Gut *Bacteroides* viruses and the ME/CFS microbiota

Fiona Newberry

Thesis submitted for the degree of Doctor of Philosophy (PhD)

University of East Anglia

Quadram Institute

November 2021

Table of Contents

Table of Tables ..... *v*

Table of Figures ..... *vi*

Abbreviations ..... *viii*

Preface ..... *xi*

Statement of Acknowledgement of the Contribution of Others ..... *xii*

Acknowledgements ..... *xiii*

Chapter 1 : General introduction *1*

1.1 The human microbiota **1**
1.1.1 Development of the intestinal microbiota 2
1.1.2 Adult intestinal microbiota in health 5
1.1.3 Intestinal microbiota in disease 10
1.1.4 Investigating the intestinal microbiota 13

1.2 Myalgic encephalomyelitis/Chronic Fatigue Syndrome **17**
1.2.1 Gut origin of ME/CFS 18

1.3 Influence of *Bacteroides* species **21**

1.4 Aims and objectives **23**

1.5 References **23**

Chapter 2 : Analysis of faecal gut microbiota in severe Myalgic Encephalomyelitis/Chronic Fatigue Syndrome *40*

2.1 Aims and objectives **40**

2.2 Methods **41**
2.2.1 Patient selection and recruitment 41
2.2.2 Faecal DNA extraction 42
2.2.3 Metagenomic sequencing 42
2.2.4 Metagenomic data processing 42
2.2.5 Statistical analysis 44

2.3 Results **45**
2.3.1 Patient demographics 45
2.3.2 Microbial gene richness 46
2.3.3 Taxonomic abundance 47
2.3.4 Functional analysis 66
2.3.5 Metagenome-assembled genomes (MAGs) 68

2.4 Discussion **73**

2.5 References **78**

Chapter 3 : Genome characterisation of Bacteroides fragilis bacteriophage vB_BfrS_23 and discovery of a novel B. fragilis phage family *82*

3.1 Introduction **82**
3.1.1 Unknown phage diversity 82
3.1.2 Phage discovery techniques 82
3.1.3 Phage phylogenetics 83

3.1.4 Virus-host prediction 84
3.1.5 Discovery of crAssphage 85
3.1.6 *Bacteroides* phage 86
3.1.7 Aims and objectives 87

3.2 Methods 88
3.2.1 Growth media constituents and buffers 88
3.2.2 Bacteriophage ϕB124-14 propagation and enumeration 88
3.2.3 *Bacteroides* strain growth dynamics 91
3.2.4 Environmental sample collection 91
3.2.5 Environmental phage screening 92
3.2.6 Phage DNA extraction 93
3.2.7 Phage sequencing 94
3.2.8 Phage physical characterisation 95
3.2.9 Phage genome assembly and annotation 97
3.2.10 Phage genome comparison 97
3.2.11 *Bacteroides* phage phylogeny 98
3.2.12 Analysis of novel *B. fragilis* phage family 101

3.3 Results 103
3.3.1 *Bacteroides* strain growth dynamics 103
3.3.2 Environmental phage screening 104
3.3.3 Phage characteristics 104
3.3.4 Phage genome characteristics and comparison 107
3.3.5 *Bacteroides* phage phylogeny 118
3.3.6 Analysis of novel *Bacteroides* phage taxonomic group 128

3.4 Discussion 144

3.5 References 147

Chapter 4 : Analysis of the Bacteroides fragilis pangenome 154

4.1 Introduction 154
4.1.1 Toxigenic and non-toxigenic strains 154
4.1.2 Potential virulence factors 155
4.1.3 Polysaccharide capsules and LPS in *B. fragilis* 156
4.1.4 *B. fragilis* prophage 158
4.1.5 Pangenome analysis of opportunistic pathogens/commensals 159
4.1.6 Aims and objectives 160

4.2 Methods 160
4.2.1 Characterisation of *B. fragilis* isolate GB-124 160
4.2.2 Selection of *B. fragilis* sequence data from literature 162
4.2.3 Collection of *B. fragilis* isolates from NCBI 162
4.2.4 Antimicrobial resistance and *Bacteroides fragilis* toxin 163
4.2.5 Generation of the pangenome 163
4.2.5.1 Phylogenetic analysis 164
4.2.6 Functional analysis of the pangenome 166
4.2.7 Analysis of co-evolving genes 166
4.2.8 Identification of prophage 168

4.3 Results 168
4.3.1 Genome characteristics of *B. fragilis* GB-124 168
4.3.2 Selection of *B. fragilis* genomes from literature 169
4.3.3 Collection of *B. fragilis* isolates from NCBI 178
4.3.4 AMR genes and BFT 187
4.3.5 *B. fragilis* pangenome 187
4.3.6 Gene cluster analysis 195
4.3.7 *rfb* gene analysis 198

4.3.8 Functional analyses of the pangenome 204
4.3.9 Analysis of co-evolving genes 208
4.3.10 Identification of prophage within *B. fragilis* genomes 208

4.4 Discussion 208

4.5 References 220

Chapter 5 : General discussion 229

5.1 Summary 229

5.2 References 236

Appendix 1 239

Appendix 2 259

Appendix 3 271

Appendix 4 272

Appendix 5 274

Appendix 6 282

Appendix 7 284

Appendix 8 287

Table of Tables

Table 2-1: Overview of patient information collected during this study 45
Table 2-2: R^2 and p value for PERMANOVA and PERDISP at all taxonomic levels 56
Table 2-3: Analysis of Similarities statistic (R) and p value 60
Table 2-4: Summary statistics for the high-quality MAGs generated in this study 69
Table 2-5: Closest phylogenetic relatives of the high-quality MAGs among the unified catalogue of genomes from the human gut microbiota 71
Table 2-6: Comparison of composition alterations in ME/CFS microbiota studies 75
Table 3-1: Media and solution recipes 89
Table 3-2: *Bacteroides* and related species and strains used for sample screening 90
Table 3-3: Sample type and collection site 92
Table 3-4: *B. fragilis* strains used for host range assay, subtype and isolation site 96
Table 3-5: Overview of publicly available *B. fragilis* phage genomes compared to vB_BfrS_23 (blastn) 98
Table 3-6: Publicly available *Bacteroides* phage 100
Table 3-7: Representative crAss-like phage and candidate genera 101
Table 3-8: Predicted coding regions and protein functions of phage vB_BfrS_23 109
Table 3-9: CheckV output of representative *Bacteroides* phage sequences and VC assignment 121
Table 3-10: CheckV quality control summary of VC_396 123
Table 3-11: CheckV quality control summary of VC_100 125
Table 3-12: NCBI reference phage used in proteomic phylogenetic tree 126
Table 3-13: Orthogroup ID and predicted protein function 135
Table 3-14: CheckV quality summary report of VC_358 138
Table 4-1: Fragilysin and fragipain protein information from NCBI used to screen genomes for Bft protein 164
Table 4-2: Protein information for each *rfb* used to create a blastp database for isolate *rfb* gene screening 167
Table 4-3: Non-clinical isolates with CheckM contamination percentage > 5 % 172
Table 4-4: Summary statistics generated from Roary pangenome analysis of non-clinical isolates 175
Table 4-5: Non-clinical isolates selected at random from each clade and subject they originated from 178
Table 4-6: *B. fragilis* genomes with ANI < 95 % against NCTC 9343^T 180
Table 4-7: CheckM output showing *B. fragilis* NCBI genomes with completeness < 90 % and contamination >5 % 180
Table 4-8: Genomes removed from further analysis according to PanarooQC due to number of genes or contigs as outliers 181
Table 4-9: Metadata for the 93 *B. fragilis* genomes used in the pangenome analysis 183
Table 4-10: Genomes encoding a *bft* gene and *bft* isotype 189
Table 4-11: Summary statistics generated from Roary pangenome analysis of 93 *B. fragilis* genomes 189
Table 4-12: Genomes belonging to each of the five outlying clusters identified in PCoA of accessory genes 193
Table 4-13: Genes identified in the majority of isolates in the main Cluster 195
Table 4-14: Unique genes identified in clusters 1,2 and 5 according to Blastp analysis 196
Table 4-15: Overview of 'missing' genes in each Cluster and the overlap between clusters 199
Table 4-16: Overview of predicted prophage identified in isolates using PhiSpy 209
Table 5-1: Overview of number of participants, diagnostic criteria, type of study and recruitment location for ME/CFS studies where the information is available 230

Table of Figures

Figure 2.1: Microbial gene richness of gut metagenomes of ME/CFS patients and controls....46
Figure 2.2: Microbiota relative abundance at phylum level for ME/CFS patients and controls....47
Figure 2.3: Microbiota relative abundance at class, order and family levels for ME/CFS patients and controls 49
Figure 2.4: Microbiota relative abundance at genus level for ME/CFS patients and controls 50
Figure 2.5: Microbiota relative abundance at phylum level for matched ME/CFS patients and controls....52
Figure 2.6: Microbiota relative abundance at class, order and family levels for matched ME/CFS patients and controls 53
Figure 2.7: Microbiota relative abundance at genus level for matched ME/CFS patients and controls 54
Figure 2.8: Shannon Index (A) and Simpson Index (B) for ME/CFS patients and controls 55
Figure 2.9: Phylum, class, order, family and genus boxplots showing the distance to the centroid for each ME/CFS patient and control datum point 57
Figure 2.10: Top contributing coefficients from PERMANOVA analysis for each taxonomic level 58
Figure 2.11: nMDS for all taxonomic levels of ME/CFS patients and controls 59
Figure 2.12: Correlation plots at class, order and family levels for ME/CFS patients and controls....62
Figure 2.13: Correlation plots at phylum level for ME/CFS patients and controls 63
Figure 2.14: Correlation plots of ME/CFS patients at genus level 64
Figure 2.15: Correlation plots of controls at genus level 65
Figure 2.16: KEGG pathway representation (L2) of metagenomes of ME/CFS patients and controls 66
Figure 2.17: PCoA of a Bray-Curtis dissimilarity matrix of L2 and L3 KEGG metabolic pathways for metagenomes of ME/CFS patients and controls....67
Figure 3.1: Growth curve of *Bacteroides* strains in BHI and BPRM broths....105
Figure 3.2: MICs of kanamycin with *Bacteroides* strains....106
Figure 3.3: Physical and biological characteristics of phage vB_BfrS_23 107
Figure 3.4:Genome characteristics of phage vB_BfrS_23 113
Figure 3.5: Genome comparison of phage vB_BfrS_23, ϕ124-14, Barc2635 and B40-8....114
Figure 3.6: Phylogeny of vB_Bfrs_23 large subunit terminase and associated metadata 116
Figure 3.7: Phylogeny of vB_BfrS_23 tail protein and associated metadata 117
Figure 3.8: Overview of gene-sharing network analysis of curated *Bacteroides* phage dataset....120
Figure 3.9:Proteomic phylogenetic tree of *Bacteroides* representative phage, reference phage, representative crAss-like phage and the VC (VC_100) of interest 127
Figure 3.10: Heatmap representing intergenomic similarity (%) within VC_100 and assigned genus....130
Figure 3.11: Proteomic phylogenetic tree generated from VC_100 131
Figure 3.12: Heatmap representing phage assignment to orthogroups and genus specificity....133
Figure 3.13: Heatmap and dendrogram representing shared orthologues within VC_100 134
Figure 3.14: Proteomic phylogenetic tree generated from VC_100 (red) and VC_358 (black) phage 137
Figure 3.15: Proteomic phylogenetic tree generated from VC_358 and uvig_314311....140
Figure 3.16: Genome comparison of phage ϕB124-14, uvig_314311 and IMGVR_UViG_461 141
Figure 3.17: Phylogenetic tree of VC_100, crAss-like phage and two *Cellulophaga* phage TerL proteins 143
Figure 4.1: Genome map of *B. fragilis* GB-124 170
Figure 4.2: Bandage map of GB-124 chromosome and plasmids 171
Figure 4.3: PanarooQC output from non-clinical pangenome quality control 174
Figure 4.4: PCoA of accessory genome of non-clinical isolates....176
Figure 4.5: Maximum likelihood phylogenetic tree generated from the core SNPs in the pangenome of the non-clinical isolates....177
Figure 4.6: ANI scores of all isolates compared to the *B. fragilis* reference genome (NCTC 9343^{T}) 179
Figure 4.7: PanarooQC output from pangenome quality control 182
Figure 4.8: AMR gene profile of each genome according to screening against CARD with Abricate....188
Figure 4.9: Number of conserved genes versus total genes in the *B. fragilis* pangenome 190
Figure 4.10: PCoA of accessory genome of the 93 *B. fragilis* genomes....191
Figure 4.11: Maximum likelihood phylogenetic tree generated from core SNPs from the pangenome analysis 194
Figure 4.12: Heatmap of *rfb* genes present within each isolate according to identification with Blastp....201
Figure 4.13: Annotated O-antigen nucleotide sugar biosynthesis KEGG pathway showing *rfb* gene involvement 203

Figure 4.14: Genome arrangement of predicted PS loci of *B. fragilis* DCMOUH0042B, BOB25 and AD135F_2B 204
Figure 4.15: Proportion (%) of annotated COGs within the accessory and core genomes 205
Figure 4.16: Stacked bar chart showing the proportion of annotated COGs of unique genes within each pangenome cluster and assigned COG category 206
Figure 4.17: Stacked bar chart showing the proportion of annotated COGs of 'missing' genes within each pangenome Cluster and assigned COG category 207
Figure 5.1: Power calculations determined for number of patients (*n*=5, ν1) and controls (*n*=14, ν2) included in the ME/CFS study 231

Abbreviations

aa – amino acid

AAHC – non-*C. difficile* antibiotic-associated haemorrhagic colitis

AAI – average amino acid identity

AMR – antimicrobial resistance

ANI – average nucleotide identity

ANOSIM – analysis of similarities

ASD – autism spectrum disorder

Bf – *Bacteroides fragilis*

BfPAI – *Bacteroides fragilis* pathogenicity island

BFT – *Bacteroides fragilis* toxin

BHI – brain heart infusion

bp – base pair

BPRM – *Bacteroides* phage recovery medium

CARD – Comprehensive Antibiotic Resistance Database

CD – Crohn's disease

CDD – Conserved Domain Database

CDS – coding sequence

COG – cluster of orthologous group

CPS – capsular polysaccharide

CRISPR – clustered regularly interspaced short palindromic repeat

DMM – demethylmenaquinone methyltransferase

dsDNA – double-stranded DNA

dsRNA – double-stranded RNA

DTR – direct terminal repeat

ETBF – enterotoxigenic *B. fragilis*

FAB – fastidious anaerobe broth

FMT – faecal microbiota transplant

GI – gastrointestinal

GPD – Gut Phage Database

HADS – hospital anxiety depression scale

HFD – high-fat diet

HGT – horizontal gene transfer

HMO – human milk oligosaccharide

IBS – irritable bowel syndrome

IBD – inflammatory bowel disease

ICC – International Consensus Criteria

ICTV – International Committee on Taxonomy of Viruses

ImP – imidazole propionate

IS – insertion sequence

kbp – kilobase pair

KEGG – Kyoto Encyclopedia of Genes and Genomes

LOS – lipooligosaccharide

LPS – lipopolysaccharide

MAG – metagenome-assembled genome

MIUViG – minimum information about an uncultivated virus genome

MCP – major capsid protein

MDR – multi-drug-resistant

ME/CFS – myalgic encephalomyelitis/chronic fatigue syndrome

MIC – minimum inhibitory concentration

mRNA – messenger ribonucleic acid

NCBI – National Centre for Biotechnology Information

NICE – National Institute for Health and Care Excellence

nt – nucleotide

nMDS – non-metric multidimensional scaling

NTBF – non-toxigenic *Bacteroides fragilis*

OD_{620} – optical density at wavelength of 620 nm

OMV – outer membrane vesicle

ORF – open reading frame

OTU – operational taxonomic unit

PCoA – principal coordinate analysis

PERMANOVA – permutational multivariate analysis of variance

PERMDISP – permutational analysis of multivariate dispersions

PS – polysaccharide

PSA – polysaccharide A

PUL – polysaccharide utilisation locus

QIB – Quadram Institute Bioscience

REML – restricted maximum likelihood linear model

rRNA – ribosomal ribonucleic acid

SCFA – short-chain fatty acid

SNP – single nucleotide polymorphism

ssDNA – single-stranded DNA

ssRNA – single-stranded RNA

T1DM – type I diabetes mellitus

T2DM – type II diabetes mellitus

TEDDY – The Environmental Determinants of Diabetes in the Young

TEM – transmission electron microscopy

TerL – large terminase subunit

tRNA – transfer ribonucleic acid

UC – ulcerative colitis

US CDC – United States Centre for Disease Control

UviG – uncultivated viral genome

VFDB – Virulence Factor Database

VIRIDIC – virus intergenomic distance calculator

VC – viral cluster

VLP – virus-like particle

Preface

The human intestinal microbiome has long been associated with health and disease. Advances in sequencing and computational approaches have enabled more detailed studies investigating the intestinal microbiome. The intestinal microbiome has been implicated in the development of ME/CFS, a multi-faceted disease mainly characterised by persistent unexplained fatigue. A high percentage of patients exhibit irritable bowel syndrome-like symptoms; this has led researchers to investigate the intestinal microbiome as a contributing factor of disease onset. Chapter 2 details the differences in the microbiota between severe ME/CFS patients versus controls and the heterogenous microbiota composition within the patient group.

Additionally, several intestinal microbiota studies have highlighted the differing abundances of *Bacteroides* spp. within patients compared to controls. *Bacteroides* spp. play a pivotal role in the maturation of the infant gut microbiome and are believed to contribute towards a healthy adult microbiome. Several *Bacteroides* spp. have been shown to mediate immune tolerance and maintain inter-species relationships with other bacteria within the microbiome, contributing towards the overall health of the human host. Prokaryotic viruses (bacteriophage) are thought to indirectly influence human host health via the gut microbiota (taxonomic and functional alterations). However, the bacteriophage of *Bacteroides* spp. have not been extensively investigated compared to other medically relevant bacteria. Chapter 3 details the isolation and characterisation of a novel *Bacteroides fragilis* bacteriophage and discovery of a novel *B. fragilis* bacteriophage family through gene-sharing network analysis.

Several *Bacteroides* spp. are opportunistic pathogens due to their ability to cause extraintestinal infection. Specific *B. fragilis* isolates can cause intestinal inflammation via secretion of an enterotoxin. The population structure in relation to different *B. fragilis* types (i.e. enterotoxigenic, clinical, non-clinical) has not been studied to date. Chapter 4 shows the pangenome of phenotypically diverse *B. fragilis* isolates and reveals a large accessory genome with no clustering according to isolate type.

Statement of Acknowledgement of the Contribution of Others

Newberry F, Hsieh SY, Wileman T, Carding SR. Does the microbiome and virome contribute to myalgic encephalomyelitis/chronic fatigue syndrome? Clin Sci (Lond). 2018 Mar 9;132(5):523-542. doi: 10.1042/CS20171330. PMID: 29523751; PMCID: PMC5843715.

Refer to Appendix 1

The information in this review was used for part of Chapter 1. The three co-authors of this review proofread and offered comments on the publication prior to submission. However, the research and bulk of the writing was undertaken by me. Some of the metagenomic sequencing analysis (Section 2.2.2) in Chapter 2 was performed by Professor Lesley Hoyles. The metagenomic sequencing quality control, taxonomic annotation, functional annotation, microbial gene richness, metagenome assembly and generation of MAGs was analysed by Professor Lesley Hoyles. The statistical analysis, interpretation of data and generation of figures was performed by myself.

Tariq MA, Newberry F, Haagmans R, Booth C, Wileman T, Hoyles L, Clokie MRJ, Ebdon J, Carding SR. Genome characterization of a novel wastewater *Bacteroides fragilis* bacteriophage (vB_BfrS_23) and its host GB124. Front Microbiol. 2020 Oct 23;11:583378. doi: 10.3389/fmicb.2020.583378. PMID: 33193224; PMCID: PMC7644841.

Refer to Appendix 2

Some of the results from this publication appear in Chapter 3 and Chapter 4. All authors contributed towards writing and proofreading the manuscript. All computational analysis was undertaken by me and Dr Tariq with guidance from Professor Lesley Hoyles. Furthermore, several of the results in Chapter 3 and Chapter 4 were generated in collaboration with Dr. Adnan Tariq and Rik Haagmans. In Chapter 3, Dr Tariq and I isolated, characterised and sequenced the phage genome together with both having an equal role in the generation of data. Additionally, Dr. Tariq performed the one-step growth curve and phage pH tolerance experiments. Myself, Rik Haagmans and Dr Tariq generated the data for the phage temperature assays. Dr Catharine Booth imaged the phage and edited any images for publication. The data generated from Section 3.2.11 (Bacteroides phage phylogeny) was performed by myself. Dr James Ebdon provided the host bacterium that the phage was isolated with. In Chapter 4, Dr Tariq and I sequenced the host bacterium and characterised the genome. All other data in this chapter was generated by myself.

Acknowledgements

Firstly, I'd like to thank my supervisors Professor Simon Carding and Professor Tom Wileman for their continued guidance during this project. I'd also like to express gratitude to Dr Mohammed Adnan Tariq and the whole Carding lab group (past and present) for their encouragement and insightful comments.

Besides from my supervisors, I would like to thank Professor Shelia Patrick, Dr Maria Rosa Doming-Sananes, Dr Evelien Adriaenssens and Dr Lesley Ogilvie for sharing their knowledge and answering any questions I had. I am also grateful to my family and son for offering their support and encouragement, especially during the tough write-up time.

Lastly, the completion of my thesis would not have been possible without the unwavering support and guidance of Professor Lesley Hoyles. Her endless patience and assistance greatly helped throughout the project, particularly during thesis writing.

Chapter 1 : General introduction

1.1 The human microbiota

Microbial cells colonise almost every surface of the human body and are believed to be as abundant as somatic cells[1,2]. The true number of microbial species co-existing on/in the human body is unknown but it is estimated 500-1000 bacteria species are present at any one time[3]. The human microbiota is composed of a wide variety of bacteria, eukaryotic viruses, prokaryotic viruses, protozoa, archaea and fungi[4-8]. Each body site displays a unique composition of microbes, even if present on the same body surface[9-11]. For example, skin physiology highly influences the microbial species present[12]. Lipophilic taxa, such as *Propionibacterium* species, are highly abundant in sebaceous areas of the skin[13]. Whereas humidity-loving bacteria, such as *Staphylococcus* and *Corynebacterium* spp., are dominant in moist regions like the feet or back of the knees[9,11,14]. The human microbiota is a constantly fluctuating ecosystem that is influenced by extrinsic (e.g. lifestyle, diet, medications, birth mode) and intrinsic (e.g. genetics, and local pH, nutrients and oxygen availability) factors[15-18]. However, there are microbes within the microbiota that are maintained over a lifetime and the ecosystem remains robust to perturbation[14,19,20]. Several studies have attempted to identify the 'core' human microbiota, a microbial profile that is similar between individuals[21,22]. A 2009 study examined bacterial diversity of 27 body sites from seven individuals at four time points[9]. A high interpersonal variability was found across all body sites, but individuals experienced minimal temporal diversity. The Human Microbiome Project sampled 242 healthy adults at 18 body sites and discovered that each habitat is characterised by a small number of highly abundant taxa but the relative abundance of these taxa varies between individuals[23]. A longitudinal study sampled two individuals at four body sites over 396 time points (~ 15 months) and reported stable differences between body sites and individuals. Variability in individual body sites was observed across months, weeks, and days. This indicates that the complexity and temporal variability of microbial communities are site-dependent and that the microbiota is highly variable within and between individuals. An ever-increasing number of studies suggest that the microbiota plays a significant role in the maintenance of human health and development of disease[24-27]. However, it is still unknown if alternations in microbiota composition are causative or simply correlated with disease.

Interest in the human microbiota has increased greatly, particularly due to advancements in sequencing technology and bioinformatic tools[28]. The importance of the human microbiota in health and disease is apparent but its true role is still unknown. There are complex microbe-

microbe and host-microbe interactions within each human microbiota that researchers are just starting to understand. Much of the research to date has focused on the intestinal (faecal) microbiota as this consortium of microbes is believed to have the greatest influence on human health.

1.1.1 Development of the intestinal microbiota

The first years of life represent a crucial window for intestinal microbiota development. A large body of research suggests that the establishment and maturation of the intestinal microbiota in the first 1000 days of life are critical[29]. The environmental influences within this short time can affect the intestinal microbiota through adulthood and may contribute to lifelong health and disease incidence[30,31].

Vertical transmission of microbes from the maternal microbiota is considered the most significant contribution to the infant microbiota[29,32]. During birth the infant's digestive tract, respiratory tract, urogenital tract, and skin are also colonised by microbes from the hospital and birthing environment[30]. It was previously believed that colonisation of the infant microbiota did not occur until birth, through passage via the birth canal or caesarean delivery (C-section), as the *in-utero* environment was sterile[33]. The idea of a placental microbiota is highly controversial; several studies have identified microbes within the placenta and faecal meconium prior to delivery[34,35]. The placental microbiota was found to contain commensal microbes from *Firmicutes, Tenericutes, Proteobacteria, Bacteroidetes* and *Fusobacteria.* This study also found similarities between the neonatal gut microbiota within the first 7 days of life and the placental microbiota, further suggesting that the *in-utero* environment is not sterile[35]. Additionally, the bacterial species found within the neonate's meconium sample and amniotic fluid were shared[36]. Therefore, the ingestion of amniotic fluid during development, especially during the 3rd trimester, may be seeding the infant microbiota prior to birth[37,38]. However, it is believed the detection of microbes within the placenta is due to bacterial contamination (laboratory reagents or delivery of the placenta). The theory has been further discounted by a 2021 study of 76 full-term pregnancies that found no evidence of a placental microbiota[39].

The intestinal microbiotas of infants born via vaginal delivery and C-section have differing microbial profiles[40-42]. Depending on mode of delivery, the infant microbiota is similar to the maternal stool, vagina, and skin microbiota. The intestinal microbiota of vaginally delivered infants contains microbes associated with the maternal vaginal microbiota, such as *Prevotella* and *Lactobacillus*[42-44]. Whereas the intestinal microbiota of C-section infants is reflective of the

maternal skin microbiota, comprising bacteria such as *Propionibacterium, Clostridium, Staphylococcus* and *Corynebacterium*. Additionally, these infants have a decreased abundance of anaerobes, particularly *Bacteroides* and *Bifidobacterium,* compared to their vaginally delivered counterparts[42,45-48]. C-section infant microbiotas have been shown to share a closer similarity to the hospital environment's microbe profile compared to vaginally birthed infants and are more likely to harbour antimicrobial-resistant pathogens[42,49,50]. Furthermore, it is estimated geographical differences may influence the maternal vaginal microbiota and, by association, the first colonising microbes[51,52]. The vaginal microbiota during pregnancy in urbanised high-income locations is dominated by *Lactobacillus*[53,54]. However, a study reported a high occurrence of *Lactobacillus*-deficient vaginal microbiotas in rural Malawian women; suggesting that external factors affecting the mother may also affect the infant microbiota[55]. Therefore, mode of delivery can significantly affect the colonisation of the microbiota, which can persist for months, or even years. C-section infants may exhibit delayed gut colonisation by *Bacteroides* spp. that can persist for up to a year after birth[56]. Additional studies have reported intestinal microbiota differences between delivery modes in children as old as 7 years of age[57]. The gut microbiota in preterm infants shows less stability compared to full-term infants and exhibits delays until an adult microbiota is established[58]. Additionally, preterm infant microbiotas show reduced microbial diversity and increased colonisation by pathogenic organisms[47,59]. A recent study reported higher abundance of facultative anaerobes, such as *Enterococcus, Enterobacter* and *Lactobacillus*, and decreased prevalence of obligate anaerobes (*Bifidobacterium* and *Bacteroides* spp.) in preterm infants compared to their full-term counterparts[60]. Additionally, full-term breastfed infants are colonised by *Bifidobacterium* spp. at day 7 of life, whereas the same is not seen in preterm infants[61].

During the first 6 months of life, facultative anaerobes are commonly the first colonisers of the infant gut microbiota followed by obligate anaerobes, such as *Bacteroides, Bifidobacterium,* and *Clostridium* spp.[62-64]. Microbial diversity is relatively low and in breastfed babies contains a high prevalence of microbes involved in metabolism of human milk oligosaccharides (HMOs). Additionally, it is thought that breastmilk introduces 25-30 % of all bacteria to the infant microbiota[65]. Successful establishment of the microbiota within the first couple years of life is imperative for development of functioning mucosal immunity and the endocrine and central nervous systems[66-68]. Breastmilk contains a plethora of carbohydrates, fatty acids, nutrients, anti-inflammatory proteins (e.g. lactoferrin) and maternal immune cells (e.g. IgA) essential for infant survival and microbiota development[69]. Several constituents are thought to promote *Bifidobacterium* growth, such as glycoconjugates and oligosaccharides, and prevent enteric

pathogen infection. For example, pathogen binding to host cells is thought to be prevented by various milk oligosaccharides and HMOs have been proven to interact directly with pathogenic bacteria[70,71]. Furthermore, supplementation of preterm infants with *Bifidobacterium* and *Lactobacillus* spp. decreased the abundance of pathobionts[72]. Several bacteria, such as *Staphylococcus, Streptococcus, Bifidobacterium,* and *Lactobacillus* spp., are thought to transfer from the maternal faecal microbiota to the breast milk through the enteromammary pathway; although this theory is somewhat controversial[73]. Growth of commensals is further promoted by fermentation of breastmilk-derived non-digestible carbohydrates in the colon[74]. In breastfed infants, transmission of maternal secretory IgA is thought to confer protection from infection by pathobionts and prevent the infant immune system from becoming overstimulated by microbes in the intestinal microbiota[74,75].

Maternal breastmilk and infant stool harbour viral assemblages that are significantly different from one another[76]. Infant faeces is dominated by *Siphoviridae* bacteriophage, whereas maternal breastmilk has a high prevalence of *Myoviridae* bacteriophage. These virus differences reflect bacterial composition within each sample. There are, however, a significant number of shared viruses between maternal breastmilk and infant faeces[76].

The intestinal microbiota of formula-fed infants has a different colonisation pattern to breastfed infants, mainly due to the alternate composition of infant formula compared to breast milk[77,78]. For example, oligosaccharides within infant formula are structurally different from HMOs, and therefore unlikely to play the exact role HMOs do in breastfed infants[79]. Formula-fed infants show a much more diverse microbiota compared to breastfed infants, comprising *Escherichia coli, Clostridioides difficile, Bacteroides, Prevotella* and *Lactobacillus* spp.[74,80,81]. Whereas the microbiota of breastfed infants is dominated by *Bifidobacterium* and has reduced abundance of *Enterobacteriaceae*[81]. These differences in colonisation patterns between formula-fed and breastfed infant gut microbiotas is thought to affect host health throughout adulthood[48]. For example, the link between infant formula use and adulthood obesity has been suggested[82,83].

Following withdrawal of breast and/or formula milk from the diet and introduction of solid foods, the taxonomic and functional diversity of the infant intestinal microbiota increases rapidly[62]. *Bifidobacterium* abundances steadily decrease and other *Actinobacteria* and *Proteobacteria* members become dominant [84,85]. Over the first few years of life, the intestinal microbiota undergoes significant changes and reaches a state of relative stability by 3 years of age[62]. As well as mode of delivery and milk source, geographical location, antibiotic use and other medications,

family lifestyle and host genetics also influence microbiota colonisation patterns in the early years of life[15,86].

The most significant infant microbiota study to date, The Environmental Determinants of Diabetes in the Young (TEDDY) study, examined longitudinal stool samples from 903 children between 3 and 46 months of age[15]. This study concluded that the progression of the infant intestinal microbiota occurs in three distinct phases (developmental, transitional, stable), where microbial diversity increases and dominant taxa shift until stability is achieved between months 31 and 46 of life. The receipt of breast milk was the most important factor associated with microbiota structure. Environmental factors, such as geographical location and household exposure (to siblings and/or pets), were also important contributors to microbiota composition[15].

The development and maintenance of the phageome (bacteriophage component of microbiota) within the first years of life has also been extensively studied[87-89]. As with the bacteriome, the phageome develops from infancy, particularly in the first 2 years of life. A longitudinal study examined faecal microbiota of twenty full-term infants and reported significantly differing phage profiles at 0 months and 24 months[90]. At 0 months, the authors observed low bacteria – high phage diversity and high bacteria – low phage diversity at 24 months. However, virus-like particles (VLPs) are almost undetectable in the infant meconium. The infant phageome exhibits higher diversity compared to the adult phageome, but it is considerably less stable[87]. The exact factors influencing intestinal phage colonisation are unclear but birth mode, feeding mode and weaning are believed to have less influence as seen with the bacterial components[87,91,92]. It has been suggested that the infant gut bacteriophage diversity is introduced via prophage induction in coloniser bacterial species[93]. A study estimated that approximately 63 % of bacteria are obtained from the mother during birth, whereas only 15 % of viruses were obtained via maternal transmission[91]. Furthermore, twin infants share more of their phageome than non-twin siblings, suggesting intrinsic and extrinsic factors influence colonisation[94]. During the first months of life, the phageome is dominated by *Caudovirales* and, by 2 years of age, has shifted to a *Microviridae*-abundant phageome.

1.1.2 Adult intestinal microbiota in health

The adult intestinal microbiota has been extensively studied due its importance in human health[2,17,18,22,23]. It has been attributed to a variety of roles that directly and indirectly benefit the human host, such as food digestion, nutrient extraction, host immune modification/modulation, host metabolism and pathogen protection[23-27,48]. The microbial composition of the gut microbiota

changes along the intestinal tract laterally within the lumen and vertically along the lumen, due to for example local nutrient, pH and oxygen conditions[19]. The microbial compositions in the small and large intestine differ significantly[95]. Although the small intestine is a nutrient-rich environment, the microbial density and diversity is relatively low mainly due to fast transit time and high pH[95]. Species belonging to *Lactobacillus, Streptococcus, Bacteroides, Veillonella* and *Clostridiales* spp. are dominant within the small intestine[96-98]. While the large intestine has a higher bacterial load, fermentation potential and abundance of obligate anaerobes[95,99]. The most common, and easier, method for studying the microbiota is sampling of faecal samples[100]. However, it should be noted that a faecal sample is not a true representation of the microbiota and contains microbes mainly residing within the colon and those sloughed off from other sections of the intestine during transit[101]. Several studies have attempted to biopsy the small intestine by recruiting patients undergoing invasive procedures such as colonoscopies, intestinal resections, or small-bowel transplantation[101,102]. However, samples obtained this way are subject to contamination from other sections of the intestine or oropharyngeal cavity. A similar issue is encountered for sampling of the large intestine further than the sigmoid, descending colon or the mucosal layer[103]. Additionally, the composition within a single stool sample can vary according to sample site[104].

A 'healthy' faecal microbiota is considered one that is highly diverse, with an abundance of obligate anaerobes belonging to the phyla *Bacteroidetes* and *Firmicutes*; however, researchers have yet to define what exactly a 'healthy' microbiota consists of[2,21-23]. The presence of a 'core' intestinal microbiota is disputed but the faeces of healthy adults is dominated by varying abundances and species belonging to the phyla *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, *Proteobacteria* and *Verrucomicrobia*[105,106]. It has been suggested that although the taxonomic composition of the faecal microbiota of metabolically healthy individuals is highly variable, the functional profile is shared and fulfilled by differing microbial communities[107]. Several studies have attempted to characterise a 'healthy' microbiota; however, this is extremely difficult given the numerous factors that can shape microbiota composition[16,106]. It has been further suggested that a healthy microbiota cannot be defined by the taxonomic profile but by its ability to maintain homeostasis, particularly at times of stress[107,108]. Furthermore, several metabolic diseases (obesity, type 2 diabetes mellitus, cardiometabolic disease, metabolic syndrome and liver disease) exhibit a clear reduction in functional richness within the intestinal microbiota versus metabolically healthy individuals. The reduction in number of unique microbial genes (i.e. reduced microbial gene richness) in a microbiota could influence disease presentation and outcome[26,109-111].

Several factors contribute towards the alteration and maintenance of a healthy microbiota, such as aging, immune status, host genetics, diet, and lifestyle[16]. It has been reported that with advancing age the proportions of *Firmicutes*, *Bifidobacterium* and *Faecalibacterium prausnitzii* decrease and abundances of *Escherichia coli*, *Proteobacteria* and staphylococci increase[112]. The microbiota of elderly individuals has been described as exhibiting a proinflammatory phenotype due to the higher potential for immune-system weakening and lower potential for vitamin B12 synthesis[112,113]. Though it is important to note that the intestinal microbiota of old age is poorly studied in comparison with the infant and adult gut microbiota in health and disease. Lifestyle is also thought to have a strong influence on microbiota composition and includes amount of exercise, living environment and pet ownership[114-117].

Although the intestinal microbiota plays an important role in immune modulation and maturation, the dense intestinal mucosal layer provides a barrier from physical contact and prevents significant immune stimulation and inflammation[118,119]. Additionally, phage adherence to the host via Ig-like domains has been shown to reduce bacterial abundance within the mucus layer. Furthermore, approximately 25 % of tailed dsDNA phage (*Caudovirales*) possess Ig-like domains, suggesting phage-mediated control of bacterial colonisation within the mucus layer[120]. Extensive research into the complex relationship between intestinal microbes and the host immune system is ongoing[121]. The mucus layer is created from MUC2 mucin that is secreted from the goblet cells[122]. A study using *muc2*-deficient mice that lacked the protective inner mucus layer developed severe colitis[123]. Furthermore, mice with mutations in the *Muc2* gene developed spontaneously inflammation compared to wild-type mice[124]. Several studies have observed an active role of intestinal microbes in the structural development of several gut-immune components such as T cells, B cells and lymphoid tissue[125-128]. For example, a germ-free mouse study reported dysfunctional intestine lymphoid development compared to conventionally house mice with healthy microbiomes[128]. Furthermore, the secretion of anti- and pro-inflammatory cytokines has been shown to be partially modulated by intestinal microbes[125]. The role of the intestinal microbiota in immune-system modulation and maintenance is further supported by the observation that responses to certain oral vaccinations are heavily dependent on living in a developed versus developing country[129]. For instance, the immune response to the oral rotavirus vaccine is significantly lower in children living in rural areas compared to children within Western countries[130,131].

Host genetics are also believed to play a significant role in the maintenance of the intestinal microbiota[132]. Studies have attempted to identify which microbial taxa are heritable using

monozygotic and dizygotic twin models[21,133]. While some of these studies did identify the same heritable taxa, the degree of heritability was different, and it is unknown if the microbial composition observed in these studies was truly attributable to shared host genetics[134]. However, it is extremely challenging to deduce a clear trend regarding the role of host genetics due to the numerous factors also shaping the microbiota. Additionally, numerous studies have shown that environmental factors can explain a greater proportion of microbiota variability than host genetics[16].

One of the main external variables that has the potential to alter the compositional and functional capacities of the intestinal microbiota is the host's diet[135]. Dietary modifications, particularly consumption of plant-based dietary fibre, significantly change microbiota composition[18,136]. Ingestion of resistant starch or non-starch polysaccharides significantly increases the abundance of specific microbes, such as *Ruminoccocus bromii* and *Eubacterium rectale*, known to be associated with production of short-chain fatty acids (SCFAs)[137]. SCFAs (acetate, propionate, butyrate) are produced by microbes through the fermentation of complex carbohydrates and polysaccharides (glycans). Within the intestinal lumen, SCFAs contribute towards epithelial barrier maintenance, production of antimicrobial peptides and induction of anti-inflammatory mediators (e.g., IL-10)[138-140]. However, SCFAs are able to cross the intestinal epithelium via passive diffusion or absorption[141-143]. Butyrate, propionate and acetate are found in differing proportions in various locations; butyrate is mainly metabolised in the intestinal epithelial, propionate is mainly utilised in the liver, whereas high concentrations of acetate are found in the plasma. Studies have shown that SCFAs contribute to brown adipose tissue activation, regulation of liver mitochondrial function, maintenance of homeostasis, appetite control and improved sleep[144-147]. Furthermore, SCFAs are believed to play a significant role in microbiota-gut-brain crosstalk and neurological disease onset. Studies in germ-free mice showed a reduction in blood-brain barrier integrity due to reduced expression of tight-junction proteins. Additionally, introduction of an adult microbiota to these germ-free mice improved blood brain-barrier integrity and decreased its permeability.

Glycans can be introduced to the intestinal microbiota via the host diet or host mucus[68,148]. Specific bacteria, such as *Eubacterium rectale*, *Roseburia* spp., *Faecalibacterium prausnitzii, Clostridium leptum* and *Rumminococcaceae*, can use host- or diet-derived glycans, depending on which is more readily available[149-152].

An important study examined the intestinal microbiota differences between children living in two locations: rural Burkina Faso and Florence, Italy[153]. The children living within these locations had

different diets, which was reflected in the microbiota composition. Children in Burkina Faso regularly ate a high-fibre diet with carbohydrates and non-animal protein, whereas those living in Italy consumed a typical Western diet that was high in sugar, starch, animal protein and low in fibre. The intestinal microbiota of Burkina Faso children showed a greater microbial richness, higher prevalence of *Prevotella* spp. and lower abundance of *Bacteroides* spp. compared to their European counterparts. A further study concluded that a consistent diet low in fats and high in carbohydrates produced a microbiota with high *Prevotella* abundance, whereas *Bacteroides* spp. were dominant in the intestinal microbiota of those consuming a high protein, high animal fat diet[136]. Differences in microbiota composition have also been observed in individuals consuming a plant-based (vegan or vegetarian) or animal-based diet (omnivorous)[18]. Of particular note was a decrease in members of the phylum *Firmicutes* with the ability to metabolise plant-based polysaccharides within the omnivorous cohort. Functional differences were also noted in individuals consuming an omnivorous diet, including increased expression of genes involved in vitamin biosynthesis. The baseline microbiota profiles reappeared within 3 days of the individuals resuming their typical diet[18]. Dietary interventions do not appear to have drastic effects on the composition of the faecal virome, with studies to date reinforcing that there is interindividual variation among and intra-personal stability of faecal viromes[154]. Diet may permanently change the metabolic potential of the intestinal microbiota by introducing genes. For example, several populations within Japan harbour bacteria within their intestinal microbiota that can metabolise marine red-algae and the gene(s) associated with this function have been transferred from marine bacteria to intestinal bacteria[155].

The adult intestinal phageome is dominated by dsDNA tailed *Caudovirales* (*Siphoviridae*, *Myoviridiae* and *Podoviriade*) and crAssphage, although there is a high degree of inter-individualisation[156-159]. It is believed that temperate phages comprise the majority of the phageome, compared to lytic phage. However, the true diversity of intestinal bacteriophages is yet to be determined due to the difficulty in studying the complete phageome. As seen with the bacteriome, phage diversity is driven by environmental factors (such as lifestyle and diet)[91]. Although disputed, there is increasing evidence for the presence of a core phageome[160-162]. A 2016 study identified 23 phage groups that were shared between ~ 50 % of healthy microbiotas studied[157,163]. Furthermore, 44 additional phage groups were present in 20-50 % of healthy subjects. Studies have also found that the adult phageome remains relatively stable over time, with one study showing ~ 95 % of viral genotypes were retained after one year and ~ 80 % after 2 years[159,161,164]. The stability of the faecal phageome is believed to be due to the predominance of temperate phage and the low mutation rate[164]. A 2020 study generated a Gut Virome Database

(GVD) from 2,697 microbial metagenomes from 1,986 individuals from 16 different countries and revealed age-dependent patterns of the virome among healthy Western individuals. The authors reported intestinal phage richness significantly increased between childhood and adulthood and decreased as age progressed into adulthood (65 + years of age)[162].

1.1.3 Intestinal microbiota in disease

The role of the intestinal microbiota in the development and/or progression of disease is a heavily researched topic and links have been found to numerous conditions such as asthma, diabetes, obesity, allergies, inflammatory bowel disease (IBD), autoimmune disorders and neurodegenerative diseases[165-171]. The microbial imbalance linked to these diseases is termed dysbiosis and refers to a decline in microbial diversity (at the taxonomic composition level) compared to healthy controls.

The role of dysbiosis has been extensively studied in IBDs such as Crohn's Disease (CD) and Ulcerative Colitis (UC)[168,172,173]. Both exhibit an overall loss of bacterial diversity and increase in specific bacteria such as *Enterobacteriaceae*[174,175]. Mouse studies have shown that expansion of *Enterobacteriaceae* populations is associated with new-onset CD and a reduction in intestinal inflammation can be achieved through selected removal of these bacteria[176]. Additionally, the loss of SCFA-producing bacteria, such as *F. prausnitzii,* has been associated with CD recurrence. An induced-colitis model within mice showed that supplementation with *F. prausnitzii* reduced inflammation and hints at an anti-inflammatory role within the gut[177]. Furthermore, microbial-related products from the stool of UC patients are able in induce inflammation in *in vitro* models. For example, stimulation of human dendritic cells with faecal metabolites from UC patients was enough to initiate inflammation. A specific metabolite pattern was noted in the most severe patients and was mainly associated with increased expansion of *Bacteroides* and *Candida* spp.[174]. Additional studies have also associated a reduced abundance of *Lactobacilllus* spp. with intestinal inflammation and dysbiosis[178,179].

Another well studied disease associated with microbial imbalances is atopic asthma[171]. In industrialised countries there has been a rapid increase in the incidence of childhood asthma and this is believed to be partly attributed to lifestyle (and indirectly the intestinal microbiota)[180]. For example, treatment of neonatal mice with antibiotics reduced microbiota diversity, exacerbated Th2 responses and increased susceptibility of allergic lung inflammation[181]. Additionally, atopy (genetic tendency to develop allergy disease) was reduced in children where a dog was a household member[115]. A mouse model reported supplementation with *Lactobacillus johnsonii*

conferred protection to airway allergy challenge[182]. A high-fibre diet may also reduce allergic airway inflammation due to the increase in microbially-produced SCFAs[183]. Multiple studies have reported consistent reduction in *Lachnospira, Faecalibacterium,* and *Akkermansia* abundance in children at risk of atopy or asthma[184-187]. These results suggest that early-life microbiota colonisation is associated with the risk of developing childhood asthma.

Obesity and type II diabetes mellitus (T2DM) are intertwined diseases and associated with dysbiosis[26,111,188]. The most compelling evidence for the role of the intestinal microbiota in the development of obesity is seen in mouse models[189-191]. A faecal microbiota transplant (FMT) from obese and lean littermates transferred the phenotype (lean or obese)[192,193]. Additionally, an FMT from obese mice to germ-free mice resulted in significant weight gain compared to an FMT from lean donors. Cohabitation of lean and obese mice reduced adiposity and other obesity-related characteristics, suggesting that the microbiota may play a role[193]. Several studies have reported a correlation between increased *Akkermansia muciniphila* and improved metabolic health in obese patients[194-197]. A high-fat diet (HFD) mouse model showed supplementation with *A. muciniphila* improved glucose tolerance, reduced circulating lipopolysaccharide levels and reduced systemic inflammation[198]. Additionally, use of a plant-based prebiotic that enriched for *A. muciniphila* showed beneficial effects[195]. However, these results are not consistent across all studies and could be attributed to environmental or strain differences[199]. HFD mouse models have been used extensively to study metabolic diseases[200-203]. For example, prebiotics protected mice from HFD-induced metabolic syndrome. This was not attributed to SCFA production but to microbiota-regulated IL-22 production and returned enterocyte function[204,205]. Additionally, supplementation of *Bifidobacterium longum* in HFD-fed mice stimulated mucin production thereby reversing intestinal mucus abnormalities[205].

As with childhood asthma, there has been a rise in the occurrence of autoimmune disorders in Western countries[206-209]. Type I diabetes mellitus (T1DM) has been studied extensively in relation to its onset and intestinal dysbiosis[210]. A 2015 study reported a reduction of bacterial diversity prior to T1DM onset[211]. Additionally, a European study observed increased abundance of *Bacteroides* spp., especially *B. dorei*, in Estonian and Finnish infants where a high occurrence of T1DM is reported[212].

It should be noted that it is extremely difficult to confidently associate a microbial alteration with the onset or progression of disease. This is due to the vast number of confounding variables that contribute to microbiota colonisation and stability[212,213]. However, an association with the

intestinal microbiota in a limited number of diseases has been proven: *C. difficile* infection and non-*C. difficile* antibiotic-associated haemorrhagic colitis (AAHC). *C. difficile* infection is the most common cause for antibiotic-associated diarrhoea, with approximately 25 % of patients exhibiting prolonged to recurrent infections that do not respond to antibiotic treatment[214-216]. Due to the high clinical cure rate (92-93 %), FMT is recommended by several national health agencies for treatment of recurring *C. difficile* infection[217-219]. *Klebsiella oxytoca* has been associated with AAHC. Transfer of a *K. oxytoca* strain isolated from an AAHC patient was able to induce the same disease within a rat model, fulfilling Koch's postulates. All isolates collected from patients produced cytotoxin, subsequently shown to contribute to disease[220].

The contribution of the microbiota towards metabolic disease phenotype is only just being uncovered. A 2018 study showed that study participants with T2DM exhibited higher concentrations of microbially produced imidazole propionate in their blood compared to subjects without T2DM. The authors identified 28 bacterial imidazole propionate-producers that were more abundant in T2DM subjects than healthy controls. Furthermore, germ-free mouse models showed that increased circulating imidazole propionate impaired glucose tolerance and insulin signalling, further suggesting that microbial metabolites contribute towards T2DM presentation[221]. A further complex study showed patients with hepatic steatosis had decreased microbial gene richness, increased hepatic inflammation, and dysfunctional aromatic and branched-chain amino acid metabolism (i.e. increased levels of related metabolites in their blood and urine). The authors reported steatosis was induced in human primary hepatic cells and in mice via treatment with a microbial product of aromatic amino acid metabolism, phenylacetic acid[111]. These studies show that a systems biology approach, with animal and *in vitro* models complementing human work, are needed to characterise the contribution of microbiota constituents or microbially produced metabolites to disease onset and progression.

The role of the phageome in disease onset and progression has been studied in several diseases, particularly IBD. A higher abundance of *Caudovirales* phage have been reported in paediatric UC and CD patients[222]. Furthermore, mucosal samples from 167 individuals with UC showed an increase in *Caudovirales* abundance but decrease in diversity, richness and evenness compared to controls[223]. The authors proposed that the alteration in phage diversity and abundance may contribute towards the inflammatory cascade observed in CD patients. It has also been suggested that the virome is responsible for the successful recovery rate of patients *C. difficile* infection with FMT. A recent study reported improvement of five patients following FMT with a filtered faecal

suspension. The authors showed no viable bacteria were present in the suspension and suggested phage may be responsible for the improvement[224].

1.1.4 Investigating the intestinal microbiota

In recent years, the advancement of sequencing technologies and bioinformatic tools have allowed researchers to begin to examine the composition and functional potential of the microbiota[225]. Prior to the advent of 454, Ion Torrent and Illumina sequencing, profiling microbial communities at scale was difficult, laborious and consisted of cultivating microbes or small-scale, expensive clone-based analyses reliant on Sanger sequencing[226,227]. The two main approaches to studying the microbiota currently are amplicon sequencing or shotgun metagenomics. Typically, amplicon sequencing is used to study the bacterial components of the microbiota via amplification of the universally conserved 16S rRNA gene. However, amplicon sequencing can be used to study fungi using intergenic transcribed spacer sequences and/or the 18S rRNA gene[228]. Although 16S rRNA gene sequencing has several advantages, such as being economically advantageous and easy to use bioinformatic tools, there are several caveats[229,230]. For example, it is difficult to achieve the resolution needed to differentiate species, and sometimes genera, within the microbiota[231]. Therefore, it may only be possible to confidently assign bacterial Operational Taxonomic Units (OTUs) to as low as family level. Lower taxonomic assignment could be achieved by using multiple V regions within the 16S rRNA gene; however, this is typically only done if a specific bacterial group wants to be investigated within the microbiota[232]. A further disadvantage of 16S rRNA gene sequencing is it is limited to only microbes that contain the gene. Therefore, viruses and fungi are excluded from investigations.

A 2015 study curated a large database containing all species isolated from the human body. The authors reported that, to date, only 2172 different prokaryotes had been isolated at least once[233]. Microbial culturomics, with improved cultivation methods and targeted approaches, is being used by an increasing number of studies to improve representation of the culturable intestinal microbiota and taxonomic assignments[234,235]. A 2016 study used multiple culture conditions to identify 1,057 prokaryotic species and added 531 species to the human gut repertoire[236]. This included 146 bacterial species which were previously isolated in humans (outside of the intestinal microbiota), one archaeon, 187 bacterial species which were not previously identified in humans, and 197 new species[236]. It is estimated 77 % of the 1525 prokaryotes identified in the human intestine have been cultured due to the increased use of culturomics. Furthermore, novel intestinal species continue to be described[235,237]. The introduction of shotgun metagenomic sequencing significantly increased the proportion of the intestinal microbiota that could be

characterised and made it possible to predict the functional capacity of the microbiota as well as its taxonomic composition. This approach is not limited to specific microbial kingdoms (detecting the bacteria, archaea, fungi, protozoa and fungi) and uses total DNA obtained from a sample. Following DNA sequencing, the reads are processed, and various quality control steps undertaken (removal of sequencing adapters, quality trimming, and removal of duplicates)[238]. The analysis is a combination of read-based and assembly-based approaches. Reads are mapped to reference databases (available microbial proteins, genomes and annotated metabolic pathways) for metabolic and taxonomic profiling. Additionally, contigs are generated via metagenome assembly. The contigs are 'binned' to attempt to group contigs of the same species together. This can be performed using supervised (i.e. with reference databases) or unsupervised (i.e. without reference databases) methods. Metagenomic studies should aim to use read and assembly-based techniques as additional inferences can be drawn from the data compared to using one technique alone[238]. The ever-decreasing cost of metagenomic sequencing has increased the number of studies using this method for microbiota profiling. It should be noted that to obtain a complete picture of the intestinal microbiota, deep shotgun metagenomic sequencing is needed to achieve adequate resolution to capture microbes with small genomes or low abundance microbes (e.g. bacteriophages). The depth of sequencing refers to the amount of data output from sequencing (normally in giga-base pairs, Gb) required to achieve coverage of each microbe above a relative abundance threshold. For example, one study concluded ~ 7 Gb of sequencing would be needed to achieve >20 x coverage of all microbes above 1 % relative abundance within the microbiota[239]. This approach vastly enhances the inferences that can be made in relation to the microbiota and health or disease. However, the accurate characterisation of the intestinal microbiota relies heavily on the databases used for taxonomic and functional assignments. Therefore, it is imperative that researchers strive to increase the databases to capture the full diversity of the intestinal microbiota. In response to the Human Microbiome Project (HMP), The Integrated Gene Catalogue (IGC) was created in 2014[23,240]. This database is a cumulation of genes derived from hundreds of bacterial genomes without sequenced representatives from HMP and Metagenomics of the Human Intestinal Tract (MetaHIT) consortium[241]. The IGC has provided pivotal in unveiling disease-associated microbial signatures in obesity, T2DM and other diseases[242-244]. However, the IGC only contains genetic information and not the organism the gene originated from. Therefore, it is not possible to achieve high-resolution taxonomic identification or examine complete functional pathways using IGC-derived data. A current method for increasing the diversity in intestinal microbe databases is the generation of metagenome-assembled genomes (MAGs)[245,246]. MAGs represent new members of existing taxa and 'unculturable' bacteria within the microbiota. MAGs are created through binning of *de novo*-assembled contigs into putative genomes. MAGs

can be used to improve taxonomic assignment of microbes within the study samples they were assembled from or added to relevant public or self-curated databases. However, the generation of high-quality MAGs relies heavily on accurate metagenome assembly and correctly binned contigs. The use of incorrect MAGs could significantly affect any taxonomic or functional conclusions drawn from the data. Therefore, careful consideration should be taken when using MAGs in microbiota studies and MAGs should be accurately vetted to ensure they are high-quality[247]. However, the correct use of MAGs has played an important role in deducing the uncultured aspect of the microbiota. For example, recent studies have generated between 60,000 and 150,000 MAGs from publicly available human microbiome studies, the majority of these genomes representing uncultured species[248-250]. One of these studies used 9,248 metagenomes from multiple body sites (stool, vagina, skin and oral cavity) and a variety of geographic locations. Through the generation of MAGs, taxonomically unexplored species were identified that were associated with non-Westernised populations[248]. A 2021 study generated a non-redundant intestinal genome database (Unified Human Gastrointestinal Genome, UHGG) and protein database (Unified Human Gastrointestinal Protein, UHGP) from 204,938 genomes generated from human microbiome studies[251]. Implementation of these databases in future microbiota studies could uncover microbial signatures that otherwise would have been missed. MAGs have also been generated from viral data, highlighting the diversity within the microbiota and bolstering the phage databases. For example, the Gut Phage Database (GPD) was recently released and contained approximately 142,000 viral genomes of > 10 kb in size[252]. This was achieved by mining 28,060 publicly available human gut metagenomes and 2,898 genomes from gut-derived bacteria for viral genomes with VirSorter and VirFinder[253,254]. The authors stated the generation of GPD significantly enhanced the current known diversity of phage within the human intestinal microbiota. Additionally, a novel viral clade (named Gubaphage) has been described, with several *Bacteroides* and *Parabacteroides* predicted as the bacterial host. Another study used MAGs for identification of viral contigs within whole metagenome samples[255]. The authors identified 3,738 complete phage genomes representing 451 putative genera from 5,742 whole-community faecal metagenome assemblies. This led to the proposal of three novel phage families: "*Quimbyviridae*" and "*Flandersviridae*" containing phage infecting abundant members of the genera *Bacteroides, Parabacteroides* and *Prevotella*, and "*Gratiaviridae*" including phage that are loosely related to the phage families *Autographiviridae*, *Drexlerviridae* and *Chaseviridae*.

One of the main disadvantages with metagenomic sequencing is the depth of sequencing needed to achieve strain-level resolution[256]. It is becoming clear that the microbiota displays a vast level of strain diversity and is highly individualised. Therefore, strain-level resolution is needed to

accurately examine microbial diversity and potential correlation of specific members of the microbiota with health and/or disease. Several programs have been developed to allow researchers to profile strains within metagenomic datasets (e.g. StrainPhlAn, PanPhlAn and InStrain)[257,258]. StrainPhlAn maps species-specific markers from reference genomes to metagenomic reads to obtain strain information. Additionally, the most abundant strain from each species is reconstructed and analysis of single nucleotide polymorphisms is used to determine if non-dominant strains are present[257]. Pangenome information can be obtained from metagenomic data using PanPhlAn. This approach can be used to identify unique strain-specific genomic traits. A recent metagenomic study used both StrainPhlAn and PanPhlAn to determine the population structure of *Ruminococcus bromii* from 4,077 available metagenomes. Despite being prevalent within the human intestine, only 15 *R. bromii* isolates have been sequenced[259,260]. Strain-level analysis allowed the authors to detect two genetically distinct clades that exhibited different functional gene annotations[261]. A further issue with metagenomic sequencing is the lack of functional information obtained. While it is possible to predict the functional potential of the microbiota using genes present and the associated metabolic pathways, the presence of a gene does not mean it is actively expressed and microbe undergoing the predicted function. Therefore, microbiota studies in recent years have used metabolomics and metatranscriptomics to investigate the intestinal metabolome and actively expressed genes, respectively[262,263]. These approaches complement metagenomic investigation with a functional "read-out" of the microbiota and provide important insights into the microbiota-metabolite-host relationship[264]. For example, shotgun metagenomic sequencing data and untargeted faecal metabolomic data from 1,004 from twins revealed a higher number of microbial metabolic pathways were shared between individuals (82 %) compared to microbial species (43 %)[265]. Furthermore, a recent study determined the core and variable portion of the metatranscriptome in a cohort of adult men[262]. The authors reported a difference between the core and variable sections of the metagenome and metatranscriptome. It was also highlighted that the metatranscriptome was more dynamic and species-specific than the metagenome.

An additional approach used to improve characterisation of intestinal microbiota diversity is viromics[266]. This involves the separation of virus-like particles (VLPs) from other microbial components in faecal samples using a variety of size and density filtration steps[267]. This allows high-resolution characterisation of viral genomes within the microbiome[268,269]. However, as with the bacterial components of the microbiome, virus databases are lagging behind sequencing technology advances[270]. However, as stated above, the mining of public metagenomes and

generation of subsequent viral databases (such as GPD) is greatly improving characterisation of the viral component of the microbiota. This is further discussed within Chapter 3.

1.2 Myalgic encephalomyelitis/Chronic Fatigue Syndrome

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a disabling and debilitating illness with an unknown ethology[271,272]. It is characterised by unexplained fatigue and a wide range of symptoms including post-exertional malaise, neurocognitive impairment, autonomic dysfunction, recurrent flu-like symptoms and gastrointestinal (GI) disturbances[273,274]. The symptoms reported can vary by patient and hint to the heterogeneous nature of the disease[274]. The United States Centre for Disease Control (US CDC) and World Health Organisation have classified the disease as a neurological disorder[275]. It is estimated that between 0.2 and 0.4 % of the UK population is affected by ME/CFS, according to National Institute for Health and Care Excellence (NICE)[276]. An additional study in 2011 estimated the prevalence at 0.2 %[277]. The peak age of onset is estimated between 20 and 45 years of age and the condition predominantly affects women[278,279]. The symptoms of ME/CFS can last for several years and most patients never return to premorbid levels of health[280,281]. Based on the severity of symptoms, patients are classified as mild, moderate or severe[282]. In the UK, there is no official guideline for designating disease severity and several different scales can be used in the clinical setting[274]. The International Consensus Criteria (ICC) define mild as ~50 % reduction in daily activity, moderate as mainly housebound, severe as mainly bedbound and very severe as bedbound, with significant dependence on help for physical functions. The ICC report also recommended that future ME/CFS research improve patient homogeneity by defining disease severity within studies[283].

The first outbreak of ME/CFS was recorded in 1934 in Los Angeles, California and was initially suspected as a poliomyelitis outbreak[284,285]. However, the presentation and age prevalence were atypical of poliomyelitis, which typically affects infants and children under 5 years of age. During this outbreak, the majority of cases were older children and younger adults[286]. Additionally, polio is commonly defined by the presence of flaccid paralysis, which was not present during the mystery outbreak[287]. Individuals with this disease presented with acute upper respiratory infection, muscle weakness, malaise, pain, fever, and photophobia. Furthermore, recurrent fever after apparent recovery was also reported[285]. Between 1934 and 1990, 62 similar outbreaks of atypical poliomyelitis were reported worldwide[284]. An outbreak in Akureyri, Iceland shared symptoms and occurrence of relapse with the Los Angeles outbreak[288]. In 1955, the Royal Free

Hospital in London reported a major outbreak where 292 staff were affected with similar symptoms as the previous outbreak. Additionally, cases exhibited neurological involvement. The disease was renamed ME and later extended to ME/CFS[289,290].

Approximately 80 % of ME/CFS sufferers reported a flu-like illness during disease onset and most patients are diagnosed following a viral, bacterial or parasitic infection[272]. Additionally, the majority of patients report delayed onset of symptoms after physical or mental activity; the severity and type of symptom can vary daily or weekly[290]. Currently there is no known cause and no specific diagnostic test available[291]. Diagnosis relies upon symptom-specific criteria after all relevant differential diagnoses have been excluded[272,292]. Currently there are 20 sets of case definitions or diagnostic criteria[273,293]. The most common diagnostic criteria used are CDC Fukuda 1994 and ICC 2011[283,294]. However, these criteria differ slightly in what symptoms define ME/CFS and what illnesses are excluded (e.g. depression). There are no universal or specific drugs for ME/CFS treatment and therapy options available, such as painkillers and antidepressants, focus on symptom relief[272,295-297].

1.2.1 Gut origin of ME/CFS

The occurrence of GI symptoms in ME/CFS patients is well reported[298,299]. For example, 92 % of ME/CFS patients reported irritable bowel syndrome (IBS)-like symptoms since onset of the disease[300]. Additionally, an increase in mucosal and systemic proinflammatory cytokines (IL-6, IL-8, IL-1β and TNFα) has been found in patients with IBS comorbidity[301]. The proposed mucosal inflammation and co-occurrence of GI symptoms in a high proportion of patients led researchers to investigate the intestinal microbiota as a possible origin of disease. In recent years, several studies have reported marked alterations in the gut microbiota of ME/CFS patients versus controls[298,299,302-306].

Although several studies have investigated the gut microbiota of ME/CFS patients, no microbes have been consistently identified as contributing towards disease onset or progression. It is difficult to directly compare studies due to inconsistencies in study design (diagnosis criteria, sample size) and microbial sequencing technology (shotgun metagenomics, 16S rRNA gene sequencing)[307]. However, multiple studies have reported an altered microbiota composition and reduced microbial diversity in patients when compared to controls[298,303,305]. A 2018 systematic review assessed the microbial composition of seven ME/CFS microbiota studies[308]. Of these seven studies, alterations in the microbiota composition of ME/CFS patients was noted in six. Similarly, an additional 2018 review compared the microbial composition of nine studies[307]. Although

similarities between study results were found, conflicting results were also discovered. For example, a decrease in overall bacterial abundance was noted in two studies but increased in another study[305,306,309]. Interestingly, relative abundance of several groups of butyrate-producing bacteria was decreased across multiple studies (*Faecalibacterium, Ruminococcaceae* and *Bacteroides*)[298,299,303-305]. Butyrate is an SCFA synthesised by microbial fermentation of dietary fibres in the large intestine. This SCFA is believed to have several beneficial properties, as described above.

The differing results reported from various ME/CFS studies can most likely be attributed to the study design (e.g. recruitment criteria, sequencing technology, etc) and subject genetic background (refer to information above and Appendix 1 for further information)[307,310]. Confounding factors such as influence of living environment and lifestyle habits could also be contributing to the alterations[311]. A recent study examining the oral microbiota of ME/CFS patients reported an increased relative abundance of *Leptotrichia*, *Prevotella* and *Fusobacterium* spp. and lower abundance of genera *Haemophilus*, *Veillonella* and *Porphyromonas* spp.[312] A 2017 shotgun metagenomic study examined the microbiota of 50 American ME/CFS patients and controls[298]. This study reported decreased abundance of *Dorea, Faecalibacterium* and *Coprococcus* spp. in ME/CFS patients compared to controls. Additionally, *Clostridium* and *Coprobacillus* spp. were higher in ME/CFS patients compared to controls. The authors stated the strongest predictors for ME/CFS were a decrease in *Faecalibacterium* spp. and an increase in *Alistipes*. This study also examined the microbiota in ME/CFS patients with IBS co-morbidity and revealed the microbiota of ME/CFS with IBS symptoms was altered compared to those without IBS. The patients with IBS showed a decrease in relative abundance of *Faecalibacterium* spp., *Ruminococcus obeum, Eubacterium hallii* and *Coprococcus comes*. Additionally, an increase in relative abundance of unclassified *Bacteroides*, *Pseudoflavonifractor capillosis* and *Eggerthella lenta* and a decrease in relative abundance of *Parabacteroides distasonis* were revealed as microbial signatures for ME/CFS patients without IBS. The authors reported certain bacterial abundance changes were attributed to differences between ME/CFS patients and controls; and additional bacterial abundance alterations separated ME/CFS patients with and without IBS. Furthermore, a handful of studies have attempted to characterise the faecal metabolome and identify metabolite indictive of ME/CFS[306,313]. A 2017 study reported an increase in SCFAs butyrate, isovalerate and valerate[306]. However, in additional microbiome studies, a decrease in SCFA-producing bacteria was consistently noted (in particular *Faecalibacterium*, eubacteria, *Roseburia* and *Ruminococcus* spp.)[298,303-305].

A recent 16S rRNA gene amplicon study offered the most comprehensive microbiota analysis of ME/CFS by examining the intestinal and oral microbiotas of 35 patients, 35 patients' relatives without ME/CFS and 35 healthy subjects not belonging to the patients' families[299].The authors reported significant alterations in the ME/CFS microbiota, compared to relative and non-relative controls. ME/CFS patients were characterised by a decrease in *Firmicutes* abundance and an increase in *Bacteroidetes* abundance, compared to controls. The relatives also showed a slight alteration in these microbial abundances when compared to controls. A decrease in several taxa of butyrate-producing bacteria was also noted. The authors also examined the faecal metabolome and showed a marked difference between the patients and external controls. As noted with the metagenome, the metabolome shared more similarity with relatives than controls, most likely due to similar lifestyles and diets. Furthermore, the authors reported specific metabolic markers within the ME/CFS patient cohort; namely, glutamic acid and argininosuccinic acid. Glutamic acid, primarily derived from dietary proteins, has been implicated in the microbiota-gut-brain axis and can act as a neurotransmitter and/or neuromodulator[314,315]. Furthermore, glutamatergic transmission alternations in the microbiota-gut-brain axis may influence physiological function. Accumulation of glutamic acid is thought to produce excitotoxicity and can result in significant neurological damage[316]. These results warrant further investigation into the microbiota-gut-brain axis and possible involvement in ME/CFS onset and progression.

Alterations in the abundance of *Bacteroides* spp. has been noted in several studies, although these alterations are not consistent across studies. To date, only four studies have noted an alternation in *Bacteroides* spp. within the intestinal microbiota of ME/CFS patients[298,299,306,313]. These studies used a variety of techniques to investigate the faecal microbiota: anaerobic culture, metabolic analysis (^{1}H-NMR spectroscopy), 16S rRNA gene amplicon sequencing and shotgun metagenomics. Two studies reported a decrease in *Bacteroides* spp. in ME/CFS patients compared to controls and a shotgun metagenomic study reported a decrease in *Bacteroides vulgatus* in ME/CFS patients without IBS[298,306,313]. The same study observed an increase in *Bacteroides* spp. (except *B. vulgatus*) in ME/CFS patients without IBS. However, a recent 16S rRNA gene-based study observed an increase in *Bacteroides* spp., especially *Bacteroides uniformis* and *Bacteroides ovatus,* ME/CFS patients and patient first relative compared to healthy controls[299]. Due to the differing patient cohorts and microbial identification techniques, it is almost impossible to determine if *Bacteroides* spp. are truly altered within the ME/CFS patient group.

1.3 Influence of *Bacteroides* species

Bacteroides spp. are dominant members of the adult intestinal microbiota and represent the most abundant commensals in the gut, but they can occasionally be opportunistic pathogens[86,317,318]. Colonisation of the intestinal tract with *Bacteroides* spp. begins at birth and their abundance is partly dependent on feeding mode[32,47]. Formula-fed infants have a higher proportion of *Bacteroides* spp. compared to their breastfed counterparts[319]. A 2015 study attributed an increase in the expression of complex sugar degradation genes within the 12-month-old infant gut microbiota to increased abundances of *Bacteroides thetaotaiomicron*[63]. As the gut microbiota reaches stability at 3 years of age, the abundance of *Bacteroides* spp. within the intestinal tract increases and they eventually become dominant members of the microbiota[86]. A 16S rRNA gene-based amplicon study examined the microbiota of children in Texas USA between 7 and 12 years of age and reported *Bacteroides* members account for nearly 40 % of a healthy child bacteriome[320]. The prevalence of specific *Bacteroides* spp. varied between individuals. A further study investigated the faecal microbiota of 281 school-age children in the Netherlands and discovered the most prevalent *Bacteroides* spp. according to detected annotated genes was *B. ovatus,* followed by *Bacteroides fragilis, Bacteroides thetaotaiomicron* and *Bacteroides xylanisolvens*[321]. The abundance of *Bacteroides* spp. within the adult intestinal microbiota varies according to different factors such as diet, environment, antibiotic use and lifestyle; particularly dietary patterns[19,322]. The prevalence of various *Bacteroides* spp. can vary according to eating habits, such as vegan, vegetarian or omnivorous diets[18,323]. For example, *B. fragilis* is less prevalent in vegan and vegetarian individuals than in omnivorous individuals[324]. Additionally, *Bacteroides salanitronis* (since reclassified as *Phocaeicola salanitronis*) and *Bacteroides coprocola* were common in omnivorous eaters, while *Bacteroides salyersiae* was present in high abundance in vegans[325]. The *Bacteroides* spp. patterns within individuals also vary geographically, and higher prevalence is noted in North American and European individuals[114]. This has been attributed to the Western diet, which is often high in fat and protein content. However, *Bacteroides* spp. are also common within Asian intestinal microbiotas[326,327]. For example, a study examined the microbiota of participants from Japan and India and reported a higher abundance of *Bacteroides* spp., *Bacteroides uniformis, B. ovatus* and *B. fragilis*, within the Japanese microbiota[328]. This was attributed to the differences in diet between the cultures as Japanese participants consumed an animal-based diet and Indian participants ate a more plant-based diet.

