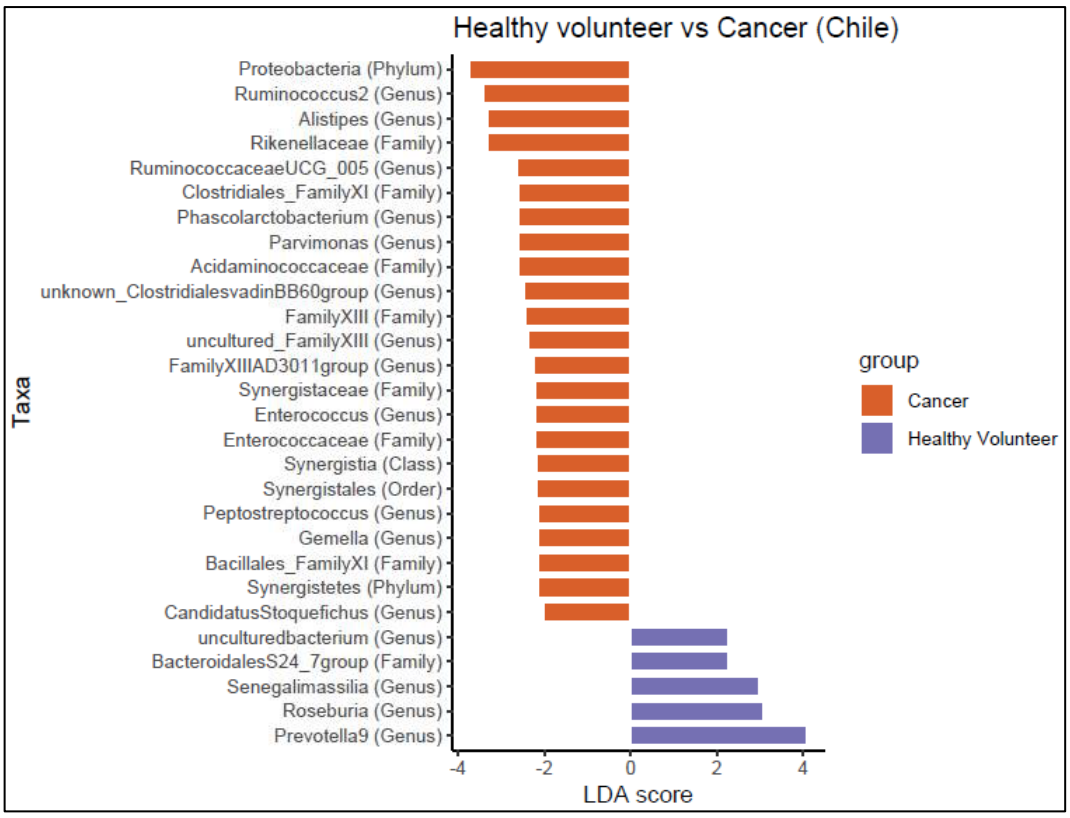


Figure 153. LEfSe plot of Chile samples (healthy volunteer compared with CRC). LEfSe plot indicates taxa which are significantly enriched in CRC samples (orange) and healthy volunteer samples (purple), ranked according to effect size.

The genera from Figure 153 are displayed in Table 48 for clarity.

Table 48. Genera enriched/depleted in CRC compared with healthy volunteers (Chile). Taxa which have been identified as being differentially abundant between CRC and controls by meta-analyses of faecal studies are shaded grey (121, 166, 477, 496, 534, 597). Taxa which are consistent with the results of these studies are marked (+); those that conflict with the results of these studies are marked (-).

Genera enriched in CRC compared with healthy volunteers (Chile)	Genera depleted in CRC compared with healthy volunteers (Chile)
Ruminococcus2	*Prevotella9*
Alistipes (+)	*Roseburia* (+)
RuminococcaceaeUCG_005	*Senegalimassilia*
Phascolarctobacterium	Uncultured bacterium
Parvimonas (+)	
Unknown *ClostridialesvadinBB60 group*	
Uncultured *FamilyXIII*	
FamilyXIIIAD3011 group	
Enterococcus	
Peptostreptococcus (+)	
Gemella (+)	
CandidatusStoquefichus	

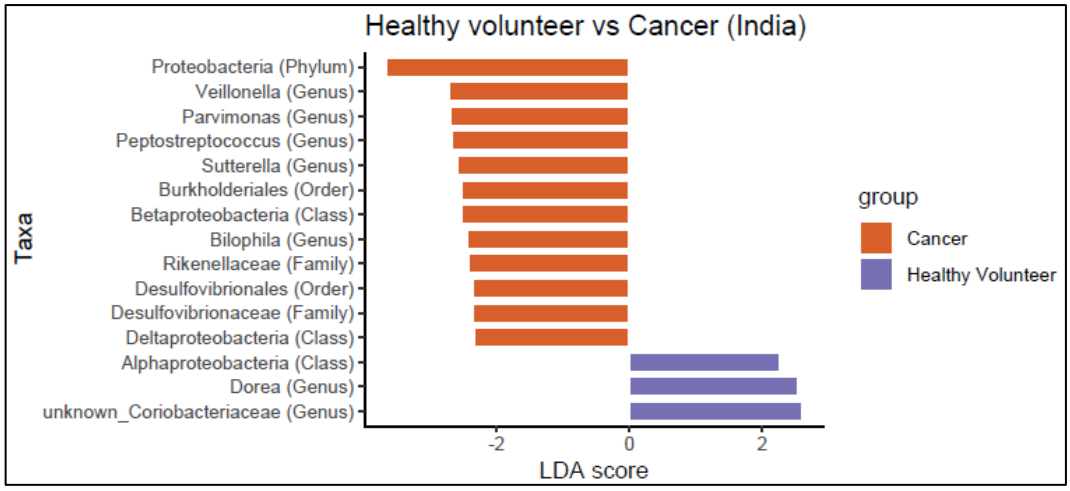


Figure 154. LEfSe plot of India samples (healthy volunteer compared with CRC). LEfSe plot indicates taxa which are significantly enriched in CRC samples (orange) and healthy volunteer samples (purple), ranked according to effect size.

The genera from Figure 154 are displayed in Table 49 for clarity.

Table 49. Genera enriched/depleted in CRC compared with healthy volunteers (India). Taxa which have been identified as being differentially abundant between CRC and controls by meta-analyses of faecal studies are shaded grey (121, 166, 477, 496, 534, 597). Taxa which are consistent with the results of these studies are marked (+); those that conflict with the results of these studies are marked (-).

Genera enriched in CRC compared with healthy volunteers (India)	Genera depleted in CRC compared with healthy volunteers (India)
Veillonella	Unknown *Coriobacteriaceae*
Parvimonas (+)	*Dorea*
Peptostreptococcus (+)	
Sutterella	
Bilophila	

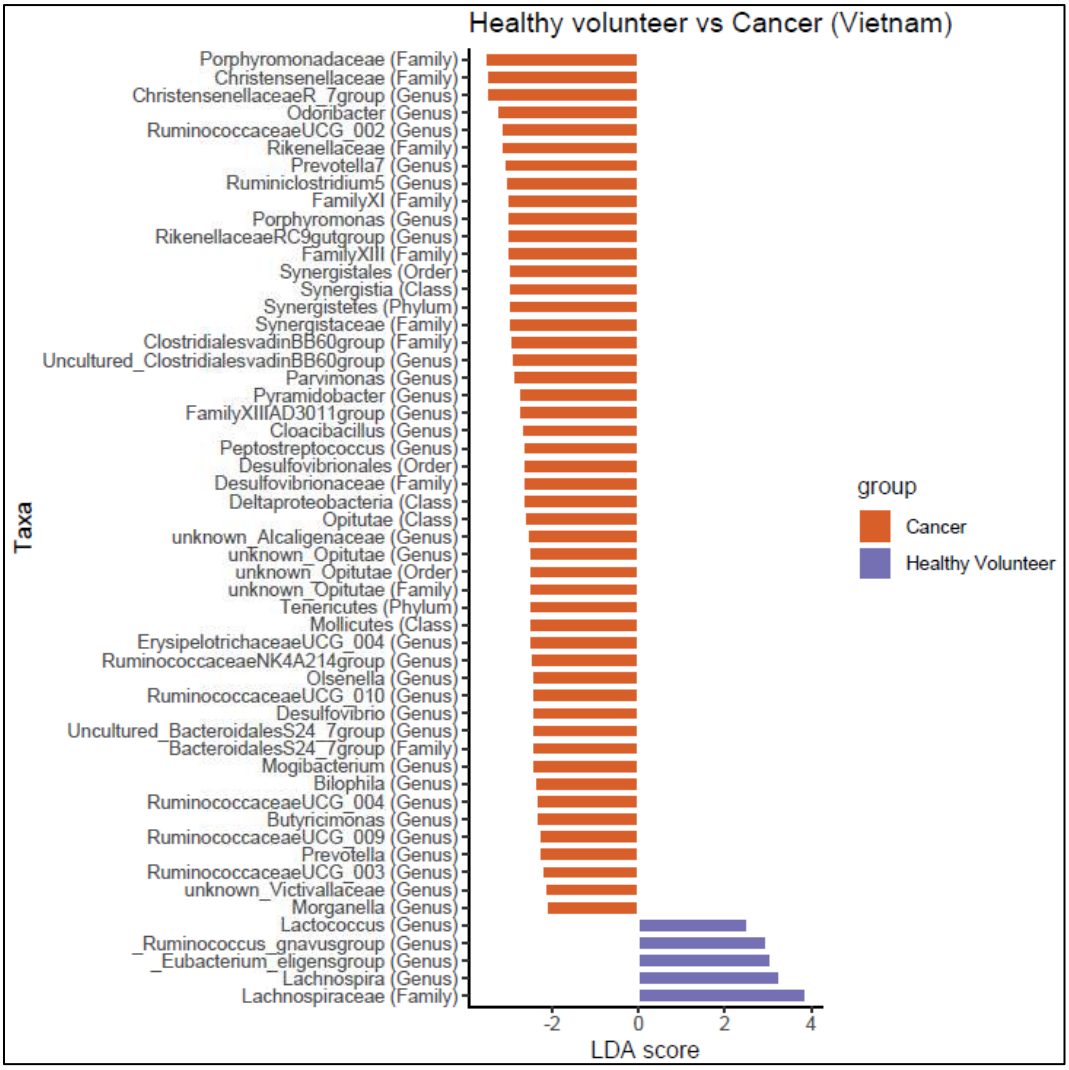


Figure 155. LEfSe plot of Vietnam samples (healthy volunteer compared with CRC). LEfSe plot indicates taxa which are significantly enriched in CRC samples (orange) and healthy volunteer samples (purple), ranked according to effect size.

The genera from Figure 155 are displayed in Table 50 for clarity.

Table 50. Genera enriched/depleted in CRC compared with healthy volunteers (Vietnam). Taxa which have been identified as being differentially abundant between CRC and controls by meta-analyses of faecal studies are shaded grey (121, 166, 477, 496, 534, 597). Taxa which are consistent with the results of these studies are marked (+); those that conflict with the results of these studies are marked (-).

Genera enriched in CRC compared with healthy volunteers (Vietnam)	Genera depleted in CRC compared with healthy volunteers (Vietnam)
ChristensenellaceaeR_7 group	*Lachnospira*
Odoribacter	*Eubacterium_eligens group*
RuminococcaceaeUCG_002	*Ruminococcus_gnavus group*
Prevotella7 (+)	*Lactococcus*
Ruminiclostridium5	
Porphyromonas (+)	
RikenellaceaeRC9gut group	
Uncultured *ClostridialesvadinBB60 group*	
Parvimonas (+)	
Pyramidobacter	
FamilyXIIIAD3011 group	
Cloacibacillus	
Peptostreptococcus (+)	
Unknown *Alcaligenaceae*	
Unknown *Opitutae*	
ErysipelotrichaceaeUCG_004	
RuminococcaceaeNK4A214 group	
Olsenella	
RuminococcaceaeUCG_010	
Desulfovibrio	
Uncultured *BacteroidalesS24_7 group*	

Genera enriched in CRC compared with healthy volunteers (Vietnam)	**Genera depleted in CRC compared with healthy volunteers (Vietnam)**
Mogibacterium	
Bilophila	
RuminococcaceaeUCG_004	
Butyricimonas	
RuminococcaceaeUCG_009	
Prevotella	
RuminococcaceaeUCG_003	
Unknown *Victivallaceae*	
Morganella	

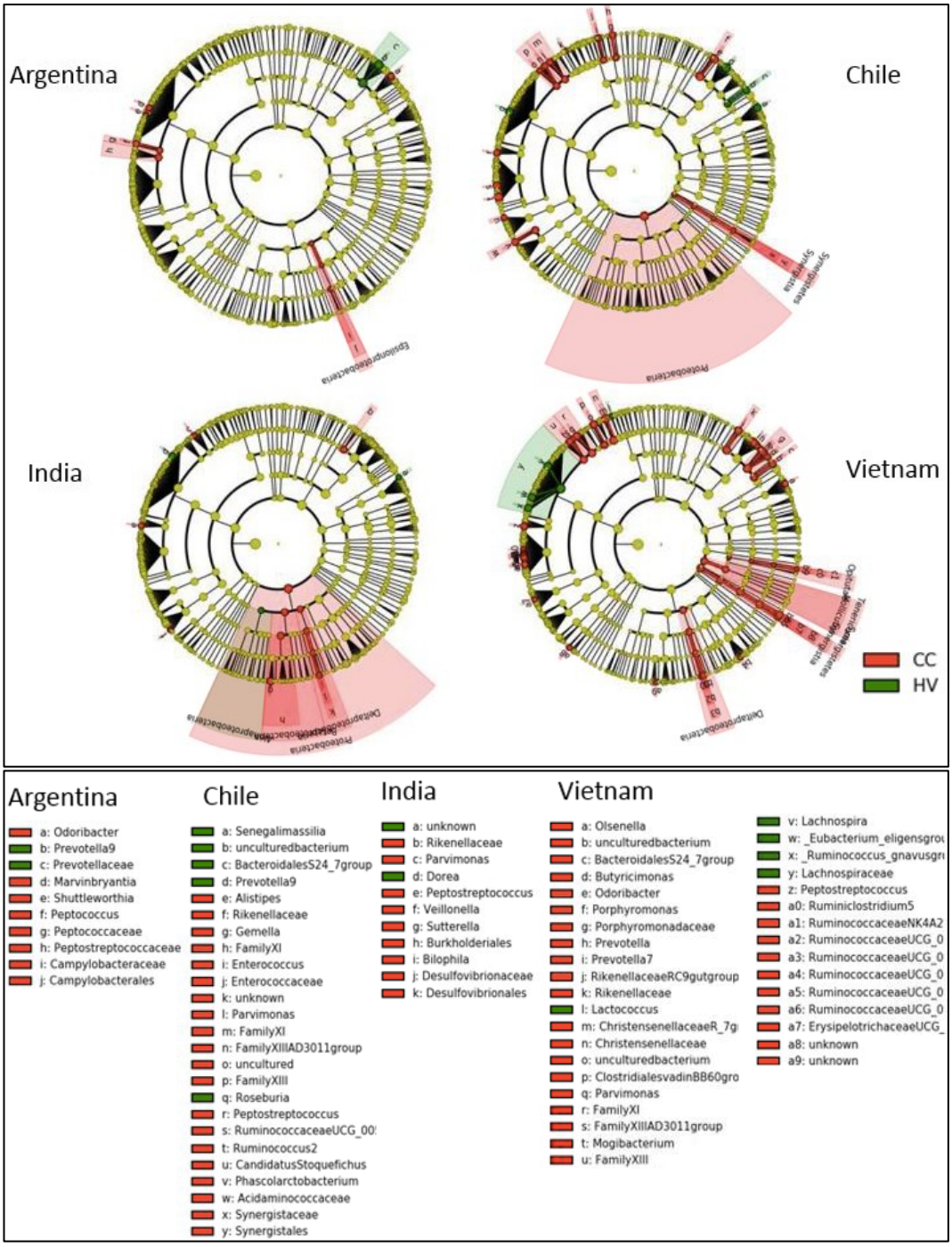


Figure 156. Cladograms of samples from the four countries (healthy volunteer compared with CRC). The cladograms indicate the phylogenetic relationship between taxa. Working outwards, the five rings denote phylum, class, order, family, and genus. Taxa are represented by circles. Circle diameter is proportional to abundance. Circle colour indicates taxa which are significantly enriched in CRC samples (red) and healthy volunteer samples (green).

The distribution of relative abundances of CRC-associated bacteria (identified as enriched in CRC by the current study's LEfSe analysis, the Random Forest models described in Chapter 3 or the aforementioned meta-analyses of faecal studies) is indicated by Figure 157 to Figure 164. The results need confirming in a larger cohort, but there appears to be a difference between countries in the maximum relative abundance of certain taxa. For example the maximum relative abundance of *Peptostreptococcus* and *Fusobacterium* is higher for samples from Vietnam compared with the other countries, and the maximum relative abundance of *Parvimonas* and *Gemella* is higher for Vietnam and Chile than India or Argentina. This suggests that country-specific thresholds of relative abundance may be required for a microbiome based CRC-screening test.

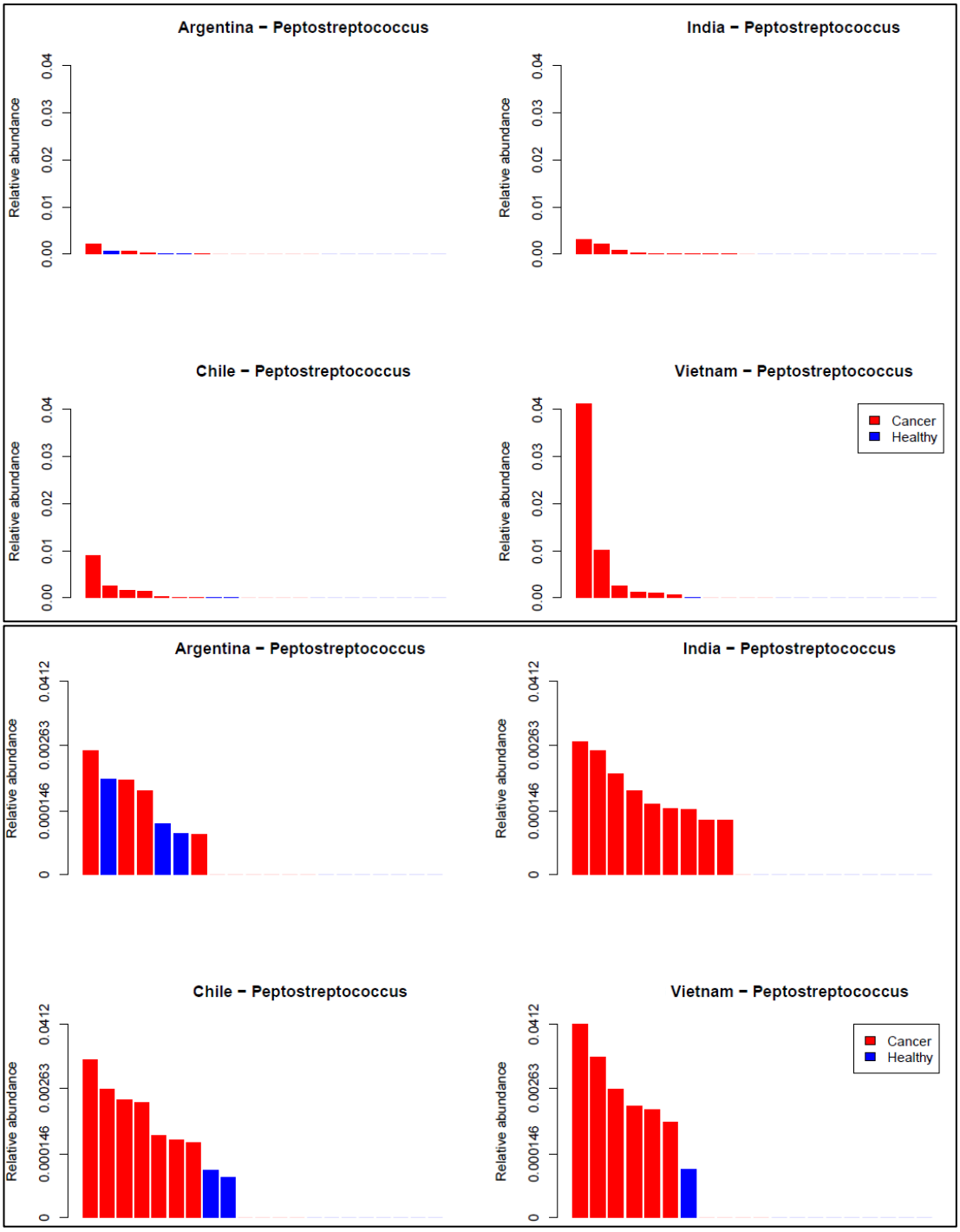


Figure 157. Waterfall plots of the relative abundance of *Peptostreptococcus* for samples derived from healthy volunteers and CRC patients from the network. The upper four plots have a normal axis; the lower four plots have a logarithmic axis to enable visualisation of low-abundance samples.

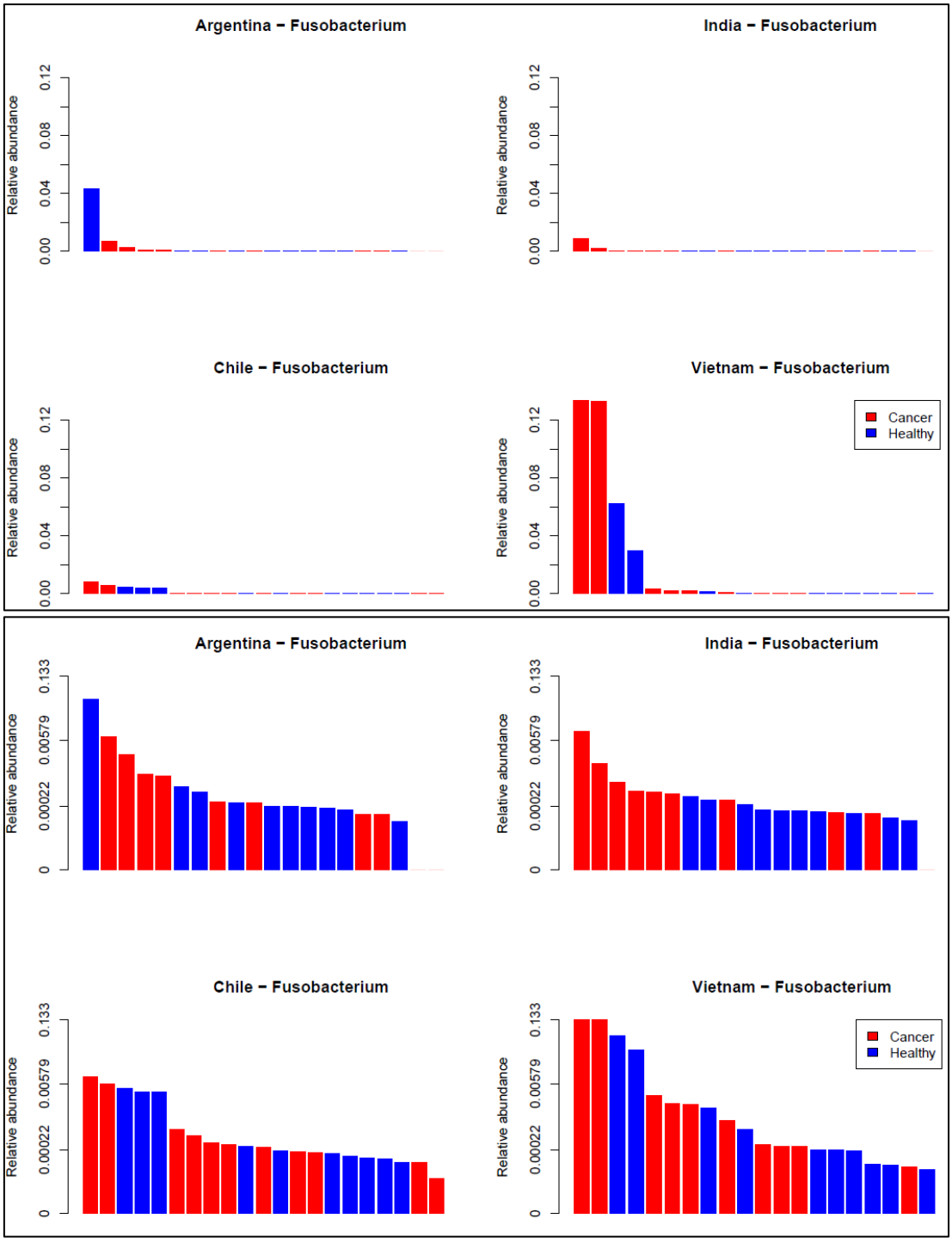


Figure 158. Waterfall plots of the relative abundance of *Fusobacterium* for samples derived from healthy volunteers and CRC patients from the network. The upper four plots have a normal axis; the lower four plots have a logarithmic axis to enable visualisation of low-abundance samples.

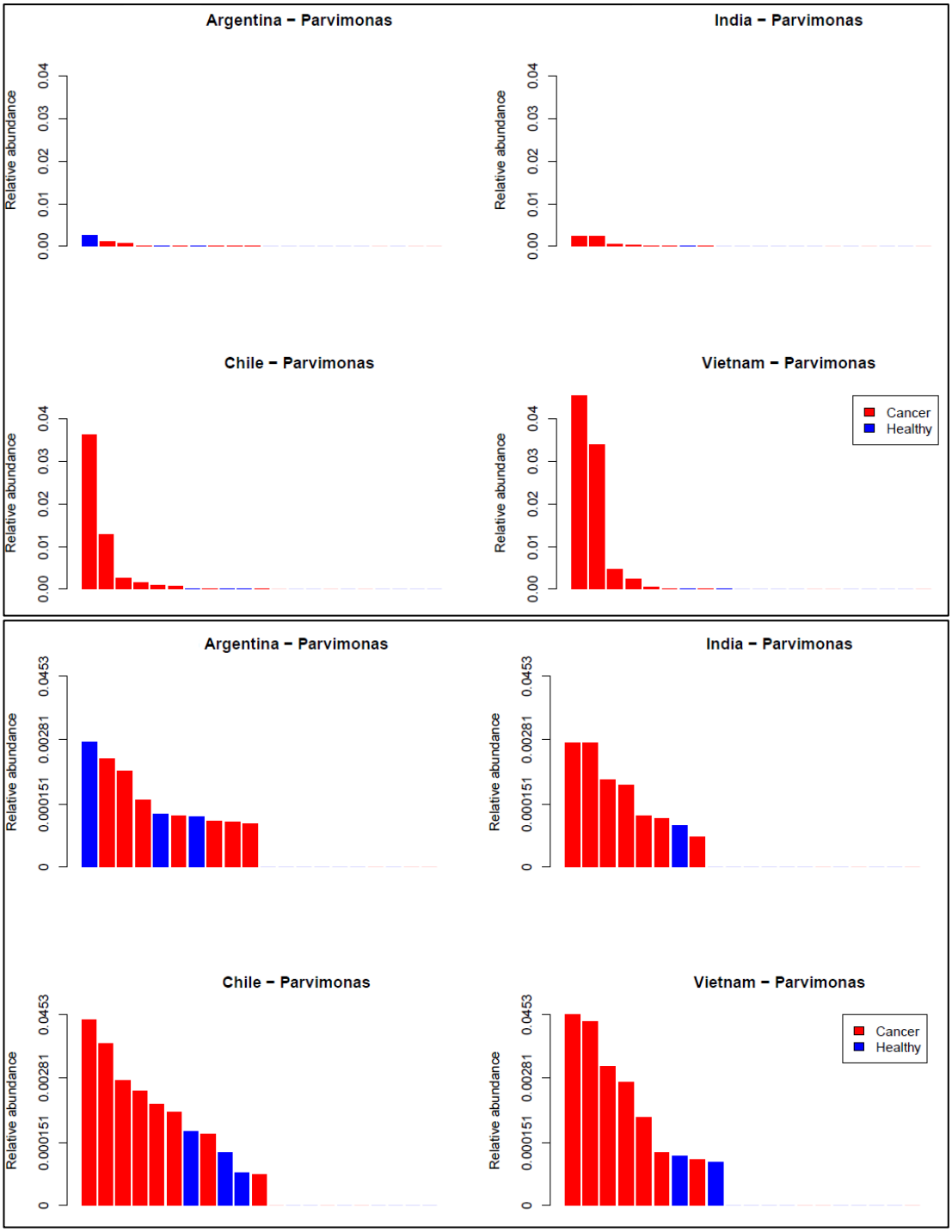


Figure 159. Waterfall plots of the relative abundance of *Parvimonas* for samples derived from healthy volunteers and CRC patients from the network. The upper four plots have a normal axis; the lower four plots have a logarithmic axis to enable visualisation of low-abundance samples.

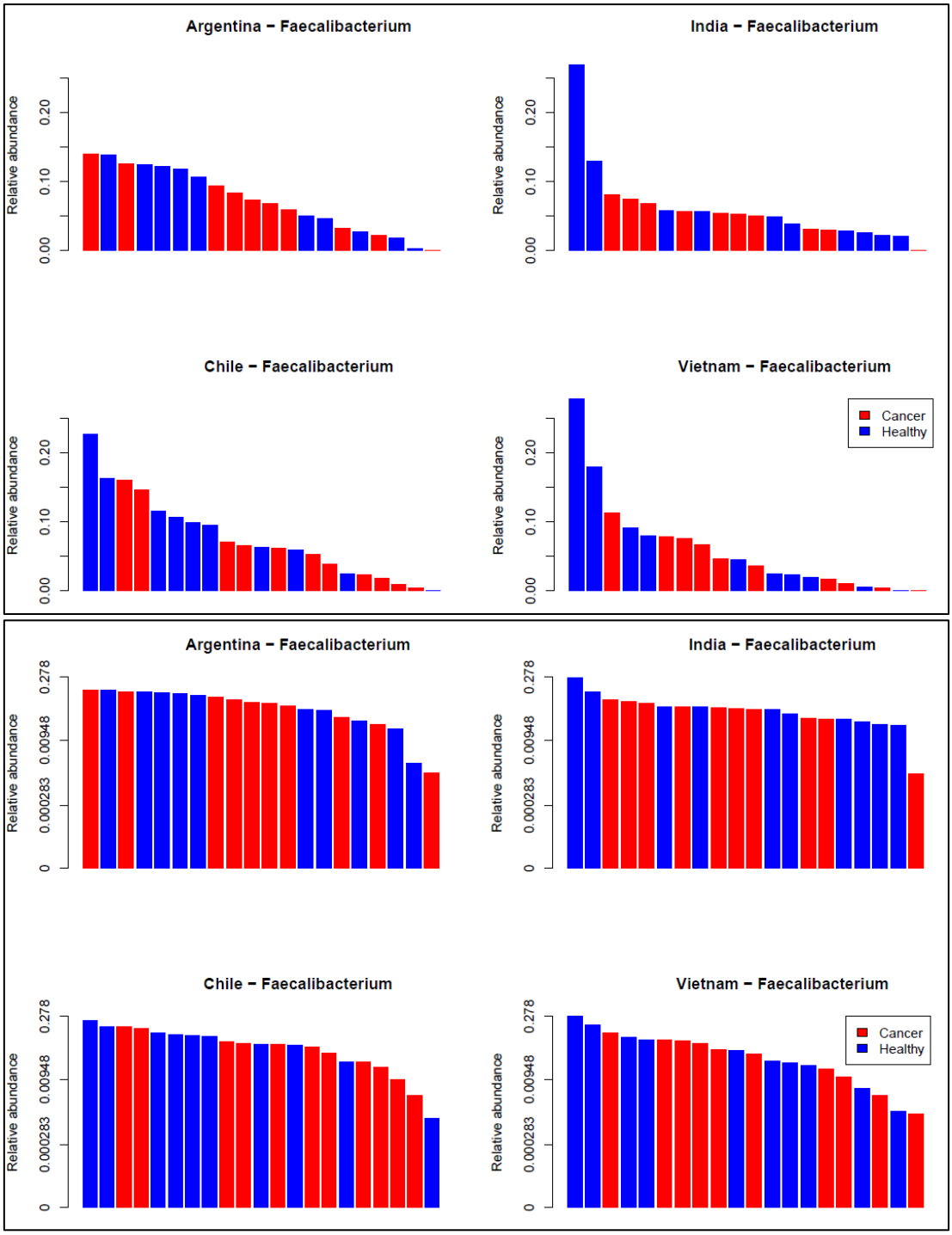


Figure 160. Waterfall plots of the relative abundance of *Faecalibacterium* for samples derived from healthy volunteers and CRC patients from the network. The upper four plots have a normal axis; the lower four plots have a logarithmic axis to enable visualisation of low-abundance samples.

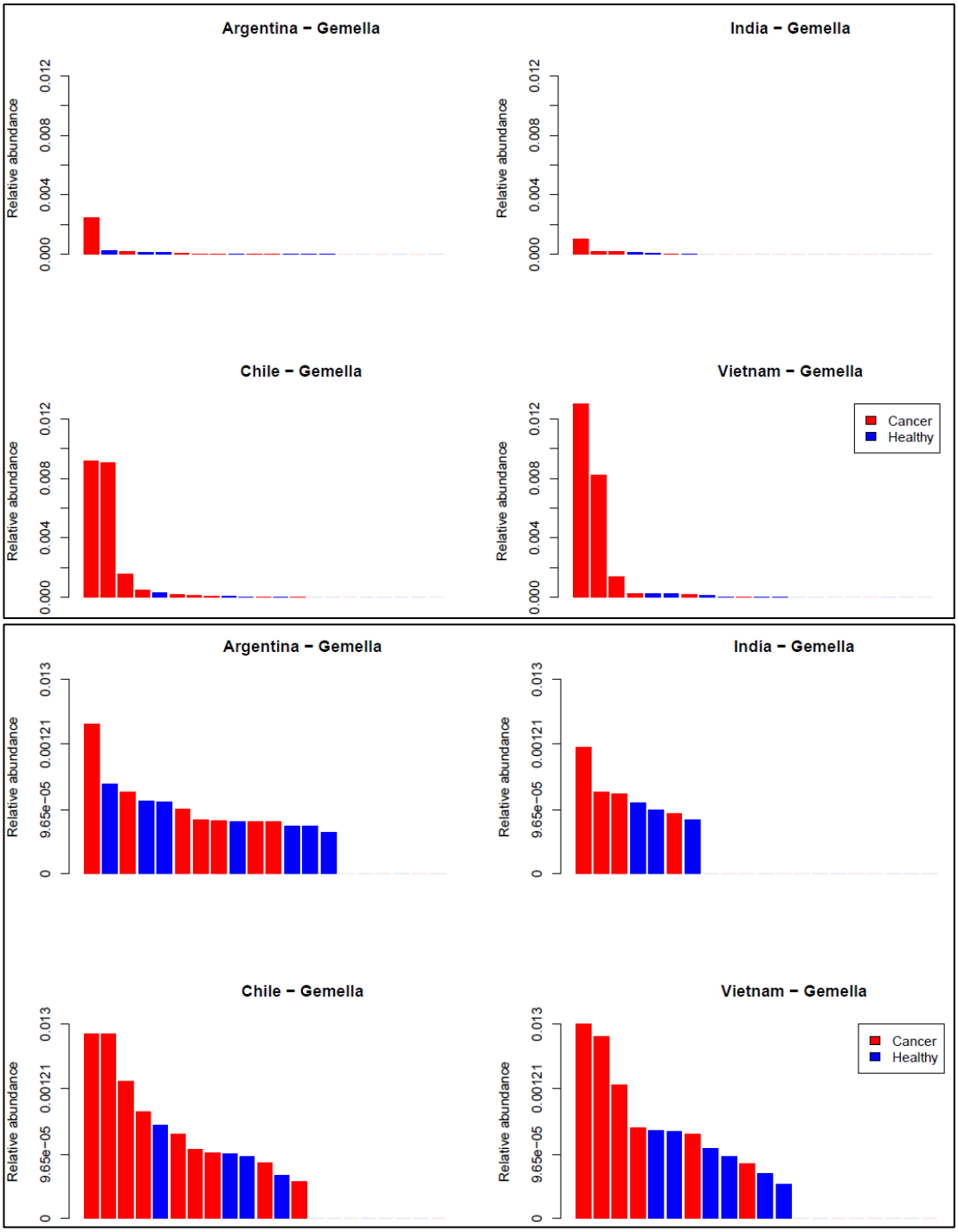


Figure 161. Waterfall plots of the relative abundance of *Gemella* for samples derived from healthy volunteers and CRC patients from the network. The upper four plots have a normal axis; the lower four plots have a logarithmic axis to enable visualisation of low-abundance samples. e = x 10^.

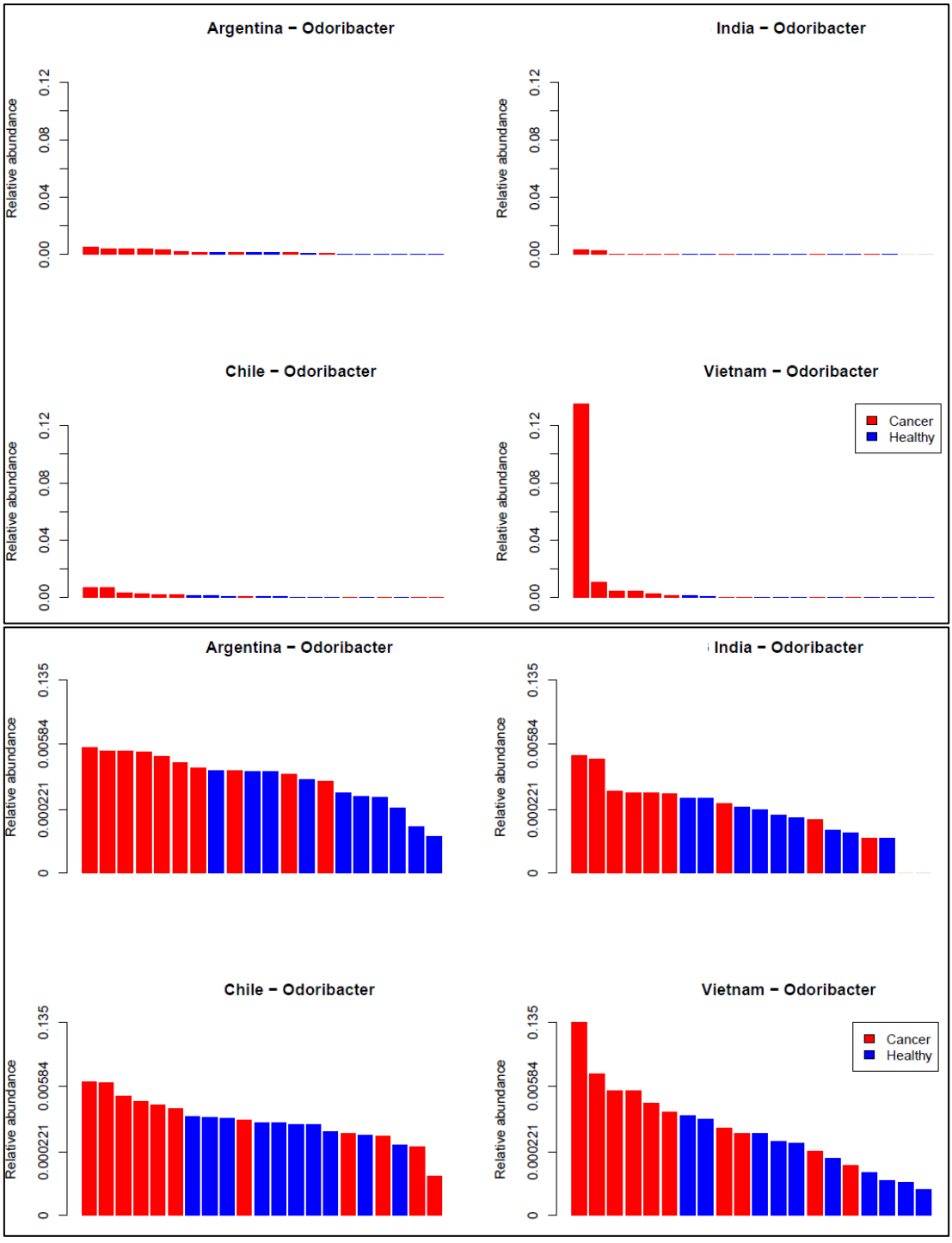


Figure 162. Waterfall plots of the relative abundance of *Odoribacter* for samples derived from healthy volunteers and CRC patients from the network. The upper four plots have a normal axis; the lower four plots have a logarithmic axis to enable visualisation of low-abundance samples.

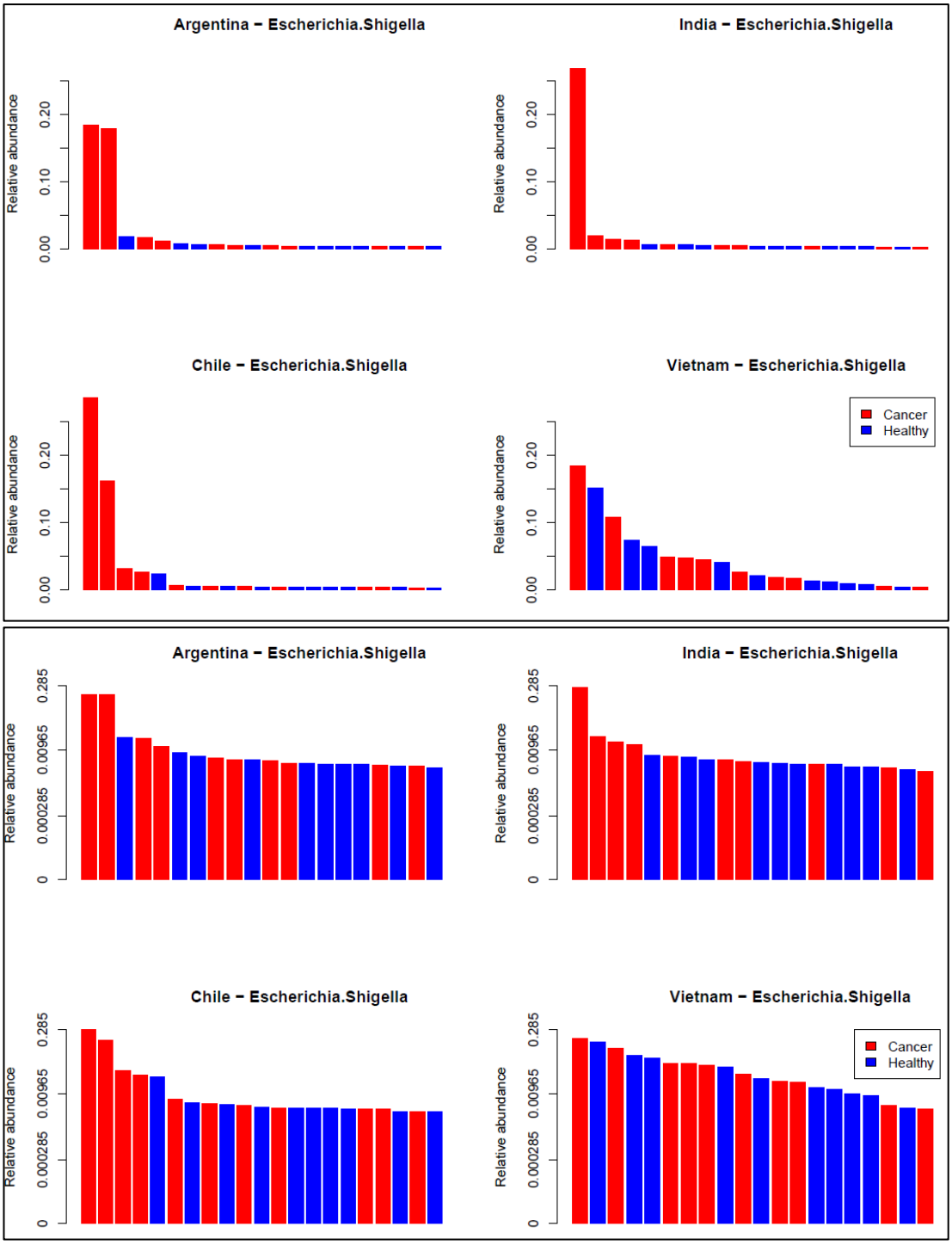


Figure 163. Waterfall plots of the relative abundance of *Escherichia-Shigella* for samples derived from healthy volunteers and CRC patients from the network. The upper four plots have a normal axis; the lower four plots have a logarithmic axis to enable visualisation of low-abundance samples.

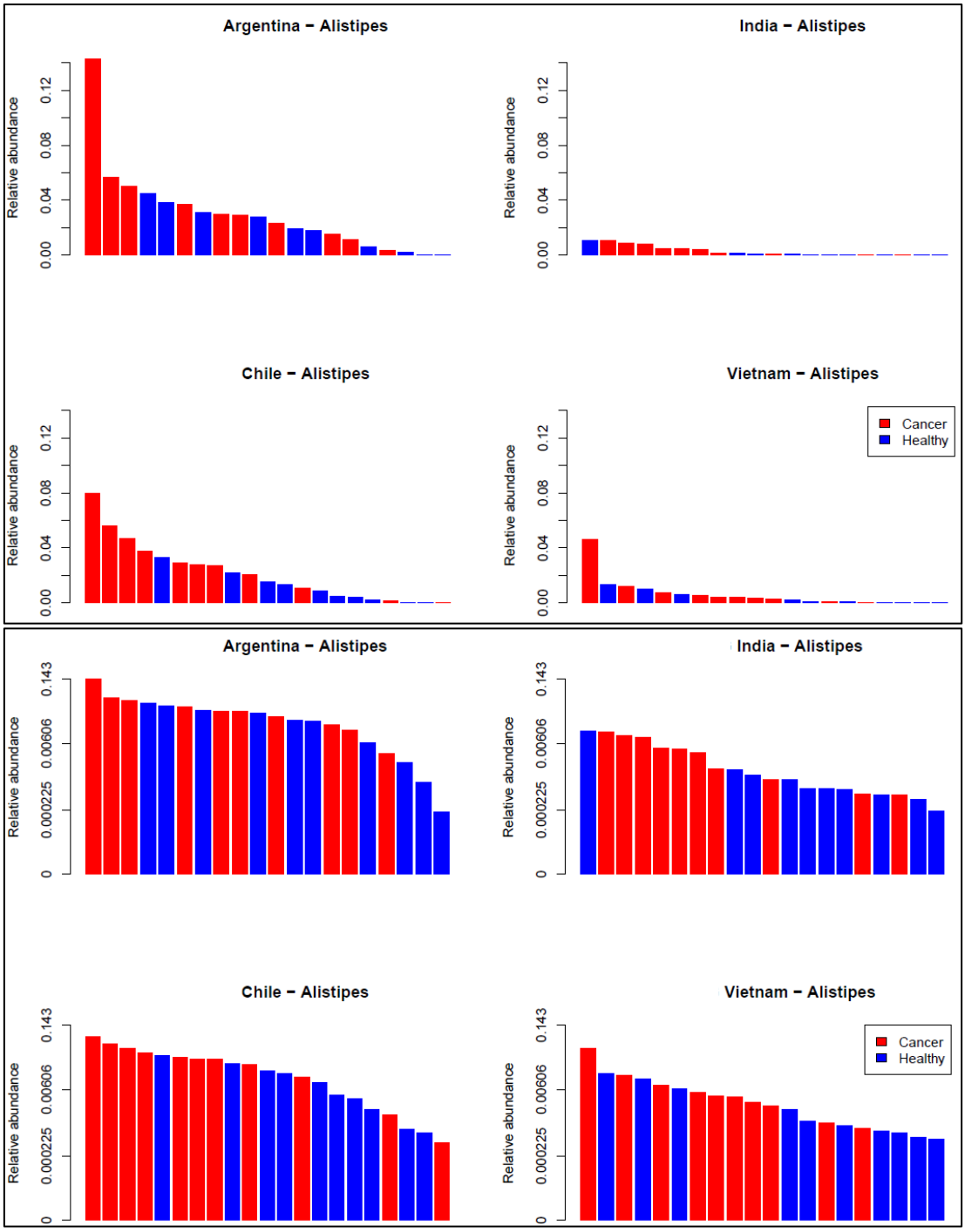


Figure 164. Waterfall plots of the relative abundance of *Alistipes* for samples derived from healthy volunteers and CRC patients from the network. The upper four plots have a normal axis; the lower four plots have a logarithmic axis to enable visualisation of low-abundance samples.

4.5 Discussion

4.5.1 Microbiome analysis of cohorts from non-Western countries using gFOBT

This is the first study to establish a global microbiome research network to investigate the CRC-associated microbiome of non-Western countries (Argentina, Chile, India and Vietnam). To date, limited microbiome research has been conducted in these countries due to difficulties of collection, storage and transport of frozen whole stool samples and the expense of sample processing.

One technical microbiome study had demonstrated that FTA cards, which are similar to gFOBT but contain a nucleic acid stabiliser, afforded high stability up to eight weeks despite marked fluctuations in temperature (4-40°C) (435) and another demonstrated that gFOBT had high stability after four days storage at ambient temperature in a technical microbiome study conducted in Bangladesh (436). Both studies were conducted using stool from healthy volunteers and did not assess the effect of international transport on the stability of the microbiome. This study aimed to assess whether the microbiome of gFOBT samples would be stable during prolonged storage abroad and international transport, both at ambient temperature.

The CRC-associated microbiome has been minimally investigated in non-Western countries. This study aimed to address this by performing a preliminary analysis of the microbiomes of ten healthy volunteers and ten CRC patients from each country, using a single standardised methodology to mitigate technical biases.

4.5.1.1 It is possible to perform microbiome analysis of gFOBT samples collected from Argentina, Chile, India and Vietnam

The majority of samples underwent successful library preparation and NGS at the first attempt. Of the first batch of samples, only one sample (1%) failed NGS with fewer than 10,000 reads. The same amplicon failed subsequent sequencing, but the original extracted DNA successfully underwent library preparation and sequencing, indicating that the fault lay with the amplicon rather than the sample. Of the second batch of samples, five samples (3%)

failed library preparation due to an inadequate concentration of PCR amplicons; they will be sequenced on the next available sequencing run. These results indicate that V4 16SrRNA sequencing can be successfully performed on DNA extracted from gFOBT samples received from Argentina, Chile, India and Vietnam.

4.5.1.2 Storage of gFOBT samples at ambient temperature abroad and transport to the UK has minimal effect on microbiome results

The microbiome of UK control samples that were sent abroad showed a similar amount of variability to the UK control samples that remained in the UK, indicating that microbiome variability was secondary to technical factors such as stool subsampling or laboratory processing, rather than due to storage abroad and transport to the UK. Importantly storage abroad and transport to the UK had no appreciable effect on the relative abundance of CRC-associated taxa. Once the Random Forest model (described in Chapter 3) is validated, these samples will be run through it to determine whether storage abroad and transport to the UK has any effect on sample classification. It should be borne in mind that only the CRC-associated taxa which were of greatest importance to the NHSBCSP-based Random Forest models described in Chapter 3 were assessed. Future work will increase the number of samples collected from these Institutes and then create Random Forest models to determine which taxa are most discriminatory for CRC for the non-Western cohort as a whole and for each individual country. The effect on these taxa of storage abroad and transport to the UK will then be assessed. Transport and storage abroad had no appreciable effect on the relative abundance of *Escherichia-Shigella*.

A potential limitation is that control samples were derived from healthy volunteers from the UK; the microbiomes and range of relative abundances of CRC-associated taxa encompassed by the control samples may therefore not be representative of samples derived from CRC patients from abroad. The control samples were stored in the UK for some time prior to transport abroad, they may therefore not be representative of samples collected and subsequently transported within a short period of time; however, this is appropriate for the current study as the majority of samples collected abroad were stored for approximately two months prior to transport to the UK. These limitations will be partly addressed by Phase 2 of the project, currently

underway, in which a subset of healthy volunteer and CRC samples processed in India will also be sent to the UK for processing. As part of the expansion-phase of the project, each Institute will be asked to create replicate gFOBT cards from CRC and healthy volunteers. This will allow confirmation that international transport has minimal effect on the relative abundances of non-Western CRC-associated taxa and assessment of the effect of long term storage at ambient temperature within each Institute.

It should be borne in mind that the maximum laboratory temperatures recorded by the four Institutes ranged from 20°C to 27°C; unfortunately it was not possible to record temperature during transport. The conclusion that storage of gFOBT samples at ambient temperature abroad has minimal effect on microbiome results is therefore confined to laboratories with maximum temperatures within this range, the results cannot necessarily be generalised to the use of gFOBT in hotter climates. The same technical validation will therefore need to be performed for all Institutes which subsequently join the network.

4.5.1.3 Prolonged storage of gFOBT samples at ambient temperature in the UK has minimal effect on microbiome results

Once samples were received by the Leeds laboratory, DNA was extracted within three days. When the project expands, it may not be possible to process large numbers of samples as quickly. It was therefore important to assess the effect on the microbiome of prolonged storage at ambient temperature in the UK. Further rationale for this investigation is the fact that only three squares from each sample were dissected, leaving the remaining squares available for future analysis provided that the microbiome remains stable. Storage at -20°C or -80°C was not performed as this would require considerable freezer space when the project expands and would not be possible in some non-Western countries (the aim is for microbiome analysis to ultimately be performed independently abroad.)

Three squares of faecally-loaded card, one from beneath each of the three flaps, were dissected and processed as a single combined sample; after prolonged storage at ambient temperature the alternate three squares were processed similarly. Whilst each gFOBT card was loaded with a single stool

sample, the stool was not homogenised prior to creating the gFOBT cards and therefore a limitation of this approach is that any variability in the microbiome between replicates could be due to a combination of biological variation (due to subsampling) and technical variation (storage duration). Stool homogenisation was not included in the study protocol, as participants were given the option of loading the gFOBT card themselves or providing the abroad Institute with a stool sample which was loaded onto a gFOBT card by laboratory staff; it would not have been appropriate to ask participants to homogenise their stool samples. Furthermore, it is unclear to what extent manual homogenisation negates variability between replicates; the gFOBT samples used for the FIT experiment (Chapter 2) which were derived from manually homogenised stool still exhibited a degree of microbiome variability. In the expansion phase of the current project, laboratories that prepare gFOBT cards using stool samples will be asked, after they have created the sample for the study, to manually homogenise the remaining stool and to create replicate gFOBT samples. This will allow better assessment of the temporal stability of the microbiome of gFOBT samples stored for prolonged periods either within the abroad Institutes or the UK.

Replicates had similar microbiomes and relative abundances of CRC-associated taxa. These findings will be confirmed by running the samples through the Random Forest model (described in Chapter 3) once it is validated, to determine whether there is any effect on sample classification. As outlined above, the aim of future work is to generate Random Forest models to determine which taxa are most discriminatory for CRC for the non-Western cohort as a whole and for each individual country. The effect on these taxa of prolonged storage abroad or in the UK will then be assessed.

LEfSe analysis did not detect any significant difference in taxa between a group comprising one member from each replicate pair and another group containing the second member from each replicate pair. This indicates that storage at room temperature in the UK up to 252 days after sample receipt does not systematically alter microbiome results compared with extraction within three days of sample receipt.

These findings are only valid for storage within the UK and should be confirmed in each Institute as described above. However, given that the

recorded laboratory temperature of each Institute was similar to that expected of the UK (excepting perhaps India), it is likely that the results will be similar. It will be important to confirm this during the expansion-phase of the project as samples will be stored abroad for longer given that a greater number of samples will be collected.

4.5.2 Analysis of the microbiome of participants from Argentina, Chile, India and Vietnam

4.5.2.1 Differences exist between the microbiomes of participants from the four countries

Differences between the microbiome of Western and non-Western populations have been described in the existing literature, but differences between non-Western countries have been less well studied. The advantage of this study was that a single standardised methodology was used.

Due to small sample numbers, for each country samples from healthy volunteers and CRC patients were combined to create a collective set. In future, with larger sample numbers it will be possible to investigate inter-country differences of the microbiome for healthy volunteers and CRC patients separately.

The alpha diversities of the Vietnamese and Indian samples were significantly lower than the alpha diversities of the Argentinian and Chilean samples. The alpha diversity of the Indian samples was significantly lower than the Vietnamese samples. This was an unexpected finding, as Argentinian and Chilean diets are more Westernised than those of India or Vietnam and, as outlined in the Chapter introduction, Western microbiomes have been associated with lower alpha diversity. This finding may be secondary to the small number of samples, the fact that results from CRC patients and healthy volunteers were combined, or due to inter-country differences in participant or tumour characteristics. However, two studies have also demonstrated a lower alpha diversity (as measured by Shannon diversity index) of faecal samples from healthy Indians compared with cohorts from the USA, Denmark, China, Japan and Finland (678, 681). Whilst the findings from these studies could

potentially be due to technical differences between cohorts, if they represent a genuine biological difference this would suggest that the concept that non-Western microbiomes have higher alpha diversity than Western microbiomes may be overly-simplistic, and may only apply to certain measures of alpha diversity or certain countries.

There was no significant difference in alpha diversity between the combined healthy volunteer group and the combined CRC group. These findings will need to be confirmed with a larger sample size, but are in keeping with the results of the NHSBCSP study described in Chapter 3, in which no difference in alpha diversity was identified between CRC and blood-negative samples. However, when comparing results between the two studies the following should be borne in mind:

- The NHSBCSP gFOBT were collected from bowel preparation-naïve individuals, whereas many of the abroad samples were collected post-bowel preparation (albeit after a minimum of 14 days).
- The gFOBT status of the NHSBCSP samples was known; faecal blood was not assessed for the abroad samples.
- The NHSBCSP samples were a combination of subsamples from three separate stools, whereas the abroad samples were subsamples from a single stool.

The majority of variation in beta diversity (as measured by Bray-Curtis distance) was unaccounted for, reflecting inter-individual variation. Country of origin contributed to variation in Bray-Curtis distance with $R^2 = 0.14$. Disease status and gender also contributed to variation in Bray-Curtis distance, each with $R^2 = 0.02$. The amount of variation accounted for by disease status is similar to that of the NHSBCSP samples (described in Chapter 3), although gender accounts for more variation in the current study than it did in the NHSBCSP study. The findings are similar to a study which compared the microbiomes of subjects from Western and non-Western countries (Columbia, America, Europe, South Korea and Japan), in which country of origin accounted for a greater amount of variation ($R^2 = 0.22$) than BMI or gender ($R^2 = 0.04$ and 0.05 respectively) (685).

The finding that samples from Argentina and Chile appeared closer together on the PCA of Bray-Curtis distances is a similar finding to that of two other

studies which demonstrated greater similarity between the faecal microbiomes of healthy volunteers from Argentina and Chile compared with healthy volunteers from the USA, Italy, Papua New Guinea and the Matses of Peru (684, 686).

Samples from India and Vietnam had on average a higher relative abundance of *Prevotella* and lower relative abundance of *Bacteroides* than samples from Chile and Argentina. This finding is in keeping with the expected dietary differences between the countries; as outlined in the Chapter introduction, plant-based diets associate with a higher relative abundance of *Prevotella* whereas meat-consumption associates with a higher relative abundance of *Bacteroides*. Studies which have analysed the microbiomes of healthy Indians have similarly demonstrated high relative abundances of *Prevotella* (676, 678, 680-682). Studies of healthy Argentinians and Chileans have similarly demonstrated high relative abundances of *Bacteroides* (683, 684). Surprisingly, the questionnaire data indicated that there were very few vegetarians, of which the majority were healthy volunteers from India. Future work will collect more detailed dietary information through the use of a validated food frequency questionnaire.

4.5.2.2 CRC-associated taxa are identified for each country

For each country, samples from healthy volunteers appeared on average to have a higher relative abundance of *Prevotella* and lower relative abundance of *Bacteroides* compared with samples from CRC patients. This is an interesting observation which needs to be confirmed with a larger sample size of matched samples, as differences between the cases and controls (such as age, gender or colonoscopy-history) could be partly accountable. The difference in the relative abundance of *Prevotella* and *Bacteroides* between cases and controls was not identified by LEfSe, perhaps because the genus *Prevotella* contains several members, some of which have been described as CRC-enriched by meta-analyses of faecal studies (*Prevotella_7*, *Prevotella intermedia* and *Prevotella nigrescens*)(121, 166, 477, 496, 534, 597).

The CRC-associated taxa identified by LEfSe analysis of the entire cohort and the individual Chilean, Indian and Vietnamese cohorts included taxa described as CRC-enriched by meta-analyses of faecal studies conducted in the USA,

Canada, Ireland, Austria, Germany, France, Spain, Italy, China and Japan (121, 166, 477, 496, 534, 597). This suggests that the CRC-associated microbiome may be a global phenomenon, or that in the existing meta-analyses, the non-Western microbiome is adequately represented by China and Japan. The fact that the CRC-enriched taxa from the Argentinian cohort did not include taxa described as CRC-enriched by meta-analyses is an interesting finding which requires confirmation with a larger sample size.

For the entire cohort, the genus which was the most significantly enriched in CRC was *Escherichia-Shigella* although the results from Chapter 2 indicate that this may not be appropriate for use as a CRC-screening marker. *Peptostreptococcus* and *Parvimonas* were enriched in CRC for all of the cohorts except Argentina, suggesting that these two taxa may have the potential to be used as a universal CRC-screening marker. However, the maximum relative abundances of CRC-associated taxa varied by country, suggesting that either a country-specific CRC-screening test or country-specific thresholds for a universal CRC-screening test may be required. Once the Random Forest model developed in Chapter 3 has been validated, the non-Western CRC and healthy volunteer samples will be run through it to determine whether the model is able to accurately classify them; high classification accuracy would suggest a universal CRC-associated microbiome, whereas low classification accuracy could be due either to differences between the Western/non-Western CRC-associated microbiome or methodological differences between the studies.

Differences in the CRC-associated taxa between the different countries may be secondary to small sample size, technical differences between the cohorts (for example bowel preparation regimen or laboratory ambient temperature), differences in participant characteristics between the cohorts, or may represent a genuine biological difference. Some evidence for this comes from a study which has indicated that CRC-enriched taxa show a degree of difference according to enterotype (*Bacteroides, Prevotella* or *Escherichia* predominant) (687). Disease-related factors (such as mean age of presentation) and tumour-related factors (such as mean size) appeared to show differences between the countries, although these findings need to be confirmed with larger sample sizes.

Within each of the four countries there is great diversity pertaining to geography, culture, ethnicity, lifestyle, diet and health, all of which have been shown to associate with differences of the microbiome (665, 676-679, 688-691). Such intra-country diversity may be greater for certain non-Western countries, for example India (675), compared with certain Western countries, for example the American Gut Project reported weak associations between the American microbiome and geographical location (413). In order to develop a microbiome-based CRC screening model which is not overfitted to limited geographical regions, future work should expand the geographical catchment of samples within each of the four countries. In India, the LogMPIE study has recently been published; this analysed faecal samples from 1000 healthy volunteers from 14 different geographical regions (676). Unfortunately the collection device used (OMNIgene.GUT) differs from the current study, but it may be possible to collaborate with this project in order to repeat sampling using gFOBT.

In order to expand the range of non-Western CRC-associated microbiomes investigated, institutes in South Africa and Russia have been invited to join the network. Once adequate numbers of samples have been collected (to be determined by a power calculation), LEfSe analysis and Random Forest modelling will be performed to identify CRC-associated taxa and investigate the potential of the faecal microbiome to be used as a CRC-screening tool. Both a universal model and country-specific models will be generated. As described in Chapter 3, additional data such as age, gender, FIT haemoglobin concentration and potentially faecal-mutation, bacterial virulence-factor or toxin testing may improve the accuracy of the models and will be investigated. However, it should be borne in mind that FIT positivity may be affected by the presence of parasites and ambient temperature (692) and CRC-associated faecal bacterial toxins may differ by country (693).

4.5.3 Chapter Summary

- The microbiome of gFOBT samples is stable when samples are stored and transported from abroad (Argentina, Chile, India and Vietnam) at ambient temperature.

- Prolonged storage of gFOBT samples at ambient temperature in the UK has minimal effect on microbiome results

- The microbiomes of samples from Argentina, Chile, India and Vietnam demonstrate differences in alpha and beta diversity.

- The CRC-associated microbiome from these countries contains CRC-associated bacteria described in Western populations, suggesting that certain taxa may be universally associated with CRC.

Chapter 5
Discussion

This thesis has addressed three of the major challenges facing the current field of CRC microbiome research by:

- Confirming that gFOBT cards transported and stored at ambient temperature represent a suitable method of sample collection, compatible with conducting large-scale microbiome research.
- Demonstrating the potential of the faecal microbiome to improve the accuracy of NHSBCSP screening.
- Investigating the CRC-associated microbiome of non-Western countries.

The major findings from the thesis will be briefly discussed and plans to further develop the work will be outlined. A final summary of the results will then be presented (section 5.4).

5.1 Analysing the microbiome from NHSBCSP samples

This thesis has demonstrated that 16SrRNA microbiome analysis can be performed from the faeces of processed NHSBCSP gFOBT samples using a standardised high-throughput methodology. It has demonstrated that microbiome results are stable despite prolonged storage of samples at ambient temperature. This finding is important, as to date the majority of microbiome research has been conducted using whole stool samples transported and stored refrigerated/frozen, limiting study size and subsequent power.

Technical studies, including one conducted by the Leeds group, had indicated that gFOBT might be suitable as a method of sample collection. However, such studies were conducted using samples from small numbers of healthy volunteers and in most studies stability was assessed after only a short period of storage. In contrast, this study has confirmed the stability of the microbiome on gFOBT collected from large numbers of NHSBCSP participants, transported and stored for prolonged periods at ambient temperature. Future work will investigate the underlying mechanism, by attempting to culture

gFOBT samples. If bacteria on gFOBT are shown to be non-viable, this would indicate that temporal variability is secondary to technical factors or contamination rather than bacterial overgrowth, which would reduce the need for technical replicates.

The current study identified differentially-abundant bacteria between clinical groups which, unexpectedly, were found to be affected by the choice of control group (gFOBT blood-negative or colonoscopy-normal). This finding will be further investigated by creating a Random Forest model to discriminate between these samples and by investigating microbiome differences between FIT blood-negative and colonoscopy-normal samples collected contemporaneously.

Marked variability of the relative abundance of *Escherichia-Shigella* was an unexpected finding with potentially important implications; *Escherichia-Shigella* has been identified as enriched in CRC by many faecal studies, including the study described in Chapter 4, yet variability in relative abundance would suggest that it is unsuitable as a screening biomarker. High relative abundances of *Escherichia-Shigella* and *F. nucleatum* were observed for some samples; this raises questions as to whether such high abundances are biologically genuine or technical, which will be further investigated by qPCR, and how they should be treated at analysis.

The potential for the faecal microbiome to improve the accuracy of screening using NHSBCSP samples was investigated for the first time. Prior technical studies had indicated the screening potential of the faecal microbiome. However, they suffered a number of limitations which meant that results were not generalisable or translatable to a national bowel cancer screening programme. These limitations included small study size, study participants not being representative of a screening-eligible cohort, collecting samples post bowel-preparation and collecting samples in a manner which would not be compatible with national screening.

The current study overcame these limitations by performing microbiome analysis of large numbers of processed NHSBCSP gFOBT samples. Microbiome-based screening models improved accuracy over that of gFOBT;

combining gFOBT and microbiome data produced AUC 0.855 (95% CI: 0.832-0.877) for the detection of CRC and AUC 0.868 (95% CI: 0.848-0.886) for the detection of CRC/adenoma. Microbiome-based screening models also showed potential as a second-tier of screening; currently all participants with a blood-positive gFOBT result are referred for colonoscopy; CRC is detected at 10% of colonoscopies, adenoma at 40%, and 50% of colonoscopies reveal a normal bowel or non-neoplastic condition. As a second-tier test, microbiome-based models for the prediction of CRC produced AUC 0.752 (95% CI: 0.714-0.787); microbiome-based models for the prediction of neoplasm produced AUC 0.704 (95% CI: 0.669-0.738); and microbiome-based models for the prediction of a normal-colonoscopy result produced AUC 0.729 (95% CI: 0.692-0.767). The models will be validated shortly (sequencing results are currently pending). Provided that accuracy of the models remains adequate, they confirm the potential of the faecal microbiome to improve the accuracy of NHSBCSP screening.

Once the models are validated, replicate samples described in the previous chapters will be run through the models in order to assess whether differences in the following experimental conditions affect sample classification: sample type, subsampling, sequencing run, or duration of storage at ambient temperature. Once available, the data regarding lesion location and past screening history will be analysed and the effect on the microbiome of season of sample collection will be explored.

Given the ultimate aim of incorporating microbiome-based analysis into the NHSBCSP, a method of automated DNA extraction using the QIAcube HT instrument was investigated and shown to improve DNA-extraction throughput. Sequencing results from replicate samples extracted manually or automatedly are pending. Based on the results, a decision will be made as to whether automated DNA extraction should be adopted. If this is the case, an assessment of well-well contamination with the QIAcube HT will first need to be performed.

As the NHSBCSP is currently replacing gFOBT with FIT, the ability to analyse the microbiome from FIT samples, exposed to the conditions that NHSBCSP FIT samples will experience, was investigated. Technical studies had

indicated that microbiome analysis could be performed from FIT, however no study had simulated NHSBCSP conditions.

After simulated short-term storage and transport, no significant difference between the microbiome of gFOBT or FIT was observed, indicating that microbiome analysis can be performed from NHSBCSP FIT samples. Sequencing results are pending from samples extracted after eight weeks storage; these results will determine how NHSBCSP FIT samples will be stored and transported from the Screening Hub to the processing laboratory. Collection and processing of NHSBCSP FIT samples will then begin. Ethical and NHSBCSP Research Advisory Committee approval has been granted for the analysis of up to 9000 samples (exact numbers to be determined by a power calculation), to be funded as part of a current CRUK Grand Challenge.

An assessment will first be performed to determine the likelihood of sample-sample contamination during FIT processing at the Screening Hub. Providing sample-sample contamination is not identified, 16SrRNA analysis of NHSBCSP FIT samples will be performed according to the current study. As it will not be possible to confirm accuracy, stability or reproducibility of FIT compared with whole stool (as this would disrupt routine screening), this will be assessed using paired FIT and frozen whole stool samples collected as part of a symptomatic FIT trial (currently being planned). Ideally samples will be collected from each of the five Screening Hubs, so that results will be generalisable to the whole of the UK NHSBCSP-eligible population. The possibility of analysing the microbiome of FIT samples collected by the Scottish BCSP will also be explored.

Random Forest models will be created and their accuracy compared with those built using the NHSBCSP gFOBT samples. This will determine whether further development of a microbiome-based screening test should use gFOBT or FIT. Whether the Random Forest models are improved by the addition of age, gender, FIT haemoglobin concentration, or faecal-mutation, bacterial virulence-factor or toxin testing will be assessed. The findings will be used to design qPCR-based screening tests; specificity will be confirmed by testing samples from patients with other diseases.

In all future work, rigorous assessment of contamination, labelling cross-checks or sample barcoding, and matched timing of sample collection will be performed. Guidance on achieving reproducible, replicable, robust and generalisable microbiome results will be followed (694).

5.2 Investigation of the CRC-associated microbiome of non-Western countries

There is currently inequity of CRC-associated microbiome research, with the majority being conducted in Western countries. This is likely due in large part to the cost and logistical implications of cold-chain collection, storage and transport of whole stool samples. This thesis has demonstrated that gFOBT can be used to collect microbiome samples from non-Western countries (Argentina, Chile, India and Vietnam) and that the microbiome is stable when gFOBT are transported to and stored in the UK at ambient temperature.

CRC-associated taxa described by meta-analyses of faecal studies conducted predominantly in Western countries were identified, which suggests that certain taxa may be universally associated with CRC. This is an important finding for two reasons: firstly it indicates that it may be possible to create a universal microbiome-based CRC screening test; and secondly it focuses future research investigating the mechanisms underlying CRC-microbiome associations to a limited number of taxa.

Future work aims to expand the number of samples collected (to be determined by a power calculation) and the range of non-Western microbiomes analysed, both within each country and by expanding the network to include Institutes in South Africa and Russia. This work will form part of a current CRUK Grand Challenge. The network will meet in Leeds in November 2019 to discuss the results of the current study and to plan this expansion phase of the project. Random Forest modelling will be performed to identify CRC-associated taxa which are unique to each country and those which are present across the network. The potential of universal or country-specific microbiome-based screening tests will be explored.

It will be important to perform ongoing sample collection from the non-Western countries as their level of Westernisation and incidence of CRC changes.

The second phase of the project, in which a subset of healthy volunteer and CRC samples processed in India will also be sent to the UK for processing, is currently underway. If there is no appreciable difference in results, this would suggest that sample processing and analysis may be undertaken by the respective Institutes; promoting independent microbiome research within these countries is an important means to address the current disparity in CRC-associated microbiome research.

5.3 Additional studies

The studies described in this thesis performed 16SrRNA analysis of the microbiome, owing to the reduced price/sample compared with metagenomic analysis, and as the majority of previous studies investigating the screening potential of the microbiome performed 16SrRNA analysis. Three squares from each gFOBT sample remain available for alternative analysis (e.g. metabolomic analysis or investigation of the virome/mycobiome). These samples therefore serve as a form of 'microbiome biobank'; a valuable resource to test hypotheses as knowledge of the CRC-associated microbiome progresses.

Three other related areas of investigation are currently being undertaken by the author and will be briefly outlined:

- Investigation of the microbiome of patients who develop CRC at a younger age (less than 50)
- Investigation of the microbiome of small intestinal adenocarcinoma
- Investigation of the microbiome associated with tumours arising in patients with Lynch syndrome

The incidence of CRC diagnosed in people aged less than 50 has increased in the USA over the past forty years (695). CRC is typically diagnosed at a later stage in this age group and has a poor prognosis (696). Although it is unclear why the incidence of CRC is increasing in younger adults, a birth cohort effect has been reported which suggests that an environmental

component is contributing. The CRC-associated microbiome has not yet been investigated in younger adults; one study performed LEfSe analysis of the off-tumour mucosal microbiome of young onset CRC patients compared with controls and identified differentially abundant taxa, however the numbers in each group were small (689).

As early onset CRC demonstrates a number of differences compared with later onset CRC, research investigating the microbiome in patients aged over 50 may not necessarily translate to patients aged less than 50 (697, 698). Funding has been received from Bowel Cancer UK to investigate the faecal microbiome of CRC patients aged under 50, compared with CRC patients aged over 50 and controls. The study aims to provide insight into the mechanism of CRC development in this age group and explore the potential of a microbiome-based diagnostic test. gFOBT samples, questionnaire and food diary data will be analysed. The study has NIHR portfolio status and recruitment is currently underway.

Whether bacteria influence the development of small intestinal adenocarcinoma has not yet been investigated. Low-coverage copy number whole genome sequencing has been performed on DNA extracted from archival FFPE small intestinal tumours; results are pending. The results will indicate whether there is a small intestinal adenocarcinoma-associated microbiome and if so, how it compares with CRC-associated taxa identified using the same methodology.

The microbiome of patients with Lynch Syndrome has been little investigated; a small pilot study has been conducted by another group, however a limitation is that faecal samples were collected at least one year post CRC resection (699). A study is currently being planned to compare the tissue and faecal microbiome of Lynch Syndrome CRC with the tissue and faecal microbiome of sporadic CRC, in order to determine whether the CRC-associated microbiome differs between the two. This will broaden understanding of the CRC microbiome association, with potential implications for microbiome diagnostics and therapeutics.

5.4 Summary of findings

In summary, this thesis has advanced CRC-associated microbiome research by confirming a method to conduct large-scale, single-methodology faecal microbiome studies (gFOBT); by demonstrating that microbiome analysis can be integrated into the NHSBCSP to improve screening accuracy; and by illustrating a method to investigate the CRC-associated microbiome in non-Western countries, with preliminary results indicating that certain CRC-associated taxa may be universal.

The results from this thesis are summarised:

Investigating the potential to use NHSBCSP samples for microbiome analysis

- The microbiome can be successfully analysed from processed NHSBCSP gFOBT samples.

- The microbiome is stable if NHSBCSP gFOBT samples are stored for prolonged periods at room temperature.

- The relative abundances of CRC-associated taxa demonstrate minimal temporal variation. The relative abundance of *Escherichia-Shigella* demonstrates marked variation potentially secondary to technical factors such as subsampling or temporal variation; this suggests that *Escherichia-Shigella* is not a useful CRC screening biomarker.

- The microbiome can be successfully analysed from the FIT devices which the NHSBCSP will use, after simulation of the conditions that NHSBCSP FIT samples will be exposed to.

- Microbiome analysis of NHSBCSP samples can be performed at scale.

Investigating the potential of the microbiome to improve the accuracy of CRC screening

- The microbiome was successfully analysed from NHSBCSP gFOBT samples; this represents a large cohort of bowel preparation-naïve individuals with confirmed colonoscopy diagnosis.

- Small but significant differences of alpha and beta diversity were identified between different clinical groups.

- CRC and adenoma-associated bacteria described in the existing literature were identified.

- Choice of control group (blood-negative or colonoscopy-normal) was found to affect the taxa identified as enriched/depleted in CRC or adenoma.

- A high relative abundance of *Fusobacterium* was identified in a small number of samples.

- Microbiome-based screening models generated using NHSBCSP samples compared favourably with those described in the existing literature and were shown to improve the accuracy of screening.

Investigating the CRC-associated microbiome of non-Western countries

- The microbiome of gFOBT samples is stable when samples are stored and transported from abroad (Argentina, Chile, India and Vietnam) at ambient temperature.

- Prolonged storage of gFOBT samples at ambient temperature in the UK has minimal effect on microbiome results

- The microbiomes of samples from Argentina, Chile, India and Vietnam demonstrate differences in alpha and beta diversity.

- The CRC-associated microbiome from these countries contains CRC-associated bacteria described in Western populations, suggesting that certain taxa may be universally associated with CRC.

Appendix A: Ethical Approvals

2015-2019 Studying the microbiome using FFPE material IRAS Project ID: 187973 REC:15/SW/0355.

2018 University of Leeds Large bowel microbiome disease network. Creation of a proof of principle exemplar in colorectal cancer across three continents MREC17-077 (plus amendment).

2016-2020 Can microbiome data improve the NHS Bowel Cancer Screening Programme? IRAS Project ID: 188007 REC: 16/NE/0210 ODR reference: ODR1617_126 (plus three approved amendments).

2018-2020 Understanding bowel cancer in people aged less than 50 years – investigating changes to the microbiome. IRAS Project ID: 247212 REC: 18/NW/0647

Appendix B: Grants

2018-2019 Bowel Cancer UK pilot grant: Understanding bowel cancer in people aged less than 50 years – investigating changes to the microbiome. £25,000. **C Young**, P Quirke, H Wood, N West, M Morris, E Morris, P Wheatstone, J Whelpton.

2018-2019 Academy of Medical Sciences Global Challenges Research Fund Networking Grant: Large bowel microbiome disease network. Creation of a proof of principle exemplar in colorectal cancer across three continents. £24,910. S Ramakrishnan, P Quirke, **C Young**, H Wood, P Nang, L Contreras, C Vaccaro.

2018 Pathological Society of Great Britain & Ireland Visiting Fellowship to the Meyerson and Huttenhower Laboratories to learn how to perform bioinformatic and statistical analysis of complex microbiome datasets. £4125. **C Young**.

2017-2018 Cancer Research UK Leeds Centre, Future Leader Award. Research Training for Cancer Healthcare Professionals. £10,000. **C Young**.

2017-2018 International Research Collaboration Award, University of Leeds. Award to visit members of the International Network for Cancer Screening Evaluation (INCaSE). £2056. **C Young**.

2016-2019 Wellcome Trust Research Training Fellowship: Investigating the potential of the microbiome to improve the accuracy of the NHS Bowel Cancer Screening Programme (NHSBCSP). £196,056. **C Young**.

2016-2017 Funding to undertake three modules of the University of Manchester's 'Genomic medicine' masters from the Health Education England Genomics Education Programme. £3000. **C Young**.

Appendix C: Publications, presentations, abstracts and prizes

Publications

Young, C., Wood, H., Quirke, P. A new approach to bowel cancer: could bowel cancer be a bacterial disease? Recommendations to reshape policy making. Colorectal Cancer Report. Government Gazette. 2018, vol 3, pp. 60-61.

Young, C., Quirke, P. The potential of the microbiome for colorectal cancer screening. Clinical Laboratory International. Dec 2016/Jan 2017, vol 40, pp. 14-17.

Presentations

2019 Investigating the potential of the faecal microbiome to improve colorectal cancer screening. **C Young**; H Wood; A Fuentes Balaguer; S Benton; C Burtonwood; M Brealey; P Quirke. Leeds Pathology 2019. 12th Joint Meeting of the British Division of the International Academy of Pathology and the Pathological Society of Great Britain & Ireland, 2–4 July 2019. The Journal of Pathology. 249(S1), pp.S1-S59.

2019 The colorectal microbiome: its potential to change practice. Leeds Pathology 2019. 12th Joint Meeting of the British Division of the International Academy of Pathology and the Pathological Society of Great Britain & Ireland, 2–4 July 2019. Invited presentation.

2019 Presentation of results from the thesis. Faculty of Medicine and Health Postgraduate Research Conference, University of Leeds. Third prize.

2019 Creating a global microbiome colorectal cancer (CRC) research network. The 108th Annual meeting of the Japanese Society of Pathology, Tokyo.

2019 Presentation of results from the thesis. Postgraduate Research Symposium at St James', University of Leeds.

2019 Presentation of results from the thesis. Inspiring the Next Generation Yorkshire & Humber Academic Trainee Network.

2018 Presentation of results from the thesis. NHSBCSP Southern Hub Annual Conference.

2018 Presentation of microbiome research proposal. The Junior Academy of the German Society of Pathology, Germany.

2018 Presentation of results from the thesis to the NHSBCSP Southern Hub.

2018 Presentation of results from the thesis to the European Society of Pathology Junior Academy Meeting, Belgium.

2018 Presentation of results from the thesis to the Leeds Medico-Chirurgical Society. Awarded the Charles Chadwick medal.