An important nutritional factor within the intestinal microbiome is the presence of glycans[149]. These glycans, or the human gut glycome, are derived from several locations; glycan introduced

from the host diet, host-secreted glycans (from the mucus) or microbially produced glycans[148]. *Bacteroides* spp. are able to use glycans as food sources, which contributes to the symbiotic relationship between host and these bacteria[329-331]. The microbial fermentation of indigestible glycans produces SCFAs, such as propionate, that directly benefit the host[332-334]. Propionate is an important anti-inflammatory mediator and contributes toward intestinal and immune system homeostasis[335]. This SCFA can inhibit the release of pro-inflammatory cytokines from neutrophils and macrophages[334]. The degradation of host-derived glycan by *B. thetaotaiomicron* assists in synthesis of the bacterium's outer capsule[336]. *B. fragilis* possesses similar genes to *B. thetaotaiomicron* and uses glycans for capsular polysaccharide synthesis, which contributes to overall colonisation and survival of the bacterium [337]. *Bacteroides* spp. encode complex polysaccharide utilisation loci (PULs) that are involved in the degradation of long-chain polysaccharides and oligosaccharides that are not absorbed by intestinal epithelial cells[150,338]. These loci are involved in complex carbohydrate acquisition and contribute towards the overall metabolism of *Bacteroides* spp. These PULs are relatively conserved across the genus *Bacteroides*[339]. PULs also contribute towards inter-species cross-feeding and overall maintenance of the gut ecosystem[340]. For example, quercetin is a well-known flavonoid present in nature and has various proposed health benefits including anti-inflammatory properties[341]. *B. thetaiotaomicron* lacks the ability to use quercetin but can degrade starch (via PULs) to maltose and glucose. In the presence of these sugars *Eubacterium ramulus* is able to degrade quercetin while fermenting glucose to butyrate; producing beneficial effects for both the human host and its commensal bacteria[340].

Bacteroides spp. are major producers of outer membrane vesicles (OMVs) and play important roles in communication with other bacteria and host tissue[342,343]. OMVs contribute to a wide range of functions including nutrient uptake, transfer of genetic material, biofilm formation and protection from antimicrobials[344]. The OMVs of *B. thetaotaomicron* contain glycosyl hydrolases that assist in levin degradation, a common carbohydrate derived from plants. The by-products of levan degradation are important for the growth of other *Bacteroides* spp.[345,346]. *Bacteroides* spp. are considered important players in the regulation and maintenance of the host immune system[347,348]. For example, capsular polysaccharide A (PSA) of *B. fragilis* has been shown to assist in host immune system homeostasis and prevent bacterial/viral infection[349,350]. This is achieved through PSA-induced $CD4^+$ T cell-dependent immune responses[351]. Additionally, treatment of herpes simplex virus 1-infected mice with PSA increased survival rates and decreased brainstem inflammation[349].

Despite their beneficial roles within the gut microbiome, *Bacteroides* spp. – especially *B. fragilis* – are also important opportunistic pathogens[317,352]. For example, *B. fragilis* can cause extra-intestinal abscesses and bacteraemia if allowed to cross the epithelial layer through physical translocation or extensive abdominal surgery[353]. Additionally, overabundance of *Bacteroides* spp. in the intestinal tract due to poor diet can also directly affect the host. A lack of bacterial competition can allow overgrowth of bacteria such as *Bacteroides caccae* and can result in the degradation of intestinal mucus and thereby increased intestinal inflammation[354]. Additionally, *B. fragilis* can produce a metalloprotease toxin (*B. fragilis* toxin) that can degrade intestinal tight junctions and increase intestinal hyperpermeability[355-357]. The isolates possessing this toxin can cause inflammatory diarrheal disease and have been associated with colon cancer risk[358,359]. Increasing incidence of antimicrobial resistance in this pathogen requires alternate therapies to antibiotics for the treatment of infections[360,361]. Phage may represent one such approach, though lytic *Bacteroides* phage are poorly represented in the literature (Chapter 3). The potential pathogenicity of *B. fragilis* (non-enterotoxigenic and enterotoxigenic) is discussed in Chapter 4.

1.4 Aims and objectives

As outlined above, the human intestinal microbiota is a complex ecosystem influenced by microbe-microbe and host-microbe interactions. Due to advancements in sequencing technologies and computational analyses, researchers are only just beginning to fully appreciate the important role the intestinal microbiome plays in human health and disease. The aims for the Thesis are to:

i. Examine the intestinal microbiota of severe ME/CFS patients compared to controls (Chapter 2);
ii. Characterise novel *B. fragilis* phage in relation to all known phage and metagenome-assembled phage genomes (Chapter 3);
iii. Investigate the pangenome of the opportunistic pathogen *B. fragilis* to determine if significant genomic differences are observed between non-enterotoxigenic and enterotoxigenic isolates (Chapter 4).

1.5 References

1 Sender, R., Fuchs, S. & Milo, R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell* **164**, 337-340, doi:10.1016/j.cell.2016.01.013 (2016).

2 Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61-66, doi:10.1038/nature23889 (2017).

3 Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449**, 804-810, doi:10.1038/nature06244 (2007).

4 Bang, C. & Schmitz, R. A. Archaea: forgotten players in the microbiome. *Emerging Topics in Life Sciences* **2**, 459-468, doi:10.1042/etls20180035 (2018).

5 Nash, A. K. *et al.* The gut mycobiome of the Human Microbiome Project healthy cohort. *Microbiome* **5**, 153, doi:10.1186/s40168-017-0373-4 (2017).

6 Chabé, M., Lokmer, A. & Ségurel, L. Gut Protozoa: Friends or Foes of the Human Gut Microbiota? *Trends in Parasitology* **33**, 925-934, doi:https://doi.org/10.1016/j.pt.2017.08.005 (2017).

7 Tisza, M. J. & Buck, C. B. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proceedings of the National Academy of Sciences* **118**, e2023202118, doi:10.1073/pnas.2023202118 (2021).

8 Kennedy, M. S. & Chang, E. B. in *Progress in Molecular Biology and Translational Science* Vol. 176 (ed Lora J. Kasselman) 1-42 (Academic Press, 2020).

9 Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science* **326**, 1694-1697, doi:10.1126/science.1177486 (2009).

10 Kuczynski, J. *et al.* Direct sequencing of the human microbiome readily reveals community differences. *Genome Biology* **11**, 210, doi:10.1186/gb-2010-11-5-210 (2010).

11 Grice, E. A. *et al.* Topographical and temporal diversity of the human skin microbiome. *Science (New York, N.Y.)* **324**, 1190-1192, doi:10.1126/science.1171700 (2009).

12 Grice, E. A. *et al.* A diversity profile of the human skin microbiota. *Genome research* **18**, 1043-1050, doi:10.1101/gr.075549.107 (2008).

13 Gribbon, E. M., Cunliffe, W. J. & Holland, K. T. Interaction of Propionibacterium acnes with skin lipids in vitro. *J Gen Microbiol* **139**, 1745-1751, doi:10.1099/00221287-139-8-1745 (1993).

14 Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature* **514**, 59-64, doi:10.1038/nature13786 (2014).

15 Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583-588, doi:10.1038/s41586-018-0617-x (2018).

16 Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210-215, doi:10.1038/nature25973 (2018).

17 Turpin, W. *et al.* Association of host genome with intestinal microbial composition in a large healthy cohort. *Nature Genetics* **48**, 1413-1417, doi:10.1038/ng.3693 (2016).

18 David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559-563, doi:10.1038/nature12820 (2014).

19 Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology* **14**, 20-32, doi:10.1038/nrmicro3552 (2016).

20 Proctor, D. M. & Relman, D. A. The Landscape Ecology and Microbiota of the Human Nose, Mouth, and Throat. *Cell Host Microbe* **21**, 421-432, doi:10.1016/j.chom.2017.03.011 (2017).

21 Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484, doi:10.1038/nature07540 (2009).

22 Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449**, 804-810, doi:10.1038/nature06244 (2007).

23 Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214, doi:10.1038/nature11234 (2012).

24 Zeevi, D. *et al.* Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43-48, doi:10.1038/s41586-019-1065-y (2019).

25 Flint, H. J., Scott, K. P., Louis, P. & Duncan, S. H. The role of the gut microbiota in nutrition and health. *Nat Rev Gastroenterol Hepatol* **9**, 577-589, doi:10.1038/nrgastro.2012.156 (2012).

26 Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541-546, doi:10.1038/nature12506 (2013).

27 Mohajeri, M. H. *et al.* The role of the microbiome for human health: from basic science to clinical applications. *Eur J Nutr* **57**, 1-14, doi:10.1007/s00394-018-1703-4 (2018).

28 Malla, M. A. *et al.* Exploring the Human Microbiome: The Potential Future Role of Next-Generation Sequencing in Disease Diagnosis and Treatment. *Frontiers in Immunology* **9**, doi:10.3389/fimmu.2018.02868 (2019).

29 Robertson, R. C., Manges, A. R., Finlay, B. B. & Prendergast, A. J. The Human Microbiome and Child Growth - First 1000 Days and Beyond. *Trends Microbiol* **27**, 131-147, doi:10.1016/j.tim.2018.09.008 (2019).

30 Gensollen, T., Iyer, S. S., Kasper, D. L. & Blumberg, R. S. How colonization by microbiota in early life shapes the immune system. *Science* **352**, 539-544, doi:10.1126/science.aad9378 (2016).

31 Tamburini, S., Shen, N., Wu, H. C. & Clemente, J. C. The microbiome in early life: implications for health outcomes. *Nature Medicine* **22**, 713-722, doi:10.1038/nm.4142 (2016).

32 Milani, C. *et al.* The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. *Microbiol Mol Biol Rev* **81**, doi:10.1128/mmbr.00036-17 (2017).

33 Jiménez, E. *et al.* Is meconium from healthy newborns actually sterile? *Res Microbiol* **159**, 187-193, doi:10.1016/j.resmic.2007.12.007 (2008).

34 Hansen, R. *et al.* First-Pass Meconium Samples from Healthy Term Vaginally-Delivered Neonates: An Analysis of the Microbiota. *PLoS One* **10**, e0133320, doi:10.1371/journal.pone.0133320 (2015).

35 Aagaard, K. *et al.* The placenta harbors a unique microbiome. *Sci Transl Med* **6**, 237ra265, doi:10.1126/scitranslmed.3008599 (2014).

36 Ardissone, A. N. *et al.* Meconium microbiome analysis identifies bacteria correlated with premature birth. *PLoS One* **9**, e90784, doi:10.1371/journal.pone.0090784 (2014).

37 Ross, M. G. & Nijland, M. J. Fetal swallowing: relation to amniotic fluid regulation. *Clin Obstet Gynecol* **40**, 352-365, doi:10.1097/00003081-199706000-00011 (1997).

38 He, Q. *et al.* The meconium microbiota shares more features with the amniotic fluid microbiota than the maternal fecal and vaginal microbiota. *Gut Microbes* **12**, 1794266, doi:10.1080/19490976.2020.1794266 (2020).

39 Sterpu, I. *et al.* No evidence for a placental microbiome in human pregnancies at term. *Am J Obstet Gynecol* **224**, 296.e291-296.e223, doi:10.1016/j.ajog.2020.08.103 (2021).

40 Del Chierico, F. *et al.* Phylogenetic and Metabolic Tracking of Gut Microbiota during Perinatal Development. *PLOS ONE* **10**, e0137347, doi:10.1371/journal.pone.0137347 (2015).

41 Mitchell, C. M. *et al.* Delivery Mode Affects Stability of Early Infant Gut Microbiota. *Cell Reports Medicine* **1**, 100156, doi:https://doi.org/10.1016/j.xcrm.2020.100156 (2020).

42 Biasucci, G. *et al.* Mode of delivery affects the bacterial community in the newborn gut. *Early Hum Dev* **86 Suppl 1**, 13-15, doi:10.1016/j.earlhumdev.2010.01.004 (2010).

43 Chu, D. M. *et al.* Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nature Medicine* **23**, 314-326, doi:10.1038/nm.4272 (2017).

44 Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences* **107**, 11971-11975, doi:10.1073/pnas.1002601107 (2010).

45 Di Mauro, A. *et al.* Gastrointestinal function development and microbiota. *Ital J Pediatr* **39**, 15-15, doi:10.1186/1824-7288-39-15 (2013).

46 Torrazza, R. M. & Neu, J. The altered gut microbiome and necrotizing enterocolitis. *Clin Perinatol* **40**, 93-108, doi:10.1016/j.clp.2012.12.009 (2013).

47 Scholtens, P. A. M. J., Oozeer, R., Martin, R., Amor, K. B. & Knol, J. The Early Settlers: Intestinal Microbiology in Early Life. *Annual Review of Food Science and Technology* **3**, 425-447, doi:10.1146/annurev-food-022811-101120 (2012).

48 Madan, J. C., Farzan, S. F., Hibberd, P. L. & Karagas, M. R. Normal neonatal microbiome variation in relation to environmental factors, infection and allergy. *Curr Opin Pediatr* **24**, 753-759, doi:10.1097/MOP.0b013e32835a1ac8 (2012).

49 Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci Transl Med* **8**, 343ra382, doi:10.1126/scitranslmed.aad7121 (2016).

50 Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117-121, doi:10.1038/s41586-019-1560-1 (2019).

51 MacIntyre, D. A. *et al.* The vaginal microbiome during pregnancy and the postpartum period in a European population. *Scientific Reports* **5**, 8988, doi:10.1038/srep08988 (2015).

52 DiGiulio, D. B. *et al.* Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences* **112**, 11060-11065, doi:10.1073/pnas.1502875112 (2015).

53 Vásquez, A., Jakobsson, T., Ahrné, S., Forsum, U. & Molin, G. Vaginal Lactobacillus Flora of Healthy Swedish Women. *Journal of Clinical Microbiology* **40**, 2746-2749, doi:doi:10.1128/JCM.40.8.2746-2749.2002 (2002).

54 Aagaard, K. *et al.* A Metagenomic Approach to Characterization of the Vaginal Microbiome Signature in Pregnancy. *PLOS ONE* **7**, e36466, doi:10.1371/journal.pone.0036466 (2012).

55 Doyle, R. *et al.* A Lactobacillus-Deficient Vaginal Microbiota Dominates Postpartum Women in Rural Malawi. *Applied and environmental microbiology* **84**, e02150-02117, doi:10.1128/AEM.02150-17 (2018).

56 Jakobsson, H. E. *et al.* Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section. *Gut* **63**, 559-566, doi:10.1136/gutjnl-2012-303249 (2014).

57 Salminen, S., Gibson, G. R., McCartney, A. L. & Isolauri, E. Influence of mode of delivery on gut microbiota composition in seven year old children. *Gut* **53**, 1388-1389, doi:10.1136/gut.2004.041640 (2004).

58 Berrington, J. E., Stewart, C. J., Embleton, N. D. & Cummings, S. P. Gut microbiota in preterm infants: assessment and relevance to health and disease. *Arch Dis Child Fetal Neonatal Ed* **98**, F286-290, doi:10.1136/archdischild-2012-302134 (2013).

59 Stewart, C. J. *et al.* Development of the preterm gut microbiome in twins at risk of necrotising enterocolitis and sepsis. *PLoS One* **8**, e73465, doi:10.1371/journal.pone.0073465 (2013).

60 Arboleya, S. *et al.* Establishment and development of intestinal microbiota in preterm neonates. *FEMS Microbiol Ecol* **79**, 763-772, doi:10.1111/j.1574-6941.2011.01261.x (2012).

61 Butel, M. J. *et al.* Conditions of bifidobacterial colonization in preterm infants: a prospective analysis. *J Pediatr Gastroenterol Nutr* **44**, 577-582, doi:10.1097/MPG.0b013e3180406b20 (2007).

62 Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences* **108**, 4578-4585, doi:10.1073/pnas.1000081107 (2011).

63 Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 690-703, doi:https://doi.org/10.1016/j.chom.2015.04.004 (2015).

64 Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci Transl Med* **8**, 343ra381, doi:10.1126/scitranslmed.aad0917 (2016).

65 Pannaraj, P. S. *et al.* Association Between Breast Milk Bacterial Communities and Establishment and Development of the Infant Gut Microbiome. *JAMA Pediatr* **171**, 647-654, doi:10.1001/jamapediatrics.2017.0378 (2017).

66 Macpherson, A. J., de Agüero, M. G. & Ganal-Vonarburg, S. C. How nutrition and the maternal microbiota shape the neonatal immune system. *Nature Reviews Immunology* **17**, 508-517, doi:10.1038/nri.2017.58 (2017).

67 Yan, J. *et al.* Gut microbiota induce IGF-1 and promote bone formation and growth. *Proceedings of the National Academy of Sciences* **113**, E7554-E7563, doi:10.1073/pnas.1607235113 (2016).

68 Braniste, V. *et al.* The gut microbiota influences blood-brain barrier permeability in mice. *Sci Transl Med* **6**, 263ra158, doi:10.1126/scitranslmed.3009759 (2014).

69 Ballard, O. & Morrow, A. L. Human milk composition: nutrients and bioactive factors. *Pediatr Clin North Am* **60**, 49-74, doi:10.1016/j.pcl.2012.10.002 (2013).

70 Lawson, M. A. E. *et al.* Breast milk-derived human milk oligosaccharides promote Bifidobacterium interactions within a single ecosystem. *The ISME Journal* **14**, 635-648, doi:10.1038/s41396-019-0553-2 (2020).

71 Oozeer, R. *et al.* Intestinal microbiology in early life: specific prebiotics can have similar functionalities as human-milk oligosaccharides. *Am J Clin Nutr* **98**, 561s-571s, doi:10.3945/ajcn.112.038893 (2013).

72 Alcon-Giner, C. *et al.* Microbiota Supplementation with Bifidobacterium and Lactobacillus Modifies the Preterm Infant Gut Microbiota and Metabolome: An Observational Study. *Cell Rep Med* **1**, 100077, doi:10.1016/j.xcrm.2020.100077 (2020).

73 Rodríguez, J. M. The origin of human milk bacteria: is there a bacterial entero-mammary pathway during late pregnancy and lactation? *Adv Nutr* **5**, 779-784, doi:10.3945/an.114.007229 (2014).

74 Jain, N. & Walker, W. A. Diet and host–microbial crosstalk in postnatal intestinal immune homeostasis. *Nature Reviews Gastroenterology & Hepatology* **12**, 14-25, doi:10.1038/nrgastro.2014.153 (2015).

75 Czosnykowska-Łukacka, M., Lis-Kuberka, J., Królak-Olejnik, B. & Orczyk-Pawiłowicz, M. Changes in Human Milk Immunoglobulin Profile During Prolonged Lactation. *Frontiers in Pediatrics* **8**, doi:10.3389/fped.2020.00428 (2020).

76 Pannaraj, P. S. *et al.* Shared and Distinct Features of Human Milk and Infant Stool Viromes. *Frontiers in Microbiology* **9**, doi:10.3389/fmicb.2018.01162 (2018).

77 Ma, J. *et al.* Comparison of gut microbiota in exclusively breast-fed and formula-fed babies: a study of 91 term infants. *Scientific Reports* **10**, 15792, doi:10.1038/s41598-020-72635-x (2020).

78 Roger, L. C., Costabile, A., Holland, D. T., Hoyles, L. & McCartney, A. L. Examination of faecal Bifidobacterium populations in breast- and formula-fed infants during the first 18 months of life. *Microbiology (Reading)* **156**, 3329-3341, doi:10.1099/mic.0.043224-0 (2010).

79 Ninonuevo, M. R. & Bode, L. Infant Formula Oligosaccharides Opening the Gates (for Speculation): Commentary on the article by Barrat et al. on page 34. *Pediatric Research* **64**, 8-10, doi:10.1203/PDR.0b013e3181752c2f (2008).

80 Guaraldi, F. & Salvatori, G. Effect of breast and formula feeding on gut microbiota shaping in newborns. *Front Cell Infect Microbiol* **2**, 94-94, doi:10.3389/fcimb.2012.00094 (2012).

81 Gomez-Llorente, C. *et al.* Three main factors define changes in fecal microbiota associated with feeding modality in infants. *J Pediatr Gastroenterol Nutr* **57**, 461-466, doi:10.1097/MPG.0b013e31829d519a (2013).

82 Koletzko, B. *et al.* Can infant feeding choices modulate later obesity risk? *Am J Clin Nutr* **89**, 1502s-1508s, doi:10.3945/ajcn.2009.27113D (2009).

83 Kong, K. L. *et al.* Association Between Added Sugars from Infant Formulas and Rapid Weight Gain in US Infants and Toddlers. *The Journal of Nutrition* **151**, 1572-1580, doi:10.1093/jn/nxab044 (2021).

84 Fallani, M. *et al.* Determinants of the human infant intestinal microbiota after the introduction of first complementary foods in infant samples from five European centres. *Microbiology (Reading)* **157**, 1385-1392, doi:10.1099/mic.0.042143-0 (2011).

85 Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the Human Infant Intestinal Microbiota. *PLOS Biology* **5**, e177, doi:10.1371/journal.pbio.0050177 (2007).

86 Rodríguez, J. M. *et al.* The composition of the gut microbiota throughout life, with an emphasis on early life. *Microb Ecol Health Dis* **26**, 26050, doi:10.3402/mehd.v26.26050 (2015).

87 Lim, E. S., Wang, D. & Holtz, L. R. The Bacterial Microbiome and Virome Milestones of Infant Development. *Trends Microbiol* **24**, 801-810, doi:10.1016/j.tim.2016.06.001 (2016).

88 Lim, E. S. *et al.* Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nature Medicine* **21**, 1228-1234, doi:10.1038/nm.3950 (2015).

89 Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23**, 111-120, doi:10.1101/gr.142315.112 (2013).

90 Lim, E. S. *et al.* Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat Med* **21**, 1228-1234, doi:10.1038/nm.3950 (2015).

91 Maqsood, R. *et al.* Discordant transmission of bacteria and viruses from mothers to babies at birth. *Microbiome* **7**, 156, doi:10.1186/s40168-019-0766-7 (2019).

92 Siranosian, B. A., Tamburini, F. B., Sherlock, G. & Bhatt, A. S. Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. *Nat Commun* **11**, 280, doi:10.1038/s41467-019-14103-3 (2020).

93 Liang, G. *et al.* The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* **581**, 470-474, doi:10.1038/s41586-020-2192-1 (2020).

94 Reyes, A. *et al.* Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc Natl Acad Sci U S A* **112**, 11941-11946, doi:10.1073/pnas.1514285112 (2015).

95 Hillman, E. T., Lu, H., Yao, T. & Nakatsu, C. H. Microbial Ecology along the Gastrointestinal Tract. *Microbes Environ* **32**, 300-313, doi:10.1264/jsme2.ME17017 (2017).

96 van den Bogert, B. *et al.* Diversity of human small intestinal Streptococcus and Veillonella populations. *FEMS Microbiol Ecol* **85**, 376-388, doi:10.1111/1574-6941.12127 (2013).

97 Wang, X., Heazlewood, S. P., Krause, D. O. & Florin, T. H. Molecular characterization of the microbial species that colonize human ileal and colonic mucosa by using 16S rDNA sequence analysis. *J Appl Microbiol* **95**, 508-520, doi:10.1046/j.1365-2672.2003.02005.x (2003).

98 Wang, M., Ahrné, S., Jeppsson, B. & Molin, G. Comparison of bacterial diversity along the human intestinal tract by direct cloning and sequencing of 16S rRNA genes. *FEMS Microbiol Ecol* **54**, 219-231, doi:10.1016/j.femsec.2005.03.012 (2005).

99 Fallingborg, J. *et al.* pH-profile and regional transit times of the normal gut measured by a radiotelemetry device. *Aliment Pharmacol Ther* **3**, 605-613, doi:10.1111/j.1365-2036.1989.tb00254.x (1989).

100 Tang, Q. *et al.* Current Sampling Methods for Gut Microbiota: A Call for More Precise Devices. *Front Cell Infect Microbiol* **10**, doi:10.3389/fcimb.2020.00151 (2020).

101 Zoetendal, E. G. *et al.* 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* **68**, 3401-3407, doi:10.1128/aem.68.7.3401-3407.2002 (2002).

102 Hartman, A. L. *et al.* Human gut microbiome adopts an alternative state following small bowel transplantation. *P Natl Acad Sci USA* **106**, 17187-17192, doi:10.1073/pnas.0904847106 (2009).

103 Kastl, A. J., Jr., Terry, N. A., Wu, G. D. & Albenberg, L. G. The Structure and Function of the Human Small Intestinal Microbiota: Current Understanding and Future Directions. *Cell Mol Gastroenterol Hepatol* **9**, 33-45, doi:10.1016/j.jcmgh.2019.07.006 (2020).

104 Gorzelak, M. A. *et al.* Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool. *PLOS ONE* **10**, e0134802, doi:10.1371/journal.pone.0134802 (2015).

105 Manor, O. *et al.* Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat Commun* **11**, 5206, doi:10.1038/s41467-020-18871-1 (2020).

106 Shanahan, F., Ghosh, T. S. & O'Toole, P. W. The Healthy Microbiome-What Is the Definition of a Healthy Gut Microbiome? *Gastroenterology* **160**, 483-494, doi:10.1053/j.gastro.2020.09.057 (2021).

107 Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220-230, doi:10.1038/nature11550 (2012).

108 Fassarella, M. *et al.* Gut microbiome stability and resilience: elucidating the response to perturbations in order to modulate gut health. *Gut* **70**, 595-605, doi:10.1136/gutjnl-2020-321747 (2021).

109 Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59-64, doi:10.1038/nature13568 (2014).

110 Cotillard, A. *et al.* Dietary intervention impact on gut microbial gene richness. *Nature* **500**, 585-588, doi:10.1038/nature12480 (2013).

111 Hoyles, L. *et al.* Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nature Medicine* **24**, 1070-1080, doi:10.1038/s41591-018-0061-3 (2018).

112 Claesson, M. J. *et al.* Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proceedings of the National Academy of Sciences* **108**, 4586-4591, doi:10.1073/pnas.1000097107 (2011).

113 Ragonnaud, E. & Biragyn, A. Gut microbiota as the key controllers of "healthy" aging of elderly people. *Immunity & Ageing* **18**, 2, doi:10.1186/s12979-020-00213-w (2021).

114 Conlon, M. A. & Bird, A. R. The impact of diet and lifestyle on gut microbiota and human health. *Nutrients* **7**, 17-44, doi:10.3390/nu7010017 (2014).

115 Ownby, D. R. & Johnson, C. C. Does exposure to dogs and cats in the first year of life influence the development of allergic sensitization? *Curr Opin Allergy Clin Immunol* **3**, 517-522, doi:10.1097/00130832-200312000-00015 (2003).

116 Song, S. J. *et al.* Cohabiting family members share microbiota with one another and with their dogs. *Elife* **2**, e00458, doi:10.7554/eLife.00458 (2013).

117 Lax, S. *et al.* Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* **345**, 1048-1052, doi:10.1126/science.1254529 (2014).

118 Johansson, M. E., Sjövall, H. & Hansson, G. C. The gastrointestinal mucus system in health and disease. *Nat Rev Gastroenterol Hepatol* **10**, 352-361, doi:10.1038/nrgastro.2013.35 (2013).

119 Vaishnava, S. *et al.* The antibacterial lectin RegIIIgamma promotes the spatial segregation of microbiota and host in the intestine. *Science (New York, N.Y.)* **334**, 255-258, doi:10.1126/science.1209791 (2011).

120 Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non–host-derived immunity. *Proceedings of the National Academy of Sciences* **110**, 10771-10776 (2013).

121 Johansson, M. E. V. & Hansson, G. C. Immunological aspects of intestinal mucus and mucins. *Nature Reviews Immunology* **16**, 639-649, doi:10.1038/nri.2016.88 (2016).

122 Ambort, D. *et al.* Calcium and pH-dependent packing and release of the gel-forming MUC2 mucin. *Proceedings of the National Academy of Sciences* **109**, 5645-5650, doi:10.1073/pnas.1120269109 (2012).

123 Hansson, G. C. & Johansson, M. E. The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Gut Microbes* **1**, 51-54, doi:10.4161/gmic.1.1.10470 (2010).

124 Heazlewood, C. K. *et al.* Aberrant mucin assembly in mice causes endoplasmic reticulum stress and spontaneous inflammation resembling ulcerative colitis. *PLoS Med* **5**, e54, doi:10.1371/journal.pmed.0050054 (2008).

125 Kamada, N. & Núñez, G. Regulation of the immune system by the resident intestinal bacteria. *Gastroenterology* **146**, 1477-1488, doi:10.1053/j.gastro.2014.01.060 (2014).

126 Mora, J. R. *et al.* Generation of gut-homing IgA-secreting B cells by intestinal dendritic cells. *Science* **314**, 1157-1160, doi:10.1126/science.1132742 (2006).

127 Gaboriau-Routhiau, V. *et al.* The key role of segmented filamentous bacteria in the coordinated maturation of gut helper T cell responses. *Immunity* **31**, 677-689, doi:10.1016/j.immuni.2009.08.020 (2009).

128 Bouskra, D. *et al.* Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis. *Nature* **456**, 507-510, doi:10.1038/nature07450 (2008).

129 de Jong, S. E., Olin, A. & Pulendran, B. The Impact of the Microbiome on Immunity to Vaccination in Humans. *Cell Host Microbe* **28**, 169-179, doi:https://doi.org/10.1016/j.chom.2020.06.014 (2020).

130 Armah, G. E. *et al.* Efficacy of pentavalent rotavirus vaccine against severe rotavirus gastroenteritis in infants in developing countries in sub-Saharan Africa: a randomised, double-blind, placebo-controlled trial. *Lancet* **376**, 606-614, doi:10.1016/s0140-6736(10)60889-6 (2010).

131 Zaman, K. *et al.* Efficacy of pentavalent rotavirus vaccine against severe rotavirus gastroenteritis in infants in developing countries in Asia: a randomised, double-blind, placebo-controlled trial. *Lancet* **376**, 615-623, doi:10.1016/s0140-6736(10)60755-6 (2010).

132 Cahana, I. & Iraqi, F. A. Impact of host genetics on gut microbiome: Take-home lessons from human and mouse studies. *Animal Model Exp Med* **3**, 229-236, doi:10.1002/ame2.12134 (2020).

133 Goodrich, J. K. *et al.* Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* **19**, 731-743, doi:10.1016/j.chom.2016.04.017 (2016).

134 Goodrich, J. K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789-799, doi:10.1016/j.cell.2014.09.053 (2014).

135 Albenberg, L. G. & Wu, G. D. Diet and the intestinal microbiome: associations, functions, and implications for health and disease. *Gastroenterology* **146**, 1564-1572, doi:10.1053/j.gastro.2014.01.058 (2014).

136 Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105-108, doi:10.1126/science.1208344 (2011).

137 Walker, A. W. *et al.* Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *Isme j* **5**, 220-230, doi:10.1038/ismej.2010.118 (2011).

138 Cox, M. A. *et al.* Short-chain fatty acids act as antiinflammatory mediators by regulating prostaglandin E(2) and cytokines. *World journal of gastroenterology* **15**, 5549-5557, doi:10.3748/wjg.15.5549 (2009).

139 Zhao, Y. *et al.* GPR43 mediates microbiota metabolite SCFA regulation of antimicrobial peptide expression in intestinal epithelial cells via activation of mTOR and STAT3. *Mucosal Immunol* **11**, 752-762, doi:10.1038/mi.2017.118 (2018).

140 Yan, H. & Ajuwon, K. M. Butyrate modifies intestinal barrier function in IPEC-J2 cells through a selective upregulation of tight junction proteins and activation of the Akt signaling pathway. *PLoS One* **12**, e0179586, doi:10.1371/journal.pone.0179586 (2017).

141 Miyauchi, S., Gopal, E., Fei, Y. J. & Ganapathy, V. Functional identification of SLC5A8, a tumor suppressor down-regulated in colon cancer, as a Na(+)-coupled transporter for short-chain fatty acids. *J Biol Chem* **279**, 13293-13296, doi:10.1074/jbc.C400059200 (2004).

142 Salvi, P. S. & Cowles, R. A. Butyrate and the Intestinal Epithelium: Modulation of Proliferation and Inflammation in Homeostasis and Disease. *Cells* **10**, 1775, doi:10.3390/cells10071775 (2021).

143 Rechkemmer, G. & von Engelhardt, W. Concentration- and pH-dependence of short-chain fatty acid absorption in the proximal and distal colon of guinea pig (Cavia porcellus). *Comp Biochem Physiol A Comp Physiol* **91**, 659-663, doi:10.1016/0300-9629(88)90944-9 (1988).

144 Szentirmai, É., Millican, N. S., Massie, A. R. & Kapás, L. Butyrate, a metabolite of intestinal bacteria, enhances sleep. *Scientific Reports* **9**, 7035, doi:10.1038/s41598-019-43502-1 (2019).

145 De Vadder, F. *et al.* Microbiota-generated metabolites promote metabolic benefits via gut-brain neural circuits. *Cell* **156**, 84-96, doi:10.1016/j.cell.2013.12.016 (2014).

146 Mollica, M. P. *et al.* Butyrate Regulates Liver Mitochondrial Function, Efficiency, and Dynamics in Insulin-Resistant Obese Mice. *Diabetes* **66**, 1405-1418, doi:10.2337/db16-0924 (2017).

147 Li, Z. *et al.* Butyrate reduces appetite and activates brown adipose tissue via the gut-brain neural circuit. *Gut* **67**, 1269-1279, doi:10.1136/gutjnl-2017-314050 (2018).

148 Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P. & Forano, E. Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* **3**, 289-306, doi:10.4161/gmic.19897 (2012).

149 Koropatkin, N. M., Cameron, E. A. & Martens, E. C. How glycan metabolism shapes the human gut microbiota. *Nature Reviews Microbiology* **10**, 323-335, doi:10.1038/nrmicro2746 (2012).

150 Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* **4**, 447-457, doi:10.1016/j.chom.2008.09.007 (2008).

151 Louis, P. & Flint, H. J. Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. *FEMS Microbiol Lett* **294**, 1-8, doi:10.1111/j.1574-6968.2009.01514.x (2009).

152 Guo, C. *et al.* Deficient butyrate-producing capacity in the gut microbiome of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome patients is associated with fatigue symptoms. *medRxiv*, 2021.2010.2027.21265575, doi:10.1101/2021.10.27.21265575 (2021).

153 De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences* **107**, 14691-14696, doi:10.1073/pnas.1005963107 (2010).

154 Carding, S. R., Davis, N. & Hoyles, L. Review article: the human intestinal virome in health and disease. *Aliment Pharmacol Ther* **46**, 800-815, doi:10.1111/apt.14280 (2017).

155 Hehemann, J.-H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908-912, doi:10.1038/nature08937 (2010).

156 Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications* **5** (2014).

157 Moreno-Gallego, J. L. *et al.* Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins. *Cell Host Microbe* **25**, 261-272.e265, doi:10.1016/j.chom.2019.01.019 (2019).

158 Kim, M. S., Park, E. J., Roh, S. W. & Bae, J. W. Diversity and abundance of single-stranded DNA viruses in human feces. *Appl Environ Microbiol* **77**, 8062-8070, doi:10.1128/aem.06331-11 (2011).

159 Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334-338, doi:10.1038/nature09199 (2010).

160 Minot, S. *et al.* The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res* **21**, 1616-1625, doi:10.1101/gr.122705.111 (2011).

161 Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, 527-541 e525, doi:10.1016/j.chom.2019.09.009 (2019).

162 Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* **28**, 724-740.e728, doi:https://doi.org/10.1016/j.chom.2020.08.003 (2020).

163 Manrique, P. *et al.* Healthy human gut phageome. *P Natl Acad Sci USA* **113**, 10400-10405 (2016).

164 Minot, S. *et al.* Rapid evolution of the human gut virome. *Proc Natl Acad Sci U S A* **110**, 12450-12455, doi:10.1073/pnas.1300833110 (2013).

165 Clapp, M. *et al.* Gut microbiota's effect on mental health: The gut-brain axis. *Clin Pract* **7**, 987-987, doi:10.4081/cp.2017.987 (2017).

166 Xu, H. *et al.* The Dynamic Interplay between the Gut Microbiota and Autoimmune Diseases. *J Immunol Res* **2019**, 7546047-7546047, doi:10.1155/2019/7546047 (2019).

167 Han, P. *et al.* The Association Between Intestinal Bacteria and Allergic Diseases—Cause or Consequence? *Front Cell Infect Microbiol* **11**, doi:10.3389/fcimb.2021.650893 (2021).

168 Glassner, K. L., Abraham, B. P. & Quigley, E. M. M. The microbiome and inflammatory bowel disease. *J Allergy Clin Immunol* **145**, 16-27, doi:10.1016/j.jaci.2019.11.003 (2020).

169 Davis, C. D. The Gut Microbiome and Its Role in Obesity. *Nutr Today* **51**, 167-174, doi:10.1097/NT.0000000000000167 (2016).

170 Chen, Z. *et al.* Association of Insulin Resistance and Type 2 Diabetes With Gut Microbial Diversity: A Microbiome-Wide Analysis From Population Studies. *JAMA Network Open* **4**, e2118811-e2118811, doi:10.1001/jamanetworkopen.2021.18811 (2021).

171 Frati, F. *et al.* The Role of the Microbiome in Asthma: The Gut–Lung Axis. *Int J Mol Sci* **20**, 123, doi:10.3390/ijms20010123 (2018).

172 Dalal, S. R. & Chang, E. B. The microbial basis of inflammatory bowel diseases. *J Clin Invest* **124**, 4190-4196, doi:10.1172/JCI72330 (2014).

173 Schirmer, M. *et al.* Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nature Microbiology* **3**, 337-346, doi:10.1038/s41564-017-0089-z (2018).

174 Mar, J. S. *et al.* Disease Severity and Immune Activity Relate to Distinct Interkingdom Gut Microbiome States in Ethnically Distinct Ulcerative Colitis Patients. *mBio* **7**, e01072-01016, doi:doi:10.1128/mBio.01072-16 (2016).

175 Gevers, D. *et al.* The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host Microbe* **15**, 382-392, doi:https://doi.org/10.1016/j.chom.2014.02.005 (2014).

176 Zhu, W. *et al.* Precision editing of the gut microbiota ameliorates colitis. *Nature* **553**, 208-211, doi:10.1038/nature25172 (2018).

177 Sokol, H. *et al.* Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci U S A* **105**, 16731-16736, doi:10.1073/pnas.0804812105 (2008).

178 Lamas, B. *et al.* CARD9 impacts colitis by altering gut microbiota metabolism of tryptophan into aryl hydrocarbon receptor ligands. *Nature Medicine* **22**, 598-605, doi:10.1038/nm.4102 (2016).

179 Tang, C. *et al.* Inhibition of Dectin-1 Signaling Ameliorates Colitis by Inducing Lactobacillus-Mediated Regulatory T Cell Expansion in the Intestine. *Cell Host Microbe* **18**, 183-197, doi:10.1016/j.chom.2015.07.003 (2015).

180 Fujimura, K. E. & Lynch, S. V. Microbiota in allergy and asthma and the emerging relationship with the gut microbiome. *Cell Host Microbe* **17**, 592-602, doi:10.1016/j.chom.2015.04.007 (2015).

181 Russell, S. L. *et al.* Early life antibiotic-driven changes in microbiota enhance susceptibility to allergic asthma. *EMBO Rep* **13**, 440-447, doi:10.1038/embor.2012.32 (2012).

182 Fonseca, W. *et al.* Lactobacillus johnsonii supplementation attenuates respiratory viral infection via metabolic reprogramming and immune cell modulation. *Mucosal Immunol* **10**, 1569-1580, doi:10.1038/mi.2017.13 (2017).

183 Trompette, A. *et al.* Gut microbiota metabolism of dietary fiber influences allergic airway disease and hematopoiesis. *Nat Med* **20**, 159-166, doi:10.1038/nm.3444 (2014).

184 Stokholm, J. *et al.* Maturation of the gut microbiome and risk of asthma in childhood. *Nature Communications* **9**, 141, doi:10.1038/s41467-017-02573-2 (2018).

185 Durack, J. *et al.* Delayed gut microbiota development in high-risk for asthma infants is temporarily modifiable by Lactobacillus supplementation. *Nature Communications* **9**, 707, doi:10.1038/s41467-018-03157-4 (2018).

186 Fujimura, K. E. *et al.* Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nature medicine* **22**, 1187-1191, doi:10.1038/nm.4176 (2016).

187 Arrieta, M. C. *et al.* Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci Transl Med* **7**, 307ra152, doi:10.1126/scitranslmed.aab2271 (2015).

188 Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99-103, doi:10.1038/nature12198 (2013).

189 Wang, S. *et al.* Gut microbiota mediates the anti-obesity effect of calorie restriction in mice. *Scientific Reports* **8**, 13037, doi:10.1038/s41598-018-31353-1 (2018).

190 Murphy, E. F. *et al.* Composition and energy harvesting capacity of the gut microbiota: relationship to diet, obesity and time in mouse models. *Gut* **59**, 1635-1642, doi:10.1136/gut.2010.215665 (2010).

191 Ridaura, V. K. *et al.* Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**, 1241214, doi:10.1126/science.1241214 (2013).

192 de Groot, P. *et al.* Donor metabolic characteristics drive effects of faecal microbiota transplantation on recipient insulin sensitivity, energy expenditure and intestinal transit time. *Gut* **69**, 502-512, doi:10.1136/gutjnl-2019-318320 (2020).

193 Vrieze, A. *et al.* Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. *Gastroenterology* **143**, 913-916.e917, doi:10.1053/j.gastro.2012.06.031 (2012).

194 Zhou, Q. *et al.* Gut bacteria Akkermansia is associated with reduced risk of obesity: evidence from the American Gut Project. *Nutrition & Metabolism* **17**, 90, doi:10.1186/s12986-020-00516-1 (2020).

195 Anhê, F. F. *et al.* A polyphenol-rich cranberry extract protects from diet-induced obesity, insulin resistance and intestinal inflammation in association with increased Akkermansia spp. population in the gut microbiota of mice. *Gut* **64**, 872-883 (2015).

196 Shin, N.-R. *et al.* An increase in the Akkermansia spp. population induced by metformin treatment improves glucose homeostasis in diet-induced obese mice. *Gut* **63**, 727-735 (2014).

197 Dao, M. C. *et al.* Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut* **65**, 426-436, doi:10.1136/gutjnl-2014-308778 (2016).

198 Everard, A. *et al.* Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proc Natl Acad Sci U S A* **110**, 9066-9071, doi:10.1073/pnas.1219451110 (2013).

199 Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55-60, doi:10.1038/nature11450 (2012).

200 Kim, K.-A., Gu, W., Lee, I.-A., Joh, E.-H. & Kim, D.-H. High Fat Diet-Induced Gut Microbiota Exacerbates Inflammation and Obesity in Mice via the TLR4 Signaling Pathway. *PLOS ONE* **7**, e47713, doi:10.1371/journal.pone.0047713 (2012).

201 Everard, A. *et al.* Microbiome of prebiotic-treated mice reveals novel targets involved in host response during obesity. *The ISME Journal* **8**, 2116-2130, doi:10.1038/ismej.2014.45 (2014).

202 Singh, R. P., Halaka, D. A., Hayouka, Z. & Tirosh, O. High-Fat Diet Induced Alteration of Mice Microbiota and the Functional Ability to Utilize Fructooligosaccharide for Ethanol Production. *Front Cell Infect Microbiol* **10**, doi:10.3389/fcimb.2020.00376 (2020).

203 Liu, Y. *et al.* Gut microbiome alterations in high-fat-diet-fed mice are associated with antibiotic tolerance. *Nature Microbiology* **6**, 874-884, doi:10.1038/s41564-021-00912-0 (2021).

204 Zou, J. *et al.* Fiber-Mediated Nourishment of Gut Microbiota Protects against Diet-Induced Obesity by Restoring IL-22-Mediated Colonic Health. *Cell Host Microbe* **23**, 41-53.e44, doi:10.1016/j.chom.2017.11.003 (2018).

205 Schroeder, B. O. *et al.* Bifidobacteria or Fiber Protects against Diet-Induced Microbiota-Mediated Colonic Mucus Deterioration. *Cell Host Microbe* **23**, 27-40.e27, doi:https://doi.org/10.1016/j.chom.2017.11.004 (2018).

206 Koch-Henriksen, N., Thygesen, L. C., Stenager, E., Laursen, B. & Magyari, M. Incidence of MS has increased markedly over six decades in Denmark particularly with late onset and in women. *Neurology* **90**, e1954-e1963, doi:10.1212/wnl.0000000000005612 (2018).

207 Hunter, T. M. *et al.* Prevalence of rheumatoid arthritis in the United States adult population in healthcare claims databases, 2004-2014. *Rheumatol Int* **37**, 1551-1557, doi:10.1007/s00296-017-3726-1 (2017).

208 Patterson, C. C. *et al.* Trends in childhood type 1 diabetes incidence in Europe during 1989-2008: evidence of non-uniformity over time in rates of increase. *Diabetologia* **55**, 2142-2147, doi:10.1007/s00125-012-2571-8 (2012).

209 Okada, H., Kuhn, C., Feillet, H. & Bach, J. F. The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. *Clinical and experimental immunology* **160**, 1-9, doi:10.1111/j.1365-2249.2010.04139.x (2010).

210 Zhou, H. *et al.* Evaluating the Causal Role of Gut Microbiota in Type 1 Diabetes and Its Possible Pathogenic Mechanisms. *Frontiers in Endocrinology* **11**, doi:10.3389/fendo.2020.00125 (2020).

211 Kostic, A. D. *et al.* The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260-273, doi:10.1016/j.chom.2015.01.001 (2015).

212 Vatanen, T. *et al.* Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* **165**, 842-853, doi:10.1016/j.cell.2016.04.007 (2016).

213 Chaudhari, S. N., McCurry, M. D. & Devlin, A. S. Chains of evidence from correlations to causal molecules in microbiome-linked diseases. *Nature Chemical Biology* **17**, 1046-1056, doi:10.1038/s41589-021-00861-z (2021).

214 Alfayyadh, M. *et al.* Recurrence of Clostridium difficile infection in the Western Australian population. *Epidemiol Infect* **147**, e153, doi:10.1017/s0950268819000499 (2019).

215 Eyre, D. W. *et al.* Predictors of first recurrence of Clostridium difficile infection: implications for initial management. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **55 Suppl 2**, S77-S87, doi:10.1093/cid/cis356 (2012).

216 Lessa, F. C., Winston, L. G. & McDonald, L. C. Burden of Clostridium difficile infection in the United States. *N Engl J Med* **372**, 2369-2370, doi:10.1056/NEJMc1505190 (2015).

217 Mullish, B. H. *et al.* The use of faecal microbiota transplant as treatment for recurrent or refractory Clostridium difficile infection and other potential indications: joint British Society of Gastroenterology (BSG) and Healthcare Infection Society (HIS) guidelines. *J Hosp Infect* **100 Suppl 1**, S1-s31, doi:10.1016/j.jhin.2018.07.037 (2018).

218 McDonald, L. C. *et al.* Clinical Practice Guidelines for Clostridium difficile Infection in Adults and Children: 2017 Update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA). *Clinical Infectious Diseases* **66**, e1-e48, doi:10.1093/cid/cix1085 (2018).

219 Hvas, C. L. *et al.* Fecal Microbiota Transplantation Is Superior to Fidaxomicin for Treatment of Recurrent Clostridium difficile Infection. *Gastroenterology* **156**, 1324-1332.e1323, doi:10.1053/j.gastro.2018.12.019 (2019).

220 Högenauer, C. *et al.* Klebsiella oxytoca as a causative organism of antibiotic-associated hemorrhagic colitis. *N Engl J Med* **355**, 2418-2426, doi:10.1056/NEJMoa054765 (2006).

221 Koh, A. *et al.* Microbially Produced Imidazole Propionate Impairs Insulin Signaling through mTORC1. *Cell* **175**, 947-961.e917, doi:https://doi.org/10.1016/j.cell.2018.09.055 (2018).

222 Norman, J. M. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447-460, doi:10.1016/j.cell.2015.01.002 (2015).

223 Zuo, T. *et al.* Gut mucosal virome alterations in ulcerative colitis. *Gut* **68**, 1169-1179, doi:10.1136/gutjnl-2018-318131 (2019).

224 Ott, S. J. *et al.* Efficacy of Sterile Fecal Filtrate Transfer for Treating Patients With Clostridium difficile Infection. *Gastroenterology* **152**, 799-811.e797, doi:10.1053/j.gastro.2016.11.010 (2017).

225 Wang, W.-L. *et al.* Application of metagenomics in the human gut microbiome. *World journal of gastroenterology* **21**, 803-814, doi:10.3748/wjg.v21.i3.803 (2015).

226 Lagier, J.-C. *et al.* Current and past strategies for bacterial culture in clinical microbiology. *Clinical microbiology reviews* **28**, 208-236, doi:10.1128/CMR.00110-14 (2015).

227 Roberfroid, M. *et al.* Prebiotic effects: metabolic and health benefits. *Br J Nutr* **104 Suppl 2**, S1-63, doi:10.1017/s0007114510003363 (2010).

228 Tonge, D. P., Pashley, C. H. & Gant, T. W. Amplicon –Based Metagenomic Analysis of Mixed Fungal Samples Using Proton Release Amplicon Sequencing. *PLOS ONE* **9**, e93849, doi:10.1371/journal.pone.0093849 (2014).

229 Petrosino, J. F., Highlander, S., Luna, R. A., Gibbs, R. A. & Versalovic, J. Metagenomic pyrosequencing and microbial identification. *Clin Chem* **55**, 856-866, doi:10.1373/clinchem.2008.107565 (2009).

230 Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. & Polz, M. F. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol* **186**, 2629-2635, doi:10.1128/jb.186.9.2629-2635.2004 (2004).

231 Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications* **10**, 5029, doi:10.1038/s41467-019-13036-1 (2019).

232 Bukin, Y. S. *et al.* The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data* **6**, 190007, doi:10.1038/sdata.2019.7 (2019).

233 Hugon, P. *et al.* A comprehensive repertoire of prokaryotic species identified in human beings. *Lancet Infect Dis* **15**, 1211-1219, doi:10.1016/s1473-3099(15)00293-5 (2015).

234 Diakite, A. *et al.* Extensive culturomics of 8 healthy samples enhances metagenomics efficiency. *PLoS One* **14**, e0223543 (2019).

235 Browne, H. P. *et al.* Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543-546, doi:10.1038/nature17645 (2016).

236 Lagier, J.-C. *et al.* Culture of previously uncultured members of the human gut microbiota by culturomics. *Nature Microbiology* **1**, 16203, doi:10.1038/nmicrobiol.2016.203 (2016).

237 Fleming, E. *et al.* Cultivation of common bacterial species and strains from human skin, oral, and gut microbiota. *BMC Microbiol* **21**, 278, doi:10.1186/s12866-021-02314-y (2021).

238 Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* **35**, 833-844, doi:10.1038/nbt.3935 (2017).

239 Ni, J., Yan, Q. & Yu, Y. How much metagenomic sequencing is enough to achieve a given goal? *Scientific reports* **3**, 1968, doi:10.1038/srep01968 (2013).

240 Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology* **32**, 834-841, doi:10.1038/nbt.2942 (2014).

241 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65, doi:10.1038/nature08821 (2010).

242 Wu, H. *et al.* Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat Med* **23**, 850-858, doi:10.1038/nm.4345 (2017).

243 Armour, C. R., Nayfach, S., Pollard, K. S. & Sharpton, T. J. A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome. *mSystems* **4**, doi:10.1128/mSystems.00332-18 (2019).

244 Liu, R. *et al.* Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nature Medicine* **23**, 859-868, doi:10.1038/nm.4358 (2017).

245 Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**, 725-731, doi:10.1038/nbt.3893 (2017).

246 Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl Environ Microbiol* **87**, doi:10.1128/aem.02593-20 (2021).

247 Chen, L. X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res* **30**, 315-333, doi:10.1101/gr.258640.119 (2020).

248 Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e620, doi:https://doi.org/10.1016/j.cell.2019.01.001 (2019).

249 Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499-504, doi:10.1038/s41586-019-0965-1 (2019).

250 Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505-510, doi:10.1038/s41586-019-1058-x (2019).

251 Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* **39**, 105-114, doi:10.1038/s41587-020-0603-3 (2021).

252 Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *bioRxiv*, 2020.2009.2003.280214, doi:10.1101/2020.09.03.280214 (2020).

253 Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Z. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5** (2017).

254 Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985, doi:10.7717/peerj.985 (2015).

255 Benler, S. *et al.* Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome* **9**, 78, doi:10.1186/s40168-021-01017-w (2021).

256 Segata, N. On the Road to Strain-Resolved Comparative Metagenomics. *mSystems* **3**, e00190-00117, doi:10.1128/mSystems.00190-17 (2018).

257 Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res* **27**, 626-638, doi:10.1101/gr.216242.116 (2017).

258 Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology* **39**, 727-736, doi:10.1038/s41587-020-00797-0 (2021).

259 Ze, X., Duncan, S. H., Louis, P. & Flint, H. J. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *The ISME Journal* **6**, 1535-1543, doi:10.1038/ismej.2012.4 (2012).

260 Mukhopadhya, I. *et al.* Sporulation capability and amylosome conservation among diverse human colonic and rumen isolates of the keystone starch-degrader Ruminococcus bromii. *Environ Microbiol* **20**, 324-336, doi:10.1111/1462-2920.14000 (2018).

261 Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, doi:10.7554/eLife.65088 (2021).

262 Abu-Ali, G. S. *et al.* Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nature Microbiology* **3**, 356-366, doi:10.1038/s41564-017-0084-4 (2018).

263 Vernocchi, P., Del Chierico, F. & Putignani, L. Gut Microbiota Profiling: Metabolomics Based Approach to Unravel Compounds Affecting Human Health. *Frontiers in Microbiology* **7**, doi:10.3389/fmicb.2016.01144 (2016).

264 Turnbaugh, P. J. & Gordon, J. I. An invitation to the marriage of metagenomics and metabolomics. *Cell* **134**, 708-713, doi:10.1016/j.cell.2008.08.025 (2008).

265 Visconti, A. *et al.* Interplay between the human gut microbiome and host metabolism. *Nature Communications* **10**, 4505, doi:10.1038/s41467-019-12476-z (2019).

266 Guerin, E. & Hill, C. Shining Light on Human Gut Bacteriophages. *Front Cell Infect Microbiol* **10**, doi:10.3389/fcimb.2020.00481 (2020).

267 Kleiner, M., Hooper, L. V. & Duerkop, B. A. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* **16**, 7, doi:10.1186/s12864-014-1207-4 (2015).

268 Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, 527-+ (2019).

269 Clooney, A. G. *et al.* Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host Microbe* **26**, 764-778.e765, doi:https://doi.org/10.1016/j.chom.2019.10.009 (2019).

270 Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host Microbe* **25**, 195-209, doi:10.1016/j.chom.2019.01.017 (2019).

271 Comerford, B. B. & Podell, R. Medically Documenting Disability in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) Cases. *Front Pediatr* **7**, 231, doi:10.3389/fped.2019.00231 (2019).

272 Bested, A. C. & Marshall, L. M. Review of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: an evidence-based approach to diagnosis and management by clinicians. *Rev Environ Health* **30**, 223-249, doi:10.1515/reveh-2015-0026 (2015).

273 Lim, E. J. & Son, C. G. Review of case definitions for myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). *J Transl Med* **18**, 289, doi:10.1186/s12967-020-02455-0 (2020).

274 Cortes Rivera, M., Mastronardi, C., Silva-Aldana, C. T., Arcos-Burgos, M. & Lidbury, B. A. Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: A Comprehensive Review. *Diagnostics (Basel)* **9**, doi:10.3390/diagnostics9030091 (2019).

275 Maes, M. Inflammatory and oxidative and nitrosative stress cascades as new drug targets in myalgic encephalomyelitis and chronic fatigue syndrome. *Mod Trends Pharmacopsychiatry* **28**, 162-174, doi:10.1159/000343982 (2013).

276 Lim, E. J. *et al.* Systematic review and meta-analysis of the prevalence of chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME). *J Transl Med* **18**, 100, doi:10.1186/s12967-020-02269-0 (2020).

277 Nacul, L. C. *et al.* Prevalence of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) in three regions of England: a repeated cross-sectional study in primary care. *BMC Med* **9**, 91, doi:10.1186/1741-7015-9-91 (2011).

278 Chu, L., Valencia, I. J., Garvert, D. W. & Montoya, J. G. Onset Patterns and Course of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. *Front Pediatr* **7**, 12, doi:10.3389/fped.2019.00012 (2019).

279 Bakken, I. J. *et al.* Two age peaks in the incidence of chronic fatigue syndrome/myalgic encephalomyelitis: a population-based registry study from Norway 2008-2012. *BMC Med* **12**, 167, doi:10.1186/s12916-014-0167-5 (2014).

280 Cairns, R. & Hotopf, M. A systematic review describing the prognosis of chronic fatigue syndrome. *Occup Med (Lond)* **55**, 20-31, doi:10.1093/occmed/kqi013 (2005).

281 Griffith, J. P. & Zarrouf, F. A. A systematic review of chronic fatigue syndrome: don't assume it's depression. *Prim Care Companion J Clin Psychiatry* **10**, 120-128, doi:10.4088/pcc.v10n0206 (2008).

282 Baraniuk, J. N. *et al.* A Chronic Fatigue Syndrome (CFS) severity score based on case designation criteria. *Am J Transl Res* **5**, 53-68 (2013).

283 Carruthers, B. M. *et al.* Myalgic encephalomyelitis: International Consensus Criteria. *J Intern Med* **270**, 327-338, doi:10.1111/j.1365-2796.2011.02428.x (2011).

284 Underhill, R. A. Myalgic encephalomyelitis, chronic fatigue syndrome: An infectious disease. *Medical Hypotheses* **85**, 765-773, doi:https://doi.org/10.1016/j.mehy.2015.10.011 (2015).

285 Meals, R. W., Hauser, V. F. & Bower, A. G. Poliomyelitis-The Los Angeles Epidemic of 1934 : Part I. *Cal West Med* **43**, 123-125 (1935).

286 HALL, W. J., NATHANSON, N. & LANGMUIR, A. D. THE AGE DISTRIBUTION OF POLIOMYELITIS IN THE UNITED STATES IN 19552. *American Journal of Epidemiology* **66**, 214-234, doi:10.1093/oxfordjournals.aje.a119896 (1957).

287 Mateen, F. J. & Black, R. E. Expansion of acute flaccid paralysis surveillance: beyond poliomyelitis. *Trop Med Int Health* **18**, 1421-1422, doi:10.1111/tmi.12181 (2013).

288 Sigurdsson, B., Sigurjonsson, J., Sigurdsson, J. H., Thorkelsson, J. & Gudmundsson, K. R. A disease epidemic in Iceland simulating poliomyelitis. *Am J Hyg* **52**, 222-238, doi:10.1093/oxfordjournals.aje.a119421 (1950).

289 The Medical Staff Of The Royal Free, H. AN OUTBREAK of encephalomyelitis in the Royal Free Hospital Group, London, in 1955. *Br Med J* **2**, 895-904 (1957).

290 Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/Chronic Fatigue, S., Board on the Health of Select, P. & Institute of, M. in *Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Redefining an Illness* (National Academies Press (US) Copyright 2015 by the National Academy of Sciences. All rights reserved., 2015).

291 Rowe, P. C. *et al.* Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Diagnosis and Management in Young People: A Primer. *Front Pediatr* **5**, 121, doi:10.3389/fped.2017.00121 (2017).

292 Clayton, E. W. Beyond myalgic encephalomyelitis/chronic fatigue syndrome: an IOM report on redefining an illness. *JAMA* **313**, 1101-1102, doi:10.1001/jama.2015.1346 (2015).

293 Brurberg, K. G., Fonhus, M. S., Larun, L., Flottorp, S. & Malterud, K. Case definitions for chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME): a systematic review. *BMJ Open* **4**, e003973, doi:10.1136/bmjopen-2013-003973 (2014).

294 Fukuda, K. *et al.* The chronic fatigue syndrome: a comprehensive approach to its definition and study. International Chronic Fatigue Syndrome Study Group. *Ann Intern Med* **121**, 953-959, doi:10.7326/0003-4819-121-12-199412150-00009 (1994).

295 Smith, M. E. *et al.* Treatment of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: A Systematic Review for a National Institutes of Health Pathways to Prevention Workshop. *Ann Intern Med* **162**, 841-850, doi:10.7326/M15-0114 (2015).

296 Richman, S. *et al.* Pharmaceutical Interventions in Chronic Fatigue Syndrome: A Literature-based Commentary. *Clin Ther* **41**, 798-805, doi:10.1016/j.clinthera.2019.02.011 (2019).

297 Castro-Marrero, J., Saez-Francas, N., Santillo, D. & Alegre, J. Treatment and management of chronic fatigue syndrome/myalgic encephalomyelitis: all roads lead to Rome. *Br J Pharmacol* **174**, 345-369, doi:10.1111/bph.13702 (2017).

298 Nagy-Szakal, D. *et al.* Fecal metagenomic profiles in subgroups of patients with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome* **5**, 44, doi:10.1186/s40168-017-0261-y (2017).

299 Lupo, G. F. D. *et al.* Potential role of microbiome in Chronic Fatigue Syndrome/Myalgic Encephalomyelits (CFS/ME). *Sci Rep* **11**, 7043, doi:10.1038/s41598-021-86425-6 (2021).

300 Aaron, L. A., Burke, M. M. & Buchwald, D. Overlapping conditions among patients with chronic fatigue syndrome, fibromyalgia, and temporomandibular disorder. *Arch Intern Med* **160**, 221-227, doi:10.1001/archinte.160.2.221 (2000).

301 Quigley, E. M. The enteric microbiota in the pathogenesis and management of constipation. *Best Pract Res Clin Gastroenterol* **25**, 119-126, doi:10.1016/j.bpg.2011.01.003 (2011).

302 Nagy-Szakal, D. *et al.* Insights into myalgic encephalomyelitis/chronic fatigue syndrome phenotypes through comprehensive metabolomics. *Sci Rep* **8**, 10056, doi:10.1038/s41598-018-28477-9 (2018).

303 Fremont, M., Coomans, D., Massart, S. & De Meirleir, K. High-throughput 16S rRNA gene sequencing reveals alterations of intestinal microbiota in myalgic encephalomyelitis/chronic fatigue syndrome patients. *Anaerobe* **22**, 50-56, doi:10.1016/j.anaerobe.2013.06.002 (2013).

304 Giloteaux, L., Hanson, M. R. & Keller, B. A. A Pair of Identical Twins Discordant for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Differ in Physiological Parameters and Gut Microbiome Composition. *Am J Case Rep* **17**, 720-729, doi:10.12659/ajcr.900314 (2016).

305 Giloteaux, L. *et al.* Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome* **4**, 30, doi:10.1186/s40168-016-0171-4 (2016).

306 Armstrong, C. W., McGregor, N. R., Lewis, D. P., Butt, H. L. & Gooley, P. R. The association of fecal microbiota and fecal, blood serum and urine metabolites in myalgic encephalomyelitis/chronic fatigue syndrome. *Metabolomics* **13**, 8, doi:10.1007/s11306-016-1145-z (2016).

307 Newberry, F., Hsieh, S. Y., Wileman, T. & Carding, S. R. Does the microbiome and virome contribute to myalgic encephalomyelitis/chronic fatigue syndrome? *Clin Sci (Lond)* **132**, 523-542, doi:10.1042/CS20171330 (2018).

308 Du Preez, S. *et al.* A systematic review of enteric dysbiosis in chronic fatigue syndrome/myalgic encephalomyelitis. *Syst Rev* **7**, 241, doi:10.1186/s13643-018-0909-0 (2018).

309 Sheedy, J. R. *et al.* Increased d-lactic Acid intestinal bacteria in patients with chronic fatigue syndrome. *In Vivo* **23**, 621-628 (2009).

310 Dibble, J. J., McGrath, S. J. & Ponting, C. P. Genetic risk factors of ME/CFS: a critical review. *Hum Mol Genet* **29**, R117-R124, doi:10.1093/hmg/ddaa169 (2020).

311 Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448-454, doi:10.1038/s41586-020-2881-9 (2020).

312 Wang, T. *et al.* Chronic fatigue syndrome patients have alterations in their oral microbiome composition and function. *PLoS One* **13**, e0203503, doi:10.1371/journal.pone.0203503 (2018).

313 Butt, H. *et al.* in *Proceedings of the AHMF International Clinical and Scientific Conference.* 12-14.

314 Baj, A. *et al.* Glutamatergic Signaling Along The Microbiota-Gut-Brain Axis. *Int J Mol Sci* **20**, doi:10.3390/ijms20061482 (2019).

315 Tomé, D. The Roles of Dietary Glutamate in the Intestine. *Ann Nutr Metab* **73 Suppl 5**, 15-20, doi:10.1159/000494777 (2018).

316 Mazzoli, R. & Pessione, E. The Neuro-endocrinological Role of Microbial Glutamate and GABA Signaling. *Frontiers in microbiology* **7**, 1934-1934, doi:10.3389/fmicb.2016.01934 (2016).

317 Yoshino, Y. *et al.* Clinical features of Bacteroides bacteremia and their association with colorectal carcinoma. *Infection* **40**, 63-67, doi:10.1007/s15010-011-0159-8 (2012).

318 Wexler, H. M. Bacteroides: the good, the bad, and the nitty-gritty. *Clinical microbiology reviews* **20**, 593-621, doi:10.1128/CMR.00008-07 (2007).

319 Wang, M. *et al.* Fecal microbiota composition of breast-fed infants is correlated with human milk oligosaccharides consumed. *J Pediatr Gastroenterol Nutr* **60**, 825-833, doi:10.1097/mpg.0000000000000752 (2015).

320 Hollister, E. B. *et al.* Structure and function of the healthy pre-adolescent pediatric gut microbiome. *Microbiome* **3**, 36-36, doi:10.1186/s40168-015-0101-x (2015).

321 Zhong, H. *et al.* Impact of early events and lifestyle on the gut microbiota and metabolic phenotypes in young school-age children. *Microbiome* **7**, 2, doi:10.1186/s40168-018-0608-z (2019).

322 Kurilshikov, A., Wijmenga, C., Fu, J. & Zhernakova, A. Host Genetics and Gut Microbiome: Challenges and Perspectives. *Trends Immunol* **38**, 633-647, doi:10.1016/j.it.2017.06.003 (2017).

323 Tomova, A. *et al.* The Effects of Vegetarian and Vegan Diets on Gut Microbiota. *Frontiers in Nutrition* **6**, doi:10.3389/fnut.2019.00047 (2019).

324 Ferrocino, I. *et al.* Fecal Microbiota in Healthy Subjects Following Omnivore, Vegetarian and Vegan Diets: Culturable Populations and rRNA DGGE Profiling. *PLoS One* **10**, e0128669, doi:10.1371/journal.pone.0128669 (2015).

325 García-López, M. *et al.* Analysis of 1,000 Type-Strain Genomes Improves Taxonomic Classification of Bacteroidetes. *Frontiers in Microbiology* **10**, doi:10.3389/fmicb.2019.02083 (2019).

326 Zhang, W. *et al.* Gut microbiota community characteristics and disease-related microorganism pattern in a population of healthy Chinese people. *Scientific Reports* **9**, 1594, doi:10.1038/s41598-018-36318-y (2019).

327 Lu, J. *et al.* Chinese gut microbiota and its associations with staple food type, ethnicity, and urbanization. *npj Biofilms and Microbiomes* **7**, 71, doi:10.1038/s41522-021-00245-0 (2021).

328 Pareek, S. *et al.* Comparison of Japanese and Indian intestinal microbiota shows diet-dependent interaction between bacteria and fungi. *npj Biofilms and Microbiomes* **5**, 37, doi:10.1038/s41522-019-0110-9 (2019).

329 Luis, A. S. *et al.* Dietary pectic glycans are degraded by coordinated enzyme pathways in human colonic Bacteroides. *Nature Microbiology* **3**, 210-219, doi:10.1038/s41564-017-0079-1 (2018).

330 Koropatkin, N. M., Martens, E. C., Gordon, J. I. & Smith, T. J. Starch Catabolism by a Prominent Human Gut Symbiont Is Directed by the Recognition of Amylose Helices. *Structure* **16**, 1105-1115, doi:https://doi.org/10.1016/j.str.2008.03.017 (2008).

331 Briliūtė, J. *et al.* Complex N-glycan breakdown by gut Bacteroides involves an extensive enzymatic apparatus encoded by multiple co-regulated genetic loci. *Nature Microbiology* **4**, 1571-1581, doi:10.1038/s41564-019-0466-x (2019).

332 Parada Venegas, D. *et al.* Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases. *Frontiers in Immunology* **10**, doi:10.3389/fimmu.2019.00277 (2019).

333 Silva, Y. P., Bernardi, A. & Frozza, R. L. The Role of Short-Chain Fatty Acids From Gut Microbiota in Gut-Brain Communication. *Frontiers in Endocrinology* **11**, doi:10.3389/fendo.2020.00025 (2020).

334 Vinolo, M. A., Rodrigues, H. G., Nachbar, R. T. & Curi, R. Regulation of inflammation by short chain fatty acids. *Nutrients* **3**, 858-876, doi:10.3390/nu3100858 (2011).

335 Shimizu, J. *et al.* Propionate-producing bacteria in the intestine may associate with skewed responses of IL10-producing regulatory T cells in patients with relapsing polychondritis. *PLOS ONE* **13**, e0203657, doi:10.1371/journal.pone.0203657 (2018).

336 Martens, E. C., Roth, R., Heuser, J. E. & Gordon, J. I. Coordinate regulation of glycan degradation and polysaccharide capsule biosynthesis by a prominent human gut symbiont. *J Biol Chem* **284**, 18445-18457, doi:10.1074/jbc.M109.008094 (2009).

337 Coyne, M. J., Chatzidaki-Livanis, M., Paoletti, L. C. & Comstock, L. E. Role of glycan synthesis in colonization of the mammalian gut by the bacterial symbiont Bacteroides fragilis. *Proc Natl Acad Sci U S A* **105**, 13099-13104, doi:10.1073/pnas.0804220105 (2008).

338 Grondin, J. M., Tamura, K., Déjean, G., Abbott, D. W. & Brumer, H. Polysaccharide Utilization Loci: Fueling Microbial Communities. *Journal of bacteriology* **199**, e00860-00816, doi:10.1128/JB.00860-16 (2017).

339 Lapébie, P., Lombard, V., Drula, E., Terrapon, N. & Henrissat, B. Bacteroidetes use thousands of enzyme combinations to break down glycans. *Nature Communications* **10**, 2043, doi:10.1038/s41467-019-10068-5 (2019).

340 Rodriguez-Castaño, G. P. *et al.* Bacteroides thetaiotaomicron Starch Utilization Promotes Quercetin Degradation and Butyrate Production by Eubacterium ramulus. *Frontiers in microbiology* **10**, 1145-1145, doi:10.3389/fmicb.2019.01145 (2019).

341 Zhao, L. *et al.* Quercetin Ameliorates Gut Microbiota Dysbiosis That Drives Hypothalamic Damage and Hepatic Lipogenesis in Monosodium Glutamate-Induced Abdominal Obesity. *Frontiers in Nutrition* **8**, doi:10.3389/fnut.2021.671353 (2021).

342 Bryant, W. A. *et al.* In Silico Analysis of the Small Molecule Content of Outer Membrane Vesicles Produced by Bacteroides thetaiotaomicron Indicates an Extensive Metabolic Link between Microbe and Host. *Front Microbiol* **8**, 2440, doi:10.3389/fmicb.2017.02440 (2017).