2018 Delivery of material at the educational microbiome workshop of the GCRFNG Global Microbiome Network, University of Leeds.

2017 Three minute thesis. National Academic Trainee Network meeting, Pathological Society of Great Britain & Ireland. Distinction.

2016 The Microbiome. Leeds Microbiology Group, Leeds General Infirmary.

Abstracts

The creation and validation of a global microbiome colorectal cancer research network. **C Young**; H Wood; AS Ramakrishnan; PV Nang; C Vaccaro; L Contreras Melendez; M Bose; M Doi; T Piñero; C Tapia Valladares; J Arguero; A Fuentes Balaguer; P Quirke. Leeds Pathology 2019. 12th Joint Meeting of the British Division of the International Academy of Pathology and the Pathological Society of Great Britain & Ireland, 2–4 July 2019. The Journal of Pathology. 249(S1), pp.S1-S59. This poster was also presented at the 'Mutographs' team CRUK Grand Challenge Meeting, at the International Agency for Research on Cancer, 2019.

Investigating the potential of the faecal microbiome to improve colorectal cancer screening. **C Young**, H Wood, A Fuentes Balaguer, S Benton, C Burtonwood, M Brealey, P Quirke. Yorkshire and the Humber Academic Presentation Day 2019. First prize.

Investigating the effects of radiotherapy on the bowel cancer microbiome: reanalysing the MRC CR07 trial. HM Wood; **C Young**; D Bottomley; NP West; A Meade; D Sebag-Montefiore; P Quirke. Leeds Pathology 2019. 12th Joint Meeting of the British Division of the International Academy of Pathology and the Pathological Society of Great Britain & Ireland, 2–4 July 2019. The Journal of Pathology. 249(S1), pp.S1-S59.

Comparison of two methods to analyse components of the microbiome from FFPE CRC tissue: low coverage WGS and qPCR. **C Young**; H Wood; A Fuentes Balaguer; S Richman; E Tinkler-Hundal; K Southward; P Quirke. Leeds Pathology 2019. 12th Joint Meeting of the British Division of the International Academy of Pathology and the Pathological Society of Great Britain & Ireland, 2–4 July 2019. The Journal of Pathology. 249(S1), pp.S1-S59.

Quantification of *Fusobacterium* in formalin fixed paraffin embedded colorectal carcinoma tissue from the QUASAR and FOCUS4 clinical trials. **Caroline Young**, Alba Fuentes Balaguer, Susan Bullman, Henry Wood, Susan Richman, Gemma Hemmings, Gordon Hutchins, Richard Gray, Tim Maughan, Matthew Meyerson, Philip Quirke; Yorkshire and the Humber Academic Presentation Day 2018. Third prize.

Prizes

2019 First prize in the Out of Programme Researcher/Clinical Lecturer poster category of the Yorkshire and the Humber Academic Presentation Day.

2019 Third prize for an oral presentation of results from the thesis at the Faculty of Medicine and Health Postgraduate Research Conference, University of Leeds.

2018 Awarded the Charles Chadwick gold medal for the presentation of research to the Leeds Medico-Chirurgical Society.

2018 Third prize in the Out of Programme Researcher poster category of the Yorkshire and the Humber Academic Presentation Day.

2018 Selected to attend the European Society of Pathology Junior Academy Meeting, Belgium.

2018 Selected to attend the Junior Academy of the German Society of Pathology, Germany.

2017 Distinction for ‘three minute thesis’ presentation at the January 2017 National Academic Trainee Network Meeting of the Pathological Society of Great Britain & Ireland.

Appendix D: Summary Feedback from the Global Challenges Research Fund Network Grant (GCRFNG) sponsored Microbiome Network Workshop

Attendance

The following members of the network attended the workshop:

India: Dr Ramakrishnan (CRC surgeon) and Dr Bose (Biologist)

Chile: Dr Contreras Melendez (Pathologist) and Mr Tapia Valladares (Biologist)

Argentina: Dr Vaccaro (CRC surgeon) and Dr Piñero (Biologist)

Vietnam: Dr Nang (CRC surgeon) and Dr Doi (CRC surgeon)

UK: Professor Quirke (Pathologist), Dr Wood (Bioinformatician), Dr Young (Pathologist) and Dr Keigo Murakami (Pathologist visiting from Japan)

Workshop content

The workshop comprised a mixture of presentations, practical demonstrations and interactive workshops. The following was covered:

Introduction: Outlined the role and scope of the GCRFN grant and the current state of microbiome-CRC research.

Laboratory skills workshops: The group was given a tour of the laboratory and health & safety induction and then given the chance to familiarise with basic laboratory skills (as some of the participants did not have previous laboratory experience). The group then split into two and was shown how samples are processed in the laboratory (DNA extraction from the gFOBT cards, DNA quantification using the NanoDrop, PCR amplification of the 16SrRNA gene, visualisation and quantification of the PCR products, pooling of the PCR products and purification of the pool ready for NGS.

Bioinformatic skills workshop: This was an interactive workshop held at the University and lead by Dr Wood. A number of presentations outlined what happens during sequencing, how data is analysed using QIIME2 and how to create mapping files. The group then worked through commands in real-time, learning how to take data from the NGS sequencer, process it and analyse it

(alpha diversity, beta diversity, taxonomy and modelling within QIIME2 were covered).

Tour of the sequencing facility: The group were given a tour of the Sequencing Facility at the University of Leeds. The advantages and disadvantages of different types of sequencing machine were described and there was an explanation of how samples are checked for quality control and what happens to them during sequencing.

Microbiome knowledge: There were many presentations relating to microbiome research. These covered: how microbiome studies are conducted and potential pitfalls/flaws to consider when reading the microbiome literature or designing microbiome studies; background to the laboratory techniques the network will use to study the microbiome; alternative methods to study the microbiome (including qPCR, IHC, metagenomics and metabolomics); how the microbiome is portrayed in the popular press (this information could be used to educate patients or policy-makers); large-scale microbiome projects to be aware of including the EMP, the Human Microbiome Project and the American and British Gut projects.

Brainstorming session to identify barriers to microbiome research being conducted by the network: After reading two articles which described the paucity of microbiome research being conducted in India and Latin America and why this might be, the group brainstormed potential barriers and solutions to microbiome research being conducted in their own countries. The barriers identified were: widespread and unmonitored antibiotic use in many of the countries; diversity of populations (including ethnicity and diet) within some of the countries; a lack of resources (as policy makers do not prioritise microbiome research due to a lack of awareness); a lack of knowledge about the microbiome (among researchers, the Ministry of Health and policy makers, the public). The GCRFN network was felt to be key to overcoming some of these barriers by providing access to a greater number and diversity of samples, funding (and the potential to apply for further funding), prestige, scientific support and resources. There is also the possibility to develop collaborations *within* each member country (as well as between countries) using the knowledge and skills gained.

Study document review: In this session, the study documents and protocol were reviewed and optimised following feedback from the group about what would be feasible and desirable. The study documents have subsequently been updated.

Conclusions and plans for the future: During this session Dr Vaccaro and Prof Quirke presented ideas for future research which the network could conduct. These ideas were developed by the other members of the network and a plan agreed to take them forward (see below).

Research taking place at the University of Leeds: Prof Quirke, Dr West, Dr Hutchins and Dr Brockmoeller gave presentations outlining CRC research currently taking place at the University of Leeds. A presentation on Lynch screening within the Yorkshire region was also requested.

Networking: There was ample opportunity for networking during the workshop, with sessions being interactive. There was a welcome dinner held at the University on the Monday evening, a trip on Wednesday afternoon to visit historic Yorkshire cities and landmarks and a farewell dinner at Prof Quirke's house on Friday evening. Delegates had Tuesday and Thursday evenings free to explore Leeds.

Material available to the network

Presentations from the workshop, updated study documents, laboratory protocols, bioinformatics commands and photos have been uploaded to a shared dropbox account and are available to all members of the network.

Feedback questionnaire responses

What were the strengths of the workshop?

- *Adaptation to different levels of knowledge and interest of the attendees.*
- *Effort to make application of knowledge feasible locally.*
- *Opportunity to interact with colleagues from different parts of the world.*
- *Planning, duration, topics, practical sessions and social functions.*
- *Well organised, sequential workshop.*

- *Practical demonstrations and the opportunity to talk with each other.*
- *Organisation, getting to know each other, planning future collaboration.*
- *Organisation, content (all the sessions were very educational and informative), the possibility of inter-relating despite differences.*
- *All the sessions.*
- *Knowledge about laboratory skills, sample collection, processing etc.*

What were the weaknesses of the workshop? What could be done to improve the workshop?

- *Nothing.*
- *Provide a brief CV of participants beforehand to facilitate interactions.*
- *If it were held once the ethical approvals were in place, we could have worked on the real samples.*
- *A hands-on workshop could be done in future.*
- *Idiomatic (language) barriers were a limiting thing.*

What have you learnt from the workshop?

- *Current and future state of CRC and the microbiome.*
- *Basics of microbiome research and appraisal of current microbiome research worldwide.*
- *Importance of microbiome studies.*
- *Networking.*
- *Future scope of microbiome research.*
- *NGS and analysis of data.*
- *Different realities, common problems, similar solutions.*
- *Basis of bioinformatics analysis with QIIME2, knowledge about protocols that we don't use in my lab, lots of knowledge about CRC (course, treatment, techniques etc).*
- *A lot: from collecting samples to qPCR etc.*
- *qPCR, 16SrRNA, metabolomics and metagenomics etc.*

What will you take back from the workshop to your home institution?

- *Many ideas to improve the implementation of microbiome research.*
- *Ideas and skills with respect to microbiome research.*
- *A lot of knowledge, information and memories.*
- *Information about the microbiome.*

- *The importance of teamwork.*
- *Many ideas to develop new lines of microbiome research.*
- *Knowledge on the microbiome in CRC and healthy people. Proposals for conducting research in the network.*
- *Information about sample collection.*

Table 51. Feedback from the GCRFNG Network Workshop.

Sessions	Excellent	Very useful	Useful
Laboratory skills demonstrations.	6	2	
Sample collection, processing, packaging and patient consent brainstorming session.	7	1	
NGS bioinformatics workshop.	6	1	1
Presentations on research taking place at the University of Leeds	7	1	
Tour of the NGS sequencing facility.	6	1	1
Presentation on alternative microbiome techniques.	5	2	1
Brainstorming session: barriers to conducting microbiome research and solutions.	6	2	
Microbiome in the popular press.	5	3	

Overall	**Definitely agree**	**Mostly agree**	**Neither agree nor disagree**
The workshop was well organised.	8		
The workshop was useful.	8		
The workshop provided a good opportunity to network.	8		
I am clear what the aims of the workshop were.	8		
I learnt a lot during the workshop.	7	1	
The workshop will help me undertake microbiome research in my home institution.	8		
I am clear about what I need to do to collect and process samples.	8		

List of References

1. Willyard, C. Could baby's first bacteria take root before birth? *Nature.* 2018, **553**(7688), pp.264-266.
2. Stinson, L.F., Payne, M.S. and Keelan, J.A. A Critical Review of the Bacterial Baptism Hypothesis and the Impact of Cesarean Delivery on the Infant Microbiome. *Frontiers in medicine.* 2018, **5**, pp.135-135.
3. Stewart, C.J., Ajami, N.J., O'Brien, J.L., Hutchinson, D.S., Smith, D.P., Wong, M.C., Ross, M.C., Lloyd, R.E., Doddapaneni, H., Metcalf, G.A., Muzny, D., Gibbs, R.A., Vatanen, T., Huttenhower, C., Xavier, R.J., Rewers, M., Hagopian, W., Toppari, J., Ziegler, A.G., She, J.X., Akolkar, B., Lernmark, A., Hyoty, H., Vehik, K., Krischer, J.P. and Petrosino, J.F. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature.* 2018, **562**(7728), pp.583-588.
4. Shao, Y., Forster, S.C., Tsaliki, E., Vervier, K., Strang, A., Simpson, N., Kumar, N., Stares, M.D., Rodger, A., Brocklehurst, P., Field, N. and Lawley, T.D. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature.* 2019, p.[no pagination].
5. Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., Heath, A.C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J.G., Lozupone, C.A., Lauber, C., Clemente, J.C., Knights, D., Knight, R. and Gordon, J.I. Human gut microbiome viewed across age and geography. *Nature.* 2012, **486**, p.222.
6. Moossavi, S., Sepehri, S., Robertson, B., Bode, L., Goruk, S., Field, C.J., Lix, L.M., de Souza, R.J., Becker, A.B., Mandhane, P.J., Turvey, S.E., Subbarao, P., Moraes, T.J., Lefebvre, D.L., Sears, M.R., Khafipour, E. and Azad, M.B. Composition and Variation of the Human Milk Microbiota Are Influenced by Maternal and Early-Life Factors. *Cell host & microbe.* 2019, **25**(2), pp.324-335.e324.
7. Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., Khan, Muhammad T., Zhang, J., Li, J., Xiao, L., Al-Aama, J., Zhang, D., Lee, Ying S., Kotowska, D., Colding, C., Tremaroli, V., Yin, Y., Bergman, S., Xu, X., Madsen, L., Kristiansen, K., Dahlgren, J. and Wang, J. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell host & microbe.* 2015, **17**(5), pp.690-703.
8. Bezirtzoglou, E., Tsiotsias, A. and Welling, G.W. Microbiota profile in feces of breast- and formula-fed newborns by using fluorescence in situ hybridization (FISH). *Anaerobe.* 2011, **17**(6), pp.478-482.
9. Dash, N.R., Khoder, G., Nada, A.M. and Al Bataineh, M.T. Exploring the impact of Helicobacter pylori on gut microbiome composition. *PloS one.* 2019, **14**(6), pp.e0218274-e0218274.
10. Franzosa, E.A., Morgan, X.C., Segata, N., Waldron, L., Reyes, J., Earl, A.M., Giannoukos, G., Boylan, M.R., Ciulla, D., Gevers, D., Izard, J., Garrett, W.S., Chan, A.T. and Huttenhower, C. Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci U S A.* 2014, **111**(22), pp.E2329-2338.

11. Schmidt, T.S., Hayward, M.R., Coelho, L.P., Li, S.S., Costea, P.I., Voigt, A.Y., Wirbel, J., Maistrenko, O.M., Alves, R.J., Bergsten, E., de Beaufort, C., Sobhani, I., Heintz-Buschart, A., Sunagawa, S., Zeller, G., Wilmes, P. and Bork, P. Extensive transmission of microbes along the gastrointestinal tract. *eLife.* 2019, **8**, p.[no pagination].
12. Sender, R., Fuchs, S. and Milo, R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell.* 2016, **164**(3), pp.337-340.
13. Sender, R., Fuchs, S. and Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS biology.* 2016, **14**(8), p.e1002533.
14. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S.D. and Wang, J. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010, **464**(7285), pp.59-65.
15. Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H.B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E.G., Wang, J., Guarner, F., Pedersen, O., de Vos, W.M., Brunak, S., Doré, J., Meta, H.I.T.C., Antolín, M., Artiguenave, F., Blottiere, H.M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K.U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Kristiansen, K., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Mérieux, A., Melo Minardi, R., M'Rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S.D. and Bork, P. Enterotypes of the human gut microbiome. *Nature.* 2011, **473**, p.174.
16. Gorvitovskaia, A., Holmes, S.P. and Huse, S.M. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. *Microbiome.* 2016, **4**, p.15.
17. Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., Sharma, V.K., Srivastava, T.P., Taylor, T.D., Noguchi, H., Mori, H., Ogura, Y., Ehrlich, D.S., Itoh, K., Takagi, T., Sakaki, Y., Hayashi, T. and Hattori, M. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* 2007, **14**(4), pp.169-181.
18. Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P.,

Egholm, M., Henrissat, B., Heath, A.C., Knight, R. and Gordon, J.I. A core gut microbiome in obese and lean twins. *Nature.* 2008, **457**, p.480.

19. Claussen, J.C., Skiecevičienė, J., Wang, J., Rausch, P., Karlsen, T.H., Lieb, W., Baines, J.F., Franke, A. and Hütt, M.-T. Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. *PLoS computational biology.* 2017, **13**(6), p.e1005361.

20. The Human Microbiome Project, C., Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S., Giglio, M.G., Hallsworth-Pepin, K., Lobos, E.A., Madupu, R., Magrini, V., Martin, J.C., Mitreva, M., Muzny, D.M., Sodergren, E.J., Versalovic, J., Wollam, A.M., Worley, K.C., Wortman, J.R., Young, S.K., Zeng, Q., Aagaard, K.M., Abolude, O.O., Allen-Vercoe, E., Alm, E.J., Alvarado, L., Andersen, G.L., Anderson, S., Appelbaum, E., Arachchi, H.M., Armitage, G., Arze, C.A., Ayvaz, T., Baker, C.C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M.J., Bloom, T., Bonazzi, V., Paul Brooks, J., Buck, G.A., Buhay, C.J., Busam, D.A., Campbell, J.L., Canon, S.R., Cantarel, B.L., Chain, P.S.G., Chen, I.M.A., Chen, L., Chhibba, S., Chu, K., Ciulla, D.M., Clemente, J.C., Clifton, S.W., Conlan, S., Crabtree, J., Cutting, M.A., Davidovics, N.J., Davis, C.C., DeSantis, T.Z., Deal, C., Delehaunty, K.D., Dewhirst, F.E., Deych, E., Ding, Y., Dooling, D.J., Dugan, S.P., Michael Dunne, W., Scott Durkin, A., Edgar, R.C., Erlich, R.L., Farmer, C.N., Farrell, R.M., Faust, K., Feldgarden, M., Felix, V.M., Fisher, S., Fodor, A.A., Forney, L.J., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D.C., Fronick, C.C., Fulton, L.L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M.Y., Goldberg, J.M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Kinder Haake, S., Haas, B.J., Hamilton, H.A., Harris, E.L., Hepburn, T.A., Herter, B., Hoffmann, D.E., Holder, M.E., Howarth, C., Huang, K.H., Huse, S.M., Izard, J., Jansson, J.K., Jiang, H., Jordan, C., Joshi, V., Katancik, J.A., Keitel, W.A., Kelley, S.T., Kells, C., King, N.B., Knights, D., Kong, H.H., Koren, O., Koren, S., Kota, K.C., Kovar, C.L., Kyrpides, N.C., La Rosa, P.S., Lee, S.L., Lemon, K.P., Lennon, N., Lewis, C.M., Lewis, L., Ley, R.E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.-C., Lozupone, C.A., Dwayne Lunsford, R., Madden, T., Mahurkar, A.A., Mannon, P.J., Mardis, E.R., Markowitz, V.M., Mavromatis, K., McCorrison, J.M., McDonald, D., McEwen, J., McGuire, A.L., McInnes, P., Mehta, T., Mihindukulasuriya, K.A., Miller, J.R., Minx, P.J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S.M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K.S., Pop, M., Priest, M.E., Proctor, L.M., Qin, X., Raes, J., Ravel, J., Reid, J.G., Rho, M., Rhodes, R., Riehle, K.P., Rivera, M.C., Rodriguez-Mueller, B., Rogers, Y.-H., Ross, M.C., Russ, C., Sanka, R.K., Sankar, P., Fah Sathirapongsasuti, J., Schloss, J.A., Schloss, P.D., Schmidt, T.M., Scholz, M., Schriml, L., Schubert, A.M., Segata, N., Segre, J.A., Shannon, W.D., Sharp, R.R., Sharpton, T.J., Shenoy, N., Sheth, N.U., Simone, G.A., Singh, I., Smillie, C.S., Sobel, J.D., Sommer, D.D., Spicer, P., Sutton, G.G., Sykes, S.M., Tabbaa, D.G., Thiagarajan, M., Tomlinson, C.M., Torralba, M., Treangen, T.J., Truty, R.M., Vishnivetskaya, T.A., Walker, J., Wang, L., Wang, Z., Ward, D.V., Warren, W., Watson, M.A.,

Wellington, C., Wetterstrand, K.A., White, J.R., Wilczek-Boney, K., Wu, Y., Wylie, K.M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B.P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J.D., Birren, B.W., Gibbs, R.A., Highlander, S.K., Methé, B.A., Nelson, K.E., Petrosino, J.F., Weinstock, G.M., Wilson, R.K. and White, O. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012, **486**, p.207.

21. Koskinen, K., Pausan, M.R., Perras, A.K., Beck, M., Bang, C., Mora, M., Schilhabel, A., Schmitz, R. and Moissl-Eichinger, C. First Insights into the Diverse Human Archaeome: Specific Detection of Archaea in the Gastrointestinal Tract, Lung, and Nose and on Skin. *mBio.* 2017, **8**(6), pp.e00824-00817.
22. Nash, A.K., Auchtung, T.A., Wong, M.C., Smith, D.P., Gesell, J.R., Ross, M.C., Stewart, C.J., Metcalf, G.A., Muzny, D.M., Gibbs, R.A., Ajami, N.J. and Petrosino, J.F. The gut mycobiome of the Human Microbiome Project healthy cohort. *Microbiome.* 2017, **5**(1), p.153.
23. Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D. and Bushman, F.D. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* 2011, **21**(10), pp.1616-1625.
24. Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M., Amundson, R., Thomas, B.C. and Banfield, J.F. A new view of the tree of life. *Nature microbiology.* 2016, **1**, p.16048.
25. Dridi, B., Henry, M., El Khéchine, A., Raoult, D. and Drancourt, M. High prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae detected in the human gut using an improved DNA detection protocol. *PloS one.* 2009, **4**(9), pp.e7063-e7063.
26. Hoffmann, C., Dollive, S., Grunberg, S., Chen, J., Li, H., Wu, G.D., Lewis, J.D. and Bushman, F.D. Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents. *PloS one.* 2013, **8**(6), p.e66019.
27. Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F. and Gordon, J.I. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature.* 2010, **466**(7304), pp.334-338.
28. Krogsgaard, L.R., Andersen, L.O.B., Johannesen, T.B., Engsbro, A.L., Stensvold, C.R., Nielsen, H.V. and Bytzer, P. Characteristics of the bacterial microbiome in association with common intestinal parasites in irritable bowel syndrome. *Clinical and Translational Gastroenterology.* 2018, **9**(6), p.161.
29. Audebert, C., Even, G., Cian, A., The Blastocystis Investigation, G., Safadi, D.E., Certad, G., Delhaes, L., Pereira, B., Nourrisson, C., Poirier, P., Wawrzyniak, I., Delbac, F., Morelle, C., Bastien, P., Lachaud, L., Bellanger, A.-P., Botterel, F., Candolfi, E., Desoubeaux, G., Morio, F., Pomares, C., Rabodonirina, M., Loywick, A., Merlin, S., Viscogliosi, E. and Chabé, M. Colonization with the enteric protozoa Blastocystis is associated with increased diversity of human gut bacterial microbiota. *Scientific reports.* 2016, **6**, p.25255.
30. Lee, S.C., Tang, M.S., Lim, Y.A.L., Choy, S.H., Kurtz, Z.D., Cox, L.M., Gundra, U.M., Cho, I., Bonneau, R., Blaser, M.J., Chua, K.H. and Loke,

P.n. Helminth Colonization Is Associated with Increased Diversity of the Gut Microbiota. *PLOS Neglected Tropical Diseases.* 2014, **8**(5), p.e2880.

31. Morton, E.R., Lynch, J., Froment, A., Lafosse, S., Heyer, E., Przeworski, M., Blekhman, R. and Ségurel, L. Variation in Rural African Gut Microbiota Is Strongly Correlated with Colonization by Entamoeba and Subsistence. *PLoS Genetics.* 2015, **11**(11), p.e1005658.
32. Burgess, S.L. and Petri, W.A., Jr. The Intestinal Bacterial Microbiome and E. histolytica Infection. *Current tropical medicine reports.* 2016, **3**, pp.71-74.
33. Bär, A.-K., Phukan, N., Pinheiro, J. and Simoes-Barbosa, A. The Interplay of Host Microbiota and Parasitic Protozoans at Mucosal Interfaces: Implications for the Outcomes of Infections and Diseases. *PLOS Neglected Tropical Diseases.* 2015, **9**(12), p.e0004176.
34. Padilla-Vaca, F., Ankri, S., Bracha, R., Koole, L.A. and Mirelman, D. Down Regulation of Entamoeba histolytica Virulence by Monoxenic Cultivation with Escherichia coli O55 Is Related to a Decrease in Expression of the Light (35-Kilodalton) Subunit of the Gal/GalNAc Lectin. *Infection and immunity.* 1999, **67**(5), p.2096.
35. Mendoza-Macias, C.L., Barrios-Ceballos, M.P., de la Pena, L.P., Rangel-Serrano, A., Anaya-Velazquez, F., Mirelman, D. and Padilla-Vaca, F. Entamoeba histolytica: effect on virulence, growth and gene expression in response to monoxenic culture with Escherichia coli 055. *Exp Parasitol.* 2009, **121**(2), pp.167-174.
36. Verma, A.K., Verma, R., Ahuja, V. and Paul, J. Real-time analysis of gut flora in Entamoeba histolytica infected patients of Northern India. *BMC microbiology.* 2012, **12**(1), p.183.
37. Cooper, P., Walker, A.W., Reyes, J., Chico, M., Salter, S.J., Vaca, M. and Parkhill, J. Patent Human Infections with the Whipworm, Trichuris trichiura, Are Not Associated with Alterations in the Faecal Microbiota. *PloS one.* 2013, **8**(10), p.e76573.
38. O'Hara, A.M. and Shanahan, F. The gut flora as a forgotten organ. *EMBO reports.* 2006, **7**(7), pp.688-693.
39. van de Guchte, M., Blottière, H.M. and Doré, J. Humans as holobionts: implications for prevention and therapy. *Microbiome.* 2018, **6**(1), pp.81-81.
40. Rowland, I., Gibson, G., Heinken, A., Scott, K., Swann, J., Thiele, I. and Tuohy, K. Gut microbiota functions: metabolism of nutrients and other food components. *European journal of nutrition.* 2018, **57**(1), pp.1-24.
41. Bäckhed, F., Ding, H., Wang, T., Hooper, L.V., Koh, G.Y., Nagy, A., Semenkovich, C.F. and Gordon, J.I. The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences of the United States of America.* 2004, **101**(44), pp.15718-15723.
42. Hague, A., Elder, D.J., Hicks, D.J. and Paraskeva, C. Apoptosis in colorectal tumour cells: induction by the short chain fatty acids butyrate, propionate and acetate and by the bile salt deoxycholate. *Int J Cancer.* 1995, **60**(3), pp.400-406.
43. Smith, P.M., Howitt, M.R., Panikov, N., Michaud, M., Gallini, C.A., Bohlooly, Y.M., Glickman, J.N. and Garrett, W.S. The microbial

metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. *Science.* 2013, **341**(6145), pp.569-573.

44. Kim, M., Friesen, L., Park, J., Kim, H.M. and Kim, C.H. Microbial metabolites, short-chain fatty acids, restrain tissue bacterial load, chronic inflammation, and associated cancer in the colon of mice. *Eur J Immunol.* 2018, **48**(7), pp.1235-1247.
45. Corrêa-Oliveira, R., Fachi, J.L., Vieira, A., Sato, F.T. and Vinolo, M.A.R. Regulation of immune cell function by short-chain fatty acids. *Clin Transl Immunology.* 2016, **5**(4), pp.e73-e73.
46. Chambers, E.S., Preston, T., Frost, G. and Morrison, D.J. Role of Gut Microbiota-Generated Short-Chain Fatty Acids in Metabolic and Cardiovascular Health. *Current Nutrition Reports.* 2018, **7**(4), pp.198-206.
47. Erny, D., Hrabe de Angelis, A.L., Jaitin, D., Wieghofer, P., Staszewski, O., David, E., Keren-Shaul, H., Mahlakoiv, T., Jakobshagen, K., Buch, T., Schwierzeck, V., Utermohlen, O., Chun, E., Garrett, W.S., McCoy, K.D., Diefenbach, A., Staeheli, P., Stecher, B., Amit, I. and Prinz, M. Host microbiota constantly control maturation and function of microglia in the CNS. *Nat Neurosci.* 2015, **18**(7), pp.965-977.
48. Ridlon, J.M., Kang, D.J., Hylemon, P.B. and Bajaj, J.S. Bile acids and the gut microbiome. *Curr Opin Gastroenterol.* 2014, **30**(3), pp.332-338.
49. Wilson, I.D. and Nicholson, J.K. Gut microbiome interactions with drug metabolism, efficacy, and toxicity. *Translational research : the journal of laboratory and clinical medicine.* 2017, **179**, pp.204-222.
50. Koppel, N., Maini Rekdal, V. and Balskus, E.P. Chemical transformation of xenobiotics by the human gut microbiota. *Science.* 2017, **356**(6344), p.eaag2770.
51. Alexander, J.L., Wilson, I.D., Teare, J., Marchesi, J.R., Nicholson, J.K. and Kinross, J.M. Gut microbiota modulation of chemotherapy efficacy and toxicity. *Nature Reviews Gastroenterology Hepatology.* 2017, **14**, p.356.
52. Sicard, J.F., Le Bihan, G., Vogeleer, P., Jacques, M. and Harel, J. Interactions of Intestinal Bacteria with Components of the Intestinal Mucus. *Front Cell Infect Microbiol.* 2017, **7**, p.387.
53. Sekirov, I., Russell, S.L., Antunes, L.C.M. and Finlay, B.B. Gut Microbiota in Health and Disease. *Physiological reviews.* 2010, **90**(3), pp.859-904.
54. Belkaid, Y. and Hand, Timothy W. Role of the Microbiota in Immunity and Inflammation. *Cell.* 2014, **157**(1), pp.121-141.
55. Schirmer, M., Smeekens, S.P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E.A., Ter Horst, R., Jansen, T., Jacobs, L., Bonder, M.J., Kurilshikov, A., Fu, J., Joosten, L.A.B., Zhernakova, A., Huttenhower, C., Wijmenga, C., Netea, M.G. and Xavier, R.J. Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell.* 2016, **167**(4), pp.1125-1136.e1128.
56. Yano, J.M., Yu, K., Donaldson, G.P., Shastri, G.G., Ann, P., Ma, L., Nagler, C.R., Ismagilov, R.F., Mazmanian, S.K. and Hsiao, E.Y. Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell.* 2015, **161**(2), pp.264-276.

57. Kamada, N., Chen, G.Y., Inohara, N. and Núñez, G. Control of pathogens and pathobionts by the gut microbiota. *Nature immunology.* 2013, **14**(7), pp.685-690.
58. Donaldson, G.P., Lee, S.M. and Mazmanian, S.K. Gut biogeography of the bacterial microbiota. *Nature reviews. Microbiology.* 2016, **14**(1), pp.20-32.
59. Saffarian, A., Mulet, C., Regnault, B., Amiot, A., Tran-Van-Nhieu, J., Ravel, J., Sobhani, I., Sansonetti, P.J. and Pedron, T. Crypt- and Mucosa-Associated Core Microbiotas in Humans and Their Alteration in Colon Cancer Patients. *mBio.* 2019, **10**(4), p.[no pagination].
60. Haggar, F.A. and Boushey, R.P. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clin Colon Rectal Surg.* 2009, **22**(4), pp.191-197.
61. Dong, Y., Zhou, J., Zhu, Y., Luo, L., He, T., Hu, H., Liu, H., Zhang, Y., Luo, D., Xu, S., Xu, L., Liu, J., Zhang, J. and Teng, Z. Abdominal obesity and colorectal cancer risk: systematic review and meta-analysis of prospective studies. *Biosci Rep.* 2017, **37**(6), p.[no pagination].
62. Choi, Y.J., Myung, S.K. and Lee, J.H. Light Alcohol Drinking and Risk of Cancer: A Meta-Analysis of Cohort Studies. *Cancer Res Treat.* 2018, **50**(2), pp.474-487.
63. Wang, J.L., Chang, C.H., Lin, J.W., Wu, L.C., Chuang, L.M. and Lai, M.S. Infection, antibiotic therapy and risk of colorectal cancer: A nationwide nested case-control study in patients with Type 2 diabetes mellitus. *International journal of cancer.* 2014, **135**(4), pp.956-967.
64. Boursi, B., Haynes, K., Mamtani, R. and Yang, Y.X. Impact of antibiotic exposure on the risk of colorectal cancer. *Pharmacoepidemiology and drug safety.* 2015, **24**(5), pp.534-542.
65. Dik, V.K., van Oijen, M.G.H., Smeets, H.M. and Siersema, P.D. Frequent Use of Antibiotics Is Associated with Colorectal Cancer Risk: Results of a Nested Case-Control Study. *Digestive diseases and sciences.* 2016, **61**(1), pp.255-264.
66. Cao, Y., Wu, K., Mehta, R., Drew, D.A., Song, M., Lochhead, P., Nguyen, L.H., Izard, J., Fuchs, C.S., Garrett, W.S., Huttenhower, C., Ogino, S., Giovannucci, E.L. and Chan, A.T. Long-term use of antibiotics and risk of colorectal adenoma. *Gut.* 2018, **67**(4), pp.672-678.
67. Kaur, K., Saxena, A., Debnath, I., O'Brien, J.L., Ajami, N.J., Auchtung, T.A., Petrosino, J.F., Sougiannis, A.J., Depaep, S., Chumanevich, A., Gummadidala, P.M., Omebeyinje, M.H., Banerjee, S., Chatzistamou, I., Chakraborty, P., Fayad, R., Berger, F.G., Carson, J.A. and Chanda, A. Antibiotic-mediated bacteriome depletion in ApcMin/+ mice is associated with reduction in mucus-producing goblet cells and increased colorectal cancer progression. *Cancer Medicine.* 2018, **7**(5), pp.2003-2012.
68. Wu, S.C., Chen, W.T.L., Muo, C.H., Ke, T.W., Fang, C.W. and Sung, F.C. Association between appendectomy and subsequent colorectal cancer development: An asian population study. *PloS one.* 2015, **10**(2), p.e0118411.
69. Momen-Heravi, F., Babic, A., Tworoger, S.S., Zhang, L., Wu, K., Smith-Warner, S.A., Ogino, S., Chan, A.T., Meyerhardt, J., Giovannucci, E.,

Fuchs, C., Cho, E., Michaud, D.S., Stampfer, M.J., Yu, Y.-H., Kim, D. and Zhang, X. Periodontal disease, tooth loss and colorectal cancer risk: Results from the Nurses' Health Study. *International journal of cancer.* 2017, **140**(3), pp.646-652.

70. Klimosch, S.N., Forsti, A., Eckert, J., Knezevic, J., Bevier, M., Von Schonfels, W.V., Heits, N., Walter, J., Hinz, S., Lascorz, J., Hampe, J., Hartl, D., Frick, J.S., Hemminki, K., Schafmayer, C. and Weber, A.N.R. Functional TLR5 genetic variants affect human colorectal cancer survival. *Cancer research.* 2013, **73**(24), pp.7232-7242.
71. Kopp, T.I., Vogel, U., Tjonneland, A. and Andersen, V. Meat and fiber intake and interaction with pattern recognition receptors (TLR1, TLR2, TLR4, and TLR10) in relation to colorectal cancer in a Danish prospective, case-cohort study. *Am J Clin Nutr.* 2018, **107**(3), pp.465-479.
72. Kwong, T.N.Y., Wang, X., Nakatsu, G., Chow, T.C., Tipoe, T., Dai, R.Z.W., Tsoi, K.K.K., Wong, M.C.S., Tse, G., Chan, M.T.V., Chan, F.K.L., Ng, S.C., Wu, J.C.Y., Wu, W.K.K., Yu, J., Sung, J.J.Y. and Wong, S.H. Association Between Bacteremia From Specific Microbes and Subsequent Diagnosis of Colorectal Cancer. *Gastroenterology.* 2018, **155**(2), p.383.
73. Yoshino, Y., Kitazawa, T., Ikeda, M., Tatsuno, K., Yanagimoto, S., Okugawa, S., Ota, Y. and Yotsuyanagi, H. Clinical features of Bacteroides bacteremia and their association with colorectal carcinoma. *Infection.* 2012, **40**(1), pp.63-67.
74. Zheng, X., Wu, K., Song, M., Ogino, S., Fuchs, C.S., Chan, A.T., Giovannucci, E.L., Cao, Y. and Zhang, X. Yogurt consumption and risk of conventional and serrated precursors of colorectal cancer. *Gut.* 2019, pp.gutjnl-2019-318374.
75. Aune, D., Chan, D.S.M., Lau, R., Vieira, R., Greenwood, D.C., Kampman, E. and Norat, T. Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose-response meta-analysis of prospective studies. *BMJ.* 2011, **343**, p.d6617.
76. *Cancer Research UK Bowel Cancer Statistics.* [Online]. [Accessed 11.2.19]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer
77. Shanahan, F. and O'Toole, P.W. Host-microbe interactions and spatial variation of cancer in the gut. *Nature Reviews Cancer.* 2014, **14**(8), pp.511-512.
78. Zoetendal, E.G., Raes, J., van den Bogert, B., Arumugam, M., Booijink, C.C.G.M., Troost, F.J., Bork, P., Wels, M., de Vos, W.M. and Kleerebezem, M. The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. *The ISME journal.* 2012, **6**(7), pp.1415-1426.
79. Shen, X.J., Rawls, J.F., Randall, T., Burcal, L., Mpande, C.N., Jenkins, N., Jovov, B., Abdo, Z., sandler, R.S. and Keku, T.O. Molecular characterization of mucosal adherent bacteria and associations with colorectal adenomas. *Gut microbes.* 2010, **1**(3), pp.1-10.
80. Sanapareddy, N., Legge, R.M., Jovov, B., McCoy, A., Burcal, L., Araujo-Perez, F., Randall, T.A., Galanko, J., Benson, A., Sandler, R.S., Rawls, J.F., Abdo, Z., Fodor, A.A. and Keku, T.O. Increased rectal

microbial richness is associated with the presence of colorectal adenomas in humans. *ISME Journal.* 2012, **6**(10), pp.1858-1868.

81. McCoy, A.N., Araujo-Perez, F., Azcarate-Peril, A., Yeh, J.J., Sandler, R.S. and Keku, T.O. Fusobacterium Is Associated with Colorectal Adenomas. *PloS one.* 2013, **8**(1), p.e53653.
82. Brim, H., Yooseph, S., Lee, E., Sherif, Z., Abbas, M., Laiyemo, A.O., Varma, S., Torralba, M., Dowd, S.E., Nelson, K.E., Pathmasiri, W., Sumner, S., De Vos, W., Liang, Q., Yu, J., Zoetendal, E. and Ashktorab, H. A microbiomic analysis in African Americans with colonic lesions reveals streptococcus sp.VT162 as a marker of neoplastic transformation. *Genes.* 2017, **8**(11), p.314.
83. Nugent, J.L., McCoy, A.N., Addamo, C.J., Jia, W., Sandler, R.S. and Keku, T.O. Altered tissue metabolites correlate with microbial dysbiosis in colorectal adenomas. *Journal of proteome research.* 2014, **13**(4), pp.1921-1929.
84. Lu, Y., Chen, J., Zheng, J., Hu, G., Wang, J., Huang, C., Lou, L., Wang, X. and Zeng, Y. Mucosal adherent bacterial dysbiosis in patients with colorectal adenomas. *Scientific reports.* 2016, **6**, p.26337.
85. Peters, B.A., Dominianni, C., Shapiro, J.A., Church, T.R., Wu, J., Miller, G., Yuen, E., Freiman, H., Lustbader, I., Salik, J., Friedlander, C., Hayes, R.B. and Ahn, J. The gut microbiota in conventional and serrated precursors of colorectal cancer. *Microbiome.* 2016, **4**, p.69.
86. Chen, W., Liu, F., Ling, Z., Tong, X. and Xiang, C. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PloS one.* 2012, **7**(6), p.e39743.
87. Zhang, H., Chang, Y., Zheng, Q., Zhang, R., Hu, C. and Jia, W. Altered intestinal microbiota associated with colorectal cancer. *Front Med.* 2019, **13**(4), pp.461-470.
88. Saito, K., Koido, S., Odamaki, T., Kajihara, M., Kato, K., Horiuchi, S., Adachi, S., Arakawa, H., Yoshida, S., Akasu, T., Ito, Z., Uchiyama, K., Saruta, M., Xiao, J.Z., Sato, N. and Ohkusa, T. Metagenomic analyses of the gut microbiota associated with colorectal adenoma. *PloS one.* 2019, **14**(2), p.e0212406.
89. Xu, K. and Jiang, B. Analysis of Mucosa-Associated Microbiota in Colorectal Cancer. *Med Sci Monit.* 2017, **23**, pp.4422-4430.
90. Thomas, A.M., Jesus, E.C., Lopes, A., Aguiar, S., Jr., Begnami, M.D., Rocha, R.M., Carpinetti, P.A., Camargo, A.A., Hoffmann, C., Freitas, H.C., Silva, I.T., Nunes, D.N., Setubal, J.C. and Dias-Neto, E. Tissue-Associated Bacterial Alterations in Rectal Carcinoma Patients Revealed by 16S rRNA Community Profiling. *Front Cell Infect Microbiol.* 2016, **6**, p.179.
91. Yang, Y., Misra, B.B., Liang, L., Bi, D., Weng, W., Wu, W., Cai, S., Qin, H., Goel, A., Li, X. and Ma, Y. Integrated microbiome and metabolome analysis reveals a novel interplay between commensal bacteria and metabolites in colorectal cancer. *Theranostics.* 2019, **9**(14), pp.4101-4114.
92. Flemer, B., Lynch, D.B., Brown, J.M., Jeffery, I.B., Ryan, F.J., Claesson, M.J., O'Riordain, M., Shanahan, F. and O'Toole, P.W. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut.* 2017, **66**(4), pp.633-643.