343 Zakharzhevskaya, N. B. *et al.* Outer membrane vesicles secreted by pathogenic and nonpathogenic Bacteroides fragilis represent different metabolic activities. *Scientific Reports* **7**, 5008, doi:10.1038/s41598-017-05264-6 (2017).

344 Jones, E. J. *et al.* The Uptake, Trafficking, and Biodistribution of Bacteroides thetaiotaomicron Generated Outer Membrane Vesicles. *Frontiers in Microbiology* **11**, doi:10.3389/fmicb.2020.00057 (2020).

345 Poeker, S. A. *et al.* Understanding the prebiotic potential of different dietary fibers using an in vitro continuous adult fermentation model (PolyFermS). *Scientific Reports* **8**, 4318, doi:10.1038/s41598-018-22438-y (2018).

346 Adamberg, S. *et al.* Degradation of Fructans and Production of Propionic Acid by Bacteroides thetaiotaomicron are Enhanced by the Shortage of Amino Acids. *Frontiers in nutrition* **1**, 21-21, doi:10.3389/fnut.2014.00021 (2014).

347 Shen, Y. *et al.* Outer membrane vesicles of a human commensal mediate immune regulation and disease protection. *Cell Host Microbe* **12**, 509-520, doi:10.1016/j.chom.2012.08.004 (2012).

348 Troy, E. B. & Kasper, D. L. Beneficial effects of Bacteroides fragilis polysaccharides on the immune system. *Front Biosci (Landmark Ed)* **15**, 25-34, doi:10.2741/3603 (2010).

349 Ramakrishna, C. *et al.* Bacteroides fragilis polysaccharide A induces IL-10 secreting B and T cells that prevent viral encephalitis. *Nature Communications* **10**, 2153, doi:10.1038/s41467-019-09884-6 (2019).

350 Sommese, L. *et al.* Evidence of Bacteroides fragilis Protection from Bartonella henselae-Induced Damage. *PLOS ONE* **7**, e49653, doi:10.1371/journal.pone.0049653 (2012).

351 Mogensen, T. H. Pathogen recognition and inflammatory signaling in innate immune defenses. *Clinical microbiology reviews* **22**, 240-273, doi:10.1128/CMR.00046-08 (2009).

352 Nguyen, M. H. *et al.* Antimicrobial Resistance and Clinical Outcome of Bacteroides Bacteremia: Findings of a Multicenter Prospective Observational Trial. *Clinical Infectious Diseases* **30**, 870-876, doi:10.1086/313805 (2000).

353 Goldstein, E. J. C. Intra-Abdominal Anaerobic Infections: Bacteriology and Therapeutic Potential of Newer Antimicrobial Carbapenem, Fluoroquinolone, and Desfluoroquinolone Therapeutic Agents. *Clinical Infectious Diseases* **35**, S106-S111, doi:10.1086/341930 (2002).

354 Desai, M. S. *et al.* A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility. *Cell* **167**, 1339-1353.e1321, doi:10.1016/j.cell.2016.10.043 (2016).

355 Ulger Toprak, N. *et al.* The distribution of the bft alleles among enterotoxigenic Bacteroides fragilis strains from stool specimens and extraintestinal sites. *Anaerobe* **12**, 71-74, doi:10.1016/j.anaerobe.2005.11.001 (2006).

356 Wu, S. *et al.* The Bacteroides fragilis toxin binds to a specific intestinal epithelial cell receptor. *Infect Immun* **74**, 5382-5390, doi:10.1128/iai.00060-06 (2006).

357 Valguarnera, E. & Wardenburg, J. B. Good Gone Bad: One Toxin Away From Disease for Bacteroides fragilis. *J Mol Biol* **432**, 765-785, doi:10.1016/j.jmb.2019.12.003 (2020).

358 Chung, L. *et al.* Bacteroides fragilis Toxin Coordinates a Pro-carcinogenic Inflammatory Cascade via Targeting of Colonic Epithelial Cells. *Cell Host Microbe* **23**, 203-214.e205, doi:https://doi.org/10.1016/j.chom.2018.01.007 (2018).

359 Sears, C. L. *et al.* Association of enterotoxigenic Bacteroides fragilis infection with inflammatory diarrhea. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **47**, 797-803, doi:10.1086/591130 (2008).

360 Raplee, I. *et al.* Emergence of nosocomial associated opportunistic pathogens in the gut microbiome after antibiotic treatment. *Antimicrobial Resistance & Infection Control* **10**, 36, doi:10.1186/s13756-021-00903-0 (2021).

361 Jasemi, S. *et al.* Antibiotic resistance pattern of Bacteroides fragilis isolated from clinical and colorectal specimens. *Annals of Clinical Microbiology and Antimicrobials* **20**, 27, doi:10.1186/s12941-021-00435-w (2021).

Chapter 2 : Analysis of faecal gut microbiota in severe Myalgic Encephalomyelitis/Chronic Fatigue Syndrome

2.1 Aims and objectives

As outlined in Chapter 1, interest in the faecal microbiota of ME/CFS patients has increased in recent years and several studies examining the faecal microbiota have been undertaken. However, it is difficult to compare across studies due to differences in diagnostic criteria, inconsistent use of household controls, patient disease severity and experimental design. This Chapter reports the analysis of the faecal microbiota from 14 severe ME/CFS patients and 5 controls. The taxonomic and functional profiles of patients and controls were compared to determine significant microbial differences between the groups.

It should be noted that the original research topic for this PhD was to examine the faecal microbiota of severe ME/CFS patients and household controls recruited from the Southeast of England. However, completion of this plan was not possible due to the slow recruitment of patients with severe disease and appropriate controls. The project was further delayed by a 9-month maternity leave and numerous lockdowns due to COVID-19. Therefore, the decision was made to investigate the faecal microbiota of a small group of severe patients collected by a previous PhD student (Dr Daniel Vipond). Due to the small sample size, the study was underpowered with respect to statistical power.

To ensure the submission of a complete thesis, research was undertaken that was beyond the scope of the original PhD plan; however, this work does revolve around the human intestinal microbiota. Due to the March 2020 COVID-19 lockdown, this research mainly involved bioinformatics and work that could be completed off-site. Following easing of the lockdown rules, I was unable to return to on-site working due to living with two clinically vulnerable individuals. Therefore, this thesis shows research involving the ME/CFS microbiota, *Bacteroides* phage discovery (Chapter 3) and *Bacteroides fragilis* pangenome analysis (Chapter 4).

2.2 Methods

2.2.1 Patient selection and recruitment

2.2.1.1 Ethics

Ethical approval for the collection of faecal samples from ME/CFS patients and household controls was obtained by a previous PhD student (Dr Daniel Vipond). The study ("A role for a leaky gut and the intestinal microbiota in the pathophysiology of ME/CFS") was a collaboration between the University of East Anglia, Quadram Institute Bioscience (formerly the Institute of Food Research) and Epsom and St Helier University Hospital NHS Trust.

2.2.1.2 Patient and control selection

The patients for the above study were selected by Dr Amolak Bansal, a consultant immunologist and Director of Chronic Fatigue Service at Epsom and St Helier University Hospital NHS Trust. Patients were diagnosed with ME/CFS by Dr Bansal if they fulfilled the Fukada, Canadian and Oxford diagnostic criteria[1-3]. Additionally, patients were excluded based on clinical depression and anxiety (using clinical history) and The Hospital Anxiety Depression Scale (HADS)[4]. A disease severity was assigned using The Chadler Fatigue Scale according to the following criteria[5]:

- **Mild** – mobile, self-caring, light domestic duties, may be working but to detriment of social, family and leisure activities;
- **Moderate** – Reduced motility, not working, reduced activities of daily life, sleeping in daytime, peaks and troughs of activity;
- **Severe** – Few activities of daily life, severe cognitive difficulties, wheelchair dependent for mobility, rarely leave house, often significant worsening of symptoms with any mental or physical exertion;
- **Very severe** – No activities of daily life, bed-bound, unable to tolerate noise, light sensitive, require someone else to watch, toilet and feed them.

For the study reported here, only severe and very severe ME/CFS patient samples were used for analysis. Healthy household controls were recruited (where possible) and were defined as family/non-family members that shared a living environment with the ME/CFS patient.

2.2.1.3 Sample collection and processing

The patient and household control were sent a faecal collection kit and a home visit was arranged to collect the sample. The faecal sample was collected in a FECOTAINER (Excretas Medical) and stored at 4 ˚C with an Oxoid™ AnaeroGen™ 2.5L anaerobic sachet (Thermo Scientific™ AN0025A) for maximum of 24 h before transport to the laboratory.

2.2.2 Faecal DNA extraction

For each sample, approximately 250 mg of faeces was thawed at room temperature and DNA extracted using MP Biomedicals™ FastDNA™ SPIN Kit for Soil (CAT:11492400) according to the manufacturer's instructions. Briefly, the faecal samples were homogenised in Lysing Matrix E tubes (CAT: 11452420) using FastPrep® 24 Classic Instrument (CAT: 116004500). Proteins and impurities were removed, and pure DNA eluted via a column-based method. The resulting DNA was eluted into DNase/Pyrogen-Free Water and stored at 4 ˚C. DNA from *Bacteroides thetaiotaomicron* VPI-5482 (GenBank accession PRJNA399) and *Lactococcus cactus* subsp. *cremoris* MG1363 (GenBank accession AM406671.1) was obtained from Dr Regis Stentz and used as a positive control for sequencing.

2.2.3 Metagenomic sequencing

The faecal DNA was sequenced by Novogene on the Illumina HiSeq (2 x 150 bp PE) and the library was prepared using the TruSeq DNA PCR-free kit.

2.2.4 Metagenomic data processing

The metagenomic data were processed by Lesley Hoyles at Imperial College London (UK Med-Bio hardware, MRC MR/L01632X/1). The quality of the sequence data was assessed for all samples using fastqc (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The data were trimmed by the sequence provider and were of high quality (Q≥30) and did not need further clean up. Reads were processed as described previously[6].

Human DNA was removed from samples by mapping reads against the human genome (hg38; GRCh38) using BWA-MEM (v. 0.7.17-r1188) with default settings for paired-end read data[7]. Taxonomic abundance and read count data for archaea and bacteria were generated using Kraken2 2.0.8-beta and the pre-complied Kraken2 GTBD_r89_54k index (downloaded May 2020) available from https://bridges.monash.edu/ndownloader/files/16378439[8,9]. DNA from *Bacteroides thetaiotaomicron* VPI-5482 (GenBank accession PRJNA399) and *Lactococcus cactus* subsp. *cremoris* MG1363 (GenBank accession AM406671.1) was included with the patient samples for a positive control. Kraken2 showed the positive-control sample to contain only reads from these two species (data not shown); consequently, this sample was not further examined.

The human-filtered, paired-end read data for this project were deposited at DDBJ/ENA/GenBank under BioProject accession PRJNA4788719. The mean number of read pairs per sample was 18,024,144 +/- SE 306,710.

2.2.4.1 Microbial gene richness

Microbial gene richness was determined according to Hoyles at al.[6]. To account for differing sequencing depth and technical variability, 10 million randomly selected reads from each sample were mapped to the non-redundant gene catalogue (of 6,091,137 genes). The mean number of genes was calculated over 30 random drawings.

2.2.4.2 Metagenome assembly

Metagenome assembly was carried out in two rounds using SPAdes (v.3.11.1), with an initial assembly carried out for each sample[10]. Representation of low-abundance sequences was improved through the use of pooled unassembled reads to complete a second round of assembly[6]. *Ab initio* gene prediction was carried out on assembled contigs using MetaGeneMark (v.3.38)[11,12]. The predicted genes were translated, and the protein sequences clustered using the cluster-fast method of UCLUST (v.7.0.10.90_i86linux64) with a 95 % identity cut off[13]. A non-redundant gene catalogue was generated from the centroid sequences of each cluster for downstream analysis. The reads were aligned against the non-redundant gene catalogue using BWA-MEM to generate gene abundance, determining the number of reads mapped to each gene sequence and normalising as described previously[7]. Functional (Kyoto Encyclopedia of Genes and Genomes (KEGG)) annotation was achieved by mapping the non-redundant gene catalogue to eggNOG-mapper (v.4.5.1) with the default settings[14-17].

2.2.4.3 Creation of metagenome-assembled genomes (MAGs)

MAGs were created by Lesley Hoyles. All forward and all reverse reads for the metagenomic dataset were concatenated. The two read files were assembled using MEGAHIT (v.1.2.9; --min-contig-len 500), generating a total of 1,140,008 contigs[18]. MAGs were created using MetaBAT 2 (v.2.12.1; -t 20 -m 1500 -v –unbinned -minContigDepth 2)[19]. Summary statistics (e.g. completeness, contamination, taxonomy) of the 668 MAGs were generated using MAGpy[20]. For each MAG, the majority taxonomic assignment in the diamond report generated by MAGpy was identified; only contigs affiliated with this taxonomic assignment were retained. Quality of the filtered MAGs (i.e. completeness, contamination) was assessed using CheckM (v.1.0.18), while tentative taxonomic assignments were made using sourmash (v.3.3.0) following guidelines at https://sourmash.readthedocs.io/en/stable/tutorials-lca.html (it should be noted that sourmash was not able to assign taxonomy to several of the MAGs using genbank-k31.lca.json.gz)[21,22]. MAGs were designated as low (n=437), medium (n=199) or high (n=32) quality with respect to completeness and contamination according to the recommendations of Bower et al.[23].

High-quality MAGs were compared against the representative MAGs (n=4,545) generated by Almeida et al. (2021) for the unified catalogue of genomes from the human gut microbiota[24]. The high-quality MAGs were annotated using Prokka (v.1.13), and the annotated representative MAGs were downloaded from ftp.ebi.ac.uk/pub.databases/metagenomics/mgnify_genomes/human-gut/v1.0/uhgg_catalogue/[25,26]. PhyloPhlAn (v.0.99) was used to generate a phylogenetic tree, which was visualised and annotated using iTol (v.4), with additional editing done with Adobe Illustrator[27].

Similarity of the high-quality MAGs to their closest phylogenetic relatives was determined by assessing average nucleotide identity (ANI) using FastANI[28].

2.2.5 Statistical analysis

Normality of the data was tested using Shapiro-Wilks normality test (stats v.3.6.2.) and visualised using histograms in R (v.3.5.2) to confirm non-parametric tests were appropriate for the data. To determine the difference in microbial gene richness, a boxplot was generated and Wilcoxon signed rank test (stats v.3.6.2) with Hochberg post-adjustment performed. The taxonomic abundances were filtered to remove all taxa representing less than 1 % abundance across all samples. Taxonomic abundance was displayed as stacked bar charts using ggplot2 (v.3.2), reshape2 (v.0.8.8) and scales (1.0). Alpha diversity was determined using Shannon index and Simpson index (vegan v. 2.5.6)[29,30]. Beta diversity was assessed using permutational multivariate analysis of variance (PERMANOVA) with Bray-Curtis distance matrix with 999 permutations and permutational analysis of multivariate dispersions (PERMDISP) (vegan v. 2.5.6)[31,32]. Principal coordinate analysis (PCoA) was used to visualise the dispersion of data and distance to centroid determined. Analysis of similarities (ANOSIM) with 999 permutations was also performed (vegan v.2.5.6)[33]. Non-metric multidimensional scaling (nMDS) with Bray-Curtis distance matrix plots were created to visualise the data (vegan v. 2.5.6). Restricted maximum likelihood linear model (REML) with Satterthwaitre approximation and Wilcoxon signed rank with Hochberg post-adjustment was used to assess taxonomic abundance differences between groups (lme4 v.27.1., stats v.3.6.2.) and accounting for age differences. Only specific taxa of interest were assessed using Wilcoxon signed rank test. Correlation plots were created using corrplot (v. 0.9). PERMANOVA/PERMDISP, ANOSIM and Wilcoxon signed rank with Hochberg post adjustment were also performed on the functional KEGG data.

2.3 Results

2.3.1 Patient demographics

A total of 19 samples were collected between 2016 and 2017. Of these 19 samples, 14 were severe ME/CFS patients and five were controls, with four household matched pairs (Table 2.1). There were 18 female participants and one male participant (house-matched control). The patients' ages ranged from18 to 61 years (mean: 42 ± SD 16.2). The control groups' ages ranged from 55 to 64 (mean: 59.4 ± SD
3.7).

Table 2-1: Overview of patient information collected during this study

Sample ID	Participant Age	Year of Collection	Participant Gender	Status	Matching Pair ID
C1	55	2017	Female	Control	1
C2	60	2017	Female	Control	2
C3	64	2017	Female	Control	4
C4	60	2017	Male	Control	3
C5	59	2016	Female	Control	
P1	61	2017	Female	Patient	
P2	38	2017	Female	Patient	
P3	44	2017	Female	Patient	
P4	63	2017	Female	Patient	
P5	18	2017	Female	Patient	
P6	37	2017	Female	Patient	4
P7	21	2017	Female	Patient	1
P8	27	2017	Female	Patient	2
P9	58	2017	Female	Patient	3
P10	56	2017	Female	Patient	
P11	54	2017	Female	Patient	
P12	57	2016	Female	Patient	
P13	35	2016	Female	Patient	
P14	20	2016	Female	Patient	

2.3.2 Microbial gene richness

Microbial gene richness assesses the number of unique microbial genes present in a metagenome[6]. Wilcoxon signed rank test revealed microbial gene richness was significantly decreased in the patient cohort compared with the controls (Figure 2.1), indicative of reduced microbial diversity within the microbiota of ME/CFS patients. Previous studies reported a lower species richness in ME/CFS patients compared to controls but did not comment on reduced functional richness[34,35].

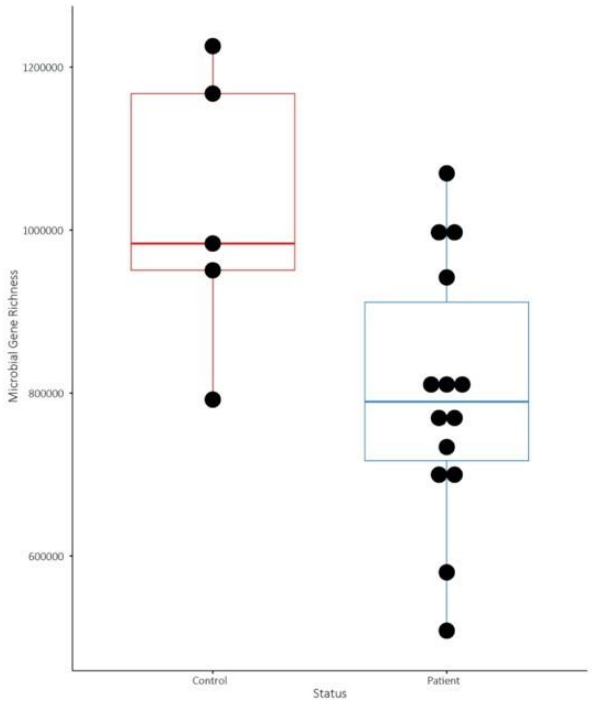

Figure 2.1: Microbial gene richness of gut metagenomes of ME/CFS patients and controls

The microbial gene richness was generated by determining the number of microbial genes from a subsample of each sample. The y axis shows the number of genes and the x axis shows the status of the individual (control or patient). The individual data points are also shown. Red, controls;blue, patients.

2.3.3 Taxonomic abundance

To assess the overall relative abundance of each taxon with the microbiota, stacked bar charts for each taxonomic level were created (phylum to species). For easier visualisation, only the top taxa (according to relative abundance) were shown, and the remaining taxa were grouped together into 'Other'. It should be noted that several of the taxa within this section have alphabetical suffixing (e.g. *Firmicutes_A, Bacillus_A,* etc). This is due taxa naming in the GTDB used for taxonomic annotation in this study. Parks et al (2018) suggested a standardized bacterial taxonomy using genome-based phylogeny and determined many current taxonomic ranks are polyphyletic[74]. Therefore, polyphyletic taxa retained the name with alphabetical suffixing (e.g. *Bacillus_A, Bacillus_B,* etc) until extensive phylogenetic can be performed to resolve the issue. At phylum level across all samples the most abundant taxa were *Firmicutes_A* and *Bacteroidota* (Figure 2.2). Five patients appeared to have a high relative abundance of *Actinobacteriota.*

Interestingly, one patient (P12) had a lower relative abundance of *Bacteroidota* (4.75 %) and higher *Euyarchaeota* (3.34 %) compared to the other patients.

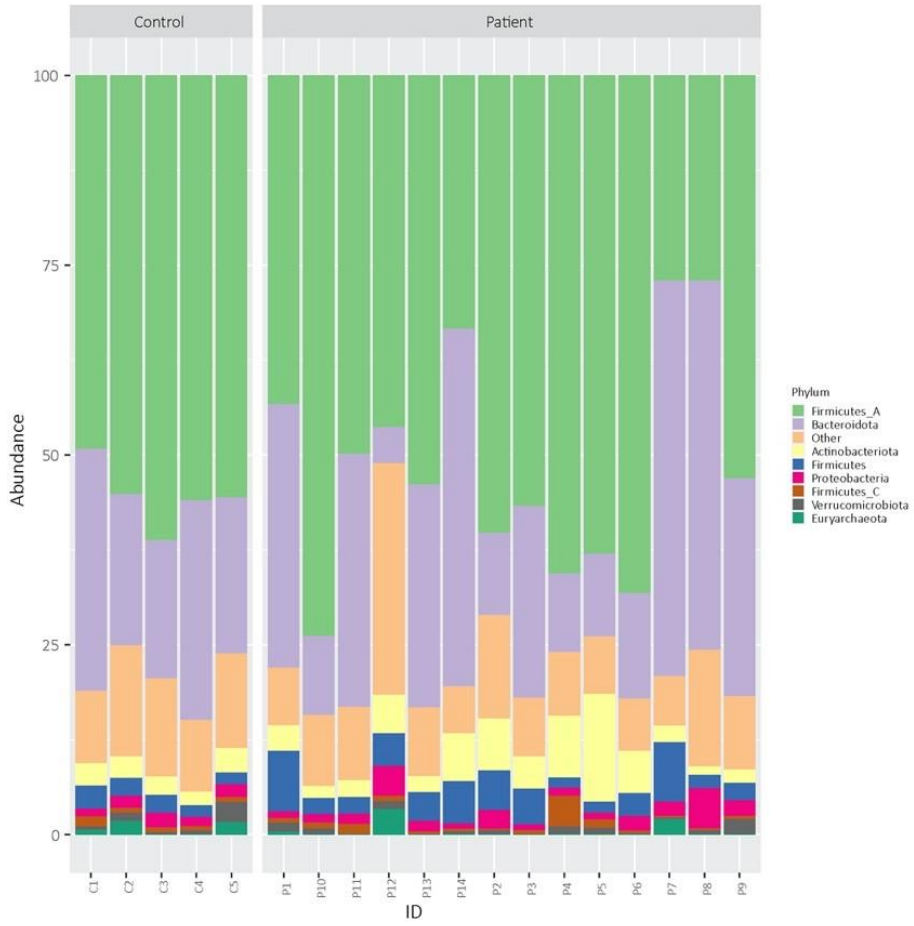

Figure 2.2: Microbiota relative abundance at phylum level for ME/CFS patients and controls

The y axis shows the relative abundance (%) of each taxon present in a sample and x axis shows the individual sample. The 'Other' portion represents low relative abundance taxa grouped together for easier visualisation. The colour of the bars corresponds to taxa found within the legend.

Clostridia and *Bacteroidia* were the most abundant taxa across all samples at the class level, with no distinct differences between groups (Figure 2.3). The patient group appeared more heterogenous whereas the control group showed consistent abundances. For example, the patient group showed differing abundances of *Bacilli* (1.37 - 7.99 %) and *Actinobacteria* (0.4 - 12.21 %). P12 also exhibited a vastly different taxonomic profile to the other patients and had the highest relative abundance of *Methanobacteria* (3.33 %).

At the order level, the three most abundant taxa across all groups were *Lachnospirales*, *Bacteroidales,* and *Oscillospirales* (Figure 2.3). As observed at class level, the patient group appeared to show more relative abundance diversity and was less homogenous than the control group. Patients showed a wide range of abundances for *Actinomycetales* (0.2 - 11.96 %), *Lactobacillales* (0.12 - 7.43
%) and *Coriobacteriales (*0.62 - 4.86 %).

Lacnospiraceae and *Bacteroidaceae* were the most abundant taxa across all samples (except P12) at the family level (Figure 2.3). No taxa appeared to be significantly increased or decreased in the patient group compared to the controls. However, the patient group appeared to be more heterogenous as noted above for higher taxonomic levels. The patient group showed a wide range of abundances for *Oscillospiraceae* (1.29 - 14.22 %), *Rickenellaceae* (0.02 - 12.26 %) and *Bifidobacteriaeae* (0.02 - 11.81 %). The highest taxonomic group in P12 was classified as the 'Other' category and showed a low relative abundance level for various other taxonomic groups not displayed in other controls or patients.

The most abundant genus across all samples (except P12) was *Bacteroides* (Figure 2.4). The control group showed overall consistency with the *Blautia_A*, *Bacteroides_B*, *Faecalibacterium* and *Alistipes* being found in similar abundances. As noted previously, the patient group exhibited a high level of heterogeneity. For example, P11 showed the highest relative abundance of *Prevotella* (13.94 %) compared to the remaining patients (0.04 - 3.92 %). Furthermore, *Agathobacter* in P6 was present at a relativeabundance of 7.6 % and ranged from 0.07 to 3.29 % in the remaining controls.

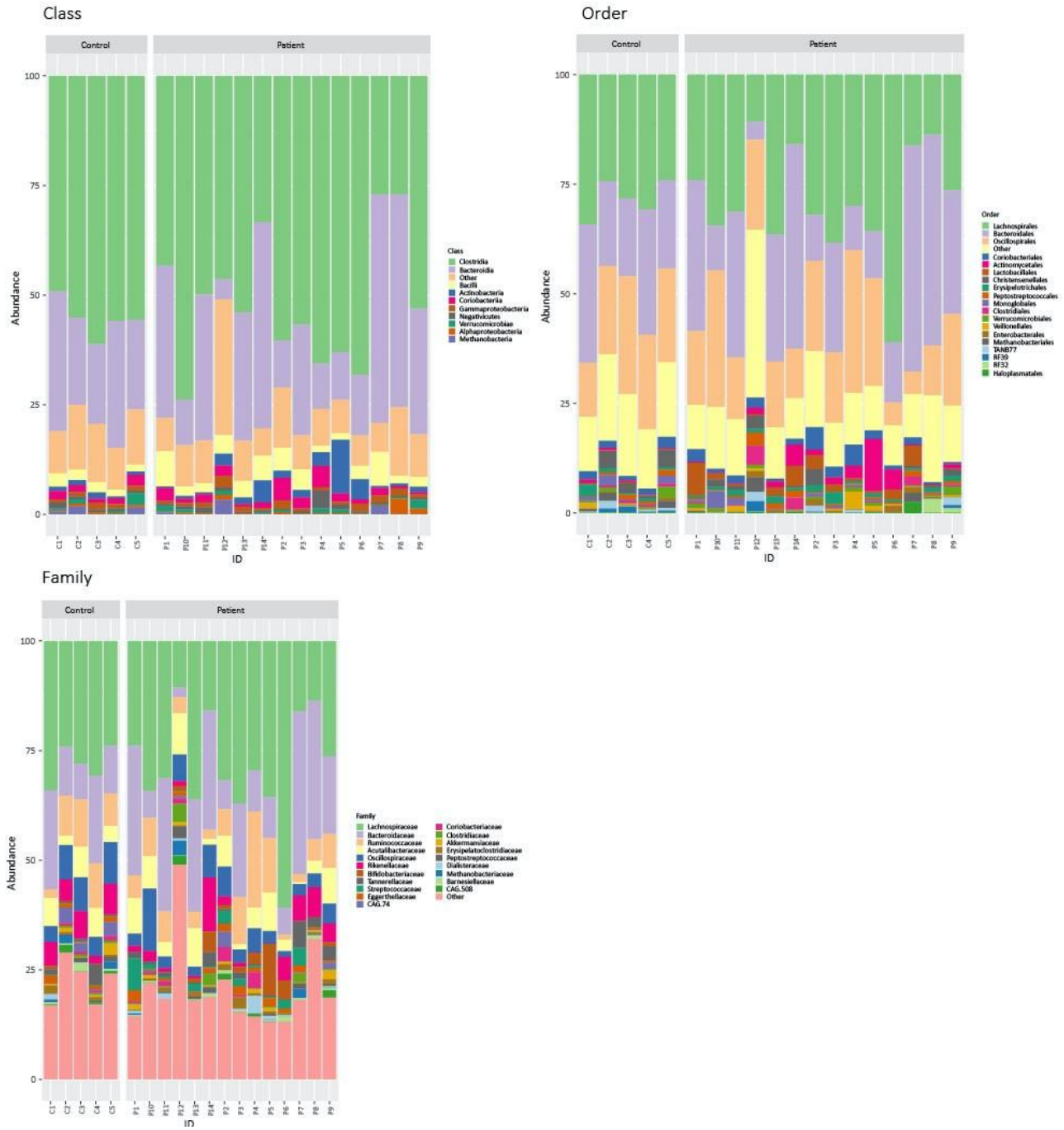

Figure 2.3: Microbiota relative abundance at class, order and family levels for ME/CFS patients andcontrols

The y axis shows the relative abundance (%) of each taxon present in a sample and x axis shows the individual sample. The 'Other' portion represents low relative abundance taxa grouped together for easier visualisation. The colour of the bars corresponds to taxa found within the legend.

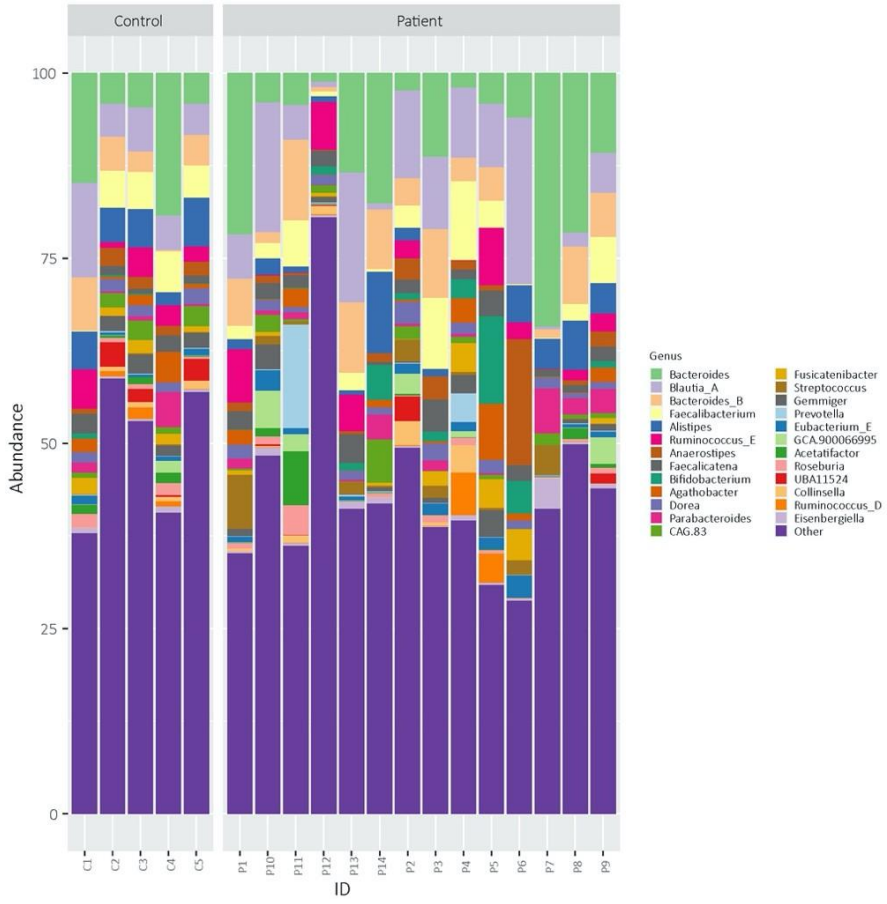

Figure 2.4: Microbiota relative abundance at genus level for ME/CFS patients and controls

The y axis shows the relative abundance (%) of each taxon present in a sample and x axis shows the individual sample. The 'Other' portion represents low relative abundance taxa grouped together for easier visualisation. The colour of the bars corresponds to taxa found within the legend.

Additionally, stacked bar charts were created for the four patient samples to the matched household control. At phylum level, matched pair 1 and 2 (P7 and P8) showed an increased relative abundance of *Bacteroidota* and decreased relative abundance of *Firmicutes_A* compared to the controls (Figure 2.5).
Similarly, matched pair 1 and 2 (P7 and P8) exhibited a noticeable decrease in *Clostridia* and an increase in *Bacteroidia* compared to controls (Figure 2.6).

Order level analysis showed an increase in *Bacteroidales* in matched pair 1 and 2 patients and *Lachnospirales* in matched pair 1, 2 and 3 (P7, P8 and P9). The patient in matched pair 4 (P6) exhibited an increased relative abundance of *Lachnospirales* compared to the control (Figure 2.6).

Matched pair 1, 2 and 3 (P7, P8 and P9) exhibited an increase in *Bacteroidaceae* and decrease in *Lachnospiraceae* compared to the matched controls (Figure 2.6). Whereas P6 in matched pair 4 showed a marked increase in *Lachnospiraceae* compared to the matched control. As noted in the higher taxonomic levels, matched pair 1 and 2 (P7 and P8) showed an increase in *Bacteroides* relative abundance compared to the matched controls (Figure 2.7). The patient in matched pair 4 (P6) exhibited a slight increase in *Bacteroides* but also a large increase in *Blautia_A* compared to the matched control. This large increase in *Blautia_A* was not seen in other matched patient samples. Matched pair 2 and 4 controls (C2 and C3) had a higher *Faecalibacterium* compared to the matchedpatients.

2.3.3.1 Alpha diversity

The alpha diversity within samples was investigated using Shannon index (*H*') and Simpson index (*D*). Shannon's and Simpson's diversity indexes aim to quantify the diversity of a single community sample, while considering both richness and relative abundance[36]. The Simpson index places more emphasis on dominant taxa and the Shannon index places more emphasis on richness (i.e. the number of unique species present in a sample)[29,30].

Statistical testing showed a significant difference for Shannon index (p value = 0.0258), while Simpson Index was not significant (p value = 0.08703) (Figure 2.8). For the patient group Shannon index ranged from 0.78 to 0.94 and the control group ranged from 0.90 to 0.95. Shannon diversity for the patient group showed a larger range (2.3 – 3.3) than the control group (2.8 – 3.4). Due to the conflicting results, it is unclear if the group microbial compositions are significantly different. However, the wide range of diversity indexes seen within the patient group suggest a heterogenous

patient cohort.

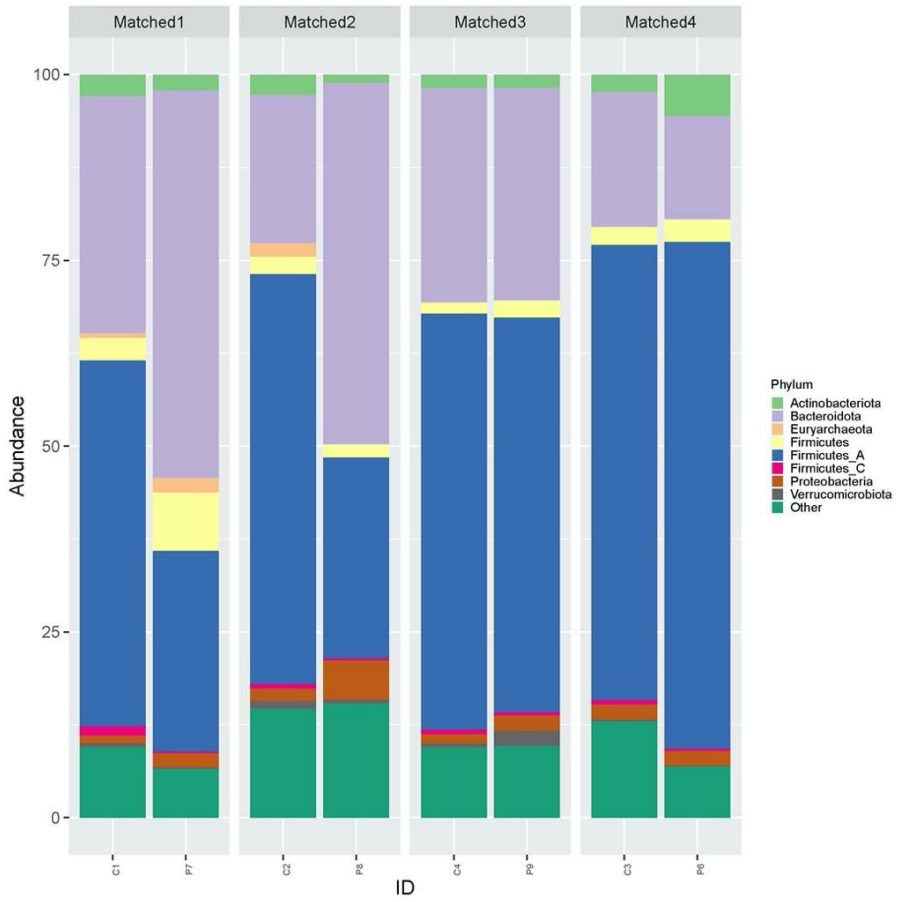

Figure 2.5: Microbiota relative abundance at phylum level for matched ME/CFS patients and controls

The y axis shows the relative abundance (%) of each taxon present in a sample. The x axis shows each matched pair and corresponding individual IDs. The 'Other' portion represents low relative abundancetaxa grouped together for easier visualisation. The colour of the bars corresponds to taxa found within the legend.

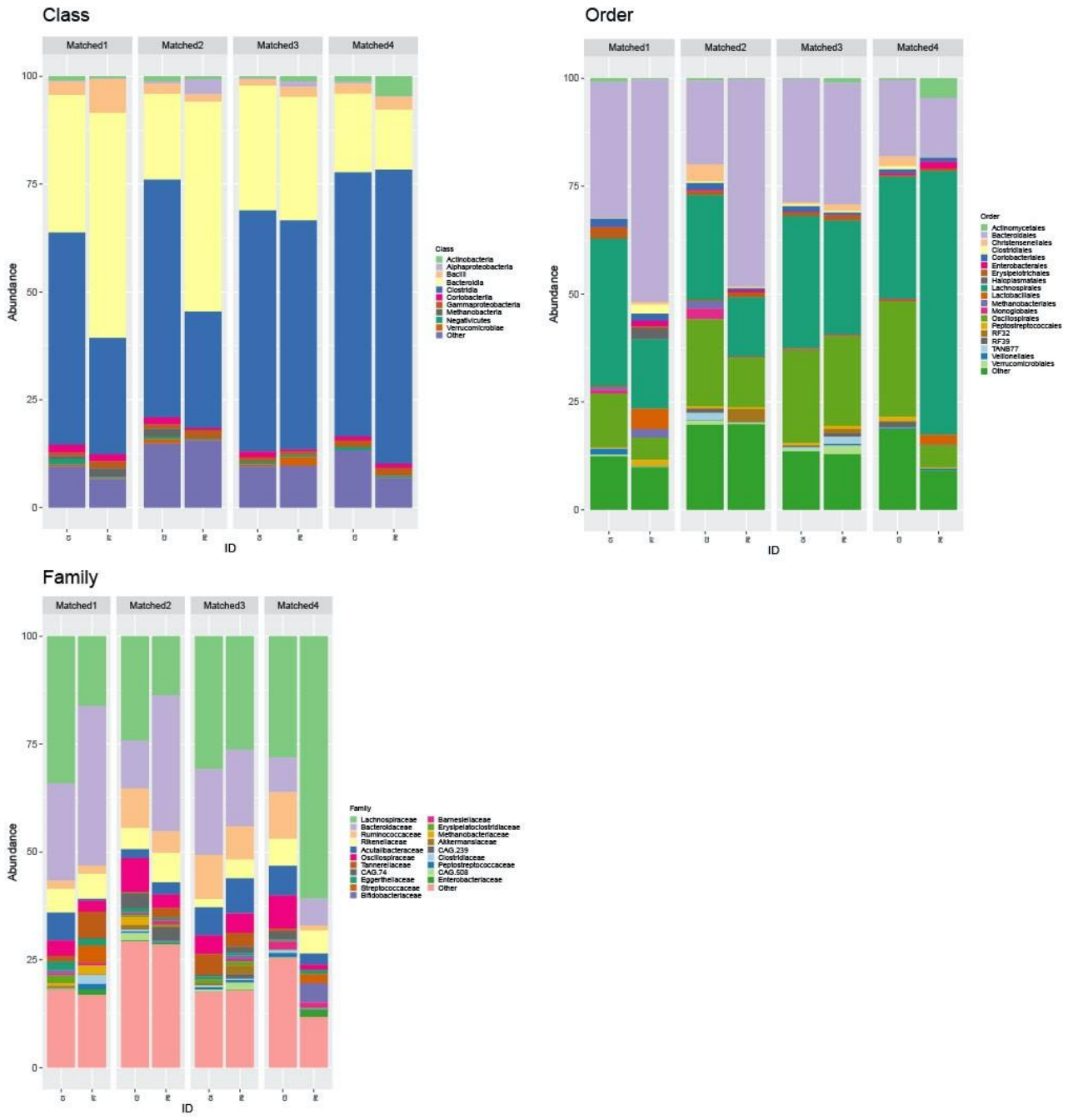

Figure 2.6: Microbiota relative abundance at class, order and family levels for matched ME/CFS patientsand controls

The y axis shows the relative abundance (%) of each taxon present in a sample. The x axis shows each matched pair and corresponding individual IDs. The 'Other' portion represents low relative abundancetaxa grouped together for easier visualisation. The colour of the bars corresponds to taxa found within the legend.

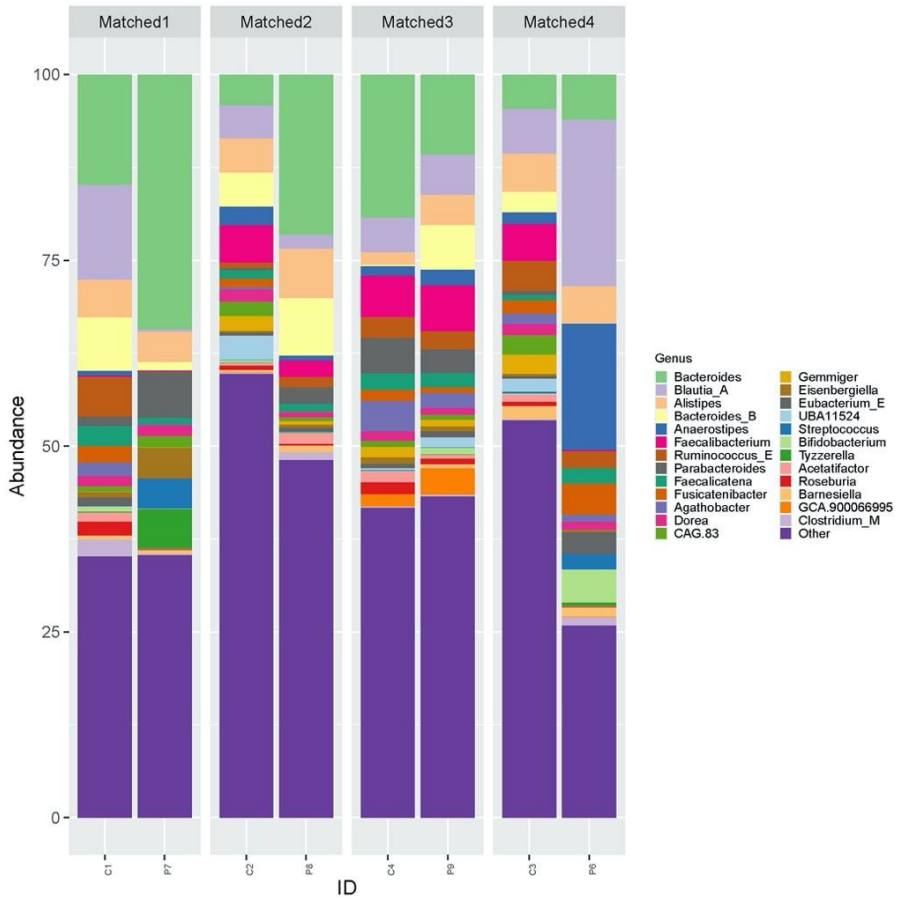

Figure 2.7: Microbiota relative abundance at genus level for matched ME/CFS patients and controls

The y axis shows the relative abundance (%) of each taxon present in a sample. The x axis shows each matched pair and corresponding individual IDs. The 'Other' portion represents low relative abundancetaxa grouped together for easier visualisation. The colour of the bars corresponds to taxa found within the legend.

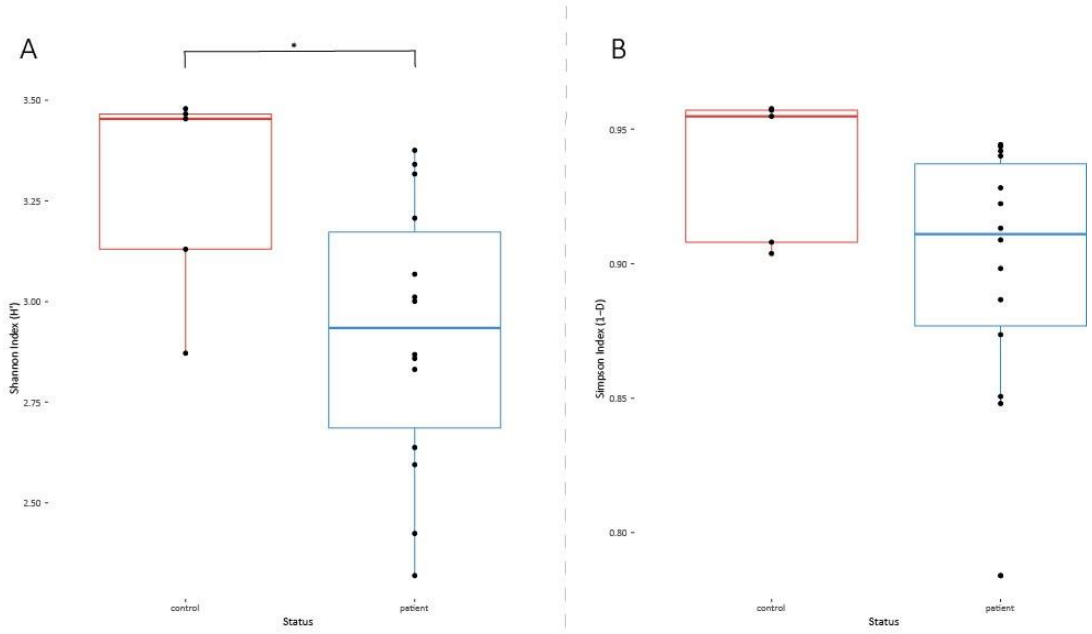

Figure 2.8: Shannon Index (A) and Simpson Index (B) for ME/CFS patients and controls

The Shannon Index and Simpson Index was generated for each sample to assess the alpha diversity between patients and controls. The Shannon Index and Simpson Index are shown on the y axis and individual status (control or patient) on the x axis. The individual data points are also shown. A significant difference in alpha diversity between the groups is represented by an asterisk.

2.3.3.2 Beta diversity

The beta diversity was assessed using PERMANOVA with Bray-Curtis distance matrix and PERMDISP for each taxonomic level. The beta diversity represents the variation of the microbial communities between ME/CFS patients and controls[30]. The distance matrix was created with Bray-Curtis as this metric takes into consideration the abundance[32]. PERMANOVA is used to assess the variance within the group and PERMDISP assess the homogeneity of group variances[31]. These results are reported as adonis (R^2 and p value) and betadisper (p value) (Table 2.2). The differences in beta diversity and beta dispersion among groups was tested by PERMANOVA (adonis) and PERMDISP. The effect size is shown by R^2 and reported as the amount of variance that can be explained by the participant status (control or patient).

Table 2-2: R^2 and p value for PERMANOVA and PERDISP at all taxonomic levels

	Adonis		**Betadisper**
Taxonomic level	**R^2**	**p value**	**p value**
Phylum	0.04	0.483	0.01*
Class	0.16	0.026*	0.002*
Order	0.06	0.323	0.015*
Family	0.06	0.3	0.03*
Genus	0.06	0.296	0.11*

These results show that very little of the variation within the data can be explained by the status of the participant, suggesting there are additional factors influencing the microbiota composition. The data were visualised using PCoA; however, nMDS was chosen for ordination visualisation. The distance to the centroid from each data point was used to determine the within group variability. The smaller the range for each group shows smaller within group variation. The distance to the centroid for each group was not significantly different at all taxonomic levels, except class (Figure 2.9). Additionally, the within group diversity appears to be low and only the class level produced a statistically significant result. Each taxonomic level displayed a high beta dispersion (PERMDISP), suggesting that while the within group diversity is low, the within group dispersion is relatively high. This is consistent with the patient heterogenous taxonomic relative abundance observed in the stacked bar charts. These results suggest that the two groups do not differ in overall composition but differ in overall heterogeneity of the composition.

Coefficient plots were created to determine which taxa contributed most to the community differences observed. Although the overall community differences were low, the top contributing taxa at each level were consistent with observations made from the stacked bar charts. For example, family level analysis showed a wide range in abundances in the patient group for *Rikenellaceae*, *Bifidobacteriaceae* and *Oscillospiraceae* and these were among the top taxa that contributed to the variation observed in PERMANOVA (Figure 2.10). Similarly, a wide variation in patient relative abundance was observed in *Prevotella, Bifidobacterium* and *Streptococcus* in the stacked bar chart and these taxa were among the top contributing coefficient (Figure 2.10).

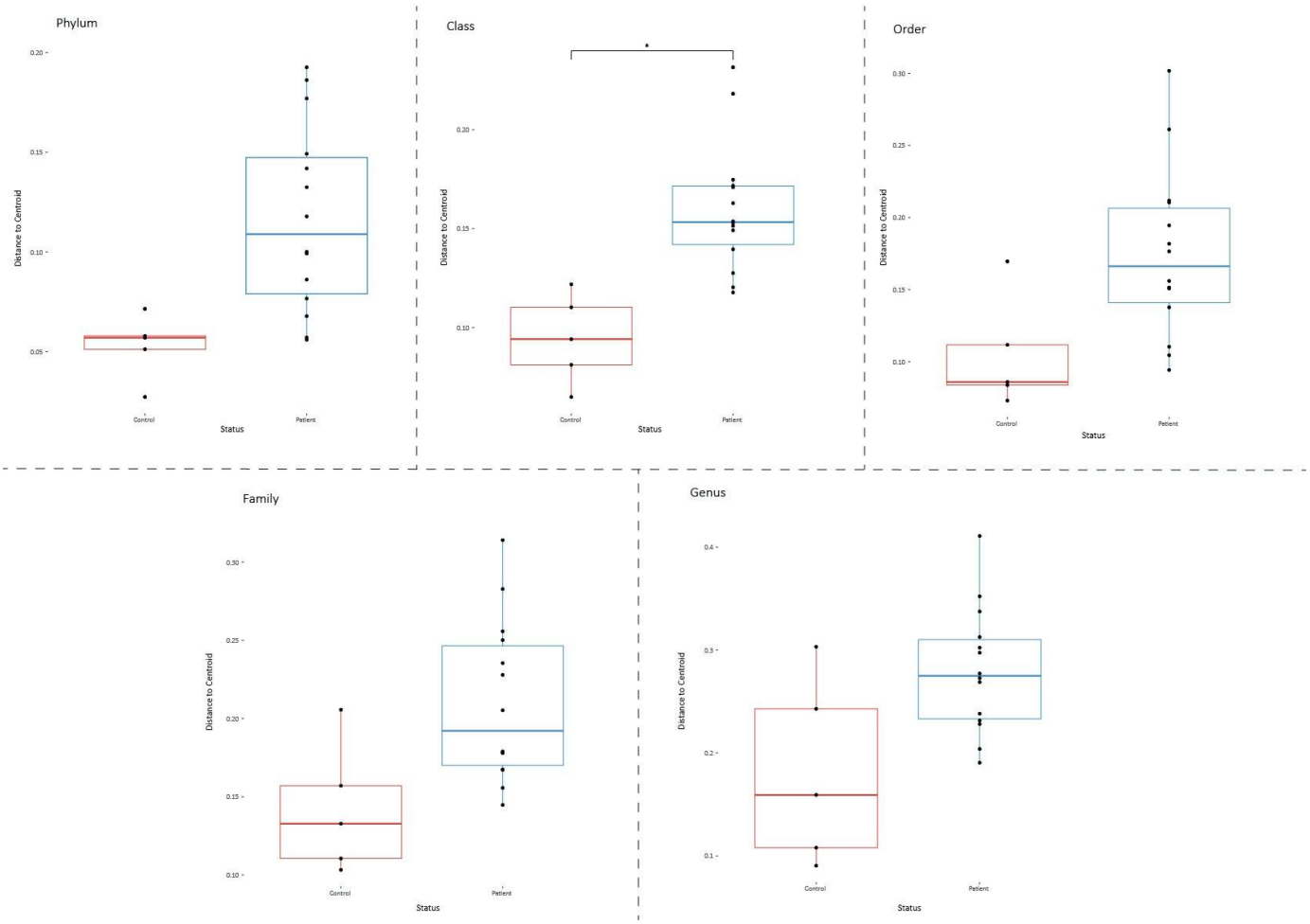

Figure 2.9: Phylum, class, order, family and genus boxplots showing the distance to the centroid for each ME/CFS patient and control datum point

The distance to the centroid from each individual sample point was generated from nMDS to investigate the beta diversity within the patient and control group. The distance is visualised in a box plot for each taxonomic level (displaced on the y axis). The smaller the boxplot spread represents a lower beta diversity within the group and low level of inter-group variability. The status of the individual is shown on the x axis (control or patient) and individual data points also shown. A significant difference of the distance to the centroid between patients and controls is represented by an asterisk in the figure.

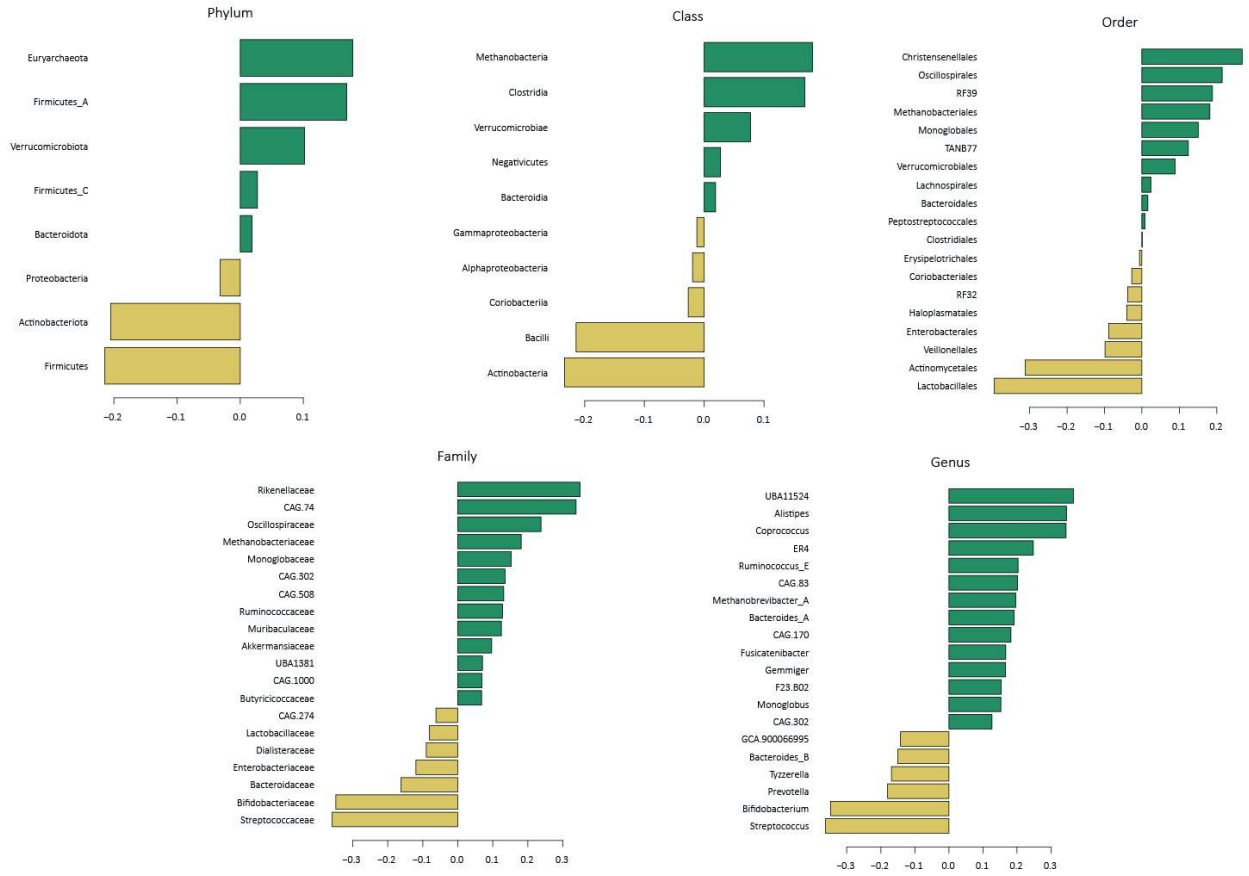

Figure 2.10: Top contributing coefficients from PERMANOVA analysis for each taxonomic level. The taxa that contribute the most to the variation observed in the PERMANOVA are represented by the bar height. The green bars represent a positive coefficient correlation and yellow bars represent a negative coefficient correlation. The x axis shows the correlation coefficient for each taxa of interest.

A nMDS plot was used to visualise the dissimilarity of the participants in a low-dimensional space. This ordination technique was chosen over the other multidimensional scaling techniques (e.g. PCoA) as it is a non-parametric and based on rank-order correlation. This type of ordination suits the data used due to the low sample size and non-parametric nature. Additionally, nMDS reports a measurement of rank-order disagreement between observed and fitted distances (stress). This value relates to the 'goodness of fit' of the multivariate data to a low dimensional space. A stress level < 0.05 is considered a good fit and high confidence in inferences made. A stress value > 0.2 suggests there are risks in interpretation. The nMDS plots revealed little to no clustering of the different groups and confirmed a high microbial heterogeneity with the patient groups (Figure 2.11). The control groups appeared to loosely cluster together at all levels (except genus). At the phylum and

class levels, three patients (P3, P11 and P13) appear to group together but this is not observed at lower taxonomic levels (Figure 2.11). Additionally, matched patient and controls were connected on the nMDS plots but no consistent grouping pattern was observed. However, matched pair 3 (P9 and C4) were grouped together at genus level but this clustering was not reported at other taxonomic levels. The stress value was relatively high (range: 0.15 to 0.195) and decreased the confidence in accurate interpretation.

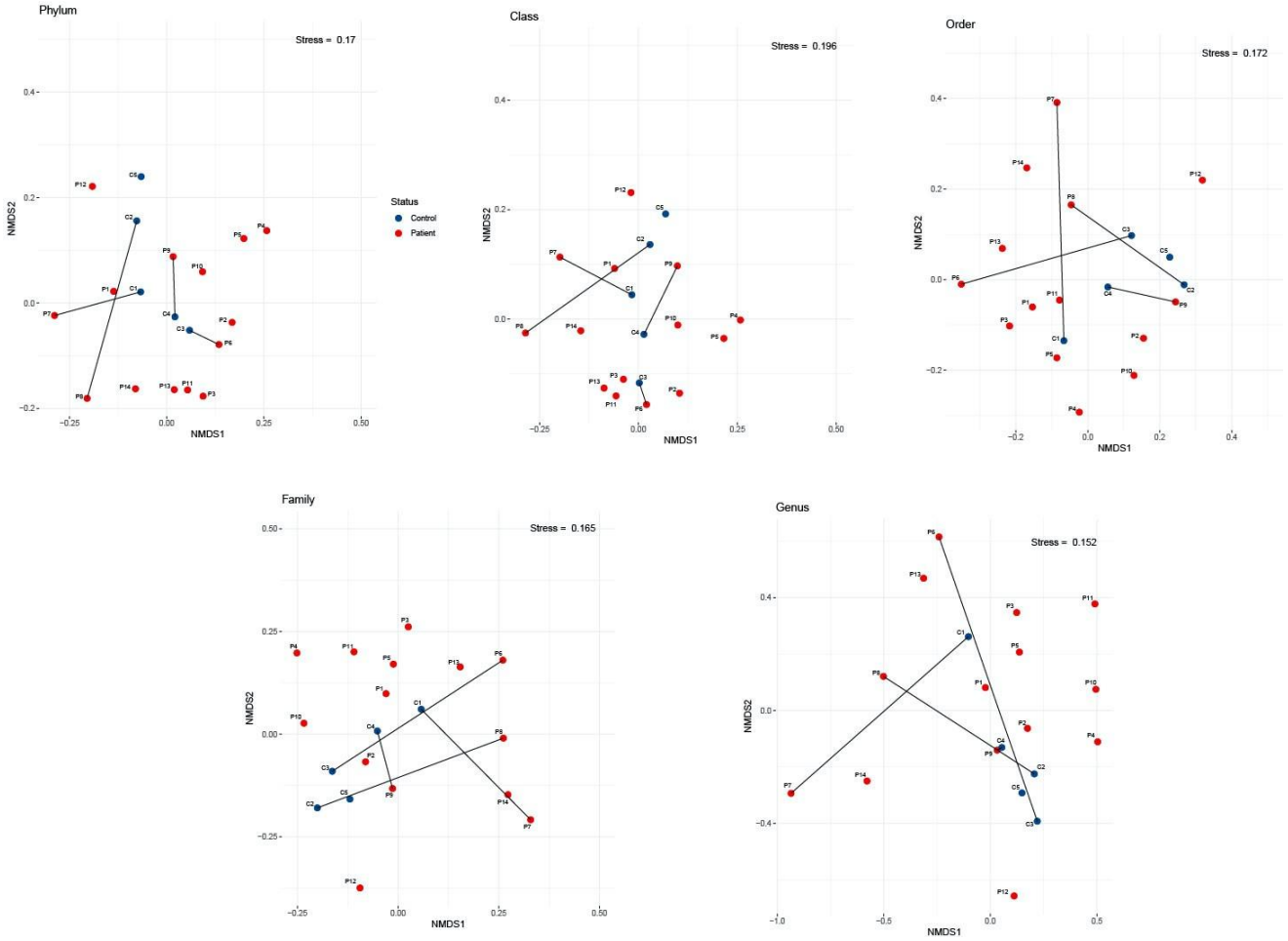

Figure 2.11: Non-metric multidimensional scaling (nMDS) plot for diversity patterns of ME/CFS patients and controls at each taxonomic level

The nMDS plot illustrates the separation of the samples determined by the differences within the intestinal microbial community. Each individual subject is represented by a dot and the sample ID displayed (patients = red dot and controls = blue dot). The lines on each plot connect the patient to its household matched control. The stress value for each plot is shown in the upper right corner and indicates how well the data is represented in reduced dimensions.

2.3.3.3 Analysis of similarities

To complement analysis undertaken in the previous section, ANOSIM was also undertaken[33]. Similar to nMDS, ANOSIM (*R*) is based on rank dissimilarities and ideal for non-parametric data. The purpose of ANOSIM is to determine whether distances between groups are greater than distances within groups. The mean values of ranked dissimilarities within and between groups are compared. The closer the *R* value to 1, the higher the dissimilarity between groups. While an R value close to 0 suggests an even distribution within and between groups. A negative R value suggests the dissimilarity is higher within groups than between groups.

The *R* value and p value were determined for each taxonomic level and confirmed the high heterogeneity within the patient group previously noted (Table 2.3). The *R* values for all taxonomic levels were negative and no significant p value was reported.

Table 2-3: Analysis of Similarities statistic (R) and p value

Taxa	***R* value**	**p value**
Phylum	-0.1771	0.927
Class	-0.2124	0.975
Order	-0.1392	0.84
Family	-0.1595	0.887
Genus	-0.2322	0.972

2.3.3.4 Linear mixed model

A REML was used to assess if the relative abundances of taxa between groups was statistically significant. This model takes into account random and fixed effects and how much variance can be captured by the random effects. Due to the low sample number, a REML was chosen over a univariate Wilcoxon test. Additionally, due to the large age variation within the cohort, this model could take into account the effects of age and participant status. The slope, intercept and p value for all taxa were reported for status and age. Any taxa with a significant REML approximation were tested using Wilcoxon test with Hochberg post hoc adjustment to confirm statistical significance. A

total of seven taxa reported a significant REML approximation for participant status. *Rikenellaceae*, *Alistipes*, *Bacteroides_A*, CAG.177 and *Coprococcus_B* were decreased within the patient group. CAG.274 and GCA.9000066995 were increased within in the patient group. However, none were statistically significant following Wilcoxon signed rank test with post hoc adjustment. Interestingly, 16 taxa showed a statistically significant REML approximation for participant age, with the relative abundance of the majority of taxa decreasing with increasing age (*Alistipes_A*, *Bifidobacterium*, CAG.177, *Coprococcus_B*, *Hungatella*, *Tyzzerella*, *Actinomycetales*, *Rikenellaceae* and *Bifidobacteriaceae*). Relative abundances of CAG.110, CAG.269, CAG.41, ER4, GCA.9000066995 and UBA1381 increased with decreasing age. These taxa were not followed up with Wilcoxon signed rank tests.

2.3.3.5 Taxonomic correlation

To visualise the correlation of taxa within the patient and control groups individually, correlation plots were created for each taxonomic level (Figure 2.12). Statistically significant correlations are marked by an asterisk and the size of the circle shows the absolute value of corresponding correlation coefficients.

A significant negative correlation between *Bacteroidota* and *Firmicutes_A* was observed in both control and patient groups at phylum level, as observed in the stacked bar charts (Figure 2.13). A negative correlation between *Proteobacteria* and *Bacteroidota* was noted in the control group but not present in the patient group. At the class level, significant correlations were similar between patient and class groups, particularly the negative correlation of *Clostridia* and *Bacteroidia* (Figure 2.12). Interestingly, the control group showed a non-significant positive correlation between *Bacilli* and *Negativicutes,* while a negative correlation was observed in patients. Overall the control group showed stronger correlations at the order level compared to the patients; however, both groups had identical significant correlations (Figure 2.12). For example, a clear negative correlation can be seen in the control group between *Bacteroidales* and RF39, and *Christensenallales* and *Lachnospirales.* Additionally, a similar observation was noted at family level as the control group correlations appeared stronger. A significant positive correlation was observed in the patient group between *Clostridiaceae* and *Methanobacteriaceae* but was a significant negative correlation in the control group (Figure 2.12). However, there were many similar significant correlations within the control and patient groups; for example, *Clostridiaceae* was negatively correlated with *Lachnospiraceae* and *Lactobacillaceae* in both groups.

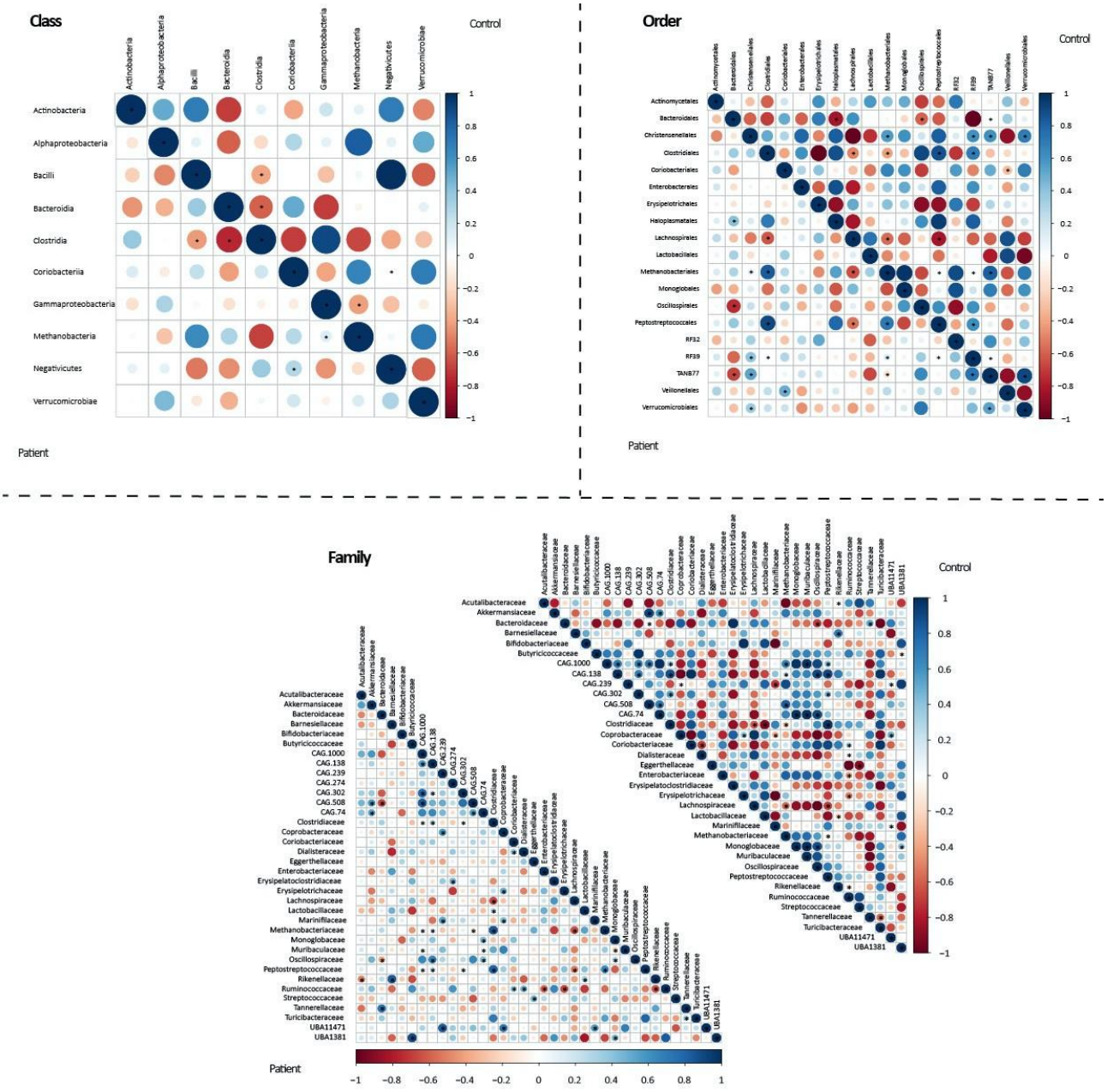

Figure 2.12: Correlation plots at class, order and family levels for ME/CFS patients and controls

The correlation plot for controls are shown in the upper right triangle and patients shown in the lower left triangle. A positive correlation is represented by a blue circle, negative correlation by a red circle and no correlation by white. Statistically significant correlations are represented by an asterisk. The larger the circle shows a stronger correlation between the two taxa within the group.