93. Flemer, B., Herlihy, M., O'Riordain, M., Shanahan, F. and O'Toole, P.W. Tumour-associated and non-tumour-associated microbiota: Addendum. *Gut microbes.* 2018, **9**(4), pp.369-373.
94. Burns, M.B., Lynch, J., Starr, T.K., Knights, D. and Blekhman, R. Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. *Genome medicine.* 2015, **7**(1), p.55.
95. Geng, J., Fan, H., Tang, X., Zhai, H. and Zhang, Z. Diversified pattern of the human colorectal cancer microbiome. *Gut Pathog.* 2013, **5**(1), p.2.
96. Sobhani, I., Tap, J., Roudot-Thoraval, F., Roperch, J.P., Letulle, S., Langella, P., Gerard, C., van Nhieu, J.T. and Furet, J.P. Microbial dysbiosis in colorectal cancer (CRC) patients. *PloS one.* 2011, **6**(1), p.e16393.
97. Allali, I., Delgado, S., Marron, P.I., Astudillo, A., Yeh, J.J., Ghazal, H., Amzazi, S., Keku, T. and Azcarate-Peril, M.A. 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), pp.161-172.
98. Flemer, B., Warren, R.D., Barrett, M.P., Cisek, K., Das, A., Jeffery, I.B., Hurley, E., O'Riordain, M., Shanahan, F. and O'Toole, P.W. The oral microbiota in colorectal cancer is distinctive and predictive. *Gut.* 2018, **67**(8), pp.1454-1463.
99. Ai, D., Pan, H., Li, X., Wu, M. and Xia, L.C. Association network analysis identifies enzymatic components of gut microbiota that significantly differ between colorectal cancer patients and healthy controls. *PeerJ.* 2019, **7**, p.e7315.
100. Minot, S.S. and Willis, A.D. Clustering co-abundant genes identifies components of the gut microbiome that are reproducibly associated with colorectal cancer and inflammatory bowel disease. *Microbiome.* 2019, **7**(1), pp.110-110.
101. Keku, T.O., Dulal, S., Deveaux, A., Jovov, B. and Han, X. The gastrointestinal microbiota and colorectal cancer. *American Journal of Physiology - Gastrointestinal and Liver Physiology.* 2015, **308**(5), pp.G351-G363.
102. Lucas, C., Barnich, N. and Nguyen, H.T.T. Microbiota, Inflammation and Colorectal Cancer. *Int J Mol Sci.* 2017, **18**(6), p.[no pagination].
103. Borges-Canha, M., Portela-Cidade, J.P., Dinis-Ribeiro, M., Leite-Moreira, A.F. and Pimentel-Nunes, P. Role of colonic microbiota in colorectal carcinogenesis: a systematic review. *Rev Esp Enferm Dig.* 2015, **107**(11), pp.659-671.
104. Sun, J. and Kato, I. Gut microbiota, inflammation and colorectal cancer. *Genes Dis.* 2016, **3**(2), pp.130-143.
105. Sobhani, I., Amiot, A., Le Baleur, Y., Levy, M., Auriault, M.-L., Van Nhieu, J.T. and Delchier, J.C. Microbial dysbiosis and colon carcinogenesis: could colon cancer be considered a bacteria-related disease? *Therapeutic Advances in Gastroenterology.* 2013, **6**(3), pp.215-229.
106. Balamurugan, R., Rajendiran, E., George, S., Samuel, G.V. and Ramakrishna, B.S. Real-time polymerase chain reaction quantification of specific butyrate-producing bacteria, Desulfovibrio and

Enterococcus faecalis in the feces of patients with colorectal cancer. *J Gastroenterol Hepatol.* 2008, **23**(8 Pt 1), pp.1298-1303.

107. Hale, V.L., Chen, J., Johnson, S., Harrington, S.C., Yab, T.C., Smyrk, T.C., Nelson, H., Boardman, L.A., Druliner, B.R., Levin, T.R., Rex, D.K., Ahnen, D.J., Lance, P., Ahlquist, D.A. and Chia, N. Shifts in the fecal microbiota associated with adenomatous polyps. *Cancer Epidemiology Biomarkers and Prevention.* 2017, **26**(1), pp.85-94.
108. Weir, T.L., Manter, D.K., Sheflin, A.M., Barnett, B.A., Heuberger, A.L. and Ryan, E.P. Stool Microbiome and Metabolome Differences between Colorectal Cancer Patients and Healthy Adults. *PloS one.* 2013, **8**(8), p.e70803.
109. Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., Jia, W., Cai, S. and Zhao, L. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *Isme j.* 2012, **6**(2), pp.320-329.
110. Wu, N., Yang, X., Zhang, R., Li, J., Xiao, X., Hu, Y., Chen, Y., Yang, F., Lu, N., Wang, Z., Luan, C., Liu, Y., Wang, B., Xiang, C., Wang, Y., Zhao, F., Gao, G.F., Wang, S., Li, L., Zhang, H. and Zhu, B. Dysbiosis Signature of Fecal Microbiota in Colorectal Cancer Patients. *Microbial ecology.* 2013, **66**(2), pp.462-470.
111. Weaver, G.A., Krause, J.A., Miller, T.L. and Wolin, M.J. Short chain fatty acid distributions of enema samples from a sigmoidoscopy population: An association of high acetate and low butyrate ratios with adenomatous polyps and colon cancer. *Gut.* 1988, **29**(11), pp.1539-1543.
112. Chen, H.M., Yu, Y.N., Wang, J.L., Lin, Y.W., Kong, X., Yang, C.Q., Yang, L., Liu, Z.J., Yuan, Y.Z., Liu, F., Wu, J.X., Zhong, L., Fang, D.C., Zou, W. and Fang, J.Y. Decreased dietary fiber intake and structural alteration of gut microbiota in patients with advanced colorectal adenoma. *American Journal of Clinical Nutrition.* 2013, **97**(5), pp.1044-1052.
113. Ohigashi, S., Sudo, K., Kobayashi, D., Takahashi, O., Takahashi, T., Asahara, T., Nomoto, K. and Onodera, H. Changes of the intestinal microbiota, short chain fatty acids, and fecal pH in patients with colorectal cancer. *Dig Dis Sci.* 2013, **58**(6), pp.1717-1726.
114. Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Jie, Z., Su, L., Li, X., Li, J., Xiao, L., Huber-Schonauer, U., Niederseer, D., Xu, X., Al-Aama, J.Y., Yang, H., Kristiansen, K., Arumugam, M., Tilg, H., Datz, C. and Wang, J. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nature communications.* 2015, **6**, p.6528.
115. Sinha, R., Ahn, J., Sampson, J.N., Shi, J., Yu, G., Xiong, X., Hayes, R.B. and Goedert, J.J. Fecal Microbiota, Fecal Metabolome, and Colorectal Cancer Interrelations. *PloS one.* 2016, **11**(3), p.e0152126.
116. Wang, X., Wang, J., Rao, B. and Deng, L.I. Gut flora profiling and fecal metabolite composition of colorectal cancer patients and healthy individuals. *Experimental and Therapeutic Medicine.* 2017, **13**(6), pp.2848-2854.
117. Warren, R.L., Freeman, D.J., Pleasance, S., Watson, P., Moore, R.A., Cochrane, K., Allen-Vercoe, E. and Holt, R.A. Co-occurrence of

anaerobic bacteria in colorectal carcinomas. *Microbiome.* 2013, **1**(1), p.16.

118. Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., Goedert, J.J., Hayes, R.B. and Yang, L. Human gut microbiome and risk for colorectal cancer. *J Natl Cancer Inst.* 2013, **105**(24), pp.1907-1911.
119. Kasai, C., Sugimoto, K., Moritani, I., Tanaka, J., Oya, Y., Inoue, H., Tameda, M., Shiraki, K., Ito, M., Takei, Y. and Takase, K. Comparison of human gut microbiota in control subjects and patients with colorectal carcinoma in adenoma: Terminal restriction fragment length polymorphism and next-generation sequencing analyses. *Oncology reports.* 2016, **35**(1), pp.325-333.
120. Nakatsu, G., Li, X., Zhou, H., Sheng, J., Wong, S.H., Wu, W.K., Ng, S.C., Tsoi, H., Dong, Y., Zhang, N., He, Y., Kang, Q., Cao, L., Wang, K., Zhang, J., Liang, Q., Yu, J. and Sung, J.J. Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat Commun.* 2015, **6**, p.8727.
121. Sze, M.A. and Schloss, P.D. Leveraging Existing 16S rRNA Gene Surveys To Identify Reproducible Biomarkers in Individuals with Colorectal Tumors. *mBio.* 2018, **9**(3), pp.e00630-00618.
122. Hall-Stoodley, L., Costerton, J.W. and Stoodley, P. Bacterial biofilms: from the Natural environment to infectious diseases. *Nature Reviews Microbiology.* 2004, **2**(2), pp.95-108.
123. Jefferson, K.K. What drives bacteria to produce a biofilm? *FEMS microbiology letters.* 2004, **236**(2), pp.163-173.
124. Huang, R., Li, M. and Gregory, R.L. Bacterial interactions in dental biofilm. *Virulence.* 2011, **2**(5), pp.435-444.
125. Dejea, C.M., Wick, E.C., Hechenbleikner, E.M., White, J.R., Mark Welch, J.L., Rossetti, B.J., Peterson, S.N., Snesrud, E.C., Borisy, G.G., Lazarev, M., Stein, E., Vadivelu, J., Roslani, A.C., Malik, A.A., Wanyiri, J.W., Goh, K.L., Thevambiga, I., Fu, K., Wan, F., Llosa, N., Housseau, F., Romans, K., Wu, X., McAllister, F.M., Wu, S., Vogelstein, B., Kinzler, K.W., Pardoll, D.M. and Sears, C.L. Microbiota organization is a distinct feature of proximal colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America.* 2014, **111**(51), pp.18321-18326.
126. Yu, J., Chen, Y., Fu, X., Zhou, X., Peng, Y., Shi, L., Chen, T. and Wu, Y. Invasive Fusobacterium nucleatum may play a role in the carcinogenesis of proximal colon cancer through the serrated neoplasia pathway. *Int J Cancer.* 2016, **139**(6), pp.1318-1326.
127. Dejea, C.M., Fathi, P., Craig, J.M., Boleij, A., Taddese, R., Geis, A.L., Wu, X., DeStefano Shields, C.E., Hechenbleikner, E.M., Huso, D.L., Anders, R.A., Giardiello, F.M., Wick, E.C., Wang, H., Wu, S., Pardoll, D.M., Housseau, F. and Sears, C.L. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Science.* 2018, **359**(6375), pp.592-597.
128. Drewes, J.L., White, J.R., Dejea, C.M., Fathi, P., Iyadorai, T., Vadivelu, J., Roslani, A.C., Wick, E.C., Mongodin, E.F., Loke, M.F., Thulasi, K., Gan, H.M., Goh, K.L., Chong, H.Y., Kumar, S., Wanyiri, J.W. and Sears, C.L. High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *NPJ Biofilms Microbiomes.* 2017, **3**, p.34.

129. Raskov, H., Kragh, K.N., Bjarnsholt, T., Alamili, M. and Gogenur, I. Bacterial biofilm formation inside colonic crypts may accelerate colorectal carcinogenesis. *Clin Transl Med.* 2018, **7**(1), p.30.
130. Johnson, C.H., Dejea, C.M., Edler, D., Hoang, L.T., Santidrian, A.F., Felding, B.H., Ivanisevic, J., Cho, K., Wick, E.C., Hechenbleikner, E.M., Uritboonthai, W., Goetz, L., Casero, R.A., Jr., Pardoll, D.M., White, J.R., Patti, G.J., Sears, C.L. and Siuzdak, G. Metabolism links bacterial biofilms and colon carcinogenesis. *Cell metabolism.* 2015, **21**(6), pp.891-897.
131. Tomkovich, S., Dejea, C.M., Winglee, K., Drewes, J.L., Chung, L., Housseau, F., Pope, J.L., Gauthier, J., Sun, X., Muhlbauer, M., Liu, X., Fathi, P., Anders, R.A., Besharati, S., Perez-Chanona, E., Yang, Y., Ding, H., Wu, X., Wu, S., White, J.R., Gharaibeh, R.Z., Fodor, A.A., Wang, H., Pardoll, D.M., Jobin, C. and Sears, C.L. Human colon mucosal biofilms from healthy or colon cancer hosts are carcinogenic. *J Clin Invest.* 2019, **130**, pp.1699-1712.
132. Gao, Z., Guo, B., Gao, R., Zhu, Q. and Qin, H. Microbiota disbiosis is associated with colorectal cancer. *Frontiers in Microbiology.* 2015, **6**(FEB), p.20.
133. Rezasoltani, S., Asadzadeh Aghdaei, H., Dabiri, H., Akhavan Sepahi, A., Modarressi, M.H. and Nazemalhosseini Mojarad, E. The association between fecal microbiota and different types of colorectal polyp as precursors of colorectal cancer. *Microb Pathog.* 2018, **124**, pp.244-249.
134. Wu, Y., Shi, L., Li, Q., Wu, J., Peng, W., Li, H., Chen, K., Ren, Y. and Fu, X. Microbiota Diversity in Human Colorectal Cancer Tissues Is Associated with Clinicopathological Features. *Nutr Cancer.* 2019, **71**(2), pp.214-222.
135. Kosumi, K., Hamada, T., Koh, H., Borowsky, J., Bullman, S., Twombly, T.S., Nevo, D., Masugi, Y., Liu, L., da Silva, A., Chen, Y., Du, C., Gu, M., Li, C., Li, W., Liu, H., Shi, Y., Mima, K., Song, M., Nosho, K., Nowak, J.A., Nishihara, R., Baba, H., Zhang, X., Wu, K., Wang, M., Huttenhower, C., Garrett, W.S., Meyerson, M.L., Lennerz, J.K., Giannakis, M., Chan, A.T., Meyerhardt, J.A., Fuchs, C.S. and Ogino, S. The Amount of Bifidobacterium Genus in Colorectal Carcinoma Tissue in Relation to Tumor Characteristics and Clinical Outcome. *Am J Pathol.* 2018, **188**(12), pp.2839-2852.
136. Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., Hosoda, F., Rokutan, H., Matsumoto, M., Takamaru, H., Yamada, M., Matsuda, T., Iwasaki, M., Yamaji, T., Yachida, T., Soga, T., Kurokawa, K., Toyoda, A., Ogura, Y., Hayashi, T., Hatakeyama, M., Nakagama, H., Saito, Y., Fukuda, S., Shibata, T. and Yamada, T. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat Med.* 2019, **25**(6), pp.968-976.
137. Bundgaard-Nielsen, C., Baandrup, U.T., Nielsen, L.P. and Sorensen, S. The presence of bacteria varies between colorectal adenocarcinomas, precursor lesions and non-malignant tissue. *BMC cancer.* 2019, **19**(1), p.399.

138. Purcell, R.V., Visnovska, M., Biggs, P.J., Schmeier, S. and Frizelle, F.A. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Sci Rep.* 2017, **7**(1), p.11590.
139. Burns, M.B., Montassier, E., Abrahante, J., Priya, S., Niccum, D.E., Khoruts, A., Starr, T.K., Knights, D. and Blekhman, R. Colorectal cancer mutational profiles correlate with defined microbial communities in the tumor microenvironment. *PLoS Genetics.* 2018, **14**(6), p.e1007376.
140. Wang, Q., Li, L. and Xu, R. A systems biology approach to predict and characterize human gut microbial metabolites in colorectal cancer. *Sci Rep.* 2018, **8**(1), p.6225.
141. Yuan, C., Burns, M.B., Subramanian, S. and Blekhman, R. Interaction between Host MicroRNAs and the Gut Microbiota in Colorectal Cancer. *mSystems.* 2018, **3**(3), p.[no pagination].
142. Hale, V.L., Jeraldo, P., Chen, J., Mundy, M., Yao, J., Priya, S., Keeney, G., Lyke, K., Ridlon, J., White, B.A., French, A.J., Thibodeau, S.N., Diener, C., Resendis-Antonio, O., Gransee, J., Dutta, T., Petterson, X.-M., Sung, J., Blekhman, R., Boardman, L., Larson, D., Nelson, H. and Chia, N. Distinct microbes, metabolites, and ecologies define the microbiome in deficient and proficient mismatch repair colorectal cancers. *Genome medicine.* 2018, **10**(1), p.78.
143. Zhang, Y., Yu, X., Yu, E., Wang, N., Cai, Q., Shuai, Q., Yan, F., Jiang, L., Wang, H., Liu, J., Chen, Y., Li, Z. and Jiang, Q. Changes in gut microbiota and plasma inflammatory factors across the stages of colorectal tumorigenesis: A case-control study. *BMC microbiology.* 2018, **18**(1), p.92.
144. Kawano, A., Ishikawa, H., Mutoh, M., Kubota, H., Matsuda, K., Tsuji, H., Matsumoto, K., Nomoto, K., Tanaka, R., Nakamura, T., Wakabayashi, K. and Sakai, T. Higher enterococcus counts indicate a lower risk of colorectal adenomas: A prospective cohort study. *Oncotarget.* 2018, **9**(30), pp.21459-21467.
145. Ying, J., Zhou, H.Y., Liu, P., You, Q., Kuang, F., Shen, Y.N. and Hu, Z.Q. Aspirin inhibited the metastasis of colon cancer cells by inhibiting the expression of toll-like receptor 4. *Cell and Bioscience.* 2018, **8**(1), p.1.
146. Li, Y., Kundu, P., Seow, S.W., de matos, C.T., Aronsson, L., Chin, K.C., Karre, K., Pettersson, S. and Greicius, G. Gut microbiota accelerate tumor growth via c-jun and STAT3 phosphorylation in APCMin/+ mice. *Carcinogenesis.* 2012, **33**(6), pp.1231-1238.
147. Cremonesi, E., Governa, V., Garzon, J.F.G., Mele, V., Amicarella, F., Muraro, M.G., Trella, E., Galati-Fournier, V., Oertli, D., Daster, S.R., Droeser, R.A., Weixler, B., Bolli, M., Rosso, R., Nitsche, U., Khanna, N., Egli, A., Keck, S., Slotta-Huspenina, J., Terracciano, L.M., Zajac, P., Spagnoli, G.C., Eppenberger-Castori, S., Janssen, K.P., Borsig, L. and Iezzi, G. Gut microbiota modulate T cell trafficking into human colorectal cancer. *Gut.* 2018, pp.1984-1994.
148. Sittipo, P., Lobionda, S., Choi, K., Sari, I.N., Kwon, H.Y. and Lee, Y.K. Toll-like receptor 2-mediated suppression of colorectal cancer pathogenesis by polysaccharide A from Bacteroides fragilis. *Frontiers in Microbiology.* 2018, **9**(JUL), p.1588.

149. Dennis, K.L., Wang, Y., Blatner, N.R., Wang, S., Saadalla, A., Trudeau, E., Roers, A., Weaver, C.T., Lee, J.J., Gilbert, J.A., Chang, E.B. and Khazaie, K. Adenomatous polyps are driven by microbe-instigated focal inflammation and are controlled by IL-10-producing T cells. *Cancer research.* 2013, **73**(19), pp.5905-5913.
150. Zackular, J.P., Baxter, N.T., Iverson, K.D., Sadler, W.D., Petrosino, J.F., Chen, G.Y. and Schloss, P.D. The Gut Microbiome Modulates Colon Tumorigenesis. *mBio.* 2013, **4**(6), pp.e00692-00613.
151. Zackular, J.P., Baxter, N.T., Chen, G.Y. and Schloss, P.D. Manipulation of the Gut Microbiota Reveals Role in Colon Tumorigenesis. *mSphere.* 2016, **1**(1), p.[no pagination].
152. Wu, Y., Wu, J., Chen, T., Li, Q., Peng, W., Li, H., Tang, X. and Fu, X. Fusobacterium nucleatum Potentiates Intestinal Tumorigenesis in Mice via a Toll-Like Receptor 4/p21-Activated Kinase 1 Cascade. *Digestive diseases and sciences.* 2018, **63**(5), pp.1210-1218.
153. Song, X., Gao, H., Lin, Y., Yao, Y., Zhu, S., Wang, J., Liu, Y., Yao, X., Meng, G., Shen, N., Shi, Y., Iwakura, Y. and Qian, Y. Alterations in the microbiota drive interleukin-17C production from intestinal epithelial cells to promote tumorigenesis. *Immunity.* 2014, **40**(1), pp.140-152.
154. Hattori, N., Niwa, T., Ishida, T., Kobayashi, K., Imai, T., Mori, A., Kimura, K., Mori, T., Asami, Y. and Ushijima, T. Antibiotics suppress colon tumorigenesis through inhibition of aberrant DNA methylation in an azoxymethane and dextran sulfate sodium colitis model. *Cancer Sci.* 2019, **110**(1), pp.147-156.
155. Sethi, V., Kurtom, S., Tarique, M., Lavania, S., Malchiodi, Z., Hellmund, L., Zhang, L., Sharma, U., Giri, B., Garg, B., Ferrantella, A., Vickers, S.M., Banerjee, S., Dawra, R., Roy, S., Ramakrishnan, S., Saluja, A. and Dudeja, V. Gut Microbiota Promotes Tumor Growth in Mice by Modulating Immune Response. *Gastroenterology.* 2018, **155**(1), pp.33-37.e36.
156. Zhu, W., Miyata, N., Winter, M.G., Arenales, A., Hughes, E.R., Spiga, L., Kim, J., Sifuentes-Dominguez, L., Starokadomskyy, P., Gopal, P., Byndloss, M.X., Santos, R.L., Burstein, E. and Winter, S.E. Editing of the gut microbiota reduces carcinogenesis in mouse models of colitis-associated colorectal cancer. *J Exp Med.* 2019, p.[no pagination].
157. Bishehsari, F., Engen, P.A., Preite, N.Z., Tuncil, Y.E., Naqib, A., Shaikh, M., Rossi, M., Wilber, S., Green, S.J., Hamaker, B.R., Khazaie, K., Voigt, R.M., Forsyth, C.B. and Keshavarzian, A. Dietary Fiber Treatment Corrects the Composition of Gut Microbiota, Promotes SCFA Production, and Suppresses Colon Carcinogenesis. *Genes (Basel).* 2018, **9**(2), p.102.
158. Grivennikov, S.I., Wang, K., Mucida, D., Stewart, C.A., Schnabl, B., Jauch, D., Taniguchi, K., Yu, G.Y., Osterreicher, C.H., Hung, K.E., Datz, C., Feng, Y., Fearon, E.R., Oukka, M., Tessarollo, L., Coppola, V., Yarovinsky, F., Cheroutre, H., Eckmann, L., Trinchieri, G. and Karin, M. Adenoma-linked barrier defects and microbial products drive IL-23/IL-17-mediated tumour growth. *Nature.* 2012, **491**(7423), pp.254-258.
159. Holtorf, A., Conrad, A., Holzmann, B. and Janssen, K.P. Cell-type specific MyD88 signaling is required for intestinal tumor initiation and progression to malignancy. *OncoImmunology.* 2018, **7**(8), p.e1466770.

160. Triner, D., Devenport, S.N., Ramakrishnan, S.K., Ma, X., Frieler, R.A., Greenson, J.K., Inohara, N., Nunez, G., Colacino, J.A., Mortensen, R.M. and Shah, Y.M. Neutrophils Restrict Tumor-Associated Microbiota to Reduce Growth and Invasion of Colon Tumors in Mice. *Gastroenterology.* 2019, **156**(5), pp.1467-1482.
161. Miyata, N., Morris, L.L., Chen, Q., Thorne, C., Singla, A., Zhu, W., Winter, M., Melton, S.D., Li, H., Sifuentes-Dominguez, L., Llano, E., Huff-Hardy, K., Starokadomskyy, P., Lopez, A., Reese, T.A., Turer, E., Billadeau, D.D., Winter, S.E. and Burstein, E. Microbial Sensing by Intestinal Myeloid Cells Controls Carcinogenesis and Epithelial Differentiation. *Cell Rep.* 2018, **24**(9), pp.2342-2355.
162. Coleman, O.I., Lobner, E.M., Bierwirth, S., Sorbie, A., Waldschmitt, N., Rath, E., Berger, E., Lagkouvardos, I., Clavel, T., McCoy, K.D., Weber, A., Heikenwalder, M., Janssen, K.P. and Haller, D. Activated ATF6 Induces Intestinal Dysbiosis and Innate Immune Response to Promote Colorectal Tumorigenesis. *Gastroenterology.* 2018, **155**(5), pp.1539-1552.e1512.
163. Wong, S.H., Zhao, L., Zhang, X., Nakatsu, G., Han, J., Xu, W., Xiao, X., Kwong, T.N.Y., Tsoi, H., Wu, W.K.K., Zeng, B., Chan, F.K.L., Sung, J.J.Y., Wei, H. and Yu, J. Gavage of Fecal Samples From Patients With Colorectal Cancer Promotes Intestinal Carcinogenesis in Germ-Free and Conventional Mice. *Gastroenterology.* 2017, **153**(6), p.1621.
164. Baxter, N.T., Zackular, J.P., Chen, G.Y. and Schloss, P.D. Structure of the gut microbiome following colonization with human feces determines colonic tumor burden. *Microbiome.* 2014, **2**, p.20.
165. Son, J.S., Khair, S., Pettet, D.W., 3rd, Ouyang, N., Tian, X., Zhang, Y., Zhu, W., Mackenzie, G.G., Robertson, C.E., Ir, D., Frank, D.N., Rigas, B. and Li, E. Altered Interactions between the Gut Microbiome and Colonic Mucosa Precede Polyposis in APCMin/+ Mice. *PloS one.* 2015, **10**(6), p.e0127985.
166. Dai, Z., Coker, O.O., Nakatsu, G., Wu, W.K.K., Zhao, L., Chen, Z., Chan, F.K.L., Kristiansen, K., Sung, J.J.Y., Wong, S.H. and Yu, J. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome.* 2018, **6**(1), p.70.
167. Brennan, C.A. and Garrett, W.S. Fusobacterium nucleatum — symbiont, opportunist and oncobacterium. *Nature Reviews Microbiology.* 2019, **17**(3), pp.156-166.
168. Yang, N.-Y., Zhang, Q., Li, J.-L., Yang, S.-H. and Shi, Q. Progression of periodontal inflammation in adolescents is associated with increased number of Porphyromonas gingivalis, Prevotella intermedia, Tannerella forsythensis, and Fusobacterium nucleatum. *International Journal of Paediatric Dentistry.* 2014, **24**(3), pp.226-233.
169. Fujii, R., Saito, Y., Tokura, Y., Nakagawa, K.I., Okuda, K. and Ishihara, K. Characterization of bacterial flora in persistent apical periodontitis lesions. *Oral microbiology and immunology.* 2009, **24**(6), pp.502-505.
170. Yamamura, K., Baba, Y., Miyake, K., Nakamura, K., Shigaki, H., Mima, K., Kurashige, J., Ishimoto, T., Iwatsuki, M., Sakamoto, Y., Yamashita, Y., Yoshida, N., Watanabe, M. and Baba, H. Fusobacterium nucleatum in gastroenterological cancer: Evaluation of measurement methods

using quantitative polymerase chain reaction and a literature review. *Oncol Lett.* 2017, **14**(6), pp.6373-6378.

171. Kostic, A.D., Chun, E., Robertson, L., Glickman, J.N., Gallini, C.A., Michaud, M., Clancy, T.E., Chung, D.C., Lochhead, P., Hold, G.L., El-Omar, E.M., Brenner, D., Fuchs, C.S., Meyerson, M. and Garrett, W.S. Fusobacterium nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment. *Cell Host and Microbe.* 2013, **14**(2), pp.207-215.
172. Castellarin, M., Warren, R.L., Freeman, J.D., Dreolini, L., Krzywinski, M., Strauss, J., Barnes, R., Watson, P., Allen-Vercoe, E., Moore, R.A. and Holt, R.A. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome research.* 2012, **22**(2), pp.299-306.
173. Kostic, A.D., Gevers, D., Pedamallu, C.S., Michaud, M., Duke, F., Earl, A.M., Ojesina, A.I., Jung, J., Bass, A.J., Tabernero, J., Baselga, J., Liu, C., Shivdasani, R.A., Ogino, S., Birren, B.W., Huttenhower, C., Garrett, W.S. and Meyerson, M. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome research.* 2012, **22**(2), pp.292-298.
174. Lee, D.W., Han, S.W., Kang, J.K., Bae, J.M., Kim, H.P., Won, J.K., Jeong, S.Y., Park, K.J., Kang, G.H. and Kim, T.Y. Association Between Fusobacterium nucleatum, Pathway Mutation, and Patient Prognosis in Colorectal Cancer. *Annals of Surgical Oncology.* 2018, **25**(11), pp.3389-3395.
175. Flanagan, L., Schmid, J., Ebert, M., Soucek, P., Kunicka, T., Liska, V., Bruha, J., Neary, P., Dezeeuw, N., Tommasino, M., Jenab, M., Prehn, J.H.M. and Hughes, D.J. Fusobacterium nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease outcome. *European Journal of Clinical Microbiology and Infectious Diseases.* 2014, **33**(8), pp.1381-1390.
176. Yoon, H., Kim, N., Park, J.H., Kim, Y.S., Lee, J., Kim, H.W., Choi, Y.J., Shin, C.M., Park, Y.S., Lee, D.H. and Jung, H.C. Comparisons of Gut Microbiota Among Healthy Control, Patients With Conventional Adenoma, Sessile Serrated Adenoma, and Colorectal Cancer. *J Cancer Prev.* 2017, **22**(2), pp.108-114.
177. Mangifesta, M., Mancabelli, L., Milani, C., Gaiani, F., de'Angelis, N., de'Angelis, G.L., van Sinderen, D., Ventura, M. and Turroni, F. Mucosal microbiota of intestinal polyps reveals putative biomarkers of colorectal cancer. *Sci Rep.* 2018, **8**(1), p.13974.
178. Tunsjo, H.S., Gundersen, G., Rangnes, F., Noone, J.C., Endres, A. and Bemanian, V. Detection of Fusobacterium nucleatum in stool and colonic tissues from Norwegian colorectal cancer patients. *Eur J Clin Microbiol Infect Dis.* 2019, **38**(7), pp.1367-1376.
179. Leung, P.H.M., Subramanya, R., Mou, Q., Lee, K.T., Islam, F., Gopalan, V., Lu, C.T. and Lam, A.K. Characterization of Mucosa-Associated Microbiota in Matched Cancer and Non-neoplastic Mucosa From Patients With Colorectal Cancer. *Front Microbiol.* 2019, **10**, p.1317.
180. Proenca, M.A., Biselli, J.M., Succi, M., Severino, F.E., Berardinelli, G.N., Caetano, A., Reis, R.M., Hughes, D.J. and Silva, A.E. Relationship between Fusobacterium nucleatum, inflammatory

mediators and microRNAs in colorectal carcinogenesis. *World J Gastroenterol.* 2018, **24**(47), pp.5351-5365.

181. Yan, X., Liu, L., Li, H., Qin, H. and Sun, Z. Clinical significance of Fusobacterium nucleatum, epithelial-mesenchymal transition, and cancer stem cell markers in stage III/IV colorectal cancer patients. *Onco Targets Ther.* 2017, **10**, pp.5031-5046.
182. Kinross, J., Mirnezami, R., Alexander, J., Brown, R., Scott, A., Galea, D., Veselkov, K., Goldin, R., Darzi, A., Nicholson, J. and Marchesi, J.R. 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.
183. Ye, X., Wang, R., Bhattacharya, R., Boulbes, D.R., Fan, F., Xia, L., Adoni, H., Ajami, N.J., Wong, M.C., Smith, D.P., Petrosino, J.F., Venable, S., Qiao, W., Baladandayuthapani, V., Maru, D. and Ellis, L.M. Fusobacterium Nucleatum Subspecies Animalis Influences Proinflammatory Cytokine Expression and Monocyte Activation in Human Colorectal Tumors. *Cancer Prev Res (Phila).* 2017, **10**(7), pp.398-409.
184. Mjelle, R., Sjursen, W., Thommesen, L., Saetrom, P. and Hofsli, E. Small RNA expression from viruses, bacteria and human miRNAs in colon cancer tissue and its association with microsatellite instability and tumor location. *BMC cancer.* 2019, **19**(1), p.161.
185. Hamada, T., Zhang, X., Mima, K., Bullman, S., Sukawa, Y., Nowak, J.A., Kosumi, K., Masugi, Y., Twombly, T.S., Cao, Y., Song, M., Liu, L., da Silva, A., Shi, Y., Gu, M., Li, W., Koh, H., Nosho, K., Inamura, K., Keum, N., Wu, K., Meyerhardt, J.A., Kostic, A.D., Huttenhower, C., Garrett, W., Meyerson, M., Giovannucci, E.L., Chan, A.T., Fuchs, C.S., Nishihara, R., Giannakis, M. and Ogino, S. Fusobacterium nucleatum in Colorectal Cancer Relates to Immune Response Differentially by Tumor Microsatellite Instability Status. *Cancer Immunol Res.* 2018, pp.1327-1336.
186. Ito, M., Kanno, S., Nosho, K., Sukawa, Y., Mitsuhashi, K., Kurihara, H., Igarashi, H., Takahashi, T., Tachibana, M., Takahashi, H., Yoshii, S., Takenouchi, T., Hasegawa, T., Okita, K., Hirata, K., Maruyama, R., Suzuki, H., Imai, K., Yamamoto, H. and Shinomura, Y. Association of Fusobacterium nucleatum with clinical and molecular features in colorectal serrated pathway. *International journal of cancer.* 2015, **137**(6), pp.1258-1268.
187. Tahara, T., Yamamoto, E., Suzuki, H., Maruyama, R., Chung, W., Garriga, J., Jelinek, J., Yamano, H.O., Sugai, T., An, B., Shureiqi, I., Toyota, M., Kondo, Y., Estecio, M.R.H. and Issa, J.P.J. Fusobacterium in colonic flora and molecular features of colorectal carcinoma. *Cancer research.* 2014, **74**(5), pp.1311-1318.
188. Peng, B.J., Cao, C.Y., Li, W., Zhou, Y.J., Zhang, Y., Nie, Y.Q., Cao, Y.W. and Li, Y.Y. Diagnostic Performance of Intestinal Fusobacterium nucleatum in Colorectal Cancer: A Meta-Analysis. *Chinese medical journal.* 2018, **131**(11), pp.1349-1356.
189. Huang, Q., Peng, Y. and Xie, F. Fecal fusobacterium nucleatum for detecting colorectal cancer: a systematic review and meta-analysis. *Int J Biol Markers.* 2018, p.1724600818781301.

190. Mima, K., Sukawa, Y., Nishihara, R., Qian, Z.R., Yamauchi, M., Inamura, K., Kim, S.A., Masuda, A., Nowak, J.A., Nosho, K., Kostic, A.D., Giannakis, M., Watanabe, H., Bullman, S., Milner, D.A., Harris, C.C., Giovannucci, E., Garraway, L.A., Freeman, G.J., Dranoff, G., Chan, A.T., Garrett, W.S., Huttenhower, C., Fuchs, C.S. and Ogino, S. Fusobacterium nucleatum and T Cells in Colorectal Carcinoma. *JAMA Oncol.* 2015, **1**(5), pp.653-661.
191. Yamaoka, Y., Suehiro, Y., Hashimoto, S., Hoshida, T., Fujimoto, M., Watanabe, M., Imanaga, D., Sakai, K., Matsumoto, T., Nishioka, M., Takami, T., Suzuki, N., Hazama, S., Nagano, H., Sakaida, I. and Yamasaki, T. Fusobacterium nucleatum as a prognostic marker of colorectal cancer in a Japanese population. *Journal of gastroenterology.* 2018, **53**(4), pp.517-524.
192. Liu, L., Tabung, F.K., Zhang, X., Nowak, J.A., Qian, Z.R., Hamada, T., Nevo, D., Bullman, S., Mima, K., Kosumi, K., da Silva, A., Song, M., Cao, Y., Twombly, T.S., Shi, Y., Liu, H., Gu, M., Koh, H., Du, C., Chen, Y., Li, C., Li, W., Mehta, R.S., Wu, K., Wang, M., Kostic, A.D., Giannakis, M., Garrett, W.S., Hutthenhower, C., Chan, A.T., Fuchs, C.S., Nishihara, R., Ogino, S. and Giovannucci, E.L. Diets That Promote Colon Inflammation Associate With Risk of Colorectal Carcinomas That Contain Fusobacterium nucleatum. *Clinical Gastroenterology and Hepatology.* 2018, **16**(10), p.1622.
193. Oh, H.J., Kim, J.H., Bae, J.M., Kim, H.J., Cho, N.Y. and Kang, G.H. Prognostic Impact of Fusobacterium nucleatum Depends on Combined Tumor Location and Microsatellite Instability Status in Stage II/III Colorectal Cancers Treated with Adjuvant Chemotherapy. *J Pathol Transl Med.* 2019, **53**(1), pp.40-49.
194. Advani, S.M., Advani, P., DeSantis, S.M., Brown, D., VonVille, H.M., Lam, M., Loree, J.M., Mehrvarz Sarshekeh, A., Bressler, J., Lopez, D.S., Daniel, C.R., Swartz, M.D. and Kopetz, S. Clinical, Pathological, and Molecular Characteristics of CpG Island Methylator Phenotype in Colorectal Cancer: A Systematic Review and Meta-analysis. *Transl Oncol.* 2018, **11**(5), pp.1188-1201.
195. Mehta, R.S., Nishihara, R., Cao, Y., Song, M., Mima, K., Qian, Z.R., Nowak, J.A., Kosumi, K., Hamada, T., Masugi, Y., Bullman, S., Drew, D.A., Kostic, A.D., Fung, T.T., Garrett, W.S., Huttenhower, C., Wu, K., Meyerhardt, J.A., Zhang, X., Willett, W.C., Giovannucci, E.L., Fuchs, C.S., Chan, A.T. and Ogino, S. Association of Dietary Patterns With Risk of Colorectal Cancer Subtypes Classified by Fusobacterium nucleatum in Tumor Tissue. *JAMA Oncol.* 2017, **3**(7), pp.921-927.
196. Gao, R., Kong, C., Huang, L., Li, H., Qu, X., Liu, Z., Lan, P., Wang, J. and Qin, H. Mucosa-associated microbiota signature in colorectal cancer. *European Journal of Clinical Microbiology and Infectious Diseases.* 2017, **36**(11), pp.2073-2083.
197. Komiya, Y., Shimomura, Y., Higurashi, T., Sugi, Y., Arimoto, J., Umezawa, S., Uchiyama, S., Matsumoto, M. and Nakajima, A. Patients with colorectal cancer have identical strains of Fusobacterium nucleatum in their colorectal cancer and oral cavity. *Gut.* 2018, p.[no pagination].
198. Russo, E., Bacci, G., Chiellini, C., Fagorzi, C., Niccolai, E., Taddei, A., Ricci, F., Ringressi, M.N., Borrelli, R., Melli, F., Miloeva, M., Bechi, P.,

Mengoni, A., Fani, R. and Amedei, A. Preliminary comparison of oral and intestinal human microbiota in patients with colorectal cancer: A pilot study. *Frontiers in Microbiology.* 2018, **8**(JAN), p.2699.