At the genus level, the patient and control groups appeared to show differing significant correlations; in addition to the stronger correlations in the controls (Figures 2.14 and 2.15). For example, within the patient group *Roseburia* was significantly positively correlated with *Acetatifactor, Agathobaculum* and CAG 41 and there was no correlation with *Prevotella* (Figure 2.14). However, the control group showed a significant positive correlation with *Acetatifactor* but negative correlation with *Agathobaculum* and *Prevotella* (Figure 2.15). CAG.41 displayed a neutral correlation with *Roseburia* in the control group. Additionally, *Faecalibacterium* was significantly positively correlated with *Agathobaculum, Lachnospira* and *Ruminococcus_C/_D* in both groups. However, patients also showed positive correlation with *Dialister* and *Collinsella.*

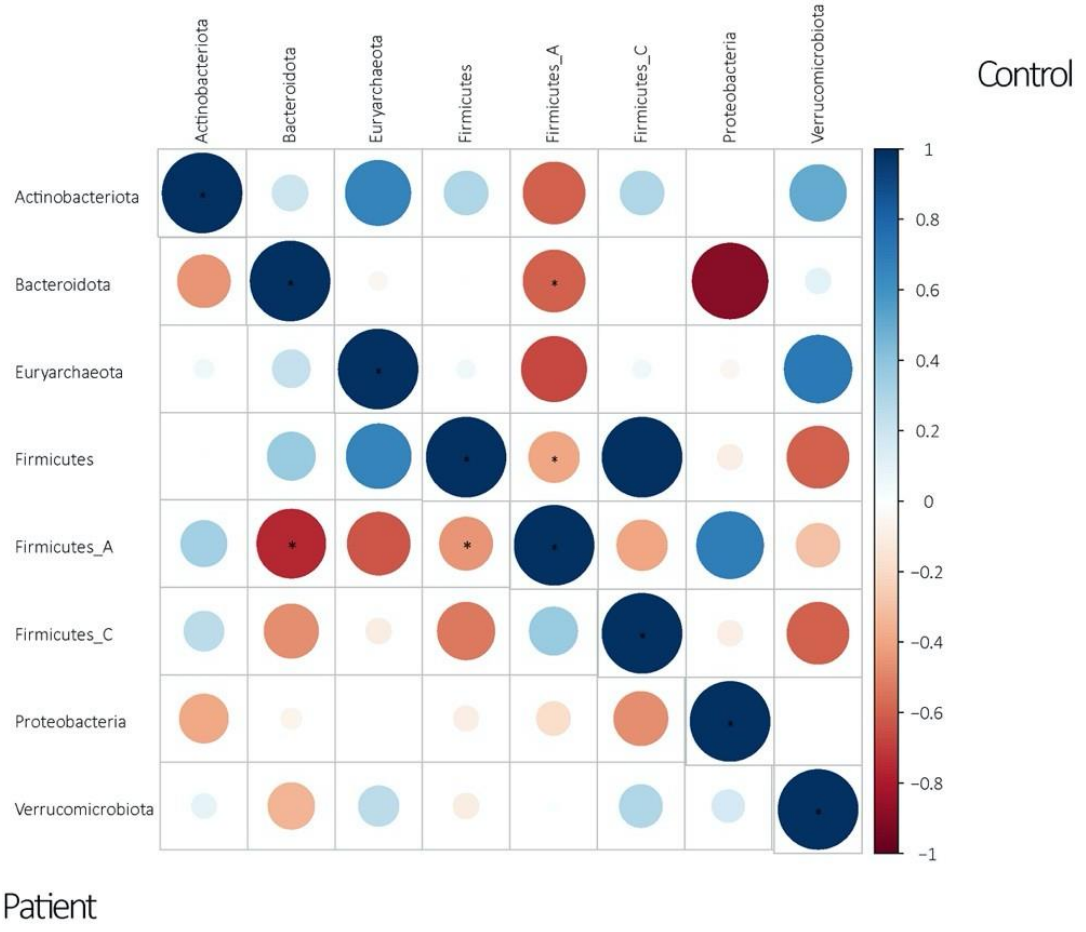

Figure 2.13: Correlation plots at phylum level for ME/CFS patients and controls

The correlation plots for controls are shown in the upper right triangle and patients shown in the lower left triangle. A positive correlation is represented by a blue circle, negative correlation by a red circle and no correlation by white. Statistically significant correlations are represented by an asterisk. The larger the circle shows a stronger correlation between the two taxa within the group.

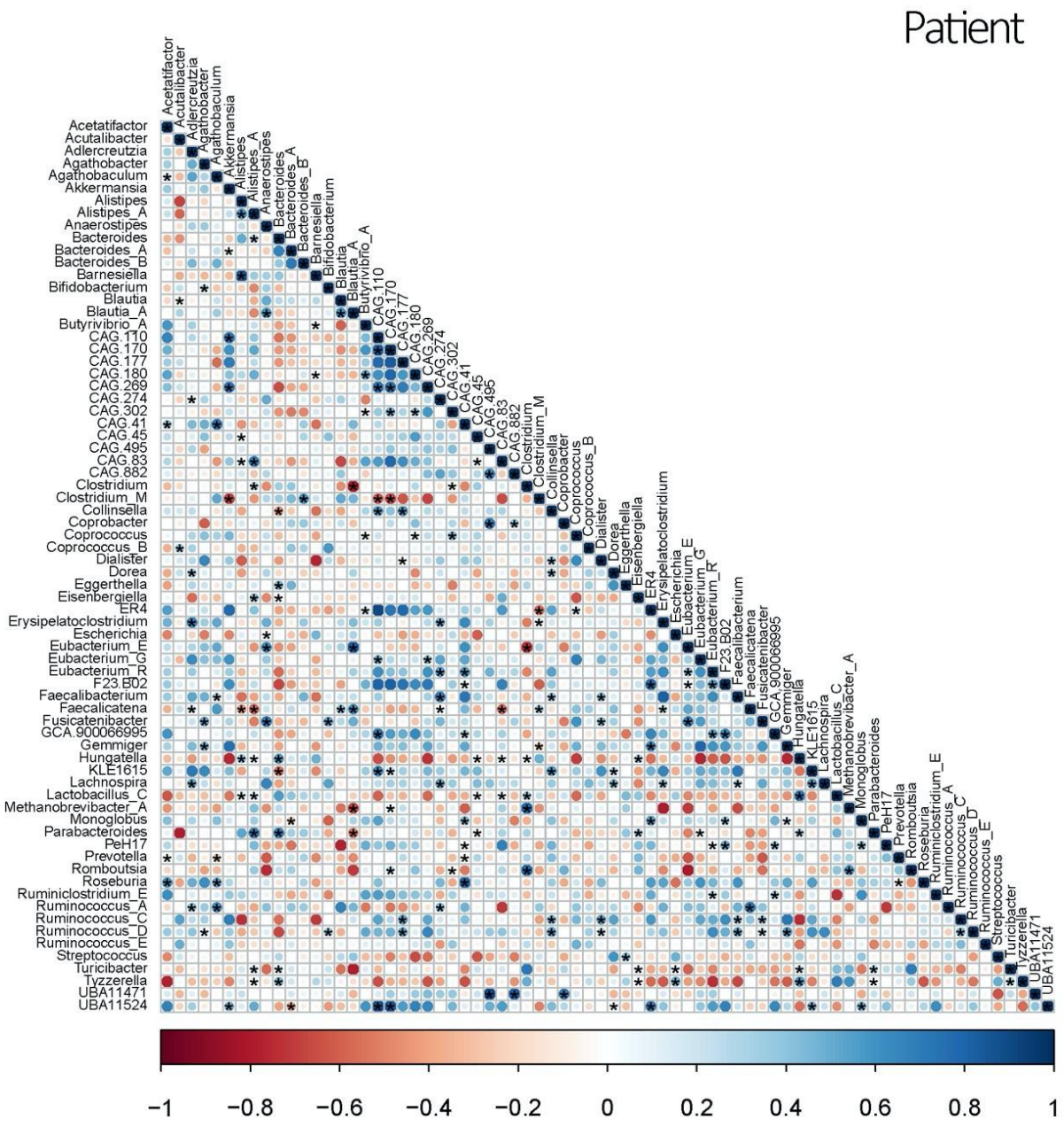

Figure 2.14: Correlation plots of ME/CFS patients at genus level

A positive correlation is represented by a blue circle, negative correlation by a red circle and no correlation by white. Statistically significant correlations are represented by an asterisk. The larger the circle shows a stronger correlation between the two taxa within the group.

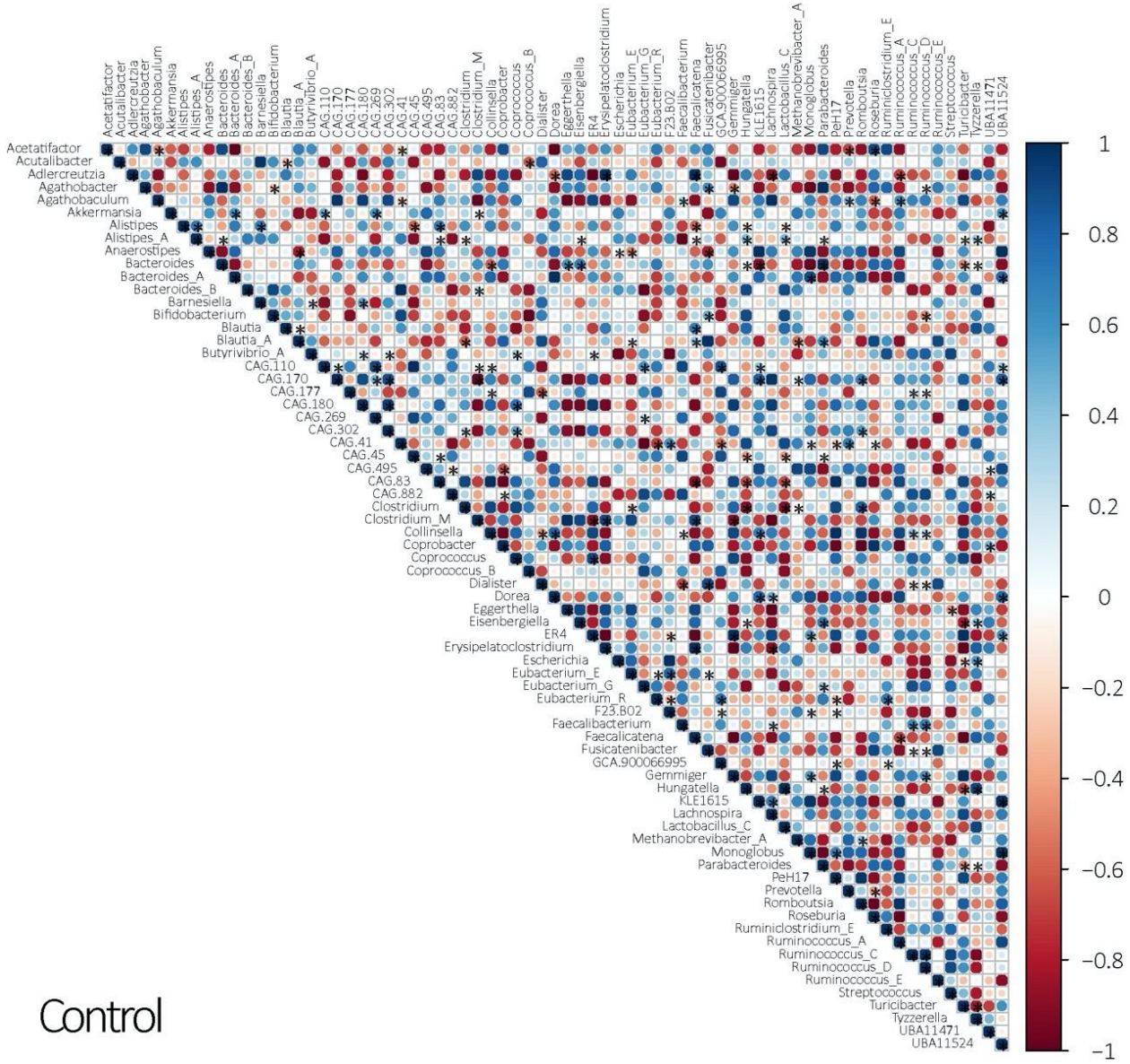

Figure 2.15: Correlation plots of controls at genus level

A positive correlation is represented by a blue circle, negative correlation by a red circle and no correlation by white. Statistically significant correlations are represented by an asterisk. The larger the circle shows a stronger correlation between the two taxa within the group.

2.3.4 Functional analysis

Functional data were created using normalised gene abundances linked with functional annotations generated using eggNOG-mapper and presented as L2 and L3 metabolism (KEGG pathway hierarchy annotations). A stacked bar chart of L2 metabolism revealed carbohydrate and amino acid metabolism were the most abundant pathways represented in all samples (Figure 2.16). As noted with the taxonomic data, the patient group showed intra-group variability.

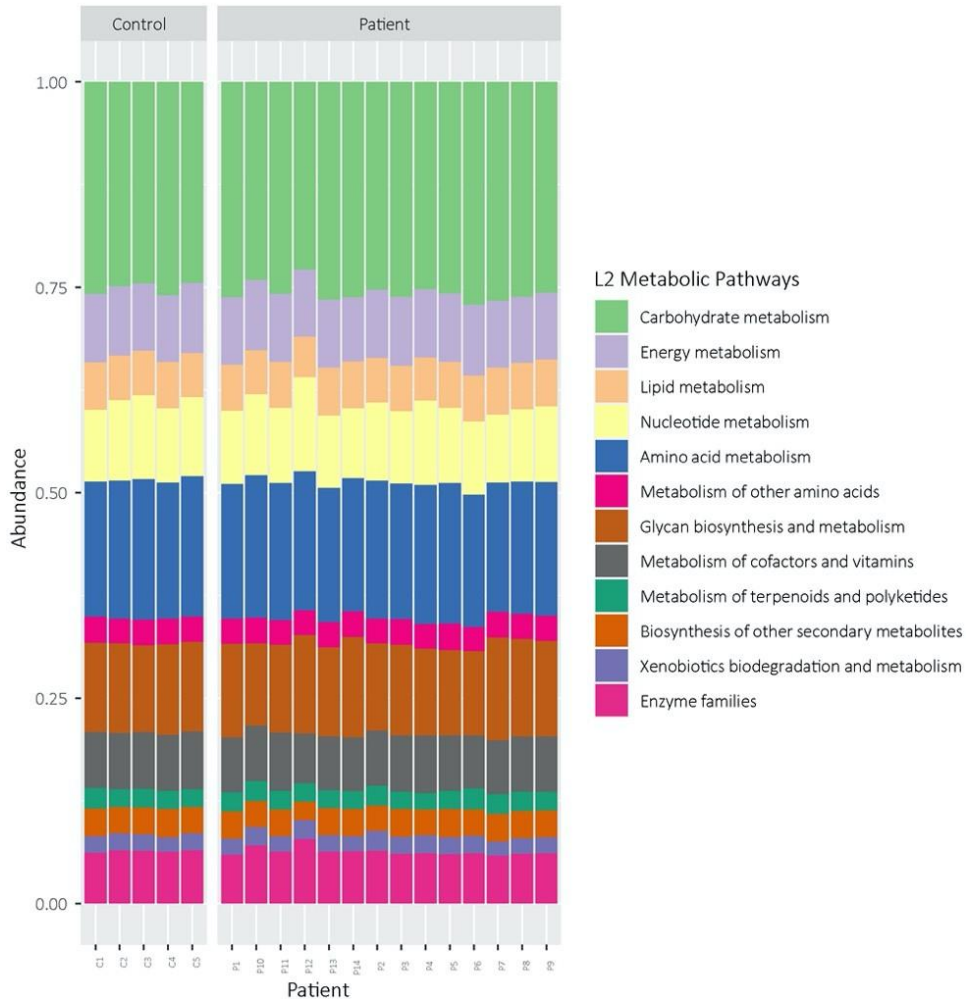

Figure 2.16: KEGG pathway representation (L2) of metagenomes of ME/CFS patients and controls
The y axis shows the relative abundance of each KEGG term represented in a sample and x axis shows the individual sample. The colour of the bars corresponds to the pathways the genes are associated with in the legend.

PERMANOVA and PERDISP revealed similar results reported with the functional data and highlighted the heterogeneity within the patient group. The dispersion of L3 metabolism within the patient group was the only statistically significant value reported (0.018) and shows the large diversity within the patient cohort (Figure 2.17). The ANOSIM *R* statistic for both L2 and L3 metabolism was negative (-0.07 and -0.0271, respectfully), suggesting diversity is larger within groups than between groups. A Wilcoxon signed rank test was performed to determine if any metabolic pathways were significantly different between patient and control groups. This was chosen over REML used for the taxonomic data due to time constraints. No metabolic pathways at L2 or L3 were significantly different between patient and control groups.

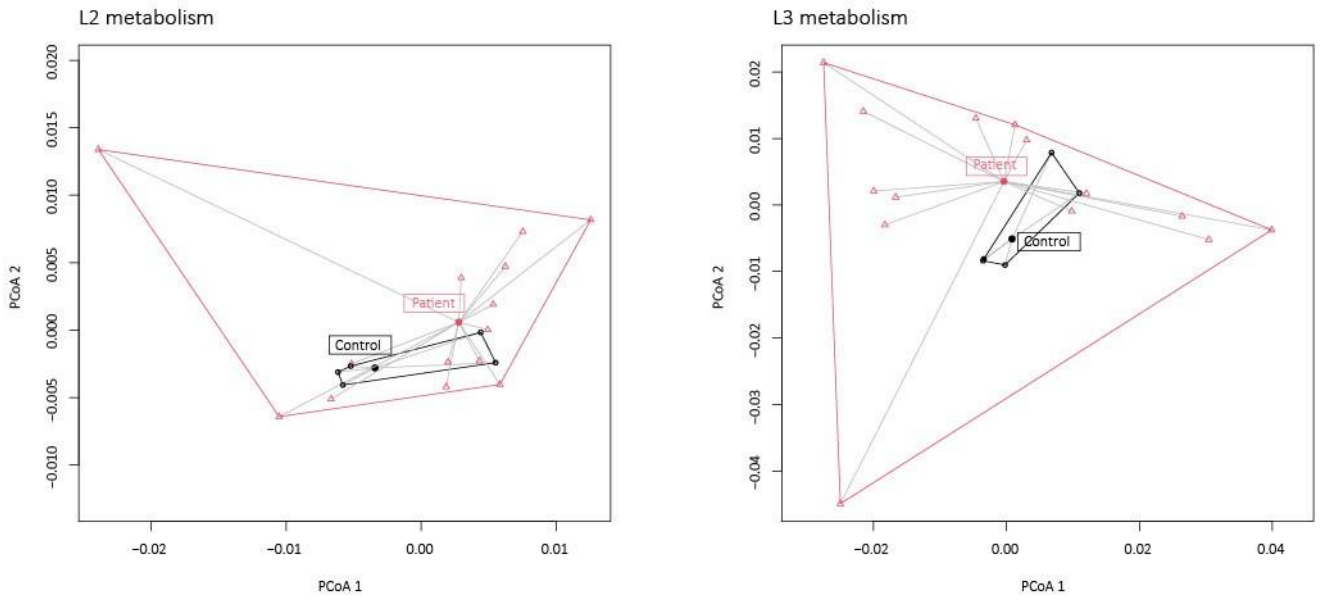

Figure 2.17: PCoA of a Bray-Curtis dissimilarity matrix of L2 and L3 KEGG metabolic pathways for metagenomes of ME/CFS patients and controls

The PCoA illustrates the diversity of metabolic pathways within the sample groups (patient and control). The centroid of each group is represented by a circle and individual data points shown by triangles.The range of each group is shown by a connecting solid line. A larger spread of the shape represents higher variation within the group.

2.3.5 Metagenome-assembled genomes (MAGs)

MetaBAT was used to generate MAGs from the concatenated metagenomic dataset[19]. A MAG is a single taxon assembly based on one or more binned metagenomes that has been determined as close representation of an existing or novel isolate[37,38]. MAGs can be used to improve metagenome taxonomic and functional annotation through addition of novel genomes to databases[38]. A total of 668 MAGs were created and the quality of MAGs was assessed with MAGpy and CheckM[20,21]. Of the 668 MAGs, 32 were high quality, 199 were medium quality and 437 were low quality. Of the high-quality MAGs (completeness > 90% and contamination < 5%; CheckM), the number of contigs ranged from 17 to 358 and genome length ranged from 1,488,234 bp to 5,809,015 bp (Table 2.4).

A phylogenetic tree was generated using the high-quality MAGs and representative MAGs to assess taxonomy (Figure 2.18). The MAGs were distributed throughout the phylogenetic tree, with the majority grouping within *Firmicutes*, *Proteobacteria* or *Bacteroidota* clades. Several MAGs clustered together within the *Proteobacteria* and *Bacteroidota* phyla, suggesting these MAGs were highly related to one another. ANI analysis revealed four MAGs represented novel species (ANI < 95% to known species). FN_CFS_73's closest relative (ANI: 92.47 %) was *Parabacteroides johnsonii*, FN_CFS_363's closest relative (ANI: 81.03 %) was CAG.353 (*Ruminococcaceae*), FN_CFS_549's closest relative (ANI: 89.01 %) was *Bacteroides_A* sp00436795 and FN_CFS_620 closest relative (ANI: 94.86 %) was *Alistipes_A ihumii* (Table 2.5).

Of the 32 high-quality MAGS, 19 belonged to the phylum *Firmicutes,* six *to Bacteroidota*, five to *Proteobacteria*, one to *Actinobacteria* and one to *Euryarchacota.* Within the phylum *Firmicutes*, nine MAGs were assigned to *Lachnospiraceae* and 3 to *Oscillospiraceae*. A MAG was also assigned to archaeal domain (FN_CFS_467) and to *Methanobrevibacter_A smithii* species.

Table 2-4: Summary statistics for the high-quality MAGs generated in this study

Bin Id	N50	nt	Contigs	Completeness (%)	Contamination (%)	tRNAs	rRNA
FN_CFS_44	53,686	1,691,571	52	94.62	0.00	38	ND*
FN_CFS_73	70,257	3,742,012	72	95.38	0.38	58	1 5S
FN_CFS_85	33,034	1,811,203	72	96.77	0.00	41	ND
FN_CFS_99	41,756	2,272,670	79	98.66	1.68	48	ND
FN_CFS_113	31,990	2,336,831	122	92.06	3.36	29	ND
FN_CFS_158	18,059	2,335,968	187	90.72	1.17	34	ND
FN_CFS_162	24,321	2,212,942	131	92.71	3.36	24	2 5S
FN_CFS_169	29,671	2,171,551	115	92.10	3.47	23	1 5S
FN_CFS_187	20,815	2,511,372	177	93.15	2.13	40	2 5S
FN_CFS_219	46,740	2,326,284	83	91.08	0.66	40	ND
FN_CFS_226	29,921	3,005,245	151	91.31	1.42	36	ND
FN_CFS_229	34,670	5,809,015	248	92.22	3.72	42	ND
FN_CFS_248	23,136	1,638,405	100	95.74	1.06	37	1 5S
FN_CFS_267	20,660	1,677,206	113	94.58	1.56	46	3 5S
FN_CFS_302	38,021	3,780,275	155	95.07	0.81	52	4 5S
FN_CFS_306	62,152	2,969,348	71	90.19	0.57	47	1 5S
FN_CFS_312	21,379	1,878,612	134	92.23	2.82	43	2 5S
FN_CFS_328	8,408	2,335,013	358	92.42	2.07	34	ND
FN_CFS_350	83,724	2,273,562	56	95.09	0.00	32	ND
FN_CFS_363	32,089	2,241,490	95	93.62	0.00	41	ND
FN_CFS_405	22,360	1,711,222	110	95.21	0.18	38	1 16S
FN_CFS_458	37,759	2,213,486	89	93.29	0.67	48	ND
FN_CFS_467	27,714	1,625,867	97	99.20	0.40	32	ND
FN_CFS_479	84,771	2,031,058	30	92.62	0.00	29	ND
FN_CFS_496	52,903	2,028,559	66	90.57	0.94	31	ND
FN_CFS_531	120,491	1,488,345	17	91.00	1.40	46	ND
FN_CFS_540	33,721	2,104,235	85	90.88	0.00	28	ND
FN_CFS_549	61,732	3,244,826	89	94.11	0.56	70	2 5S
FN_CFS_569	43,273	2,223,234	77	96.89	0.62	59	1 5S
FN_CFS_594	48,671	2,344,169	74	97.55	0.00	46	2 5S
FN_CFS_605	53,914	2,580,286	90	93.75	1.01	38	ND
FN_CFS_620	68,206	2,560,839	49	98.08	0.64	44	ND

* ND = None detected

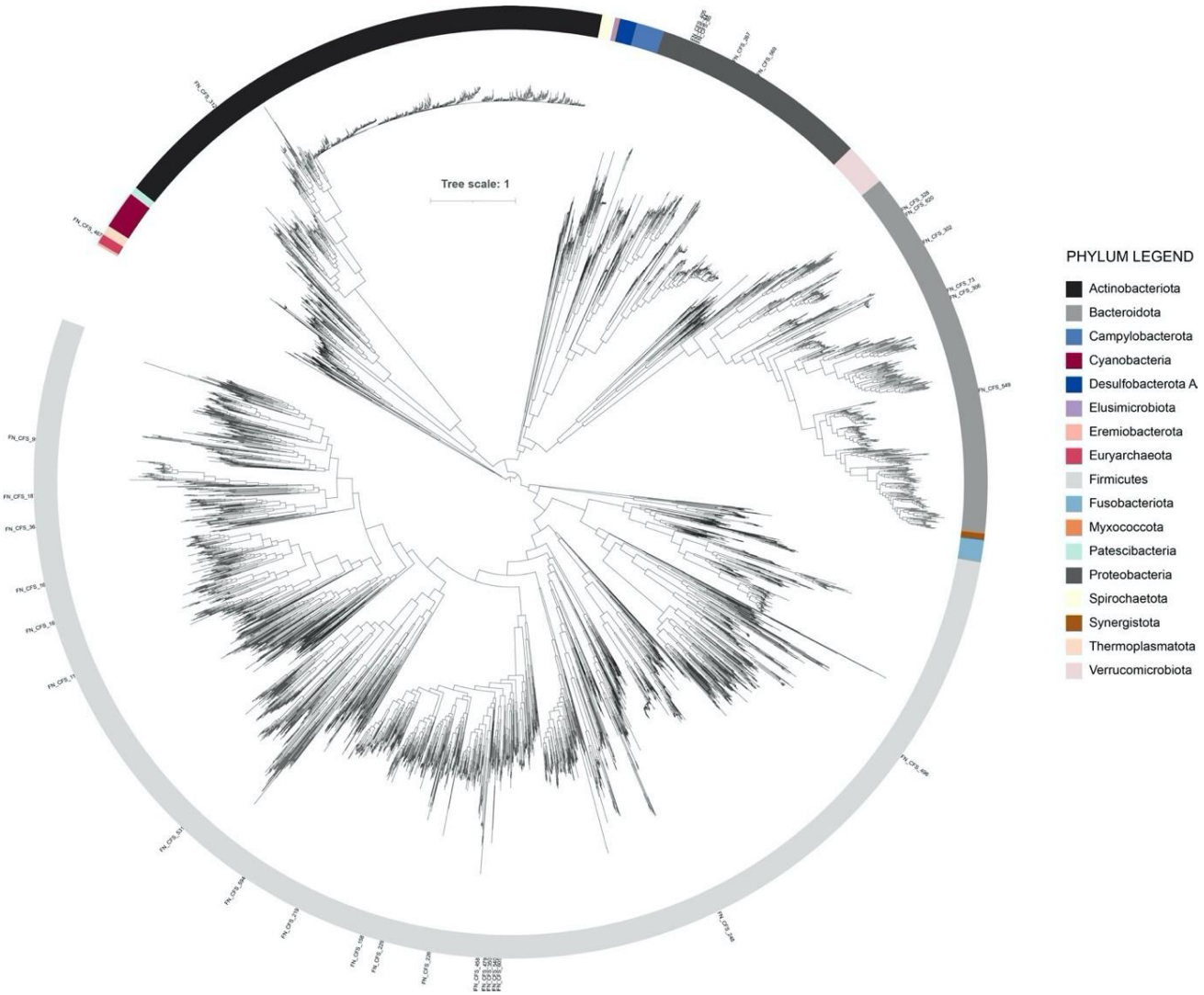

Figure 2.18: Phylogenetic tree showing the taxonomic placement of MAGs generated in this study with representative MAGs from Almeida et al.[24]

The phylogenetic tree was generated with species representative MAGs and high-quality MAGs generated in this study. PhyloPhlAn was used to determine the phylogenetic profile of each new MAG in relation to representative MAGs and phylogenetic tree visualised using iTOL. The phyla are represented by the coloured segments and displayed in the phylum legend. Each high-quality MAG generated during this study is represented by an ID in the outer circle.

Table 2-5: Closest phylogenetic relatives of the high-quality MAGs among the unified catalogue of genomes from the human gut microbiota

MAGs with pink-highlighted ANI values represent novel taxa.

Bin Id	Closest relative	ANI (%) versus closest relative	Closest relative ID (according to comparison with Almeida et al., 2021)*
FN_CFS_44	MGYG-HGUT-02616	99.20	d_Bacteria;p_Proteobacteria;c_Alphaproteobacteria;o_RF32;f_CAG-239;g_CAG-495;s_CAG-495 sp001917125
FN_CFS_73	MGYG-HGUT-00138	92.47	d_Bacteria;p_Bacteroidota;c_Bacteroidia;o_Bacteroidales;f_Tannerellaceae;g_Parabacteroides;s_Parabacteroides johnsonii
FN_CFS_85	MGYG-HGUT-02873	98.91	d_Bacteria;p_Proteobacteria;c_Alphaproteobacteria;o_RF32;f_CAG-239;g_CAG-495;s_
FN_CFS_99	MGYG-HGUT-04129	98.11	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Oscillospirales;f_Acutalibacteraceae;g_Acutalibacter;s_Acutalibacter sp000435395
FN_CFS_113	MGYG-HGUT-04341	99.05	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Oscillospirales;f_Oscillospiraceae;g_CAG-110;s_CAG-110 sp000435995
FN_CFS_158	MGYG-HGUT-00159	99.59	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Lachnospirales;f_Lachnospiraceae;g_Sellimonas;s_Sellimonas intestinalis
FN_CFS_162	MGYG-HGUT-02327	98.92	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Oscillospirales;f_Oscillospiraceae;g_Oscillibacter;s_Oscillibacter sp900066435
FN_CFS_169	MGYG-HGUT-01500	99.30	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Oscillospirales;f_Oscillospiraceae;g_Lawsonibacter;s_Lawsonibacter asaccharolyticus
FN_CFS_187	MGYG-HGUT-03876	99.48	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Oscillospirales;f_Ruminococcaceae;g_Anaerotruncus;s_
FN_CFS_219	MGYG-HGUT-02286	98.45	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Lachnospirales;f_Lachnospiraceae;g_Blautia;s_Blautia sp000436935
FN_CFS_226	MGYG-HGUT-00242	99.33	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Lachnospirales;f_Lachnospiraceae;g_Clostridium_M;s_Clostridium_M sp000431375
FN_CFS_229	MGYG-HGUT-02330	97.49	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Lachnospirales;f_Lachnospiraceae;g_Eisenbergiella;s_Eisenbergiella tayi
FN_CFS_248	MGYG-HGUT-04198	99.25	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Peptostreptococcales;f_Anaerovoracaceae;g_UBA1191;s_
FN_CFS_267	MGYG-HGUT-00567	96.49	d_Bacteria;p_Proteobacteria;c_Gammaproteobacteria;o_Burkholderiales;f_Burkholderiaceae;g_CAG-521;s_CAG-521 sp000437635
FN_CFS_302	MGYG-HGUT-00254	99.73	d_Bacteria;p_Bacteroidota;c_Bacteroidia;o_Bacteroidales;f_Marinifilaceae;g_Odoribacter;s_Odoribacter splanchnicus
FN_CFS_306	MGYG-HGUT-01391	98.39	d_Bacteria;p_Bacteroidota;c_Bacteroidia;o_Bacteroidales;f_Coprobacteraceae;g_Coprobacter;s_Coprobacter fastidiosus
FN_CFS_312	MGYG-HGUT-04041	96.58	d_Bacteria;p_Actinobacteriota;c_Coriobacteriia;o_Coriobacteriales;f_Eggerthellaceae;g_;s_
FN_CFS_328	MGYG-HGUT-03926	97.43	d_Bacteria;p_Bacteroidota;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;g_Rikenella;s_Rikenella microfusus
FN_CFS_350	MGYG-HGUT-00204	99.03	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Lachnospirales;f_Lachnospiraceae;g_Eubacterium_G;s_Eubacterium_G sp000435815
FN_CFS_363	MGYG-HGUT-00424	81.03	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Oscillospirales;f_Ruminococcaceae;g_CAG-353;s_
FN_CFS_405	MGYG-HGUT-02021	99.38	d_Bacteria;p_Proteobacteria;c_Alphaproteobacteria;o_RF32;f_CAG-239;g_51-20;s_51-20 sp001917175
FN_CFS_458	MGYG-HGUT-04317	98.88	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Lachnospirales;f_Lachnospiraceae;g_Lachnospira;s_Lachnospira sp900316325
FN_CFS_467	MGYG-HGUT-02163	98.79	d_Archaea;p_Euryarchaeota;c_Methanobacteria;o_Methanobacteriales;f_Methanobacteriaceae;g_Methanobrevibacter_A;s_Methanobrevibacter_A smithii
FN_CFS_479	MGYG-HGUT-00484	98.81	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Lachnospirales;f_Lachnospiraceae;g_Butyrivibrio_A;s_Butyrivibrio_A sp000431815

Bin Id	Closest relative	ANI (%) versus closest relative	Closest relative ID (according to comparison with Almeida et al., 2021)*
FN_CFS_496	MGYG-HGUT-01398	99.85	d_Bacteria;p_Firmicutes;c_Bacilli;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Absiella;s_Absiella sp000163515
FN_CFS_531	MGYG-HGUT-02831	96.98	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_4C28d-15;f_CAG-917;g_CAG-349;s_CAG-349 sp003539515
FN_CFS_540	MGYG-HGUT-00169	98.86	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Lachnospirales;f_Lachnospiraceae;g_Eubacterium_G;s_Eubacterium_G sp000432355
FN_CFS_549	MGYG-HGUT-03097	89.01	d_Bacteria;p_Bacteroidota;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides_A;s_Bacteroides_A sp000436795
FN_CFS_569	MGYG-HGUT-01410	98.89	d_Bacteria;p_Proteobacteria;c_Gammaproteobacteria;o_Burkholderiales;f_Burkholderiaceae;g_Sutterella;s_Sutterella wadsworthensis_A
FN_CFS_594	MGYG-HGUT-01132	98.67	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Christensenellales;f_CAG-138;g_UBA1685;s_UBA1685 sp002320595
FN_CFS_605	MGYG-HGUT-01831	98.99	d_Bacteria;p_Firmicutes_A;c_Clostridia;o_Lachnospirales;f_Lachnospiraceae;g_TF01-11;s_TF01-11 sp000436755
FN_CFS_620	MGYG-HGUT-04056	94.86	d_Bacteria;p_Bacteroidota;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;g_Alistipes_A;s_Alistipes_A ihumii

2.4 Discussion

This Chapter presents the analysis of the intestinal microbiota of severe ME/CFS patients and controls using shotgun metagenomic data. A total of 19 participants were recruited for this study (14 patients and 5 controls) and four of those patients had matched household controls. This study revealed the high level of microbial diversity within the patient cohort and further confirms the heterogenous nature of the disease. Interestingly, no significant functional differences were determined between the patient and control groups. These results suggest that the patient microbiome do not all exhibit similar 'disease markers' (e.g. a significant decrease/increase in specific bacterial taxa or function). Future microbiome studies should take into the account the heterogenous microbiome composition of patients and aim to stratify patients to decrease the wide variation seen in this study.

This study revealed a significant decrease in the microbial gene richness in the patient cohort compared to the control cohort. Armstrong (2017) and Giloteaux (2017) reported a decrease in overall bacterial relative abundance within the patient cohort, but did not report on reduced functional (microbial gene) richness[35,39]. The analysis of faecal microbial metabolism or predicted function has been utilised in several diseases to examine links to pathogenesis that may not be evident with taxonomic analysis alone[40,41]. For example, microbially-synthesised imidazole propionate (ImP) has been found to be increased in the faecal microbiome of subjects with type 2 diabetes. Additionally, the authors reported an association of high levels of ImP with unhealthy eating habits, suggesting a link between ImP and impaired glucose metabolism[40]. While this Chapter did not show any significant functional differences between patients and controls, the functional datacorrelated with the heterogeneity seen within the patient group in taxonomic analysis. Variation can be seen within Figure 2.16 with genes assigned to metabolic pathways, particularly in P10, P11 and P12. A 2021 faecal metabolomics and 16S rRNA gene-based sequencing study reported a similar metabolomic profile between patients and relative controls, compared to non-relative controls[42]. This could be attributed to the similar diet and lifestyles. The authors reported a significant association of glutamic acid in patients, compared to controls. The increase of glutamicacid in patients is of particular significance as a high accumulation of glutamate (originating from bacterial synthesis) can contribute to central nervous system damage. Several studies have attempted to characterise the faecal metabolome in ME/CFS patients and attribute a decrease/increase of various metabolites to disease manifestations[39,42-44]. For example, a 2017 study reported an increase in SCFAs in ME/CFS patients, with butyrate being statistically significant[39]. However, a 2021 study showed a significant decrease in faecal butyrate concentrations in ME/CFS[43]. Additionally, the degree of reduction in butyrate-producing bacteria in the faecal microbiota correlated with the level of patient fatigue severity. These conflicting

results highlight the current confusion within the ME/CFS-associated microbiome research field. Due to the small sample size of the cohort discussed in this Chapter, it is impossible to draw any significant conclusions from the current data regarding predicted metabolomic function.

Several studies report conflicting results with respect to taxonomic abundances and the ME/CFS faecal microbiota, and no taxon has been shown to be significantly different (in terms of abundance) between patients and controls across all ME/CFS studies (Table 2.6). The conflicting results may be in part due to differing patient selection/disease criteria, sample processing, genome sequencing and downstream bioinformatic analysis[45].

While no specific taxa were statistically different between the control and patient group in the current study, several interesting observations within the patient group were noted such as increase in intra-group diversity and weaker correlations within the patient group. Previous studies have reported a decrease in *Firmicutes* and an increase in *Bacteroidetes* in patients compared to controls, although these were not statistically significant[35,42,46]. A similar observation was noted in the current study but was not consistent across all patients as half of the patients showed a decrease in *Bacteroidetes* and an increase in *Firmicutes* or *Actinobacteria*. Interestingly, a strong negative correlation between *Bacteroidetes* and *Proteobacteria* was noted in the patient group as several patients appeared to have a reduced *Bacteroidetes* relative abundance and increased *Proteobacteria* abundance. However, this was not statistically significant. A previous study reported that the reduction of *Firmicutes* and increase of *Bacteroidetes* in patients was attributedto a decrease in several *Clostridiales* families, particularly *Lachnopiraceae*[42]. At the order level *Clostrdiales* and *Bacteroidales* were negatively correlated within the control group but slightly positively correlated within the patient group. *Oscillospirales* and *Bacteroidales* were negatively correlated in the patient group. However, a consistent reduction of *Lachnospiraceae* was not observed among the patient group.

Table 2-6: Comparison of composition alterations in ME/CFS microbiota studies

The table highlights the conflicting results reported for various ME/CFS microbiota studies adapted from Newberry et al.[45]. The up arrows represent taxa increased in patients and the down arrows represent taxa decreased in patients.

Microbial component	Lupo (2021)[42]	Armstrong (2017)[39]	Nagy-Szakal (2017)[44]	Giloteaux (2016)[35]	Giloteaux (2016)[46]	Fremont (2013)[34]	Sheedy (2009)[47]	Evangård (2007)[48]	Butt (2001)[49]	Butt (1998)[50]
Phylum *Firmicutes*				↓	↑					
Phylum *Proteobacteria*				↑	↓					
Famiy *Bacteroidaceae*	↑			↓	↑					
Family *Enterobacteriacaeae*				↑						↑
Family *Lachnospiraceae*	↓		↓							
Family *Prevotellaceae*				↑	↓					
Family *Rikenellaceae*				↓	↓					
Family *Ruminococcaceae*				↓	↓					
Genus *Anaerostipes*	↓		↑							
Genus *Bacteroides*	↑	↓								↓
Genus *Bifidobacterium*				↓	↓			↑	↓	↓
Genus *Clostridium*			↑	↓						
Genus *Coprobacillus*			↑	↑						
Genus *Faecalibacterium*			↓	↓	↓					
Genus *Haemophilus*			↓	↓						
Genus *Ruminococcus*	↓			↓		↓				
Species *Bacteroides uniformis*	↑	↓								
Species *Bacteroides ovatus*	↑		↑							
Species *Enterococcus faecalis*							↑			↑
Species *Escherichia coli*									↓	↓

The patient group appeared to show two main profiles at the family level; i) an increase in *Lachnospiraceae, Ruminococcaceae, Acualibacteraceae* and *Oscillospiraceae* with a decrease in *Bacteroidaceae* and *Rikenellaceae* compared to control groups (P2, P4, P5, P6, P10); ii) increase in *Bacteroidaceae* with an increase in *Streptococcaceae* (P1), *Rikenellaceae* (P14) or *Bifidobacterium* (P5). However, these findings are not consistent with previous studies. Multiple studies have observed a reduction in *Lachnospiraceae* and *Ruminococcaceae*[42,44]. Previous studies have reported a lower abundance of *Lachnospiraceae* in paediatric patients with ulcerative colitis and Crohn's disease[51,52]. *Ruminococcaceae* and *Lachnospiraceae* are prominent gut microbiota members and hydrolyse various sugars (such as starch) to produce SCFAs (e.g. butyrate). SCFAs play an important role in host epithelium maintenance and contribute to reduced levels of inflammatory markers[53,54]. However, the true effect on host health is disputed as several studies have reported conflicting correlations of *Lachonospiraceae* with disease status[55]. An increase in *Lachnospiraceae* has been associated with aging and could explain the increase observed in certain patients[56]. However, the age for these patients ranged from 18 to 63. Two studies (2016, 2021) also reported an increase in *Bacteroidaceae* in ME/CFS patients[42,46]. *Bacteroidaceae* are gut commensals and contribute to SCFA production[54]. Certain *Bacteroidaceae* (*Bacteroides* species) possess virulence factors (lipopolysaccharides) that can induce a high inflammatory response and potentially alter intestinal epithelium permeability[57-59]. At the genus level, the two main profiles previously mentioned were not as evident.

The patient group exhibited a high level of beta diversity, suggesting significant microbiota heterogeneity. However, it is unclear if this is an artefact of an underpowered study or a true disease trait. Due to the heterogenous nature of ME/CFS, further patient information should have been collected. A 2021 study analysed the oral and intestinal microbiota of 105 subjects (35 patients, 35 relative controls and 35 non-relative controls)[42]. Extensive patient information was collected including body mass index, diet, presence of gastrointestinal symptoms, presence of post-exertional malaise, IBS co-morbidity, Chalder Fatigue Scale and SF-36 Health Survey. Due to the lack of metadata collected and small sample size in the study presented here, it is extremely difficult to determine if the diversity seen within the patient group is due to confounding factors (e.g. diet, medication, age, sex, BMI, co-morbidities) or disease variability (e.g. presence of IBS, ME/CFS onset, etc). For example, *Lachnospiraceae* was increased in five patients and several studies have observed that *Lachnospiraceae* abundance is influenced by high non-starch polysaccharide diets[60,61]. However, diet diaries for study participants were not collected; therefore, it is unknown if diet contributed to profile seen in these patients. The varied presentation of ME/CFS is well known and includes symptoms, severity, disease onset, co-

morbidities, and family history[62-64]. The most commonly reported disease onset events, according to a 2019 study, are infection (e.g. viral or bacterial), stressful incident and/or environmental toxin exposure[65]. The study also reported that 97 % of patients also suffer from at least one co-morbidity, such as anxiety, depression, fibromyalgia, IBS or migraines. Approximately 13 % of the patients questioned confirmed at least one first-degree relative also suffered or had suffered from ME/CFS. Future studies should subset patients and collect extensive information to avoid bias introduced by confounding factors. However, recruitment of specific subgroups of patients may prove difficult.

It should be noted that species level was not investigated in this study due to the inaccurate species abundance estimate commonly encountered with Kraken2 and other tools (e.g. Centrifuge, MetaPhlAn2)[8,66,67]. Due to the extremely high inter-species variability within some genera, Kraken's classification algorithm correctly reports only the lowest common ancestor[8]. Therefore, for some species most reads might be classified at a higher level of the taxonomy and the number of true reads for a species are lower than what is classified. Bracken (Bayesian Re-estimation of Abundance after Classification with KraKEN) is able to accurately estimate species abundance in metagenomic samples by re-distributing reads in the taxonomic reads according to probability[68]. However, due to the time constraints it was not possible to perform Bracken analysis on the study dataset.

A total of 668 MAGs were generated from the study dataset, with ~ 5 % of high-quality. The creation of MAGs is becoming common practice within metagenome studies and increases the number of bacterial genomes within the reference databases[38,69]. Additionally, MAGs can be used to increase the accuracy of read classification by addition of high-quality MAGs to the original classification database or quantification of intrapopulation diversity within certain disease states[70]. It is estimated that 40-50 % of the human gut species lack a reference species; however, recent studies have expanded the known cultured and uncultured genomes of the human gut[71]. The use of MAGs in microbiome studies can provide useful information about uncultured diversity without the presence of isolate genomes. However due to incorrect contig binning, use of MAGs within a microbiome study requires careful consideration [72,73]. Most MAGs identified within this study belonged to *Firmicutes,* particularly *Lachnospiraceae* and *Oscillospiraceae*. This is unsurprising given the increase in these families within certain patients and further analysis could have been done to determine the intrapopulation diversity of *Lachnospiraceae* within the patient cohort if it had been large enough for meaningful analyses to be undertaken. Four novel MAGs

were generated during this study and the closest relatives were *Parabacteroides johnsonii*, *Ruminococcaceae*, *Bacteroides_A* and *Alistipes ihumii.*

This study aimed to investigate the intestinal microbiota of severe ME/CFS patients and household controls. While no specific taxa of interest were significantly different between groups, it has highlighted the potential heterogenous nature of the disease. Additionally, it has shown the need for future studies to subgroup patients and collect extensive metadata to potentially account for confounding factors. As with any microbiota study, it is imperative to recruit a sufficient number of participants to avoid an underpowered study (as seen in this study). This is discussed further in the General Discussion. Furthermore, generated MAGs should be used to improve read classification and study intrapopulation diversity of taxa of interest. Previous studies have suggested an altered microbiota in ME/CFS patients; however, further studies are needed to determine if the altered microbiota is due to the disease itself or simply a consequence of systemic disease.

2.5 References

1 Fukuda, K. *et al.* The chronic fatigue syndrome: a comprehensive approach to its definition and study. International ChronicFatigue Syndrome Study Group. *Ann Intern Med* **121**, 953-959, doi:10.7326/0003-4819-121-12-199412150-00009 (1994).

2 Sharpe, M. C. *et al.* A report--chronic fatigue syndrome: guidelines for research. *J R Soc Med* **84**, 118-121 (1991).

3 Carruthers, B. M. *et al.* Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. *Journal of Chronic Fatigue Syndrome* **11**, 7-115, doi:10.1300/J092v11n01_02 (2003).

4 Zigmond, A. S. & Snaith, R. P. The hospital anxiety and depression scale. *Acta Psychiatr Scand* **67**, 361-370,doi:10.1111/j.1600-0447.1983.tb09716.x (1983).

5 Chalder, T. *et al.* Development of a fatigue scale. *J Psychosom Res* **37**, 147-153, doi:10.1016/0022-3999(93)90081-p (1993).

6 Hoyles, L. *et al.* Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nat Med* **24**,1070-1080, doi:10.1038/s41591-018-0061-3 (2018).

7 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 (2013). <https://ui.adsabs.harvard.edu/abs/2013arXiv1303.3997L>.

8 Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 257,doi:10.1186/s13059-019-1891-0 (2019).

9 Méric, G., Wick, R. R., Watts, S. C., Holt, K. E. & Inouye, M. Correcting index databases improves metagenomic studies. *bioRxiv*, 712166, doi:10.1101/712166 (2019).

10 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).

11 Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research* **38**,e132-e132, doi:10.1093/nar/gkq275 (2010).

12 Besemer, J. & Borodovsky, M. Heuristic approach to deriving models for gene finding. *Nucleic Acids Research* **27**, 3911-3920, doi:10.1093/nar/27.19.3911 (1999).

13 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461,doi:10.1093/bioinformatics/btq461 (2010).

14 Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci* **28**, 1947-1951,doi:10.1002/pro.3715 (2019).

15 Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellularorganisms. *Nucleic Acids Res* **49**, D545-d551, doi:10.1093/nar/gkaa970 (2021).

16 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30,doi:10.1093/nar/28.1.27 (2000).

17 Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *MolBiol Evol* **34**, 2115-2122, doi:10.1093/molbev/msx148 (2017).

18 Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676, doi:10.1093/bioinformatics/btv033(2015).

19 Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction frommetagenome assemblies. *PeerJ* **7**, e7359-e7359, doi:10.7717/peerj.7359 (2019).

20 Stewart, R. D., Auffret, M. D., Snelling, T. J., Roehe, R. & Watson, M. MAGpy: a reproducible pipeline for the downstream analysis of metagenome-assembled genomes (MAGs). *Bioinformatics* **35**, 2150-2152, doi:10.1093/bioinformatics/bty905(2018).

21 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbialgenomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**, 1043-1055, doi:10.1101/gr.186072.114 (2015).

22 Pierce, N. T., Irber, L., Reiter, T., Brooks, P. & Brown, C. T. Large-scale sequence comparisons with sourmash. *F1000Res* **8**,1006, doi:10.12688/f1000research.19675.1 (2019).

23 Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembledgenome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**, 725-731, doi:10.1038/nbt.3893 (2017).

24 Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* **39**, 105-114, doi:10.1038/s41587-020-0603-3 (2021).

25 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069,doi:10.1093/bioinformatics/btu153 (2014).

26 Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research* **48**, D570-D578,doi:10.1093/nar/gkz1035 (2019).

27 Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic andtaxonomic placement of microbes. *Nature communications* **4**, 2304-2304, doi:10.1038/ncomms3304 (2013).

28 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryoticgenomes reveals clear species boundaries. *Nature Communications* **9**, 5114, doi:10.1038/s41467-018-07641-9 (2018).

29 Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal* **27**, 379-423 (1948).

30 Simpson, E. H. Measurement of Diversity. *Nature* **163**, 688-688, doi:10.1038/163688a0 (1949).

31 Anderson, M. J. in *Wiley StatsRef: Statistics Reference Online* 1-15.

32 Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* **27**,325-349, doi:https://doi.org/10.2307/1942268 (1957).

33 Clarke, K. R. Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology* **18**, 117-143 (1993).

34 Fremont, M., Coomans, D., Massart, S. & De Meirleir, K. High-throughput 16S rRNA gene sequencing reveals alterations ofintestinal microbiota in myalgic encephalomyelitis/chronic fatigue syndrome patients. *Anaerobe* **22**, 50-56, doi:10.1016/j.anaerobe.2013.06.002 (2013).

35 Giloteaux, L. *et al.* Reduced diversity and altered composition of the gut microbiome in individuals with myalgicencephalomyelitis/chronic fatigue syndrome. *Microbiome* **4**, 30, doi:10.1186/s40168-016-0171-4 (2016).

36 Nagendra, H. Opposite trends in response for the Shannon and Simpson indices of landscape diversity. *Applied Geography*

22, 175-186, doi:https://doi.org/10.1016/S0143-6228(02)00002-4 (2002).

37 Wilkins, L. G. E., Ettinger, C. L., Jospin, G. & Eisen, J. A. Metagenome-assembled genomes provide new insight into the microbial diversity of two thermal pools in Kamchatka, Russia. *Scientific Reports* **9**, 3059, doi:10.1038/s41598-019-39576-6 (2019).

38 Alneberg, J. *et al.* Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplifiedgenomes. *Microbiome* **6**, 173, doi:10.1186/s40168-018-0550-0 (2018).

39 Armstrong, C. W., McGregor, N. R., Lewis, D. P., Butt, H. L. & Gooley, P. R. The association of fecal microbiota and fecal,blood serum and urine metabolites in myalgic encephalomyelitis/chronic fatigue syndrome. *Metabolomics* **13**, 8, doi:10.1007/s11306-016-1145-z (2016).

40 Molinaro, A. *et al.* Imidazole propionate is increased in diabetes and associated with dietary patterns and altered microbialecology. *Nat Commun* **11**, 5881, doi:10.1038/s41467-020-19589-w (2020).

41 Chen, F. *et al.* Integrated analysis of the faecal metagenome and serum metabolome reveals the role of gut microbiome- associated metabolites in the detection of colorectal cancer and adenoma. *Gut*, gutjnl-2020-323476, doi:10.1136/gutjnl-2020-323476 (2021).

42 Lupo, G. F. D. *et al.* Potential role of microbiome in Chronic Fatigue Syndrome/Myalgic Encephalomyelits (CFS/ME). *Sci Rep* **11**, 7043, doi:10.1038/s41598-021-86425-6 (2021).

43 Guo, C. *et al.* Deficient butyrate-producing capacity in the gut microbiome of Myalgic Encephalomyelitis/Chronic FatigueSyndrome patients is associated with fatigue symptoms. *medRxiv*, 2021.2010.2027.21265575, doi:10.1101/2021.10.27.21265575 (2021).

44 Nagy-Szakal, D. *et al.* Fecal metagenomic profiles in subgroups of patients with myalgic encephalomyelitis/chronic fatiguesyndrome. *Microbiome* **5**, 44, doi:10.1186/s40168-017-0261-y (2017).

45 Newberry, F., Hsieh, S. Y., Wileman, T. & Carding, S. R. Does the microbiome and virome contribute to myalgicencephalomyelitis/chronic fatigue syndrome? *Clin Sci (Lond)* **132**, 523-542, doi:10.1042/CS20171330 (2018).

46 Giloteaux, L., Hanson, M. R. & Keller, B. A. A Pair of Identical Twins Discordant for Myalgic Encephalomyelitis/ChronicFatigue Syndrome Differ in Physiological Parameters and Gut Microbiome Composition. *Am J Case Rep* **17**, 720-729, doi:10.12659/ajcr.900314 (2016).

47 Sheedy, J. R. *et al.* Increased d-lactic Acid intestinal bacteria in patients with chronic fatigue syndrome. *In Vivo* **23**, 621-628(2009).

48 Evengård, B., Nord, C. & Sullivan, Å. P1239 Patients with chronic fatigue syndrome have higher numbers of anaerobicbacteria in the intestine compared to healthy subjects. *International Journal of Antimicrobial Agents*, S340 (2007).

49 Butt, H., Dunstan, R., McGregor, N. & Roberts, T. in *Proceedings of the AHMF international clinical and scientific conference.* 1-2.

50 Butt, H. *et al.* in *Proceedings of the AHMF International Clinical and Scientific Conference.* 12-14.

51 Schirmer, M. *et al.* Compositional and Temporal Changes in the Gut Microbiome of Pediatric Ulcerative Colitis Patients AreLinked to Disease Course. *Cell Host Microbe* **24**, 600-610.e604, doi:https://doi.org/10.1016/j.chom.2018.09.009 (2018).

52 Maukonen, J. *et al.* Altered Fecal Microbiota in Paediatric Inflammatory Bowel Disease. *J Crohns Colitis* **9**, 1088-1095,doi:10.1093/ecco-jcc/jjv147 (2015).

53 Silva, Y. P., Bernardi, A. & Frozza, R. L. The Role of Short-Chain Fatty Acids From Gut Microbiota in Gut-BrainCommunication. *Front Endocrinol (Lausanne)* **11**, 25, doi:10.3389/fendo.2020.00025 (2020).

54 Parada Venegas, D. *et al.* Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevancefor Inflammatory Bowel Diseases. *Front Immunol* **10**, 277, doi:10.3389/fimmu.2019.00277 (2019).

55 Vacca, M. *et al.* The Controversial Role of Human Gut Lachnospiraceae. *Microorganisms* **8**, 573,doi:10.3390/microorganisms8040573 (2020).

56 Odamaki, T. *et al.* Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol* **16**, 90-90, doi:10.1186/s12866-016-0708-5 (2016).

57 Moncrief, J. S. *et al.* The enterotoxin of Bacteroides fragilis is a metalloprotease. *Infection and immunity* **63**, 175-181 (1995).

58 Franco, A. A. *et al.* Cloning and characterization of the Bacteroides fragilis metalloprotease toxin gene. *Infection andimmunity* **65**, 1007-1013 (1997).

59 Wu, S., Lim, K.-C., Huang, J., Saidi, R. F. & Sears, C. L. Bacteroides fragilis enterotoxin cleaves the zonula adherens protein, E-cadherin. *Proceedings of the National Academy of Sciences* **95**, 14979-14984 (1998).

60 Salonen, A. *et al.* Impact of diet and individual variation on intestinal microbiota composition and fermentation products inobese men. *Isme j* **8**, 2218-2230, doi:10.1038/ismej.2014.63 (2014).

61 Martínez, I. *et al.* Gut microbiome composition is linked to whole grain-induced immunological improvements. *Isme j* **7**,269-280, doi:10.1038/ismej.2012.104 (2013).

62 Brurberg, K. G., Fonhus, M. S., Larun, L., Flottorp, S. & Malterud, K. Case definitions for chronic fatigue syndrome/myalgicencephalomyelitis (CFS/ME): a systematic review. *BMJ Open* **4**, e003973, doi:10.1136/bmjopen-2013-003973 (2014).

63 Lim, E. J. *et al.* Systematic review and meta-analysis of the prevalence of chronic fatigue syndrome/myalgicencephalomyelitis (CFS/ME). *J Transl Med* **18**, 100, doi:10.1186/s12967-020-02269-0 (2020).

64 Lim, E. J. & Son, C. G. Review of case definitions for myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). *J TranslMed* **18**, 289, doi:10.1186/s12967-020-02455-0 (2020).

65 Chu, L., Valencia, I. J., Garvert, D. W. & Montoya, J. G. Onset Patterns and Course of Myalgic Encephalomyelitis/ChronicFatigue Syndrome. *Front Pediatr* **7**, 12, doi:10.3389/fped.2019.00012 (2019).

66 Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**, 902-903,doi:10.1038/nmeth.3589 (2015).

67 Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* **26**, 1721-1729, doi:10.1101/gr.210641.116 (2016).

68 Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerjComput Sci*, doi:ARTN e104 10.7717/peerj-cs.104 (2017).

69 Kim, C. Y. *et al.* Human reference gut microbiome catalog including newly assembled genomes from under-representedAsian metagenomes. *Genome Medicine* **13**, 134, doi:10.1186/s13073-021-00950-7 (2021).

70 Teh, J. J. *et al.* Novel strain-level resolution of Crohn's disease mucosa-associated microbiota via an ex vivo combination ofmicrobe culture and metagenomic sequencing. *The ISME Journal* **15**, 3326-3338, doi:10.1038/s41396-021-00991-1 (2021).

71 Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the globalhuman gut microbiome. *Nature* **568**, 505-510, doi:10.1038/s41586-019-1058-x (2019).

72 Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insightsfrom Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl Environ Microbiol* **87**, doi:10.1128/aem.02593-20 (2021).

73 Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes frommetagenomes. *Genome research* **30**, 315-333, doi:10.1101/gr.258640.119 (2020).

74 Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A. & Hugenholtz, P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology* **36,** 996-1004, doi:10.1038/nbt.4229 (2018).

Chapter 3 : Genome characterisation of *Bacteroides fragilis* bacteriophage vB_BfrS_23 and discovery of a novel *B. fragilis* phage family

Part of this work has been published. Refer to Appendix 2.

3.1 Introduction

3.1.1 Unknown phage diversity

According to the National Centre for Biotechnology Information (NCBI) Virus database, as of December 2020, there were 19,663 complete bacteriophage genomes separated into 12 families (according to the International Committee on Taxonomy of Viruses (ICTV); September 2019)[1]. However, this represents a minute fraction of the potential phage diversity on the planet[2]. *Caudovirales* phage are the most abundant in public databases with the majority of these phage belonging to families *Siphoviridae, Myoviridae* and *Podoviridae*[2,3]. Additionally, there is an over-representation of phage from certain genera (*Mycobacterium, Streptococcus, Escherichia, Pseudomonas, Gordonia, Lactococcus* and *Salmonella*) due to the medical relevance of the host bacteria[2]. Phage exhibit a variety of morphological traits (e.g. tailed, non-tailed), genetic material (dsDNA, ssDNA, dsRNA or ssRNA), genome size (2,435 bp to 735 kbp), host range and environment (e.g. human gut, soil, ocean)[2,4-8]. Additionally, little to no sequence similarity is typically seen between phage infecting different hosts, with phage infecting the same host displaying significant sequence differences[4,9,10]. The majority (97 %) of nucleotide pairwise comparisons of 2,333 phages reported no detectable homology[11]. In recent years, significant advances in sequencing technology and bioinformatic tools have increased the understanding and importance of phage. Additionally, phage discovery and classification has increased exponentially.

3.1.2 Phage discovery techniques

The huge diversity of phage physical and genomic traits makes these entities difficult to study and characterise. A culture-based method was traditionally used for virus discovery and involves co-culture of potential hosts with the source sample (e.g. sewage water)[12]. While this method allows for the physical isolation of the phage, it has several caveats. It is low-throughput, time-consuming and restricted to culturable bacterial hosts and lytic phage[13]. A portion of the human gut microbiota remains uncultured; however, this is changing with the introduction of bacterial/archaeal culturomics[14,15]. Despite these challenges, the isolation and characterisation of

novel phage has increased and begun to address the low sequence diversity and host taxonomy seen within public databases[16-18].

The number and diversity of phage within public databases have been greatly increased by viral metagenomics[18]. A study using a collection of viral protein families expanded the number of viral genes 16-times and discovered > 125,000 viral genomes from 3,042 global metagenomes. This study highlighted the vast undiscovered phage diversity[16]. This is particularly noted in viral analysis of environmental and human microbiota metagenome studies. The majority of viral metagenomic sequences, sometimes up to 90 % of reads, remain unknown; which is a major obstacle to obtaining an accurate picture of microbiota diversity[6,16,19,20]. Additionally, metagenome studies rarely have the high resolution needed to correctly reconstruct closely related viral genomes, resulting in viral-population microdiversity being ignored[21-23]. It can also be difficult to predict the host range of metagenome-assembled phage (discussed below)[24]. Therefore, it is necessary to integrate culture-based phage isolation with metagenome phage discovery to uncover the true level of phage diversity.

3.1.3 Phage phylogenetics

Following isolation of a novel phage, it is necessary to determine its relatedness and evolutionary history to currently known phage through phylogenetic analysis. However, viral phylogeny is challenged by the lack of universal genetic markers, lateral gene transfer and rapid mutation rates[25-28]. New methods for viral phylogeny are being developed as new phage are discovered. Traditionally, classification was based on phage morphology[29]. Genetically diverse phage can share physical characteristics such as the major capsid protein conserved between all tailed phage. However, a high level of structural conservation is rarely observed at amino acid and nucleotide sequence levels[30-32]. A common technique used to determine genomic similarities is pairwise comparison of phage genomes and is primarily used for classification by ICTV primary classification. The overall nucleic acid sequence identity thresholds demarcate phage into species (95 %) and genera (~ 70 %)[33]. However, as mentioned previously, phage exhibit large diversity in nucleotide sequence. Therefore, protein-/orthologue-based techniques are also used to determined relatedness of phage[26,27,34,35].

While traditional phylogenetic trees can be beneficial for phage with orthologous proteins, they cannot accommodate the fluidity of the phage genome[27]. The use of networks has recently been implemented in visualising phage phylogeny and allows for the connection of nodes (phage genomes) via edges (gene, genome or protein similarity)[26]. A 2008 study produced a network with

306 temperate and virulent phage genomes. The network allowed the authors to visualise the high similarity of temperate phage depicted as a tightly formed cluster and relatedness of the virulent phage dispersed on the periphery in distinct clusters[26]. Viral network-based phylogenetic analysis was further advanced by the creation of vConTACT[36-38]. Viral predicted proteins are extracted, used to create viral protein clusters and pairwise genome similarities generated. Intergenomic similarity thresholds are used to determine which viruses are linked by an edge. The authors demonstrated that viruses can be accurately grouped at genus level and this allowed for discovery of novel phage families.

3.1.4 Virus-host prediction

An important advantage to physical phage isolation is confirmation of the phage's host. This is not easily determined for metagenome-assembled phage. Phage and host genomes give insights into virus-host interactions and can be used to predict virus-host relationships[24,39]. Host information for all viral sequences from NCBI Viral RefSeq (release 99) are present in the Virus-Host Database[40]. Several computational approaches are used for virus host prediction such as sequence homology, abundance profiles and k-mer frequency[24,41-43].

Both virus and host genomes are used to search for sequence homology. Virus genomes are queried for bacterial auxiliary metabolic genes or tRNAs. A common sequence homology search site is the Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) spacers present in most prokaryotic genomes[44-46]. CRISPRs are used by prokaryotes to evade viral invasion by integrating short segments of virus DNA (25-50 bp) into the prokaryotic genome ("spacers")[44]. Virus hosts can be predicted by aligning CRISPR spacer sequences to viral reads in the same metagenome[47,48]. However, multiple host prediction techniques should be used as some prokaryotes do not possess complete CRISPR-Cas defence systems[44,49]. The number of prokaryotes without CRISPR-Cas is disputed[50]. For example, *Staphylococcus aureus* was originally believed to not contain a CRISPR-Cas system. However, a 2018 study identified 57 CRISPR loci in 38 *S. aureus* strains but only 4 *cas* genes were located near the CRISPR loci. It should be noted that the *cas* gene is not required for identification of spacers[51]. Abundance profiles of virus and host sequences can also be used for host prediction and produce the most accurate results if used across multiples samples (e.g. longitudinal studies)[52-54]. This is based on the idea that viruses generally mimic host abundance patterns; such as infection type (lytic vs lysogenic), number of prophage in host and predator-prey dynamics[43]. However, in general this method produces relatively few correct host predictions due to host variation and temperate phage[55-57]. K-mer frequency profile-based host prediction (e.g. VirFinder) is typically less accurate than spacer

sequence homology due to low specificity but produces more potential virus-host pairs[41,58]. A similar k-mer frequency profile is commonly shared between phage and the host[59]. These distances between tetranucleotide (4-mer) frequency profile of viruses and hosts are predicted and most likely host shown by closest Euclidean distance[57,59-62]. This method is often best at predicting viral hosts above genus level as there may not be enough differentiation at species level[24]. Accurate host prediction is imperative for metagenome-derived phage and, if correct, greatly assists in successful isolation of a closely-related phage (e.g. crAssphage).

3.1.5 Discovery of crAssphage

The most successful case of integrating culture-based and metagenomic methods for phage discovery is the discovery of the most abundant human gut-associated viruses, crAssphage[63]. crAssphage have been reported in multiple metagenomes from a variety of geographical locations and are believed to be a core component of the healthy human gut microbiota[53,64-67]. In some individuals, crAssphage account for < 22 % of reads in whole-community metagenomes and < 90 % of viral reads in the virus-enriched portion[63,64,68]. Interestingly, Old World monkeys, New World monkeys and great apes were found to harbour divergent crAssphage, hinting at sustained co-evolution of these viruses with primates[69].

Despite their abundance and global distribution, crAssphage have only recently been discovered, mainly due to their dissimilarities to known viral genomes. Due to these dissimilarities, very little was known about their evolutionary relationships, predicted gene functions and comparison to known phage[63,70]. crAssphage were declared members of a novel viral clade and the putative crAss-like family divided into four subfamilies with 10 candidate genera[71]. A recent proposal to ICTV has further characterised crAss-like phage to a new order (*Crassvirales*) comprising six families, 10 subfamilies, 78 genera and 279 species[72]. Several approaches were used to predict the bacterial host as a member of phylum *Bacteroidetes* including read co-occurrence and presence of *Bacteroides*-related carbohydrate-binding BACON domains[58,63,73]. Further evidence was discovered when two *Bacteroides* species CRISPR spacers partially matched two crAss-like phage genomes[70,74]. These host predictions were confirmed following successful isolation of ϕcrAss001 with *Bacteroides intestinalis*[68]. However, crAss-like phage most likely infect other members of *Bacteroidetes* due to the crAss-like phage genome diversity observed. A 2020 study isolated two additional crAss-like phage (DAC15 and DAC17) from wastewater effluent using *Bacteroides thetaiotaomicron*[75]. Structural module genes (major capsid protein (MCP), portal protein, large terminase subunit, tail proteins) and several proteins without known function are conserved throughout crAss-like phage[63]. The MCP, portal protein and large terminase subunit of

crAss-like phage were used for phylogenetic analysis with known phage and revealed a relationship with three phage: *Azobacteroides* phage ProJPt-Bp from termite gut, *Flavobacterium psychrophilum* phage Fpv3 isolated from fish and *Cellulophaga* phage phi14:2 from the ocean[76-78]. Interestingly, these phage have no known association with the human gut microbiota. These findings highlight the vast phage diversity and that major new groups of phage remain to be discovered; particularly gut-resident phage.

3.1.6 *Bacteroides* phage

Despite the importance of *Bacteroides* within the human gut microbiota, only 38 *Bacteroides* phage are present on NCBI Virus (four partial, 34 complete) and isolated from different geographical locations and sample sources (sewage and faeces). The genome size ranges from 335 bp (partial *Bacteroides* phage ATCC 700786-B1) to 179,283 bp (*Bacteroides* phage DAC22). The majority of phage were isolated using *Bacteroides thetaiotaomicron* VPI-5482 (27 phage), with other hosts including *Bacteroides uniformis*, *B. intestinalis* and *Bacteroides fragilis*[1]. All but two phage were isolated within the past two years, highlighting the recent increased rate of phage discovery and characterisation. Phage-host relationships in most commensal gut-associated bacteria remain mainly unexplored, particularly among *Bacteroides* species.

A recent study isolated 27 *B. thetaiotaomicron*-infecting phage from two continents and, through network-based phylogeny, discovered the phage split into three distinct clusters. One cluster shared extensive phams (shared gene phamily membership) and genome organisation with ϕcrAss001, reinforcing the previous crAss-like phage host predictions. Low protein homology existed between other isolated *B. thetaiotaomicron* phage and previously isolated *Bacteroides* phage[75]. Additionally, capsular polysaccharide (CPS) mediated *Bacteroides* phage interactions were studied using several USA-isolated phage from the previous study. *B. thetaiotaomicron* CPSs were involved in phage host tropism and CPS variants that allowed escape from phage predation were actively selected for[79]. CPSs play an important role in host immune evasion and modulation; however, it is possible CPSs have multiple roles due to the diversity of CPS synthesis loci in gut bacteria[80-83]. Interestingly, *B. thetaiotaomicron* without CPSs were able to escape phage predation by modifying eight phase-variable lipoproteins[79]. A recent study explored phage-host dynamics of ϕcrAss001 and *B. intestinalis* and ϕcrAss001 persistence within a monoxenic mouse model. The authors reported acquisition of phage resistance depending on host CPS phase variation. Continuous phage invasion resulted in one of two locus changes with opposite effects; switching off PVR9 CPS locus correlated with phage adsorption or phage protection by increased expression of alternative CPSs (PVR7, PVR8, PVR11 and PVR12). The authors proposed the long term

persistence of ϕcrAss001 is partially due to host CPS switching; allowing for an equilibrium between phage-sensitive and -resistant host cells[84]. These results reveal the complexity of phage-host relationships and highlight the need for similar studies to truly understand their ecological roles within the gut microbiota.

Bacteroides phage are currently used for surveillance of faecal contamination in treated and untreated water systems (microbial source tracking) due to the specificity of *Bacteroides* to the human gut[85-87]. Microbial source tracking with phage is a relatively cheap and easy technique to accurately detect faecal pollution in environmental waters. *B. fragilis*-infecting phage, particularly using strain GB-124 as host, have been described as potential markers of human faecal contamination in water sources[88,89]. *B. fragilis* GB-124-infecting phage are ideal for tracking human faecal pollution due to their morphology, environmental persistence and resistance to treatment processes. They have geographical stability and have been used for microbial source tracking in municipal wastewaters worldwide[85,90].

3.1.7 Aims and objectives

The recent increase in metagenome and virome studies highlights the unexplored potential phage diversity within all biomes[16,91]. It displays the necessity of combining metagenome-based phage discovery and phage isolation to fully characterise and understand the fundamental roles phage play in their environment and interactions with the host. The discovery of crAss-like phage and isolation of ϕcrAss001 highlight the success of metagenome-based phage discovery[63,68]. This Chapter reports the isolation and characterisation of a novel *B. fragilis* phage from sewage using *B. fragilis* GB-124. The relatedness of this phage and three published *B. fragilis* phage were explored within the context of currently known phage and metagenome-assembled *Bacteroides* phage. The exploration of a large *Bacteroides*-infecting phage dataset revealed the presence of a novel *B. fragilis* phage family consisting of five genera and 37 species, with little protein or gene sequence identity to currently known phage. A genus also displayed specific geographical occurrence within metagenomes.

3.2 Methods

3.2.1 Growth media constituents and buffers

3.2.1.1 Sterilisation

All glassware and reagents were sterilised in an autoclave for 20 min at 121 ˚C and 15 psi pressure.

3.2.1.2 Media and buffers

Full details of media and buffers used in this work are provided in Table 3.1.

3.2.1.3 B. fragilis growth conditions

B. fragilis (Bf) strains were grown anaerobically (5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure; MACS MG 1000 Anaerobic Workstation) in liquid BPRM or BHI, BPRM agar (15 %) or BPRM semi-soft overlay (3.5 %) (Table 3.1). Kanamycin was added to liquid medium and semi-soft agar when stated. Liquid medium was placed into the anaerobic cabinet at least 24 h prior to inoculation to allow removal of oxygen.

3.2.1.4 Storage of strains

All strains were stored in liquid medium and 40 % glycerol at -80 ˚C and 100 µl of freezer stock used for inoculation.

3.2.1.5 Strains

B. fragilis strains were obtained from Dr Regis Stentz, Quadram Institute Bioscience (QIB), and Dr James Ebdon, University of Brighton (Table 3.2).

3.2.2 Bacteriophage ϕB124-14 propagation and enumeration

B. fragilis strain GB-124 and its phage ϕB124-14 were supplied by Dr James Ebdon (University of Brighton). The phage was used as a positive control for environmental and water screening assays. Prior to screening assays, it was necessary to increase the phage stock volume and determine the phage titre.

Table 3-1: Media and solution recipes

Reagents highlighted grey were added following autoclaving.

Medium/buffer	Constituent	Weight/volume
Bacteroides phage recovery media (BPRM) broth, pH 7, stored at 4 ˚C in the dark	Peptone	10 g
	Tryptone	10 g
	Yeast Extract	2 g
	NaCl	5 g
	L-cysteine	0.5 mg
	Glucose	1.8 g
	$MgSO_4 \bullet 7H_2O$	0.12 g
	$CaCl_2$ (0.45 M)	1 mL
	MilliQ H_2O	965 mL
	Na_2CO_3 (1 mol/L) filter sterilised	25 mL
	Hemin (0.1 % wt/vol) filter sterilised	10 mL
BPRM agar, pH 7, stored at 4 ˚C in the dark	Bacteriological agar (1.5 % wt/v)	15 g
BPRM semi-soft agar, pH 7, stored at 4 ˚C in the dark	Bacteriological agar (0.35 % wt/v)	3.5 g
$CaCl_2$ (0.45 M), stored at room temperature	$CaCl_2 \bullet H_2O$	5 g
	Sterile MilliQ H_2O	95 mL
Na_2CO_3 (1 mol/L), filter-sterilised (0.22 μm) and autoclaved, stored at 4 ˚C	Na_sCO_3	10.6 g
	Sterile MilliQ H_2O	89.4 mL
Hemin (0.1 % wt/vol), filter-sterilised (0.22 μm) and autoclaved, stored at 4 ˚C	Hemin	0.1 g
	NaOH solution (1 mol/L)	0.5 mL
	MilliQ H_2O	99.4 mL
Brain Heart Infusion (BHI) Broth, stored at 4 ˚C (Oxoid, CM1135)	BHI powder	37 g
	MilliQ H_2O	Up to 1000 mL
Fastidious Anaerobe Broth (FAB; Neogen LabM, LAB071)	FAB powder	29.7g
	MilliQ H_2O	Up to 1000 mL
SM buffer, autoclaved and stored at room temperature	NaCl	5.8 g (final concentration: 100 mM)
	$MgSO_4 \bullet 7H_2O$	2 g (final concentration 8 mM)
	Tris-Cl (1 M, pH 7.5)	50 ml (final concentration 50 mM)
	MilliQ H_2O	Up to 1000 mL

Table 3-2: *Bacteroides* and related species and strains used for sample screening

Species	Strain
Bacteroides thetaiotaomicron	VPI 5482^T
Bacteroides ovatus	V975
Bacteroides stercoris	DSM 19555^T
*Phocaeicola dorei**	DSM 17855^T
Bacteroides xylanisolvens	XBIA DSM 18836^T
Bacteroides fragilis	NCTC 9343^T
	GB-124

*Previously *Bacteroides dorei*.

3.2.2.1 Phage propagation

The host strain was inoculated into BPRM broth and incubated anaerobically (5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure) for 12-16 h. This was sub-cultured in BPRM broth to exponential phase (OD_{620} 0.3-0.33). A soft-agar overlay phage assay was used to determine the phage titre. BPRM semi-soft agar (5 mL aliquots) were melted in a 95˚C water bath and stored at 55 ˚C until needed. Serial dilutions (10^{-1} to 10^{-9}) of ϕB124-14 freezer stock were mixed with GB-124 in BPRM semi-soft agar at a ratio of 1:2 (100 µl:200 µl). The molten overlay was poured onto a room temperature BPRM agar plate and allowed to cool before anaerobic incubation for 16-24 h.

The plate with the highest plaque count was selected for phage harvesting. Plates were checked for consistent ϕB124-14 morphology prior to harvesting. Approximately 5-8 mL of phage disruption buffer was added to the plate, which was then gently shaken on a mini gyratory shaker SSM3 (Stuart UK) for 1 h. The liquid and semi-soft agar were collected into a 50 mL falcon tube and centrifuged at 3,000 ***g*** for 5 min. The supernatant was filtered through a 0.45 µm PES membrane syringe filter (Sartorius UK Ltd) and stored at 4 ˚C until needed.

3.2.2.2 Phage enumeration

To determine the phage stock titre, a plaque assay with ϕB124-14 and GB-124 was performed (as above). Following incubation, the dilution with the clearest plaques were chosen to count. The

dilution above and below this plate was also counted. The phage stock titre was determined by accounting for dilution and volume used in assay.

3.2.3 *Bacteroides* strain growth dynamics

A freezer stock of each *Bacteroides* strain (except GB-124) was inoculated into BHI or BPRM broth and incubated anaerobically (5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure) for 12-16 h. The strains were sub-cultured into BHI or BPRM broth and the OD_{620} measured using a spectrophotometer every hour until the OD_{620} was > 0.1.

3.2.3.1 Growth media

The OD_{620} was normalised to 0.1 (final volume: 200µl in starting liquid media) and aliquoted into a flat bottom 96-well EIA/RIA Assay Microplate (Corning®). BHI and BPRM broths were used as negative controls. The OD_{595} was measured every 15 min over a 24-h period anaerobically (Tecan Infinite F50 Absorbance Microplate Reader; 5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure). Data were exported into Excel format from Magellan Data Analysis Software. Three biological and technical replicates were obtained. The averaged OD values were used to plot a growth curve for each strain in BHI and BPRM media.

3.2.3.2 Minimum Inhibitory Concentration (MIC)

The OD_{620} of each bacterial strain culture was normalised to 0.1 and differing concentrations of kanamycin (1000 µg/mL, 200 µg/mL, 100 µg/mL, 50 µg/mL, 10 µg/mL, 1 µg/mL and 0.1 µg/mL) added prior to adjusting to a final volume of 200 µl and aliquoting to a flat bottom 96-well EIA/RIA Assay Microplate (Corning®). A positive and negative control were used: bacteria without antibiotics (positive control) and BPRM broth (negative control). Additionally, an *Escherichia coli* strain (DH5α) was used as a positive control at kanamycin concentrations 50 µg/mL and 100 µg/mL. The OD_{595} was measured every 15 min over a 24-h period anaerobically (Tecan Infinite F50 Absorbance Microplate Reader; 5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure). Data were exported into Excel format from Magellan Data Analysis Software. Three biological and technical replicates were obtained and MICs determined.

3.2.4 Environmental sample collection

3.2.4.1 Freshwater and sewage water collection and concentration

A total of eight freshwater samples (50 mL each) were collected from ponds in and around Titchwell Marsh Norfolk (52.962569 N˚, 0.608813 E˚), UK. Raw (untreated) municipal wastewater

(100 mL) was collected from a UK-based sewage treatment plant. The freshwater samples were centrifuged at 5,000 ***g*** for 5 min to pellet large debris. All samples were filtered through a 0.45 µm PES membrane syringe filter (Sartorius UK Ltd) and concentrated by centrifugation using Amicon Ultra-15 10K centrifugal units (15 min at 5,000 ***g***). Filtrate was stored at 4 ˚C until used for phage screening.

3.2.4.2 Animal faeces collection and concentration

A total of six samples of animal faeces (30-40 g each) were collected from four different locations (Table 3.3). The faecal samples were homogenised and approximately 3 g of each faecal sample was diluted 1:10 in sterile Milli-Q H_2O (Milli-Q® Reference Water Purification System). Following a brief vortex, the samples were left on ice for 2 h to allow diffusion of viral particles from solid material.

The samples were centrifuged twice at 11,200 ***g*** for 30 min, with the supernatant retained following each centrifugation step. The samples were filtered and concentrated according to the previous section. The concentrated faecal water was stored at 4 ˚C until needed for phage screening assays.

Table 3-3: Sample type and collection site

Sample number	Sample type	Location
1-3	Fresh horse faeces	Norwich (52.627739 N˚, 1.218993 E˚)
4	Horse manure	Norfolk (52.505429 N˚, 1.101968 E˚)
5	Horse manure	Norfolk (52.503755 N˚, 1.087619 E˚)
6	Pig faeces	Norfolk (Private residence)

3.2.5 Environmental phage screening

A 16-18 h culture of each *Bacteroides* strain was sub-cultured anaerobically in BPRM broth until mid-exponential phase was reached (OD_{620} 0.3-0.33). Kanamycin (final conc. 100 µg/mL) was added to BPRM semi-soft agar and solid agar during preparation to reduce potential contamination introduced from the environmental samples. An aliquot (1 mL) of each

environmental sample was mixed with 1 mL of each sub-cultured *Bacteroides* strain, allowed 5 min for adsorption, mixed in molten BPRM semi-soft agar (final conc. 0.35 %) and poured onto BPRM agar plates. After 16-24 h anaerobic incubation (5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure), the plates were screened for plaques. Plaques with a distinct morphology were picked with a sterile pipette tip and stored in 10 mL BPRM medium with sub-cultured host (OD_{620} 0.3-0.4), incubated for 18 h to allow further propagation of the phages and filtered through 0.22 µm PES membrane filter (Sartorius UK Ltd). The procedure was performed three times to ensure a pure phage stock.

3.2.5.1 Phage purification

It should be noted that only one phage completed all three purification steps mentioned above and was named vB_BfrS_23. Further propagation was necessary to increase the phage stock titre. Several dilutions of 50 µL phage with 200 µL mid-exponential phase bacterial host (OD_{620} 0.3-0.4) were mixed in semi-soft BPRM agar (0.35 %) and poured onto BPRM agar plates. Following 16 h of anaerobic incubation (5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure), 5 mL SM buffer was added and gently shook on a mini gyratory shaker SSM3 (Stuart UK) for 1 h. The top agar and buffer were harvested, centrifuged at 3,000 ***g*** for 10 min and supernatant filtered through a PES membrane bottle top vacuum filter using ~ 100 psi pressure (Millipore Millivac, Merck UK). The phage stock titre was determined using dilutions 10^{-1} to 10^{-9} was stored at 4 ˚C until needed.

3.2.6 Phage DNA extraction

DNA was extracted from a phage stock (>10^9 PFU/mL) for Illumina and Oxford Nanopore MinION sequencing. These required different DNA extraction techniques due to the differences in sequencing platforms. These extractions were performed by Dr Mohammad Tariq, QIB. The quality of DNA was assessed by a Nanodrop™ Spectrophotometer and Qubit™ dsDNA HS Assay Kit (Invitrogen™).

3.2.6.1 For Illumina sequencing

The phage stock was incubated with RNase A (100 U Ambion™) and Turbo DNase (2 U Invitrogen™) for 30 min at 37 ˚C to remove bacterial chromosomal RNA and DNA, respectively. The sample was heat treated at 65 ˚C for 10 min with 15 mM EDTA to inactivate nucleases. The Norgen Phage DNA isolation kit (Geneflow Limited, Lichfield, UK) was used to extract phage DNA and resulting DNA stored at 4 ˚C.

3.2.6.2 For MinION sequencing

The phage stock was PEG-precipitated (10 % (w/v) PEG 8000 and 6 % (w/v) NaCl) and treated with DNase (4 U Turbo DNase Invitrogen™) and RNase A (100 U Ambion™). SDS (0.5 % w/v) and 4 µL proteinase K (Ambion™ 80 µg, 20 mg/mL) were added and the sample heated at 55 ˚C for 1 h, followed by heat inactivation at 75 ˚C for 15 min. Lipids and proteins were removed by mixing the sample 1:1 with chloroform and vigorously shaken for a few seconds. It was centrifuged at 15,000 g for 5 min at 20 ˚C. The upper aqueous phase was carefully removed and treated with NaCl (final conc. 0.2 M) prior to mixing 1:1 with isopropanol and left at -20 ˚C for 16 h. The sample was centrifuged at 13,000 ***g*** at 20 ˚C for 1 h followed by two washes with fresh 70 % EtOH wash. The pellet was resuspended in nuclease-free water (Invitrogen™) and stored at 4 ˚C.

3.2.7 Phage sequencing

The phage DNA was sequenced using Illumina and MinION Oxford Nanopore Technologies platforms.

3.2.7.1 For Illumina sequencing

The DNA was sequenced by David Baker at QIB Sequencing Service using the Illumina MiSeq system. The sequencing library was prepared with Illumina Nextera XT (Illumina, Saffron Walden, UK) library preparation kit, sequenced on Illumina MiSeq 2 x 150-cycle v2 chemistry and paired-end reads provided as FASTQ files. The adapters of the raw reads were removed using Trimmomatic (v.0.39) before quality control trimming with Sickle (v.1.33) at --q 30 and --l 15[92,93].

3.2.7.2 For MinION sequencing

The manufacturer's protocol was followed and native barcoding kit EXP-NDB104 with the ligation sequencing kit SQK-LSK109 were used. MinION sequencing was performed with Dr Mohammad Tariq. Briefly, the NEBNext FFPE Repair Mix (M6630) and NEBNext End Repair/dA-tailing (E7546) were mixed with 1 µg of high-quality phage DNA for end-repair and dA-tailing. The native barcode kit (EXP-NBD104) was used to barcode and ligated using NEB Blunt/TA Ligase Master Mix (M0367). Sequence adapters were ligated using the NEBNext Quick Ligation Module (E6056) and samples primed and loaded using the Flow Cell Priming Kit (EXP-FLP001) on MinION R9 4.1 FLO-MIN106. Samples were sequenced for 72 h and the FAST5 files saved for base-calling and any future use. The raw reads were base-called using Guppy (v3.5.1; downloaded from https://nanoporetech.com) and adapters removed using PoreChop (v0.2.3; https://github.com/rrwick/Porechop).

3.2.8 Phage physical characterisation

Following sequencing, vB_BfrS_23 was determined to be novel. Section 3.2.10 details how vB_BfrS_23 was determined to be a novel phage.

3.2.8.1 Transmission Electron Microscopy (TEM)

TEM imaging was performed by Dr Catharine Booth, QIB, at the John Innes Centre Bioimaging Facility. Briefly, a small droplet of phage stock (~$1x10^7$ PFU/mL) was added to a formvar/carbon-coated copper TEM grid (Agar Scientific, Stansted, UK) and adsorption allowed for 1 min. Filter paper was used to remove excess liquid. A small droplet of 2 % uranyl acetate (BDH 10288) was added to the grid surface, left for 1 min and excess liquid removed with filter paper. Grids were allowed to dry fully prior to imaging using a Talos F200c TEM with Gatan Oneview digital camera.

3.2.8.2 Host range assay

A total of eight *B. fragilis* strains were used to determine the host range and specificity of vB_BfrS_23. The strains were selected from a freeze-dried collection curated by Dr Ella Bond of Institute of Food Research (Table 3.4). The ampoules were carefully opened in an anaerobic cabinet (5 % CO_2 and 37 ˚C) using a scoring stylus. The freeze-dried cells were rehydrated using ~ 100 µL FAB, inoculated into FAB and incubated anaerobically (5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure) for 12-16 h. Each strain was streak-diluted onto BPRM agar plates and incubated anaerobically (5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure) for 12-16 h. A colony from each plate was cultured in BPRM broth to exponential phase (OD_{620} 0.3-0.33) prior to incorporation into double BPRM agar overlays. Dilutions of vB_BfrS_23 were spot onto the double agar overlay and incubated anaerobically for 16 h (5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure). Plates were observed for plaques following incubation.

Table 3-4: *B. fragilis* strains used for host range assay, subtype and isolation site

Strain	DNA homology subtype	Isolated from
NCTC 9343^{T}	I	Appendix abscess
VPI 2362	I	Liver abscess
VPI 2557	I	Septic arthritic joint
VPI 2360	II	Pus from shell fragment wound
VPI 2393	II	Unknown
VPI 2361	Unknown	Liver abscess
NCTC 8560	Unknown	Post appendectomy abscess
NCTC 9344	Unknown	Septic operation wound
GB-124*	Unknown	Sewage

*, Positive control.

3.2.8.3 One-step growth curve

A one-step growth curve was performed by Dr Mohammed Tariq to determine latency period and burst size of vB_BfrS_23[94]. Firstly, 9.9 mL of mid-exponential (OD_{620} 0.5) host strain *B. fragilis* GB-124 was mixed with 0.1 mL of 1x10^{7} PFU/mL of vB_BfrS_23 and phage adsorption allowed for 5 min. A ten-fold dilution (final dilution: 1x10^{1}) of the inoculum was made from 0.1 mL. An adsorption control was created from 1 mL of the 1x10^{3} flask dilution and added to 50 µL of $CHCl_3$. It was kept on ice for the duration of the experiment (less than 4 h). At set time points, 0.1 mL of each dilution was mixed with 200 µL of bacterial host in BPRM in 0.35 % (w/v) BPRM agar and poured onto BPRM agar plates. Plaques were observed following anaerobic incubation (5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure for 12-16 h). The data were normalised by multiplying the adsorption control and the value adjusted by dilution factor. The burst size was determined as previously described[94].

The eclipse period was determined by taking 475 µL of suspension at each time point, mixing with 25 µL of chloroform (5% v/v) and keeping on ice following a brief vortex. An aliquot (100 µL) from each time point was added to 200 µL of bacterial host in 0.35 % (w/v) BPRM agar and poured onto BPRM agar plates. Plaques were observed following anaerobic incubation (5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure for 12-16 h. The one-step growth curve and eclipse experiment were repeated to produce three biological replicates.

3.2.8.4 Thermal assay

The thermal assay was performed with Dr Mohammed Tariq and Rik Haagmans. A thermal assay was used to assess the stability of vB_BfrS_23 at 4, 24, 30, 37, 40, 45, 60 or 80 ˚C for 15, 30 or 60 min (out of direct sunlight). This temperature range was selected due to the environmental origin of the phage. Following exposure to the differing temperatures and times, the tubes were cooled to room temperature and pulse-centrifuged to remove condensation from their walls. The bacterial host strain was grown to exponential phase (OD_{620} 0.3-0.33) as mentioned previously. Serial dilutions of the temperature-exposed phage (100 µL) were mixed with 200 µL of host strain culture in 5 mL of BPRM semi-soft agar (0.35 % w/v) and poured onto BPRM agar plates. The plates were anaerobically (5 % CO_2, 5 % H_2 and 90 % N at 37 ˚C and ~ 25 psi pressure) incubated for 18 h and plaques counted on plates between 30 and 300 PFU. The thermal assay was repeated three times.

3.2.9 Phage genome assembly and annotation

The Illumina MiSeq- and MinION-generated reads were co-assembled using UniCycler (v.0.4.8) and annotated using RAST[95-98]. The putative functions of the coding sequences (CDSs) were predicted using NCBI-nr (accessed: 15th June 2020) and Conserved Domain Database (CDD; accessed 15th June 2020) searches using Blastp and tBlastn. Hits were considered significant for Blastp and tBlastn if the e-values were lower than $1e^{-5}$ at ≥60% protein identity[99]. For CDD searches, hits were considered significant if they had an e-value of 0.01 or lower[100,101]. All hits were manually checked for accuracy.

3.2.10 Phage genome comparison

The genome of vB_BfrS_23 was compared to other *B. fragilis* phage (Barc2635, B40-8 and ϕB124-14; Table 3.5)[101,102]. The GenBank and fasta files of these phage were downloaded from NCBI. Due to the orientation of B40-8 and Barc2635, a reverse complement of their genomes was generated using Artemis (v.18.1.0) and a new fasta and GenBank file produced[103]. The nucleotide sequences were aligned using ClustalW 2.1 (default parameters) and a fasta alignment file generated by EMBOSS seqret[104,105]. The alignment file was input to Gubbins-FastTree (v. 2.3.4) to generate a newick phylogenetic tree (v.2.3.4)[106,107]. Blastn suite-2sequences was used to generate a base comparison table[99]. The comparison and GenBank files were imported into R (v.3.5.2) and a genome comparison plot generated using GenoPlotR (v.0.8.9). Annotations relating to predicted product were retained and coloured according to function (structure, replication and regulation, DNA packaging and lysis).

Table 3-5: Overview of publicly available *B. fragilis* phage genomes compared to vB_BfrS_23 (blastn)

Phage	Genome size (bp)	GC %	Identity with vB_BfrS_23 (%)	Query coverage to vB_BfrS_23 (%)	Accession
B40-8	44,929	38.6	95.59	73	FJ008913.1
Barc2635	45,990	38.9	95.90	85	MN078104.1
ϕB124-14	47,159	38.7	97.41	86	HE608841.1

3.2.10.1 Phylogenetic tree of large terminase subunit and tail fibre

The large terminase subunit and tail fibre of vB_BfrS_23 and other phage were used to construct a phylogenetic tree. Briefly, the coding region for the large terminase subunit and tail fibre of vB_BfrS_23 were submitted to blastp (default parameters)[99]. The amino acid sequences of the top ten hits (sorted by E-value) were downloaded and aligned with clustalW 2.1 (default parameters)[104]. The alignment file was inputted to Gubbins-FastTree (v.2.3.4) to generate a newick format maximum likelihood phylogenetic tree[106,107]. Metadata from the blastp results were used to generate a heatmap in R (v.3.5.2) with ggtree (v. 1.14.6) and phangorn (v. 2.5.5). Prophage regions were predicted using PHASTER web server (https://phaster.ca/)[108].

3.2.11 *Bacteroides* phage phylogeny

3.2.11.1 Creation of a Bacteroides phage dataset

All complete publicly available *Bacteroides* phage genomes were downloaded from NCBI Virus (accessed: 17/09/2020; https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/), including vB_BfrS_23 (Table 3.6). The IMG/VR database (v.3) was searched for Uncultivated Viral Genomes (UviGs) with *Bacteroides* as the predicted host[91]. Genomes were filtered by completeness (> 70 %) and non-prophage genomes retained for further analysis. Additionally, Gut Phage Database (GPD) was also searched using similar parameters as above[109]. Only category 1, 2 and 3 GPD VirSorter-identified phage sequences were retained. VirSorter assigns predicted phage sequences to categories (1-5) based on presence of *Caudovirales* hallmark genes and represents the confidence of the program in accurately determining a phage sequence. Category 1 phage sequences represent the most confident predictions. A detailed explanation of the categories can be found within the VirSorter article[42]. A representative sequence from each crAss-like phage genera, as stated in Guerin *et al.*, was also included in the analysis (Table 3.7). CD-HIT-EST-2D (v.4.8.1, cut off threshold: 1) was

used to remove redundant sequences between IMG/VR and GPD[110]. The curated dataset was annotated using prokka (v.1.14.6) with the metagenome option used[111].

3.2.11.2 Gene-sharing networks with vConTACT

faa and tsv files were generated from the prokka output using a python script developed by Dr. Alikhan, QIB (v.0.1.0; prokka2vcontact.py). These files were used as input for vConTACT (v.2.0) with the following parameters: *--rel-mode 'Diamond' --db 'ProkaryoticViralRefSeq94-Merged'--pcs-mode MCL --vcs-mode ClusterONE*[36,37]. The network (c1.ntw) and annotation file (genome_by_genome_overview.csv) from vConTACT were input to Cytoscape (v.3.7.2) for network visualisation[112]. The cluster(s) containing *B. fragilis* phage genomes (vB_BfrS_23, ϕB124-14, B40-8 and Barc2635) were identified in the genome_by_genome_overview.csv file. Due to the size of the dataset, singletons, outliers and overlap genomes were ignored and removed from further analyses.

3.2.11.3 Selection of representative and reference genomes

The completeness and contamination of sequences within each cluster were assessed with checkV (v.0.7.0)[113]. One sequence from each cluster was selected according to the following criteria: i) complete genome; ii) contamination < 5%; iii) no warnings. If no complete genomes were available within the cluster, the highest quality genome with the lowest contamination was selected. Additionally, if there were multiple genomes that met the above criteria a genome was selected at random.

The representative genome from each cluster was inputted to ViPTree server (v.1.9) and a proteomic tree generated based on genome-wide sequence similarities computed by tBLASTx[27]. Default parameters were used. Reference genomes within the same clade as a representative genome were selected for generation of the phylogenetic tree.

Table 3-6: Publicly available *Bacteroides* phage

Accession	Species	Genome length (bp)	Isolation location	Isolation source	Host
NC_049977	crAss001	102679	Ireland	Human faeces	*B. intestinalis* 919/174
NC_016770	B124-14	47159	United Kingdom (Sussex)	Raw sewage	*B. fragilis* GB-124
NC_011222	B40-8	44929	-	Raw sewage	*B. fragilis* HSP40
MT635598	Bacuni_F1	40421	-	Unknown faeces	*Bacteroides* sp.
MT630433	vB_BfrS_23	48011	United Kingdom	Wastewater effluent	*B. fragilis* GB-124
MT074134	ARB14	37476	USA: Ann Arbor, MI	Wastewater effluent	*B. thetaiotaomicron*
MT074135	ARB25	37389	USA: Ann Arbor, MI	Wastewater effluent	*B. thetaiotaomicron*
MT074136	DAC15	99494	Bangladesh: Dhaka	Sewer-adjacent pond water	*B. thetaiotaomicron*
MT074137	DAC16	178147	Bangladesh: Dhaka	Sewer-adjacent pond water	*B. thetaiotaomicron*
MT074138	DAC17	98900	Bangladesh: Dhaka	Sewer-adjacent pond water	*B. thetaiotaomicron*
MT074139	DAC19	178921	Bangladesh: Dhaka	Sewer-adjacent pond water	*B. thetaiotaomicron*
MT074140	DAC20	178920	Bangladesh: Dhaka	Sewer-adjacent pond water	*B. thetaiotaomicron*
MT074141	DAC22	179283	Bangladesh: Dhaka	Sewer-adjacent pond water	*B. thetaiotaomicron*
MT074142	DAC23	179161	Bangladesh: Dhaka	Sewer-adjacent pond water	*B. thetaiotaomicron*
MT074143	HNL05	37887	USA: Honolulu, HI	Wastewater effluent	*B. thetaiotaomicron*
MT074144	HNL35	37928	USA: Honolulu, HI	Wastewater effluent	*B. thetaiotaomicron*
MT074145	SJC01	38129	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074146	SJC03	166827	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074147	SJC09	38149	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074148	SJC10	37392	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074149	SJC11	38137	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074150	SJC12	38328	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074151	SJC13	38497	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074152	SJC14	38202	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074153	SJC15	38150	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074154	SJC16	38138	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074155	SJC17	38127	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074156	SJC18	37398	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074157	SJC20	37449	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074158	SJC22	38120	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074159	SJC23	38546	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MT074160	SJC25	38175	USA: San Jose, CA	Wastewater effluent	*B. thetaiotaomicron*
MN078104	Barc2635	45990	Spain: Barcelona	Raw sewage	*B. fragilis*
BK010646	p00	42831	Unknown	Unknown	*P. dorei* CL02T12C06

Table 3-7: Representative crAss-like phage and candidate genera

Candidate Genera	Phage ID	Length (bp)	Location
I	p-crassphage[63]	97065	Unknown
II	cs_ms_21[70]	97421	Ireland
III	HvCF_A6_ms_4[70]	91332	Ireland
IV	SRR4295175[70]	96082	USA
V	Sib1_ms_5[70]	92132	Ireland
VI	Fferm_ms_11[70]	104564	Ireland
VII	Inf125_s_2[70]	102169	Ireland
VIII	Eld241_T0_s_1[70]	103133	Ireland
IX	ERR975045_s_1[70]	94037	Malawi
X	ERR844030_ms_1[70]	100426	USA

3.2.11.4 Generation of phylogenetic tree

A *Bacteroides* phage phylogenetic tree was generated using ViPTreeGen (v.1.1.2) with default settings to produce a maximum likelihood tree[27]. A representative genome from each cluster identified by vConTACT, reference genomes identified by VipTree and all sequences from the cluster (VC_100) containing isolated *B. fragilis* phage (vB_BfrS_23, ϕB124-14, B40-8 and Barc2635) were used as inputs[36,37]. The resulting maximum likelihood tree file was inputted to FigTree (v.1.4.4)[114]. The tree was rooted at the midpoint and annotated in Adobe Illustrator (v.24.0.6).

3.2.11.5 Identification of orthologous proteins

The sequences used to generate the phylogenetic tree in the above section were searched for orthologous proteins. Orthofinder (v.2.2.6) was used to identify orthogroups using faa files generated by prokka (v. 1.4.6)[111,115].

3.2.12 Analysis of novel *B. fragilis* phage family

3.2.12.1 Pairwise intergenomic similarity

VIRIDIC (Virus Intergenomic Distance Calculator) was used to determine genomic similarities between all sequences within VC_100 (http://rhea.icbm.uni-oldenburg.de/VIRIDIC/; accessed August 2020; default settings)[33]. Putative genus clusters were identified according to

intergenomic similarity scores (95 % for species and 70 % for genera). The intergenomic similarity scores and genus clusters were used to generate a heatmap in R (v.3.5.2), ggdendro (v.1.20) and ggplot2 (v.3.3.2). Adobe Illustrator (v.24.0.6) was used to finalise the heatmap. The closest reference sequences to VC_100 were determined using VipTree server (v.1.9) and intergenomic similarities between all VC_100 and collected reference sequences investigated[27].

3.2.12.2 Phylogenetic tree

A phylogenetic tree of all sequences within the cluster was generated from the fasta files using VipTreeGen (v.1.1.2) with default settings[27]. The resulting asc newick file was input to FigTree (v.1.4.4) and tree rooted at the midpoint[114]. The tree was annotated in Adobe Illustrator (v.24.0.6).

3.2.12.3 Identification of orthologous proteins

Orthogroups were identified using OrthoFinder (v.2.2.6) with default settings[115]. A heatmap of gene count per orthogroup by phage was generated using R (v.3.5.2) and ggplot2 (v.3.3.2). Adobe Illustrator was used to annotate the heatmap (v.24.0.6). Orthologues shared across the viral cluster (VC) and within each genus cluster were identified and putative protein function determined using NCBI blastp (https://blast.ncbi.nlm.nih.gov/Blast.cgi; accessed: November 2020)[99]. Hits were considered significant for blastp if the e-values were lower than $1e^{-5}$ at ≥ 40 % protein identity. The percentage of shared orthologues between each phage genome in VC_100 was calculated using data generated by OrthoFinder. A heatmap of percentage shared orthologues within VC_100 was generated using R (v.3.5.2), ggdendro (v.1.20) and ggplot2 (v.3.3.2).

Orthologous proteins between VC_100 and all representative sequences were also identified using the above method with OrthoFinder (v.2.2.6).

3.2.12.4 Comparison to additional VC

It was noted in the phylogenetic tree created in Section 3.2.11 that a representative sequence from another VC was placed within VC_100. A phylogenetic tree of VC_100 and VC_358 was generated using VipTreeGen (v.1.1.2) with default settings[27]. The resulting asc newick file was input to FigTree (v.1.4.4) and the tree rooted at phage uvig_314311.

Additionally, any orthologues between VC_100 and VC_358 were identified using OrthoFinder (v.2.2.6) with default settings[115]. A genome comparison map was made of the closest relative to uvig_314311 from VC_100 and VC_358 to determine the regions of homology. The sequence selected from VC_358 was reverse-complemented using Artemis (v.18.1.0) and a new GenBank and fasta file produced[103]. Blastn suit-2sequences was used to generate base comparison table. The comparison and GenBank files were imported into R (v.3.5.2) and a genome comparison plot generated using GenoPlotR (v.0.8.9). Predicted protein products for the three genomes were obtained using blastp (accessed: November 2020)[99]. Hits were considered significant for blastp if the e-values were lower than $1e^{-5}$ at ≥ 40 % protein identity. Annotations relating to predicted product were retained and coloured according to function (structure, replication and regulation, DNA packaging and lysis).

3.2.12.5 Comparison to crAss-like phage large terminase subunit

The large terminase subunit (TerL) was used to determine the phylogenetic relationship between VC_100 and crAss-like phage. The TerL protein sequence was extracted from the crAss-like phage and from all sequences with an identifiable protein in VC_100. The TerL protein from *Cellulophaga* phage phi18:2 (accession KC821627) and phi12:1 (accession KC821613) were also included as outliers. The protein sequences were aligned using MUSCLE (v.3.8.1551) and TerL phylogeny inferred using IQTree (v.1.6.10, maximum bootstrap: 1000) with default settings and best-fit model determined using ModelFinder[116-119]. The resulting maximum likelihood tree was visualised in FigTree (v.1.4.4), rooted at the *Cellulophaga* phage and bootstrap percentage determined. The figure was finalised in Adobe Illustrator (v.24.0.6).

3.3 Results

3.3.1 *Bacteroides* strain growth dynamics

BHI medium is frequently used for the culture of *Bacteroides* species[120]. However, the current protocol established by Dr James Ebdon for phage screening cultures the host species (GB-124) in BPRM[121]. This medium provides nutrients for rapid bacterial growth and increased phage infectivity. Growth curves of the six additional *Bacteroides* strains used for environmental screening were created to determine growth dynamics in BHI and BPRM (Figure 3.1).

All strains showed a shorter lag phase and quicker exponential phage in BPRM compared to BHI. Additionally, strains reached a higher OD in the BPRM medium. The bacterial host is required to

reach an OD_{620} between 0.3 and 0.33 before incubation with the phage. Based on the growth curves above, all strains reached the optimum OD within 5 h. Therefore, BPRM was preferred over BHI for phage screening assays.

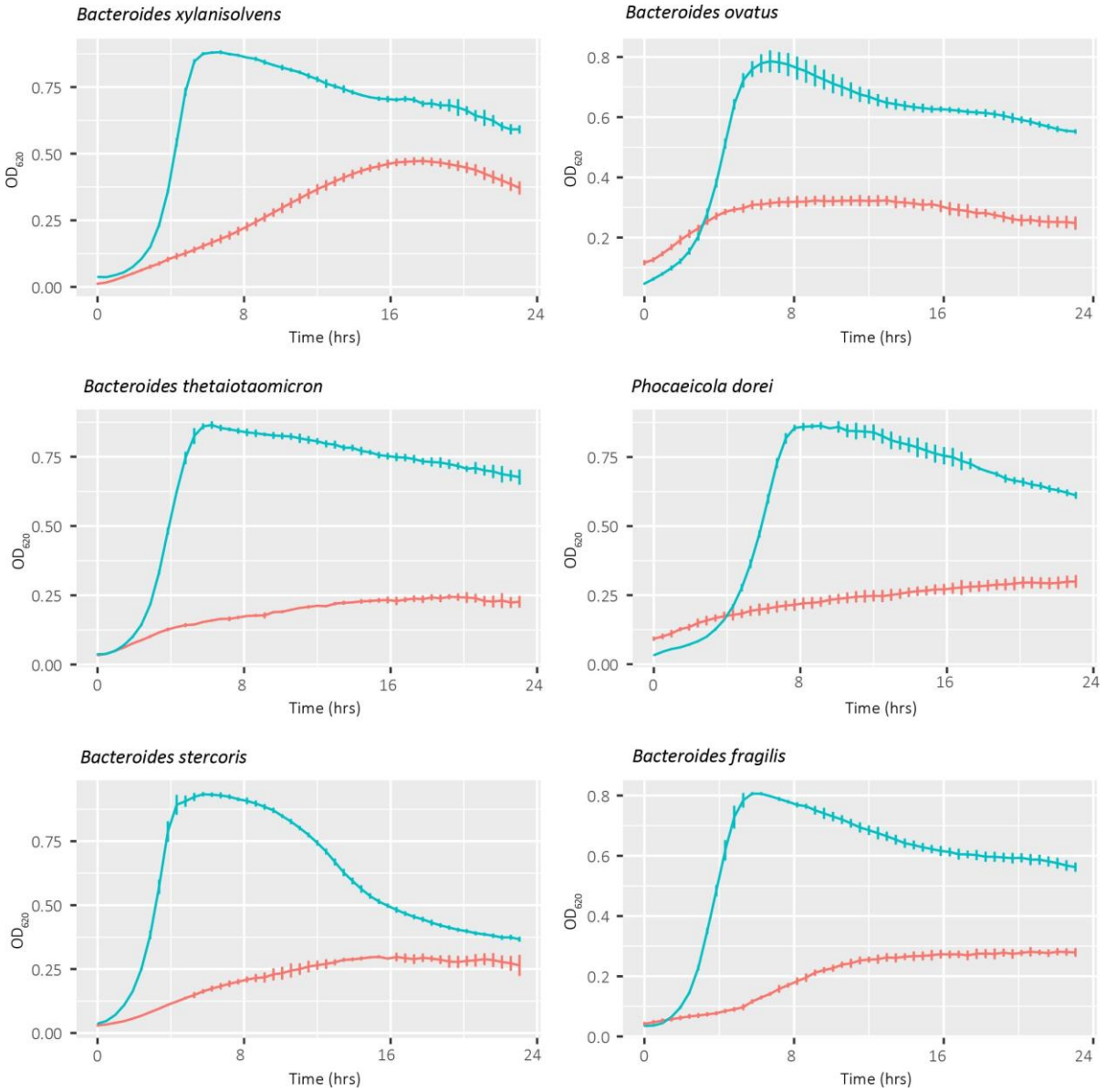

Figure 3.1: Growth curve of *Bacteroides* strains in BHI and BPRM broths

Anaerobic growth curve over 24 h read at OD_{620} of six *Bacteroides* strains in BHI (orange line) and BPRM (blue line) (n = 3 technical replicates). The OD_{620} is shown on the y axis and time in hours shown on the x axis.

For the screening of complex environmental samples (e.g. sewage and faeces), the addition of kanamycin (100 µg/mL) to BPRM plates and overlays was required to reduce bacterial contamination. The susceptibility of the six additional *Bacteroides* strains to kanamycin has not been tested previously. A 1974 paper reported resistance of *B. fragilis* clinical isolates to 1000 µg/mL kanamycin discs in BHI[122]. A MIC assay was performed to determine the susceptibility of the *Bacteroides* strains to kanamycin (1000 µg/mL, 200 µg/mL, 100 µg/mL, 50 µg/mL, 10 µg/mL, 1 µg/mL and 0.1 µg/mL) and if the concentration affected the growth conditions. None of the kanamycin concentrations tested produced a reduction in OD_{620} in the seven *Bacteroides* strains (Figure 3.2). Therefore, a concentration of 100 µg/mL kanamycin was used in all environmental phage screening assays.

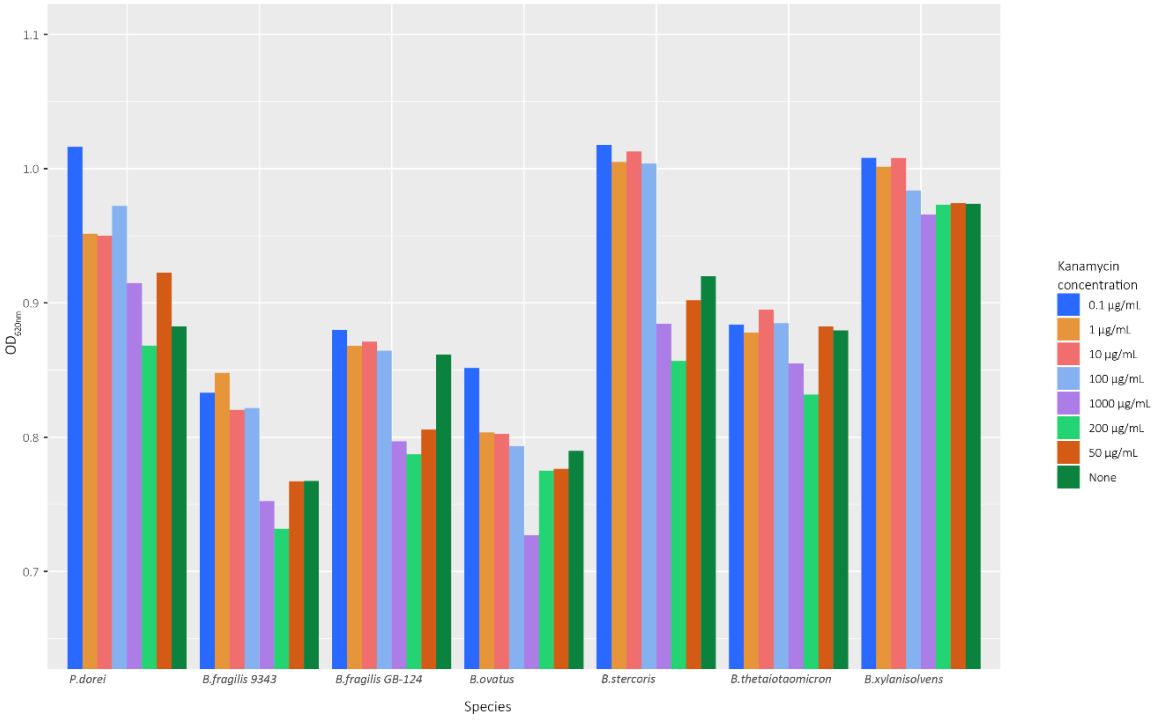

Figure 3.2: **MICs of kanamycin with *Bacteroides* strains**

Anaerobic growth over 15 h of seven *Bacteroides* strains with different kanamycin concentrations (n = 2 biological replicates). The OD_{620} is shown on the axis and each Bacteroides species shown on the x axis. The coloured bars represent a different kanamycin concentration and correspond to the legend colours.

3.3.2 Environmental phage screening

A total of 15 environmental samples were screened for phage using seven *Bacteroides* strains. Plaques were only observed on *B. fragilis* GB-124 (39 plaques) and *B. thetaiotaomicron* VPI-5482 (one plaque) plates incubated with sewage filtrate. All plaques showed clear edges and diameter ranged from 0.1 mm to 3 mm. Plaques with unique morphology were collected for future analysis.

3.3.3 Phage characteristics

Only one novel phage (vB_BfrS_23) was identified and isolated from the environmental sample screening. vB_Bfrs_23 is a lytic phage capable of infecting GB-124 generating different plaque sizes ranging from 0.5 mm to 2 mm (Figure 3.3a). TEM identified vB_BfrS_23 as belonging to the family *Siphoviridae* of the order *Caudovirales* and was ~200 nm in length with a ~150 nm non-contractile tail and ~50 nm icosahedral head (Figure 3.3b). A host range assay with nine *B. fragilis* strains (Table 3.4) revealed vB_BfrS_23 was only able to infect GB-124. A burst size of ~44 and latency period of ~37 min was determined from a one-step growth curve (Figure 3.3c). The eclipse period was determined to be ~23 min. Additionally, the phage remained stable to temperatures between 4 ˚C and 45 ˚C (Figure 3.3d). A reduction in viability was observed at 65 ˚C and significant reduction at 70 ˚C. No plaques were observed at 80 ˚C, suggesting complete loss of vB_BfrS_23 viability. A slight increase in PFU/mL was seen between 40 ˚C and 45 ˚C, with plaques being of uniform size (0.5 mm).

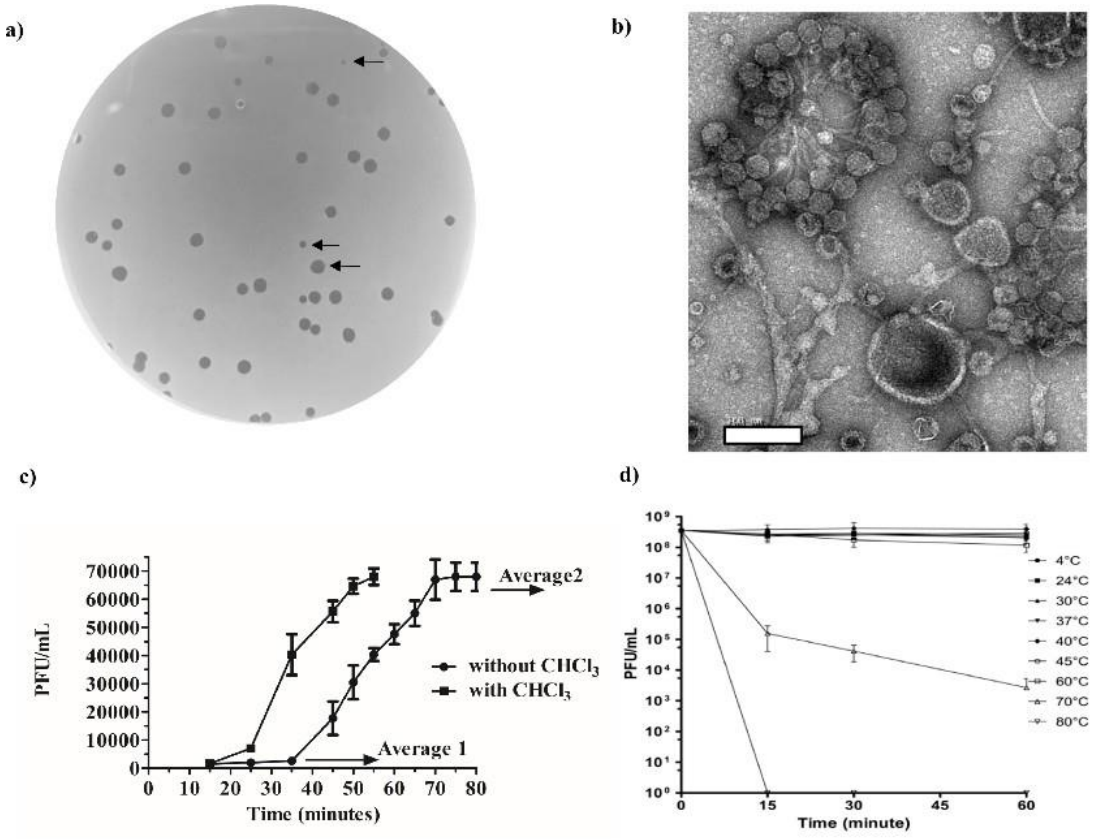

Figure 3.3: Physical and biological characteristics of phage vB_BfrS_23

a) Differing plaque sizes (0.5 mm to 2 mm) seen on a lawn of *B. fragilis* GB-124. b) Negatively stained TEM image of vB_BfrS_23 aggregates and single phage. Scale bar, 200 nm. c) One-step growth curve of vB_BfrS_23 with error bars showing SEM values (n = 3 biological replicates). d) Thermal stability of vB_BfrS_23 at temperatures ranging from 4 ˚C to 80 ˚C with error bars showing SEM values (n = 3 biological replicates). This figure is reproduced from Tariq et al. (2018) (see Appendix 2) under terms of the Creative Commons Attribution License (CC BY) of *Frontiers in Microbiology*[135].

3.3.4 Phage genome characteristics and comparison

vB_BfrS_23 is a dsDNA phage of 48,011 bp with a GC content of 38.6 %, encoding 73 putative CDSs (Table 3.8). Of these 73 CDSs, 14 had a putative function, eight contained conserved domain signatures and 10 showed no significant homology to any protein within the database. A total of 27 CDSs shared highest homology to genes in ϕB124-14, 27 to Barc2635, eight to B40-8 and one to *B. ovatus* (Figure 3.4, Table 3.8). Most CDSs with assignable function were associated with genome structure, and replication and regulation (Figure 3.4c).

As of June 2020, three additional *B. fragilis* phage have been isolated and characterised: Barc2635, B40-8 and ϕB124-14 (Table 3.5). vB_BfrS_23 shared highest nucleotide sequence similarity with ϕB124-14 and least with B40-8.

Genome comparison of *B. fragilis* phage revealed significant similarity in genome organisation with four distinct modules: replication and regulation, lysis, DNA packaging and genome structure (Figure 3.5). The vB_BfrS_23 genome is lacking homology to five putative proteins when compared to the ϕB124-14 genome; including a capsid associated protein, mismatch repair protein, resolvase, nuclease and an additional anti-repressor. The Barc2635 genome encodes two additional proteins not located within the vB_BfrS_23 genome: tail assembly chaperone protein and capsid-associated protein.

Similar to ϕB124-14 and ϕB40-8, vB_BfrS_23 lacks an obvious module related to phage lifestyle and contains only one putative protein that eludes to a lytic life cycle (CDS18). A lytic module is defined by the absence of a recognizable integrase gene; a ubiquitous gene utilized by prophage for integration into the bacterial chromosome. The lack of an integrase gene and method of phage isolation without any obvious prophage induction highly suggests vB_BfrS_23 is a lytic phage. This CDS showed closest homology to a putative peptidase in ϕB124-14 and contained a peptidase superfamily domain. The peptidase sits within a cluster of unassignable protein function; suggesting it may be a putative lytic module.

Seven CDSs were assigned a predicted function relating to virus replication and regulation: recombination protein, thymidylate synthase, exoribonuclease, anti-repressor, DNA replication protein, HNH endonuclease and ssDNA binding protein. CDS11, encoding a putative thymidylate synthase, is present in all *B. fragilis* phages[101]. This protein is a key enzyme in the synthesis of 2'-deoxythymidine-5', an essential precursor for DNA replication[123]. Additionally, a conserved domain region encodes for ThyA-like enzyme. CDS7 (recombination protein) and CDS70 (anti-repressor) were also encoded within the replication and regulation genome module. These are involved in prophage insertion, formation and re-entry into a lytic lifestyle[124]. CDS61 encoded for a HNH endonuclease protein, present in many phage and prophage. Phage HNH endonucleases are commonly located close to the large terminase CDS and are highly conserved[125].

Table 3-8: Predicted coding regions and protein functions of phage vB_BfrS_23

CDS	Start	End	Size (aa)	Predicted Function	Putative product	E value*	aa identity (%)*
1	61	2	20	-	No significant hits	-	-
2	158	48	36	-	No significant hits	-	-
3	337	155	60	-	No significant hits	-	-
4	592	347	81	Unknown	Hypothetical protein F3B42_14490 [*Bacteroides ovatus*]	9.00E-66	71/81 (88 %)
5	1563	619	314	Unknown	Hypothetical protein B124-14_003 [*Bacteroides* phage B124-14]	0	305/314 (97 %)
6	2304	1576	242	Unknown	Hypothetical protein B40-8019 [*Bacteroides* phage B40-8]	0	211/233 (91 %)
7	2995	2363	210	Replication & Regulation	Putative essential recombination protein [*Bacteroides* phage B124-14]; ERF superfamily (Pfam 04404)	0	199/210 (95 %)
8	3252	3001	83	Unknown	Hypothetical protein B124-14_005 [*Bacteroides* phage B124-14]	2.00E-63	64/65 (98 %)
9	3644	3249	131	Unknown	Hypothetical protein B124-14_006 [*Bacteroides* phage B124-14]	1.00E-131	125/131 (95 %)
10	4088	3915	57	Unknown	Hypothetical protein B124-14_007 [*Bacteroides* phage B124-14]	2.00E-57	56/57 (98 %)
11	4882	4085	265	Replication & Regulation	Putative thymidylate synthase [*Bacteroides* phage B40-8]; Thymidylate synthase (Pfam 00303)	0	261/265 (98 %)
12	5793	4948	281	Unknown	Hypothetical protein B124-14_011 [*Bacteroides* phage B124-14]	0	273/278 (98 %)
13	5983	5786	65	Unknown	Hypothetical protein B124-14_012 [*Bacteroides* phage B124-14]	4.00E-52	59/65 (91 %)
14	6446	6030	138	Unknown	Hypothetical protein B40-8013 [*Bacteroides* phage B40-8]	1.00E-144	134/138 (97 %)
15	6745	6464	93	Unknown	Hypothetical protein B124-14_014 [*Bacteroides* phage B124-14]	3.00E-96	92/93 (99 %)
16	7009	6764	81	Unknown	Hypothetical protein B124-14_015 [*Bacteroides* phage B124-14]	5.00E-78	78/81 (96 %)
17	8021	7416	201	Unknown	Hypothetical protein B124-14_016 [*Bacteroides* phage B124-14]	0	187/201 (93 %)
18	8413	8018	131	Lysis	Putative peptidase [*Bacteroides* phage B124-14]; peptidase M15 (Pfam 08291)	9.00E-126	123/131 (94 %)

CDS	Start	End	Size (aa)	Predicted Function	Putative product	E value*	aa identity (%)*
19	8947	8456	163	Unknown	Hypothetical protein B124-14_018 [*Bacteroides* phage B124-14]	3.00E-176	161/163 (99 %)
20	9315	8947	122	Unknown	Hypothetical protein B124-14_019 [*Bacteroides* phage B124-14]	5.00E-130	122/122 (100 %)
21	15462	9397	2021	Structure	Putative phage tail fibre protein [*Bacteroides* phage B124-14]	0	1006/1153 (87 %)
22	17668	15494	724	Unknown	Hypothetical protein B124-14_021 [*Bacteroides* phage B124-14]	0	718/724 (99 %)
23	20793	17668	1041	DNA packaging	Putative DNA segregation protein [*Bacteroides* phage B124-14]	0	1030/1041 (99 %)
24	21557	20805	250	Unknown	Hypothetical protein B124-14_023 [*Bacteroides* phage B124-14]	0	248/250 (99 %)
25	21831	21544	95	Unknown	Hypothetical protein B124-14_024 [*Bacteroides* phage B124-14]	3.00E-91	87/95 (92 %)
26	221326	21831	101	Unknown	Hypothetical protein B124-14_025 [*Bacteroides* phage B124-14]	3.00E-100	94/101 (93 %)
27	22657	22277	126	Unknown	Hypothetical protein B124-14_026 [*Bacteroides* phage B124-14]	1.00E-98	101/126 (80 %)
28	22904	22650	84	Unknown	Hypothetical protein B124-14_027 [*Bacteroides* phage B124-14]	6.00E-82	81/84 (96 %)
29	23278	23069	69	Unknown	Hypothetical protein B124-14_028 [*Bacteroides* phage B124-14]	9.00E-58	61/69 (88 %)
30	23577	23275	100	-	No significant hits	-	-
31	23792	23628	54	-	No significant hits	-	-
32	24009	23779	76	Unknown	Hypothetical protein B124-14_031 [*Bacteroides* phage B124-14]	2.00E-67	68/70 (97 %)
33	24265	24017	82	-	No significant hits	-	-
34	24984	24310	224	-	No significant hits	-	-
35	25128	25015	37	Unknown	Hypothetical protein B124-14_031 [*Bacteroides* phage B124-14]	6.00E-22	31/33 (94 %)
36	25381	25121	86	Unknown	Hypothetical protein B124-14_032 [*Bacteroides* phage B124-14]	2.00E-87	85/86 (99 %)
37	26641	25424	405	-	No significant hits	-	-
38	26883	26674	69	Unknown	Hypothetical protein [*Parabacteroides* sp. ZJ-118]	1.00E-15	37/59 (63 %)
39	27029	26889	46	Unknown	Hypothetical protein B124-14_032 [*Bacteroides* phage B124-14]	1.00E-44	46/46 (100 %)
40	27363	27127	78	Unknown	Hypothetical protein [*Bacteroides fragilis];* Glyco_tranf_GTA_type superfamily	2.00E-27	40/60 (67 %)
41	27555	27373	60	Unknown	Hypothetical protein [*Bacteroides fragilis*]	3.00E-20	35/59 (59 %)
42	27807	27700	35	-	No significant hits	-	-

CDS	Start	End	Size (aa)	Predicted Function	Putative product	E value*	aa identity (%)*
43	28409	27840	189	Unknown	Hypothetical protein B124-14_035 [*Bacteroides* phage B124-14]	0	189/189 (100 %)
44	28851	28399	150	Unknown	Hypothetical protein B40-8045 [*Bacteroides* phage B40-8]	5.00E-164	150/150 (100 %)
45	30161	28848	437	Unknown	Hypothetical protein B124-14_037 [*Bacteroides* phage B124-14]	0	434/437 (99 %)
46	31546	30230	438	Structure	Major protein 1 [*Bacteroides* phage B40-8]	0	415/438 (95 %)
47	32515	31598	305	Unknown	Hypothetical protein B40-8042 [*Bacteroides* phage B40-8]	0	304/305 (99 %)
48	32913	32515	132	Unknown	Hypothetical protein B124-14_040 [*Bacteroides* phage B124-14]	3.00E-136	129/132 (98 %)
49	34609	32903	568	Structure	Major protein 3 [*Bacteroides* phage B40-8]	0	566/568 (99 %)
50	35401	34757	214	Structure	Putative capsid protein, major protein 2 [*Bacteroides* phage B124-14]	0	206/214 (96 %)
51	37001	35487	504	DNA packaging	Putative phage terminase large subunit [*Bacteroides* phage B124-14]; Terminase_6 family (Pfam 03237)	0	436/445 (98 %)
52	37588	36998	196	Unknown	Hypothetical protein B40-8037 [*Bacteroides* phage B40-8]	0	193/196 (98 %)
53	38410	37745	221	Unknown	Hypothetical protein B124-14_045 [*Bacteroides* phage B124-14]	0	200/221 (90 %)
54	39023	38421	200	Unknown	Hypothetical protein B40-8035 [*Bacteroides* phage B40-8]; NTP-PPase superfamily (cd11542)	0	179/200 (90 %)
55	39349	39047	100	Unknown	Hypothetical protein B124-14_047 [*Bacteroides* phage B124-14]	1.00E-98	96/100 (96 %)
56	39551	39336	71	Unknown	Hypothetical protein B124-14_048 [*Bacteroides* phage B124-14]	3.00E-71	69/71 (97 %)
57	39918	39544	124	Unknown	Hypothetical protein B40-8033 [*Bacteroides* phage B40-8]	1E-135	123/124 (99 %)
58	40188	39955	77	Unknown	Hypothetical protein B124-14_050 [*Bacteroides* phage B124-14]	2.00E-51	61/77 (79 %)
59	40794	40339	151	Replication & Regulation	Putative single-stranded DNA binding protein [*Bacteroides* phage B124-14]; SSB protein family (Pfam 00436)	3.00E-161	150/151 (99 %)
60	41057	40794	87	Unknown	Hypothetical protein B124-14_052 [*Bacteroides* phage B124-14]	6.00E-41	59/99 (60 %)
61	41485	41054	143	Replication & Regulation	Putative HNH endonuclease [*Bacteroides* phage B124-14]; HNH endonuclease (Pfam 01844)	5.00E-163	142/143 (99 %)

CDS	Start	End	Size (aa)	Predicted Function	Putative product	E value*	aa identity (%)*
62	42122	41472	216	Unknown	Hypothetical protein B124-14_054 [*Bacteroides* phage B124-14]	0	211/216 (98 %)
63	42585	42271	104	Unknown	Hypothetical protein B124-14_055 [*Bacteroides* phage B124-14]	5.00E-108	101/104 (97 %)
64	42956	42582	124	Unknown	Hypothetical protein B40-8028 [*Bacteroides* phage B40-8]	3.00E-129	119/124 (96 %)
65	43338	42979	119	Unknown	Hypothetical protein B40-8027 [*Bacteroides* phage B40-8]	1.00E-106	106/119 (89 %)
66	44213	43599	204	Unknown	Hypothetical protein B124-14_059 [*Bacteroides* phage B124-14]	0	204/204 (100 %)
67	44425	44204	73	Unknown	Hypothetical protein B124-14_060 [*Bacteroides* phage B124-14]	4.00E-76	73/73 (100 %)
68	45381	44437	314	Unknown	Hypothetical protein B124-14_061 [*Bacteroides* phage B124-14]	0	314/314 (100 %)
69	46382	45432	316	Unknown	Hypothetical protein B124-14_062 [*Bacteroides* phage B124-14]	0	308/316 (97 %)
70	487134	46379	251	Replication & Regulation	Putative phage antirepressor [*Bacteroides* phage B124-14]; Phage Rha family (Pfam 09669)	0	200/257 (78 %)
71	47426	47214	70	Unknown	Hypothetical protein B124-14_068 [*Bacteroides* phage B124-14]; Helix-turn-helix domain (Pfam 12728)	1.00E-07	29/67 (43 %)
72	47881	47561	106	Unknown	Hypothetical protein B40-8021 [*Bacteroides* phage B40-8]	4.00E-38	61/99 (62 %)
73	48009	47875	44	-	No significant hits	-	-

*E value and per cent amino acid identity according to blastp shown.

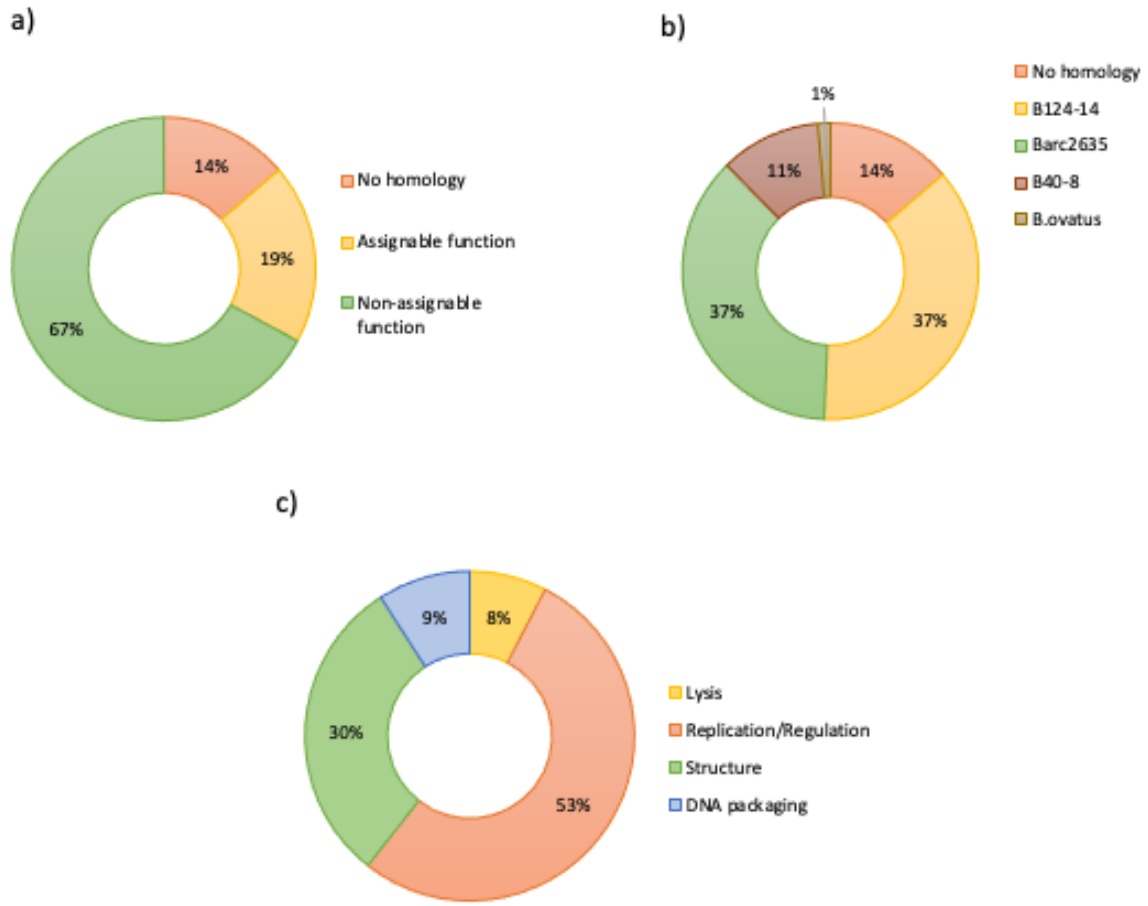

Figure 3.4: **Genome characteristics of phage vB_BfrS_23**

Percentage of vB_BfrS_23 coding region: a) with no significant homology, assignable or non-assignable function, b) with predicted protein homology origin, and c) assigned to four predicted functionalmodules. The coloured sections of each graph are described in the legend. No significant homology is defined as no significant homologous protein hits within the Blast database. Assignable function is defined as a significant hit within the Blast database where protein function could be inferred (e.g. putative phage large terminase subunit). Non-assignable function is defined as a significant hit within the Blast database to a hypothetical protein where the protein function can not be predicted. The origin of these assignable an d non-assignable predicted proteins is shown in b and includes 3 *B.fragilis* phage (B124-14, Barc2635, B40-8) and *Bacteroides ovatus*. Figure 3.4 c shows the percentage of proteins with predicted function as being involved in lysis, replication/regulation, structure or DNA packaging.

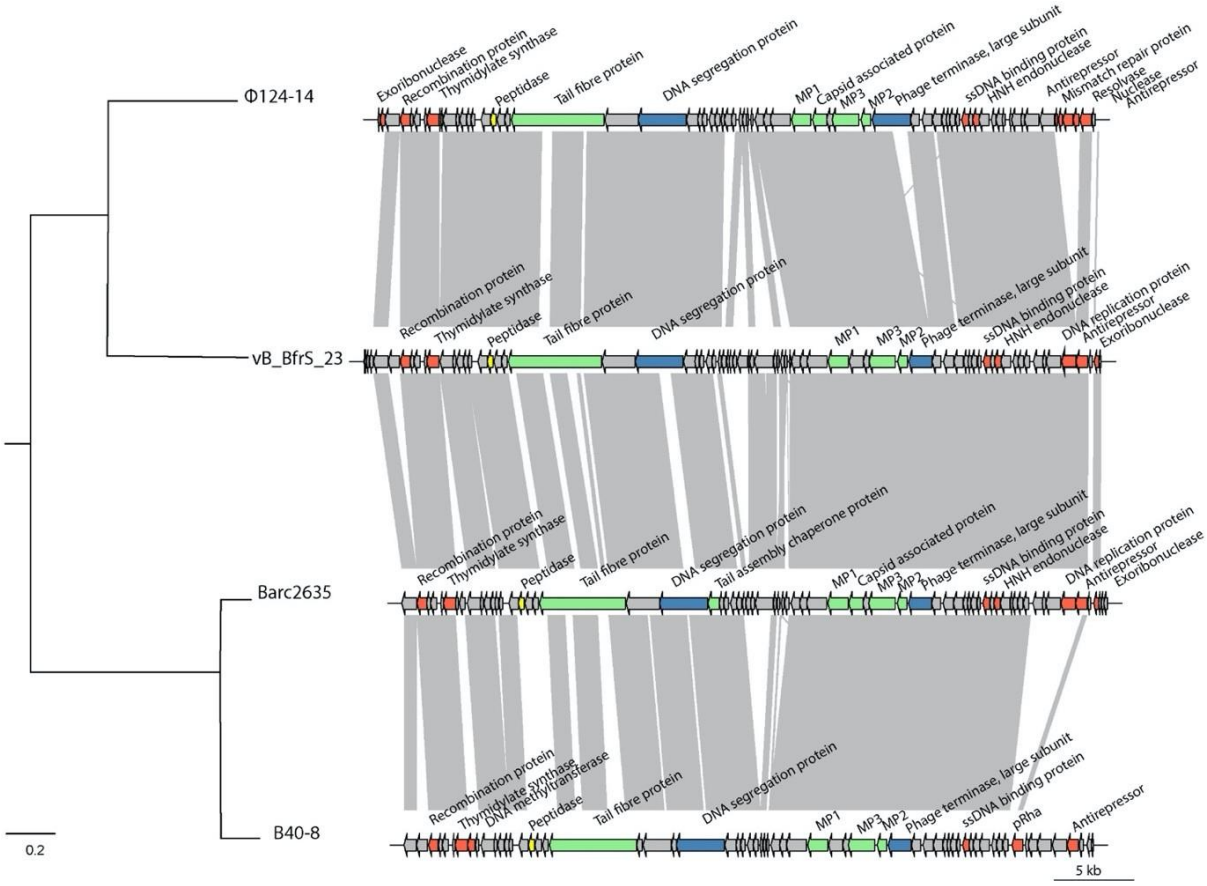

Figure 3.5: Genome comparison of phage vB_BfrS_23, ϕ124-14, Barc2635 and B40-8.

Position and orientation of each predicted coding region show for each genome. Colours of thearrows represent differing predicted protein function: red, replication/regulation; yellow, lysis;green, structure; blue, DNA packaging. Scale bar on the right-hand side of the image, genome size. Gray bars connecting phage genomes represent protein similarity according to blastp e-values. The phylogenetic tree shows the nucleotide relationship of the phage genomes. Scale barof the left-hand side of the image, substitutions per site.

The structure and DNA packaging genome module contain six CDSs, comprising 61 % of the assignable phage genome. The DNA segregation protein (CDS23) and large terminase subunit (CDS51) constitute the only structural proteins identified within the *B. fragilis* phage genomes. The large terminase subunit is involved in packaging phage DNA into the empty phage capsid and is normally a heteromultimer composed of one large and one small subunit; however, a small subunit was not identified in the *B. fragilis* phage genomes[126,127]. Four CDSs relating to phage structure were identified within vB_BfrS_23 genome: tail fibre protein (CDS21), MP1 (CDS46), MP2 (CDS49) and MP3 (CDS50). Barc2635 and ϕB124-14 also exhibit an additional structural protein: capsid-associated protein. The terminase large subunit and the tail fibre protein were used to generate phylogenetic trees and heatmap with metadata (Figure 3.6 and Figure 3.7). The terminase large subunit showed the highest identity to ϕB124-14 (98.41 %), followed by Barc2635 (98.02 %) and B40-8 (94.96 %) (Figure 3.6). Interestingly, this is not represented in the phylogenetic tree generated. Only the top 10 blastp results were used to generate the phylogenetic tree and heatmap. The remaining hits show closest homology to prophage regions within host bacteria genomes. Five of these prophage regions are within *Alistipes* species; however, the percentage identity is between 50 and 60 %.

The phylogeny of vB_BfrS_23 tail fibre protein (Figure 3.7) was investigated and revealed closest percentage identity to ϕB124-14 (82.14 %), followed by B40-8 (73.26 %) and Barc2635 (70.08 %). Interestingly, ϕB124-14 and vB_Bfrs_23 were not closest on the phylogenetic tree. No other phage tail fibre proteins were identified from the blastp search and the other hits were to hypothetical proteins within *B. fragilis* genomes. However, these were suspected prophage regions. The percentage identity and query coverage to these were relatively low (60-71 %).