199. Kageyama, S., Takeshita, T., Takeuchi, K., Asakawa, M., Matsumi, R., Furuta, M., Shibata, Y., Nagai, K., Ikebe, M., Morita, M., Masuda, M., Toh, Y., Kiyohara, Y., Ninomiya, T. and Yamashita, Y. Characteristics of the Salivary Microbiota in Patients With Various Digestive Tract Cancers. *Front Microbiol.* 2019, **10**, p.1780.
200. Yang, Y., Cai, Q., Shu, X.O., Steinwandel, M.D., Blot, W.J., Zheng, W. and Long, J. Prospective study of oral microbiome and colorectal cancer risk in low-income and African American populations. *Int J Cancer.* 2019, **144**(10), pp.2381-2389.
201. Guo, S., Li, L., Xu, B., Li, M., Zeng, Q., Xiao, H., Xue, Y., Wu, Y., Wang, Y., Liu, W. and Zhang, G. A simple and novel fecal biomarker for colorectal cancer: Ratio of Fusobacterium nucleatum to probiotics populations, based on their antagonistic effect. *Clinical chemistry.* 2018, **64**(9), pp.1327-1337.
202. Abed, J., Maalouf, N., Parhi, L., Chaushu, S., Mandelboim, O. and Bachrach, G. Tumor targeting by Fusobacterium nucleatum: A Pilot Study and future Perspectives. *Frontiers in cellular and infection microbiology.* 2017, **7**(JUN), p.295.
203. Abed, J., Emgard, J.E., Zamir, G., Faroja, M., Almogy, G., Grenov, A., Sol, A., Naor, R., Pikarsky, E., Atlan, K.A., Mellul, A., Chaushu, S., Manson, A.L., Earl, A.M., Ou, N., Brennan, C.A., Garrett, W.S. and Bachrach, G. Fap2 Mediates Fusobacterium nucleatum Colorectal Adenocarcinoma Enrichment by Binding to Tumor-Expressed Gal-GalNAc. *Cell Host Microbe.* 2016, **20**(2), pp.215-225.
204. Bullman, S., Pedamallu, C.S., Sicinska, E., Clancy, T.E., Zhang, X., Cai, D., Neuberg, D., Huang, K., Guevara, F., Nelson, T., Chipashvili, O., Hagan, T., Walker, M., Ramachandran, A., Diosdado, B., Serna, G., Mulet, N., Landolfi, S., Ramon, Y.C.S., Fasani, R., Aguirre, A.J., Ng, K., Elez, E., Ogino, S., Tabernero, J., Fuchs, C.S., Hahn, W.C., Nuciforo, P. and Meyerson, M. Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. *Science.* 2017, **358**(6369), pp.1443-1448.
205. Flynn, K.J., Baxter, N.T. and Schloss, P.D. Metabolic and Community Synergy of Oral Bacteria in Colorectal Cancer. *mSphere.* 2016, **1**(3), p.[no pagination].
206. Xue, Y., Xiao, H., Guo, S., Xu, B., Liao, Y., Wu, Y. and Zhang, G. Indoleamine 2,3-dioxygenase expression regulates the survival and proliferation of Fusobacterium nucleatum in THP-1-derived macrophages. *Cell Death and Disease.* 2018, **9**(3), p.355.
207. Park, H.E., Kim, J.H., Cho, N.Y., Lee, H.S. and Kang, G.H. Intratumoral Fusobacterium nucleatum abundance correlates with macrophage infiltration and CDKN2A methylation in microsatellite-unstable colorectal carcinoma. *Virchows Arch.* 2017, **471**(3), pp.329-336.
208. Chen, T., Li, Q., Zhang, X., Long, R., Wu, Y., Wu, J. and Fu, X. TOX expression decreases with progression of colorectal cancers and is associated with CD4 T-cell density and Fusobacterium nucleatum infection. *Hum Pathol.* 2018, **79**, pp.93-101.

209. Chen, T., Li, Q., Wu, J., Wu, Y., Peng, W., Li, H., Wang, J., Tang, X., Peng, Y. and Fu, X. Fusobacterium nucleatum promotes M2 polarization of macrophages in the microenvironment of colorectal tumours via a TLR4-dependent mechanism. *Cancer Immunol Immunother.* 2018, **67**(10), pp.1635-1646.
210. Gur, C., Ibrahim, Y., Isaacson, B., Yamin, R., Abed, J., Gamliel, M., Enk, J., Bar-On, Y., Stanietsky-Kaynan, N., Coppenhagen-Glazer, S., Shussman, N., Almogy, G., Cuapio, A., Hofer, E., Mevorach, D., Tabib, A., Ortenberg, R., Markel, G., Miklic, K., Jonjic, S., Brennan, C.A., Garrett, W.S., Bachrach, G. and Mandelboim, O. Binding of the Fap2 protein of Fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. *Immunity.* 2015, **42**(2), pp.344-355.
211. Gur, C., Maalouf, N., Shhadeh, A., Berhani, O., Singer, B.B., Bachrach, G. and Mandelboim, O. Fusobacterium nucleatum supresses anti-tumor immunity by activating CEACAM1. *OncoImmunology.* 2019, **8**(6), p.e1581531.
212. Rubinstein, M.R., Wang, X., Liu, W., Hao, Y., Cai, G. and Han, Y.W. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/beta-catenin signaling via its FadA adhesin. *Cell Host Microbe.* 2013, **14**(2), pp.195-206.
213. Rubinstein, M.R., Baik, J.E., Lagana, S.M., Han, R.P., Raab, W.J., Sahoo, D., Dalerba, P., Wang, T.C. and Han, Y.W. Fusobacterium nucleatum promotes colorectal cancer by inducing Wnt/beta-catenin modulator Annexin A1. *EMBO Rep.* 2019, **20**(4), p.[no pagination].
214. Ma, C.T., Luo, H.S., Gao, F., Tang, Q.C. and Chen, W. Fusobacterium nucleatum promotes the progression of colorectal cancer by interacting with e-cadherin. *Oncology Letters.* 2018, **16**(2), pp.2606-2612.
215. Yang, Y., Weng, W., Peng, J., Hong, L., Yang, L., Toiyama, Y., Gao, R., Liu, M., Yin, M., Pan, C., Li, H., Guo, B., Zhu, Q., Wei, Q., Moyer, M.-P., Wang, P., Cai, S., Goel, A., Qin, H. and Ma, Y. Fusobacterium nucleatum Increases Proliferation of Colorectal Cancer Cells and Tumor Development in Mice by Activating Toll-Like Receptor 4 Signaling to Nuclear Factor-kappaB, and Up-regulating Expression of MicroRNA-21. *Gastroenterology.* 2017, **152**(4), pp.851-866.e824.
216. Noguti, J., Chan, A.A., Bandera, B., Brislawn, C.J., Protic, M., Sim, M.S., Jansson, J.K., Bilchik, A.J. and Lee, D.J. Both the intratumoral immune and microbial microenvironment are linked to recurrence in human colon cancer: Results from a prospective, multicenter nodal ultrastaging trial. *Oncotarget.* 2018, **9**(34), pp.23564-23576.
217. Mima, K., Nishihara, R., Qian, Z.R., Cao, Y., Sukawa, Y., Nowak, J.A., Yang, J., Dou, R., Masugi, Y., Song, M., Kostic, A.D., Giannakis, M., Bullman, S., Milner, D.A., Baba, H., Giovannucci, E.L., Garraway, L.A., Freeman, G.J., Dranoff, G., Garrett, W.S., Huttenhower, C., Meyerson, M., Meyerhardt, J.A., Chan, A.T., Fuchs, C.S. and Ogino, S. Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis. *Gut.* 2016, **65**(12), pp.1973-1980.
218. Wei, Z., Cao, S., Liu, S., Yao, Z., Sun, T., Li, Y., Li, J., Zhang, D. and Zhou, Y. Could gut microbiota serve as prognostic biomarker associated with colorectal cancer patients' survival? A pilot study on relevant mechanism. *Oncotarget.* 2016, **7**(29), pp.46158-46172.

219. Yu, T., Guo, F., Yu, Y., Sun, T., Ma, D., Han, J., Qian, Y., Kryczek, I., Sun, D., Nagarsheth, N., Chen, Y., Chen, H., Hong, J., Zou, W. and Fang, J.-Y. Fusobacterium nucleatum Promotes Chemoresistance to Colorectal Cancer by Modulating Autophagy. *Cell.* 2017, **170**(3), pp.548-563.e516.
220. Kunzmann, A.T., Proenca, M.A., Jordao, H.W., Jiraskova, K., Schneiderova, M., Levy, M., Liska, V., Buchler, T., Vodickova, L., Vymetalkova, V., Silva, A.E., Vodicka, P. and Hughes, D.J. Fusobacterium nucleatum tumor DNA levels are associated with survival in colorectal cancer patients. *Eur J Clin Microbiol Infect Dis.* 2019, p.[no pagination].
221. Zhang, S., Yang, Y., Weng, W., Guo, B., Cai, G., Ma, Y. and Cai, S. 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.
222. Sears, C.L. Enterotoxigenic Bacteroides fragilis: a rogue among symbiotes. *Clin Microbiol Rev.* 2009, **22**(2), pp.349-369.
223. Pierce, J.V. and Bernstein, H.D. Genomic diversity of enterotoxigenic strains of Bacteroides fragilis. *PloS one.* 2016, **11**(6), p.e0158171.
224. Chen, L.A., Van Meerbeke, S., Albesiano, E., Goodwin, A., Wu, S., Yu, H., Carroll, K. and Sears, C. Fecal detection of enterotoxigenic Bacteroides fragilis. *European Journal of Clinical Microbiology and Infectious Diseases.* 2015, **34**(9), pp.1871-1877.
225. Mootien, S. and Kaplan, P.M. Monoclonal antibodies specific for Bacteroides fragilis enterotoxins BFT1 and BFT2 and their use in immunoassays. *PloS one.* 2017, **12**(3), p.e0173128.
226. Boleij, A., Hechenbleikner, E.M., Goodwin, A.C., Badani, R., Stein, E.M., Lazarev, M.G., Ellis, B., Carroll, K.C., Albesiano, E., Wick, E.C., Platz, E.A., Pardoll, D.M. and Sears, C.L. The bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clinical Infectious Diseases.* 2015, **60**(2), pp.208-215.
227. Ulger Toprak, N., Yagci, A., Gulluoglu, B.M., Akin, M.L., Demirkalem, P., Celenk, T. and Soyletir, G. A possible role of Bacteroides fragilis enterotoxin in the aetiology of colorectal cancer. *Clinical Microbiology and Infection.* 2006, **12**(8), pp.782-786.
228. Keenan, J.I., Aitchison, A., Purcell, R.V., Greenlees, R., Pearson, J.F. and Frizelle, F.A. Screening for enterotoxigenic Bacteroides fragilis in stool samples. *Anaerobe.* 2016, **40**, pp.50-53.
229. Purcell, R.V., Pearson, J., Aitchison, A., Dixon, L., Frizelle, F.A. and Keenan, J.I. Colonization with enterotoxigenic Bacteroides fragilis is associated with early-stage colorectal neoplasia. *PloS one.* 2017, **12**(2), p.e0171602.
230. Sanfilippo, L., Baldwin, T.J., Menozzi, M.G., Borriello, S.P. and Mahida, Y.R. Heterogeneity in responses by primary adult human colonic epithelial cells to purified enterotoxin of Bacteroides fragilis. *Gut.* 1998, **43**(5), p.651.
231. Obiso, R.J., Jr., Azghani, A.O. and Wilkins, T.D. The Bacteroides fragilis toxin fragilysin disrupts the paracellular barrier of epithelial cells. *Infect Immun.* 1997, **65**(4), pp.1431-1439.
232. Riegler, M., Lotz, M., Sears, C., Pothoulakis, C., Castagliuolo, I., Wang, C.C., Sedivy, R., Sogukoglu, T., Cosentini, E., Bischof, G., Feil, W.,

Teleky, B., Hamilton, G., LaMont, J.T. and Wenzl, E. Bacteroides fragilis toxin 2 damages human colonic mucosa in vitro. *Gut.* 1999, **44**(4), pp.504-510.

233. Wick, E.C., Rabizadeh, S., Albesiano, E., Wu, X., Wu, S., Chan, J., Rhee, K.-J., Ortega, G., Huso, D.L., Pardoll, D., Housseau, F. and Sears, C.L. Stat3 activation in murine colitis induced by enterotoxigenic Bacteroides fragilis. *Inflammatory bowel diseases.* 2014, **20**(5), pp.821-834.
234. Wells, C.L., van de Westerlo, E.M., Jechorek, R.P., Feltis, B.A., Wilkins, T.D. and Erlandsen, S.L. Bacteroides fragilis enterotoxin modulates epithelial permeability and bacterial internalization by HT-29 enterocytes. *Gastroenterology.* 1996, **110**(5), pp.1429-1437.
235. Sears, C.L., Islam, S., Saha, A., Arjumand, M., Alam, N.H., Faruque, A.S., Salam, M.A., Shin, J., Hecht, D., Weintraub, A., Sack, R.B. and Qadri, F. Association of enterotoxigenic Bacteroides fragilis infection with inflammatory diarrhea. *Clin Infect Dis.* 2008, **47**(6), pp.797-803.
236. Rhee, K.J., Wu, S., Wu, X., Huso, D.L., Karim, B., Franco, A.A., Rabizadeh, S., Golub, J.E., Mathews, L.E., Shin, J., Balfour Sartor, R., Golenbock, D., Hamad, A.R., Gan, C.M., Housseau, F. and Sears, C.L. Induction of persistent colitis by a human commensal, enterotoxigenic Bacteroides fragilis, in wild-type C57BL/6 mice. *Infection and immunity.* 2009, **77**(4), pp.1708-1718.
237. Chung, L., Thiele Orberg, E., Geis, A.L., Chan, J.L., Fu, K., DeStefano Shields, C.E., Dejea, C.M., Fathi, P., Chen, J., Finard, B.B., Tam, A.J., McAllister, F., Fan, H., Wu, X., Ganguly, S., Lebid, A., Metz, P., Van Meerbeke, S.W., Huso, D.L., Wick, E.C., Pardoll, D.M., Wan, F., Wu, S., Sears, C.L. and Housseau, F. Bacteroides fragilis Toxin Coordinates a Pro-carcinogenic Inflammatory Cascade via Targeting of Colonic Epithelial Cells. *Cell host & microbe.* 2018, **23**(2), pp.203-214.e205.
238. Wu, S., Rhee, K.J., Albesiano, E., Rabizadeh, S., Wu, X., Yen, H.R., Huso, D.L., Brancati, F.L., Wick, E., McAllister, F., Housseau, F., Pardoll, D.M. and Sears, C.L. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nat Med.* 2009, **15**(9), pp.1016-1022.
239. Geis, A.L., Fan, H., Wu, X., Wu, S., Huso, D.L., Wolfe, J.L., Sears, C.L., Pardoll, D.M. and Housseau, F. Regulatory T-cell Response to Enterotoxigenic Bacteroides fragilis Colonization Triggers IL17-Dependent Colon Carcinogenesis. *Cancer discovery.* 2015, **5**(10), pp.1098-1109.
240. Hecht, A.L., Casterline, B.W., Choi, V.M. and Bubeck Wardenburg, J. A Two-Component System Regulates Bacteroides fragilis Toxin to Maintain Intestinal Homeostasis and Prevent Lethal Disease. *Cell Host and Microbe.* 2017, **22**(4), p.443.
241. Kim, J.M., Lee, J.Y. and Kim, Y.J. Inhibition of apoptosis in Bacteroides fragilis enterotoxin-stimulated intestinal epithelial cells through the induction of c-IAP-2. *Eur J Immunol.* 2008, **38**(8), pp.2190-2199.
242. Kim, J.M., Oh, Y.K., Kim, Y.J., Oh, H.B. and Cho, Y.J. Polarized secretion of CXC chemokines by human intestinal epithelial cells in response to Bacteroides fragilis enterotoxin: NF-kappa B plays a major

role in the regulation of IL-8 expression. *Clin Exp Immunol.* 2001, **123**(3), pp.421-427.

243. Kim, J.M., Cho, S.J., Oh, Y.K., Jung, H.Y., Kim, Y.J. and Kim, N. Nuclear factor-kappa B activation pathway in intestinal epithelial cells is a major regulator of chemokine gene expression and neutrophil migration induced by Bacteroides fragilis enterotoxin. *Clin Exp Immunol.* 2002, **130**(1), pp.59-66.
244. Wu, S., Powell, J., Mathioudakis, N., Kane, S., Fernandez, E. and Sears, C.L. Bacteroides fragilis enterotoxin induces intestinal epithelial cell secretion of interleukin-8 through mitogen-activated protein kinases and a tyrosine kinase-regulated nuclear factor-kappaB pathway. *Infect Immun.* 2004, **72**(10), pp.5832-5839.
245. Wu, S., Lim, K.C., Huang, J., Saidi, R.F. and Sears, C.L. Bacteroides fragilis enterotoxin cleaves the zonula adherens protein, E-cadherin. *Proceedings of the National Academy of Sciences of the United States of America.* 1998, **95**(25), pp.14979-14984.
246. Wu, S., Rhee, K.J., Zhang, M., Franco, A. and Sears, C.L. Bacteroides fragilis toxin stimulates intestinal epithelial cell shedding and gamma-secretase-dependent E-cadherin cleavage. *J Cell Sci.* 2007, **120**(Pt 11), pp.1944-1952.
247. Wu, S., Shin, J., Zhang, G., Cohen, M., Franco, A. and Sears, C.L. The Bacteroides fragilis toxin binds to a specific intestinal epithelial cell receptor. *Infect Immun.* 2006, **74**(9), pp.5382-5390.
248. Wu, S., Morin, P.J., Maouyo, D. and Sears, C.L. Bacteroides fragilis enterotoxin induces c-Myc expression and cellular proliferation. *Gastroenterology.* 2003, **124**(2), pp.392-400.
249. Destefano Shields, C.E., Van Meerbeke, S.W., Housseau, F., Wang, H., Huso, D.L., Casero, R.A., O'Hagan, H.M. and Sears, C.L. Reduction of Murine Colon Tumorigenesis Driven by Enterotoxigenic Bacteroides fragilis Using Cefoxitin Treatment. *Journal of Infectious Diseases.* 2016, **214**(1), pp.122-129.
250. Buc, E., Dubois, D., Sauvanet, P., Raisch, J., Delmas, J., Darfeuille-Michaud, A., Pezet, D. and Bonnet, R. High prevalence of mucosa-associated E. coli producing cyclomodulin and genotoxin in colon cancer. *PloS one.* 2013, **8**(2), p.e56964.
251. Schmidt, H. and Hensel, M. Pathogenicity Islands in Bacterial Pathogenesis. *Clinical microbiology reviews.* 2004, **17**(1), p.14.
252. Prorok-Hamon, M., Friswell, M.K., Alswied, A., Roberts, C.L., Song, F., Flanagan, P.K., Knight, P., Codling, C., Marchesi, J.R., Winstanley, C., Hall, N., Rhodes, J.M. and Campbell, B.J. Colonic mucosa-associated diffusely adherent afaC+ Escherichia coli expressing lpfA and pks are increased in inflammatory bowel disease and colon cancer. *Gut.* 2014, **63**(5), pp.761-770.
253. Arthur, J.C., Perez-Chanona, E., Muhlbauer, M., Tomkovich, S., Uronis, J.M., Fan, T.J., Campbell, B.J., Abujamel, T., Dogan, B., Rogers, A.B., Rhodes, J.M., Stintzi, A., Simpson, K.W., Hansen, J.J., Keku, T.O., Fodor, A.A. and Jobin, C. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science.* 2012, **338**(6103), pp.120-123.
254. Bronowski, C., Smith, S.L., Yokota, K., Corkill, J.E., Martin, H.M., Campbell, B.J., Rhodes, J.M., Hart, C.A. and Winstanley, C. A subset

of mucosa-associated Escherichia coli isolates from patients with colon cancer, but not Crohn's disease, share pathogenicity islands with urinary pathogenic E. coli. *Microbiology.* 2008, **154**(Pt 2), pp.571-583.

255. Li, Y., Zhang, X., Wang, L., Zhou, Y., Hassan, J.S. and Li, M. Distribution and gene mutation of enteric flora carrying beta-glucuronidase among patients with colorectal cancer. *International Journal of Clinical and Experimental Medicine.* 2015, **8**(4), pp.5310-5316.
256. Kohoutova, D., Smajs, D., Moravkova, P., Cyrany, J., Moravkova, M., Forstlova, M., Cihak, M., Rejchrt, S. and Bures, J. Escherichia coli strains of phylogenetic group B2 and D and bacteriocin production are associated with advanced colorectal neoplasia. *BMC infectious diseases.* 2014, **14**(1), p.733.
257. Bonnet, M., Buc, E., Sauvanet, P., Darcha, C., Dubois, D., Pereira, B., Dechelotte, P., Bonnet, R., Pezet, D. and Darfeuille-Michaud, A. Colonization of the human gut by E. coli and colorectal cancer risk. *Clinical Cancer Research.* 2014, **20**(4), pp.859-867.
258. Arthur, J.C., Gharaibeh, R.Z., Muhlbauer, M., Perez-Chanona, E., Uronis, J.M., McCafferty, J., Fodor, A.A. and Jobin, C. Microbial genomic analysis reveals the essential role of inflammation in bacteria-induced colorectal cancer. *Nature communications.* 2014, **5**, p.4724.
259. Swidsinski, A., Khilkin, M., Kerjaschki, D., Schreiber, S., Ortner, M., Weber, J. and Lochs, H. Association between intraepithelial Escherichia coli and colorectal cancer. *Gastroenterology.* 1998, **115**(2), pp.281-286.
260. Martin, H.M., Campbell, B.J., Hart, C.A., Mpofu, C., Nayar, M., Singh, R., Englyst, H., Williams, H.F. and Rhodes, J.M. Enhanced Escherichia coli adherence and invasion in Crohn's disease and colon cancer. *Gastroenterology.* 2004, **127**(1), pp.80-93.
261. Ambrosi, C., Sarshar, M., Aprea, M.R., Pompilio, A., Di Bonaventura, G., Strati, F., Pronio, A., Nicoletti, M., Zagaglia, C., Palamara, A.T. and Scribano, D. Colonic adenoma-associated Escherichia coli express specific phenotypes. *Microbes Infect.* 2019, p.[no pagination].
262. Zarei, O., Arabestan, M.R., Majlesi, A., Mohammadi, Y. and Alikhani, M.Y. Determination of virulence determinants of Escherichia coli strains isolated from patients with colorectal cancer compared to the healthy subjects. *Gastroenterol Hepatol Bed Bench.* 2019, **12**(1), pp.52-59.
263. Cuevas-Ramos, G., Petit, C.R., Marcq, I., Boury, M., Oswald, E. and Nougayrede, J.P. Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc Natl Acad Sci U S A.* 2010, **107**(25), pp.11537-11542.
264. Nougayrede, J.P., Homburg, S., Taieb, F., Boury, M., Brzuszkiewicz, E., Gottschalk, G., Buchrieser, C., Hacker, J., Dobrindt, U. and Oswald, E. Escherichia coli induces DNA double-strand breaks in eukaryotic cells. *Science.* 2006, **313**(5788), pp.848-851.
265. Wilson, M.R., Jiang, Y., Villalta, P.W., Stornetta, A., Boudreau, P.D., Carra, A., Brennan, C.A., Chun, E., Ngo, L., Samson, L.D., Engelward, B.P., Garrett, W.S., Balbo, S. and Balskus, E.P. The human gut bacterial genotoxin colibactin alkylates DNA. *Science.* 2019, **363**(6428), p.[no pagination].

266. Dalmasso, G., Cougnoux, A., Delmas, J., Darfeuille-Michaud, A. and Bonnet, R. The bacterial genotoxin colibactin promotes colon tumor growth by modifying the tumor microenvironment. *Gut microbes.* 2015, **5**(5), pp.675-680.
267. Maddocks, O.D., Short, A.J., Donnenberg, M.S., Bader, S. and Harrison, D.J. Attaching and effacing Escherichia coli downregulate DNA mismatch repair protein in vitro and are associated with colorectal adenocarcinomas in humans. *PloS one.* 2009, **4**(5), p.e5517.
268. Maddocks, O.D.K., Scanlon, K.M. and Donnenberg, M.S. An escherichia coli effector protein promotes host mutation via depletion of DNA mismatch repair proteins. *mBio.* 2013, **4**(3), pp.e00152-00113.
269. Klein, R.S., Recco, R.A., Catalano, M.T., Edberg, S.C., Casey, J.I. and Steigbigel, N.H. Association of Streptococcus bovis with carcinoma of the colon. *New England Journal of Medicine.* 1977, **297**(15), pp.800-802.
270. Paritsky, M., Pastukh, N., Brodsky, D., Isakovich, N. and Peretz, A. Association of Streptococcus bovis presence in colonic content with advanced colonic lesion. *World journal of gastroenterology.* 2015, **21**(18), pp.5663-5667.
271. Kumar, R., Herold, J.L., Schady, D., Davis, J., Kopetz, S., Martinez-Moczygemba, M., Murray, B.E., Han, F., Li, Y., Callaway, E., Chapkin, R.S., Dashwood, W.-M., Dashwood, R.H., Berry, T., Mackenzie, C. and Xu, Y. Streptococcus gallolyticus subsp. gallolyticus promotes colorectal tumor development. *PLoS pathogens.* 2017, **13**(7), p.e1006440.
272. Abdulamir, A.S., Hafidh, R.R. and Bakar, F.A. Molecular detection, quantification, and isolation of Streptococcus gallolyticus bacteria colonizing colorectal tumors: inflammation-driven potential of carcinogenesis via IL-1, COX-2, and IL-8. *Mol Cancer.* 2010, **9**, p.249.
273. Boleij, A., Roelofs, R., Schaeps, R.M.J., Schulin, T., Glaser, P., Swinkels, D.W., Kato, I. and Tjalsma, H. Increased exposure to bacterial antigen RpL7/L12 in early stage colorectal cancer patients. *Cancer.* 2010, **116**(17), pp.4014-4022.
274. Butt, J., Jenab, M., Willhauck-Fleckenstein, M., Michel, A., Pawlita, M., Kyro, C., Tjonneland, A., Boutron-Ruault, M.C., Carbonnel, F., Severi, G., Kaaks, R., Kuhn, T., Boeing, H., Trichopoulou, A., la Vecchia, C., Karakatsani, A., Panico, S., Tumino, R., Agnoli, C., Palli, D., Sacerdote, C., Bueno-de-Mesquita, H.B., Weiderpass, E., Sanchez, M.J., Bonet Bonet, C., Huerta, J.M., Ardanaz, E., Bradbury, K., Gunter, M., Murphy, N., Freisling, H., Riboli, E., Tsilidis, K., Aune, D., Waterboer, T. and Hughes, D.J. Prospective evaluation of antibody response to Streptococcus gallolyticus and risk of colorectal cancer. *International journal of cancer.* 2018, **143**(2), pp.245-252.
275. Aymeric, L., Donnadieu, F., Mulet, C., Du Merle, L., Nigro, G., Saffarian, A., Berard, M., Poyart, C., Robine, S., Regnault, B., Trieu-Cuot, P., Sansonetti, P.J. and Dramsi, S. Colorectal cancer specific conditions promote Streptococcus gallolyticus gut colonization. *Proceedings of the National Academy of Sciences of the United States of America.* 2017, **115**(2), pp.E283-E291.
276. Abdulamir, A.S., Hafidh, R.R. and Abu Bakar, F. The association of Streptococcus bovis/gallolyticus with colorectal tumors: the nature and

the underlying mechanisms of its etiological role. *Journal of experimental & clinical cancer research : CR.* 2011, **30**(1), pp.11-11.
277. Li, Y.X., Zhang, L., Simayi, D., Zhang, N., Tao, L., Yang, L., Zhao, J., Chen, Y.Z., Li, F. and Zhang, W.J. Human papillomavirus infection correlates with inflammatory stat3 signaling activity and IL-17 level in patients with colorectal cancer. *PloS one.* 2015, **10**(2), p.e0118391.
278. Hannigan, G.D., Duhaime, M.B., Ruffin, M.T.t., Koumpouras, C.C. and Schloss, P.D. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. *mBio.* 2018, **9**(6), p.[no pagination].
279. Bedri, S., Sultan, A.A., Alkhalaf, M., Al Moustafa, A.E. and Vranic, S. Epstein-Barr virus (EBV) status in colorectal cancer: a mini review. *Hum Vaccin Immunother.* 2019, **15**(3), pp.603-610.
280. Jarzynski, A., Zajac, P., Zebrowski, R., Boguszewska, A. and Polz-Dacewicz, M. Occurrence of BK Virus and Human Papilloma Virus in colorectal cancer. *Ann Agric Environ Med.* 2017, **24**(3), pp.440-445.
281. Kleist, B., Bagdonas, M., Oommen, P., Schoenhardt, I., Levermann, J. and Poetsch, M. The association between clinical outcome and CD8(+) lymphocytic infiltration in advanced stages of colorectal cancer differs by latent virus infection in tumour tissue. *Histopathology.* 2018, **72**(2), pp.201-215.
282. Nakatsu, G., Zhou, H., Wu, W.K.K., Wong, S.H., Coker, O.O., Dai, Z., Li, X., Szeto, C.-H., Sugimura, N., Lam, T.Y.-T., Yu, A.C.-S., Wang, X., Chen, Z., Wong, M.C.-S., Ng, S.C., Chan, M.T.V., Chan, P.K.S., Chan, F.K.L., Sung, J.J.-Y. and Yu, J. Alterations in Enteric Virome Are Associated With Colorectal Cancer and Survival Outcomes. *Gastroenterology.* 2018, **155**(2), pp.529-541.e525.
283. zur Hausen, H. Red meat consumption and cancer: reasons to suspect involvement of bovine infectious factors in colorectal cancer. *Int J Cancer.* 2012, **130**(11), pp.2475-2483.
284. zur Hausen, H. and de Villiers, E.M. Dairy cattle serum and milk factors contributing to the risk of colon and breast cancers. *Int J Cancer.* 2015, **137**(4), pp.959-967.
285. zur Hausen, H., Bund, T. and de Villiers, E.M. Infectious Agents in Bovine Red Meat and Milk and Their Potential Role in Cancer and Other Chronic Diseases. *Curr Top Microbiol Immunol.* 2017, **407**, pp.83-116.
286. Eilebrecht, S., Hotz-Wagenblatt, A., Sarachaga, V., Burk, A., Falida, K., Chakraborty, D., Nikitina, E., Tessmer, C., Whitley, C., Sauerland, C., Gunst, K., Grewe, I. and Bund, T. Expression and replication of virus-like circular DNA in human cells. *Sci Rep.* 2018, **8**(1), p.2851.
287. zur Hausen, H., Bund, T. and de Villiers, E.M. Specific nutritional infections early in life as risk factors for human colon and breast cancers several decades later. *Int J Cancer.* 2019, **144**(7), pp.1574-1583.
288. Gao, R., Kong, C., Li, H., Huang, L., Qu, X., Qin, N. and Qin, H. Dysbiosis signature of mycobiota in colon polyp and colorectal cancer. *European Journal of Clinical Microbiology and Infectious Diseases.* 2017, **36**(12), pp.2457-2468.
289. Coker, O.O., Nakatsu, G., Dai, R.Z., Wu, W.K.K., Wong, S.H., Ng, S.C., Chan, F.K.L., Sung, J.J.Y. and Yu, J. Enteric fungal microbiota

dysbiosis and ecological alterations in colorectal cancer. *Gut.* 2019, **68**(4), p.654.

290. Richard, M.L., Liguori, G., Lamas, B., Brandi, G., da Costa, G., Hoffmann, T.W., Pierluigi Di Simone, M., Calabrese, C., Poggioli, G., Langella, P., Campieri, M. and Sokol, H. Mucosa-associated microbiota dysbiosis in colitis associated cancer. *Gut microbes.* 2018, **9**(2), pp.131-142.
291. Malik, A., Sharma, D., Malireddi, R.K.S., Guy, C.S., Chang, T.C., Olsen, S.R., Neale, G., Vogel, P. and Kanneganti, T.D. SYK-CARD9 Signaling Axis Promotes Gut Fungi-Mediated Inflammasome Activation to Restrict Colitis and Colon Cancer. *Immunity.* 2018, **49**(3), pp.515-530.e515.
292. Wang, T., Fan, C., Yao, A., Xu, X., Zheng, G., You, Y., Jiang, C., Zhao, X., Hou, Y., Hung, M.C. and Lin, X. The Adaptor Protein CARD9 Protects against Colon Cancer by Restricting Mycobiota-Mediated Expansion of Myeloid-Derived Suppressor Cells. *Immunity.* 2018, **49**(3), pp.504-514.e504.
293. Sulżyc-Bielicka, V., Kołodziejczyk, L., Jaczewska, S., Bielicki, D., Safranow, K., Bielicki, P., Kładny, J. and Rogowski, W. Colorectal cancer and Cryptosporidium spp. infection. *PloS one.* 2018, **13**(4), pp.e0195834-e0195834.
294. Osman, M., Benamrouz, S., Guyot, K., Baydoun, M., Frealle, E., Chabe, M., Gantois, N., Delaire, B., Goffard, A., Aoun, A., Jurdi, N., Dabboussi, F., Even, G., Slomianny, C., Gosset, P., Hamze, M., Creusy, C., Viscogliosi, E. and Certad, G. High association of Cryptosporidium spp. infection with colon adenocarcinoma in Lebanese patients. *PloS one.* 2017, **12**(12), pp.e0189422-e0189422.
295. Toychiev, A., Abdujapparov, S., Imamov, A., Navruzov, B., Davis, N., Badalova, N. and Osipova, S. Intestinal helminths and protozoan infections in patients with colorectal cancer: prevalence and possible association with cancer pathogenesis. *Parasitol Res.* 2018, **117**(12), pp.3715-3723.
296. Mohamed, A.M., Ahmed, M.A., Ahmed, S.A., Al-Semany, S.A., Alghamdi, S.S. and Zaglool, D.A. Predominance and association risk of Blastocystis hominis subtype I in colorectal cancer: a case control study. *Infect Agent Cancer.* 2017, **12**, p.21.
297. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J.A., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalova, L., Pevsner-Fischer, M., Bikovsky, R., Halpern, Z., Elinav, E. and Segal, E. Personalized Nutrition by Prediction of Glycemic Responses. *Cell.* 2015, **163**(5), pp.1079-1094.
298. Appleyard, C.B., Cruz, M.L., Isidro, A.A., Arthur, J.C., Jobin, C. and de Simone, C. Pretreatment with the probiotic VSL#3 delays transition from inflammation to dysplasia in a rat model of colitis-associated cancer. *American Journal of Physiology - Gastrointestinal and Liver Physiology.* 2011, **301**(6), pp.G1004-G1013.
299. Gianotti, L., Morelli, L., Galbiati, F., Rocchetti, S., Coppola, S., Beneduce, A., Gilardini, C., Zonenschain, D., Nespoli, A. and Braga, M. A randomized double-blind trial on perioperative administration of

probiotics in colorectal cancer patients. *World journal of gastroenterology.* 2010, **16**(2), pp.167-175.
300. Hibberd, A.A., Lyra, A., Ouwehand, A.C., Rolny, P., Lindegren, H., Cedgard, L. and Wettergren, Y. Intestinal microbiota is altered in patients with colon cancer and modified by probiotic intervention. *BMJ Open Gastroenterology.* 2017, **4**(1), p.e000145.
301. Mendes, M.C.S., Paulino, D.S.M., Brambilla, S.R., Camargo, J.A., Persinoti, G.F. and Carvalheira, J.B.C. Microbiota modification by probiotic supplementation reduces colitis associated colon cancer in mice. *World journal of gastroenterology.* 2018, **24**(18), pp.1995-2008.
302. Rong, J., Liu, S., Hu, C. and Liu, C. Single probiotic supplement suppresses colitis-associated colorectal tumorigenesis by modulating inflammatory development and microbial homeostasis. *J Gastroenterol Hepatol.* 2019, **34**(7), pp.1182-1192.
303. Gao, C., Ganesh, B.P., Shi, Z., Shah, R.R., Fultz, R., Major, A., Venable, S., Lugo, M., Hoch, K., Chen, X., Haag, A., Wang, T.C. and Versalovic, J. Gut Microbe-Mediated Suppression of Inflammation-Associated Colon Carcinogenesis by Luminal Histamine Production. *Am J Pathol.* 2017, **187**(10), pp.2323-2336.
304. Saito, Y., Hinoi, T., Adachi, T., Miguchi, M., Niitsu, H., Kochi, M., Sada, H., Sotomaru, Y., Sakamoto, N., Sentani, K., Oue, N., Yasui, W., Tashiro, H. and Ohdan, H. Synbiotics suppress colitis-induced tumorigenesis in a colon-specific cancer mouse model. *PloS one.* 2019, **14**(6), p.e0216393.
305. Ho, C.L., Tan, H.Q., Chua, K.J., Kang, A., Lim, K.H., Ling, K.L., Yew, W.S., Lee, Y.S., Thiery, J.P. and Chang, M.W. Engineered commensal microbes for diet-mediated colorectal-cancer chemoprevention. *Nat Biomed Eng.* 2018, **2**(1), pp.27-37.
306. Zheng, D.W., Dong, X., Pan, P., Chen, K.W., Fan, J.X., Cheng, S.X. and Zhang, X.Z. Phage-guided modulation of the gut microbiota of mouse models of colorectal cancer augments their responses to chemotherapy. *Nat Biomed Eng.* 2019, p.[no pagination].
307. Kabwe, M., Brown, T.L., Dashper, S., Speirs, L., Ku, H., Petrovski, S., Chan, H.T., Lock, P. and Tucci, J. Genomic, morphological and functional characterisation of novel bacteriophage FNU1 capable of disrupting Fusobacterium nucleatum biofilms. *Sci Rep.* 2019, **9**(1), p.9107.
308. Vrieze, A., Van Nood, E., Holleman, F., Salojarvi, J., Kootte, R.S., Bartelsman, J.F., Dallinga-Thie, G.M., Ackermans, M.T., Serlie, M.J., Oozeer, R., Derrien, M., Druesne, A., Van Hylckama Vlieg, J.E., Bloks, V.W., Groen, A.K., Heilig, H.G., Zoetendal, E.G., Stroes, E.S., de Vos, W.M., Hoekstra, J.B. and Nieuwdorp, M. Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. *Gastroenterology.* 2012, **143**(4), pp.913-916.e917.
309. van Nood, E., Dijkgraaf, M.G. and Keller, J.J. Duodenal infusion of feces for recurrent Clostridium difficile. *N Engl J Med.* 2013, **368**(22), p.2145.
310. Lee, C.H., Steiner, T., Petrof, E.O., Smieja, M., Roscoe, D., Nematallah, A., Weese, J.S., Collins, S., Moayyedi, P., Crowther, M., Ropeleski, M.J., Jayaratne, P., Higgins, D., Li, Y., Rau, N.V. and Kim,

P.T. Frozen vs Fresh Fecal Microbiota Transplantation and Clinical Resolution of Diarrhea in Patients With Recurrent Clostridium difficile Infection: A Randomized Clinical TrialFrozen Fecal Microbiota Transplantation and C difficile InfectionFrozen Fecal Microbiota Transplantation and C difficile Infection. *JAMA.* 2016, **315**(2), pp.142-149.

311. Zuo, T., Wong, S.H., Lam, K., Lui, R., Cheung, K., Tang, W., Ching, J.Y.L., Chan, P.K.S., Chan, M.C.W., Wu, J.C.Y., Chan, F.K.L., Yu, J., Sung, J.J.Y. and Ng, S.C. Bacteriophage transfer during faecal microbiota transplantation in Clostridium difficile infection is associated with treatment outcome. *Gut.* 2018, **67**(4), p.634.
312. Ott, S.J., Waetzig, G.H., Rehman, A., Moltzau-Anderson, J., Bharti, R., Grasis, J.A., Cassidy, L., Tholey, A., Fickenscher, H., Seegert, D., Rosenstiel, P. and Schreiber, S. Efficacy of Sterile Fecal Filtrate Transfer for Treating Patients With Clostridium difficile Infection. *Gastroenterology.* 2017, **152**(4), pp.799-811.e797.
313. Guo, S.-H., Wang, H.-F., Nian, Z.-G., Wang, Y.-D., Zeng, Q.-Y. and Zhang, G. Immunization with alkyl hydroperoxide reductase subunit C reduces Fusobacterium nucleatum load in the intestinal tract. *Scientific reports.* 2017, **7**(1), pp.10566-10566.
314. Gopalakrishnan, V., Spencer, C.N., Nezi, L., Reuben, A., Andrews, M.C., Karpinets, T.V., Prieto, P.A., Vicente, D., Hoffman, K., Wei, S.C., Cogdill, A.P., Zhao, L., Hudgens, C.W., Hutchinson, D.S., Manzo, T., Petaccia de Macedo, M., Cotechini, T., Kumar, T., Chen, W.S., Reddy, S.M., Szczepaniak Sloane, R., Galloway-Pena, J., Jiang, H., Chen, P.L., Shpall, E.J., Rezvani, K., Alousi, A.M., Chemaly, R.F., Shelburne, S., Vence, L.M., Okhuysen, P.C., Jensen, V.B., Swennes, A.G., McAllister, F., Marcelo Riquelme Sanchez, E., Zhang, Y., Le Chatelier, E., Zitvogel, L., Pons, N., Austin-Breneman, J.L., Haydu, L.E., Burton, E.M., Gardner, J.M., Sirmans, E., Hu, J., Lazar, A.J., Tsujikawa, T., Diab, A., Tawbi, H., Glitza, I.C., Hwu, W.J., Patel, S.P., Woodman, S.E., Amaria, R.N., Davies, M.A., Gershenwald, J.E., Hwu, P., Lee, J.E., Zhang, J., Coussens, L.M., Cooper, Z.A., Futreal, P.A., Daniel, C.R., Ajami, N.J., Petrosino, J.F., Tetzlaff, M.T., Sharma, P., Allison, J.P., Jenq, R.R. and Wargo, J.A. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science.* 2018, **359**(6371), p.97.
315. Routy, B., Le Chatelier, E., Derosa, L., Duong, C.P.M., Alou, M.T., Daillère, R., Fluckiger, A., Messaoudene, M., Rauber, C., Roberti, M.P., Fidelle, M., Flament, C., Poirier-Colame, V., Opolon, P., Klein, C., Iribarren, K., Mondragón, L., Jacquelot, N., Qu, B., Ferrere, G., Clémenson, C., Mezquita, L., Masip, J.R., Naltet, C., Brosseau, S., Kaderbhai, C., Richard, C., Rizvi, H., Levenez, F., Galleron, N., Quinquis, B., Pons, N., Ryffel, B., Minard-Colin, V., Gonin, P., Soria, J.-C., Deutsch, E., Loriot, Y., Ghiringhelli, F., Zalcman, G., Goldwasser, F., Escudier, B., Hellmann, M.D., Eggermont, A., Raoult, D., Albiges, L., Kroemer, G. and Zitvogel, L. Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. *Science.* 2018, **359**(6371), p.91.
316. Matson, V., Fessler, J., Bao, R., Chongsuwat, T., Zha, Y., Alegre, M.-L., Luke, J.J. and Gajewski, T.F. The commensal microbiome is

associated with anti–PD-1 efficacy in metastatic melanoma patients. *Science.* 2018, **359**(6371), p.104.