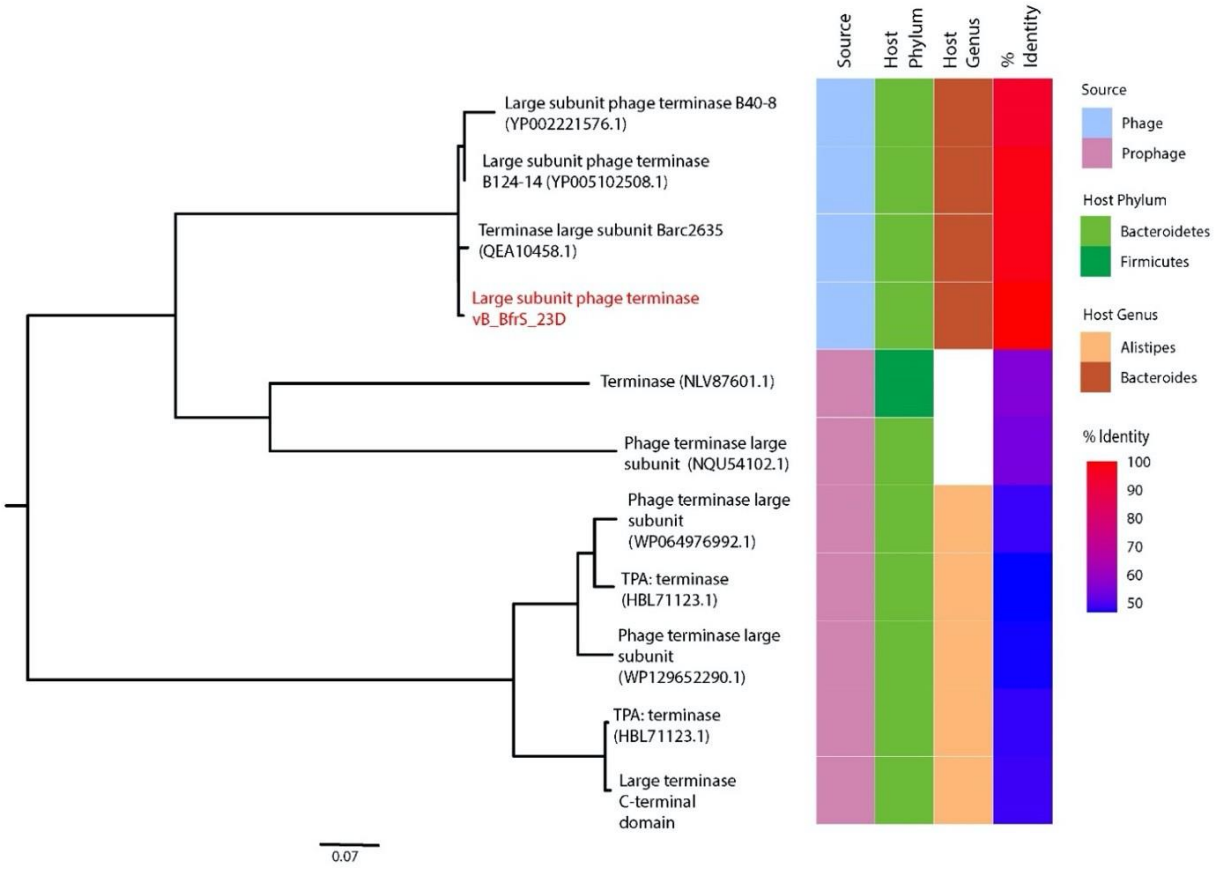

Figure 3.6: Phylogeny of vB_Bfrs_23 large subunit terminase and associated metadata.

The amino acid sequences of the top 10 blastp hits (according to e-value) were aligned with ClustalW and a maximum likelihood tree produced using FastTree[104,107]. Scale bar, amino acid substitutions. vB_BfrS_23 is in red text. Metadata were used from blastp results to create a heatmap showing the protein source (phage or prophage), host phylum (*Bacteroidetes* or *Firmicutes*), host genus (*Alistipes or Bacteroides*) and percentage identity to vB_BfrS_23 large terminase subunit.

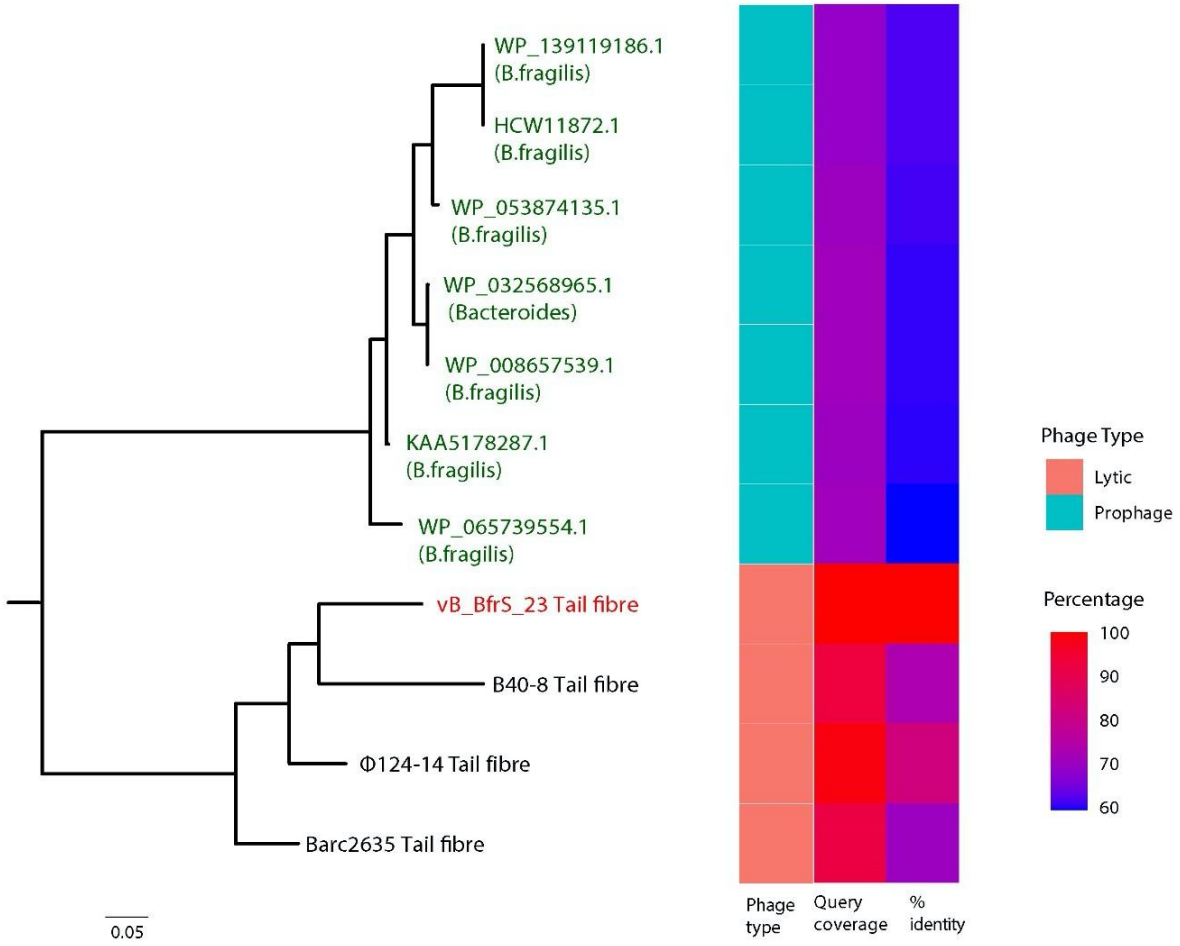

Figure 3.7: Phylogeny of vB_BfrS_23 tail protein and associated metadata

The amino acid sequences of the top 10 blastp hits (according to e-value) were aligned with ClustalW and a maximum likelihood tree produced using FastTree[104,107]. Scale bar, amino acid substitutions. vB_BfrS_23 is highlighted in red and hypothetical proteins highlighted in green. Metadata was used from blastp results to create a heatmap showing the protein source (phage or prophage), query coverage and percentage identity to vB_BfrS_23 tail fibre protein.

3.3.5 *Bacteroides* phage phylogeny

3.3.5.1 Creation of Bacteroides phage dataset

To determine the relationship between the four isolated *B. fragilis* and other *Bacteroides* phage, 2,639 predicted *Bacteroides* phage sequences were collated. Of the 39 *Bacteroides* phage genomes present on NCBI Virus, 34 were downloaded (Table 3.6). Five phage genomes were excluded due to incompleteness or duplicated genomes. The isolation hosts were *B. fragilis*, *B. thetaiotaomicron*, *P. dorei*, *B. intestinalis* and *Bacteroides* sp. The phage genome sizes ranged from 37,389 bp to 179,283 bp.

A total of 871 phage genomes with *Bacteroides* predicted host were downloaded from the IMGVR database[91]. The phage originated from a variety of sources: 829 human gut microbiota, one human oral microbiota, 25 mammal gut microbiota (foregut/large intestine), five bird gut microbiota (faecal/caeca), one environmental wetland, one mixed alcohol bioreactors, three anaerobic bioreactors, two wastewater and five unclassified. The GPD contained 2,044 predicted *Bacteroides* phage; however, 320 phage sequences were removed from further analysis due to VirSorter classification of category 4 and 5[109,128]. Redundant sequences between IMGVR and GPD were removed, resulting in 1724 phage sequences collected from GPD. All phage collected from the GPD were assembled from human faecal metagenomes or viromes. Additionally, 10 crAss-like phage (one from each proposed genera) were collected to determine taxonomic relatedness to *B. fragilis* phage (Table 3.7).

3.3.5.2 Exploration of Bacteroides phage network cluster

To examine the taxonomic classification of the curated *Bacteroides* phage dataset, a gene-sharing network was used. This newly developed software predicts genus-level groups (VCs) from the viral population used. Genus level is defined at the “sub viral cluster” level and sub-family defined at the “viral cluster” level. A network computed from 2,636 *Bacteroides* phage and 2,538 reference phage genomes (from NCBI Viral RefSeq v.85) revealed 465 VCs and 916 sub-VCs. Of these, 97 VCs were exclusively composed of *Bacteroides* phage genomes (2,340 genomes), three VCs contained genomes from both RefSeq and *Bacteroides* phage (excluding p-crAssphage, B40-8 and ϕB124-14; 10 genomes) and 365 VCs were composed of RefSeq genomes only (2,535 genomes). The three RefSeq genomes assigned to a VC with *Bacteroides* phage were *Lactococcus* phage P335 *sensu lato* (NC_004746.1; VC 358), *Clostridium* phage vB_CpeS-CP51 (KC237729; VC 220) and *Lepus americanus* faeces-associated microvirus SHP1 6472 (NC_040341, VC 411). Two *Lactococcus* phage genomes (BK5-T, NC_002796; bIL286, NC_002667) overlapped with VC 358 and VC 404 but

were not included in the VC due to the overlap with a *Lactococcus*-dominated VC. A total of 289 *Bacteroides* phage were categorised as singleton, outlier or cluster overlap and excluded from further analysis.

Visualisation of the network revealed the majority of *Bacteroides* phage formed a large distinct cluster connected to several RefSeq viral genomes (Figure 3.8). The cluster was surrounded by various other VCs but did not share a significant number of edges (or interactions) with the VCs. It was noted that five RefSeq genomes interacted with the *Bacteroides* cluster and may represent the closest known viral relatives; these included *Croceibacter* phage P2559Y/P2559S (40 and 207 interactions; NC_023614.1 and NC_018276.1), *Cellulophaga* phage phi14:2 (15 interactions; NC_021806.1) and *Riemerella* phage RAP44 (42 interactions; NC_019490.1). These results suggest that these reference phage do not share a significant proportion of their genes with the connected *Bacteroides* phage but may be related at family level. Two VCs sat outside the *Bacteroides* phage cluster and were not connected to any RefSeq genomes (VC_389 and VC_396).

A total of 27 *B.thetaiotaomicron* phage have been isolated and present on NCBI. Surprisingly, these phage did not cluster within the same VC and appeared to group according to geographic isolation location. Nineteen *B.thetaiotaomicron* isolated in the USA shared a VC (VC_388); ARB14, ARB25, HNL05, HNL35, SJC01, SJC09:18, SJC20, SJC22, SJC23, SJC25. While six additional *B.thetaiotaomicron* phage isolated in Bangladesh and USA exclusively formed a VC (VC_395); DAC16, DAC19, DAC20, DAC22, DAC23, SJC03. The two remaining *B. thetaiotaomicron* phage clustered within a VC (VC_206) with two crAss-like phage (crAss001 and Fferm_ms_11). Additionally, *B. fragilis* phage were grouped into a VC with 44 other *Bacteroides* phage genomes (VC_100).

Surprisingly, the crAss-like phage were not assigned to the same VC and further hints at the wide diversity of the crass-like phage. Inf125_s_2 and SRR4295175_ms_5 clustered within two large VCs with other genomes from the curated *Bacteroides* dataset; VC_261 and VC_266 respectively. The remaining crass-like phage were classified as outliers (eld241, ERR975045, ERR844030).

Bacteroides phage represent diverse uncharacterised taxonomic group

Due to the large dataset, a representative sequence from each of 99 *Bacteroides* phage VC was selected for production of a phylogenetic tree (Table 3.9). The genome quality of all sequences within a VC was assessed with CheckV (v.0.7.0) using the criteria specified in Section 3.2.11[113]. If the VC contained multiple sequences with high quality sequences, a sequence was randomly

selected. However, 12 VCs contained “medium quality” genome fragments. In this case, the most complete sequence was selected as a representative (74.73 - 88.02 %). Additionally, 17 representative sequences were high quality with completeness < 100 % (91.56 - 99.89 %). The remaining 70 VCs contained multiple sequences with high-quality complete genomes. The genome size of selected representative sequences ranges from 6,206 bp to 400,107 bp and number of genes from 8 to 533. Interestingly, the representative sequence with the largest genome (IMGVR_UVIG_26) was clustered within a VC (VC_396) that also contained large genomes (289,806 bp to 400,107 bp). All sequences within this cluster contained a relatively small percentage of host genes (0.77 – 2.07%), did not contain contamination and were of high quality (Table 3.10).

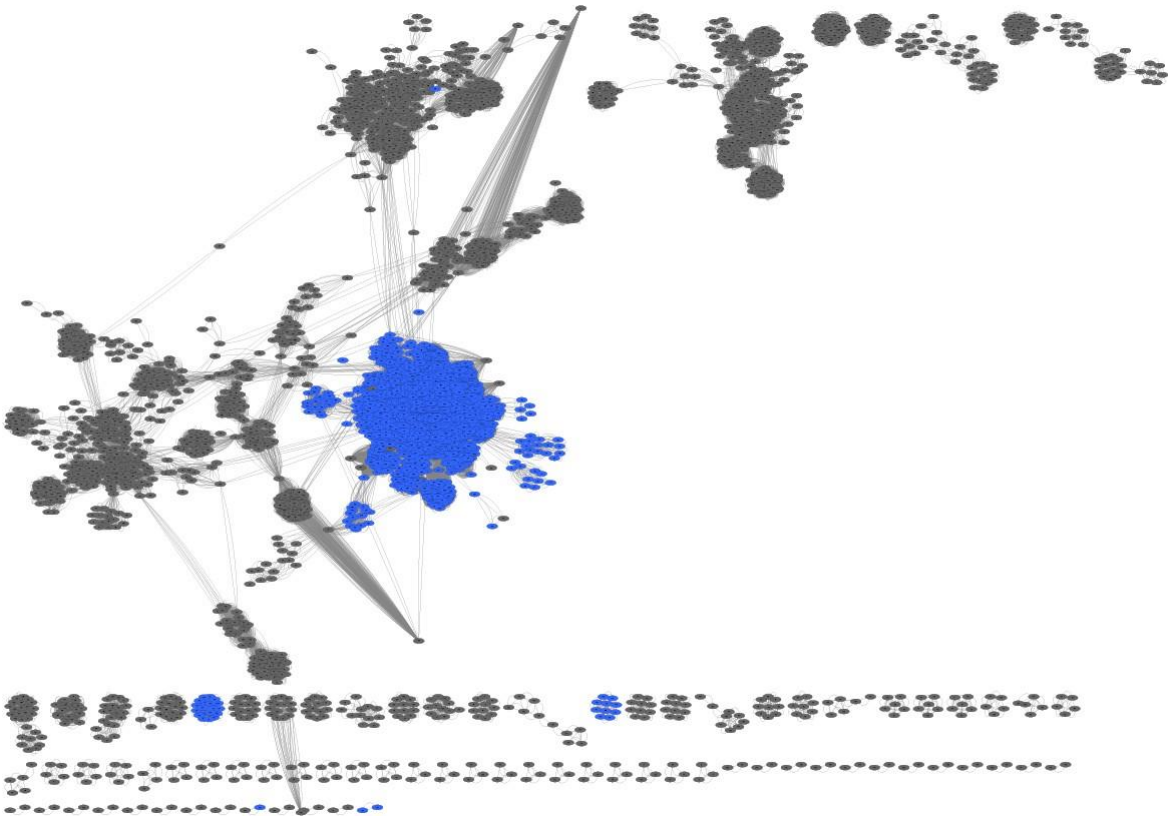

Figure 3.8: Overview of gene-sharing network analysis of curated *Bacteroides* phage dataset
The network was produced using vConTACT with 2,538 reference phage (grey nodes) and 2,639 *Bacteroides* phage from a curated dataset (blue nodes). The *Bacteroides* phage form a large cluster and appears to be connected other reference phage. Two groups of *Bacteroides* phage do not appear related to any other reference phage and form their own clusters (bottom of the figure). Additionally, several *Bacteroides* phage appear as singletons. Each circle (node)

represents one phageand each line (edge) represents shared protein content between the nodes. *B.fragilis* phage vB_BfrS_23 is located within the main cluster of *Bacteroides* phage.

Table 3-9: CheckV output of representative *Bacteroides* phage sequences and VC assignment

None of the genomes had any contamination, contained any proviruses nor generated any warnings, and with the exception of uvig_401720 (kmer frequency 1.65) had a kmer frequency of 1.

VC	Representative sequence	Contig length (nt)	Gene count	Viral genes	Host genes	Quality*		Completeness (%)	Method (confidence)
						Checkv	miuvig		
VC_101	uvig_375980	64376	79	11	2	C	H	100	DTR (high)
VC_102	IMGVR_UViG_315	46627	51	6	2	H	H	96.33	AAI-based (medium)
VC_103	IMGVR_UViG_429	39579	45	8	1	H	H	91.56	AAI-based (medium)
VC_104	uvig_193521	51803	63	7	0	C	H	100	DTR (high)
VC_106	IMGVR_UViG_18	111507	145	12	3	H	H	100	AAI-based (high)
VC_107	uvig_462209	50302	62	9	3	H	H	100	AAI-based (high)
VC_108	uvig_129546	44531	63	8	1	C	H	100	DTR (high)
VC_109	uvig_284048	90215	139	17	1	C	H	100	DTR (high)
VC_110	uvig_569056	66120	84	2	12	H	H	100	AAI-based (high)
VC_117	uvig_155226	86805	119	25	6	C	H	100	DTR (high)
VC_206	MT074136.1	99494	110	17	0	C	H	100	DTR (high)
VC_210	uvig_23839	107341	133	19	3	C	H	100	DTR (high)
VC_220	ivig_2445	31092	46	12	1	M	GF	80.07	AAI-based (medium)
VC_223	uvig_189095	33413	45	12	0	C	H	100	DTR (high)
VC_228	IMGVR_UViG_487	50890	59	13	0	H	H	98.92	AAI-based (high)
VC_229	uvig_525887	43150	57	9	0	C	H	100	DTR (high)
VC_230	uvig_252231	42585	57	9	0	C	H	100	DTR (high)
VC_231	uvig_424999	62472	81	10	0	C	H	100	DTR (high)
VC_232	IMGVR_UViG_317	32355	48	14	1	H	H	99.66	AAI-based (high)
VC_233	uvig_265347	44344	56	8	1	H	H	100	AAI-based (medium)
VC_234	uvig_126463	79809	121	16	1	C	H	100	DTR (high)
VC_235	IMGVR_UViG_389	57252	83	15	3	H	H	98.62	AAI-based (high)
VC_236	uvig_335737	39151	58	11	0	C	H	100	DTR (high)
VC_238	uvig_571635	95951	144	16	0	C	H	100	DTR (high)
VC_239	uvig_169571	84073	107	14	1	C	H	100	DTR (high)
VC_241	MT635598.1	40421	51	18	1	H	H	100	AAI-based (high)
VC_242	uvig_332742	30536	41	12	1	M	GF	75.33	AAI-based (high)
VC_258	IMGVR_UViG_324	98324	172	26	1	H	H	99.89	AAI-based (high)
VC_259	uvig_172870	60853	95	20	0	C	H	100	DTR (high)
VC_260	uvig_280224	100229	161	15	1	C	H	100	DTR (high)
VC_261	uvig_234487	100259	168	23	1	C	H	100	DTR (high)
VC_263	uvig_178134	102165	160	32	1	C	H	100	DTR (high)
VC_264	uvig_208702	95669	162	19	1	C	H	100	DTR (high)
VC_266	uvig_377659	97009	85	10	0	H	H	100	AAI-based (high)
VC_267	uvig_34710	96199	83	6	0	C	H	100	DTR (high)
VC_268	Sib1_ms_5	92132	84	10	0	H	H	98.89	AAI-based (high)
VC_269	IMGVR_UViG_718	43412	59	7	1	M	GF	74.73	AAI-based (high)
VC_333	uvig_425355	52660	64	11	0	C	H	100	ITR (high)
VC_334	uvig_51867	55171	89	19	1	C	H	100	DTR (high)
VC_336	IMGVR_UViG_699	109520	132	11	6	H	H	98.8	AAI-based (medium)
VC_337	uvig_505175	28537	43	12	0	M	GF	75.46	AAI-based (medium)
VC_338	uvig_177968	77878	134	19	3	H	H	100	AAI-based (high)
VC_340	uvig_80643	74444	115	22	3	H	H	100	AAI-based (high)
VC_341	uvig_55388	44163	68	5	0	H	H	100	AAI-based (high)

VC	Representative sequence	Contig length (nt)	Gene count	Viral genes	Host genes	Quality*		Completeness (%)	Method (confidence)
						Checkv	miuvig		
VC_342	uvig_287841	45428	57	6	2	H	H	100	AAI-based (medium)
VC_343	uvig_235031	56911	73	6	2	C	H	100	ITR (high)
VC_344	IMGVR_UViG_701	40824	52	3	1	H	H	96.76	AAI-based (high)
VC_345	uvig_327558	63240	86	16	2	C	H	100	ITR (high)
VC_347	uvig_418377	70971	101	18	3	C	H	100	DTR (high)
VC_348	uvig_590419	39847	62	13	0	C	H	100	DTR (high)
VC_349	uvig_193089	108436	142	18	3	C	H	100	DTR (high)
VC_350	IMGVR_UViG_823	34760	41	7	3	M	GF	86.17	AAI-based (medium)
VC_351	uvig_140333	64522	84	9	11	H	H	100	AAI-based (medium)
VC_352	IMGVR_UViG_698	38660	55	11	1	M	GF	79.87	AAI-based (high)
VC_353	IMGVR_UViG_752	40922	51	7	2	H	H	92.35	AAI-based (high)
VC_354	uvig_5976	64264	76	8	6	H	H	100	AAI-based (high)
VC_355	IMGVR_UViG_524	64430	75	14	1	C	H	100	DTR (high)
VC_356	IMGVR_UViG_624	61730	93	14	2	H	H	100	AAI-based (high)
VC_357	uvig_424998	46277	67	11	0	C	H	100	DTR (high)
VC_358	uvig_510143	46064	52	6	1	C	H	100	DTR (high)
VC_359	IMGVR_UViG_298	52473	67	11	2	H	H	95.19	AAI-based (high)
VC_361	uvig_264521	86314	113	13	2	C	H	100	DTR (high)
VC_362	uvig_540493	57930	86	18	1	C	H	100	DTR (high)
VC_363	uvig_254157	109113	140	24	4	C	H	100	DTR (high)
VC_364	uvig_235484	35827	41	10	2	H	H	100	AAI-based (high)
VC_366	uvig_199655	80858	133	21	0	C	H	100	DTR (high)
VC_367	uvig_445349	54959	64	15	1	H	H	100	AAI-based (medium)
VC_368	IMGVR_UViG_771	49987	64	13	1	H	H	100	AAI-based (high)
VC_370	uvig_309912	56499	75	14	2	H	H	100	AAI-based (medium)
VC_371	uvig_10477	62346	101	17	0	C	H	100	DTR (high)
VC_372	uvig_101329	55218	70	12	0	M	GF	75.95	AAI-based (high)
VC_373	uvig_71647	54317	84	12	0	M	GF	88.02	AAI-based (high)
VC_374	IMGVR_UViG_784	57973	80	13	2	H	H	99.32	AAI-based (high)
VC_375	IMGVR_UViG_35	54816	74	13	3	H	H	96.79	AAI-based (high)
VC_376	IMGVR_UViG_247	52043	75	12	3	M	GF	87.33	AAI-based (medium)
VC_378	uvig_425872	60638	89	18	2	C	H	100	DTR (high)
VC_379	uvig_63537	35052	50	3	0	C	H	100	DTR (high)
VC_381	IMGVR_UViG_792	36459	55	12	1	M	GF	87.3	AAI-based (medium)
VC_382	uvig_355263	40059	65	8	3	H	H	100	AAI-based (high)
VC_383	uvig_392724	75262	118	16	2	H	H	93.47	AAI-based (high)
VC_384	uvig_510021	106008	125	19	3	C	H	100	DTR (high)
VC_385	IMGVR_UViG_38	6206	8	3	0	H	H	100	AAI-based (high)
VC_386	IMGVR_UViG_494	36720	53	2	3	H	H	96.26	AAI-based (medium)
VC_387	uvig_438138	37126	45	11	0	C	H	100	DTR (high)
VC_388	uvig_242970	183808	225	29	3	C	H	100	DTR (high)
VC_389	IMGVR_UViG_676	32513	80	34	0	H	H	98.49	AAI-based (high)
VC_390	uvig_54817	57865	102	20	3	C	H	100	DTR (high)
VC_391	uvig_179206	54864	79	14	3	C	H	100	DTR (high)
VC_392	IMGVR_UViG_7	15205	30	10	0	H	H	100	AAI-based (high)
VC_393	IMGVR_UViG_392	59449	70	10	5	M	GF	82.92	HMM-based (lower)
VC_394	MT074142.1	179161	257	21	4	H	H	100	AAI-based (medium)
VC_395	uvig_105953	100540	124	21	1	C	H	100	DTR (high)
VC_396	IMGVR_UViG_26	400107	533	35	10	H	H	100	AAI-based (high)
VC_397	IMGVR_UViG_342	103466	143	9	2	H	H	94.97	AAI-based (high)

VC	Representative sequence	Contig length (nt)	Gene count	Viral genes	Host genes	Quality*		Completeness (%)	Method (confidence)
						Checkv	miuvig		
VC_398	IMGVR_UViG_740	37064	66	21	0	H	H	100	AAI-based (high)
VC_399	uvig_493028	35908	53	10	0	M	GF	78.49	AAI-based (high)
VC_400	IMGVR_UViG_650	186253	221	23	7	H	H	100	AAI-based (high)
VC_411	uvig_401720	11557	16	6	0	H	H	100	AAI-based (high)
VC_465	uvig_257457	36728	46	11	0	C	H	100	DTR (high)

*C, complete; GF, genome fragment; H, high; M, medium.

Table 3-10: CheckV quality control summary of VC_396

None of the genomes had any contamination, contained any proviruses nor generated any warnings, and all had a kmer frequency of 1. Completeness for all was assessed using an AAI-based method (all had high confidence).

Sequence ID	Contig length (nt)	Gene count	Viral genes	Host genes	Quality*		Completeness (%)
					Checkv	miuvig	
IMGVR_UViG_23	399500	531	35	11	H	H	100
IMGVR_UViG_26	400107	533	35	10	H	H	100
IMGVR_UViG_300	365782	499	29	8	H	H	92.51
IMGVR_UViG_358	375433	498	33	9	H	H	94.96
IMGVR_UViG_384	289806	391	25	8	M	GF	73.3
IMGVR_UViG_422	394073	530	33	9	H	H	99.65
IMGVR_UViG_447	289877	391	25	3	M	GF	73.29
IMGVR_UViG_533	335415	452	30	9	M	GF	84.86
IMGVR_UViG_566	382539	503	32	8	H	H	96.76

*GF, genome fragment; H, high; M, medium.

No sequence similarity for any phage in VC_396 was found suggesting potentially previously unrecognised jumbo phage within this *Bacteroides* phage dataset. Only sequences from VC_100 were included in further analyses as this VC contains the four *B. fragilis* phage sequences of interest. VC_100 contained 30 high-quality complete genomes and 12 high-quality genomes with completeness < 100 % (91.75 - 99.41 %) and two medium-quality genome fragments (73.85 - 89.73 %) (Table 3.11).

Following the selection of representative sequences, it was necessary to obtain related reference sequences. This was necessary to determine the relatedness of uncharacterised *Bacteroides* phage to currently recognised phage. Additionally, it is possible to explore the evolutionary relationship of characterised reference phage and *Bacteroides* phage within our dataset. Fourteen sequences were found to share the same clade as the representative sequences following analysis with VipTree server (v.1.9) (Table 3.12)[27].

A proteomic phylogenetic tree was constructed to explore the diversity of *Bacteroides* phage and relatedness to reference phage and crAss-like phage (Figure 3.9). It was discovered that there were two clades comprising 45 phage, with 11 phage in a distantly related clade and the remaining 34 phage appearing more closely related. All phage within VC_100 appeared within these two clades. These two clades were determined to be a potentially novel family when clustering the protein orthologous sequences. A representative sequence from VC_358 sat within the larger subclade that contained the VC_100 sequences. This will be explored further in Section 3.3.5.

Table 3-11: CheckV quality control summary of VC_100

None of the genomes had any contamination nor contained any proviruses. All had a kmer frequency of 1. None encoded host genes, with the exception of uvig_314311 (three host genes). Only three sequences had warnings: uvig_31439, low-confidence DTR; uvig_314311 and uvig_422350, both comprise single contigs >1.5x longer than the expected genome length.

Sequence ID	Contig length (nt)	Gene count	Viral genes	Quality*		Completeness (%)	Completeness method (confidence)
				CheckV	miuvig		
uvig_31439	34228	47	12	M	GF	73.85	AAI-based (high)
uvig_364892	41602	67	13	M	GF	89.73	AAI-based (high)
uvig_90520	42524	67	14	H	H	91.75	AAI-based (high)
uvig_266181	43775	66	15	H	H	94.44	AAI-based (high)
IMGVR_UViG_736	44670	60	13	H	H	96.4	AAI-based (high)
B40-8	44929	61	17	H	H	97.04	AAI-based (high)
IMGVR_UViG_653	44988	62	13	H	H	97.07	AAI-based (high)
uvig_294204	45329	66	19	H	H	97.93	AAI-based (high)
uvig_293893	45352	65	19	H	H	97.98	AAI-based (high)
uvig_296087	45625	62	14	H	H	98.45	AAI-based (high)
uvig_465436	45833	65	17	H	H	99	AAI-based (high)
uvig_285949	45857	67	15	H	H	99.04	AAI-based (high)
Barc2635	45990	67	14	H	H	99.22	AAI-based (high)
uvig_227632	46023	69	18	H	H	99.41	AAI-based (high)
B124-14	47159	66	15	H	H	100	AAI-based (high)
IMGVR_UViG_737	47321	68	15	H	H	100	AAI-based (high)
uvig_110769	47569	65	16	C	H	100	DTR (high)
uvig_124569	47576	66	15	H	H	100	AAI-based (high)
uvig_175686	47376	68	16	C	H	100	DTR (high)
uvig_176975	44725	60	12	C	H	100	DTR (high)
uvig_188088	45325	62	16	C	H	100	DTR (high)
uvig_233765	46609	65	18	C	H	100	DTR (high)
uvig_237530	47339	72	15	H	H	100	AAI-based (high)
uvig_259966	43985	59	14	C	H	100	DTR (high)
uvig_264822	45755	66	17	C	H	100	DTR (high)
uvig_265317	45755	65	16	C	H	100	DTR (high)
uvig_267541	45663	63	15	C	H	100	DTR (high)
uvig_268800	46396	67	16	C	H	100	DTR (high)
uvig_272641	46397	67	14	C	H	100	DTR (high)
uvig_274313	45663	63	15	C	H	100	DTR (high)
uvig_274976	45663	64	15	C	H	100	DTR (high)
uvig_291773	46571	65	16	C	H	100	DTR (high)
uvig_297825	46529	64	18	C	H	100	DTR (high)
uvig_314311	113009	143	22	H	H	100	AAI-based (high)
uvig_319905	47135	67	19	C	H	100	DTR (high)
uvig_320042	45359	62	17	C	H	100	DTR (high)
uvig_332402	47447	69	18	H	H	100	AAI-based (high)
uvig_422023	47247	65	15	C	H	100	DTR (high)
uvig_422350	77994	107	27	H	H	100	AAI-based (high)
uvig_543730	44455	65	13	C	H	100	DTR (high)

Sequence ID	Contig length (nt)	Gene count	Viral genes	Quality*		Completeness (%)	Completeness method (confidence)
				CheckV	miuvig		
uvig_82289	47448	66	14	H	H	100	AAI-based (high)
uvig_82951	46031	69	12	C	H	100	DTR (high)
uvig_93296	47348	67	14	H	H	100	AAI-based (high)
vB_BfrS_23	48011	70	15	H	H	100	AAI-based (high)

*C, complete; GF, genome fragment; H, high; M, medium.

Table 3-12: NCBI reference phage used in proteomic phylogenetic tree

Phage	Accession	In VC?
Bacillus phage BCD7	NC_019515	-*
Brevibacillus phage Emery	KC595516	-
Brucella phage BipB01	NC_031264	-
Cellulophaga phage phi39:1	NC_021804	-
Clostridium phage vB_CpeS_CP51	NC_021325	220
Croceibacter phage P2559S	NC_018276	-
Croceibacter phage P2559Y	NC_023614	-
Flavobacterium phage 1H	NC_031911	-
Flavobacterium phage 2A	NC_031926	-
Flavobacterium phage 6H	NC_021867	-
Lactococcus phage 1358	NC_027120	-
Lactococcus phage P335 *sensu lato*	NC_004746.1	358
Lepus americanus faeces-associated microvirus SHP1 6472	NC_040341	411
Riemerella phage RAP44	NC_019490	-

*-, Not present in VC as assessed by vConTACT.

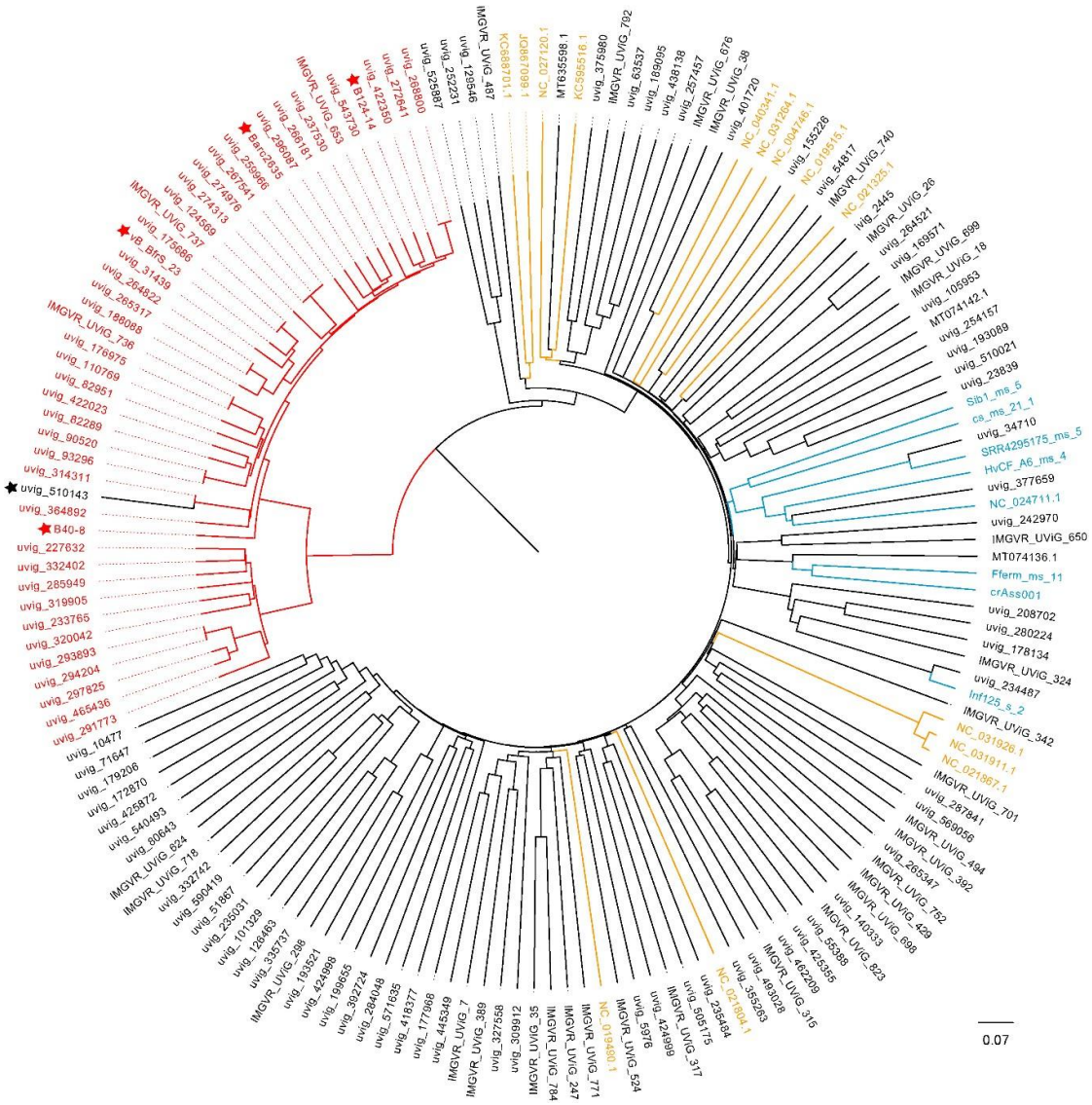

Figure 3.9:Proteomic phylogenetic tree of *Bacteroides* representative phage, reference phage, representative crAss-like phage and the VC (VC_100) of interest

The tree was generated from 164 phage sequences using VipTreeGen[27] and visualised in FigTree. Yellow, reference phage;blue, crAss-like phage; red, VC_100 phage. Red stars, the four cultured *B. fragilis* phage of interest, including vB_BfrS_23. Black star, uvig_510143 – an additional phage of interest. The treewas rooted at the midpoint. Scale bar, average number of amino acid substitutions per position. Genomes with uvig prefix are from the GPD; genomes with IMGVR prefix are from IMG/VR database.

Every other *Bacteroides* phage used to construct the tree was deemed distantly related to the VC_100-specific clades. The four *B. fragilis* phage, although within the more closely related cluster, were not paired within a subclade. Additionally, the branch lengths suggested each VC within the phylogenetic tree was distantly related to each other. The reference phage were dispersed throughout the phylogenetic tree, with *Flavobacterium* phage 2A/6H/1H forming a closely related subclade. Additionally, *Lepus americanus* faeces-associated microvirus SHP1 6472 was placed with a subclade with representative phage from its shared cluster (VC_411, uvig_401720). This was untrue for the other phage clustered within VC with *Bacteroides* phage; *Clostridium* phage vB_CpeS_CP51 (VC 220) and *Lactococcus* phage P335 *sensu lacto* (VC_358). The crAss-like phage appeared to be distributed between two subclades, highlighting the diversity of crAss-like phage. The phylogenetic tree addresses the question of how related isolated *B. fragilis* phage are to cultured and uncultured *Bacteroides* phage and shows a high level of diversity within the *Bacteroides* phage dataset.

Unsurprisingly, no protein orthologues were shared across the entire representative sequence and VC_100 dataset. A total of 1,315 orthogroups were defined using OrthoFinder (v.2.2.6) and the number of sequences in each orthogroup ranged from 2 to 136. Orthogroup 1 contained 136 sequences and was predicted to encode the phage anti-repressor protein. Orthogroup 2 contained 112 sequences and was predicted to encode the DNA segregation/tail tape protein. Additionally, these results suggest that publicly available, including vB_BfrS_23, and metagenome-assembled *Bacteroides* phage represent a diverse and uncharacterised taxonomic group with no known closely related reference sequences.

3.3.6 Analysis of novel *Bacteroides* phage taxonomic group

3.3.6.1 VC_100 represents a novel B. fragilis phage family

Generation of the phylogenetic tree (Figure 3.9) highlighted an unexplored *Bacteroides* phage taxonomic group (VC_100) that contains the *B. fragilis* phage of interest. It was necessary to determine if the phage within VC_100 formed a family with any known phage. VipTree identified three reference phage with similarity to B40-8 and ϕB124-14: *Flavobacterium* sp. phage 1/32 (genome accession KJ018210), *Croceibacter* phage P25559Y (NC_023614) and *Croceibacter* phage P2559S (NC_018276)[27]. The intergenomic similarity between VC_100 and the reference phage was < 1% and the reference phage were not within the same family as VC_100 phage. To determine the relatedness between the cluster members, nucleotide-based intergenomic similarity analysis and hierarchical clustering were undertaken using VIRIDIC (Figure 3.10)[33].

This approach defined VC_100 as a clear family (> 27.4 % intergenomic similarity across all pairwise comparisons), showed five distinct genera (1-5) and 37 species. The largest genus (Genus 1) contains 30 sequences and separated into 24 species. This genus also contains all four cultured *B. fragilis* phage (B40-8, VB_BfrS_23, ϕB124-14 and Barc2635). As confirmed previously (Section 3.3.4), B40-8 appears to be the least similar phage to VB_BfrS_23, ϕB124-14 and Barc2635 with similarity scores ranging from 76.9 to 77.2 %. Barc2635 and ϕB124-14 showed highest similarity to vB_BfrS23. Among all species within Genus 1, B40-8 showed highest similarity to uvig_259966 (77.9 %), Barc2635 to uvig_266181 (86.9 %), vB_BfrS_23 to IMGVR_UVIG_737 and uvig_175686 (90.3 %) and ϕB124-14 to uvig_543730 (84.6 %). Eleven phage determined to be the same five species as they shared 100 % intergenomic similarity (Figure 3.10).

Genus 2 contained 11 sequences and was separated into 10 defined species. This genus contains several highly related sequences, potentially grouping at subspecies or strain level (uvig_293893 vs uvig_294204: 99.9 % intergenomic similarity; uvig_465436 vs uvig_297825: 94.2 % intergenomic similarity). Genera 3, 4 and 5 contained only one species each and exhibited the lowest intergenomic similarity to all other genera. Genus 3 (containing uvig_314311) showed the least similarity with pairwise percentage with other phage, ranging from 25.7 to 48.5 %. A whole-genome proteomic phylogenetic tree was constructed (Figure 3.11) to confirm the taxonomic conclusions drawn from the VIRIDIC analysis (Figure 3.10).

Two distinct clades were observed in Figure 3.11 that agree with the genera described in Figure 3.10. The clade containing genus 2 is monophyletic, whereas genera 1, 3, 4 and 5 exist within the same clade. Additionally, B40-8 appears to separate from the main clade containing Genus 1 and can be explained by the lower intergenomic similarity scores. Interestingly, genera 3, 4 and 5 sit within the Genus 1 subclade, suggesting these sequences share a higher protein similarity to Genus 1 than nucleotide similarity generated previously. This further highlights the need to investigate nucleotide and protein similarity when generating phage taxonomy. Additionally, the longer branch length observed in Genus 2 suggests there is a higher rate of change among these phage than in the other genera.

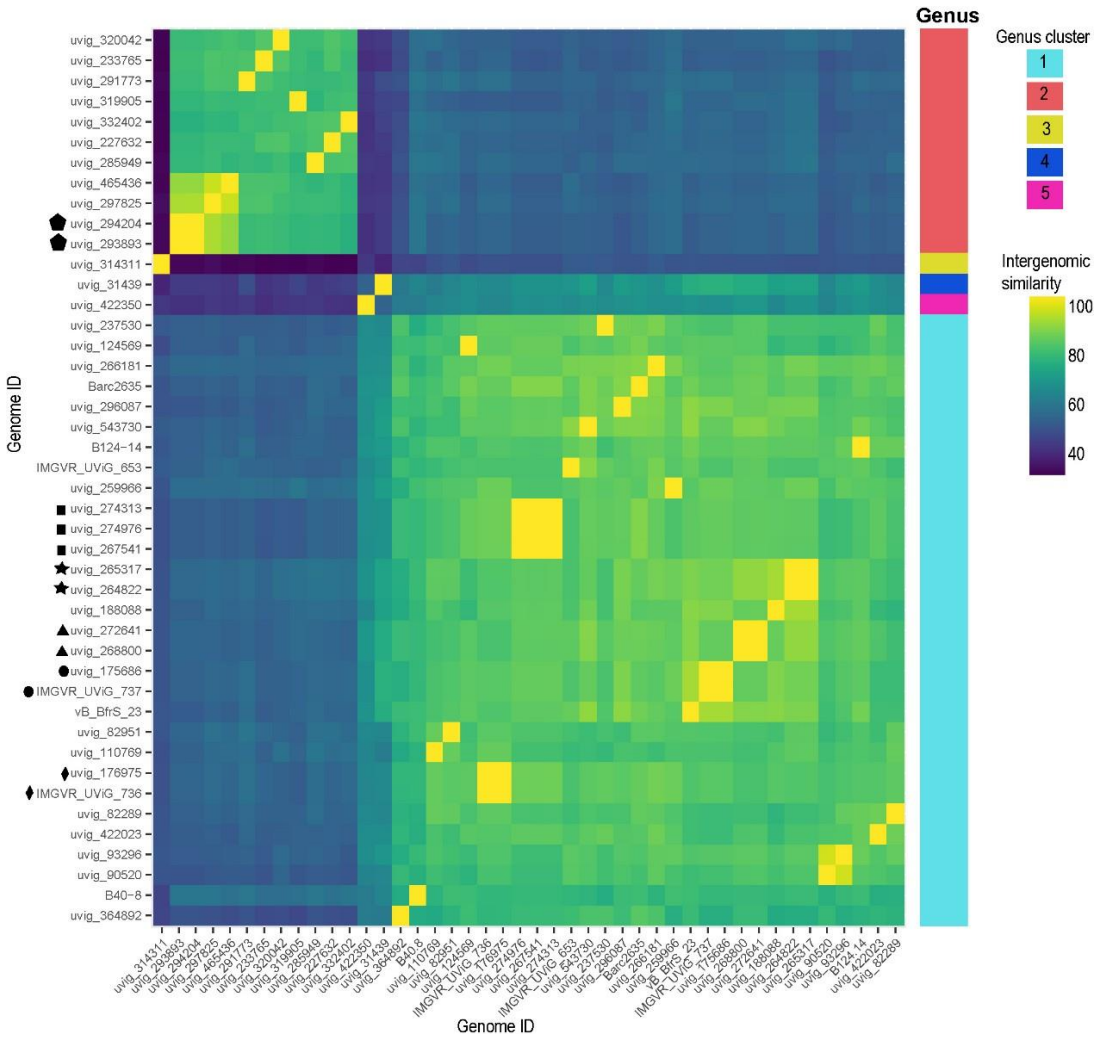

Figure 3.10: **Heatmap representing intergenomic similarity (%) within VC_100 and assigned genus**

Genomic similarities were generated using VIRIDIC and plotted in R using ggplot2[33]. Genus cluster assignment is represented by the coloured bar to the right of the plot. Six clusters of phage were found to be identical (genomic similarity of 100 %; shown in yellow) and highlighted on the plot by matching black shapes (diamond, circle, triangle, square and pentagon).

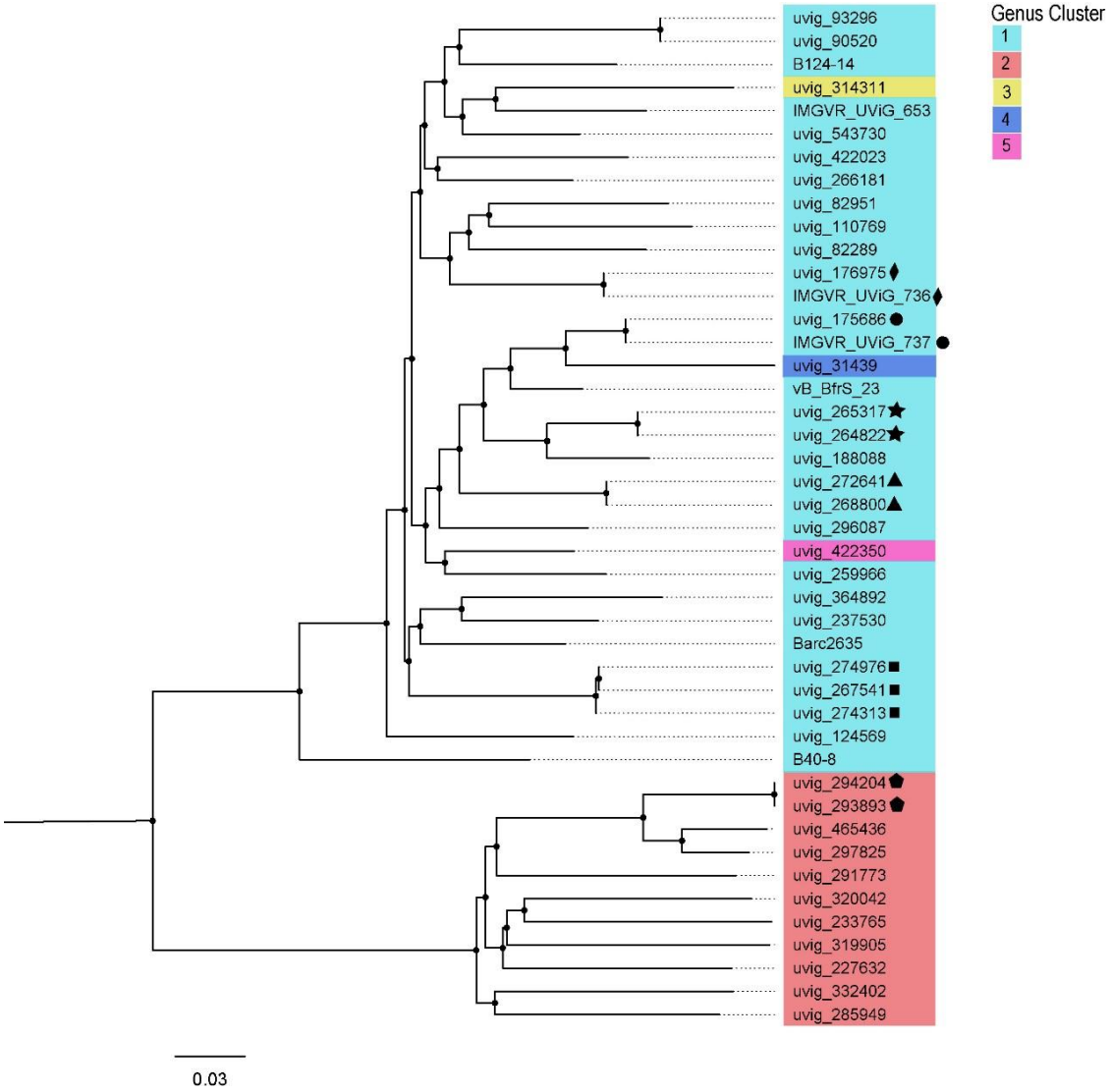

Figure 3.11: Proteomic phylogenetic tree generated from VC_100

VipTreeGen was used to produce a maximum likelihood phylogenetic tree from 44 phage sequences[27]. The assigned genus clusters are represented by coloured background and phage sharing identical genomic similarity are shown by matching black shapes. The tree was rooted at the midpoint and visualised in FigTree. Scale bar, average number of amino acid substitutions per position.

The metadata relating to these phage were explored to determine if global or health-related correlations could be inferred. Interestingly, Genus 2 phage originated only from human gut microbiota studies undertaken in China. Genus 1 displayed a global distribution from a variety of gut microbiota studies. Genera 3, 4 and 5 originated from infant gut studies, with uvig_422350 (Genus 5) appearing in the same subject at day 405 and 496 hinting at persistence within the human infant gut. Additionally, by searching the associated database (GPD or IMG/VR) metadata, it was revealed all uncultured phage within VC_100 had a predicted bacterial host of *B. fragilis*.

3.3.6.2 Several proteins universally conserved across family

A total of 107 orthogroups were identified within the family using OrthoFinder (v.2.2.6) with 95.5 % of all genes assigned to an orthogroup (Figure 3.12 and Figure 3.13)[115]. The percentage of shared orthologous proteins within the family ranged from 98.41 to 52.46 %. Genus 2 shared 81.25-98.36 % of protein orthologues and Genus 1 shared 75-98.41 % of protein orthologues (Figure 3.13). Twenty-one orthologues were conserved across the family, with the majority assigned as hypothetical proteins. However, phage anti-repressor (OG0000000), essential recombination protein (OG0000006), thymidylate synthase(OG0000009), ssDNA binding protein(OG0000018) and HNH endonuclease (OG0000019) were assigned (Table 3.13).

Structural proteins were not universally conserved across the family as uvig_31439 (Genus 4) was not assigned to large terminase subunit (OG0000037), major protein 1 (OG000032), major protein 2 (OG0000036), major protein 3 (OG0000035) or capsid-associated protein (OG0000033) orthogroups. Manual protein search of uvig_31439 revealed it does not possess recognisable structural proteins, except the tail fibre protein (OG0000013). Additionally, uvig_364892 (Genus 1) was not assigned to the tail fibre protein orthogroup. The majority of orthogroups contained one gene from the assigned sequence; however, OG0000000 (phage anti-repressor protein), OG0000001 and OG0000057 (Genus 2 specific phage anti-repressor protein) contained multiple proteins from the same genome.

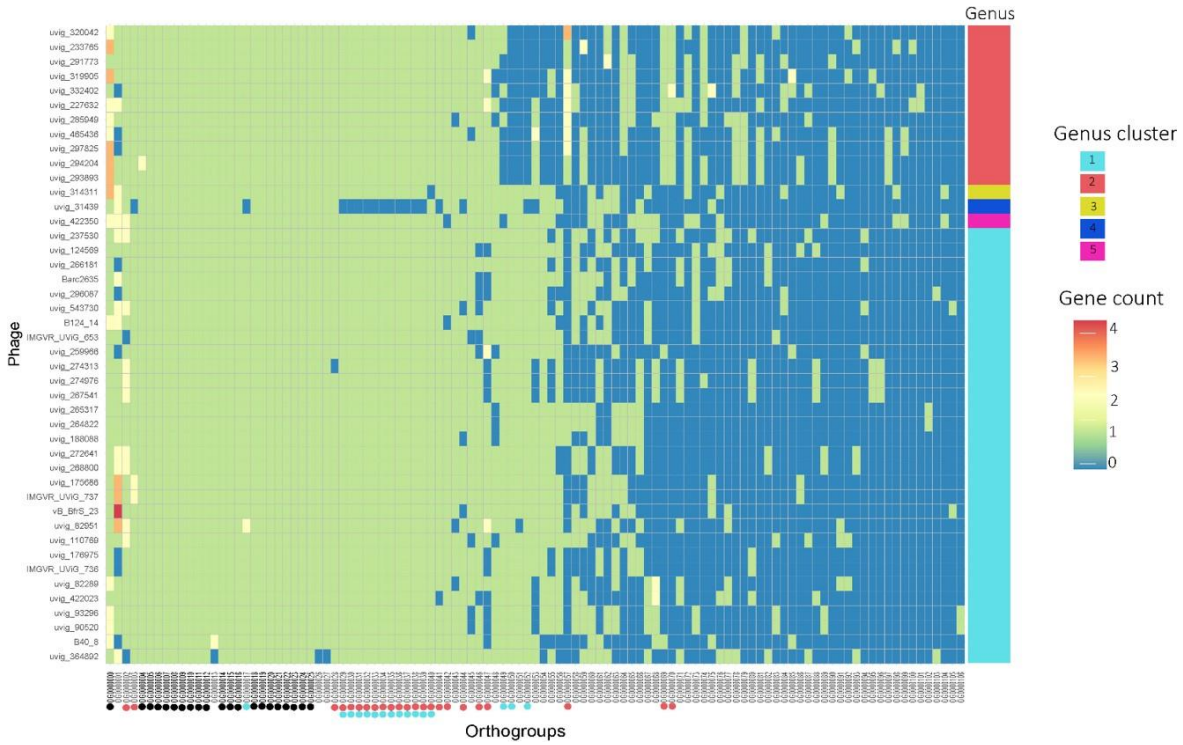

Figure 3.12: Heatmap representing phage assignment to orthogroups and genus specificity

Orthogroups were generated from 44 phage sequences from novel family VC_100 using OrthoFinder[115] and visualised in R using ggplot2. The number of genes (gene count) found in each orthogroup for each phage isrepresented by a coloured square. Genus cluster assignment represented by coloured at top right-hand side of the heatmap. The black circles along *x* axis show the orthogroups conservedacross family, blue circles show orthogroups conserved across Genus 1, and red circles show orthogroups conserved across Genus 2.

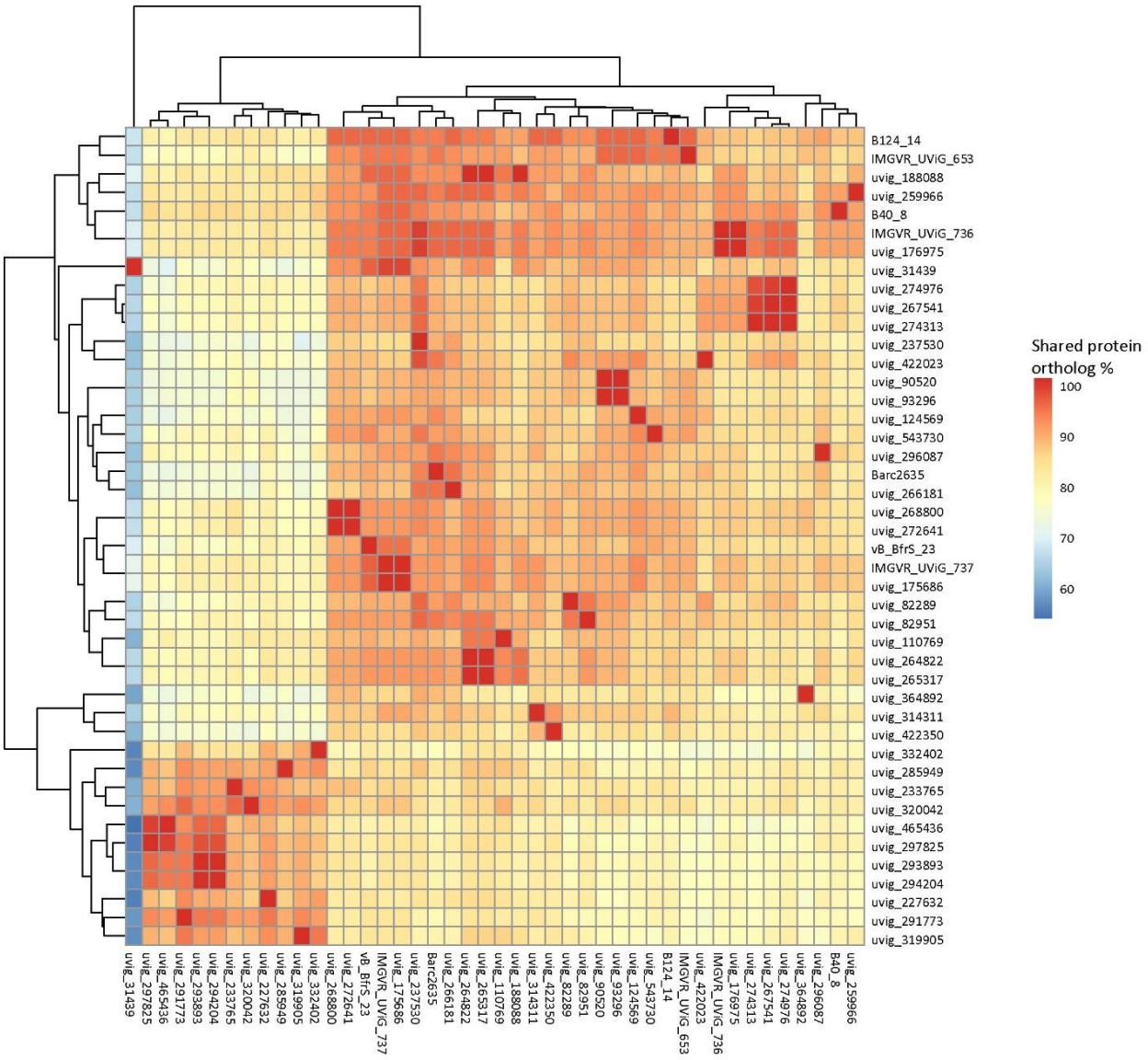

Figure 3.13: **Heatmap and dendrogram representing shared orthologues within VC_100**

Shared percentage was generated using OrthoFinder and plotted in R using ggplot2 and ggdendro. The percentage of shared protein orthologs are represented by the coloured squares and corresponds to the figure legend. The higher shared orthologues are represented by red squares and lowest by bluesquares. The dendrogram shows the separation of the phage genomes into Genus 1 and 2, as shown in previous figures.

Table 3-13: Orthogroup ID and predicted protein function

Orthogroups identified in VC_100 by OrthoFinder and protein function predicted by blastp (hits were considered significant for Blastp if the e-values were lower than $1e^{-5}$ at ≥40% protein identity)[99,115].

Orthogroup ID	Predicted Protein Function
OG0000000	Phage anti-repressor
OG0000002	Exoribonuclease
OG0000006	Essential recombination
OG0000009	Thymidylate synthase
OG0000013	Tail fibre
OG0000018	ssDNA binding
OG0000019	HNH endonuclease
OG0000028	DNA segregation/tail tape measure
OG0000032	MP1
OG0000033	Capsid-associated protein
OG0000035	MP3
OG0000036	MP2
OG0000037	Terminase large subunit
OG0000046	None known
OG0000057	Phage anti-repressor

Several orthogroups appear to be conserved across the specific genera and Genus 1 and Genus 2. Four orthogroups were conserved across Genus 1 and were classified as hypothetical proteins. Genus 2 contained 12 universally conserved orthologues and included exoribonuclease (OG0000002), tail fibre protein (OG0000013), DNA segregation/tail fibre protein (OG0000028), and an additional phage anti-repressor (OG0000057). The phage in Genus 2 appeared to possess a second phage anti-repressor protein that was absent from Genus 1. For example, a blastp search of the universally conserved phage anti-repressor orthologue in uvig_233765 (Genus 2) revealed 58 % sequence similarity (94 % query coverage) to phage anti-repressor KiLAC domain-containing protein from *B. intestinalis* (accession WP_118487259.1). A similar search with the phage anti-repressor present across all Genus 2 sequences showed 49.2 % sequence similarity (97 % query

coverage) to phage anti-repressor KiLAC domain containing protein from *Phocaeicola sartorii* (previously *Bacteroides sartorii*; accession WP_135951200.1). One orthogroup was determined to be genera-specific (Genus 2; OG000070) but no protein similarity was found.

3.3.6.3 Comparison to a VC_358 phage

As mentioned in Section 3.3.5, it was discovered a representative from VC_358 (uvig_510143) clustered within a VC_100 subclade in the phylogenetic tree (Figure 3.9). A proteomic phylogenetic tree constructed from all sequences in both VCs revealed two distinct monophyletic clades for each VC (Figure 3.14). Additionally, uvig_314311 (VC_100) appeared as an outgroup and shared a root with the two other clades, suggesting uvig_314311 does not share enough protein similarity to be assigned to either clade confidently. This highlights the need to properly investigate VCs assigned by using vConTACT (network-based analysis).

VC_358 contained 47 sequences and separated into two clades, with one clade appearing to share closer protein similarity. However, the branches on one sub-clade (containing uvig_549401, IMGVR_UViG_435, uvig_533947, uvig_3175 and uvig_371792) were longer than the surrounding sub-clades, suggesting a higher rate of substitution. To determine the shared genome regions between uvig_314311, VC_358 and VC_100, a genome comparison map was created. IMGVR_UViG_461 (VC_358) was selected for comparison due to the position to uvig_314311 in a proteomic phylogenetic tree (Table 3.14 and Figure 3.15).

Uvig_314311 shares multiple hallmark proteins for structure, regulation and replication, lysis and DNA packaging with ϕB124-14 (Figure 3.16). It shares less than half of its genome (uvig_314311: 113,009 bp) with either IMGVR_UViG_461 (42,225 bp) or ϕB124-14 (47,159 bp). Additionally, the tail fibre protein and large terminase subunit protein appear to be truncated in uvig_314311 compared to ϕB124-14. The second half of uvig_314311 genome appears to be dominated by bacterial replication and regulation genes. However, it overlaps with IMGVR_UViG_461 across phage-related tail tape measure protein. Both IMGVR_UViG_461 and uvig_314311 encoded DNA methyltransferase protein(s), suggesting the ability of these phage to resist bacterial host restriction endonucleases[129]. It is unclear if uvig_314311 is a true uncultured phage or an artefact from metagenome assembly which caused a hybrid between VC_100 phage and VC_358 phage (chimeric assembly). This highlights the need to accurately and carefully curate any metagenome-assembled phage genomes prior to drawing conclusions regarding their phylogeny.

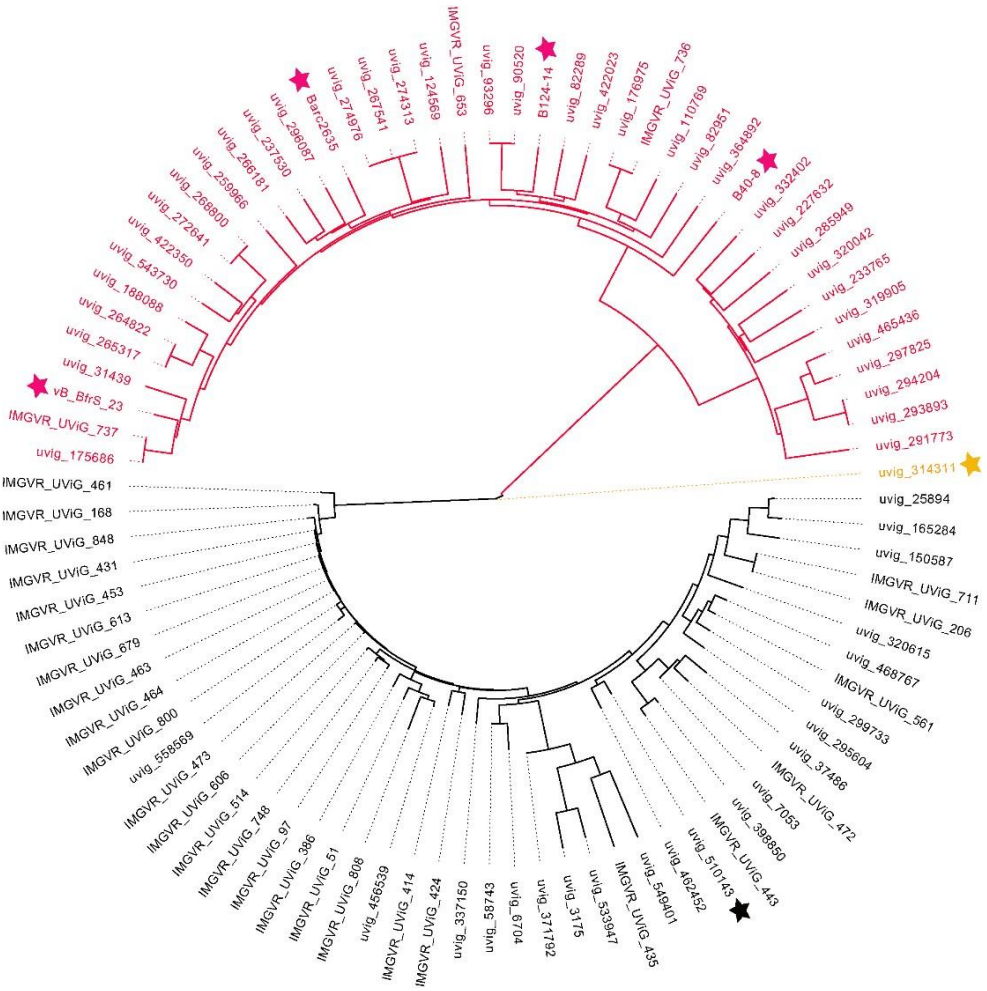

Figure 3.14: Proteomic phylogenetic tree generated from VC_100 (red) and VC_358 (black) phage

VipTreeGen was used to produce the maximum likelihood phylogenetic tree[27] and visualised in FigTree. Yellow representsthe phage with closest relation to both VCs. Red stars, cultured *B. fragilis* phage of interest. Blackstars, representative phage from VC_358. The tree was rooted at midpoint. Scale bar, average number of amino acid substitutions per residue.

Table 3-14: CheckV quality summary report of VC_358

None of the genomes had any contamination nor contained any proviruses. All had a kmer frequency of 1, with the exception of uvig_371792 (kmer frequency 1.96). Three sequences had warnings: uvig_150587, uvig_337150 and uvig_371792 all comprise single contigs >1.5x longer than the expected genome length; uvig_371792 also had a high kmer frequency that may indicate large duplication.

Contig ID	Contig length (nt)	Gene count	Virus genes	Host genes	Quality*		Completeness (%)	Completeness method (confidence)
					checkV	miuvig		
IMGVR_UViG_168	41199	46	4	1	M	GF	89.69	AAI-based (high)
IMGVR_UViG_206	45334	49	4	1	H	H	98.69	AAI-based (high)
IMGVR_UViG_386	41521	46	6	1	H	H	90.39	AAI-based (high)
IMGVR_UViG_414	44609	51	3	2	H	H	97.11	AAI-based (high)
IMGVR_UViG_424	34093	39	4	1	M	GF	74.32	AAI-based (high)
IMGVR_UViG_431	45112	50	4	2	H	H	98.2	AAI-based (high)
IMGVR_UViG_435	39820	65	8	1	M	GF	86.69	AAI-based (high)
IMGVR_UViG_443	46250	51	4	2	H	H	100	AAI-based (high)
IMGVR_UViG_453	45033	48	4	2	H	H	98.03	AAI-based (high)
IMGVR_UViG_461	42225	45	4	1	H	H	91.92	AAI-based (high)
IMGVR_UViG_463	43985	48	5	1	H	H	95.75	AAI-based (high)
IMGVR_UViG_464	43985	48	5	1	H	H	95.74	AAI-based (high)
IMGVR_UViG_472	32881	33	4	0	M	GF	71.58	AAI-based (high)
IMGVR_UViG_473	36379	38	4	1	M	GF	79.2	AAI-based (high)
IMGVR_UViG_514	45031	49	4	2	H	H	98.02	AAI-based (high)
IMGVR_UViG_51	42757	49	6	1	H	H	93.08	AAI-based (high)
IMGVR_UViG_561	46175	51	6	3	H	H	100	AAI-based (high)
IMGVR_UViG_606	41767	44	4	1	H	H	90.92	AAI-based (high)
IMGVR_UViG_613	37388	40	4	1	M	GF	81.39	AAI-based (high)
IMGVR_UViG_679	43988	48	4	1	H	H	95.75	AAI-based (high)
IMGVR_UViG_711	37177	39	4	1	M	GF	80.93	AAI-based (high)
IMGVR_UViG_748	42572	46	4	1	H	H	92.67	AAI-based (high)
IMGVR_UViG_800	44909	49	4	2	H	H	97.76	AAI-based (high)
IMGVR_UViG_808	37232	38	4	2	M	GF	81.05	AAI-based (high)
IMGVR_UViG_848	42989	46	4	1	H	H	93.58	AAI-based (high)
IMGVR_UViG_97	36414	38	4	1	M	GF	79.27	AAI-based (high)
uvig_150587	78580	92	14	1	H	H	100	AAI-based (high)
uvig_165284	41146	48	5	0	M	GF	89.61	AAI-based (high)
uvig_25894	45170	50	5	0	H	H	98.37	AAI-based (high)
uvig_295604	46589	51	4	3	C	H	100	DTR (high)
uvig_299733	44568	47	4	3	H	H	97.01	AAI-based (high)
uvig_3175	36756	59	6	0	M	GF	80.02	AAI-based (high)
uvig_320615	44974	49	4	2	H	H	97.9	AAI-based (high)
uvig_337150	94559	98	10	3	H	H	100	AAI-based (high)
uvig_371792	90363	97	8	4	H	H	100	AAI-based (high)
uvig_37486	40164	43	4	2	M	GF	87.42	AAI-based (high)
uvig_398850	46308	52	4	2	C	H	100	DTR (high)
uvig_456539	46261	51	6	2	H	H	100	AAI-based (high)

Contig ID	Contig length (nt)	Gene count	Virus genes	Host genes	Quality*		Completeness (%)	Completeness method (confidence)
					checkV	miuvig		
uvig_462452	45839	52	6	1	C	H	100	DTR (high)
uvig_468767	44781	48	4	3	H	H	97.47	AAI-based (high)
uvig_510143	46064	52	6	1	C	H	100	DTR (high)
uvig_533947	42169	64	8	1	H	H	91.8	AAI-based (high)
uvig_549401	36306	52	3	2	M	GF	79.03	AAI-based (high)
uvig_558569	43249	48	4	2	H	H	94.14	AAI-based (high)
uvig_58743	45912	49	4	2	H	H	99.94	AAI-based (high)
uvig_6704	38569	48	4	1	M	GF	83.95	AAI-based (high)
uvig_7053	45440	51	4	1	H	H	98.91	AAI-based (highs)

*C, complete; GF, genome fragment; H, high; M, medium.

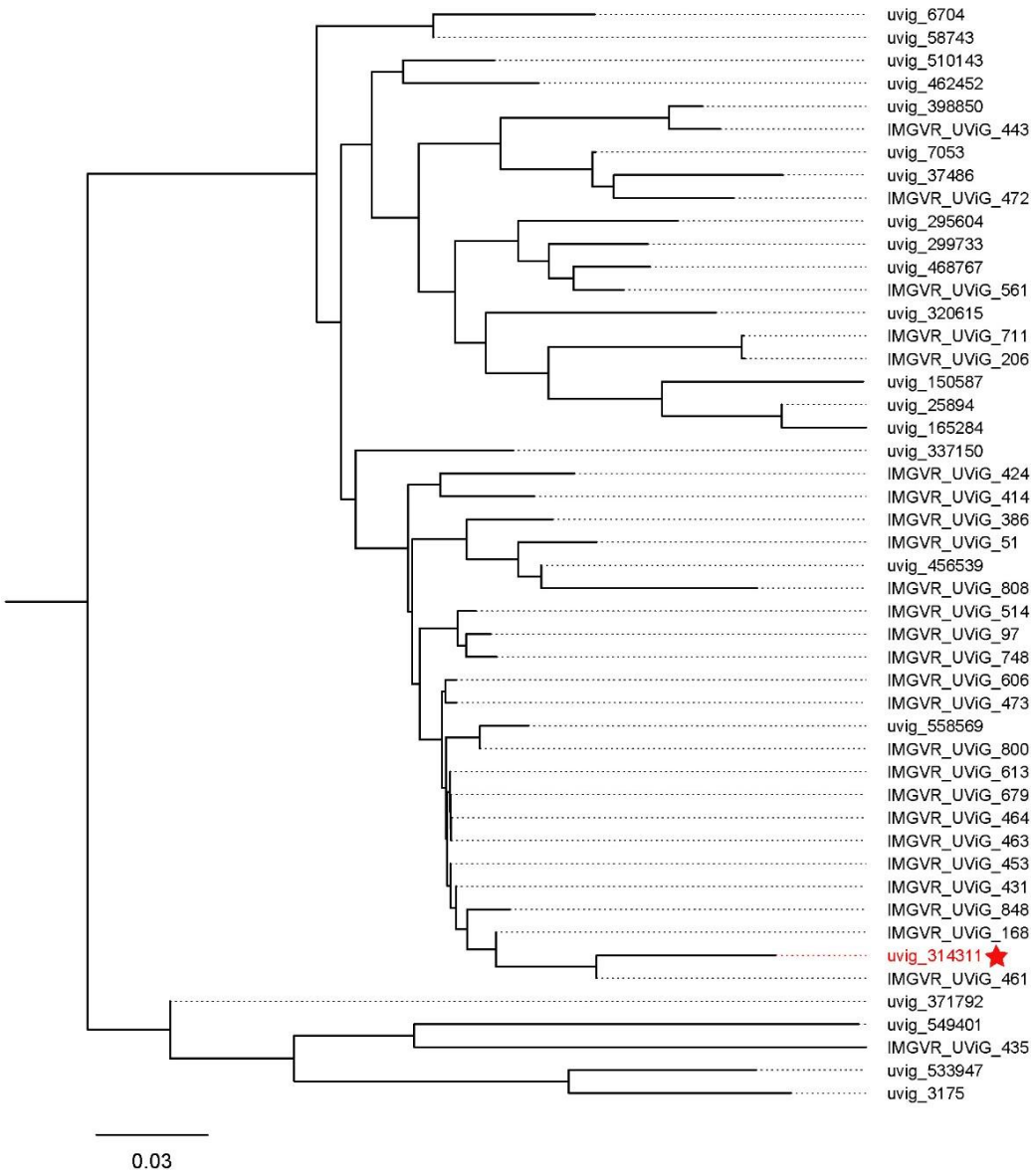

Figure 3.15: Proteomic phylogenetic tree generated from VC_358 and uvig_314311

VipTreeGen was used to produce a phylogenetic tree from uvig_314311 (VC_100) and VC_358 phage sequences[27]. The red star shows uvig_314311 from VC_100. The tree was rooted at the mid point and visualised in FigTree. Scale bar, average number of amino acid substitutions per residue.

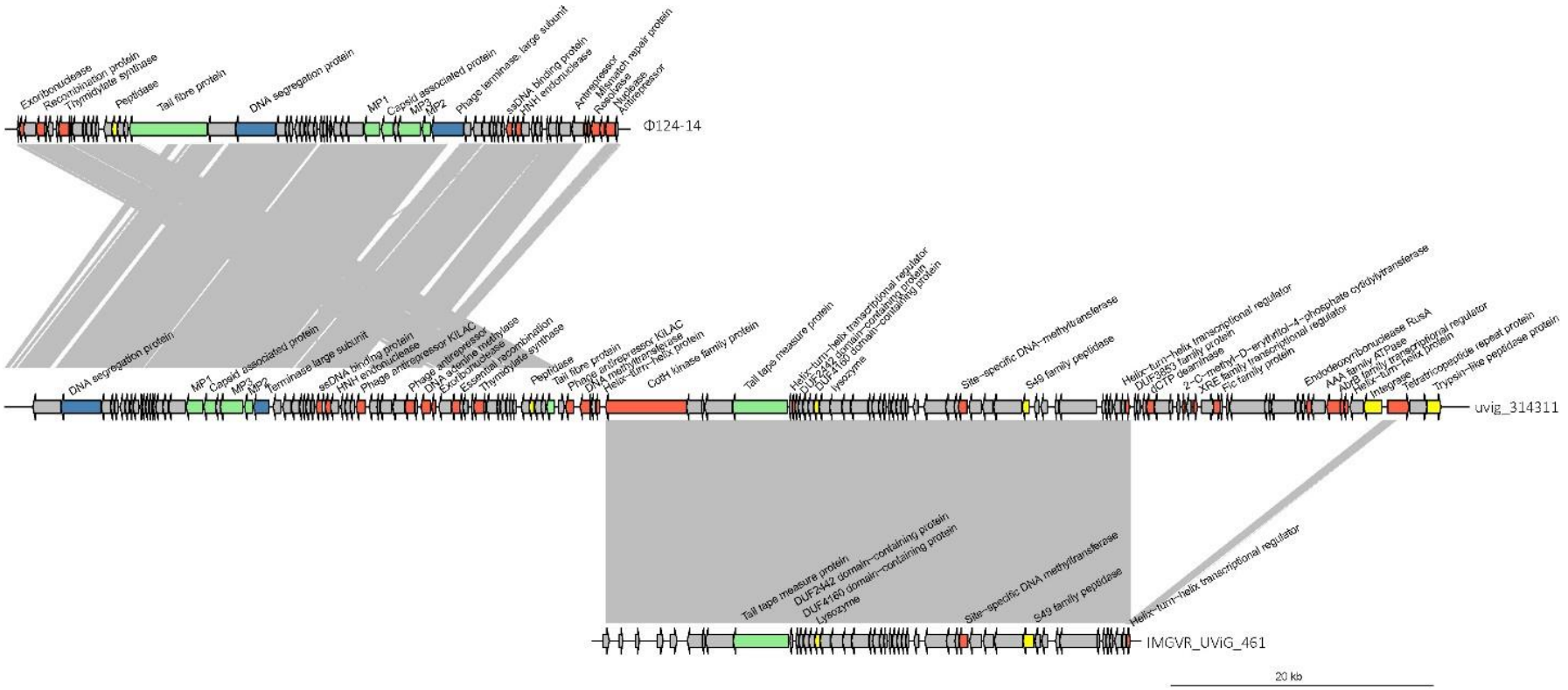

Figure 3.16: Genome comparison of phage ϕB124-14, uvig_314311 and IMGVR_UViG_461

Position and orientation of each predicted coding region shown for each genome. Colours of the arrows represent differing predicted protein function: red, replication and regulation; yellow, lysis; green, structure; blue, DNA packaging. Scale bar, genome size. Gray bars connecting phage genomes represent protein similarity according to blastp e-values. ϕB124-14 and uvig_314311 are phage from VC_100. IMGVR_UViG_461 is from VC_358.

3.3.6.4 Orthologous proteins to other Bacteroides phage

Eighteen orthogroups were discovered to contain all VC_100 and additional *Bacteroides* phage genomes when analysed with OrthoFinder (v. 2.2.6)[115]. The largest orthogroup contained 74 phage genomes and encoded for phage anti-repressor protein. Additional orthogroups ranged from containing 55 to 45 phage genome and included predicted thymidylate synthase, ssDNA binding protein, HNH endonuclease and essential recombination protein. Although not universally conserved across VC_100 (not present in uvig_364892), 32 other *Bacteroides* phage genomes contained the tail fibre protein orthologue. Similarly, 14 other *Bacteroides* phage genomes were present in the large terminase subunit orthogroup (not present in uvig_31439 from VC_100). Three phage (uvig_129546, uvig_252231, uvig_525887) from VC_108, VC_230 and VC_229, respectively, shared 12 protein orthologues with VC_100. Additionally, these phage were assigned to the same tail fibre protein, MP1, MP2, MP3 and capsid-associated protein orthogroups as the majority of VC_100 suggesting these VCs may represent a family closely related to VC_100.

3.3.6.5 Novel family TerL unrelated to crAss-like phage TerL

The crAss-like phage TerL gene is used to mine metagenome and virome datasets for novel crAss-like phage[71]. Additionally, TerL phylogeny was recently used in a ICTV classification proposal for defining crAss-like phage as *Crassvirales*[72]. Due to the apparent importance within the microbiota, global distribution of crAss-like phage and proposed *Bacteroides* host, it was important to determine if phage within VC_100 were related to crAss-like phage. The TerL protein sequence was extracted from all VC_100 sequences (except uvig_31439), 10 candidate genera crAss-like phage and isolated crAss001. The TerL amino acid sequences from *Cellulophaga* phage phi12:1 (NC_021791) and phi18:2 (KC821627) were used as outliers. The TerL phylogeny was inferred using IQTree and VT+F+G4 was determined best-fit model according to Bayesian information criterion[118,119]. The ideal tree was constructed after 200 iterations. The maximum likelihood tree revealed a distinct differentiation between crAss-like phage and VC_100 TerL protein, suggesting VC_100 phage are not related to crAss-like phage at family level or below (Figure 3.17). The TerL protein is highly conserved among VC_100 phage and displays genus-level specificity. Additionally, among crAss-like phage the TerL protein appeared to be loosely conserved. However, this is not surprising given that the crAss-like phage used in this phylogenetic tree were selected from 10 genera.

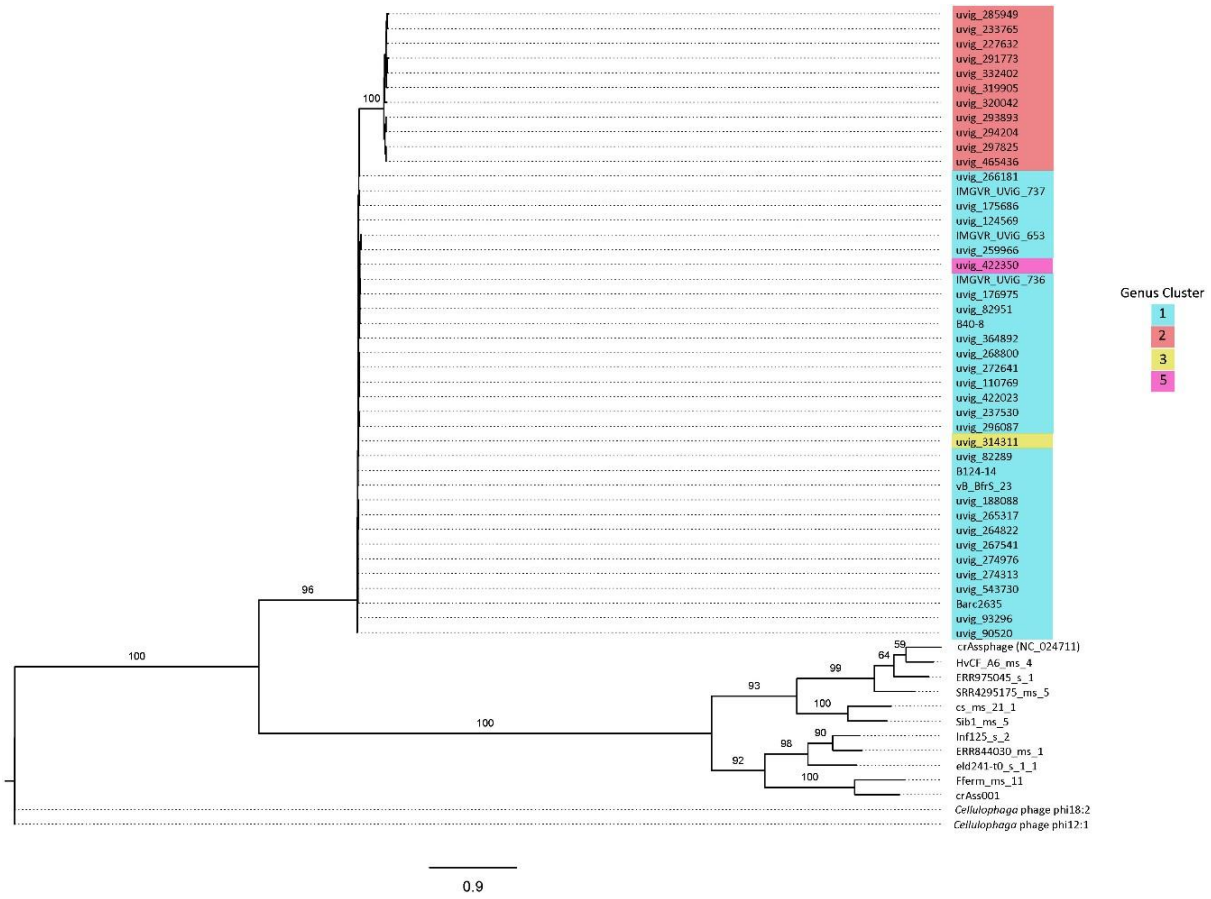

Figure 3.17: Phylogenetic tree of VC_100, crAss-like phage and two *Cellulophaga* phage TerL proteins

Amino acid sequences of large terminase protein from all phage within VC_100 was aligned with MUSCLE and the tree was produced using IQTree with 1000 bootstraps and ModelFinder[117-119]. The tree was visualised in FigTree. Bootstrap percentage shown on branches. Genus cluster assignment for VC_100 shown as background colour. The treewas rooted at the outlying *Cellulophaga* phage phi18:2. *Cellulophaga* phage phi18:2 and phi12:1 were used as outlier groups. Scale bar, amino acid substitutions.

3.4 Discussion

This Chapter presents the isolation and characterisation of a novel phage (vB_BfrS_23) isolated with *B. fragilis* GB-124 from wastewater effluent. Phage vB_BfrS_23 was described in the context of known *B. fragilis* phage and within a wider *Bacteroides* phage dataset. This revealed a potential novel *B.fragilis* phage family comprising 44 phage and five genera (Figure 3.9 and Figure 3.10).

vB_BfrS_23 is a dsDNA phage of 48,011 bp, with a GC content of 38.6 % and encoding 73 putative CDSs. The majority of these CDSs had no known predicted protein function and the genome was closely related to three other *B. fragilis* phage (ϕB124-14, Barc2635 and B40-8) (Table 3.8 and Figure 3.4). Interestingly, phylogenetic trees constructed using large terminase subunit protein and tail fibre protein produced differing evolutionary relationship, based on closest relatives, highlighting the difficulties in determining true phage phylogeny. Additionally, the tail fibre phylogenetic tree suggested these phage may be closely related to unknown temperate *B. fragilis* phage (Figure 3.6 and Figure 3.7).

The thermal assay revealed interesting information regarding the phenotypic characteristics of vB_BfrS_23. The phage was stable at lower temperatures (4, 24, 30 and 37 ˚C); however, the number of plaques increased at 40 and 45 ˚C while the plaque size decreased (from 2 mm to 0.5 mm). TEM imaging showed a large number of phage aggregates connected by the tail fibres (Figure 3.3). A 1974 paper reported a similar *B. fragilis* phage characteristic with phage isolated from animal sera[130]. It is possible the aggregation of phage < 40 ˚C is due a structural attraction of the tail fibre that is resolved at higher temperatures or an artefact of the experimental procedure, such as type of media used and duration of vortexing[131-133]. These thermal assay results are consistent with previous studies with naturally occurring *B. fragilis* GB-124 phage[88,134].

The origin of *B. fragilis* GB-124 is unknown and it was isolated from wastewater effluent in south England. It does not contain an enterotoxin and is assumed to be human commensal gut bacterium[135]. Phage vB_BfrS_23 was shown to have a narrow host range when screened using eight other *B. fragilis* isolates; consistent with results seen with ϕB124-14 (Table 3.4)[101]. However, it should be noted that non-enterotoxigenic *B. fragilis* is an opportunistic pathogen and can cause anaerobic infection outside the intestinal lumen (e.g. appendicitis, soft tissue infection, bacteremia)[136,137]. The *B. fragilis* strains used for the host assay were mainly isolated from anaerobic infections. The relationship between phage and pathogenic *B. fragilis* host remains

relatively unexplored. Therefore, to determine the true host range of vB_BfrS_23 commensal *B. fragilis* strains should be used.

A *Bacteroides* phage dataset was manually created from NCBI Virus, IMGVR and GPD databases and used to explore the relatedness of *Bacteroides* phage to known *B. fragilis* phage[91,109]. A total of 100 VCs were discovered from the dataset using network-based program vConTACT and highlighted the vast diversity of undiscovered phage[36,37]. The classification of phage depends heavily on reference databases. vConTACT uses the NCBI RefSeq database, which currently contains 2,538 reference phage genomes. This database contains an over-representation of dsDNA phage and vConTACT is currently biased towards these. Due to the constraint of NCBI RefSeq database, it is necessary to use additional viral databases to obtain a true phylogenetic profile[24]. MilliardLab (http://millardlab.org/bioinformatics/bacteriophage-genomes/phage-genomes-nov2020/) recently supplemented the NCBI RefSeq database with additional phage genomes, resulting in a new database size of 7,527[138]. Additionally, vConTACT creates monopartite networks which lack information regarding gene connection to VCs[36,37]. A monopartite gene-sharing network predicts viral proteins, translates into proteins and clusters into Markov cluster-based protein families[139]. A pairwise protein cluster comparison is applied to determine protein profiles and represented in weighted graphs (nodes being viral genome and edges being shared proteins). The graph is described as monopartite as it only uses one type of node[140]. Bipartite gene networks display connections between two sets of nodes (i.e. genomes and protein families) and can be more accurate in determining genes shared between genomes[141,142]. Additionally, bipartite networks are better applied for detecting mosaic genomes than monopartite networks[26]. However, while vConTACT creates a monopartite gene-sharing network, the output can be visualised as a bipartite network[36,37].

vB_BfrS_23 was revealed to belong to a potential novel *B. fragilis* phage family consisting of five genera (Figure 3.9 and Figure 3.10). It was not possible to confidently classify these phage as a novel family as exemplar species from other bacteriophage families were used for comparison. However, given the high intergenomic similarity and grouping of the phage within a viral cluster without a known reference phage, it is highly likely that these phage form a novel family. Genus 1 was the largest genus and contains the four known *B. fragilis*phage. Genera 3, 4 and 5 contained one phage genome each and were closely related to Genus 1.However, genomes uvig_31439, uvig_314311 and uvig_422350 were of questionable quality (either genome fragment or longer than average contig). uvig_314311 encodes regions with homology to VC_100 and VC_358; suggesting it is an artefact of metagenome assembly.
Therefore, genera 3, 4 and 5 cannot confidently be assigned without further investigation and

should not be included in proposal of a novel *B. fragilis* phage family. This highlights the need for careful and accurate quality control of metagenome-assembled phage genomes prior to addition to databases. Interestingly, the branch lengths in the generated maximum likelihood proteome tree of Genus 2 are longer than Genus 1, suggesting a higher level of divergence (Figure 3.11). Themetadata pertaining to the novel family was consulted to determine if any disease or geographical correlations could be determined. As mentioned above, Genus 2 phage genomes were only present in metagenomes originating from China, suggesting a country-specific genus. However, additional metagenomes should be screened to determine the accuracy of this claim. Interestingly, the closest known relatives of *B. fragilis* phage according to proteomic phylogenetics were phage with no known association with the human microbiota: *Croceibacter* phage P2559S (NC_018276) from surface water, *Croceibacter* phage P2559Y (NC_023614) from surface water and *Flavobacterium* sp. phage 1/32 (KJ018210) from Baltic sea ice. The hosts for these phage (*Croceibacter atlanticus* HTCC2559 and *Flavobacterium gelidilacus* LMG 21619) are regarded as marine bacteria with no known association with the human gut microbiota[143,144]. A similar observation was noted for crAss-like phage during its discovery[76].

Phage phylogenetics is rapidly changing as more phage are discovered and characterised. Although there is no accepted methodology for phage phylogenetics, most studies use protein comparison with conserved structural genes (e.g. MCP, tail fibre, TerL)[2,8,71]. The TerL is commonly chosen for phage phylogenetics due to its role in DNA packaging and low selective pressure. Furthermore, a 2021 study used 3 phage markers (terminase large subunit, major capsid protein and portal protein) to identify ~3700 unknown phage from human gut metagenomes[155]. However, as mentioned previously, there no universal gene shared among all phage. Therefore, to gain true picture of the relationship between phage within a genus or family, a phylogenetic tree should be constructed from all shared proteins. Additionally, outgroups should be used in the phylogenetic tree to give a wider context to the placement of the taxonomic group within phage taxonomy. In this study crAssphage and *Croceibacter* phage were used as an outgroup for construction of the phylogenetic tree. However, an exemplar species from each of the defined dsDNA phage families should have been used. In recent years, pangenome analysis have been used to determine the core genes within a phage taxonomic level (i.e genus). However, there is no defined cutoff for sequence similarity and sequence coverage. A 2021 paper recommended a sequence similarity and sequence coverage cut off as >30% identity and >50% coverage for genus level identification[156]. However, a 2022 study reported the core genes of 24 *Klebsiella* phage were defined using ≥95% identity and ≥70% coverage, suggesting pangenome guidelines may need modification[157]. In addition to protein-based phylogeny, pairwise genome comparisons are commonly used to determine family-, genus-or species-level thresholds. The currently accepted

cut-off for species is 90 % and genus is 70 %[33].A variety of tools are available to calculate phage genomic similarities; however, VIRIDIC was chosen for this study. VIRIDIC offers an advantage over other viral genome comparison tools (ANICalculator, OrthoANI, EMBOSS Stretcher, Gegenees, JSpeciesWS, Pairwise Sequence Comparison, Sequence Demarcation Tool, Yet Another Similarity Searcher) as it normalises to the whole genome length, whereas other tools only apply normalisation to the length of the alignment. This can generate high similarity values that are an artifact of the tool chosen [33,145-152]Additionally, some studies use percentage of shared orthologues to determine family, genus or species thresholds. However, these cut-offs are not as well defined as genome similarity[153]. It is clear that phage phylogeny is in its infancy and will become more defined as phage discovery and characterisation increases.

Creation of the *Bacteroides* phage dataset used in this Chapter revealed the vast unexplored diversity within *Bacteroides* (Figure 3.8). It also revealed a potential novel unrelated family of *Bacteroides* jumbophage. Phylogenetic tree and orthologue analysis showed *Bacteroides* phage, while they clustered together in vConTACT analysis, are distinct from one another. The closest known reference phage included marine phage (*Cellulophaga* phage and *Croceibacter* phage) and duck microbiota-associated phage (*Riemerella* phage RAP44). It is well documented that phage with taxonomically related hosts can be genetically diverse; this is noted with the lack of similaritybetween *B. fragilis* phage and *B. thetaotaiomicron* phage[75,154]. To fully understand the genetic diversity within the *Bacteroides* dataset an in-depth study of each VC will need to be undertaken, one that is outside the scope of this Chapter.

In conclusion, this Chapter combined phage isolation and metagenome-based phage discovery approaches to characterise a novel potential *Bacteroides fragilis* phage and family. Future studies should explore the complex phage-host relationship of vB_BfrS_23 within the gut microbiota.
Additionally, the presence and abundance of the novel *B. fragilis* phage family within the human gut microbiota should be investigated.

3.5 References

1 Hatcher, E. L. *et al.* Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic Acids Research* **45**,D482-D490 (2017).

2 Dion, M. B., Oechslin, F. & Moineau, S. Phage diversity, genomics and phylogeny. *Nature Reviews Microbiology* **18**, 125-138(2020).

3 Ackermann, H. W. 5500 Phages examined in the electron microscope. *Arch Virol* **152**, 227-243 (2007).

4 Hatfull, G. F. Bacteriophage genomics. *Curr Opin Microbiol* **11**, 447-453 (2008).

5 Wigington, C. H. *et al.* Re-examination of the relationship between marine virus and microbial cell abundances.

NatureMicrobiology **1** (2016).

6 Aggarwala, V., Liang, G. & Bushman, F. D. Viral communities of the human gut: metagenomic analysis of composition anddynamics. *Mob DNA* **8**, 12, doi:10.1186/s13100-017-0095-y (2017).

7 Williamson, K. E., Fuhrmann, J. J., Wommack, K. E. & Radosevich, M. Viruses in Soil Ecosystems: An Unknown QuantityWithin an Unexplored Territory. *Annu Rev Virol* **4**, 201-219 (2017).

8 Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425-+ (2020).

9 Krupovic, M., Prangishvili, D., Hendrix, R. W. & Bamford, D. H. Genomics of Bacterial and Archaeal Viruses: Dynamics withinthe Prokaryotic Virosphere. *Microbiol Mol Biol R* **75**, 610-+ (2011).

10 Grose, J. H. & Casjens, S. R. Understanding the enormous diversity of bacteriophages: The tailed phages that infect thebacterial family Enterobacteriaceae. *Virology* **468**, 421-443 (2014).

11 Mavrich, T. N. & Hatfull, G. F. Bacteriophage evolution differs by host, lifestyle and genome. *Nature Microbiology* **2** (2017).

12 Lederberg, E. M. & Lederberg, J. Genetic Studies of Lysogenicity in Escherichia-Coli. *Genetics* **38**, 51-64 (1953).

13 Clokie, M. R. J. & Kropinski, A. M. *Bacteriophages : methods and protocols*. (Humana Press, 2009).

14 Bilen, M. *et al.* The contribution of culturomics to the repertoire of isolated human bacterial and archaeal species. *Microbiome* **6** (2018).

15 Lagier, J. C. *et al.* Culturing the human microbiota and culturomics'. *Nature Reviews Microbiology* **16**, 540-550 (2018).

16 Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425-+ (2016).

17 Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-+ (2019).

18 Simmonds, P. *et al.* Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology* **15**, 161-168 (2017).

19 Aggarwala, V., Liang, G. X. & Bushman, F. D. Viral communities of the human gut: metagenomic analysis of composition anddynamics. *Mobile DNA-Uk* **8** (2017).

20 Duerkop, B. A. *et al.* Murine colitis reveals a disease-associated bacteriophage community. *Nat Microbiol* **3**, 1023-1031,doi:10.1038/s41564-018-0210-y (2018).

21 Martinez, J. M., Martinez-Hernandez, F. & Martinez-Garcia, M. Single-virus genomics and beyond. *Nature ReviewsMicrobiology* (2020).

22 Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat Commun* **8**,15892, doi:10.1038/ncomms15892 (2017).

23 Pena, M. J. D. *et al.* Deciphering the Human Virome with Single-Virus Genomics and Metagenomics. *Viruses-Basel* **10** (2018).

24 Khot, V., Strous, M. & Hawley, A. K. Computational approaches in viral ecology. *Comput Struct Biotechnol J* **18**, 1605-1612,doi:10.1016/j.csbj.2020.06.019 (2020).

25 Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol R* **84** (2020).

26 Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Reticulate representation of evolutionary and functionalrelationships between phage genomes. *Mol Biol Evol* **25**, 762-777 (2008).

27 Rohwer, F. & Edwards, R. The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of Bacteriology* **184**,4529-4535 (2002).

28 Paterson, S. *et al.* Antagonistic coevolution accelerates molecular evolution. *Nature* **464**, 275-U154 (2010).

29 Tikhonenko, A. S. *Ultrastructure of bacterial viruses*. (Plenum Press, 1970).

30 Baker, M. L., Jiang, W., Rixon, F. J. & Chiu, W. Common ancestry of herpesviruses and tailed DNA bacteriophages. *Journal ofVirology* **79**, 14967-14970 (2005).

31 Effantin, G., Boulanger, P., Neumann, E., Letellier, L. & Conway, J. F. Bacteriophage T5 structure reveals similarities withHK97 and T4 suggesting evolutionary relationships. *J Mol Biol* **361**, 993-1002 (2006).

32 Fokine, A. *et al.* Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to acommon ancestry. *P Natl Acad Sci USA* **102**, 7163-7168 (2005).

33 Moraru, C., Varsani, A. & Kropinski, A. M. VIRIDIC-A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses. *Viruses-Basel* **12** (2020).

34 Glazko, G., Makarenkov, V., Liu, J. & Mushegian, A. Evolutionary history of bacteriophages with double-stranded DNAgenomes. *Biol Direct* **2** (2007).

35 Liu, J., Glazko, G. & Mushegian, A. Protein repertoire of double-stranded DNA bacteriophages. *Virus Res* **117**, 68-80 (2006).

36 Bolduc, B. *et al.* vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**,e3243, doi:10.7717/peerj.3243 (2017).

37 Jang, H. B. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology* **37**, 632-+ (2019).

38 Barylski, J. *et al.* Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Phages. *SystematicBiology* **69**, 110-123 (2020).

39 Weitz, J. S. *et al.* Phage-bacteria infection networks. *Trends Microbiol* **21**, 82-91, doi:10.1016/j.tim.2012.11.003 (2013).

40 Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, 66, doi:10.3390/v8030066 (2016).

41 Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Z. VirFinder: a novel k-mer based tool for identifying viral sequencesfrom assembled metagenomic data. *Microbiome* **5** (2017).

42 Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *Peerj* **3** (2015).

43 Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lyticbacterial viruses in aquatic systems. *Limnol Oceanogr* **45**, 1320-1328 (2000).

44 Horvath, P. & Barrangou, R. CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science* **327**, 167-170 (2010).

45 Bland, C. *et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromicrepeats. *Bmc Bioinformatics* **8** (2007).

46 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712 (2007).

47 Camacho, C. *et al.* BLAST plus : architecture and applications. *Bmc Bioinformatics* **10** (2009).

48 Sanguino, L., Franqueville, L., Vogel, T. M. & Larose, C. Linking environmental prokaryotic viruses and their host throughCRISPRs. *FEMS Microbiol Ecol* **91**, doi:10.1093/femsec/fiv046 (2015).

49 Barrangou, R. & Oost, J. v. d. *CRISPR-Cas systems : RNA-mediated adaptive immunity in bacteria and archaea*. (Springer, 2013).

50 Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries ofspacers and repeats. *Bmc Bioinformatics* **8** (2007).

51 Zhao, X. H., Yu, Z. X. & Xu, Z. B. Study the Features of 57 Confirmed CRISPR Loci in 38 Strains of Staphylococcus aureus. *Frontiers in Microbiology* **9** (2018).

52 Van Goethem, M. W., Swenson, T. L., Trubl, G., Roux, S. & Northen, T. R. Characteristics of Wetting-Induced BacteriophageBlooms in Biological Soil Crust. *mBio* **10**, doi:10.1128/mBio.02287-19 (2019).

53 Arkhipova, K. *et al.* Temporal dynamics of uncultured viruses: a new dimension in viral diversity. *Isme Journal* **12**, 199-211(2018).

54 Reyes, A., Wu, M., McNulty, N. P., Rohwer, F. L. & Gordon, J. I. Gnotobiotic mouse model of phage-bacterial host dynamicsin the human gut. *P Natl Acad Sci USA* **110**, 20236-20241 (2013).

55 Koskella, B. Bacteria-Phage Interactions across Time and Space: Merging Local Adaptation and Time-Shift Experiments toUnderstand Phage Evolution. *Am Nat* **184**, S9-S21 (2014).

56 Needham, D. M. *et al.* Short-term observations of marine bacterial and viral communities: patterns, connections andresilience. *Isme Journal* **7**, 1274-1285 (2013).

57 Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage-hostrelationships. *Fems Microbiology Reviews* **40**, 258-272 (2016).

58 Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free d2*oligonucleotide frequency dissimilarity measureimproves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* **45**, 39-53, doi:10.1093/nar/gkw1002 (2017).

59 Pride, D. T., Wassenaar, T. M., Ghose, C. & Blaser, M. J. Evidence of host-virus co-evolution in tetranucleotide usagepatterns of bacteriophages and eukaryotic viruses. *Bmc Genomics* **7** (2006).

60 Coutinho, F. H., Gregoracci, G. B., Walter, J. M., Thompson, C. C. & Thompson, F. L. Metagenomics Sheds Light on theEcology of Marine Microbes and Their Viruses. *Trends Microbiol* **26**, 955-965, doi:10.1016/j.tim.2018.05.015 (2018).

61 Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689-693,doi:10.1038/nature19366 (2016).

62 Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol* **3**, 870-880,doi:10.1038/s41564-018-0190-y (2018).

63 Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications* **5** (2014).

64 Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *NatureMicrobiology* **4**, 1727-1736 (2019).

65 Yarygin, K. *et al.* Abundance profiling of specific gene groups using precomputed gut metagenomes yields novel biologicalhypotheses. *Plos One* **12** (2017).

66 Manrique, P. *et al.* Healthy human gut phageome. *P Natl Acad Sci USA* **113**, 10400-10405 (2016).

67 Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, 527-+(2019).

68 Shkoporov, A. N. *et al.* PhiCrAss001 represents the most abundant bacteriophage family in the human gut and infectsBacteroides intestinalis. *Nat Commun* **9**, 4781, doi:10.1038/s41467-018-07225-7 (2018).

69 Siranosian, B. A., Tamburini, F. B., Sherlock, G. & Bhatt, A. S. Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. *Nat Commun* **11**, 280, doi:10.1038/s41467-019-14103-3 (2020).

70 Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the humangut. *Nature Microbiology* **3** (2018).

71 Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell HostMicrobe* **24**, 653-664 e656, doi:10.1016/j.chom.2018.10.002 (2018).

72 Shkoporov, A., Stockdale, SR., Adriaenssens EM., Yutin N., Koonin EV., Dutilh BE., Krupovic M., Edwards RA., Tolstoy I., Hill C. . in *https://talk.ictvonline.org/files/proposals/taxonomy_proposals_prokaryote1/m/bact02/10964* (2020).

73 Jonge, P. A., Meijenfeldt, F., Rooijen, L. E. V., Brouns, S. J. J. & Dutilh, B. E. Evolution of BACON Domain Tandem Repeats incrAssphage and Novel Gut Bacteriophage Lineages. *Viruses* **11**, doi:10.3390/v11121085 (2019).

74 Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, 527-541 e525, doi:10.1016/j.chom.2019.09.009 (2019).

75 Hryckowian, A. J. *et al.* Bacteroides thetaiotaomicron-Infecting Bacteriophage Isolates Inform Sequence-Based Host RangePredictions. *Cell Host Microbe* **28**, 371-+ (2020).

76 Koonin, E. V. & Yutin, N. The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome. *Trends inMicrobiology* **28**, 349-359 (2020).

77 Pramono, A. K. *et al.* Discovery and Complete Genome Sequence of a Bacteriophage from an Obligate IntracellularSymbiont of a Cellulolytic Protist in the Termite Gut. *Microbes Environ* **32**, 112-117 (2017).

78 Holmfeldt, K. *et al.* Twelve previously unknown phage genera are ubiquitous in global oceans. *P Natl Acad Sci USA* **110**,12798-12803 (2013).

79 Porter, N. T. *et al.* Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility inBacteroides thetaiotaomicron. *Nat Microbiol* **5**, 1170-1181, doi:10.1038/s41564-020-0746-5 (2020).

80 Porter, N. T., Canales, P., Peterson, D. A. & Martens, E. C. A Subset of Polysaccharide Capsules in the Human SymbiontBacteroides thetaiotaomicron Promote Increased Competitive Fitness in the Mouse Gut. *Cell Host Microbe* **22**, 494-+ (2017).

81 Porter, N. T. & Martens, E. C. The Critical Roles of Polysaccharides in Gut Microbial Ecology and Physiology. *Annual Reviewof Microbiology, Vol 71* **71**, 349-369 (2017).

82 Peterson, D. A., McNulty, N. P., Guruge, J. L. & Gordon, J. I. IgA response to symbiotic bacteria as a mediator of guthomeostasis. *Cell Host Microbe* **2**, 328-339 (2007).

83 Patrick, S. *et al.* Twenty-eight divergent polysaccharide loci specifying within- and amongst-strain capsule diversity in threestrains of Bacteroides fragilis. *Microbiol-Sgm* **156**, 3255-3269 (2010).

84 Shkoporov, A. N. *et al.* Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroidesintestinalis. *bioRxiv*, 2020.2012.2002.408625, doi:10.1101/2020.12.02.408625 (2020).

85 Ebdon, J. E., Sellwood, J., Shore, J. & Taylor, H. D. Phages of Bacteroides (GB-124): A Novel Tool for Viral WaterborneDisease Control? *Environ Sci Technol* **46**, 1163-1169 (2012).

86 Purnell, S., Ebdon, J., Buck, A., Tupper, M. & Taylor, H. Bacteriophage removal in a full-scale membrane bioreactor (MBR) -Implications for wastewater reuse. *Water Research* **73**, 109-117 (2015).

87 Payan, A. *et al.* Method for isolation of Bacteroides bacteriophage host strains suitable for tracking sources of fecalpollution in water. *Appl Environ Microbiol* **71**, 5659-5662, doi:10.1128/AEM.71.9.5659-5662.2005 (2005).

88 McMinn, B. R., Korajkic, A. & Ashbolt, N. J. Evaluation of Bacteroides fragilis GB-124 bacteriophages as novel human-associated faecal indicators in the United States. *Lett Appl Microbiol* **59**, 115-121, doi:10.1111/lam.12252 (2014).

89 Jofre, J., Blanch, A. R., Lucena, F. & Muniesa, M. Bacteriophages infecting Bacteroides as a marker for microbial sourcetracking. *Water Research* **55**, 1-11 (2014).

90 Prado, T. *et al.* Distribution of human fecal marker GB-124 bacteriophages in urban sewage and reclaimed water of SaoPaulo city, Brazil. *J Water Health* **16**, 289-299 (2018).

91 Paez-Espino, D. *et al.* IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Res* **45**,D457-D465, doi:10.1093/nar/gkw1030 (2017).

92 Joshi, N. & Fass, J. (2011).

93 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

94 Kropinski, A. M. in *Bacteriophages: Methods and Protocols, Volume 3* (eds Martha R. J. Clokie, Andrew M. Kropinski, & Rob Lavigne) 41-47 (Springer New York, 2018).

95 Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and longsequencing reads. *PLoS computational biology* **13**, e1005595 (2017).

96 Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC genomics* **9**, 1-15 (2008).

97 Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleicacids research* **42**, D206-D214 (2014).

98 Brettin, T. *et al.* RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotationpipelines and annotating batches of genomes. *Scientific reports* **5**, 8365 (2015).

99 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410(1990).

100 Marchler-Bauer, A. *et al.* CDD: NCBI's conserved domain database. *Nucleic acids research* **43**, D222-D226 (2015).

101 Ogilvie, L. A. *et al.* Comparative (meta)genomic analysis and ecological profiling of human gut-specific bacteriophagephiB124-14. *PLoS One* **7**, e35053, doi:10.1371/journal.pone.0035053 (2012).

102 Puig, M. & Girones, R. Genomic structure of phage B40-8 of Bacteroides fragilis. *Microbiology* **145**, 1661-1670 (1999).

103 Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization andanalysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**, 464-469, doi:10.1093/bioinformatics/btr703 (2012).

104 Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr ProtocBioinformatics* **Chapter 2**, Unit 2 3, doi:10.1002/0471250953.bi0203s00 (2002).

105 Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research* **47**, W636-W641,doi:10.1093/nar/gkz268 (2019).

106 Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences usingGubbins. *Nucleic Acids Research* **43**, e15-e15, doi:10.1093/nar/gku1196 (2014).

107 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOSONE* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).

108 Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research* **44**, W16-W21(2016).

109 Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gutbacteriophage diversity. *bioRxiv*, 2020.2009.2003.280214, doi:10.1101/2020.09.03.280214 (2020).

110 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659, doi:10.1093/bioinformatics/btl158 (2006).

111 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069,doi:10.1093/bioinformatics/btu153 (2014).

112 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genomeresearch* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

113 Nayfach, S., Camargo, A. P., Eloe-Fadrosh, E., Roux, S. & Kyrpides, N. CheckV: assessing the quality of metagenome-assembled viral genomes. *bioRxiv*, 2020.2005.2006.081778, doi:10.1101/2020.05.06.081778 (2020).

114 Rambaut, A. *FigTree*, <http://tree.bio.ed.ac.uk/software/figtree/> (

115 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238,doi:10.1186/s13059-019-1832-y (2019).

116 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).

117 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast BootstrapApproximation. *Mol Biol Evol* **35**, 518-522, doi:10.1093/molbev/msx281 (2017).

118 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection foraccurate phylogenetic estimates. *Nature Methods* **14**, 587-589, doi:10.1038/nmeth.4285 (2017).

119 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol BiolEvol* **37**, 1530-1534, doi:10.1093/molbev/msaa015 (2020).

120 Bacic, M. K. & Smith, C. J. Laboratory maintenance and cultivation of bacteroides species. *Curr Protoc Microbiol* **Chapter 13**,10.1002/9780471729259.mc9780471729213c9780471729201s9780471729259-9780471729213C.9780471729251, doi:10.1002/9780471729259.mc13c01s9 (2008).

121 Ebdon, J., Muniesa, M. & Taylor, H. The application of a recently isolated strain of Bacteroides (GB-124) to identify humansources of faecal pollution in a temperate river catchment. *Water Res* **41**, 3683-3690, doi:10.1016/j.watres.2006.12.020 (2007).

122 Vargo, V., Korzeniowski, M. & Spaulding, E. H. Tryptic soy bile-kanamycin test for the indentification of Bacteroides fragilis. *Appl Microbiol* **27**, 480-483 (1974).

123 Finer-Moore, J. S., Maley, G. F., Maley, F., Montfort, W. R. & Stroud, R. M. Crystal structure of thymidylate synthase from T4phage: component of a deoxynucleoside triphosphate-synthesizing complex. *Biochemistry* **33**, 15459-15468, doi:10.1021/bi00255a028 (1994).

124 Lemire, S., Figueroa-Bossi, N. & Bossi, L. Bacteriophage crosstalk: coordination of prophage induction by trans-actingantirepressors. *PLoS Genet* **7**, e1002149, doi:10.1371/journal.pgen.1002149 (2011).

125 Kala, S. *et al.* HNH proteins are a widespread component of phage DNA packaging machines. *Proceedings of the NationalAcademy of Sciences* **111**, 6022-6027 (2014).

126 Feiss, M. & Rao, V. B. in *Viral molecular machines* 489-509 (Springer, 2012).

127 Rao, V. B. & Feiss, M. The bacteriophage DNA packaging motor. *Annual review of genetics* **42**, 647-681 (2008).

128 Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985,doi:10.7717/peerj.985 (2015).

129 Murphy, J., Mahony, J., Ainsworth, S., Nauta, A. & van Sinderen, D. Bacteriophage Orphan DNA Methyltransferases: Insightsfrom Their Bacterial Origin, Function, and Occurrence. *Applied and Environmental Microbiology* **79**, 7547-7555, doi:10.1128/aem.02229-13 (2013).

130 Keller, R. & Traub, N. The characterization of Bacteroides fragilis bacteriophage recovered from animal sera: observationson the nature of bacteroides phage carrier cultures. *Journal of General Virology* **24**, 179-189 (1974).

131 Tariq, M. A. *et al.* A metagenomic approach to characterize temperate bacteriophage populations from Cystic Fibrosis andnon-Cystic Fibrosis bronchiectasis patients. *Frontiers in microbiology* **6**, 97 (2015).

132 Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non–host-derived immunity. *Proceedings of the NationalAcademy of Sciences* **110**, 10771-10776 (2013).

133 Sathaliyawala, T. *et al.* Functional analysis of the highly antigenic outer capsid protein, Hoc, a virus decoration protein fromT4-like bacteriophages. *Molecular microbiology* **77**, 444-455 (2010).

134 Bertrand, I. *et al.* The impact of temperature on the inactivation of enteric viruses in food and water: a review. *Journal ofApplied Microbiology* **112**, 1059-1074 (2012).

135 Tariq, M. A. *et al.* Genome Characterization of a Novel Wastewater Bacteroides fragilis Bacteriophage (vB_BfrS_23) and itsHost GB124. *Frontiers in Microbiology* **11**, doi:10.3389/fmicb.2020.583378 (2020).

136 Perez-Brocal, V. *et al.* Metagenomic Analysis of Crohn's Disease Patients Identifies Changes in the Virome and MicrobiomeRelated to Disease Status and Therapy, and Detects Potential Interactions and Biomarkers. *Inflamm Bowel Dis* **21**, 2515- 2532, doi:10.1097/MIB.0000000000000549 (2015).

137 Shenoy, P. A. *et al.* Anaerobic bacteria in clinical specimens–frequent, but a neglected lot: a five year experience at atertiary care hospital. *Journal of Clinical and Diagnostic Research: JCDR* **11**, DC44 (2017).

138 Cook, R., Millard, A. *Adding More Reference Genomes to vConTACT2 Clusters*, <http://millardlab.org/2020/02/18/adding-more-reference-genomes-to-vcontact2-clusters/> (2020).

139 Fuxman Bass, J. I. *et al.* Using networks to measure similarity between genes: association index selection. *Nature methods* **10**, 1169-1176, doi:10.1038/nmeth.2728 (2013).

140 Corel, E., Lopez, P., Méheust, R. & Bapteste, E. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Trends in Microbiology* **24**, 224-237, doi:https://doi.org/10.1016/j.tim.2015.12.003 (2016).

141 Iranzo, J., Koonin, E. V., Prangishvili, D. & Krupovic, M. Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. *Journal of Virology* **90**, 11043-11055, doi:10.1128/jvi.01622-16 (2016).

142 Iranzo, J., Krupovic, M. & Koonin, E. V. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of GeneSharing. *mBio* **7**, e00978-00916, doi:10.1128/mBio.00978-16 (2016).

143 Luhtanen, A.-M. *et al.* Isolation and characterization of phage–host systems from the Baltic Sea ice. *Extremophiles* **18**, 121-130, doi:10.1007/s00792-013-0604-y (2014).

144 Oh, H.-M., Kang, I., Ferriera, S., Giovannoni, S. J. & Cho, J.-C. Complete genome sequence of Croceibacter atlanticusHTCC2559T. *Journal of bacteriology* **192**, 4796-4797, doi:10.1128/JB.00733-10 (2010).

145 Han, N., Qiang, Y. & Zhang, W. ANItools web: a web tool for fast genome comparison within multiple bacterial strains. *Database (Oxford)* **2016**, baw084, doi:10.1093/database/baw084 (2016).

146 Lee, I., Ouk Kim, Y., Park, S. C. & Chun, J. OrthoANI: An improved algorithm and software for calculating average nucleotideidentity. *Int J Syst Evol Microbiol* **66**, 1100-1103, doi:10.1099/ijsem.0.000760 (2016).

147 Ceyssens, P. J. *et al.* Comparative analysis of the widespread and conserved PB1-like viruses infecting Pseudomonasaeruginosa. *Environ Microbiol* **11**, 2874-2883, doi:10.1111/j.1462-2920.2009.02030.x (2009).

148 Ågren, J., Sundström, A., Håfström, T. & Segerman, B. Gegenees: Fragmented Alignment of Multiple Genomes for Determining Phylogenomic Distances and Genetic Signatures Unique for Specified Target Groups. *PLOS ONE* **7**, e39107,doi:10.1371/journal.pone.0039107 (2012).

149 Richter, M., Rosselló-Móra, R., Oliver Glöckner, F. & Peplies, J. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* **32**, 929-931, doi:10.1093/bioinformatics/btv681 (2015).

150 Bao, Y., Chetvernin, V. & Tatusova, T. Improvements to pairwise sequence comparison (PASC): a genome-based web toolfor virus classification. *Arch Virol* **159**, 3293-3304, doi:10.1007/s00705-014-2197-x (2014).

151 Muhire, B. M., Varsani, A. & Martin, D. P. SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment andIdentity Calculation. *PLOS ONE* **9**, e108277, doi:10.1371/journal.pone.0108277 (2014).

152 Noé, L. & Kucherov, G. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic acids research* **33**, W540-W543,doi:10.1093/nar/gki478 (2005).

153 Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic acids research* **45**, D491-D498, doi:10.1093/nar/gkw975 (2017).

154 Pope, W. H. *et al.* Bacteriophages of Gordonia spp. Display a Spectrum of Diversity and Genetic Relationships. *mBio* **8**,e01069-01017, doi:10.1128/mBio.01069-17 (2017).

155 Benler, S. *et al.* Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome* **9**, 78, doi:10.1186/s40168-021-01017-w (2021).

156 Turner, D., Kropinski, A. M. & Adriaenssens, E. M. A Roadmap for Genome-Based Phage Taxonomy. *Viruses* **13**, 506,doi:10.3390/v13030506 (2021).

157 Smith-Zaitlik, T., Shibu, P., McCartney, A.L., Foster, G., Hoyles, L. & Negs, D. Extended genomic analysis of the broad-host-range phages vB_KmiM-2Di and vB_KmiM-4Dii reveal slopekviruses have highly conserved genomes. *bioRxiv,* doi:10.1101/2022/04.06.486684 (2022)

Chapter 4 : Analysis of the *Bacteroides fragilis* pangenome

4.1 Introduction

Bacteroides species play a pivotal role within the human microbiome and are among the most prevalent anaerobic bacteria within the gastrointestinal tract[1,2]. Members of the genus *Bacteroides* are generally considered commensal; however, several species have been implicated in infection and are considered opportunistic pathogens[3-7]. *Bacteroides fragilis,* which accounts for 2 % of *Bacteroides* species within the human gut, is an important opportunistic pathogen[8,9]. Why it switches from commensal to infectious agent is unknown. It is estimated that *B. fragilis* is responsible for > 70 % of extra-intestinal infections causes by a *Bacteroides* species[10]. It is the main cause of intra-abdominal abscesses and anaerobic septicaemia. *B. fragilis* has also been associated with soft tissue infections, peritonitis, brain abscess, gynaecological infections, and surgical-site infections[10-13]. A 2005 study in Taiwan revealed *B. fragilis* was isolated in 45 % of systemic infections[14]. In addition to isolation from extra-intestinal sites and faecal samples from healthy individuals, specific *B. fragilis* isolates are associated with inflammatory diarrheal disease due to the secretion of a metalloprotease toxin (*B. fragilis* toxin (BFT))[7].

Bacteroides species are intrinsically resistant to aminoglycoside antibiotics and *B. fragilis* isolates show an increasing resistance to tetracycline antibiotics[15-17]. The rapid spread of tetracycline resistance is due to the transfer of *tetQ* gene via horizontal gene transfer (HGT), with 80 % of *B. fragilis* isolates tetracycline-resistant[17,18]. Additionally, *B. fragilis* shows resistance to penicillin, which is attributed to cephalosporinase genes (*cepA*)[19]. Resistance to cephamycins and carbapenems in a minority of *B. fragilis* strains has been attributed to *cfiA*/*ccrA*[20,21]. There are also reports of multi-drug resistant (MDR) *B. fragilis* emerging in clinical settings in the UK, USA, and Afghanistan[22]. Genome sequence analysis of an MDR strain from the UK revealed the presence of *cfiA* (metallo-β-lactamase), *ermF* (macrolides, lincosamides, streptogramines resistance) and *tetQ* (tetracycline resistance)[22].

4.1.1 Toxigenic and non-toxigenic strains

Enterotoxigenic *B. fragilis* (ETBF) and non-toxigenic *B. fragilis* (NTBF) strains can have opposing roles within the microbiota and interact with the human host[23]. For example, the NTBF *B. fragilis* reference strain (NCTC 9343^{T}) was isolated from a systemic infection and suppresses intestinal inflammation in mice[24]. Whereas the ETBF strains can initiate intestinal inflammation through disruption of the intestinal barrier via secretion of BFT[25]. Additionally, ETBF have been implicated

in the development of colorectal cancer. An isolate encoding *bft-2* was able to coordinate a pro-inflammatory pro-carcinogenic cascade in a mouse model[3]. Furthermore, several studies have reported an increased prevalence of ETBF in individuals with pre-cancerous or cancerous intestinal lesions compared to healthy controls[26,27]. The *B. fragilis* pathogenicity island (BfPAI) encodes the associated BFT genes, and the flanking mobilisation proteins (CTn86, a conjugative transposon) suggest the BfPAI is transmissible[28,29]. It is thought that the BfPAI has been obtained through independent acquisition as ETBF and NTBF strains do not form monophyletic clusters. This idea is further supported by sequence analysis of a NTBF isolate, 638R, which is more closely related to ETBF isolates than other NTBF[30].

Approximately 30 % of healthy individuals harbour ETBF isolates asymptomatically[31]. This suggests that pathogenicity of ETBF may be dependent on host susceptibility, such as intestinal barrier integrity, and that pathogenic potential may vary[32]. Additionally, ETBF strains can possess differing *bft* isoforms (*bft-1, bft-2, bft-3)* and the copy number within a strain can vary[31,33]. It has been proposed that the variability of ETBF pathogenicity could also be attributed to type of *bft* and copy number[33].

4.1.2 Potential virulence factors

The pathogenic potential of *B. fragilis* can be attributed to its currently known virulence factors, which include genes involved in attachment to host tissue, defence from host immune system and tissue destruction[34-38]. Several cell surface structures have been characterised in *B. fragilis* that are associated with pathogenicity (such as pili and fimbriae)[39,40]. For example, one study implicated peritrichous fimbriae in attachment to host tissue as inhibition of haemagglutination reduced adhesion to a human intestinal cell line[40]. Furthermore, multiple homologues of the streptococcal virulence factor (SpeB), a C10 family protease, were discovered within *B. fragilis* isolates 638R, YCH46 and NCTC 9343^{T} [41,42]. It has also been suggested that lipopolysaccharide (LPS) is involved in pathogenicity; however, the LPSs of *Bacteroides* species do not contain an O-antigen and have been shown to be significantly less virulent compared to LPS derived from *Esherichia coli*[43,44].

B. fragilis capsular polysaccharides are heavily involved in abscess formation[45-47]. Outside of the intestinal environment capsular polysaccharide A (PSA) in *B. fragilis* NCTC 9343^{T} induced abscess formation in animal models[46]. The authors reported induction of abscess formation was higher with PSA, compared to polysaccharide B or polysaccharide C. Furthermore, several secreted enzymes (e.g. hyaluronidase and chondroitin sulfatase) have been implicated in host tissue

destruction, such as degradation of the epithelial barrier[35,48]. Haemolysins are commonly secreted by pathogenic bacteria to lyse host cells and contribute towards pathogen survival[49]. It is also suggested that commensal bacteria use haemolysins for niche competition with the intestinal environment[50]. The exact role of haemolysins is unknown but a handful of studies have indicated they may be important for survival in oxygen-rich environments, like outside the large intestine[50]. A 2013 study showed that haemolysin gene expression of *B. fragilis* isolates was increased in an oxygen-rich environment[51]. Furthermore, reduced survival of *B. fragilis* mutants lacking two haemolysin genes was noted.

4.1.3 Polysaccharide capsules and LPS in *B. fragilis*

Despite its potential pathogenicity, *B. fragilis* can promote immune tolerance within the human host[24,52]. PSA expressed by *B. fragilis* NCTC 9343^T has been shown to maintain immune tolerance by promoting production of anti-inflammatory cytokine interleukin-10 via regulatory T cells[52]. The surface polysaccharides (PSs) of *Bacteroides* species exhibit extensive within- and between-strain variation[46,53,54]. Each *B. fragilis* strain has the potential to express three outer capsules: large, small and micro. These are structurally and antigenically distinct from one another[55-57]. The large capsule and small capsule are ON-OFF-ON phase-variable[57]. A PS biosynthesis locus has not been confidently assigned to large capsule expression; however, a *wbaP*-like glycosyltransferase is associated with exportation[58].

Microcapsules are also ON-OFF-ON phase-variable and a single strain can expression at least eight unique microcapsules[59,60]. Genomic analysis of *B. fragilis* NCTC 9343^T revealed eight diverse PS biosynthesis loci (PSA-PSH), with seven of these regions containing upstream invertible DNA promoter regions believed to be responsible for PS switching within the strain[61]. A similar invertible promoter region was not detected in PS locus C. Significant sequence similarity of the upstream inverted regions to an invertible promoter region within *Salmonella* was discovered. This region is associated with variable expression of differing flagella. These regions have been termed *B. fragilis* inversion crossover (fix)1 sites[62].

Although there is significant diversity among PS loci, the genomic arrangement between loci is relatively conserved. A 2001 study examined the PSA locus within 50 *B. fragilis* strains and reported a high level of conservation in the regions up- and down-stream of the locus[46]. However, the PSA locus was not conserved, and genes appeared to be diverse[53]. One study highlighted that the polymerase, flippase and other biosynthesis-associated genes are divergent within the PS loci[54]. In addition to the conserved invertible promoter region, two regulatory genes have also

been identified downstream of the invertible promoters and upstream of the PS biosynthesis loci[63,64]. These are termed *up(a-h)Y* and *up(a-h)*Z, depending on the PS loci they are located within but referred to as *upxZ* and *upxY* when not assigned to a PS locus. UpxZ proteins act as antagonists against anti-terminator UpxY proteins and only allow transcription of one PS locus at a time. The UpxZ protein from one PS locus inhibits the transcription of UpxY from another PS locus. Several studies have shown that expression of a PS locus can be halted by genetic manipulation of the *upxY* gene directly upstream[58,64].

The variety of diverse PS loci appears to be a common feature among *Bacteroides* species but not in bacteria outside the genera[53]. Significant structural variation of PSs has also been reported between strains[53]. For example, 638R PSA contains five monosaccharides compared to NCTC 9343^{T}, which contains four[65]. However, extra-intestinal NCTC 9343^{T} PSA in a mouse model was able to induce peritoneal abscess formation compared to 638R PSA[65]. Additionally, monoclonal antibodies reactive with NCTC 9343^{T} were not reactive with 638R or YCH46 due to the unique PS on each strain[54]. Genome comparison revealed only two PS loci were conserved between the three strains; NCTC 9343^{T} and YCH46, and 638R and YCH46[54]. The true extent of *Bacteroides* surface diversity is vastly complex and yet to be determined.

The outer-most layer of all Gram-negative bacteria is the LPS layer and comprises the lipid A (closest to the bacterial peptidoglycan layer), core oligosaccharide region and PS repeating region termed the O-antigen[66,67]. LPS molecules that do not contain the O-antigen are termed as 'rough' or lipooligosaccharide (LOS) and LPS with the O-antigen cap are termed 'smooth'[67]. Due to their position on the bacterial outer membrane and shared nature among Gram-negative bacteria, LPS is heavily involved in host immune system-bacteria interactions, involving host Toll-like receptors and NOD proteins[68]. The structure of LPS varies considerably among species and is believed to be mainly attributed to functional differences[67,69]. For example, the modification in the O-antigen of *Pseudomonas aeruginosa* LPS is thought to play a role in establishment of chronic infection in cystic fibrosis[70,71]. However, little is known about the biosynthesis, structure and function of LPSs of commensal bacteria – including *B. fragilis* – and their importance to host immunity.

Several studies have reported an altered LPS in *Bacteroides* species compared to conventionally 'pathogenic' LPS[72,73]. *Bacteroides thetaiotaomicron*, *B. fragilis* and *Phocaeicola* (formerly *Bacteroides*) *dorei* produce penta-acylated, monophosphorylated lipid A[74]. Compared to hexa-acylated, dephosphorylated LPS exhibited by *Escherichia coli*[75]. There is also controversy surrounding the presence of an O-antigen in *Bacteroides* LPS[76-78]. A recent study reported

Phocaeicola (formerly *Bacteroides*) *vulgatus* ATCC 8482^T exhibited a 'laddered' pattern on an SDS PAGE gel similar to *E. coli* O55:B5[79].The 'laddering' pattern is indicative of an O-antigen and, as the number of repeating units on an O-antigen is variable, the number of 'rungs' observed on an SDS PAGE gel can also be variable. As no O-antigen was observed in *B. thetaiotaomicron*, the authors suggested that *B. thetaiotaomicron* has an LOS, instead of an LPS. However, a 1994 study stating the presence of few repeating units with LPS sizes < 10 kDa is still referenced in recent published articles[77].

The common *B. fragilis* laboratory strains were isolated from clinical infections (e.g. NCTC 9343^T and 638R) but are often used to understand immune tolerance within the host intestine[52,54,80]. The immune response to clinical and non-clinical faecal isolates has not been widely studied; therefore, it is not known if isolates from differing isolation sites produce unique immune responses. For example, a study reported NCTC 9343^T did not significantly affect synthesis of TNF-α, a proinflammatory cytokine, in mice with LPS-induced intestinal inflammation[81]. Whereas an isolate (HCK-B3) from the faeces of a healthy donor was able to down-regulate TNF-α expression[82]. Furthermore, studies have reported strains show differing responses to intestinal immune regulators[83]. For example, isolates originating from faecal samples were more susceptible to human β-defensin-3, an antimicrobial peptide, compared to strains isolated from blood or extra-intestinal infections[82]. However, the genomic differences between non-clinical and clinical samples has not been studied extensively.

4.1.4 *B. fragilis* prophage

To date, only one *Bacteroides* species prophage has been characterised; BV01 in *P. vulgatus* (previously *B. vulgatus*)[84]. The authors reported that this phage was the first explored representative of the broad family *Salyersviridae* and discovered 20 potential BV01-like phage. *Bacteroides* species were assigned as the predicted host for all potential phage and included phage of differing lifestyle (lytic and temperate). The authors also reported the ability of BV01 to alter the bacterial host function, by repressing bile acid deconjugation, after integration of the phage[84]. The exact role of bile acid deconjugation within the microbiome and benefit to microbes is unclear even though bile acid deconjugation is relatively common among intestinal microbes. However, the modification of bile acids within the microbiome is thought to benefit human host metabolism and contribute to regional protection from viral pathogens[85,86]. These results suggest that intestinal phages may directly and indirectly influence the human host and microbiota environment through undiscovered mechanisms. Additionally, a 2016 study reported nine prophage regions in five ETBF strains and three prophage regions in three NTBF strains. Five of the

prophage showed the closest relationship to 6H phage *Flavobacterium psychrophilum*[30]. However, no prophage regions were found in *B. fragilis* NCTC 9343^{T}.

It is believed that bacteriophage may be responsible for the high level of diversity observed in *Bacteroides* PS. A 2021 study suggested that the phase variation of PSA in *Bacteroides intestinalis* is implicated in the long-term persistence of *Bacteroides* phage crAss001[87]. Furthermore, variation in outer membrane structures in *B. thetaiotaomicron* (capsular PS, S-layer lipoprotein, TonB-dependent nutrient receptors and OmpA-like proteins) are associated with phage resistance and sensitivity switching.

4.1.5 Pangenome analysis of opportunistic pathogens/commensals

Typically, pangenome analysis has been used for comparative genomics of pathogenic bacteria[88]. The pangenome is composed of the core genome, accessory genome and singleton genes (i.e. species- or strain-specific genes)[89,90]. The core genome is genes that are shared by all analysed genomes, and most are involved in vital roles for bacterial survival. However, some bacterial species have pathogenicity- and virulence-associated genes within the core genome. The accessory genome is defined as genes not conserved across all isolates but also found in more than one isolate. This is commonly genes found within 5-95 % of all isolates. The accessory genome is considered the flexible region of the pangenome as it mainly contains genes implicated in bacterial adaptation to environmental changes[91]. A 2015 study analysing the genome evolutionary dynamics in multiple *Klebsiella pneumoniae* clones revealed key differences among clades due to HGT using comparative and pangenome analyses[92]. Additionally, *Clostridium perfringens* shows a highly variable pangenome that appears to be driven by HGT[93].

In recent years, pangenome analysis has been applied to opportunistic pathogens such as *Staphylococcus epidermidis*, a common human skin commensal that has the ability to inhibit colonisation by pathogenic *Staphylococcus aureus*[94]. However, *S. epidermidis* is considered an opportunistic pathogen as it can cause infection if it enters the bloodstream[95]. *S. epidermidis* is able to form biofilms on medical devices and detachment from biofilms can lead to bacteremia[96]. A 2012 study showed that 80 % of genes in *S. epidermidis* isolates were in the core genome and the strains clustered into two distinct groups based on virulence[94]. Commensal bacterium *Cutibacterium* (formerly *Propionibacterium*) *acnes* is an important part of the skin microbiota and a pathogenic factor in several diseases, including acne[97,98]. The core genes of this bacterium accounted for 88 % of the pangenome and lineage-specific genetic elements were identified that could account for the differing phenotypes[99].

Pangenome analysis has also been used to identify important gene clusters and define subspecies in commensal bacteria[91]. The pangenome of *Bifidobacterium* species has been extensively studied due to the strain heterogeneity within subspecies and the importance of these bacteria within the infant gut microbiome[100-103]. For example, the pangenome of *Bifidobacterium longum* subsp. *longum* showed variation in its sugar usage profile and allowed authors to identify five gene clusters implicated in breakdown of xylo-oligosaccharides, arabinan, arabinoxylan, galactan and fucosyllactose (a human milk oligosaccharide)[104].

Therefore, the application of pangenome analysis to opportunistic pathogens and commensals can help identify genomic regions involved in pathogenesis or important genes needed for commensal colonisation.

4.1.6 Aims and objectives

B. fragilis is an important member of the human gut microbiota but the mechanisms of its pathogenesis remain elusive. To date, no extensive pangenome analysis has been undertaken to determine the genomic differences between ETBF, intestinal NTBF and systemic NTBF strains. It is unknown if these phenotypically distinct strains exhibit genetic differences related to lifestyle or predisposition of intestinal NTBF to cause systemic infection. This Chapter reports the pangenome analysis of 93 *B.fragilis* genomes (ETBF, intestinal NTBF and systemic NTBF strains) collected from NCBI, current literature and a newly sequenced isolate. Phylogenetic and comparative genomic analysis were applied to identify genomic regions involved in conversion from a commensal to pathogenic lifestyle (intestinal NTBF to systemic NTBF) and potential virulence factors. Additionally, the isolates were screened for the presence of antibiotic resistance genes, BFT and prophage.

4.2 Methods

4.2.1 Characterisation of *B. fragilis* isolate GB-124

4.2.1.1 DNA extraction and sequencing

The Promega Maxwell® RSC Instrument (AS4500) and Promega Maxwell® RSC Cultured Cell DNA Kit (AS1620) were used to extract DNA from *B. fragilis* GB-124. The bacterium's growth conditions are detailed in section 3.2.1.3. DNA quality and quantity were assessed using Nanodrop™ Spectrophotometer and Qubit™ dsDNA HS Assay Kit (Invitrogen™). Bacterial DNA was sequenced on an Illumina MiSeq and Oxford Nanopore Technologies MinION. The DNA was sequenced by David Baker at QIB Sequencing Service using the Illmina MiSeq system. The sequencing library was prepared with Illumina Nextera XT (Illumina, Saffron Walden, UK) library preparation kit,

sequenced on an Illumina MiSeq 2 x 150-cycle v2 chemistry and paired-end reads provided as

FASTQ files. The adapters of the raw reads were removed using Trimmomatic (v. 0.39) before quality control trimming with Sickle (v. 1.33) at --q 30 and --l 15[105,106]. For MinION sequencing, the manufacturer's protocol was followed and native barcoding kit EXP-NDB104 with the ligation sequencing kit SQK-LSK109 were used. MinION sequencing was performed with Dr Mohammad Tariq. Briefly, the NEBNext FFPE Repair Mix (M6630) and NEBNext End Repair/dA-tailing (E7546) were mixed with 1 µg of high-quality phage DNA for end-repair and dA-tailing. The native barcode kit (EXP-NBD104) was used to barcode and ligated using NEB Blunt/TA Ligase Master Mix (M0367). Sequence adapters were ligated using the NEBNext Quick Ligation Module (E6056) and samples primed and loaded using the Flow Cell Priming Kit (EXP-FLP001) on MinION R9 4.1 FLO-MIN106. Samples were sequenced for 72 h and the FAST5 files saved for base-calling and any future use. The raw reads were base-called using Guppy (v3.5.1; downloaded from https://nanoporetech.com) and adapters removed using PoreChop (v0.2.3; https://github.com/rrwick/Porechop).

4.2.1.2 Genome characterisation

Unicycler was used to create a hybrid-assembly of the genome from Illumina MiSeq and MinION reads[107]. Following assembly, the genome was annotated using Prokka (v.1.14.6). CheckM (v.1.0.18) was used to determine genome completeness and contamination[108,109]. FastANI (v.1.3) with *B. fragillis* NCTC 9343^{T} was used to confirm identification of GB-124[110]. The genome was visualised using Bandage (v.0.8.1) using the assembly graph file[111]. Additionally, ABRicate (v.0.9.8) was used to identify antimicrobial resistance (AMR) genes using Resfinder (database v. 2020-06-02) and NCBI (database v. 2020-05-04.1)[112,113]. ABRicate hits were considered significant if the coverage and identity were > 90 %. Insertion sequence (IS) elements were predicted using ISfinder (http://www-is.biotoul.fr/)[114]. Predicted IS elements (significant if bit score > 100 and E.value < 4e-11) were examined in the Prokka GenBank file and the protein sequence submitted to blastp for confirmation. The suspected IS elements were visualised in Artemis and investigated for downstream AMR genes[115].

PLSDB web server (v.0.1.3; https://ccb-microbe.cs.uni-saarland.de/plsdb/) (database v. 2020_03_04) was used to identify plasmids within the assembly[116]. Plasmid identity was confirmed using blastn[117]. Coding regions were found using Prokka (v.1.14.6) and the putative function manually checked using blastp (NCBI-nr and CDD)[108,117]. Blastp hits were considered significant if the e-values were lower than 1e^{-}5 at ≥ 80 % protein identity[117]. Plasmids were visualised using SnapGene Viewer (v.5.0.5). Resistance and virulence genes were predicted with ABRicate (v.0.9.8) and Resfinder (database v. 2020-06-02), NCBI (database v. 2020-05-04.1) and

VFDB (accessed 2020-06-30) databases[112,113,118]. The cut-off values mentioned previously were used.