317. Yuan, L., Zhang, S., Li, H., Yang, F., Mushtaq, N., Ullah, S., Shi, Y., An, C. and Xu, J. The influence of gut microbiota dysbiosis to the efficacy of 5-Fluorouracil treatment on colorectal cancer. *Biomedicine and Pharmacotherapy.* 2018, **108**, pp.184-193.
318. Geller, L.T., Barzily-Rokni, M., Danino, T., Jonas, O.H., Shental, N., Nejman, D., Gavert, N., Zwang, Y., Cooper, Z.A., Shee, K., Thaiss, C.A., Reuben, A., Livny, J., Avraham, R., Frederick, D.T., Ligorio, M., Chatman, K., Johnston, S.E., Mosher, C.M., Brandis, A., Fuks, G., Gurbatri, C., Gopalakrishnan, V., Kim, M., Hurd, M.W., Katz, M., Fleming, J., Maitra, A., Smith, D.A., Skalak, M., Bu, J., Michaud, M., Trauger, S.A., Barshack, I., Golan, T., Sandbank, J., Flaherty, K.T., Mandinova, A., Garrett, W.S., Thayer, S.P., Ferrone, C.R., Huttenhower, C., Bhatia, S.N., Gevers, D., Wargo, J.A., Golub, T.R. and Straussman, R. Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science.* 2017, **357**(6356), p.1156.
319. Chang, C.W., Liu, C.Y., Lee, H.C., Huang, Y.H., Li, L.H., Chiau, J.C., Wang, T.E., Chu, C.H., Shih, S.C., Tsai, T.H. and Chen, Y.J. Lactobacillus casei Variety rhamnosus probiotic preventively attenuates 5-Fluorouracil/Oxaliplatin-induced intestinal injury in a syngeneic colorectal cancer model. *Frontiers in Microbiology.* 2018, **9**(MAY), p.983.
320. Wallace, B.D., Wang, H., Lane, K.T., Scott, J.E., Orans, J., Koo, J.S., Venkatesh, M., Jobin, C., Yeh, L.-A., Mani, S. and Redinbo, M.R. Alleviating Cancer Drug Toxicity by Inhibiting a Bacterial Enzyme. *Science.* 2010, **330**(6005), p.831.
321. Sougiannis, A.T., VanderVeen, B.N., Enos, R.T., Velazquez, K.T., Bader, J.E., Carson, M., Chatzistamou, I., Walla, M., Pena, M.M., Kubinak, J.L., Nagarkatti, M., Carson, J.A. and Murphy, E.A. Impact of 5 fluorouracil chemotherapy on gut inflammation, functional parameters, and gut microbiota. *Brain Behav Immun.* 2019, **80**, pp.44-55.
322. Polakowski, C.B., Kato, M., Preti, V.B., Schieferdecker, M.E.M. and Ligocki Campos, A.C. Impact of the preoperative use of synbiotics in colorectal cancer patients: A prospective, randomized, double-blind, placebo-controlled study. *Nutrition.* 2019, **58**, pp.40-46.
323. Kostic, A.D., Howitt, M.R. and Garrett, W.S. Exploring host-microbiota interactions in animal models and humans. *Genes Dev.* 2013, **27**(7), pp.701-718.
324. Nguyen, T.L.A., Vieira-Silva, S., Liston, A. and Raes, J. How informative is the mouse for human gut microbiota research? *Disease Models & Mechanisms.* 2015, **8**(1), p.1.
325. Arrieta, M.-C., Walter, J. and Finlay, B.B. Human Microbiota-Associated Mice: A Model with Challenges. *Cell host & microbe.* 2016, **19**(5), pp.575-578.
326. Hugenholtz, F. and de Vos, W.M. Mouse models for human intestinal microbiota research: a critical evaluation. *Cellular and molecular life sciences : CMLS.* 2018, **75**(1), pp.149-160.

327. Parker, K.D., Albeke, S.E., Gigley, J.P., Goldstein, A.M. and Ward, N.L. Microbiome Composition in Both Wild-Type and Disease Model Mice Is Heavily Influenced by Mouse Facility. *Frontiers in Microbiology.* 2018, **9**(1598), p.[no pagination].
328. Ericsson, A.C., Gagliardi, J., Bouhan, D., Spollen, W.G., Givan, S.A. and Franklin, C.L. The influence of caging, bedding, and diet on the composition of the microbiota in different regions of the mouse gut. *Scientific reports.* 2018, **8**(1), p.4065.
329. Liu, L., Firrman, J., Tanes, C., Bittinger, K., Thomas-Gahring, A., Wu, G.D., Van den Abbeele, P. and Tomasula, P.M. Establishing a mucosal gut microbial community in vitro using an artificial simulator. *PloS one.* 2018, **13**(7), p.e0197692.
330. Van de Wiele, T., Van den Abbeele, P., Ossieur, W., Possemiers, S. and Marzorati, M. The Simulator of the Human Intestinal Microbial Ecosystem (SHIME((R))). In: Verhoeckx, K. et al. eds. *The Impact of Food Bioactives on Health: in vitro and ex vivo models*. Cham (CH): Springer Copyright 2015, The Author(s). 2015, pp.305-317.
331. Kim, H.J. and Ingber, D.E. Gut-on-a-Chip microenvironment induces human intestinal cells to undergo villus differentiation. *Integr Biol (Camb).* 2013, **5**(9), pp.1130-1140.
332. Kim, H.J., Huh, D., Hamilton, G. and Ingber, D.E. Human gut-on-a-chip inhabited by microbial flora that experiences intestinal peristalsis-like motions and flow. *Lab Chip.* 2012, **12**(12), pp.2165-2174.
333. Kim, H.J., Li, H., Collins, J.J. and Ingber, D.E. Contributions of microbiome and mechanical deformation to intestinal bacterial overgrowth and inflammation in a human gut-on-a-chip. *Proc Natl Acad Sci U S A.* 2016, **113**(1), pp.E7-15.
334. Kim, H.J., Lee, J., Choi, J.H., Bahinski, A. and Ingber, D.E. Co-culture of Living Microbiome with Microengineered Human Intestinal Villi in a Gut-on-a-Chip Microfluidic Device. *J Vis Exp.* 2016, (114), p.[no pagination].
335. Sato, T., Stange, D.E., Ferrante, M., Vries, R.G., Van Es, J.H., Van den Brink, S., Van Houdt, W.J., Pronk, A., Van Gorp, J., Siersema, P.D. and Clevers, H. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology.* 2011, **141**(5), pp.1762-1772.
336. Williamson, I.A., Arnold, J.W., Samsa, L.A., Gaynor, L., DiSalvo, M., Cocchiaro, J.L., Carroll, I., Azcarate-Peril, M.A., Rawls, J.F., Allbritton, N.L. and Magness, S.T. A High-Throughput Organoid Microinjection Platform to Study Gastrointestinal Microbiota and Luminal Physiology. *Cell Mol Gastroenterol Hepatol.* 2018, **6**(3), pp.301-319.
337. Blutt, S.E., Crawford, S.E., Ramani, S., Zou, W.Y. and Estes, M.K. Engineered Human Gastrointestinal Cultures to Study the Microbiome and Infectious Diseases. *Cellular and Molecular Gastroenterology and Hepatology.* 2018, **5**(3), pp.241-251.
338. Johansson, M.E., Larsson, J.M. and Hansson, G.C. The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. *Proc Natl Acad Sci U S A.* 2011, **108 Suppl 1**, pp.4659-4665.
339. Watt, E., Gemmell, M.R., Berry, S., Glaire, M., Farquharson, F., Louis, P., Murray, G.I., El-Omar, E. and Hold, G.L. Extending colonic mucosal

microbiome analysis—assessment of colonic lavage as a proxy for endoscopic colonic biopsies. *Microbiome.* 2016, **4**(1), p.61.

340. Lyra, A., Forssten, S., Rolny, P., Wettergren, Y., Lahtinen, S.J., Salli, K., Cedgård, L., Odin, E., Gustavsson, B. and Ouwehand, A.C. Comparison of bacterial quantities in left and right colon biopsies and faeces. *World journal of gastroenterology.* 2012, **18**(32), pp.4404-4411.
341. Durbán, A., Abellán, J.J., Jiménez-Hernández, N., Ponce, M., Ponce, J., Sala, T., D'Auria, G., Latorre, A. and Moya, A. Assessing Gut Microbial Diversity from Feces and Rectal Mucosa. *Microbial ecology.* 2011, **61**(1), pp.123-133.
342. Stearns, J.C., Lynch, M.D., Senadheera, D.B., Tenenbaum, H.C., Goldberg, M.B., Cvitkovitch, D.G., Croitoru, K., Moreno-Hagelsieb, G. and Neufeld, J.D. Bacterial biogeography of the human digestive tract. *Sci Rep.* 2011, **1**, p.170.
343. Zoetendal, E.G., von Wright, A., Vilpponen-Salmela, T., Ben-Amor, K., Akkermans, A.D. and de Vos, W.M. Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces. *Appl Environ Microbiol.* 2002, **68**(7), pp.3401-3407.
344. Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E. and Relman, D.A. Diversity of the human intestinal microbial flora. *Science.* 2005, **308**(5728), pp.1635-1638.
345. Ringel, Y., Maharshak, N., Ringel-Kulka, T., Wolber, E.A., Sartor, R.B. and Carroll, I.M. High throughput sequencing reveals distinct microbial populations within the mucosal and luminal niches in healthy individuals. *Gut microbes.* 2015, **6**(3), pp.173-181.
346. Albenberg, L., Esipova, T.V., Judge, C.P., Bittinger, K., Chen, J., Laughlin, A., Grunberg, S., Baldassano, R.N., Lewis, J.D., Li, H., Thom, S.R., Bushman, F.D., Vinogradov, S.A. and Wu, G.D. Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota. *Gastroenterology.* 2014, **147**(5), pp.1055-1063.e1058.
347. Flynn, K.J., Ruffin, M.T.t., Turgeon, D.K. and Schloss, P.D. Spatial Variation of the Native Colon Microbiota in Healthy Adults. *Cancer Prev Res (Phila).* 2018, **11**(7), pp.393-402.
348. Shah, M.S., DeSantis, T., Yamal, J.-M., Weir, T., Ryan, E.P., Cope, J.L. and Hollister, E.B. Re-purposing 16S rRNA gene sequence data from within case paired tumor biopsy and tumor-adjacent biopsy or fecal samples to identify microbial markers for colorectal cancer. *PloS one.* 2018, **13**(11), p.e0207002.
349. Shobar, R.M., Velineni, S., Keshavarzian, A., Swanson, G., DeMeo, M.T., Melson, J.E., Losurdo, J., Engen, P.A., Sun, Y., Koenig, L. and Mutlu, E.A. The Effects of Bowel Preparation on Microbiota-Related Metrics Differ in Health and in Inflammatory Bowel Disease and for the Mucosal and Luminal Microbiota Compartments. *Clin Transl Gastroenterol.* 2016, **7**, p.e143.
350. Araujo-Perez, F., McCoy, A.N., Okechukwu, C., Carroll, I.M., Smith, K.M., Jeremiah, K., Sandler, R.S., Asher, G.N. and Keku, T.O. Differences in microbial signatures between rectal mucosal biopsies and rectal swabs. *Gut microbes.* 2012, **3**(6), pp.530-535.

351. Aguirre de Carcer, D., Cuiv, P.O., Wang, T., Kang, S., Worthley, D., Whitehall, V., Gordon, I., McSweeney, C., Leggett, B. and Morrison, M. Numerical ecology validates a biogeographical distribution and gender-based effect on mucosa-associated bacteria along the human colon. *Isme j.* 2011, **5**(5), pp.801-809.
352. Zhang, Z., Geng, J., Tang, X., Fan, H., Xu, J., Wen, X., Ma, Z.S. and Shi, P. Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *Isme j.* 2014, **8**(4), pp.881-893.
353. Hong, P.Y., Croix, J.A., Greenberg, E., Gaskins, H.R. and Mackie, R.I. Pyrosequencing-based analysis of the mucosal microbiota in healthy individuals reveals ubiquitous bacterial groups and micro-heterogeneity. *PloS one.* 2011, **6**(9), p.e25042.
354. Smolinska, A., Tedjo, D.I., Blanchet, L., Bodelier, A., Pierik, M.J., Masclee, A.A.M., Dallinga, J., Savelkoul, P.H.M., Jonkers, D., Penders, J. and van Schooten, F.J. Volatile metabolites in breath strongly correlate with gut microbiome in CD patients. *Anal Chim Acta.* 2018, **1025**, pp.1-11.
355. Liesenfeld, D.B., Habermann, N., Toth, R., Owen, R.W., Frei, E., Bohm, J., Schrotz-King, P., Klika, K.D. and Ulrich, C.M. Changes in urinary metabolic profiles of colorectal cancer patients enrolled in a prospective cohort study (ColoCare). *Metabolomics.* 2015, **11**(4), pp.998-1012.
356. Mozdiak, E., Wicaksono, A.N., Covington, J.A. and Arasaradnam, R.P. Colorectal cancer and adenoma screening using urinary volatile organic compound (VOC) detection: early results from a single-centre bowel screening population (UK BCSP). *Techniques in coloproctology.* 2019, **23**(4), pp.343-351.
357. Wilmanski, T., Rappaport, N., Earls, J.C., Magis, A.T., Manor, O., Lovejoy, J., Omenn, G.S., Hood, L., Gibbons, S. and Price, N.D. Blood metabolome signature predicts gut microbiome α-diversity in health and disease. *bioRxiv.* 2019, p.561209.
358. Org, E., Blum, Y., Kasela, S., Mehrabian, M., Kuusisto, J., Kangas, A.J., Soininen, P., Wang, Z., Ala-Korpela, M., Hazen, S.L., Laakso, M. and Lusis, A.J. Relationships between gut microbiota, plasma metabolites, and metabolic syndrome traits in the METSIM cohort. *Genome biology.* 2017, **18**(1), p.70.
359. Pace, N.R. A molecular view of microbial diversity and the biosphere. *Science.* 1997, **276**(5313), pp.734-740.
360. Langendijk, P.S., Schut, F., Jansen, G.J., Raangs, G.C., Kamphuis, G.R., Wilkinson, M.H. and Welling, G.W. Quantitative fluorescence in situ hybridization of Bifidobacterium spp. with genus-specific 16S rRNA-targeted probes and its application in fecal samples. *Applied and environmental microbiology.* 1995, **61**(8), pp.3069-3075.
361. Suau, A., Bonnet, R., Sutren, M., Godon, J.J., Gibson, G.R., Collins, M.D. and Dore, J. Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Environ Microbiol.* 1999, **65**(11), pp.4799-4807.
362. Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R.S., Magrum, L.J., Zablen, L.B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B.J., Stahl,

D.A., Luehrsen, K.R., Chen, K.N. and Woese, C.R. The phylogeny of prokaryotes. *Science.* 1980, **209**(4455), p.457.
363. Woese, C.R. Bacterial evolution. *Microbiological reviews.* 1987, **51**(2), pp.221-271.
364. Woese, C.R., Kandler, O. and Wheelis, M.L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A.* 1990, **87**(12), pp.4576-4579.
365. Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L. and Pace, N.R. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences of the United States of America.* 1985, **82**(20), pp.6955-6959.
366. Wang, Q., Garrity, G.M., Tiedje, J.M. and Cole, J.R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol.* 2007, **73**(16), pp.5261-5267.
367. Cai, L., Ye, L., Tong, A.H.Y., Lok, S. and Zhang, T. Biased Diversity Metrics Revealed by Bacterial 16S Pyrotags Derived from Different Primer Sets. *PloS one.* 2013, **8**(1), p.e53649.
368. Engelbrektson, A., Kunin, V., Wrighton, K.C., Zvenigorodsky, N., Chen, F., Ochman, H. and Hugenholtz, P. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *Isme j.* 2010, **4**(5), pp.642-647.
369. Liu, Z., Lozupone, C., Hamady, M., Bushman, F.D. and Knight, R. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res.* 2007, **35**(18), p.e120.
370. Liu, Z., DeSantis, T.Z., Andersen, G.L. and Knight, R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* 2008, **36**(18), p.e120.
371. Baker, G.C., Smith, J.J. and Cowan, D.A. Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods.* 2003, **55**(3), pp.541-555.
372. Chakravorty, S., Helb, D., Burday, M., Connell, N. and Alland, D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods.* 2007, **69**(2), pp.330-339.
373. Wang, J., Kurilshikov, A., Radjabzadeh, D., Turpin, W., Croitoru, K., Bonder, M.J., Jackson, M.A., Medina-Gomez, C., Frost, F., Homuth, G., Rühlemann, M., Hughes, D., Kim, H.-n., Ahluwalia, T., Barkan, E., Bedrani, L., Bell, J., Bisgaard, H., Boehnke, M., Bonder, M.J., Bønnelykke, K., Boomsma, D.I., Croitoru, K., Davies, G.E., de Geus, E., Degenhardt, F., D'Amato, M., Ehli, E.A., Espin-Garcia, O., Finnicum, C.T., Fornage, M., Franke, A., Franke, L., Frost, F., Fu, J., Heinsen, F.-A., Homuth, G., Hughes, D., Ijzerman, R., Jackson, M.A., Jessen, L.E., Jonkers, D., Kacprowski, T., Kim, H.-N., Kim, H.-L., Kraaij, R., Kurilshikov, A., Laakso, M., Launer, L., Lerch, M.M., Lüll, K., Lusis, A.J., Mangino, M., Mayerle, J., Mbarek, H., Medina, M.C., Meyer, K., Mohlke, K.L., Org, E., Paterson, A., Payami, H., Radjabzadeh, D., Raes, J., Rothschild, D., Rühlemann, M., Sanna, S., Segal, E., Shah, S., Smith, M., Spector, T., Steves, C., Stokholm, J., Szopinska, J.W., Thorsen, J., Timpson, N., Turpin, W., Uitterlinden, A.G., Vasquez, A.A., Völzke, H., Vosa, U., Wallen, Z., Wang, J., Weiss, F.U., Weissbrod, O.,

Wijmenga, C., Willemsen, G., Xu, W., Yun, Y., Zhernakova, A., Spector, T.D., Bell, J.T., Steves, C.J., Timpson, N., Franke, A., Wijmenga, C., Meyer, K., Kacprowski, T., Franke, L., Paterson, A.D., Raes, J., Kraaij, R., Zhernakova, A. and MiBioGen Consortium, I. Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative. *Microbiome.* 2018, **6**(1), p.101.

374. Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G. and Huttenhower, C. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature biotechnology.* 2013, **31**, p.814.
375. Kostic, A.D., Ojesina, A.I., Pedamallu, C.S., Jung, J., Verhaak, R.G.W., Getz, G. and Meyerson, M. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nature biotechnology.* 2011, **29**, p.393.
376. Imrit, K., Goldfischer, M., Wang, J., Green, J., Levine, J., Lombardo, J. and Hong, T. Identification of Bacteria in Formalin-Fixed, Paraffin-Embedded Heart Valve Tissue via 16S rRNA Gene Nucleotide Sequencing. *Journal of clinical microbiology.* 2006, **44**(7), p.2609.
377. Xuan, C., Shamonki, J.M., Chung, A., DiNome, M.L., Chung, M., Sieling, P.A. and Lee, D.J. Microbial Dysbiosis Is Associated with Human Breast Cancer. *PloS one.* 2014, **9**(1), p.e83744.
378. Petersen, T.N., Lukjancenko, O., Thomsen, M.C.F., Maddalena Sperotto, M., Lund, O., Møller Aarestrup, F. and Sicheritz-Pontén, T. MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads. *PloS one.* 2017, **12**(5), p.e0176469.
379. Guo, M., Xu, E. and Ai, D. Inferring Bacterial Infiltration in Primary Colorectal Tumors From Host Whole Genome Sequencing Data. *Front Genet.* 2019, **10**, p.213.
380. Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Zech Xu, Z., Kightley, E.P., Thompson, L.R., Hyde, E.R., Gonzalez, A. and Knight, R. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems.* 2017, **2**(2), pp.e00191-00116.
381. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M.C., Rice, B.L., DuLong, C., Morgan, X.C., Golden, C.D., Quince, C., Huttenhower, C. and Segata, N. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell.* 2019, **176**(3), pp.649-662.e620.
382. Rappe, M.S. and Giovannoni, S.J. The uncultured microbial majority. *Annu Rev Microbiol.* 2003, **57**, pp.369-394.
383. Shin, J., Lee, S., Go, M.-J., Lee, S.Y., Kim, S.C., Lee, C.-H. and Cho, B.-K. Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Scientific reports.* 2016, **6**, p.29681.
384. Moss, E.L. and Bhatt, A.S. Generating closed bacterial genomes from long-read nanopore sequencing of microbiomes. *bioRxiv.* 2018, p.489641.
385. Lozupone, C.A. and Knight, R. Species divergence and the measurement of microbial diversity. *FEMS Microbiology Reviews.* 2008, **32**(4), pp.557-578.

386. Lozupone, C. and Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol.* 2005, **71**(12), pp.8228-8235.
387. Lozupone, C.A., Hamady, M., Kelley, S.T. and Knight, R. Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Applied and environmental microbiology.* 2007, **73**(5), pp.1576-1585.
388. Layeghifard, M., Hwang, D.M. and Guttman, D.S. Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in microbiology.* 2017, **25**(3), pp.217-228.
389. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S. and Huttenhower, C. Metagenomic biomarker discovery and explanation. *Genome biology.* 2011, **12**(6), pp.R60-R60.
390. Petriz, B.A. and Franco, O.L. Metaproteomics as a Complementary Approach to Gut Microbiota in Health and Disease. *Frontiers in chemistry.* 2017, **5**(4), p.[no pagination].
391. Zierer, J., Jackson, M.A., Kastenmüller, G., Mangino, M., Long, T., Telenti, A., Mohney, R.P., Small, K.S., Bell, J.T., Steves, C.J., Valdes, A.M., Spector, T.D. and Menni, C. The fecal metabolome as a functional readout of the gut microbiome. *Nature Genetics.* 2018, **50**(6), pp.790-795.
392. Vernocchi, P., Del Chierico, F. and Putignani, L. Gut Microbiota Profiling: Metabolomics Based Approach to Unravel Compounds Affecting Human Health. *Frontiers in Microbiology.* 2016, **7**(1144), p.[no pagination].
393. Paliy, O. and Agans, R. Application of phylogenetic microarrays to interrogation of human microbiota. *FEMS microbiology ecology.* 2012, **79**(1), pp.2-11.
394. Rieger, J., Janczyk, P., Hünigen, H. and Plendl, J. Enhancement of immunohistochemical detection of Salmonella in tissues of experimentally infected pigs. *European journal of histochemistry : EJH.* 2015, **59**(3), pp.2516-2516.
395. Wu, G.D., Lewis, J.D., Hoffmann, C., Chen, Y.Y., Knight, R., Bittinger, K., Hwang, J., Chen, J., Berkowsky, R., Nessel, L., Li, H. and Bushman, F.D. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol.* 2010, **10**, p.206.
396. Sinha, R., Chen, J., Amir, A., Vogtmann, E., Shi, J., Inman, K.S., Flores, R., Sampson, J., Knight, R. and Chia, N. Collecting Fecal Samples for Microbiome Analyses in Epidemiology Studies. *Cancer Epidemiol Biomarkers Prev.* 2016, **25**(2), pp.407-416.
397. Couch, R.D., Navarro, K., Sikaroodi, M., Gillevet, P., Forsyth, C.B., Mutlu, E., Engen, P.A. and Keshavarzian, A. The Approach to Sample Acquisition and Its Impact on the Derived Human Fecal Microbiome and VOC Metabolome. *PloS one.* 2013, **8**(11), p.e81163.
398. Mathay, C., Hamot, G., Henry, E., Georges, L., Bellora, C., Lebrun, L., de Witt, B., Ammerlaan, W., Buschart, A., Wilmes, P. and Betsou, F. Method Optimization for Fecal Sample Collection and Fecal DNA Extraction. *Biopreservation and Biobanking.* 2015, **13**(2), pp.79-93.

399. Sinha, R., Abnet, C.C., White, O., Knight, R. and Huttenhower, C. The microbiome quality control project: baseline study design and future directions. *Genome biology.* 2015, **16**, pp.276-276.
400. Fu, B.C., Randolph, T.W., Lim, U., Monroe, K.R., Cheng, I., Wilkens, L.R., Le Marchand, L., Hullar, M.A.J. and Lampe, J.W. Characterization of the gut microbiome in epidemiologic studies: the multiethnic cohort experience. *Annals of epidemiology.* 2016, **26**(5), pp.373-379.
401. Sze, M.A. and Schloss, P.D. The Impact of DNA Polymerase and Number of Rounds of Amplification in PCR on 16S rRNA Gene Sequence Data. *mSphere.* 2019, **4**(3), p.[no pagination].
402. *International Human Microbiome Standards.* [Online]. [Accessed 11.2.19]. Available from: http://www.microbiome-standards.org
403. *Earth Microbiome Project.* [Online]. [Accessed 11.2.19]. Available from: http://www.earthmicrobiome.org
404. Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G., Navas-Molina, J.A., Janssen, S., Kopylova, E., Vázquez-Baeza, Y., González, A., Morton, J.T., Mirarab, S., Zech Xu, Z., Jiang, L., Haroon, M.F., Kanbar, J., Zhu, Q., Jin Song, S., Kosciolek, T., Bokulich, N.A., Lefler, J., Brislawn, C.J., Humphrey, G., Owens, S.M., Hampton-Marcell, J., Berg-Lyons, D., McKenzie, V., Fierer, N., Fuhrman, J.A., Clauset, A., Stevens, R.L., Shade, A., Pollard, K.S., Goodwin, K.D., Jansson, J.K., Gilbert, J.A., Knight, R., The Earth Microbiome Project, C., Rivera, J.L.A., Al-Moosawi, L., Alverdy, J., Amato, K.R., Andras, J., Angenent, L.T., Antonopoulos, D.A., Apprill, A., Armitage, D., Ballantine, K., Bárta, J.í., Baum, J.K., Berry, A., Bhatnagar, A., Bhatnagar, M., Biddle, J.F., Bittner, L., Boldgiv, B., Bottos, E., Boyer, D.M., Braun, J., Brazelton, W., Brearley, F.Q., Campbell, A.H., Caporaso, J.G., Cardona, C., Carroll, J., Cary, S.C., Casper, B.B., Charles, T.C., Chu, H., Claar, D.C., Clark, R.G., Clayton, J.B., Clemente, J.C., Cochran, A., Coleman, M.L., Collins, G., Colwell, R.R., Contreras, M., Crary, B.B., Creer, S., Cristol, D.A., Crump, B.C., Cui, D., Daly, S.E., Davalos, L., Dawson, R.D., Defazio, J., Delsuc, F., Dionisi, H.M., Dominguez-Bello, M.G., Dowell, R., Dubinsky, E.A., Dunn, P.O., Ercolini, D., Espinoza, R.E., Ezenwa, V., Fenner, N., Findlay, H.S., Fleming, I.D., Fogliano, V., Forsman, A., Freeman, C., Friedman, E.S., Galindo, G., Garcia, L., Garcia-Amado, M.A., Garshelis, D., Gasser, R.B., Gerdts, G., Gibson, M.K., Gifford, I., Gill, R.T., Giray, T., Gittel, A., Golyshin, P., Gong, D., Grossart, H.-P., Guyton, K., Haig, S.-J., Hale, V., Hall, R.S., Hallam, S.J., Handley, K.M., Hasan, N.A., Haydon, S.R., Hickman, J.E., Hidalgo, G., Hofmockel, K.S., Hooker, J., Hulth, S., Hultman, J., Hyde, E., Ibáñez-Álamo, J.D., Jastrow, J.D., Jex, A.R., Johnson, L.S., Johnston, E.R., Joseph, S., Jurburg, S.D., Jurelevicius, D., Karlsson, A., Karlsson, R., Kauppinen, S., Kellogg, C.T.E., Kennedy, S.J., Kerkhof, L.J., King, G.M., Kling, G.W., Koehler, A.V., Krezalek, M., Kueneman, J., Lamendella, R., Landon, E.M., Lane-deGraaf, K., LaRoche, J., Larsen, P., Laverock, B., Lax, S., Lentino, M., Levin, I.I., Liancourt, P., Liang, W., Linz, A.M., Lipson, D.A., Liu, Y., Lladser, M.E., Lozada, M., Spirito, C.M., MacCormack, W.P., MacRae-Crerar, A., Magris, M., Martín-Platero, A.M., Martín-Vivaldi, M., Martínez, L.M., Martínez-Bueno, M.,

Marzinelli, E.M., Mason, O.U., Mayer, G.D., McDevitt-Irwin, J.M., McDonald, J.E., McGuire, K.L., McMahon, K.D., McMinds, R., Medina, M., Mendelson, J.R., Metcalf, J.L., Meyer, F., Michelangeli, F., Miller, K., Mills, D.A., Minich, J., Mocali, S., Moitinho-Silva, L., Moore, A., Morgan-Kiss, R.M., Munroe, P., Myrold, D., Neufeld, J.D., Ni, Y., Nicol, G.W., Nielsen, S., Nissimov, J.I., Niu, K., Nolan, M.J., Noyce, K., O'Brien, S.L., Okamoto, N., Orlando, L., Castellano, Y.O., Osuolale, O., Oswald, W., Parnell, J., Peralta-Sánchez, J.M., Petraitis, P., Pfister, C., Pilon-Smits, E., Piombino, P., Pointing, S.B., Pollock, F.J., Potter, C., Prithiviraj, B., Quince, C., Rani, A., Ranjan, R., Rao, S., Rees, A.P., Richardson, M., Riebesell, U., Robinson, C., Rockne, K.J., Rodriguezl, S.M., Rohwer, F., Roundstone, W., Safran, R.J., Sangwan, N., Sanz, V., Schrenk, M., Schrenzel, M.D., Scott, N.M., Seger, R.L., Seguin-Orlando, A., Seldin, L., Seyler, L.M., Shakhsheer, B., Sheets, G.M., Shen, C., Shi, Y., Shin, H., Shogan, B.D., Shutler, D., Siegel, J., Simmons, S., Sjöling, S., Smith, D.P., Soler, J.J., Sperling, M., Steinberg, P.D., Stephens, B., Stevens, M.A., Taghavi, S., Tai, V., Tait, K., Tan, C.L., Tas¸, N., Taylor, D.L., Thomas, T., Timling, I., Turner, B.L., Urich, T., Ursell, L.K., van der Lelie, D., Van Treuren, W., van Zwieten, L., Vargas-Robles, D., Thurber, R.V., Vitaglione, P., Walker, D.A., Walters, W.A., Wang, S., Wang, T., Weaver, T., Webster, N.S., Wehrle, B., Weisenhorn, P., Weiss, S., Werner, J.J., West, K., Whitehead, A., Whitehead, S.R., Whittingham, L.A., Willerslev, E., Williams, A.E., Wood, S.A., Woodhams, D.C., Yang, Y., Zaneveld, J., Zarraonaindia, I., Zhang, Q. and Zhao, H. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature.* 2017, **551**, p.457.

405. Gilbert, J.A., Jansson, J.K. and Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biology.* 2014, **12**(1), p.69.

406. Gilbert, J.A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C.T., Brown, C.T., Desai, N., Eisen, J.A., Evers, D., Field, D., Feng, W., Huson, D., Jansson, J., Knight, R., Knight, J., Kolker, E., Konstantindis, K., Kostka, J., Kyrpides, N., Mackelprang, R., McHardy, A., Quince, C., Raes, J., Sczyrba, A., Shade, A. and Stevens, R. Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Stand Genomic Sci.* 2010, **3**(3), pp.243-248.

407. Sinha, R., Goedert, J.J., Vogtmann, E., Hua, X., Porras, C., Hayes, R., Safaeian, M., Yu, G., Sampson, J., Ahn, J. and Shi, J. Quantification of Human Microbiome Stability over 6 Months: Implications for Epidemiologic Studies. *American journal of epidemiology.* 2018, **187**(6), pp.1282-1290.

408. *International Human Microbiome Consortium.* [Online]. [Accessed 10.2.19]. Available from: http://www.human-microbiome.org

409. *Human Microbiome Project.* [Online]. [Accessed 11.2.19]. Available from: https://hmpdacc.org

410. *MetaHIT Final Report Summary.* [Online]. [Accessed 10.02.19]. Available from: https://cordis.europa.eu/project/rcn/87834/reporting/en

411. *MetaHIT project.* [Online]. [Accessed 10.2.19]. Available from: http://www.metahit.eu

412. *British Gut Project.* [Online]. [Accessed 11.2.19]. Available from: http://britishgut.org

413. McDonald, D., Hyde, E., Debelius, J.W., Morton, J.T., Gonzalez, A., Ackermann, G., Aksenov, A.A., Behsaz, B., Brennan, C., Chen, Y., DeRight Goldasich, L., Dorrestein, P.C., Dunn, R.R., Fahimipour, A.K., Gaffney, J., Gilbert, J.A., Gogul, G., Green, J.L., Hugenholtz, P., Humphrey, G., Huttenhower, C., Jackson, M.A., Janssen, S., Jeste, D.V., Jiang, L., Kelley, S.T., Knights, D., Kosciolek, T., Ladau, J., Leach, J., Marotz, C., Meleshko, D., Melnik, A.V., Metcalf, J.L., Mohimani, H., Montassier, E., Navas-Molina, J., Nguyen, T.T., Peddada, S., Pevzner, P., Pollard, K.S., Rahnavard, G., Robbins-Pianka, A., Sangwan, N., Shorenstein, J., Smarr, L., Song, S.J., Spector, T., Swafford, A.D., Thackray, V.G., Thompson, L.R., Tripathi, A., Vázquez-Baeza, Y., Vrbanac, A., Wischmeyer, P., Wolfe, E., Zhu, Q. and Knight, R. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems.* 2018, **3**(3), pp.e00031-00018.
414. Malcomson, F.C., Breininger, S.P., ElGendy, K., Joel, A., Ranathunga, R., Hill, T.R., Bradburn, D.M., Turnbull, D.M., Greaves, L.C. and Mathers, J.C. Design and baseline characteristics of the Biomarkers Of Risk In Colorectal Cancer (BORICC) Follow-Up study: A 12+ years follow-up. *Nutr Health.* 2019, p.260106019866963.
415. Flores, R., Shi, J., Gail, M.H., Gajer, P., Ravel, J. and Goedert, J.J. Assessment of the human faecal microbiota: II. Reproducibility and associations of 16S rRNA pyrosequences. *Eur J Clin Invest.* 2012, **42**(8), pp.855-863.
416. Gorzelak, M.A., Gill, S.K., Tasnim, N., Ahmadi-Vand, Z., Jay, M. and Gibson, D.L. Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool. *PloS one.* 2015, **10**(8), p.e0134802.
417. Voigt, A.Y., Costea, P.I., Kultima, J.R., Li, S.S., Zeller, G., Sunagawa, S. and Bork, P. Temporal and technical variability of human gut metagenomes. *Genome biology.* 2015, **16**(1), p.73.
418. Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I. and Knight, R. Bacterial community variation in human body habitats across space and time. *Science.* 2009, **326**(5960), pp.1694-1697.
419. David, L.A., Materna, A.C., Friedman, J., Campos-Baptista, M.I., Blackburn, M.C., Perrotta, A., Erdman, S.E. and Alm, E.J. Host lifestyle affects human microbiota on daily timescales. *Genome biology.* 2014, **15**(7), pp.R89-R89.
420. Wang, Z., Zolnik, C.P., Qiu, Y., Usyk, M., Wang, T., Strickler, H.D., Isasi, C.R., Kaplan, R.C., Kurland, I.J., Qi, Q. and Burk, R.D. Comparison of Fecal Collection Methods for Microbiome and Metabolomics Studies. *Front Cell Infect Microbiol.* 2018, **8**, p.301.
421. Dominianni, C., Wu, J., Hayes, R.B. and Ahn, J. Comparison of methods for fecal microbiome biospecimen collection. *BMC Microbiol.* 2014, **14**, p.103.
422. Carroll, I.M., Ringel-Kulka, T., Siddle, J.P., Klaenhammer, T.R. and Ringel, Y. Characterization of the fecal microbiota using high-throughput sequencing reveals a stable microbial community during storage. *PloS one.* 2012, **7**(10), p.e46953.
423. Hang, J., Desai, V., Zavaljevski, N., Yang, Y., Lin, X., Satya, R.V., Martinez, L.J., Blaylock, J.M., Jarman, R.G., Thomas, S.J. and Kuschner, R.A. 16S rRNA gene pyrosequencing of reference and

clinical samples and investigation of the temperature stability of microbiome profiles. *Microbiome.* 2014, **2**, p.31.

424. Bahl, M.I., Bergstrom, A. and Licht, T.R. Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. *FEMS Microbiol Lett.* 2012, **329**(2), pp.193-197.
425. Jenkins, S.V., Vang, K.B., Gies, A., Griffin, R.J., Jun, S.-R., Nookaew, I. and Dings, R.P.M. Sample storage conditions induce post-collection biases in microbiome profiles. *BMC microbiology.* 2018, **18**(1), p.227.
426. Cardona, S., Eck, A., Cassellas, M., Gallart, M., Alastrue, C., Dore, J., Azpiroz, F., Roca, J., Guarner, F. and Manichanh, C. Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC microbiology.* 2012, **12**(1), p.158.
427. Fouhy, F., Deane, J., Rea, M.C., O'Sullivan, Ó., Ross, R.P., O'Callaghan, G., Plant, B.J. and Stanton, C. The Effects of Freezing on Faecal Microbiota as Determined Using MiSeq Sequencing and Culture-Based Investigations. *PloS one.* 2015, **10**(3), p.e0119355.
428. Tedjo, D.I., Jonkers, D.M., Savelkoul, P.H., Masclee, A.A., van Best, N., Pierik, M.J. and Penders, J. The effect of sampling and storage on the fecal microbiota composition in healthy and diseased subjects. *PloS one.* 2015, **10**(5), p.e0126685.
429. Roesch, L.F., Casella, G., Simell, O., Krischer, J., Wasserfall, C.H., Schatz, D., Atkinson, M.A., Neu, J. and Triplett, E.W. Influence of fecal sample storage on bacterial community diversity. *Open Microbiol J.* 2009, **3**, pp.40-46.
430. Amir, A., McDonald, D., Navas-Molina, J.A., Debelius, J., Morton, J.T., Hyde, E., Robbins-Pianka, A. and Knight, R. Correcting for Microbial Blooms in Fecal Samples during Room-Temperature Shipping. *mSystems.* 2017, **2**(2), pp.e00199-00116.
431. Lauber, C.L., Zhou, N., Gordon, J.I., Knight, R. and Fierer, N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiol Lett.* 2010, **307**(1), pp.80-86.
432. Vogtmann, E., Chen, J., Amir, A., Shi, J., Abnet, C.C., Nelson, H., Knight, R., Chia, N. and Sinha, R. Comparison of Collection Methods for Fecal Samples in Microbiome Studies. *Am J Epidemiol.* 2017, **185**(2), pp.115-123.
433. Wong, W.S.W., Clemency, N., Klein, E., Provenzano, M., Iyer, R., Niederhuber, J.E. and Hourigan, S.K. Collection of non-meconium stool on fecal occult blood cards is an effective method for fecal microbiota studies in infants. *Microbiome.* 2017, **5**(1), p.114.
434. Taylor, M., Wood, H.M., Halloran, S.P. and Quirke, P. Examining the potential use and long-term stability of guaiac faecal occult blood test cards for microbial DNA 16S rRNA sequencing. *Journal of clinical pathology.* 2017, **70**(7), pp.600-606.
435. Song, S.J., Amir, A., Metcalf, J.L., Amato, K.R., Xu, Z.Z., Humphrey, G. and Knight, R. Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. *mSystems.* 2016, **1**(3), p.[no pagination].
436. Vogtmann, E., Chen, J., Kibriya, M.G., Chen, Y., Islam, T., Eunes, M., Ahmed, A., Naher, J., Rahman, A., Amir, A., Shi, J., Abnet, C.C.,

Nelson, H., Knight, R., Chia, N., Ahsan, H. and Sinha, R. Comparison of Fecal Collection Methods for Microbiota Studies in Bangladesh. *Appl Environ Microbiol.* 2017, **83**(10).

437. Anderson, E.L., Li, W., Klitgord, N., Highlander, S.K., Dayrit, M., Seguritan, V., Yooseph, S., Biggs, W., Venter, J.C., Nelson, K.E. and Jones, M.B. A robust ambient temperature collection and stabilization strategy: Enabling worldwide functional studies of the human microbiome. *Scientific reports.* 2016, **6**, p.31731.
438. Szopinska, J.W., Gresse, R., van der Marel, S., Boekhorst, J., Lukovac, S., van Swam, I., Franke, B., Timmerman, H., Belzer, C. and Arias Vasquez, A. Reliability of a participant-friendly fecal collection method for microbiome analyses: a step towards large sample size investigation. *BMC microbiology.* 2018, **18**(1), p.110.
439. Hill, C.J., Brown, J.R.M., Lynch, D.B., Jeffery, I.B., Ryan, C.A., Ross, R.P., Stanton, C. and O'Toole, P.W. Effect of room temperature transport vials on DNA quality and phylogenetic composition of faecal microbiota of elderly adults and infants. *Microbiome.* 2016, **4**(1), p.19.
440. Choo, J.M., Leong, L.E.X. and Rogers, G.B. Sample storage conditions significantly influence faecal microbiome profiles. *Scientific reports.* 2015, **5**, p.16350.
441. Gudra, D., Shoaie, S., Fridmanis, D., Klovins, J., Wefer, H., Silamikelis, I., Peculis, R., Kalnina, I., Elbere, I., Radovica-Spalvina, I., Hultcrantz, R., Šķenders, Ģ., Leja, M. and Engstrand, L. A widely used sampling device in colorectal cancer screening programmes allows for large-scale microbiome studies. *Gut.* 2019, **68**(9), p.1723.
442. Couto Furtado Albuquerque, M., van Herwaarden, Y., Kortman, G.A.M., Dutilh, B.E., Bisseling, T. and Boleij, A. Preservation of bacterial DNA in 10-year-old guaiac FOBT cards and FIT tubes. *Journal of clinical pathology.* 2017, **70**(11), p.994.
443. Rounge, T.B., Meisal, R., Nordby, J.I., Ambur, O.H., de Lange, T. and Hoff, G. Evaluating gut microbiota profiles from archived fecal samples. *BMC Gastroenterol.* 2018, **18**(1), p.171.
444. Loftfield, E., Vogtmann, E., Sampson, J.N., Moore, S.C., Nelson, H., Knight, R., Chia, N. and Sinha, R. Comparison of Collection Methods for Fecal Samples for Discovery Metabolomics in Epidemiologic Studies. *Cancer Epidemiol Biomarkers Prev.* 2016, **25**(11), pp.1483-1490.
445. Imperiale, T.F., Ransohoff, D.F., Itzkowitz, S.H., Levin, T.R., Lavin, P., Lidgard, G.P., Ahlquist, D.A. and Berger, B.M. Multitarget stool DNA testing for colorectal-cancer screening. *N Engl J Med.* 2014, **370**(14), pp.1287-1297.
446. Carozzi, F.M. and Sani, C. Fecal Collection and Stabilization Methods for Improved Fecal DNA Test for Colorectal Cancer in a Screening Setting. *Journal of Cancer Research.* 2013, **2013**, p.8.
447. Nechvatal, J.M., Ram, J.L., Basson, M.D., Namprachan, P., Niec, S.R., Badsha, K.Z., Matherly, L.H., Majumdar, A.P. and Kato, I. Fecal collection, ambient preservation, and DNA extraction for PCR amplification of bacterial and human markers from human feces. *J Microbiol Methods.* 2008, **72**(2), pp.124-132.
448. Schultze, A., Akmatov, M.K., Andrzejak, M., Karras, N., Kemmling, Y., Maulhardt, A., Wieghold, S., Ahrens, W., Gunther, K., Schlenz, H.,

Krause, G. and Pessler, F. Comparison of stool collection on site versus at home in a population-based study : feasibility and participants' preference in Pretest 2 of the German National Cohort. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz.* 2014, **57**(11), pp.1264-1269.

449. Abrahamson, M., Hooker, E., Ajami, N.J., Petrosino, J.F. and Orwoll, E.S. Successful collection of stool samples for microbiome analyses from a large community-based population of elderly men. *Contemporary Clinical Trials Communications.* 2017, **7**, pp.158-162.
450. von Wagner, C., Baio, G., Raine, R., Snowball, J., Morris, S., Atkin, W., Obichere, A., Handley, G., Logan, R.F., Rainbow, S., Smith, S., Halloran, S. and Wardle, J. Inequalities in participation in an organized national colorectal cancer screening programme: results from the first 2.6 million invitations in England. *International Journal of Epidemiology.* 2011, **40**(3), pp.712-718.
451. Moss, S., Mathews, C., Day, T.J., Smith, S., Seaman, H.E., Snowball, J. and Halloran, S.P. Increased uptake and improved outcomes of bowel cancer screening with a faecal immunochemical test: results from a pilot study within the national screening programme in England. *Gut.* 2017, **66**(9), pp.1631-1644.
452. Chambers, J.A., Callander, A.S., Grangeret, R. and O'Carroll, R.E. Attitudes towards the Faecal Occult Blood Test (FOBT) versus the Faecal Immunochemical Test (FIT) for colorectal cancer screening: perceived ease of completion and disgust. *BMC cancer.* 2016, **16**, p.96.
453. *Scottish Bowel Screening Programme Statistics for invitations between 1 May 2016 and 30 April 2018.* [Online]. 2019. [Accessed 20.7.19]. Available from: https://www.isdscotland.org/Health-Topics/Cancer/Publications/2019-02-05/2019-02-05-Bowel-Screening-Publication-Summary.pdf
454. *Guidance on regulations for the transport of infectious substances 2019–2020.Geneva: World Health Organization (WHO/WHE/CPI/2019.20).* [Online]. 2019. [Accessed 28.7.19]. Available from: https://www.who.int/ihr/publications/WHO-WHE-CPI-2019.20/en/
455. *Guidance for Public Health and Commissioners. Public Health Resource Unit. BCSP Publication No 3.* [Online]. 2008. [Accessed 11.2.19].
456. Yu, Z. and Morrison, M. Improved extraction of PCR-quality community DNA from digesta and fecal samples. *BioTechniques.* 2004, **36**(5), pp.808-812.
457. Hartmann, C., Lennartz, K., Ibrahim, H., Coz, A., Kasper, Y., Lenz, C., Mathur, D. and Polidor, M. *Stable 16-year storage of DNA purified with the QIAamp® DNA Blood Mini Kit.* [Online]. 2016. [Accessed 5.8.19]. Available from: www.qiagen.com
458. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal.* 2011, **17**(1), pp.10-12.
459. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. and Holmes, S.P. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods.* 2016, **13**, p.581.
460. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai,

Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R. and Caporaso, J.G. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology.* 2019, **37**(8), pp.852-857.

461. Shannon, C.E. and Weaver, W. *The mathematical theory of communication.* University of Illonois Press, Champaign, Illonois, 1949.
462. Kruskal, W.H. and Wallis, W.A. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association.* 1952, **47**(260), pp.583-621.
463. Bray, J.R. and Curtis, J.T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs.* 1957, **27**(4), pp.325-349.
464. Halko, N., Martinsson, P.-G., Shkolnisky, Y. and Tygert, M. An Algorithm for the Principal Component Analysis of Large Data Sets. *SIAM J. Sci. Comput.* 2011, **33**(5), pp.2580-2594.
465. Anderson, M.J. A new method for non-parametric multivariate analysis of variance. *Austral Ecology.* 2001, **26**(1), pp.32-46.
466. Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.M., Szoecs, E. and Wagner, H. *vegan: Community Ecology Package. 2018. R package version 2.5-3.* [Online]. 2018. [Accessed 13.8.19]. Available from: https://CRAN.R-project.org/package=vegan.
467. Mallick H, T.T., McIver LJ, Rahnavard G, Nguyen LH, Weingart G, Ma S, Ren B, Schwager E, Subramanian A, Paulson JN, Franzosa EA, Corrada Bravo H, Huttenhower C. Multivariable Association in Population-scale Meta'omic Surveys. In Submission.
468. Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G.A. and Gregory Caporaso, J. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome.* 2018, **6**(1), p.90.

469. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. BLAST+: architecture and applications. *BMC bioinformatics.* 2009, **10**(1), p.421.
470. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F.O. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2012, **41**(D1), pp.D590-D596.
471. Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W. and Glöckner, F.O. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* 2013, **42**(D1), pp.D643-D648.
472. Glöckner, F.O., Yilmaz, P., Quast, C., Gerken, J., Beccati, A., Ciuprina, A., Bruns, G., Yarza, P., Peplies, J., Westram, R. and Ludwig, W. 25 years of serving the community with ribosomal RNA gene reference databases and tools. *Journal of Biotechnology.* 2017, **261**, pp.169-176.
473. Giavarina, D. Understanding Bland Altman analysis. *Biochemia medica.* 2015, **25**(2), pp.141-151.
474. Bland, J.M. and Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986, **1**(8476), pp.307-310.
475. Datta, D. *blandr: a Bland-Altman Method Comparison package for R.* [Online]. 2017. [Accessed 1.9.19]. Available from: https://github.com/deepankardatta/blandr
476. Peters, B.A., Dominianni, C., Shapiro, J.A., Church, T.R., Wu, J., Miller, G., Yuen, E., Freiman, H., Lustbader, I., Salik, J., Friedlander, C., Hayes, R.B. and Ahn, J. The gut microbiota in conventional and serrated precursors of colorectal cancer. *Microbiome.* 2016, **4**(1), pp.69-69.
477. Zhang, B., Xu, S., Xu, W., Chen, Q., Chen, Z., Yan, C., Fan, Y., Zhang, H., Liu, Q., Yang, J., Yang, J., Xiao, C., Xu, H. and Ren, J. Leveraging Fecal Bacterial Survey Data to Predict Colorectal Tumors. *Frontiers in Genetics.* 2019, **10**, pp.447-447.
478. Hornung, B.V.H., Zwittink, R.D. and Kuijper, E.J. Issues and current standards of controls in microbiome research. *FEMS microbiology ecology.* 2019, **95**(5), p.fiz045.
479. Pollock, J., Glendinning, L., Wisedchanwet, T. and Watson, M. The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Applied and environmental microbiology.* 2018, **84**(7), pp.e02627-02617.
480. Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J. and Walker, A.W. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 2014, **12**, p.87.
481. Laurence, M., Hatzis, C. and Brash, D.E. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PloS one.* 2014, **9**(5), p.e97876.
482. Glassing, A., Dowd, S.E., Galandiuk, S., Davis, B. and Chiodini, R.J. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* 2016, **8**, pp.24-24.

483. Eisenhofer, R., Minich, J.J., Marotz, C., Cooper, A., Knight, R. and Weyrich, L.S. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in microbiology.* 2019, **27**(2), pp.105-117.
484. Weyrich, L.S., Farrer, A.G., Eisenhofer, R., Arriola, L.A., Young, J., Selway, C.A., Handsley-Davis, M., Adler, C.J., Breen, J. and Cooper, A. Laboratory contamination over time during low-biomass sample analysis. *Molecular Ecology Resources.* 2019, **19**(4), pp.982-996.
485. Minich, J.J., Sanders, J.G., Amir, A., Humphrey, G., Gilbert, J.A. and Knight, R. Quantifying and Understanding Well-to-Well Contamination in Microbiome Research. *mSystems.* 2019, **4**(4), pp.e00186-00119.
486. Yu, S.Y., Xie, Y.H., Qiu, Y.W., Chen, Y.X. and Fang, J.Y. Moderate alteration to gut microbiota brought by colorectal adenoma resection. *J Gastroenterol Hepatol.* 2019, p.[no pagination].
487. Sze, M.A., Baxter, N.T., Ruffin, M.T.t., Rogers, M.A.M. and Schloss, P.D. Normalization of the microbiota in patients after treatment for colonic lesions. *Microbiome.* 2017, **5**(1), p.150.
488. Swidsinski, A., Loening-Baucke, V., Verstraelen, H., Osowska, S. and Doerffel, Y. Biostructure of fecal microbiota in healthy subjects and patients with chronic idiopathic diarrhea. *Gastroenterology.* 2008, **135**(2), pp.568-579.
489. Martinson, J.N.V., Pinkham, N.V., Peters, G.W., Cho, H., Heng, J., Rauch, M., Broadaway, S.C. and Walk, S.T. Rethinking gut microbiome residency and the Enterobacteriaceae in healthy human adults. *The ISME journal.* 2019, **13**(9), pp.2306-2318.
490. Fang, X., Monk, J.M., Nurk, S., Akseshina, M., Zhu, Q., Gemmell, C., Gianetto-Hill, C., Leung, N., Szubin, R., Sanders, J., Beck, P.L., Li, W., Sandborn, W.J., Gray-Owen, S.D., Knight, R., Allen-Vercoe, E., Palsson, B.O. and Smarr, L. Metagenomics-Based, Strain-Level Analysis of Escherichia coli From a Time-Series of Microbiome Samples From a Crohn's Disease Patient. *Front Microbiol.* 2018, **9**, p.2559.
491. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012, **486**(7402), pp.207-214.
492. Aagaard, K., Petrosino, J., Keitel, W., Watson, M., Katancik, J., Garcia, N., Patel, S., Cutting, M., Madden, T., Hamilton, H., Harris, E., Gevers, D., Simone, G., McInnes, P. and Versalovic, J. The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. *Faseb j.* 2013, **27**(3), pp.1012-1022.
493. Caporaso, J.G., Lauber, C.L., Costello, E.K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., Gordon, J.I. and Knight, R. Moving pictures of the human microbiome. *Genome biology.* 2011, **12**(5), p.R50.
494. Anderson, M.A., Whitlock, J.E. and Harwood, V.J. Diversity and Distribution of Escherichia coli Genotypes and Antibiotic Resistance Phenotypes in Feces of Humans, Cattle, and Horses. *Applied and environmental microbiology.* 2006, **72**(11), p.6914.
495. Mori, G., Rampelli, S., Orena, B.S., Rengucci, C., De Maio, G., Barbieri, G., Passardi, A., Casadei Gardini, A., Frassineti, G.L., Gaiarsa, S., Albertini, A.M., Ranzani, G.N., Calistri, D. and Pasca, M.R. Shifts of

Faecal Microbiota During Sporadic Colorectal Carcinogenesis. *Sci Rep.* 2018, **8**(1), p.10329.

496. Thomas, A.M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., Gandini, S., Serrano, D., Tarallo, S., Francavilla, A., Gallo, G., Trompetto, M., Ferrero, G., Mizutani, S., Shiroma, H., Shiba, S., Shibata, T., Yachida, S., Yamada, T., Wirbel, J., Schrotz-King, P., Ulrich, C.M., Brenner, H., Arumugam, M., Bork, P., Zeller, G., Cordero, F., Dias-Neto, E., Setubal, J.C., Tett, A., Pardini, B., Rescigno, M., Waldron, L., Naccarati, A. and Segata, N. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat Med.* 2019, **25**(4), pp.667-678.
497. Knights, D., Kuczynski, J., Koren, O., Ley, R.E., Field, D., Knight, R., DeSantis, T.Z. and Kelley, S.T. Supervised classification of microbiota mitigates mislabeling errors. *The ISME journal.* 2011, **5**(4), pp.570-573.
498. Flores, G.E., Caporaso, J.G., Henley, J.B., Rideout, J.R., Domogala, D., Chase, J., Leff, J.W., Vázquez-Baeza, Y., Gonzalez, A., Knight, R., Dunn, R.R. and Fierer, N. Temporal variability is a personalized feature of the human microbiome. *Genome biology.* 2014, **15**(12), p.531.
499. Koo, S., Neilson, L.J., Von Wagner, C. and Rees, C.J. The NHS Bowel Cancer Screening Program: current perspectives on strategies for improvement. *Risk management and healthcare policy.* 2017, **10**, pp.177-187.
500. Hewitson, P., Glasziou, P., Watson, E., Towler, B. and Irwig, L. Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. *Am J Gastroenterol.* 2008, **103**(6), pp.1541-1549.
501. *Bowel cancer screening: the facts (FOB test kit).* [Online]. [Accessed 24.9.19]. Available from: https://www.gov.uk/government/publications/bowel-cancer-screening-benefits-and-risks
502. Bretthauer, M. Colorectal cancer screening. *J Intern Med.* 2011, **270**(2), pp.87-98.
503. Tan, B., Qiu, Y., Zou, X., Chen, T., Xie, G., Cheng, Y., Dong, T., Zhao, L., Feng, B., Hu, X., Xu, L.X., Zhao, A., Zhang, M., Cai, G., Cai, S., Zhou, Z., Zheng, M., Zhang, Y. and Jia, W. Metabonomics identifies serum metabolite markers of colorectal cancer. *Journal of proteome research.* 2013, **12**(6), pp.3000-3009.
504. Guevarra, L.A., Afable, A.C.F., Belza, P.J.O., Dy, K.J.S., Lee, S.J.Q., Sy-Ortin, T.T. and Albano, P.M.S.P. Immunogenicity of a Fap2 peptide mimotope of Fusobacterium nucleatum and its potential use in the diagnosis of colorectal cancer. *Infectious Agents and Cancer.* 2018, **13**(1), p.11.
505. Kato, I., Boleij, A., Kortman, G.A.M., Roelofs, R., Djuric, Z., Severson, R.K. and Tjalsma, H. Partial associations of dietary iron, smoking and intestinal bacteria with colorectal cancer risk. *Nutrition and cancer.* 2013, **65**(2), pp.169-177.
506. Wang, Z., Lin, Y., Liang, J., Huang, Y., Ma, C., Liu, X. and Yang, J. NMR-based metabolomic techniques identify potential urinary biomarkers for early colorectal cancer detection. *Oncotarget.* 2017, **8**(62), pp.105819-105831.

507. Cheng, Y., Xie, G., Chen, T., Qiu, Y., Zou, X., Zheng, M., Tan, B., Feng, B., Dong, T., He, P., Zhao, L., Zhao, A., Xu, L.X., Zhang, Y. and Jia, W. Distinct urinary metabolic profile of human colorectal cancer. *Journal of proteome research.* 2012, **11**(2), pp.1354-1363.
508. Qiu, Y., Cai, G., Su, M., Chen, T., Liu, Y., Xu, Y., Ni, Y., Zhao, A., Cai, S., Xu, L.X. and Jia, W. Urinary metabonomic study on colorectal cancer. *Journal of proteome research.* 2010, **9**(3), pp.1627-1634.
509. Sonoda, H., Kohnoe, S., Yamazato, T., Satoh, Y., Morizono, G., Shikata, K., Morita, M., Watanabe, A., Morita, M., Kakeji, Y., Inoue, F. and Maehara, Y. Colorectal cancer screening with odour material by canine scent detection. *Gut.* 2011, **60**(6), p.814.
510. Amal, H., Leja, M., Funka, K., Lasina, I., Skapars, R., Sivins, A., Ancans, G., Kikuste, I., Vanags, A., Tolmanis, I., Kirsners, A., Kupcinskas, L. and Haick, H. Breath testing as potential colorectal cancer screening tool. *International journal of cancer.* 2016, **138**(1), pp.229-236.
511. Altomare, D.F., Di Lena, M., Porcelli, F., Trizio, L., Travaglio, E., Tutino, M., Dragonieri, S., Memeo, V. and de Gennaro, G. Exhaled volatile organic compounds identify patients with colorectal cancer. *Br J Surg.* 2013, **100**(1), pp.144-150.
512. de Meij, T.G., Larbi, I.B., van der Schee, M.P., Lentferink, Y.E., Paff, T., Terhaar sive Droste, J.S., Mulder, C.J., van Bodegraven, A.A. and de Boer, N.K. Electronic nose can discriminate colorectal carcinoma and advanced adenomas by fecal volatile biomarker analysis: proof of principle study. *International journal of cancer.* 2014, **134**(5), pp.1132-1138.
513. Bond, A., Greenwood, R., Lewis, S., Corfe, B., Sarkar, S., O'Toole, P., Rooney, P., Burkitt, M., Hold, G. and Probert, C. Volatile organic compounds emitted from faeces as a biomarker for colorectal cancer. *Aliment Pharmacol Ther.* 2019, **49**(8), pp.1005-1012.
514. Batty, C.A., Cauchi, M., Lourenco, C., Hunter, J.O. and Turner, C. Use of the Analysis of the Volatile Faecal Metabolome in Screening for Colorectal Cancer. *PloS one.* 2015, **10**(6), p.e0130301.
515. McFarlane, M., Millard, A., Hall, H., Savage, R., Constantinidou, C., Arasaradnam, R. and Nwokolo, C. Urinary volatile organic compounds and faecal microbiome profiles in colorectal cancer. *Colorectal Dis.* 2019, p.[no pagination].
516. Arasaradnam, R.P., McFarlane, M.J., Ryan-Fisher, C., Westenbrink, E., Hodges, P., Thomas, M.G., Chambers, S., O'Connell, N., Bailey, C., Harmston, C., Nwokolo, C.U., Bardhan, K.D. and Covington, J.A. Detection of colorectal cancer (CRC) by urinary volatile organic compound analysis. *PloS one.* 2014, **9**(9), p.e108750.
517. Widlak, M.M., Neal, M., Daulton, E., Thomas, C.L., Tomkins, C., Singh, B., Harmston, C., Wicaksono, A., Evans, C., Smith, S., Savage, R.S., Covington, J.A. and Arasaradnam, R.P. Risk stratification of symptomatic patients suspected of colorectal cancer using faecal and urinary markers. *Colorectal Dis.* 2018, **20**(12), pp.O335-o342.
518. Amiot, A., Dona, A.C., Wijeyesekera, A., Tournigand, C., Baumgaertner, I., Lebaleur, Y., Sobhani, I. and Holmes, E. (1)H NMR Spectroscopy of Fecal Extracts Enables Detection of Advanced Colorectal Neoplasia. *J Proteome Res.* 2015, **14**(9), pp.3871-3881.

519. Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Bohm, J., Brunetti, F., Habermann, N., Hercog, R., Koch, M., Luciani, A., Mende, D.R., Schneider, M.A., Schrotz-King, P., Tournigand, C., Tran Van Nhieu, J., Yamada, T., Zimmermann, J., Benes, V., Kloor, M., Ulrich, C.M., von Knebel Doeberitz, M., Sobhani, I. and Bork, P. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol.* 2014, **10**, p.766.
520. Arabameri, A., Asemani, D. and Teymourpour, P. Detection of Colorectal Carcinoma Based on Microbiota Analysis using Generalized Regression Neural Networks and Nonlinear Feature Selection. *IEEE/ACM Trans Comput Biol Bioinform.* 2018, p.[no pagination].
521. Goedert, J.J., Gong, Y., Hua, X., Zhong, H., He, Y., Peng, P., Yu, G., Wang, W., Ravel, J., Shi, J. and Zheng, Y. Fecal Microbiota Characteristics of Patients with Colorectal Adenoma Detected by Screening: A Population-based Study. *EBioMedicine.* 2015, **2**(6), pp.597-603.
522. Zackular, J.P., Rogers, M.A., Ruffin, M.T.t. and Schloss, P.D. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev Res (Phila).* 2014, **7**(11), pp.1112-1121.
523. Baxter, N.T., Ruffin, M.T.t., Rogers, M.A. and Schloss, P.D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* 2016, **8**(1), p.37.
524. Amiot, A., Mansour, H., Baumgaertner, I., Delchier, J.C., Tournigand, C., Furet, J.P., Carrau, J.P., Canoui-Poitrine, F. and Sobhani, I. The detection of the methylated Wif-1 gene is more accurate than a fecal occult blood test for colorectal cancer screening. *PloS one.* 2014, **9**(7), p.e99233.
525. Wong, S.H., Kwong, T.N.Y., Chow, T.-C., Luk, A.K.C., Dai, R.Z.W., Nakatsu, G., Lam, T.Y.T., Zhang, L., Wu, J.C.Y., Chan, F.K.L., Ng, S.S.M., Wong, M.C.S., Ng, S.C., Wu, W.K.K., Yu, J. and Sung, J.J.Y. Quantitation of faecal Fusobacterium improves faecal immunochemical test in detecting advanced colorectal neoplasia. *Gut.* 2017, **66**(8), pp.1441-1448.
526. Amitay, E.L., Werner, S., Vital, M., Pieper, D.H., Hofler, D., Gierse, I.J., Butt, J., Balavarca, Y., Cuk, K. and Brenner, H. Fusobacterium and colorectal cancer: Causal factor or passenger? Results from a large colorectal cancer screening study. *Carcinogenesis.* 2017, **38**(8), pp.781-788.
527. Suehiro, Y., Sakai, K., Nishioka, M., Hashimoto, S., Takami, T., Higaki, S., Shindo, Y., Hazama, S., Oka, M., Nagano, H., Sakaida, I. and Yamasaki, T. Highly sensitive stool DNA testing of Fusobacterium nucleatum as a marker for detection of colorectal tumours in a Japanese population. *Annals of clinical biochemistry.* 2017, **54**(1), pp.86-91.
528. Liang, Q., Chiu, J., Chen, Y., Huang, Y., Higashimori, A., Fang, J., Brim, H., Ashktorab, H., Chien Ng, S., Ng, S.S.M., Zheng, S., Chan, F.K.L., Sung, J.J.Y. and Yu, J. Fecal bacteria act as novel biomarkers for noninvasive diagnosis of colorectal cancer. *Clinical Cancer Research.* 2017, **23**(8), pp.2061-2070.
529. Yu, J., Feng, Q., Wong, S.H., Zhang, D., Liang, Q.Y., Qin, Y., Tang, L., Zhao, H., Stenvang, J., Li, Y., Wang, X., Xu, X., Chen, N., Wu, W.K.,

Al-Aama, J., Nielsen, H.J., Kiilerich, P., Jensen, B.A., Yau, T.O., Lan, Z., Jia, H., Li, J., Xiao, L., Lam, T.Y., Ng, S.C., Cheng, A.S., Wong, V.W., Chan, F.K., Xu, X., Yang, H., Madsen, L., Datz, C., Tilg, H., Wang, J., Brunner, N., Kristiansen, K., Arumugam, M., Sung, J.J. and Wang, J. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut.* 2017, **66**(1), pp.70-78.

530. Eklof, V., Lofgren-Burstrom, A., Zingmark, C., Edin, S., Larsson, P., Karling, P., Alexeyev, O., Rutegard, J., Wikberg, M.L. and Palmqvist, R. Cancer-associated fecal microbial markers in colorectal cancer detection. *International journal of cancer.* 2017, **141**(12), pp.2528-2536.
531. Rezasoltani, S., Sharafkhah, M., Asadzadeh Aghdaei, H., Nazemalhosseini Mojarad, E., Dabiri, H., Akhavan Sepahi, A., Modarressi, M.H., Feizabadi, M.M. and Zali, M.R. Applying simple linear combination, multiple logistic and factor analysis methods for candidate fecal bacteria as novel biomarkers for early detection of adenomatous polyps and colon cancer. *J Microbiol Methods.* 2018, **155**, pp.82-88.
532. Xie, Y.H., Gao, Q.Y., Cai, G.X., Sun, X.M., Zou, T.H., Chen, H.M., Yu, S.Y., Qiu, Y.W., Gu, W.Q., Chen, X.Y., Cui, Y., Sun, D., Liu, Z.J., Cai, S.J., Xu, J., Chen, Y.X. and Fang, J.Y. Fecal Clostridium symbiosum for Noninvasive Detection of Early and Advanced Colorectal Cancer: Test and Validation Studies. *EBioMedicine.* 2017, **25**, pp.32-40.
533. Zhang, X., Zhu, X., Cao, Y., Fang, J.Y., Hong, J. and Chen, H. Fecal Fusobacterium nucleatum for the diagnosis of colorectal tumor: A systematic review and meta-analysis. *Cancer Med.* 2019, **8**(2), pp.480-491.
534. Shah, M.S., DeSantis, T.Z., Weinmaier, T., McMurdie, P.J., Cope, J.L., Altrichter, A., Yamal, J.M. and Hollister, E.B. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut.* 2018, **67**(5), pp.882-891.
535. Bressler, B., Paszat, L.F., Chen, Z., Rothwell, D.M., Vinden, C. and Rabeneck, L. Rates of New or Missed Colorectal Cancers After Colonoscopy and Their Risk Factors: A Population-Based Analysis. *Gastroenterology.* 2007, **132**(1), pp.96-102.
536. Zhao, S., Wang, S., Pan, P., Xia, T., Chang, X., Yang, X., Guo, L., Meng, Q., Yang, F., Qian, W., Xu, Z., Wang, Y., Wang, Z., Gu, L., Wang, R., Jia, F., Yao, J., Li, Z. and Bai, Y. Magnitude, Risk Factors, and Factors Associated With Adenoma Miss Rate of Tandem Colonoscopy: A Systematic Review and Meta-analysis. *Gastroenterology.* 2019, **156**(6), pp.1661-1674.e1611.
537. Amitay, E.L., Krilaviciute, A. and Brenner, H. Systematic review: Gut microbiota in fecal samples and detection of colorectal neoplasms. *Gut microbes.* 2018, **9**(4), pp.293-307.
538. Nagata, N., Tohya, M., Fukuda, S., Suda, W., Nishijima, S., Takeuchi, F., Ohsugi, M., Tsujimoto, T., Nakamura, T., Shimomura, A., Yanagisawa, N., Hisada, Y., Watanabe, K., Imbe, K., Akiyama, J., Mizokami, M., Miyoshi-Akiyama, T., Uemura, N. and Hattori, M. Effects of bowel preparation on the human gut microbiome and metabolome. *Scientific reports.* 2019, **9**(1), p.4042.

539. Jalanka, J., Salonen, A., Salojärvi, J., Ritari, J., Immonen, O., Marciani, L., Gowland, P., Hoad, C., Garsed, K., Lam, C., Palva, A., Spiller, R.C. and de Vos, W.M. Effects of bowel cleansing on the intestinal microbiota. *Gut.* 2015, **64**(10), p.1562.
540. O'Brien, C.L., Allison, G.E., Grimpen, F. and Pavli, P. Impact of colonoscopy bowel preparation on intestinal microbiota. *PloS one.* 2013, **8**(5), pp.e62815-e62815.
541. Baxter, N.T., Koumpouras, C.C., Rogers, M.A., Ruffin, M.T.t. and Schloss, P.D. DNA from fecal immunochemical test can replace stool for detection of colonic lesions using a microbiota-based model. *Microbiome.* 2016, **4**(1), p.59.
542. Brown, J.P., Wooldrage, K., Wright, S., Nickerson, C., Cross, A.J. and Atkin, W.S. High test positivity and low positive predictive value for colorectal cancer of continued faecal occult blood test screening after negative colonoscopy. *Journal of Medical Screening.* 2017, **25**(2), pp.70-75.
543. Thaiss, C.A., Zeevi, D., Levy, M., Zilberman-Schapira, G., Suez, J., Tengeler, A.C., Abramson, L., Katz, M.N., Korem, T., Zmora, N., Kuperman, Y., Biton, I., Gilad, S., Harmelin, A., Shapiro, H., Halpern, Z., Segal, E. and Elinav, E. Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. *Cell.* 2014, **159**(3), pp.514-529.
544. Liang, X., Bushman, F.D. and FitzGerald, G.A. Rhythmicity of the intestinal microbiota is regulated by gender and the host circadian clock. *Proc Natl Acad Sci U S A.* 2015, **112**(33), pp.10479-10484.
545. Smits, S.A., Leach, J., Sonnenburg, E.D., Gonzalez, C.G., Lichtman, J.S., Reid, G., Knight, R., Manjurano, A., Changalucha, J., Elias, J.E., Dominguez-Bello, M.G. and Sonnenburg, J.L. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science.* 2017, **357**(6353), p.802.
546. Davenport, E.R., Mizrahi-Man, O., Michelini, K., Barreiro, L.B., Ober, C. and Gilad, Y. Seasonal Variation in Human Gut Microbiome Composition. *PloS one.* 2014, **9**(3), p.e90731.
547. Flores, R., Shi, J., Fuhrman, B., Xu, X., Veenstra, T.D., Gail, M.H., Gajer, P., Ravel, J. and Goedert, J.J. Fecal microbial determinants of fecal and systemic estrogens and estrogen metabolites: a cross-sectional study. *Journal of translational medicine.* 2012, **10**, pp.253-253.
548. Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I.N., Bar, N., Shilo, S., Lador, D., Vila, A.V., Zmora, N., Pevsner-Fischer, M., Israeli, D., Kosower, N., Malka, G., Wolf, B.C., Avnit-Sagi, T., Lotan-Pompan, M., Weinberger, A., Halpern, Z., Carmi, S., Fu, J., Wijmenga, C., Zhernakova, A., Elinav, E. and Segal, E. Environment dominates over host genetics in shaping human gut microbiota. *Nature.* 2018, **555**, p.210.
549. Dominianni, C., Sinha, R., Goedert, J.J., Pei, Z., Yang, L., Hayes, R.B. and Ahn, J. Sex, Body Mass Index, and Dietary Fiber Intake Influence the Human Gut Microbiome. *PloS one.* 2015, **10**(4), p.e0124599.
550. Haro, C., Rangel-Zuniga, O.A., Alcala-Diaz, J.F., Gomez-Delgado, F., Perez-Martinez, P., Delgado-Lista, J., Quintana-Navarro, G.M., Landa,

B.B., Navas-Cortes, J.A., Tena-Sempere, M., Clemente, J.C., Lopez-Miranda, J., Perez-Jimenez, F. and Camargo, A. Intestinal Microbiota Is Influenced by Gender and Body Mass Index. *PloS one.* 2016, **11**(5), p.e0154090.

551. Yurkovetskiy, L., Burrows, M., Khan, A.A., Graham, L., Volchkov, P., Becker, L., Antonopoulos, D., Umesaki, Y. and Chervonsky, A.V. Gender bias in autoimmunity is influenced by microbiota. *Immunity.* 2013, **39**(2), pp.400-412.
552. Bolnick, D.I., Snowberg, L.K., Hirsch, P.E., Lauber, C.L., Org, E., Parks, B., Lusis, A.J., Knight, R., Caporaso, J.G. and Svanback, R. Individual diet has sex-dependent effects on vertebrate gut microbiota. *Nat Commun.* 2014, **5**, p.4500.
553. Claesson, M.J., Cusack, S., O'Sullivan, O., Greene-Diniz, R., de Weerd, H., Flannery, E., Marchesi, J.R., Falush, D., Dinan, T., Fitzgerald, G., Stanton, C., van Sinderen, D., O'Connor, M., Harnedy, N., O'Connor, K., Henry, C., O'Mahony, D., Fitzgerald, A.P., Shanahan, F., Twomey, C., Hill, C., Ross, R.P. and O'Toole, P.W. Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc Natl Acad Sci U S A.* 2011, **108 Suppl 1**, pp.4586-4591.
554. Biagi, E., Nylund, L., Candela, M., Ostan, R., Bucci, L., Pini, E., Nikkïla, J., Monti, D., Satokari, R., Franceschi, C., Brigidi, P. and De Vos, W. Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PloS one.* 2010, **5**(5), pp.e10667-e10667.
555. Biagi, E., Franceschi, C., Rampelli, S., Severgnini, M., Ostan, R., Turroni, S., Consolandi, C., Quercia, S., Scurti, M., Monti, D., Capri, M., Brigidi, P. and Candela, M. Gut Microbiota and Extreme Longevity. *Curr Biol.* 2016, **26**(11), pp.1480-1485.
556. Wang, F., Yu, T., Huang, G., Cai, D., Liang, X., Su, H., Zhu, Z., Li, D., Yang, Y., Shen, P., Mao, R., Yu, L., Zhao, M. and Li, Q. Gut Microbiota Community and Its Assembly Associated with Age and Diet in Chinese Centenarians. *J Microbiol Biotechnol.* 2015, **25**(8), pp.1195-1204.
557. Kong, F., Hua, Y., Zeng, B., Ning, R., Li, Y. and Zhao, J. Gut microbiota signatures of longevity. *Curr Biol.* 2016, **26**(18), pp.R832-r833.
558. Odamaki, T., Kato, K., Sugahara, H., Hashikura, N., Takahashi, S., Xiao, J.Z., Abe, F. and Osawa, R. Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol.* 2016, **16**, p.90.
559. Finlay, B.B., Pettersson, S., Melby, M.K. and Bosch, T.C.G. The Microbiome Mediates Environmental Effects on Aging. *BioEssays.* 2019, p.e1800257.
560. Smith, P., Willemsen, D., Popkes, M., Metge, F., Gandiwa, E., Reichard, M. and Valenzano, D.R. Regulation of life span by the gut microbiota in the short-lived African turquoise killifish. *eLife.* 2017, **6**, p.[no pagination].
561. Thevaranjan, N., Puchta, A., Schulz, C., Naidoo, A., Szamosi, J.C., Verschoor, C.P., Loukov, D., Schenck, L.P., Jury, J., Foley, K.P., Schertzer, J.D., Larché, M.J., Davidson, D.J., Verdú, E.F., Surette, M.G. and Bowdish, D.M.E. Age-Associated Microbial Dysbiosis Promotes Intestinal Permeability, Systemic Inflammation, and Macrophage Dysfunction. *Cell host & microbe.* 2017, **21**(4), pp.455-466.e454.

562. Franceschi, C., Garagnani, P., Parini, P., Giuliani, C. and Santoro, A. Inflammaging: a new immune-metabolic viewpoint for age-related diseases. *Nat Rev Endocrinol.* 2018, **14**(10), pp.576-590.
563. Muegge, B.D., Kuczynski, J., Knights, D., Clemente, J.C., Gonzalez, A., Fontana, L., Henrissat, B., Knight, R. and Gordon, J.I. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science.* 2011, **332**(6032), pp.970-974.
564. Ou, J., Carbonero, F., Zoetendal, E.G., DeLany, J.P., Wang, M., Newton, K., Gaskins, H.R. and O'Keefe, S.J. Diet, microbiota, and microbial metabolites in colon cancer risk in rural Africans and African Americans. *Am J Clin Nutr.* 2013, **98**(1), pp.111-120.
565. Lin, D., Peters, B.A., Friedlander, C., Freiman, H.J., Goedert, J.J., Sinha, R., Miller, G., Bernstein, M.A., Hayes, R.B. and Ahn, J. Association of dietary fibre intake and gut microbiota in adults. *Br J Nutr.* 2018, **120**(9), pp.1014-1022.
566. David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., Biddinger, S.B., Dutton, R.J. and Turnbaugh, P.J. Diet rapidly and reproducibly alters the human gut microbiome. *Nature.* 2014, **505**(7484), pp.559-563.
567. Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F.D. and Lewis, J.D. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science (New York, N.Y.).* 2011, **334**(6052), pp.105-108.
568. Zaura, E., Brandt, B.W., Teixeira de Mattos, M.J., Buijs, M.J., Caspers, M.P.M., Rashid, M.-U., Weintraub, A., Nord, C.E., Savell, A., Hu, Y., Coates, A.R., Hubank, M., Spratt, D.A., Wilson, M., Keijser, B.J.F. and Crielaard, W. Same Exposure but Two Radically Different Responses to Antibiotics: Resilience of the Salivary Microbiome versus Long-Term Microbial Shifts in Feces. *mBio.* 2015, **6**(6), pp.e01693-01615.
569. Palleja, A., Mikkelsen, K.H., Forslund, S.K., Kashani, A., Allin, K.H., Nielsen, T., Hansen, T.H., Liang, S., Feng, Q., Zhang, C., Pyl, P.T., Coelho, L.P., Yang, H., Wang, J., Typas, A., Nielsen, M.F., Nielsen, H.B., Bork, P., Wang, J., Vilsbøll, T., Hansen, T., Knop, F.K., Arumugam, M. and Pedersen, O. Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nature microbiology.* 2018, **3**(11), pp.1255-1265.
570. Dethlefsen, L., Huse, S., Sogin, M.L. and Relman, D.A. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.* 2008, **6**(11), p.e280.
571. Hildebrand, F., Moitinho-Silva, L., Blasche, S., Jahn, M.T., Gossmann, T.I., Huerta-Cepas, J., Hercog, R., Luetge, M., Bahram, M., Pryszlak, A., Alves, R.J., Waszak, S.M., Zhu, A., Ye, L., Costea, P.I., Aalvink, S., Belzer, C., Forslund, S.K., Sunagawa, S., Hentschel, U., Merten, C., Patil, K.R., Benes, V. and Bork, P. Antibiotics-induced monodominance of a novel gut bacterial order. *Gut.* 2019, pp.gutjnl-2018-317715.
572. Mikkelsen, K.H., Frost, M., Bahl, M.I., Licht, T.R., Jensen, U.S., Rosenberg, J., Pedersen, O., Hansen, T., Rehfeld, J.F., Holst, J.J., Vilsbøll, T. and Knop, F.K. Effect of Antibiotics on Gut Microbiota, Gut

Hormones and Glucose Metabolism. *PloS one.* 2015, **10**(11), p.e0142352.

573. Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E.E., Brochado, A.R., Fernandez, K.C., Dose, H., Mori, H., Patil, K.R., Bork, P. and Typas, A. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature.* 2018, **555**, p.623.
574. Milanovic, V., Osimani, A., Aquilanti, L., Tavoletti, S., Garofalo, C., Polverigiani, S., Litta-Mulondo, A., Cocolin, L., Ferrocino, I., Di Cagno, R., Turroni, S., Lazzi, C., Pellegrini, N. and Clementi, F. Occurrence of antibiotic resistance genes in the fecal DNA of healthy omnivores, ovo-lacto vegetarians and vegans. *Mol Nutr Food Res.* 2017, **61**(9).
575. Zhu, Y.-G., Zhao, Y., Li, B., Huang, C.-L., Zhang, S.-Y., Yu, S., Chen, Y.-S., Zhang, T., Gillings, M.R. and Su, J.-Q. Continental-scale pollution of estuaries with antibiotic resistance genes. *Nature microbiology.* 2017, **2**, p.16270.
576. Biedermann, L., Zeitz, J., Mwinyi, J., Sutter-Minder, E., Rehman, A., Ott, S.J., Steurer-Stey, C., Frei, A., Frei, P., Scharl, M., Loessner, M.J., Vavricka, S.R., Fried, M., Schreiber, S., Schuppler, M. and Rogler, G. Smoking Cessation Induces Profound Changes in the Composition of the Intestinal Microbiota in Humans. *PloS one.* 2013, **8**(3), p.e59260.
577. Kato, I., Nechvatal, J.M., Dzinic, S., Basson, M.D., Majumdar, A.P. and Ram, J.L. Smoking and other personal characteristics as potential predictors for fecal bacteria populations in humans. *Medical Science Monitor.* 2010, **16**(1), pp.CR1-CR7.
578. Zhang, M., Sun, K., Wu, Y., Yang, Y., Tso, P. and Wu, Z. Interactions between Intestinal Microbiota and Host Immune Response in Inflammatory Bowel Disease. *Front Immunol.* 2017, **8**, p.942.
579. Ianiro, G., Eusebi, L.H., Black, C.J., Gasbarrini, A., Cammarota, G. and Ford, A.C. Systematic review with meta-analysis: efficacy of faecal microbiota transplantation for the treatment of irritable bowel syndrome. 2019, **50**(3), pp.240-248.
580. Van de Wiele, T., Van Praet, J.T., Marzorati, M., Drennan, M.B. and Elewaut, D. How the microbiota shapes rheumatic diseases. *Nature Reviews Rheumatology.* 2016, **12**, p.398.
581. Al Khodor, S., Reichert, B. and Shatat, I.F. The Microbiome and Blood Pressure: Can Microbes Regulate Our Blood Pressure? *Frontiers in pediatrics.* 2017, **5**, pp.138-138.
582. Yang, T., Santisteban, M.M., Rodriguez, V., Li, E., Ahmari, N., Carvajal, J.M., Zadeh, M., Gong, M., Qi, Y., Zubcevic, J., Sahay, B., Pepine, C.J., Raizada, M.K. and Mohamadzadeh, M. Gut dysbiosis is linked to hypertension. *Hypertension.* 2015, **65**(6), pp.1331-1340.
583. Guzman-Castaneda, S.J., Ortega-Vega, E.L., de la Cuesta-Zuluaga, J., Velasquez-Mejia, E.P., Rojas, W., Bedoya, G. and Escobar, J.S. Gut microbiota composition explains more variance in the host cardiometabolic risk than genetic ancestry. *Gut microbes.* 2019, pp.1-14.
584. Peters, B.A., Shapiro, J.A., Church, T.R., Miller, G., Trinh-Shevrin, C., Yuen, E., Friedlander, C., Hayes, R.B. and Ahn, J. A taxonomic signature of obesity in a large study of American adults. *Sci Rep.* 2018, **8**(1), p.9749.

585. Zhou, W., Sailani, M.R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S.R., Zhang, M.J., Rao, V., Avina, M., Mishra, T., Johnson, J., Lee-McMullen, B., Chen, S., Metwally, A.A., Tran, T.D.B., Nguyen, H., Zhou, X., Albright, B., Hong, B.-Y., Petersen, L., Bautista, E., Hanson, B., Chen, L., Spakowicz, D., Bahmani, A., Salins, D., Leopold, B., Ashland, M., Dagan-Rosenfeld, O., Rego, S., Limcaoco, P., Colbert, E., Allister, C., Perelman, D., Craig, C., Wei, E., Chaib, H., Hornburg, D., Dunn, J., Liang, L., Rose, S.M.S.-F., Kukurba, K., Piening, B., Rost, H., Tse, D., McLaughlin, T., Sodergren, E., Weinstock, G.M. and Snyder, M. Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature.* 2019, **569**(7758), pp.663-671.
586. Bastiaanssen, T.F.S., Cowan, C.S.M., Claesson, M.J., Dinan, T.G. and Cryan, J.F. Making Sense of ... the Microbiome in Psychiatry. *Int J Neuropsychopharmacol.* 2019, **22**(1), pp.37-52.
587. Song, S.J., Lauber, C., Costello, E.K., Lozupone, C.A., Humphrey, G., Berg-Lyons, D., Caporaso, J.G., Knights, D., Clemente, J.C., Nakielny, S., Gordon, J.I., Fierer, N. and Knight, R. Cohabiting family members share microbiota with one another and with their dogs. *eLife.* 2013, **2**, p.e00458.
588. Nuriel-Ohayon, M., Neuman, H. and Koren, O. Microbial Changes during Pregnancy, Birth, and Infancy. *Frontiers in Microbiology.* 2016, **7**(1031), p.[no pagination].
589. Benedict, C., Vogel, H., Jonas, W., Woting, A., Blaut, M., Schurmann, A. and Cedernaes, J. Gut microbiota and glucometabolic alterations in response to recurrent partial sleep deprivation in normal-weight young individuals. *Mol Metab.* 2016, **5**(12), pp.1175-1186.
590. Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A.V., Falony, G., Vieira-Silva, S., Wang, J., Imhann, F., Brandsma, E., Jankipersadsing, S.A., Joossens, M., Cenit, M.C., Deelen, P., Swertz, M.A., Weersma, R.K., Feskens, E.J., Netea, M.G., Gevers, D., Jonkers, D., Franke, L., Aulchenko, Y.S., Huttenhower, C., Raes, J., Hofker, M.H., Xavier, R.J., Wijmenga, C. and Fu, J. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science.* 2016, **352**(6285), pp.565-569.
591. Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., Kurilshikov, A., Bonder, M.J., Valles-Colomer, M., Vandeputte, D., Tito, R.Y., Chaffron, S., Rymenans, L., Verspecht, C., De Sutter, L., Lima-Mendez, G., D'Hoe, K., Jonckheere, K., Homola, D., Garcia, R., Tigchelaar, E.F., Eeckhaudt, L., Fu, J., Henckaerts, L., Zhernakova, A., Wijmenga, C. and Raes, J. Population-level analysis of gut microbiome variation. *Science.* 2016, **352**(6285), pp.560-564.
592. Logan, R.F.A., Patnick, J., Nickerson, C., Coleman, L., Rutter, M.D. and von Wagner, C. Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. *Gut.* 2011, pp.1439-1446.
593. Breiman, L. Random Forests. *Machine Learning.* 2001, **45**(1), pp.5-32.
594. Wiener, A.L.a.M. Classification and Regression by randomForest. *R News.* 2002, **2**(3), pp.18-22.
595. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. and Müller, M. pROC: an open-source package for R and S+ to

analyze and compare ROC curves. *BMC bioinformatics.* 2011, **12**(1), p.77.

596. DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988, **44**(3), pp.837-845.
597. Wirbel, J., Pyl, P.T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J.S., Voigt, A.Y., Palleja, A., Ponnudurai, R., Sunagawa, S., Coelho, L.P., Schrotz-King, P., Vogtmann, E., Habermann, N., Niméus, E., Thomas, A.M., Manghi, P., Gandini, S., Serrano, D., Mizutani, S., Shiroma, H., Shiba, S., Shibata, T., Yachida, S., Yamada, T., Waldron, L., Naccarati, A., Segata, N., Sinha, R., Ulrich, C.M., Brenner, H., Arumugam, M., Bork, P. and Zeller, G. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature medicine.* 2019, **25**(4), pp.679-689.
598. Constante, M., Fragoso, G., Calve, A., Samba-Mondonga, M. and Santos, M.M. Dietary heme induces gut dysbiosis, aggravates colitis, and potentiates the development of adenomas in mice. *Frontiers in Microbiology.* 2017, **8**(SEP), p.1809.
599. Martin, O.C.B., Olier, M., Ellero-Simatos, S., Naud, N., Dupuy, J., Huc, L., Taché, S., Graillot, V., Levêque, M., Bézirard, V., Héliès-Toussaint, C., Estrada, F.B.Y., Tondereau, V., Lippi, Y., Naylies, C., Peyriga, L., Canlet, C., Davila, A.M., Blachier, F., Ferrier, L., Boutet-Robinet, E., Guéraud, F., Théodorou, V. and Pierre, F.H.F. Haem iron reshapes colonic luminal environment: impact on mucosal homeostasis and microbiome through aldehyde formation. *Microbiome.* 2019, **7**(1), pp.72-72.
600. Kortman, G.A.M., Dutilh, B.E., Maathuis, A.J.H., Engelke, U.F., Boekhorst, J., Keegan, K.P., Nielsen, F.G.G., Betley, J., Weir, J.C., Kingsbury, Z., Kluijtmans, L.A.J., Swinkels, D.W., Venema, K. and Tjalsma, H. Microbial Metabolism Shifts Towards an Adverse Profile with Supplementary Iron in the TIM-2 In vitro Model of the Human Colon. *Frontiers in Microbiology.* 2016, **6**(1481).
601. Jaeggi, T., Kortman, G.A.M., Moretti, D., Chassard, C., Holding, P., Dostal, A., Boekhorst, J., Timmerman, H.M., Swinkels, D.W., Tjalsma, H., Njenga, J., Mwangi, A., Kvalsvig, J., Lacroix, C. and Zimmermann, M.B. Iron fortification adversely affects the gut microbiome, increases pathogen abundance and induces intestinal inflammation in Kenyan infants. *Gut.* 2015, **64**(5), p.731.
602. Yilmaz, B. and Li, H. Gut Microbiota and Iron: The Crucial Actors in Health and Disease. *Pharmaceuticals (Basel).* 2018, **11**(4), p.98.
603. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics.* 2001, **29**(5), pp.1189-1232.
604. Jackson, M.A., Verdi, S., Maxan, M.-E., Shin, C.M., Zierer, J., Bowyer, R.C.E., Martin, T., Williams, F.M.K., Menni, C., Bell, J.T., Spector, T.D. and Steves, C.J. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nature communications.* 2018, **9**(1), p.2655.
605. Sears, C.L. The who, where and how of fusobacteria and colon cancer. *eLife.* 2018, **7**, p.e28434.

606. Repass, J., Iorns, E., Denis, A., Williams, S.R., Perfito, N. and Errington, T.M. Replication Study: Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *eLife.* 2018, **7**, p.e25801.
607. Libby, G., Fraser, C.G., Carey, F.A., Brewster, D.H. and Steele, R.J.C. Occult blood in faeces is associated with all-cause and non-colorectal cancer mortality. *Gut.* 2018, **67**(12), pp.2116-2123.
608. Steele, R.J., McDonald, P.J., Digby, J., Brownlee, L., Strachan, J.A., Libby, G., McClements, P.L., Birrell, J., Carey, F.A., Diament, R.H., Balsitis, M. and Fraser, C.G. Clinical outcomes using a faecal immunochemical test for haemoglobin as a first-line test in a national programme constrained by colonoscopy capacity. *United European Gastroenterol J.* 2013, **1**(3), pp.198-205.
609. Lalkhen, A.G. and McCluskey, A. Clinical tests: sensitivity and specificity. *BJA Education.* 2008, **8**(6), pp.221-223.
610. Robertson, D.J., Lee, J.K., Boland, C.R., Dominitz, J.A., Giardiello, F.M., Johnson, D.A., Kaltenbach, T., Lieberman, D., Levin, T.R. and Rex, D.K. Recommendations on Fecal Immunochemical Testing to Screen for Colorectal Neoplasia: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer. *Gastroenterology.* 2017, **152**(5), pp.1217-1237.e1213.
611. Lee, J.K., Liles, E.G., Bent, S., Levin, T.R. and Corley, D.A. Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. *Ann Intern Med.* 2014, **160**(3), p.171.
612. Launois, R., Le Moine, J.G., Uzzan, B., Fiestas Navarrete, L.I. and Benamouzig, R. Systematic review and bivariate/HSROC random-effect meta-analysis of immunochemical and guaiac-based fecal occult blood tests for colorectal cancer screening. *Eur J Gastroenterol Hepatol.* 2014, **26**(9), pp.978-989.
613. Lin, J.S., Piper, M.A., Perdue, L.A., Rutter, C.M., Webber, E.M., O'Connor, E., Smith, N. and Whitlock, E.P. Screening for Colorectal Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA.* 2016, **315**(23), pp.2576-2594.
614. Niedermaier, T., Weigl, K., Hoffmeister, M. and Brenner, H. Diagnostic performance of flexible sigmoidoscopy combined with fecal immunochemical test in colorectal cancer screening: meta-analysis and modeling. *Eur J Epidemiol.* 2017, **32**(6), pp.481-493.
615. Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D. and Lyu, C. A reliable method for colorectal cancer prediction based on feature selection and support vector machine. *Med Biol Eng Comput.* 2019, **57**(4), pp.901-912.
616. Zhai, R.L., Xu, F., Zhang, P., Zhang, W.L., Wang, H., Wang, J.L., Cai, K.L., Long, Y.P., Lu, X.M., Tao, K.X. and Wang, G.B. The Diagnostic Performance of Stool DNA Testing for Colorectal Cancer: A Systematic Review and Meta-Analysis. *Medicine (Baltimore).* 2016, **95**(5), p.e2129.
617. Dadkhah, E., Sikaroodi, M., Korman, L., Hardi, R., Baybick, J., Hanzel, D., Kuehn, G., Kuehn, T. and Gillevet, P.M. Gut microbiome identifies risk for colorectal polyps. *BMJ Open Gastroenterol.* 2019, **6**(1), p.e000297.
618. Fraser, C.G., Digby, J., McDonald, P.J., Strachan, J.A., Carey, F.A. and Steele, R.J. Experience with a two-tier reflex gFOBT/FIT strategy in a

national bowel screening programme. *J Med Screen.* 2012, **19**(1), pp.8-13.

619. Ai, D., Pan, H., Li, X., Gao, Y., Liu, G. and Xia, L.C. Identifying Gut Microbiota Associated With Colorectal Cancer Using a Zero-Inflated Lognormal Model. *Front Microbiol.* 2019, **10**, p.826.
620. Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A. and Alm, E.J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun.* 2017, **8**(1), p.1784.
621. Pasolli, E., Truong, D.T., Malik, F., Waldron, L. and Segata, N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS computational biology.* 2016, **12**(7), p.e1004977.
622. Mancabelli, L., Milani, C., Lugli, G.A., Turroni, F., Cocconi, D., van Sinderen, D. and Ventura, M. Identification of universal gut microbial biomarkers of common human intestinal diseases by meta-analysis. *FEMS Microbiol Ecol.* 2017, **93**(12), p.[no pagination].
623. Bang, S., Yoo, D., Kim, S.J., Jhang, S., Cho, S. and Kim, H. Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data. *Sci Rep.* 2019, **9**(1), p.10189.
624. Youssef, O., Lahti, L., Kokkola, A., Karla, T., Tikkanen, M., Ehsan, H., Carpelan-Holmstrom, M., Koskensalo, S., Bohling, T., Rautelin, H., Puolakkainen, P., Knuutila, S. and Sarhadi, V. Stool Microbiota Composition Differs in Patients with Stomach, Colon, and Rectal Neoplasms. *Digestive diseases and sciences.* 2018, pp.1-9.
625. Zhu, Z., Ren, J., Michail, S. and Sun, F. MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations. *Genome Biol.* 2019, **20**(1), p.154.
626. Sze, M.A., Topcuoglu, B.D., Lesniak, N.A., Ruffin, M.T.t. and Schloss, P.D. Fecal Short-Chain Fatty Acids Are Not Predictive of Colonic Tumor Status and Cannot Be Predicted Based on Bacterial Community Structure. *mBio.* 2019, **10**(4).
627. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians.* 2018, **0**(0), p.[no pagination].
628. Arnold, M., Sierra, M.S., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F. Global patterns and trends in colorectal cancer incidence and mortality. *Gut.* 2017, **66**(4), p.683.
629. Bishehsari, F., Mahdavinia, M., Vacca, M., Malekzadeh, R. and Mariani-Costantini, R. Epidemiological transition of colorectal cancer in developing countries: Environmental factors, molecular pathways, and opportunities for prevention. *World journal of gastroenterology.* 2014, **20**(20), pp.6055-6072.
630. Marley, A.R. and Nan, H. Epidemiology of colorectal cancer. *International journal of molecular epidemiology and genetics.* 2016, **7**(3), pp.105-114.
631. Lee, J., Demissie, K., Lu, S.E. and Rhoads, G.G. Cancer incidence among Korean-American immigrants in the United States and native Koreans in South Korea. *Cancer Control.* 2007, **14**(1), pp.78-85.

632. Paszat, L., Sutradhar, R., Liu, Y., Baxter, N.N., Tinmouth, J. and Rabeneck, L. Risk of colorectal cancer among immigrants to Ontario, Canada. *BMC gastroenterology.* 2017, **17**(1), pp.85-85.
633. Thomas, D.B. and Karagas, M.R. Cancer in first and second generation Americans. *Cancer Res.* 1987, **47**(21), pp.5771-5776.
634. Vangay, P., Johnson, A.J., Ward, T.L., Al-Ghalith, G.A., Shields-Cutler, R.R., Hillmann, B.M., Lucas, S.K., Beura, L.K., Thompson, E.A., Till, L.M., Batres, R., Paw, B., Pergament, S.L., Saenyakul, P., Xiong, M., Kim, A.D., Kim, G., Masopust, D., Martens, E.C., Angkurawaranon, C., McGready, R., Kashyap, P.C., Culhane-Pera, K.A. and Knights, D. US Immigration Westernizes the Human Gut Microbiome. *Cell.* 2018, **175**(4), pp.962-972.e910.
635. Blaser, M.J. and Falkow, S. What are the consequences of the disappearing human microbiota? *Nature Reviews Microbiology.* 2009, **7**, p.887.
636. Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, J.S., Schlegel, M.L., Tucker, T.A., Schrenzel, M.D., Knight, R. and Gordon, J.I. Evolution of mammals and their gut microbes. *Science.* 2008, **320**(5883), pp.1647-1651.
637. Moeller, A.H., Caro-Quintero, A., Mjungu, D., Georgiev, A.V., Lonsdorf, E.V., Muller, M.N., Pusey, A.E., Peeters, M., Hahn, B.H. and Ochman, H. Cospeciation of gut microbiota with hominids. *Science.* 2016, **353**(6297), p.380.
638. Moeller, A.H., Li, Y., Mpoudi Ngole, E., Ahuka-Mundeke, S., Lonsdorf, E.V., Pusey, A.E., Peeters, M., Hahn, B.H. and Ochman, H. Rapid changes in the gut microbiome during human evolution. *Proc Natl Acad Sci U S A.* 2014, **111**(46), pp.16431-16435.
639. Tito, R.Y., Macmil, S., Wiley, G., Najar, F., Cleeland, L., Qu, C., Wang, P., Romagne, F., Leonard, S., Ruiz, A.J., Reinhard, K., Roe, B.A. and Lewis, C.M., Jr. Phylotyping and Functional Analysis of Two Ancient Human Microbiomes. *PloS one.* 2008, **3**(11), p.e3703.
640. Tito, R.Y., Knights, D., Metcalf, J., Obregon-Tito, A.J., Cleeland, L., Najar, F., Roe, B., Reinhard, K., Sobolik, K., Belknap, S., Foster, M., Spicer, P., Knight, R. and Lewis, C.M., Jr. Insights from Characterizing Extinct Human Gut Microbiomes. *PloS one.* 2012, **7**(12), p.e51146.
641. Schnorr, S.L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., Fiori, J., Gotti, R., De Bellis, G., Luiselli, D., Brigidi, P., Mabulla, A., Marlowe, F., Henry, A.G. and Crittenden, A.N. Gut microbiome of the Hadza hunter-gatherers. *Nature communications.* 2014, **5**, p.3654.
642. Rampelli, S., Schnorr, S.L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A.N., Henry, A.G. and Candela, M. Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr Biol.* 2015, **25**(13), pp.1682-1693.
643. Fragiadakis, G.K., Smits, S.A., Sonnenburg, E.D., Van Treuren, W., Reid, G., Knight, R., Manjurano, A., Changalucha, J., Dominguez-Bello, M.G., Leach, J. and Sonnenburg, J.L. Links between environment, diet, and the hunter-gatherer microbiome. *Gut microbes.* 2019, **10**(2), pp.216-227.
644. Obregon-Tito, A.J., Tito, R.Y., Metcalf, J., Sankaranarayanan, K., Clemente, J.C., Ursell, L.K., Zech Xu, Z., Van Treuren, W., Knight, R.,

Gaffney, P.M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarroel, O., Foster, M., Guija-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A.T. and Lewis, C.M. Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature communications.* 2015, **6**, p.6505.

645. Gomez, A., Petrzelkova, K.J., Burns, M.B., Yeoman, C.J., Amato, K.R., Vlckova, K., Modry, D., Todd, A., Jost Robinson, C.A., Remis, M.J., Torralba, M.G., Morton, E., Umana, J.D., Carbonero, F., Gaskins, H.R., Nelson, K.E., Wilson, B.A., Stumpf, R.M., White, B.A., Leigh, S.R. and Blekhman, R. Gut Microbiome of Coexisting BaAka Pygmies and Bantu Reflects Gradients of Traditional Subsistence Patterns. *Cell Rep.* 2016, **14**(9), pp.2142-2153.
646. Clemente, J.C., Pehrsson, E.C., Blaser, M.J., Sandhu, K., Gao, Z., Wang, B., Magris, M., Hidalgo, G., Contreras, M., Noya-Alarcon, O., Lander, O., McDonald, J., Cox, M., Walter, J., Oh, P.L., Ruiz, J.F., Rodriguez, S., Shen, N., Song, S.J., Metcalf, J., Knight, R., Dantas, G. and Dominguez-Bello, M.G. The microbiome of uncontacted Amerindians. *Sci Adv.* 2015, **1**(3).
647. Marini, E., Maldonado-Contreras, A.L., Cabras, S., Hidalgo, G., Buffa, R., Marin, A., Floris, G., Racugno, W., Pericchi, L.R., Castellanos, M.E., Groschl, M., Blaser, M.J. and Dominguez-Bello, M.G. Helicobacter pylori and intestinal parasites are not detrimental to the nutritional status of Amerindians. *Am J Trop Med Hyg.* 2007, **76**(3), pp.534-540.
648. Jha, A.R., Davenport, E.R., Gautam, Y., Bhandari, D., Tandukar, S., Ng, K.M., Fragiadakis, G.K., Holmes, S., Gautam, G.P., Leach, J., Sherchand, J.B., Bustamante, C.D. and Sonnenburg, J.L. Gut microbiome transition across a lifestyle gradient in Himalaya. *PLoS biology.* 2018, **16**(11), p.e2005396.
649. Winglee, K., Howard, A.G., Sha, W., Gharaibeh, R.Z., Liu, J., Jin, D., Fodor, A.A. and Gordon-Larsen, P. Recent urbanization in China is correlated with a Westernized microbiome encoding increased virulence and antibiotic resistance genes. *Microbiome.* 2017, **5**(1), pp.121-121.
650. Tyakht, A.V., Kostryukova, E.S., Popenko, A.S., Belenikin, M.S., Pavlenko, A.V., Larin, A.K., Karpova, I.Y., Selezneva, O.V., Semashko, T.A., Ospanova, E.A., Babenko, V.V., Maev, I.V., Cheremushkin, S.V., Kucheryavyy, Y.A., Shcherbakov, P.L., Grinevich, V.B., Efimov, O.I., Sas, E.I., Abdulkhakov, R.A., Abdulkhakov, S.R., Lyalyukova, E.A., Livzan, M.A., Vlassov, V.V., Sagdeev, R.Z., Tsukanov, V.V., Osipenko, M.F., Kozlova, I.V., Tkachev, A.V., Sergienko, V.I., Alexeev, D.G. and Govorun, V.M. Human gut microbiota community structures in urban and rural populations in Russia. *Nature communications.* 2013, **4**, p.2469.
651. Stagaman, K., Cepon-Robins, T.J., Liebert, M.A., Gildner, T.E., Urlacher, S.S., Madimenos, F.C., Guillemin, K., Snodgrass, J.J., Sugiyama, L.S. and Bohannan, B.J.M. Market Integration Predicts Human Gut Microbiome Attributes across a Gradient of Economic Development. *mSystems.* 2018, **3**(1), pp.e00122-00117.
652. Zhang, J., Guo, Z., Lim, A.A., Zheng, Y., Koh, E.Y., Ho, D., Qiao, J., Huo, D., Hou, Q., Huang, W., Wang, L., Javzandulam, C., Narangerel, C., Jirimutu, Menghebilige, Lee, Y.K. and Zhang, H. Mongolians core

gut microbiota and its correlation with seasonal dietary changes. *Sci Rep.* 2014, **4**, p.5001.

653. Katsidzira, L., Ocvirk, S., Wilson, A., Li, J., Mahachi, C.B., Soni, D., DeLany, J., Nicholson, J.K., Zoetendal, E.G. and O'Keefe, S.J.D. Differences in Fecal Gut Microbiota, Short-Chain Fatty Acids and Bile Acids Link Colorectal Cancer Risk to Dietary Changes Associated with Urbanization Among Zimbabweans. *Nutr Cancer.* 2019, pp.1-12.
654. Martinez, I., Stegen, J.C., Maldonado-Gomez, M.X., Eren, A.M., Siba, P.M., Greenhill, A.R. and Walter, J. The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes. *Cell Rep.* 2015, **11**(4), pp.527-538.
655. Lin, A., Bik, E.M., Costello, E.K., Dethlefsen, L., Haque, R., Relman, D.A. and Singh, U. Distinct distal gut microbiome diversity and composition in healthy children from Bangladesh and the United States. *PloS one.* 2013, **8**(1), p.e53838.
656. De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G. and Lionetti, P. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci U S A.* 2010, **107**(33), pp.14691-14696.
657. Suzuki, T.A. and Worobey, M. Geographical variation of human gut microbial composition. *Biol Lett.* 2014, **10**(2), p.20131037.
658. Mancabelli, L., Milani, C., Lugli, G.A., Turroni, F., Ferrario, C., van Sinderen, D. and Ventura, M. Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. *Environ Microbiol.* 2017, **19**(4), pp.1379-1390.
659. Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D. and Finn, R.D. A new genomic blueprint of the human gut microbiota. *Nature.* 2019, **568**(7753), pp.499-504.
660. O'Keefe, S.J.D., Ou, J., Aufreiter, S., O'Connor, D., Sharma, S., Sepulveda, J., Fukuwatari, T., Shibata, K. and Mawhinney, T. Products of the colonic microbiota mediate the effects of diet on colon cancer risk. *Journal of Nutrition.* 2009, **139**(11), pp.2044-2048.
661. O'Keefe, S.J., Li, J.V., Lahti, L., Ou, J., Carbonero, F., Mohammed, K., Posma, J.M., Kinross, J., Wahl, E., Ruder, E., Vipperla, K., Naidoo, V., Mtshali, L., Tims, S., Puylaert, P.G., DeLany, J., Krasinskas, A., Benefiel, A.C., Kaseb, H.O., Newton, K., Nicholson, J.K., de Vos, W.M., Gaskins, H.R. and Zoetendal, E.G. Fat, fibre and cancer risk in African Americans and rural Africans. *Nat Commun.* 2015, **6**, p.6342.
662. Allali, I., Boukhatem, N., Bouguenouch, L., Hardi, H., Boudouaya, H.A., Cadenas, M.B., Ouldim, K., Amzazi, S., Azcarate-Peril, M.A. and Ghazal, H. Gut microbiome of Moroccan colorectal cancer patients. *Medical Microbiology and Immunology.* 2018, **207**(3-4), pp.211-225.
663. Alomair, A.O., Masoodi, I., Alyamani, E.J., Allehibi, A.A., Qutub, A.N., Alsayari, K.N., Altammami, M.A. and Alshanqeeti, A.S. Colonic Mucosal Microbiota in Colorectal Cancer: A Single-Center Metagenomic Study in Saudi Arabia. *Gastroenterol Res Pract.* 2018, **2018**, p.5284754.
664. Faruk, M., Ibrahim, S., Adamu, A., Rafindadi, A.H., Ukwenya, Y., Iliyasu, Y., Adamu, A., Aminu, S.M., Shehu, M.S., Ameh, D.A., Mohammed, A., Ahmed, S.A., Idoko, J., Ntekim, A., Suleiman, A.M.,

Shah, K.Z. and Adoke, K.U. An analysis of dietary fiber and fecal fiber components including pH in rural Africans with colorectal cancer. *Intest Res.* 2018, **16**(1), pp.99-108.

665. Bamola, V.D., Ghosh, A., Kapardar, R.K., Lal, B., Cheema, S., Sarma, P. and Chaudhry, R. Gut microbial diversity in health and disease: experience of healthy Indian subjects, and colon carcinoma and inflammatory bowel disease patients. *Microb Ecol Health Dis.* 2017, **28**(1), p.1322447.
666. Loke, M.F., Chua, E.G., Gan, H.M., Thulasi, K., Wanyiri, J.W., Thevambiga, I., Goh, K.L., Wong, W.F. and Vadivelu, J. Metabolomics and 16S rRNA sequencing of human colorectal cancers and adjacent mucosa. *PloS one.* 2018, **13**(12), p.e0208584.
667. He, Y., Wu, W., Zheng, H.-M., Li, P., McDonald, D., Sheng, H.-F., Chen, M.-X., Chen, Z.-H., Ji, G.-Y., Zheng, Z.-D.-X., Mujagond, P., Chen, X.-J., Rong, Z.-H., Chen, P., Lyu, L.-Y., Wang, X., Wu, C.-B., Yu, N., Xu, Y.-J., Yin, J., Raes, J., Knight, R., Ma, W.-J. and Zhou, H.-W. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nature medicine.* 2018, **24**(10), pp.1532-1535.
668. Wu, G.D., Compher, C., Chen, E.Z., Smith, S.A., Shah, R.D., Bittinger, K., Chehoud, C., Albenberg, L.G., Nessel, L., Gilroy, E., Star, J., Weljie, A.M., Flint, H.J., Metz, D.C., Bennett, M.J., Li, H., Bushman, F.D. and Lewis, J.D. Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. *Gut.* 2016, **65**(1), pp.63-72.
669. Sierra, M.S. and Forman, D. Burden of colorectal cancer in Central and South America. *Cancer epidemiology.* 2016, **44**, pp.S74-S81.
670. Vuong, D.A., Velasco-Garrido, M., Lai, T.D. and Busse, R. Temporal trends of cancer incidence in Vietnam, 1993-2007. *Asian Pac J Cancer Prev.* 2010, **11**(3), pp.739-745.
671. Pathy, S., Lambert, R., Sauvaget, C. and Sankaranarayanan, R. The incidence and survival rates of colorectal cancer in India remain low compared with rising rates in East Asia. *Dis Colon Rectum.* 2012, **55**(8), pp.900-906.
672. Chung, R.Y., Tsoi, K.K.F., Kyaw, M.H., Lui, A.R., Lai, F.T.T. and Sung, J.J. A population-based age-period-cohort study of colorectal cancer incidence comparing Asia against the West. *Cancer Epidemiol.* 2019, **59**, pp.29-36.
673. Ferlay J, E.M., Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I and Bray F. *Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer.* [Online]. 2018. [Accessed 23.7.19]. Available from: https://gco.iarc.fr/today
674. Magne, F., O'Ryan, M.L., Vidal, R. and Farfan, M. The human gut microbiome of Latin America populations: a landscape to be discovered. *Curr Opin Infect Dis.* 2016, **29**(5), pp.528-537.
675. Shetty, S.A., Marathe, N.P. and Shouche, Y.S. Opportunities and challenges for gut microbiome studies in the Indian population. *Microbiome.* 2013, **1**(1), p.24.
676. Dubey, A.K., Uppadhyaya, N., Nilawe, P., Chauhan, N., Kumar, S., Gupta, U.A. and Bhaduri, A. LogMPIE, pan-India profiling of the human

gut microbiome using 16S rRNA sequencing. *Scientific data.* 2018, **5**, pp.180232-180232.

677. Chauhan, N.S., Pandey, R., Mondal, A.K., Gupta, S., Verma, M.K., Jain, S., Ahmed, V., Patil, R., Agarwal, D., Girase, B., Shrivastava, A., Mobeen, F., Sharma, V., Srivastava, T.P., Juvekar, S.K., Prasher, B., Mukerji, M. and Dash, D. Western Indian Rural Gut Microbial Diversity in Extreme Prakriti Endo-Phenotypes Reveals Signature Microbes. *Frontiers in Microbiology.* 2018, **9**, pp.118-118.
678. Dhakan, D.B., Maji, A., Sharma, A.K., Saxena, R., Pulikkan, J., Grace, T., Gomez, A., Scaria, J., Amato, K.R. and Sharma, V.K. The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *GigaScience.* 2019, **8**(3), p.giz004.
679. Das, B., Ghosh, T.S., Kedia, S., Rampal, R., Saxena, S., Bag, S., Mitra, R., Dayal, M., Mehta, O., Surendranath, A., Travis, S.P.L., Tripathi, P., Nair, G.B. and Ahuja, V. Analysis of the Gut Microbiome of Rural and Urban Healthy Indians Living in Sea Level and High Altitude Areas. *Scientific reports.* 2018, **8**(1), p.10104.
680. Bhute, S., Pande, P., Shetty, S.A., Shelar, R., Mane, S., Kumbhare, S.V., Gawali, A., Makhani, H., Navandar, M., Dhotre, D., Lubree, H., Agarwal, D., Patil, R., Ozarkar, S., Ghaskadbi, S., Yajnik, C., Juvekar, S., Makharia, G.K. and Shouche, Y.S. Molecular Characterization and Meta-Analysis of Gut Microbial Communities Illustrate Enrichment of Prevotella and Megasphaera in Indian Subjects. *Front Microbiol.* 2016, **7**, p.660.
681. Tandon, D., Haque, M.M., R, S., Shaikh, S., P, S., Dubey, A.K. and Mande, S.S. A snapshot of gut microbiota of an adult urban population from Western region of India. *PloS one.* 2018, **13**(4), p.e0195643.
682. Dehingia, M., Devi, K.T., Talukdar, N.C., Talukdar, R., Reddy, N., Mande, S.S., Deka, M. and Khan, M.R. Gut bacterial diversity of the tribes of India and comparison with the worldwide data. *Scientific reports.* 2015, **5**, pp.18563-18563.
683. Carbonetto, B., Fabbro, M.C., Sciara, M., Seravalle, A., Méjico, G., Revale, S., Romero, M.S., Brun, B., Fay, M., Fay, F. and Vazquez, M.P. Human Microbiota of the Argentine Population- A Pilot Study. *Frontiers in Microbiology.* 2016, **7**, pp.51-51.
684. Fujio-Vejar, S., Vasquez, Y., Morales, P., Magne, F., Vera-Wolf, P., Ugalde, J.A., Navarrete, P. and Gotteland, M. The Gut Microbiota of Healthy Chilean Subjects Reveals a High Abundance of the Phylum Verrucomicrobia. *Frontiers in Microbiology.* 2017, **8**, pp.1221-1221.
685. Escobar, J.S., Klotz, B., Valdes, B.E. and Agudelo, G.M. The gut microbiota of Colombians differs from that of Americans, Europeans and Asians. *BMC Microbiol.* 2014, **14**, p.311.
686. Belforte, F.S., Fernandez, N., Tonin Monzon, F., Rosso, A.D., Quesada, S., Cimolai, M.C., Millan, A., Cerrone, G.E., Frechtel, G.D., Burcelin, R., Coluccio Leskow, F. and Penas-Steinhardt, A. Getting to Know the Gut Microbial Diversity of Metropolitan Buenos Aires Inhabitants. *Front Microbiol.* 2019, **10**, p.965.
687. Yang, T.W., Lee, W.H., Tu, S.J., Huang, W.C., Chen, H.M., Sun, T.H., Tsai, M.C., Wang, C.C., Chen, H.Y., Huang, C.C., Shiu, B.H., Yang, T.L., Huang, H.T., Chou, Y.P., Chou, C.H., Huang, Y.R., Sun, Y.R.,

Liang, C., Lin, F.M., Ho, S.Y., Chen, W.L., Yang, S.F., Ueng, K.C., Huang, H.D., Huang, C.N., Jong, Y.J. and Lin, C.C. Enterotype-based Analysis of Gut Microbiota along the Conventional Adenoma-Carcinoma Colorectal Cancer Pathway. *Sci Rep.* 2019, **9**(1), p.10923.

688. Kwok, L.-y., Zhang, J., Guo, Z., Gesudu, Q., Zheng, Y., Qiao, J., Huo, D. and Zhang, H. Characterization of Fecal Microbiota across Seven Chinese Ethnic Groups by Quantitative Polymerase Chain Reaction. *PloS one.* 2014, **9**(4), p.e93631.
689. Yazici, C., Wolf, P.G., Kim, H., Cross, T.L., Vermillion, K., Carroll, T., Augustus, G.J., Mutlu, E., Tussing-Humphreys, L., Braunschweig, C., Xicola, R.M., Jung, B., Llor, X., Ellis, N.A. and Gaskins, H.R. Race-dependent association of sulfidogenic bacteria with colorectal cancer. *Gut.* 2017, **66**(11), pp.1983-1994.
690. Hester, C.M., Jala, V.R., Langille, M.G., Umar, S., Greiner, K.A. and Haribabu, B. Fecal microbes, short chain fatty acids, and colorectal cancer across racial/ethnic groups. *World J Gastroenterol.* 2015, **21**(9), pp.2759-2769.
691. Greenhill, A.R., Tsuji, H., Ogata, K., Natsuhara, K., Morita, A., Soli, K., Larkins, J.A., Tadokoro, K., Odani, S., Baba, J., Naito, Y., Tomitsuka, E., Nomoto, K., Siba, P.M., Horwood, P.F. and Umezaki, M. Characterization of the gut microbiota of Papua New Guineans using reverse transcription quantitative PCR. *PloS one.* 2015, **10**(2), p.e0117427.
692. Knapp, G.C., Sharma, A., Olopade, B., Alatise, O.I., Olasehinde, O., Arije, O.O., Castle, P.E. and Kingham, T.P. An Exploratory Analysis of Fecal Immunochemical Test Performance for Colorectal Cancer Screening in Nigeria. *World J Surg.* 2019, p.[no pagination].
693. Gomez-Moreno, R., Gonzalez-Pons, M., Soto-Salgado, M., Cruz-Correa, M. and Baerga-Ortiz, A. The Presence of Gut Microbial Genes Encoding Bacterial Genotoxins or Pro-Inflammatory Factors in Stool Samples from Individuals with Colorectal Neoplasia. *Diseases.* 2019, **7**(1), p.E16.
694. Schloss, P.D. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *mBio.* 2018, **9**(3), pp.e00525-00518.
695. Siegel, R.L., Fedewa, S.A., Anderson, W.F., Miller, K.D., Ma, J., Rosenberg, P.S. and Jemal, A. Colorectal Cancer Incidence Patterns in the United States, 1974-2013. *J Natl Cancer Inst.* 2017, **109**(8), p.[no pagination].
696. Kim, T.J., Kim, E.R., Hong, S.N., Chang, D.K. and Kim, Y.H. Long-Term Outcome and Prognostic Factors of Sporadic Colorectal Cancer in Young Patients: A Large Institutional-Based Retrospective Study. *Medicine (Baltimore).* 2016, **95**(19), p.e3641.
697. Silla, I.O., Rueda, D., Rodríguez, Y., García, J.L., de la Cruz Vigo, F. and Perea, J. Early-onset colorectal cancer: a separate subset of colorectal cancer. *World journal of gastroenterology.* 2014, **20**(46), pp.17288-17296.
698. Weinberg, B.A. and Marshall, J.L. Colon Cancer in Young Adults: Trends and Their Implications. *Curr Oncol Rep.* 2019, **21**(1), p.3.
699. Mori, G., Orena, B.S., Cultrera, I., Barbieri, G., Albertini, A.M., Ranzani, G.N., Carnevali, I., Tibiletti, M.G. and Pasca, M.R. Gut Microbiota

Analysis in Postoperative Lynch Syndrome Patients. *Front Microbiol.* 2019, **10**, p.1746.