4.2.2 Selection of *B. fragilis* sequence data from literature

4.2.2.1 Genome assembly from literature

Due to the lack of publicly available non-clinical strain assemblies, the literature was searched for isolation of *B. fragilis* from faeces. PubMed NCBI (April 2020) was searched for literature published within the last 5 years using the terms “Bacteroides fragilis” AND “healthy”. The resulting literature was screened for any studies that isolated *B. fragilis* from faeces of healthy donors and sequenced the strains[119]. The SRR paired-end reads were download using fastq-dump from sra-toolkit (v. 2.9.6.1) and assembled using SPAdes (v.3.13.1)[120]. The assembly quality of all genomes was assessed with QUAST (v.5.0.2)[121].

4.2.2.2 Quality control

Average nucleotide identity (ANI) between all genomes and that of the reference strain (NCTC 9343^{T}; GCA_000025985) was determined using FastANI (v.1.3) with default settings[110]. Any genomes with an ANI score < 95 % were removed from further analyses. CheckM (v.1.0.18) was used to check the completeness and contamination of the genomes[109]. Any genomes with completeness < 90 % or contamination > 5 % were excluded from further analyses. PanarooQC (v. 1.2.3) was used to identify outliers based on the number of genes and number of contigs in assembled genomes[122].

4.2.3 Collection of *B. fragilis* isolates from NCBI

4.2.3.1 Publicly available genome assemblies

Assembled sequences of *B. fragilis* stored in the NCBI genome database (April 2020) were collected and duplicate strains were removed[123]. Metadata were extracted for each genome from the GenBank database and BioSample entries. This included information about isolation site and host disease. The clinical relevance of the strain was inferred from the metadata and classified as “non-clinical”, “clinical” or “enterotoxigenic”. Non-clinical strains were isolated from faeces of healthy individuals or individuals without inflammatory diarrheal disease. Clinical strains were isolated from blood or soft tissue infections of individuals with bacterial infectious disease. Enterotoxigenic strains were isolated from individuals suffering from inflammatory diarrheal disease. *B. fragilis* strain GB-124 was also included in the pangenome analysis.

4.2.3.2 Genome quality control

Done as described in section 4.2.2.2.

4.2.4 Antimicrobial resistance and *Bacteroides fragilis* toxin

4.2.4.1 AMR gene identification

Abricate (v.0.9.8) was used with the Comprehensive Antibiotic Resistance Database (CARD) (v. 2019-Sep-10) to identify AMR genes[124]. The resulting summary file was used to determine presence of AMR genes (> 75 % percentage identity and > 75 % coverage) and visualised using R (v.3.5.2) and ggplot2 (v.3.3.2).

4.2.4.2 Detection of fragilysin (BFT)

The genomes were searched for the BFT (fragilysin) and toxin-activating protease (fragipain) using blastp (default settings, v.2.10.0)[125]. The fragipain and fragilysin protein sequences were collected from NCBI (accessed December 2020; Table 4.1).

Hits were considered significant if the e-value was > 2e-125 and the percentage identity was < 95 %. The correct isoform was assigned to a positive hit using the lowest e-value and highest percentage identity.

4.2.5 Generation of the pangenome

Prokka (v.1.14.6) was used to annotate all isolates that passed quality control[108]. The resulting .gff files were input to Roary (v.3.13.0; default settings; minimum 95 % percentage identity) for pangenome analysis[126]. A script (*create_pan_genome_plots*) created by Dr Andrew Page was used to generate pangenome overview plots in R (v.3.5.2) using ggplot2 (v.3.3.2). The gene_presence_absence.csv file was used to determine the number of unique genes for each classification (enterotoxigenic, clinical and non-clinical).

Table 4-1: Fragilysin and fragipain protein information from NCBI used to screen genomes for Bft protein

A complete and partial sequence for each bft isoform was used and a complete fragipain sequence.

Protein (isoform)	Protein accession	Length (aa)	Complete or partial
Fragilysin (Bft-1)	KAB5480848.1	397	Complete
Fragilysin (Bft-2)	WP_103483278.1	397	Complete
Fragilysin (Bft-3)	AAD33214.1	397	Complete
Fragipain	AMR55390	393	Complete
Fragilysin (Bft-1)	AAF72830.1	63	Partial
Fragilysin (Bft-2)	AF72838.1	63	Partial
Fragilysin (Bft-3)	AF72839.1	63	Partial

4.2.5.1 Phylogenetic analysis

A core single nucleotide polymorphism (SNP) maximum likelihood tree was generated using IQTree (v. 1.16.10, maximum bootstrap:1000) with default settings and best-fit model determined using ModelFinder[127-129]. The core genome alignment generated by Roary was input to snp-sites (v.2.5.1; default settings) and the resulting phylip file used for phylogenetic analysis with IQTree[126,130]. The resulting tree was visualised using FigTree (v.1.4.4), rooted at the midpoint and bootstrap percentage determined from 999 replicates. The figure was annotated in Adobe Illustrator (v.24.0.6).

4.2.5.2 Core and accessory genes

Core genes were defined as genes present in 99-100 % of isolates and accessory genes all remaining genes. A principal component analysis (PCoA) was performed based on the presence of common (5-95 % prevalence) accessory genes using R (v.3.5.2), FactoMineR (v.3.5.3) and factoextra (v. 3.5.3). Unique genes in each outlying Cluster were determined. Additionally, genes that were present in > 50 % of the main cluster but not present in the outlying Cluster were also determined (labelled “missing” genes). The identity of these genes was determined using blastp (v.2.10.0; default settings) and UniProtKB Bacterial database (accessed 14/04/2021)[125,131]. Blastp hits were considered significant if the e-values were lower than 0.02 at ≥ 40 % protein identity and ≥ 50 % coverage.

4.2.5.3 Cluster analysis

Unique genes from each Cluster identified in 4.2.5.2 were determined and were defined as any gene that was present in all genomes in a Cluster but not present in any other genomes. The coding regions for the unique genes were determined and protein sequence extracted from the Prokka faa files. Blastp (default settings, v.2.10.0) was used to identify the potential function of unique genes[117]. Blastp hits were considered significant if the e-values were lower than 0.02 at ≥ 40 % protein identity and ≥ 50 % coverage.

Missing genes from each Cluster identified in 4.2.5.2 were determined. A missing gene was determined as a gene that was present in > 50 % of all other genomes and not present in any genomes in the cluster. This approach was taken as no genes were missing from one Cluster that were present in all other genomes. The identities of genes that were present in > 50 % of all other genomes were determined using the pan_reference_genome.fa file generated by Roary[126]. The list of gene identities was screened against all genomes from a Cluster to determine which genes were not present in all genomes. Following identification of missing genes, the coding regions were identified, and protein sequences were extracted from faa files. Blastx (default settings, v.2.10.0) was used to identify the potential function of missing genes[125]. Blastx hits were considered significant if the e-values were lower than 0.02 at ≥ 40 % protein identity and ≥ 50 % coverage.

4.2.5.4 rfb gene identification

The gene locations of all *rfb* genes present in the roary gene_presence_absence.csv were selected from each genome. The amino acid sequences for the *rfb* genes from each genome were pulled from the faa files. An amino acid reference sequence for all *rfb* genes present in *B. fragilis* was selected from KEGG Orthologs (Table 4.2)[132].

rfbJ and *rfbX* reference sequences were taken from *Salmonella enterica* serovar *Typhimurium* and *Escherichia coli*[133,134]. The reference sequences were used to build a custom blastp database (v.2.10.0; default settings) and the *rfb* percentage identity for genomes' predicted proteins was determined by comparison to the database[125]. The percentage identity was used to create a heatmap showing the distribution of the genes within the genomes. The top blastp hit for each *rfb* gene was noted. For a few of the *rfb* genes the top blast hit differed among genomes. Therefore, to make visualisation of the heatmap easier, the displayed percentage identity to a blastp hit was kept consistent. The genomes that showed a differing top blastp hit are noted in Appendix 6. For

example, the majority of genomes showed the highest percentage identity to an *rfbG* gene from BOB25. However, six genomes had a higher percentage identity to an *rfbG* gene from 638R. Therefore, the percentage identity to BOB25 was shown on the heatmap and the genomes with a higher percentage identity to 638R *rfbG* gene are shown in Appendix 6. The heatmap was produced in R (v.3.5.2) with ggplot2 (v.3.3.2) and ggdendro (v.0.1.22). The figure was annotated in Adobe Illustrator (v.24.0.6).

The location and orientation of identified *rfb* genes was visualised in each genome using Geneious Prime (v.20.0.5). Clustering of any *rfb* genes was noted for each genome. Three genomes were selected as examples to show the variation in the *rfb* gene clusters. Genes were visualised up- and down-stream of the *rfb* genes and annotated in Adobe Illustrator (v.24.0.6).

4.2.6 Functional analysis of the pangenome

4.2.6.1 Overview of Cluster of Orthologous Groups (COGs)

The pangenome reference fasta file generated by Roary was used to analyse COG data within the core and accessory genome[126]. EggNOG-mapper server (http://eggnog-mapper.embl.de/; default settings) produced an annotation table[135-138]. The COG category for the core and accessory genome was extracted and percentage of genes per COG category was determined. A bar chart was produced using R (v.3.5.2) and annotated in Adobe Illustrator (v.24.0.6). Additionally, the COG categories for the unique and "missing" genes for each outlying Cluster identified in the PCoA were visualised via stacked bar charts using R (v.3.5.2) and annotated in Adobe Illustrator (v.24.0.6).

4.2.7 Analysis of co-evolving genes

Coinfinder (v.1.0.8; default settings) was used to identify associating and dissociating genes within the pangenome[139]. This analysis was run by Dr Maria Rosa Domingo-Sananes at Nottingham Trent University, using the gene_presence_absence.csv file generated from Roary and core gene alignment tree generated with IQTree in 4.2.5. the previous section[126,127].

Table 4-2: Protein information for each *rfb* used to create a blastp database for isolate *rfb* gene screening

For each *rfb* gene, protein sequences from all *B. fragilis* sequences on KEGG were used[132]. Note *rfbJ* and *rfbX* do not originate from *B. fragilis.* These homologs were selected as these *rfbJ* and *rfbX* genes have been characterised in *Salmonella* and *Escherichia*. A homolog belonging to a species more closely related to *B. fragilis* could not be found.

***rfb* gene**	**Species**	**Strain**	**KO**	**KEGG ID**	**UniProt ID**	**Length (aa)**
rfbG	*Bacteroides fragilis*	YCH46	K01709	BF1536	Q6W40	362
		NCTC 9343^{T}	K01709	BF9343_2521	Q5LC64	359
		638R	K01709	BF638R_0780	E1WLJ6	366
			K01709	BF638R_2596	E1WPM6	359
			K01709	BF638R_3484	E1WVN6	373
		BOB25	K01709	VU15_06455	-	359
			K01709	VU15_11520	-	359
			K01709	VU15_16420	-	359
rfbF	*Bacteroides fragilis*	YCH46	K00978	BF1534	Q64W42	270
		NCTC 9343^{T}	K00978	BF9343_2522	Q5LC63	258
		638R	K00978	BF638R_0779	E1WLJ5	277
			K00978	BF638R_2597	E1WPM7	258
			K00978	BF638R_3485	E1WPM7	258
		BOB25	K00978	VU15_06445	A0A7D4JTJ1	269
			K00978	VU15_11525	-	258
			K00978	VU15_16425	A0A0I9S7H2	258
rfbE	*Bacteroides fragilis*	NCTC 9343^{T}	K12454	BF93943_2519	Q5LC66	339
		638R	K12454	BF638R_3482	E1WVN4	336
		BOB25	K12454	VU15_06465	-	337
			K12454	VU15_11510	-	339
			K12454	VU15_16410	-	337
rfbC	*Bacteroides fragilis*	YCH46	K01790	BF0806	Q64Y69	189
			K01790	BF2296	Q64TY6	182
		NCTC 9343^{T}	K01790	BF9343_2302	Q5LCT0	182
			K01790	BF9343_3362	Q5L9T4	180
		638R	K01790	BF638R_0781	E1WLJ7	146
			K01790	BF638R_1545	E1WTC8	195
			K01790	BF638R_2397	E1WNL3	182
			K01790	BF638R_3473	E1WVM5	180
		BOB25	K01790	VU15_03385	-	189
			K01790	VU15_09970	A0A149NKA5	182
			K01790	VU15_16355	-	180
rfbB	*Bacteroides fragilis*	YCH46	K01710	BF0807	Q64Y68	356
			K01710	BF3711	Q64PX7	379
		BOB25	K01710	VU15_03390	-	356
			K01710	VU15_16620	-	379

***rfb* gene**	**Species**	**Strain**	**KO**	**KEGG ID**	**UniProt ID**	**Length (aa)**
rfbM	*Bacteroides fragilis*	YCH46	K00971	BF4322	Q64N77	260
		NCTC 9343^{T}	K00971	BF9343_4017	Q5L801	360
rfbA	*Bacteroides fragilis*	YCH46	k00973	BF0805	Q64Y70	295
			k00973	BF1094	Q64XD2	294
			k00973	BF2583	Q64T46	294
			k00973	BF3664	Q64Q24	297
			k00973	BF3712	Q64PX6	287
		638R	k00973	BF638R_1076	E1WQ49	294
			k00973	BF638R_1454	E1WSQ7	293
			k00973	BF638R_1539	E1WTC2	296
			k00973	BF638R_1864	E1WUX9	297
			k00973	BF638R_3474	E1WVM6	295
		BOB25	k00973	VU15_03380	-	295
			k00973	VU15_04710	Q9RGK4	294
			k00973	VU15_16360	A0A5M5PRV0	295
			k00973	VU15_16625	A0A0K6BXM7	287
rfbJ	*Salmonella enterica* serovar *Typhimurium*	LT2	k12455	STM2089	P0A1P4	299
rfbX	*Escherichia coli*	K-12 MG1655	k18799	B2037	P37745	415

4.2.8 Identification of prophage

A similar method for prophage screening from Crispim *et al.* was followed to identify potential prophage within the *B. fragilis* dataset[140]. PhiSpy (v. 4.2.15; default settings) was used to identify prophage sequences in the assembled genomes in GenBank format[141]. The prophage-like element candidates were manually checked for the presence of an integrase gene and structural viral genes (tail and capsid) using blastp (default settings, v.2.10.0)[125]. Candidates that did not have an integrase gene or that possessed an integrase gene in addition to no genes related to viral structure were classified as degenerate prophages. Candidates were only considered prophage if they contained multiple phage structural genes and an integrase gene.

4.3 Results

4.3.1 Genome characteristics of *B. fragilis* GB-124

Seven contigs > 100 bp in length were assembled from the short- and long-reads of *B. fragilis* GB-124 (N50: 4,986,460 bp). The genome was 99.26 % complete with no contamination and shared 99.08 % ANI with the genome of *B. fragilis* NCTC 9343^{T}. This confirmed *B. fragilis* GB-124 as an

authentic *B. fragilis* strain. Five contigs belonged to the *B. fragilis* GB-124 genome, with a genome size of 5,093,249 bp and GC content of 43.87 %. A total of 4,266 ORFs were predicted, of which 72 were tRNA, one mRNA and two CRISPR repeat regions (Figure 4.1a). Of the five IS elements predicted, only one was identified as belonging to a family (IS1182 family ISBf3). The two remaining contigs were assigned to circular plasmids: PfB1 (4,148 bp) and PfB2 (43,923 bp). PBf1 was a match to *Bacteroides xylanisolvens* strain H207 plasmid unnamed2 (NC_CP041232.1) and *Bacteroides ovatus* strain 3725 D1 iv plasmid unnamed3 (NZ_CP041398.1) according to PLSDB (Figure 4.1b, Appendix 3). PBf2 was a match to *B. ovatus* strain 3725 D1 iv plasmid unnamed2 (NZ_CP041397.1) and *B. thetaiotaomicron* F9-2 plasmid p1-F9 DNA (AP022661.1) according to PLSDB (Figure 4.1c, Appendix 4)[116]. The circularity of the plasmids was confirmed using Bandage (v.0.8.1). It should be noted that the genome of GB-124 comprises eight contigs; however, 3 contigs that are < 100 bp are shown in the Bandage image (Figure 4.2; contig 8,9 and 10). Additional information about the genome of *B. fragilis* GB-124 can be found in Appendix 2.

4.3.2 Selection of *B. fragilis* genomes from literature

PubMed NCBI was searched for any literature published within the last 5 years using the search terms “Bacteroides fragilis” and “healthy”. This search revealed one publication authored by Zhao *et al.* titled “Adaptive Evolution within Gut Microbiomes of Healthy People”[142]. The authors isolated *B. fragilis* strains from 10 individuals from multiple time points (spanning 31 to 709 days). I attempted to assemble genomes of the 601 isolates; assembly attempts from the read files failed for 204 genomes due to un-matched read pairs. A total of 397 isolates were assembled successfully with SPAdes. The average N50 was 205,997 and the number of contigs > 1000 bp in length ranged from 25 to 1777. The authors reported highly intra-individual specific *B. fragilis* populations that were dominated by a single unique lineage and diversified to form coexisting sublineages.

FastANI with the genome of *B. fragilis* NCTC 9343^{T} confirmed the identity of the selected genomes as *B. fragilis* (98-100 % ANI). The completeness and contamination of the isolates was assessed using CheckM and revealed 21 genomes with contamination > 5 % (Table 4.3). These were excluded from further analyses.

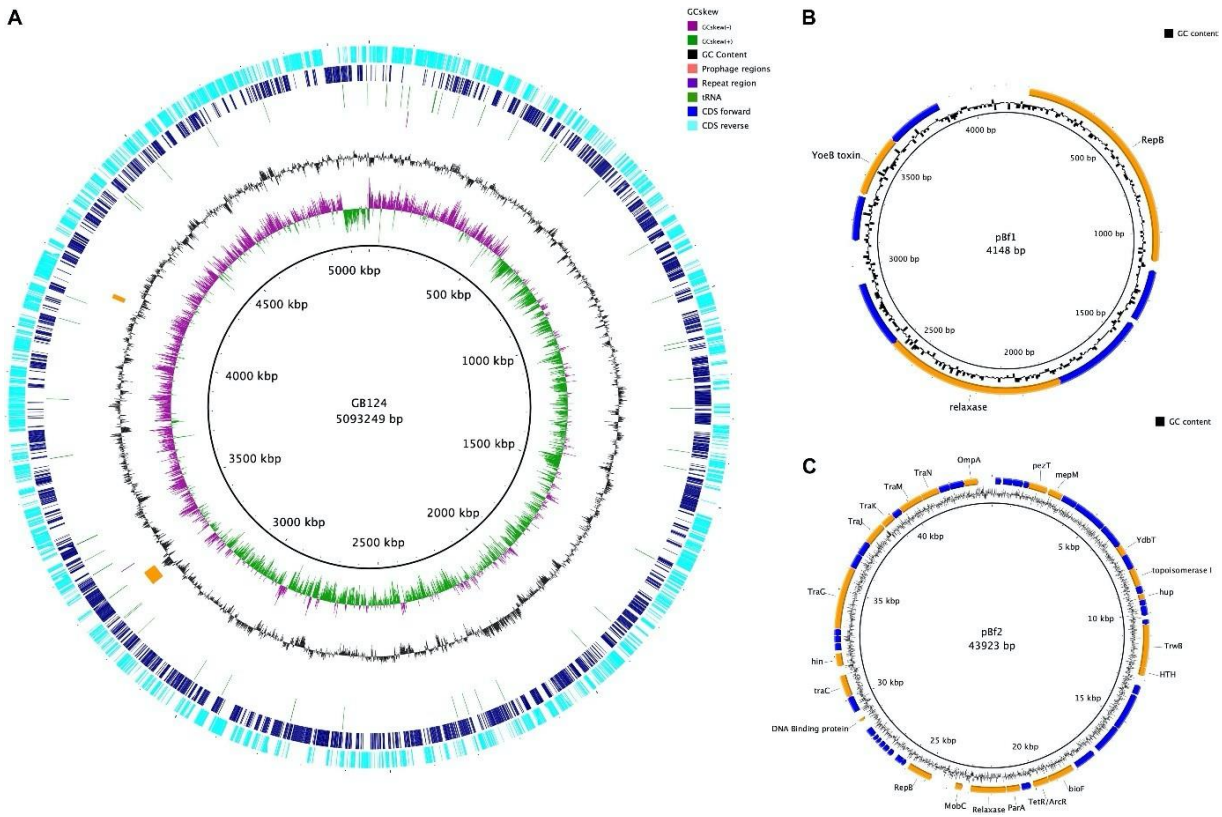

Figure 4.1: Genome map of *B. fragilis* GB-124

A: Chromosome map of GB-124. The outer ring shows the coding sequences in anti-clockwise (aqua) and clockwise (blue) direction. The tRNAs are shown as green arcs and purple arcs depict CRISPR repeat regions. The orange arcs show two predicted prophage regions. The GC content is shown in black and GC skewing by green and purple. B: Map of plasmid pBf1. C: Map of plasmid pBf2. B and C: Hypothetical proteins with no known function shown with blue coding regions and putative function shown with orange coding regions. This figure is reproduced from Tariq et al., 2018 (Appendix 2) under terms of the Creative Commons Attribution License (CC BY) of *Frontiers in Microbiology*.

Figure 4.2: Bandage map of GB-124 chromosome and plasmids

Contigs represented by coloured blocks. Contigs 3 and 5 show circular plasmids pBf1 and pBf2. All remaining contigs represent circular Gb-124 genome. It should be noted that contigs 8,9 and 10 are < 100 bp.

Table 4-3: Non-clinical isolates with CheckM contamination percentage > 5 %

Isolate SRR identification	Contamination (%)
SRR9713631	55.31
SRR9713630	87.05
SRR9713609	87.04
SRR9713516	15.52
SRR9713377	40.86
SRR9713267	87.73
SRR9713266	128.30
SRR9713225	75.66
SRR9713224	43.23
SRR9713781	6.13
SRR9713747	6.13
SRR9713730	10.22
SRR9713713	6.13
SRR9713689	5.95
SRR9713688	17.47
SRR9713557	8.92
SRR9713514	5.95
SRR9713508	6.13
SRR9713505	6.13
SRR9713487	6.12
SRR9713482	8.55

Completeness of genomes ranged from 100 % to 98.05 % according to CheckM; therefore no further genomes were removed. PanarooQC identified 41 genomes that were outliers with the number of contigs and 69 genomes that were outliers with the number of genes; although some genomes were outliers in both categories (Figure 4.3). Following the quality control steps, a total of 273 genomes remained that were suitable for pangenome analysis.

The pangenome was created using Roary with gff files generated with Prokka. This revealed a total of 10,765 genes with 3014 (27.5 %) in the core genome (Table 4.4).

A PCoA plot was generated using the accessory genes (5-95 %) to visualise the clusters. As reported with the original paper, the subjects showed distinct *B. fragilis* populations and formed obvious clusters in the PCoA plot (Figure 4.4).

A core SNP maximum likelihood tree was generated from the core genome to confirm the clusters observed in the PCoA plot (Figure 4.5).

The structure of the resulting tree was consistent with what the authors of the original paper reported: the *B. fragilis* populations from each subject formed distinct clades. A total of 3,451,890 SNPs was reported. For the wider pangenome analysis, it was necessary to randomly select one isolate from each individual lineage (Table 4.5). This removed any bias that would have been introduced using multiple isolates from an individual and likely would have skewed the pangenome results.

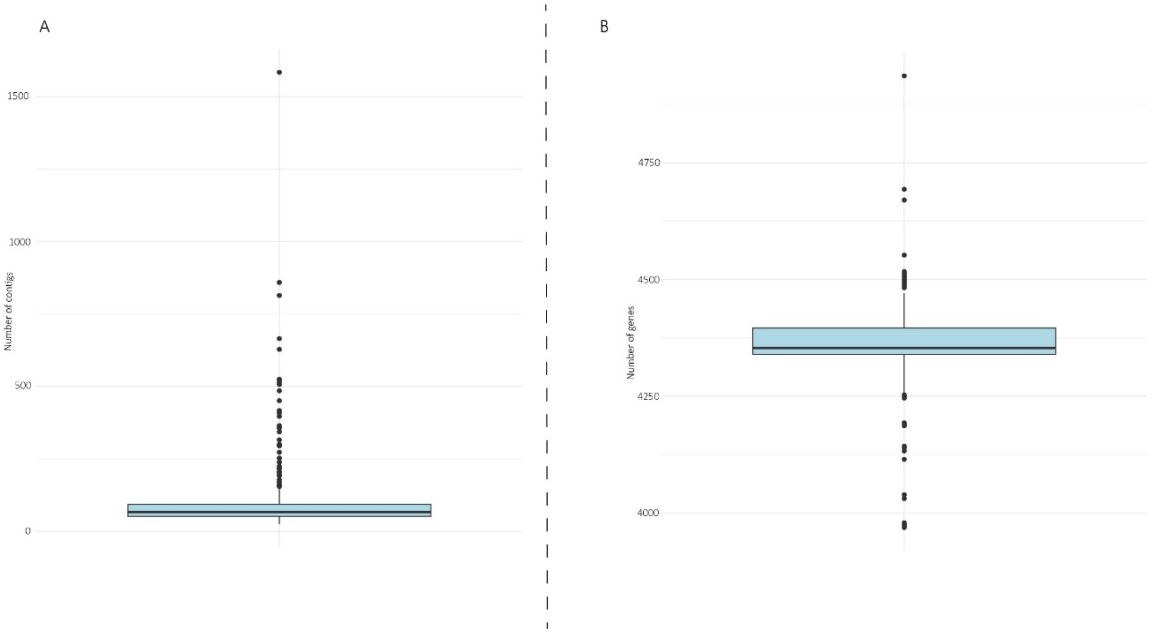

Figure 4.3: PanarooQC output from non-clinical pangenome quality control

A: Boxplot showing number of contigs across all isolates. B: Boxplot showing number of genes across all isolates. The interquartile range of the boxplots is shown by the rectangular box, with the 1st and 3rd quartile being the lower and upper ranges. The median is shown by the black line in the middle of the interquartile range. The maximum upper and lower limit are represented by whiskers. Outliers shown by points at the end of the boxplot whiskers and were excluded from further analysis. The y axis shows A: the number of contigs and B: the number of genes.

Table 4-4: Summary statistics generated from Roary pangenome analysis of non-clinical isolates

Pangenome component*	**Present in strains**	**Number of genes**	**Proportion of genes (%)**
Core	99% <= strains <= 100%	3014	27.5
Soft core	95% <= strains < 99%	46	0.4
Shell	15% <= strains < 95%	2642	25.1
Cloud	0% <= strains < 15%	5063	47
Total	0% <= strains <= 100%	10765	100

* The accessory genome comprises the soft core, shell and cloud pangenome components.

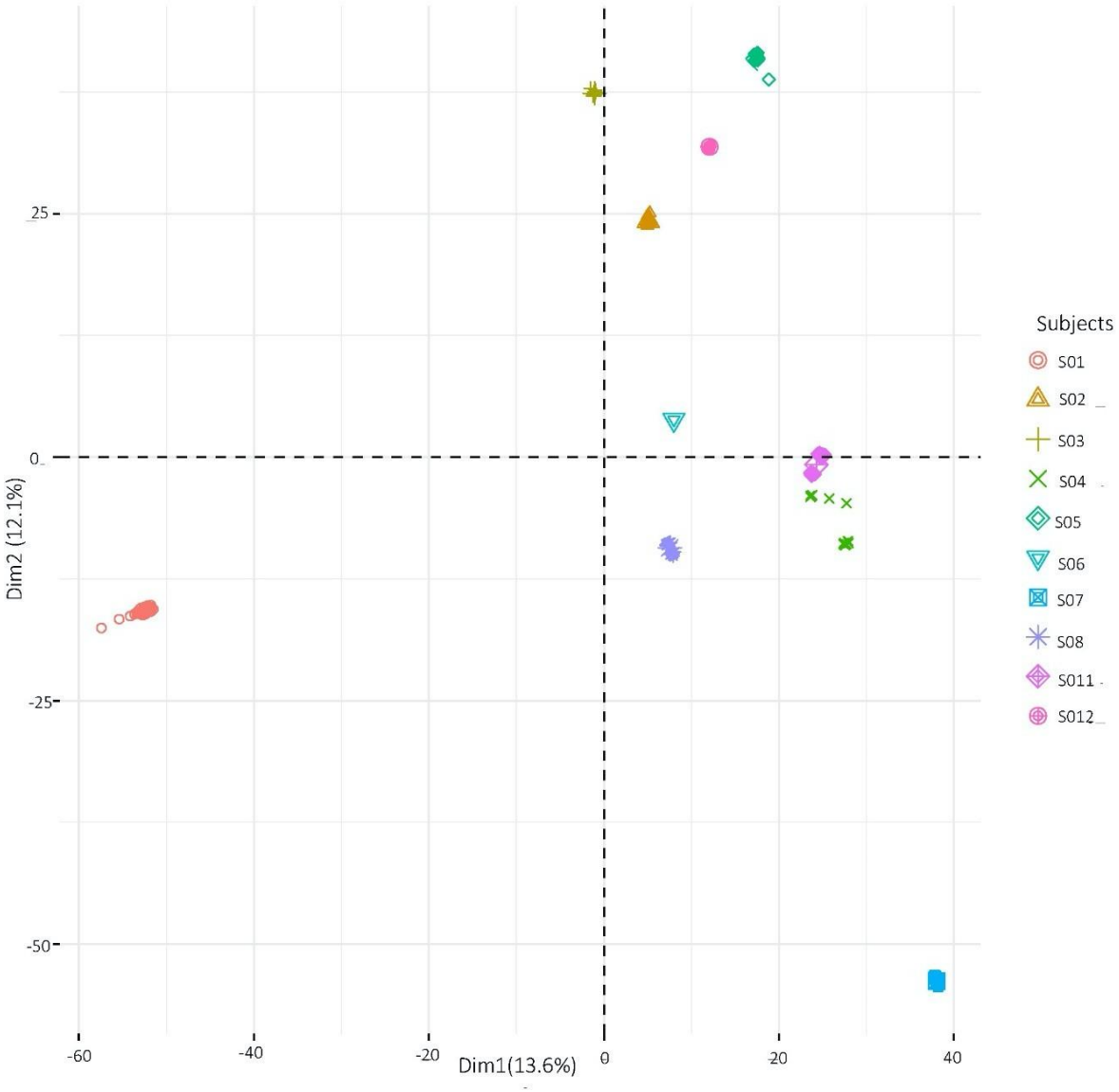

Figure 4.4: PCoA of accessory genome of non-clinical isolates

A PCoA plot was generated from the number of accessory genes detected in each genome to examine the variation of each genome within a subject. As shown by the original publication, the genomes clustered according to sample origin. Dimension 1 (Dim1) explains 13.6 % of the variation within the dataset and Dimension 2 (Dim2) explains 12.1 % of thevariation within the dataset. Each point represents a genome, with points coloured according to subject, as shown in the legend.

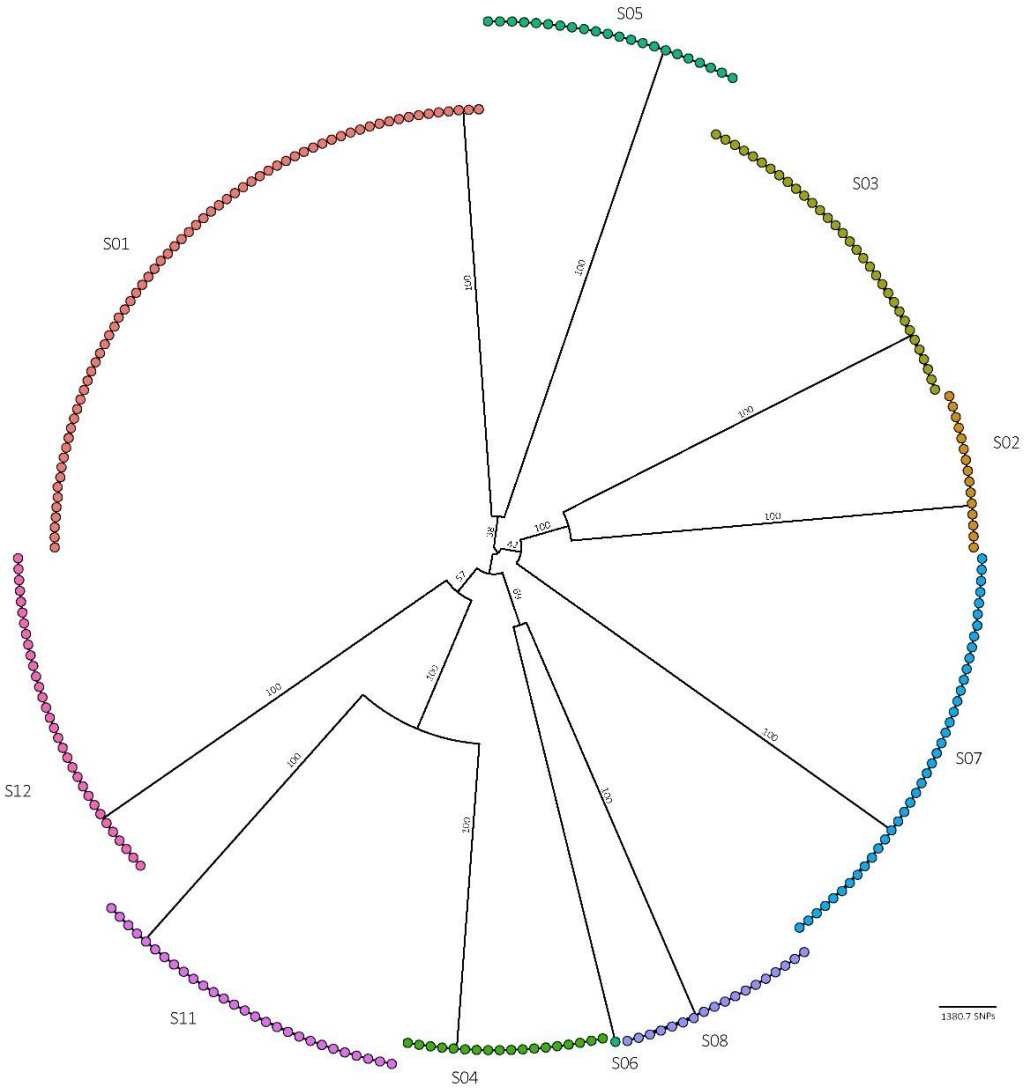

Figure 4.5: Maximum likelihood phylogenetic tree generated from the core SNPs in the pangenome of the non-clinical isolates

The core single nucleotide polymorphisms (SNPs) from each genome were used to generate a phylogenetic tree and confirms conclusions draw from Figure 4.4. Each subject's identifier is displayed around the outside, with corresponding coloured isolate (shown by a circle). IQTree used GTR+F+R4 model and 1001 bootstraps to generate the tree (bootstrap values expressed as a percentage on branches). The bootstrap values are shown on each branch. The SNP scale is shown in the lower right corner. The tree was rooted according tothe midpoint and visuliased in FigTree. Scale indicates nucleotide substitutions.

Table 4-5: Non-clinical isolates selected at random from each clade and subject they originated from

Genome SRR accession	Subject
SRR9713233	S01
SRR9713383	S02
SRR9713745	S03
SRR9713221	S04
SRR9713692	S05
SRR9713457	S06
SRR9713736	S07
SRR9713365	208
SRR9713536	S11
SRR9686280	S12

4.3.3 Collection of *B. fragilis* isolates from NCBI

A total of 116 strains were collected from the NCBI genome database; 25 non-clinical strains, 19 clinical strains and 73 enterotoxigenic strains. It should be noted that some non-clinical strains were isolated from patients in intensive care units and children with cystic fibrosis. However, these genomes were classified as non-clinical as they were not isolated from an individual with inflammatory diarrheal disease or a bacterial infection.

ANI analysis with the reference genome (NCTC 9343^{T}) revealed eight isolates with a score < 95 % (Figure 4.6). Seven of these were clinical and one was enterotoxigenic (Table 4.6).

Five of the genomes originated from the same study and represented multidrug-resistant strains. A blastn search revealed that one of the isolates (3725D9ii) was classified as *Parabacteroides distasonis*. The completeness and contamination of remaining 109 isolates was assessed with CheckM. A total of 10 isolates failed this step of the quality control due to a contamination

percentage > 5 %. The completeness for all isolates was > 90 %. Of the isolates that failed quality control, eight were enterotoxigenic isolates and two were non-clinical isolates (Table 4.7). These were excluded from further analyses.

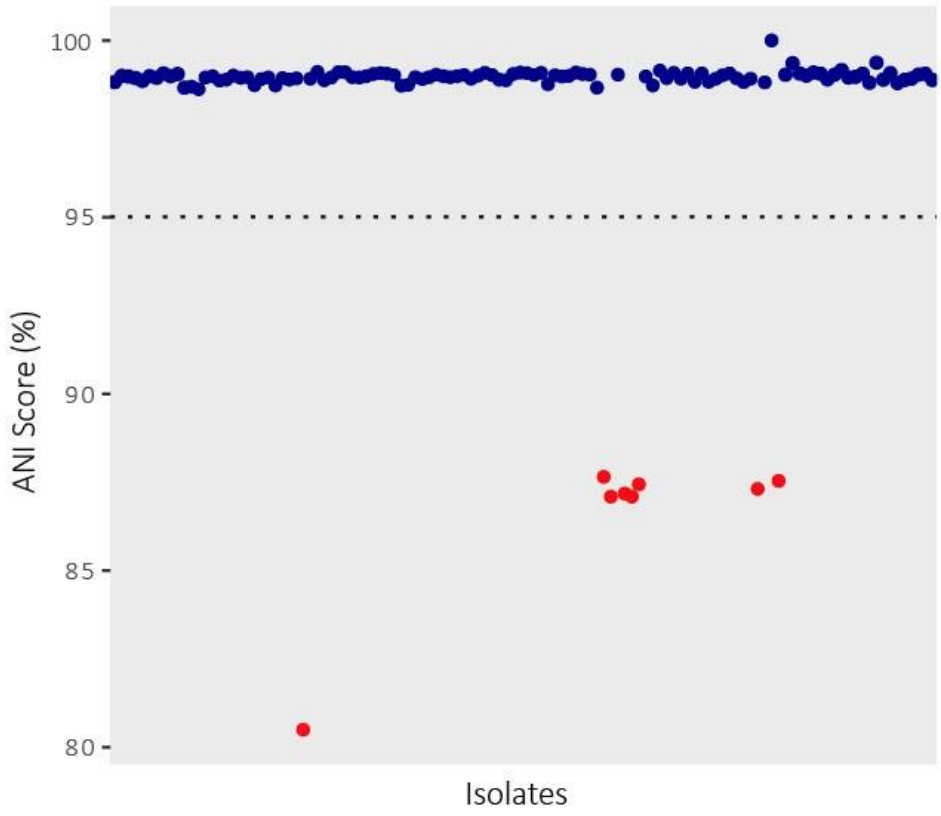

Figure 4.6: ANI scores of all isolates compared to the *B. fragilis* reference genome (NCTC 9343^{T}). The average nucleotide identity percentage for each isolate was determined compared to NCTC 9343. Each dot represents an isolate. All isolates under the 95 % cut-off (dotted line) were excluded from analysis (red dots). The y axis represents the ANI score percentage.

Table 4-6: *B. fragilis* genomes with ANI < 95 % against NCTC 9343^T

Genome	ANI	Classification	Assembly accession
3725D9ii	80 %	Enterotoxigenic	GCF_000699685
DCMOUH0017B*	87.6 %	Clinical	GCF_000710375
DCMOUH0018B*	87 %	Clinical	GCF_000724665
DCMOUH0067B*	87.1 %	Clinical	GCF_000724805
DCMOUH0085B*	87 %	Clinical	GCF_000724815
DCMSKEJBY001B*	87.4 %	Clinical	GCF_000710365
JIM10	87.3 %	Clinical	GCF_001692695
QIF2	87.5 %	Clinical	GCF_002849695

* Genomes originated from the same study.

Table 4-7: CheckM output showing *B. fragilis* NCBI genomes with completeness < 90 % and contamination >5 %

Genome	Completeness (%)	Contamination (%)	Classification	Assembly accession
2d2A	98.96	12.88	Non-Clinical	GCF_000944095
915_BFRA	96.37	10.91	Non-Clinical	GCF_001077245
3783N1-8	100.00	10.64	Enterotoxigenic	GCF_000598605
1009-4-F#7	100.00	6.56	Enterotoxigenic	GCF_000599285
3998TB3	99.82	8.77	Enterotoxigenic	GCF_000598485
3986TB9	92.73	6.46	Enterotoxigenic	GCF_000598465
B1(UDC16-1)	97.52	78.79	Enterotoxigenic	GCF_000598625
S6L3	100.00	55.28	Enterotoxigenic	GCF_000599225
S23L24	100.00	25.89	Enterotoxigenic	GCF_000599305
S38L5	100.00	9.22	Enterotoxigenic	GCF_000599365

PanarooQC reported nine genomes that were outliers in the number of genes and/or contigs (Table 4.8). It should be noted that an additional seven genomes were removed at this stage of analysis due to my incorrect interpretation of the PanarooQC output (which I noticed at a later stage of the analyses).

Table 4-8: Genomes removed from further analysis according to PanarooQC due to number of genes or contigs as outliers

Isolate	Number of genes	Number of contigs	Classification	Assembly accession
AF33-30*	4605	390	Non-Clinical	GCA_003475745
884_BFRA*	4372	359	Non-Clinical	GCA_001058745
3998T(B)4*	4570	359	Enterotoxigenic	GCF_000598385
3988T1*	4404	396	Enterotoxigenic	GCF_000598205
3976T7*	4271	394	Enterotoxigenic	GCF_000598165
3783N2-1*	4264	389	Enterotoxigenic	GCF_000598345
3783N1-2*	4346	412	Enterotoxigenic	GCF_000598325
S36L11	3849	666	Enterotoxigenic	GCF_000599125
AD135F_1B	4953		Non-Clinical	GCF_007896575
S6R5	4987		Enterotoxigenic	GCF_000599045
BFR_KZ02	5005		Clinical	GCA_004798515
JCM11017	5291		Non-Clinical	GCA_000613425
3397T10	6014	2131	Enterotoxigenic	GCA_000598405
DS-233		497	Enterotoxigenic	GCF_000598805
34-F-2#3		651	Enterotoxigenic	GCF_000598425
894_BFRA		992	Non-Clinical	GCF_001058775

* Genomes that were accidentally removed from analysis.

Two genomes (S36L11 and 3397T10) were outliers in both the number of contigs and number of genes (Figure 4.7). A total of 10 enterotoxigenic isolates, one clinical isolate and five non-clinical isolates failed the PanarooQC step.

A total of 93 isolates remained following the above quality control steps and were used for generation of the pangenome. This was curated from NCBI, the selected publication and GB-124. Of these isolates, 29 were classified as non-clinical, 53 as enterotoxigenic and 11 as clinical. The average genome size was 5.3 Mb and GC content was 43.3 % (Table 4.9).

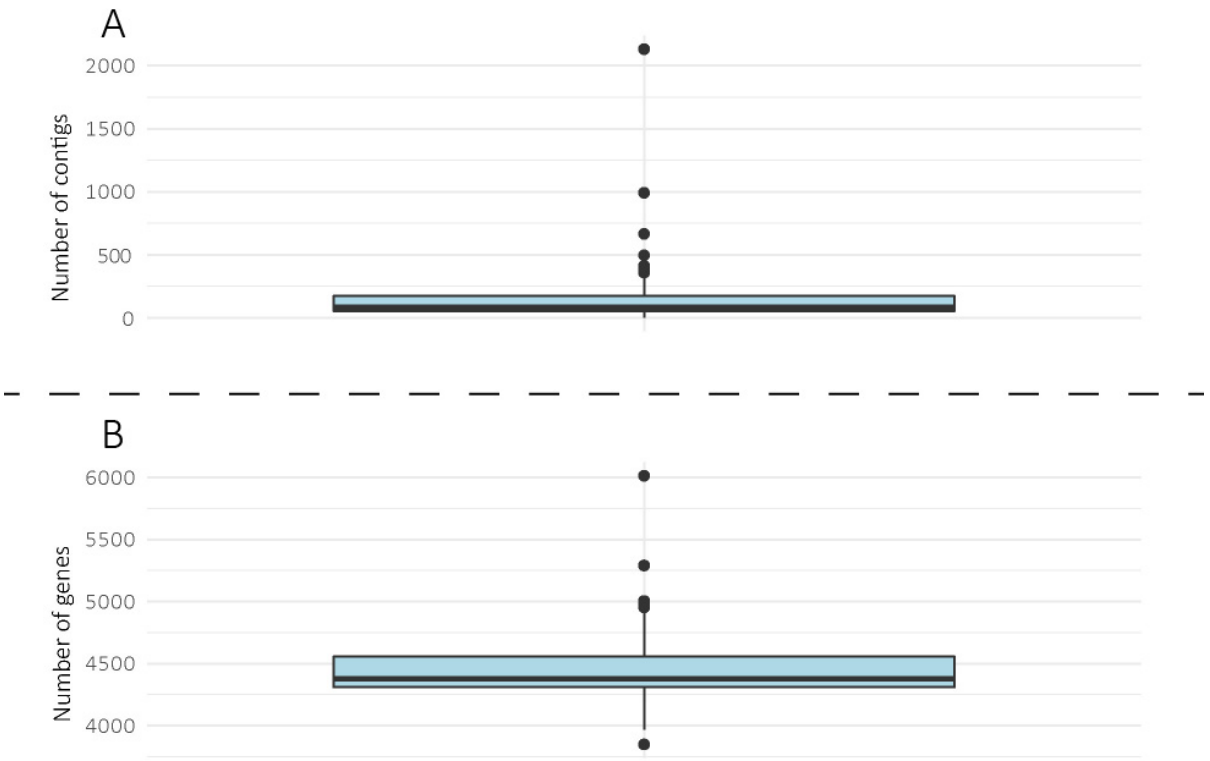

Figure 4.7: PanarooQC output from pangenome quality control

A: Boxplot showing number of contigs across all isolates. B: Boxplot showing number of genes across all isolates. The interquartile range of the boxplots is shown by the rectangular box, with the 1st and 3rd quartile being the lower and upper ranges. The median is shown by the black line in the middle of the interquartile range. The maximum upper and lower limit are represented by whiskers. Outliers shown by points at the ends of the boxplot whiskers and were excluded from further analysis. The y axis shows A: the number of contigs and B: the number of genes.

Table 4-9: Metadata for the 93 *B. fragilis* genomes used in the pangenome analysis

N50 determined using Quast, with CheckM completeness and contamination reported.

Strain	Assembly accession	BioSample	Isolation site*	Host information†	Level‡	BioProject	Size (Mb)	GC%	No. of contigs	No. of CDS	N50	Completeness (%)	Contamination (%)	Reference	Classified as§
1007-1-F #10	GCA_000598685	SAMN02314435	F	IDD, ETBF	Cg	PRJNA206138	5.538	43.2	83	4716	211815	100	0.53		ETBF
1007-1-F #3	GCA_000599265	SAMN02315074	F	IDD, ETBF	Cg	PRJNA206180	5.695	43.2	106	4847	186808	100	1.89		ETBF
1007-1-F #4	GCA_000598545	SAMN02315075	F	IDD, ETBF	Cg	PRJNA206181	5.413	43.0	167	4610	121404	100	0.97		ETBF
1007-1-F #5	GCA_000601035	SAMN02315076	F	IDD, ETBF	Cg	PRJNA206182	5.492	43.3	157	4670	93690	100	0.53		ETBF
1007-1-F #6	GCA_000601095	SAMN02315077	F	IDD, ETBF	Cg	PRJNA206183	5.603	43.2	87	4774	193412	100	0.53		ETBF
1007-1-F #7	GCA_000599145	SAMN02314431	F	IDD, ETBF	Cg	PRJNA206135	5.553	43.2	130	4721	51465	100	0.18		ETBF
1007-1-F #8	GCA_000598265	SAMN02314432	F	IDD, ETBF	Cg	PRJNA206136	5.493	43.2	315	4671	44035	100	0.65		ETBF
1007-1-F #9	GCA_000598885	SAMN02314433	F	IDD, ETBF	Cg	PRJNA206137	5.521	43.2	66	4700	212634	100	0.89		ETBF
1009-4-F #10	GCA_000598705	SAMN02314519	F	IDD, ETBF	Cg	PRJNA206140	5.117	43.2	63	4229	283732	100	0.53		ETBF
20656-2-1	GCA_001699875	SAMN03839335	F	DD	Cg	PRJNA288885	4.896	43.5	68	4087	181621	100	0.35	[30]	ETBF
2-078382-3	GCA_001699865	SAMN03839333	F	DD	Cg	PRJNA288885	5.211	43.1	140	4373	175950	100	0.35	[30]	ETBF
20793-3	GCA_001699855	SAMN03839334	F	DD	Cg	PRJNA288885	5.213	43.2	52	4367	346660	100	0.12	[30]	ETBF
2-F-2 #4	GCA_000598825	SAMN02314421	F	IDD, ETBF	Cg	PRJNA206111	5.577	43.4	213	4660	103460	100	2.98		ETBF
2-F-2 #5	GCA_000598285	SAMN02314526	F	IDD, ETBF	Cg	PRJNA206112	5.774	43.5	250	4824	72386	97.52	4.4		ETBF
2-F-2 #7	GCA_000598145	SAMN02314427	F	IDD, ETBF	Cg	PRJNA206113	5.656	43.5	363	4710	56909	100	0		ETBF
320_BFRA	GCA_001054865	SAMN03197511	-	ICUP	Cg	PRJNA267549	5.500	43.7	150	4419	79562	100	0	[143]	NC
322_BFRA	GCA_001054895	SAMN03197513	-	ICUP	Cg	PRJNA267549	5.450	43.6	193	4386	58847	100	0	[143]	NC
3397 N2	GCA_000598565	SAMN02314521	F	IDD, ETBF	Cg	PRJNA206143	5.117	43.3	93	4292	138404	99.27	0.47		ETBF
3397 N3	GCA_000598925	SAMN02314529	F	IDD, ETBF	Cg	PRJNA206144	5.205	43.3	102	4349	197071	99.48	1.5		ETBF
3397 T14	GCA_000599165	SAMN02314520	F	IDD, ETBF	Cg	PRJNA206142	5.238	43.3	94	4400	167179	100	0.65		ETBF
3719 A10	GCA_000598845	SAMN02314524	F	IDD, ETBF	Cg	PRJNA206150	5.245	43.3	117	4297	135120	100	0		ETBF
3719 T6	GCA_000598725	SAMN02314530	F	IDD, ETBF	Cg	PRJNA206149	4.986	43.2	64	4217	338507	100	0.35		ETBF
3725 D9(v)	GCA_000598585	SAMN02317049	F	IDD, ETBF	Cg	PRJNA206141	5.596	43.2	75	4880	292374	100	0		ETBF
3774 T13	GCA_000598305	SAMN02314525	F	IDD, ETBF	Cg	PRJNA206151	5.422	43.8	340	4596	41659	100	0.58		ETBF
3783N1-6	GCA_000599065	SAMN02314528	F	IDD, ETBF	Cg	PRJNA206153	5.430	43.3	57	4589	204746	100	0.83		ETBF
3976T8	GCA_000599185	SAMN02314538	F	IDD, ETBF	Cg	PRJNA206157	5.408	43.7	73	4500	148019	100	0		ETBF
3986 N(B)19	GCA_000598445	SAMN02317048	F	IDD, ETBF	Cg	PRJNA206120	5.312	44.1	792	4308	23358	99.82	3.51		ETBF
3986 N(B)22	GCA_000598945	SAMN02314545	F	IDD, ETBF	Cg	PRJNA206148	5.051	43.2	31	4184	116865	100	0		ETBF

Strain	Assembly accession	BioSample	Isolation site*	Host information†	Level‡	BioProject	Size (Mb)	GC%	No. of contigs	No. of CDS	N50	Completeness (%)	Contamination (%)	Reference	Classified as§
3986 N3	GCA_000601115	SAMN02314522	F	IDD, ETBF	Cg	PRJNA206147	5.094	43.2	36	4227	544932	100	1.95		ETBF
3986 T(B)13	GCA_000598965	SAMN02314539	F	IDD, ETBF	Cg	PRJNA206146	5.063	43.2	44	4192	565709	100	0		ETBF
3988T(B)14	GCA_000598365	SAMN02314558	F	IDD, ETBF	Cg	PRJNA206158	5.190	43.4	408	4359	46575	97.87	0.53		ETBF
3996 N(B) 6	GCA_000598225	SAMN02314532	F	IDD, ETBF	Cg	PRJNA206114	5.654	43.4	298	4857	71055	99.65	4.34		ETBF
3-F-2 #6	GCA_000598865	SAMN02315072	F	IDD, ETBF	Cg	PRJNA206178	5.300	43.1	201	4495	84284	100	0.3		ETBF
4g8B	GCA_001373095	SAMEA3217991	F	-	Cg	PRJEB8297	4.599	44.9	1402	3547	184948	100	0		NC
638R	GCA_000210835	SAMEA3138381	AA	-	Co	PRJNA50405	5.373	43.4	1	4241	5373121	100	0.18	[54]	C
885_BFRA	GCA_001058755	SAMN03198091	-	ICUP	Cg	PRJNA267549	5.314	43.5	435	4310	42215	100	0	[143]	NC
A7 (UDC12-2)	GCA_000598985	SAMN02314413	F	IDD, ETBF	Cg	PRJNA206105	5.201	43.2	87	4312	378186	100	0		ETBF
AD126T_1B	GCA_007896685	SAMN12414675	F	CF+ child	Cg	PRJNA557692	5.260	43.5	49	4449	488392	100	0		NC
AD126T_2B	GCA_007896675	SAMN12414676	F	CF+ child	Cg	PRJNA557692	5.261	43.5	47	4451	581360	100	0		NC
AD135F_2B	GCA_009024655	SAMN12414678	F	CF+ child	Cg	PRJNA557692	5.615	43.4	101	4773	527045	100	0		NC
AD135F_3B	GCA_007896605	SAMN12414679	F	CF+ child	Cg	PRJNA557692	5.601	43.4	81	4760	581360	100	0		NC
am_0171	GCA_004167855	SAMN10239568	F	Healthy	Cg	PRJNA496358	5.210	43.3	36	4441	122658	90.15	4.32	[144]	NC
BE1	GCA_001286525	SAMEA3494626	B	IAI	Co	PRJEB10044	5.189	43.1	1	4298	5188967	100	0		C
BF8	GCA_001695355	SAMN03921828	I	BI	Cg	PRJNA290835	5.239	43.3	5	4235	301080	100	0	[22]	C
BFR_KZ01	GCA_004798445	SAMN11371866	P	AG	Cg	PRJNA531645	5.553	42.9	840	4211	254934	98.7	0		C
BFR_KZ03	GCA_004798525	SAMN11371868	P	DFP	Cg	PRJNA531645	5.701	42.8	530	4449	415978	100	0.35		C
BOB25	GCA_000965785	SAMN03420872	F	Dysbiosis; ETBF	Co	PRJNA278510	5.282	43.2	1	4137	5282232	100	0.12	[145]	ETBF
CF01-8	GCA_003463555	SAMN09736660	F	-	Cg	PRJNA482748	5.045	43.3	139	3720	48935	100	0.75		NC
CFPLTA004_1B	GCA_007896595	SAMN12414692	F	CF– child	Cg	PRJNA557692	5.684	43.5	64	4976	415978	100	0.35		NC
CL05T00C42	GCA_000269525	SAMN02463923	-	-	Cg	PRJNA64815	5.301	43.5	12	4304	1275491	100	0		NC
DCMOUH0042B	GCA_000724795	SAMN02892979	B	BI	Co	PRJNA253771	5.156	43.4	3	4272	5141257	100	0.35		C
DS-166	GCA_000598245	SAMN02314419	F	IDD, ETBF	Cg	PRJNA206109	5.167	43.4	124	4338	113088	100	1.06		ETBF
DS-208	GCA_000598505	SAMN02314417	F	IDD, ETBF	Cg	PRJNA206107	5.051	43.6	271	4199	56245	100	0.02		ETBF

Strain	Assembly accession	BioSample	Isolation site*	Host information†	Level‡	BioProject	Size (Mb)	GC%	No. of contigs	No. of CDS	N50	Completeness (%)	Contamination (%)	Reference	Classified as§
DS-71	GCA_000599085	SAMN02314418	F	IDD, ETBF	Cg	PRJNA206108	5.039	43.3	308	4141	343583	100	0.35		ETBF
GB-124	GCA_008369705	SAMN12675660	S	-	Cg	PRJNA224116	5.141	43.4	6	4069	5093249	100	0	[146]	NC
GUT04	GCA_008369705	SAMN12675660	F	-	Co	PRJNA563525	5.420	43.1	2	4569	5330584	100	0		NC
HAP130N_1B	GCA_009025695	SAMN12414695	F	CF+ child	Cg	PRJNA557692	5.236	43.3	54	4434	479474	100	0		NC
HAP130N_2B	GCA_007896745	SAMN12414696	F	CF+ child	Cg	PRJNA557692	5.218	43.3	33	4418	600021	100	0		NC
HAP130N_3B	GCA_009025705	SAMN12414697	F	CF+ child	Cg	PRJNA557692	5.225	43.3	42	4427	479474	100	0		NC
HCK-B3	GCA_003363115	SAMN09729823	F	-	Cg	PRJNA483264	5.253	43.4	75	4456	96938	99.13	0.18		C
I1345	GCA_000598785	SAMN02314404	F	IDD, ETBF	Cg	PRJNA206101	5.318	43.5	76	4469	215678	100	0.18		ETBF
ISCST1982	GCA_003852685	SAMN09780485	X	AP	Cg	PRJNA485001	5.206	43.0	458	4630	515154	100	0		C
J-143-4	GCA_000598525	SAMN02314403	F	IDD, ETBF	Cg	PRJNA206102	5.521	43.1	270	4674	76735	98.94	0.89		ETBF
J38-1	GCA_000598645	SAMN02314412	F	IDD, ETBF	Cg	PRJNA206103	5.145	43.4	120	4294	156598	100	0.18		ETBF
Korea 419	GCA_000599205	SAMN02363658	F	IDD, ETBF	Cg	PRJNA206100	7.286	43.0	246	6192	175435	100	0.09		ETBF
NCTC 9343	GCA_000025985	SAMEA1705957	PI	-	Co	PRJNA224116	5.242	43.1	2	4395	5205140	100	0.71		C
S01_NC	-	SAMN12302038	F	-	Cg	PRJNA524913	5.376	43.3	49	4495	479739	98.07	0	[142]	NC
S02_NC	-	SAMN12302209	F	-	Cg	PRJNA524913	5.278	43.3	29	4459	483759	99.72	0.36	[142]	NC
S03_NC	-	SAMN12302231	F	-	Cg	PRJNA524913	5.320	43.1	55	4440	492753	99.94	0.27	[142]	NC
S04_NC	-	SAMN12302305	F	-	Cg	PRJNA524913	5.221	43.3	47	4423	495720	99.78	0.36	[142]	NC
S05_NC	-	SAMN12302338	F	-	Cg	PRJNA524913	5.271	43.5	52	4444	492759	100	0	[142]	NC
S06_NC	-	SAMN12302382	F	-	Cg	PRJNA524913	5.147	43.3	46	4332	490582	99.07	0.37	[142]	NC
S07_NC	-	SAMN12302398	F	-	Cg	PRJNA524913	5.211	43.1	45	4425	500107	100	0.35	[142]	NC
S08_NC	-	SAMN12302444	F	-	Cg	PRJNA524913	5.257	43.4	50	4370	492740	100	0	[142]	NC
S11_NC	-	SAMN12302530	F	-	Cg	PRJNA524913	5.276	43.1	47	4397	492756	98.47	0.77	[142]	NC
S12_NC	-	SAMN12276365	F	-	Cg	PRJNA524913	5.309	43.1	43	4399	486629	98.48	0.48	[142]	NC
S13 L11	GCA_000599105	SAMN02314429	F	IDD, ETBF	Cg	PRJNA206121	4.868	43.3	790	4091	61167	95.12	1.24		ETBF
S14	GCA_001682215	SAMN03921941	-	BI	Co	PRJNA290855	4.902	43.2	1	4059	4902215	100	0.35		C
S23 R14	GCA_000598665	SAMN02314430	F	IDD, ETBF	Cg	PRJNA206122	5.262	43.1	263	4507	85652	97.16	4.59		ETBF
S23L17	GCA_000601055	SAMN02315067	F	IDD, ETBF	Cg	PRJNA206172	5.333	43.4	133	4525	105766	100	0.51		ETBF

Strain	Assembly accession	BioSample	Isolation site*	Host information†	Level‡	BioProject	Size (Mb)	GC%	No. of contigs	No. of CDS	N50	Completeness (%)	Contamination (%)	Reference	Classified as§
S24L15	GCA_000599005	SAMN02314564	F	IDD, ETBF	Cg	PRJNA206166	5.239	43.2	147	4371	198514	100	0.53		ETBF
S24L26	GCA_000598745	SAMN02314565	F	IDD, ETBF	Cg	PRJNA206167	5.274	43.1	72	4408	286209	100	0.18		ETBF
S24L34	GCA_000599325	SAMN02315064	F	IDD, ETBF	Cg	PRJNA206168	5.250	43.1	65	4387	147625	100	0.34		ETBF
S36L12	GCA_000599345	SAMN02363971	F	IDD, ETBF	Cg	PRJNA206171	6.184	43.6	90	5383	260420	100	0.53		ETBF
S36L5	GCA_000599025	SAMN02315065	F	IDD, ETBF	Cg	PRJNA206169	5.744	43.6	105	5007	124735	100	0.35		ETBF
S38L3	GCA_000598765	SAMN02315069	F	IDD, ETBF	Cg	PRJNA206174	4.955	43.2	40	4096	398504	100	1.42		ETBF
S6L5	GCA_000601015	SAMN02363972	F	IDD, ETBF	Cg	PRJNA206161	6.389	42.6	230	5806	164018	100	0.85		ETBF
S6L8	GCA_000599385	SAMN02314560	F	IDD, ETBF	Cg	PRJNA206162	5.226	43.4	100	4462	213485	100	0.35		ETBF
S6R6	GCA_000599245	SAMN02314563	F	IDD, ETBF	Cg	PRJNA206164	5.247	43.5	84	4470	251060	100	0.47		ETBF
S6R8	GCA_000601075	SAMN02314562	F	IDD, ETBF	Cg	PRJNA206165	5.215	43.4	133	4425	131161	100	1.18		ETBF
TL139C_1B	GCA_007896795	SAMN12414700	F	CF+ child	Cg	PRJNA557692	5.254	43.3	26	4520	405616	100	0.38		NC
TL139C_2B	GCA_009025495	SAMN12414701	F	CF+ child	Cg	PRJNA557692	5.276	43.3	53	4532	558150	100	0.35		NC
YCH46	GCA_000009925	SAMD00061068	B	BI	Co	PRJNA13067	5.311	43.2	1	4873	5277274	100	0	[147]	C

* AA, abdominal abscess; AG, acute gangrenous perforated appendicitis; AP, acute appendicitis, perforated. Secondary peritonitis; B, blood; BI, bacterial infection; DFP, diffuse fibrinopuruent peritonitis; F, faeces; I, infection site; IAI, intra-abdominal infection; P, purulent sample; PI, peritoneal infection; S, sewage; X, appendix wall; –, unknown.

† CF–, cystic fibrosis-negative; CF+, cystic fibrosis-positive; DD, diarrheal disease; ETBF, enterotoxigenic *B. fragilis*; ICUP, intensive care unit patient; IDD, Inflammatory diarraheal disease.

‡ Cg, contig; Co, complete.

§ C, clinical; ETBF, enterotoxigenic B. fragilis; NC, non-clinical.

4.3.4 AMR genes and BFT

4.3.4.1 AMR genes

All genomes were screened against CARD using Abricate to search for AMR genes. Only hits > 75 % were considered significant. Ninety-two of the genomes encoded the *cepA* beta-lactamase gene; only am_0171 did not encode this gene (Figure 4.8). am_0171 had the *ErmG* and *Mef(En2)* genes present. Additionally, 56 isolates were positive for *tetQ* and, of those, 16 encoded *Mef(En2)*. DCMOUH0042B, a clinical isolate, encoded the most AMR genes; *EreD*, *ErmF*, *Mef(En2)*, *Oxa-347*, *aads*, *cepA* and *tetQ*. Thirty-two isolates encoded only one AMR gene: *cepA*. The *cepA* gene confers resistance to penicillins and cephalosporins through the production of β-lactamases, and has only been found in *B. fragilis* strains[148].

4.3.4.2 BFT

All isolates were screened for the fragipain protein and BFT protein using blastp. The fragipain protein was present in all genomes. Only 13 of the 53 enterotoxigenic genomes encoded a *bft* gene (Table 4.10). Furthermore, three non-clinical isolates and one clinical isolate were also positive for a *bft* gene (Table 4.10).

Isoform *bft-3* was originally detected in blood isolates from Korea and is detected in a lower proportion of ETBF strains[149]. While rare, it is mainly found in isolates from East Asia. One study detected *bft-3* in two isolates from Great Britain[33]. However, no *bft* gene was detected during the blastp search. BOB25 possessed two copies of the *bft-2* gene. The majority of *bft*-positive isolates possessed *bft-1* (12 isolates), and the remaining five isolates possessed *bft-2*. The current classifications applied to isolates during this study were based on information collected from NCBI. The lack of *bft* gene in 40 “enterotoxigenic” isolates confuses the current classification. However, the original classifications assigned at the beginning of the study will remain.

4.3.5 *B. fragilis* pangenome

Roary analysis revealed a total of 24,471 genes in the pangenome of the 93 genomes. The core genome accounted for 6.42 % (present in 99-100 % of isolates) of the total pangenome and contained 1571 genes (Table 4.11).

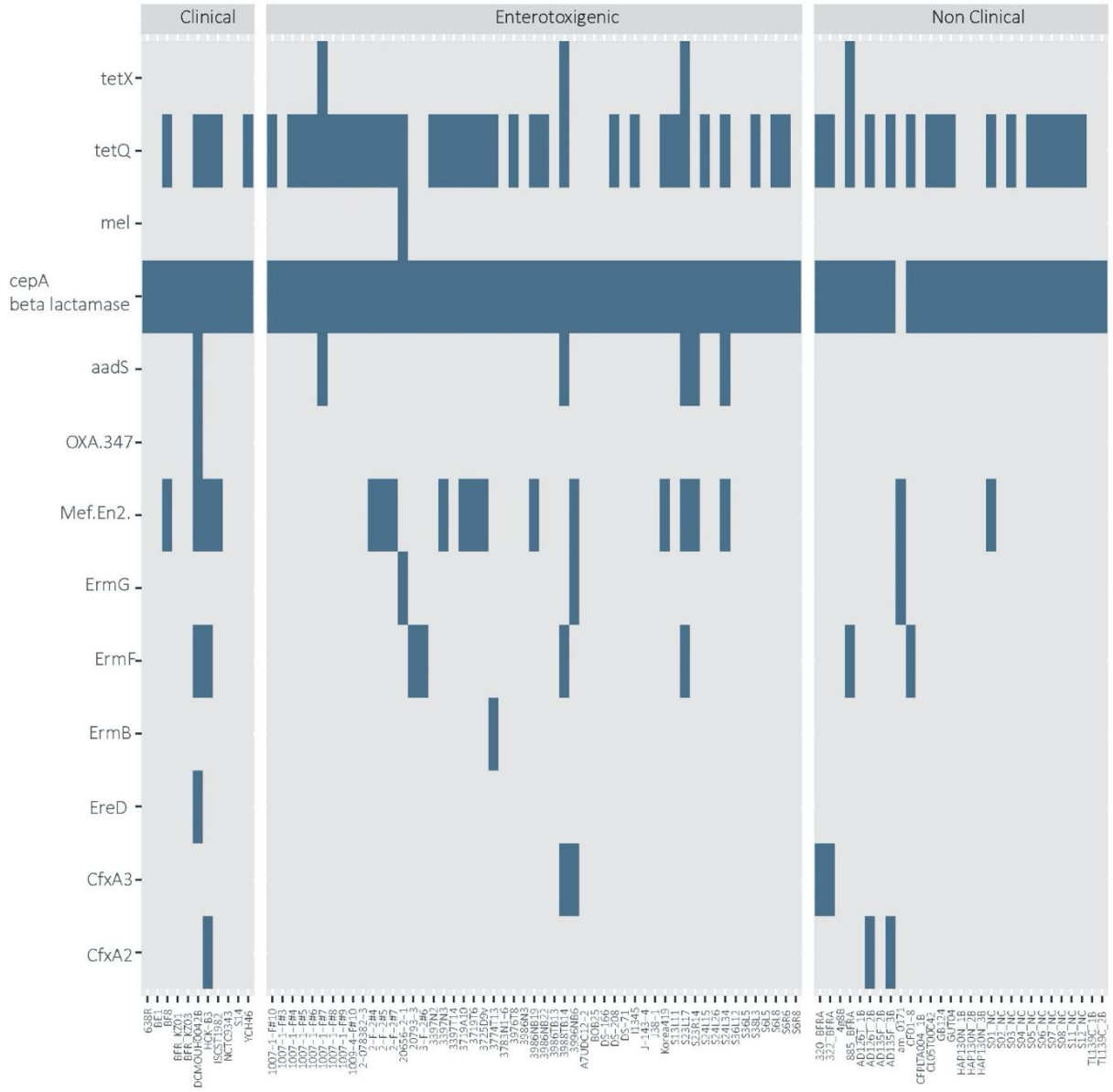

Figure 4.8: AMR gene profile of each genome according to screening against CARD with Abricate

The AMR profile of each isolate was determined using Abricate against CARD with > 75 % percentage identity and > 75 % coverage. The presence of an AMR gene within the isolate was represented by a blue rectangle. The AMR genes shown along the y axis and genome identities on the x axis. The isolates are grouped according to their classification (clinical, enterotoxigenic or non-clinical) and shown at the top of the plot.

Table 4-10: Genomes encoding a *bft* gene and *bft* isotype

Genome	Bft region*	Bft isoform	Classification	Assembly accession
BF8	BF8_02931	bft-2	Clinical	GCA_001695355
2-078382-3	2-078382-3_01925	bft-1	Enterotoxigenic	GCA_001699865
2-F-2#4	2-F-2#4_03712	bft-1	Enterotoxigenic	GCA_000598825
3397N2	3397N2_03946	bft-1	Enterotoxigenic	GCA_000598565
3719A10	3719A10_01627	bft-1	Enterotoxigenic	GCA_000598845
3976T8	3976T8_02608	bft-1	Enterotoxigenic	GCA_000599185
3986NB22	3986NB22_02706	bft-1	Enterotoxigenic	GCA_000598945
DS-166	DS-166_01594	bft-1	Enterotoxigenic	GCA_000598245
J38-1	J38-1_01437	bft-1	Enterotoxigenic	GCA_000598645
S24L15	S24L15_04131	bft-1	Enterotoxigenic	GCA_000599005
20793-3	20793-3_02155	bft-2	Enterotoxigenic	GCA_001699855
20656-2-1	20656-2-1_01566	bft-2	Enterotoxigenic	GCA_001699875
3397N3	3397N3_01537	bft-2	Enterotoxigenic	GCA_000598925
BOB25	BOB25_02978, 03882	bft-2	Enterotoxigenic	GCA_000965785
HAP130N_2B	HAP130N_2B_01190	bft-1	Non_clinical	GCA_007896745
CL05T00C42	CL05T00C42_00174	bft-1	Non_clinical	GCA_000269525
AD135F_2B	AD135F_2B_01438	bft-1	Non_clinical	GCA_009024655

* The genome region in which each *bft* gene was found.

Table 4-11: Summary statistics generated from Roary pangenome analysis of 93 *B. fragilis* genomes

Pangenome component*	Present in strains	No. of genes	Proportion of genes (%)
Core	99% <= strains <= 100%	1571	6.42%
Soft core	95% <= strains < 99%	988	4.04%
Shell	15% <= strains < 95%	2949	12.05%
Cloud	0% <= strains < 15%	18963	77.49%
Total	0% <= strains <= 100%	24471	100%

* The accessory genome comprises the soft core, shell and cloud pangenome components.

A high number of total genes versus a low number of conserved genes was observed from the Roary output, suggesting there are a high number of unique genes and that the *B. fragilis* pangenome is open (Figure 4.9).

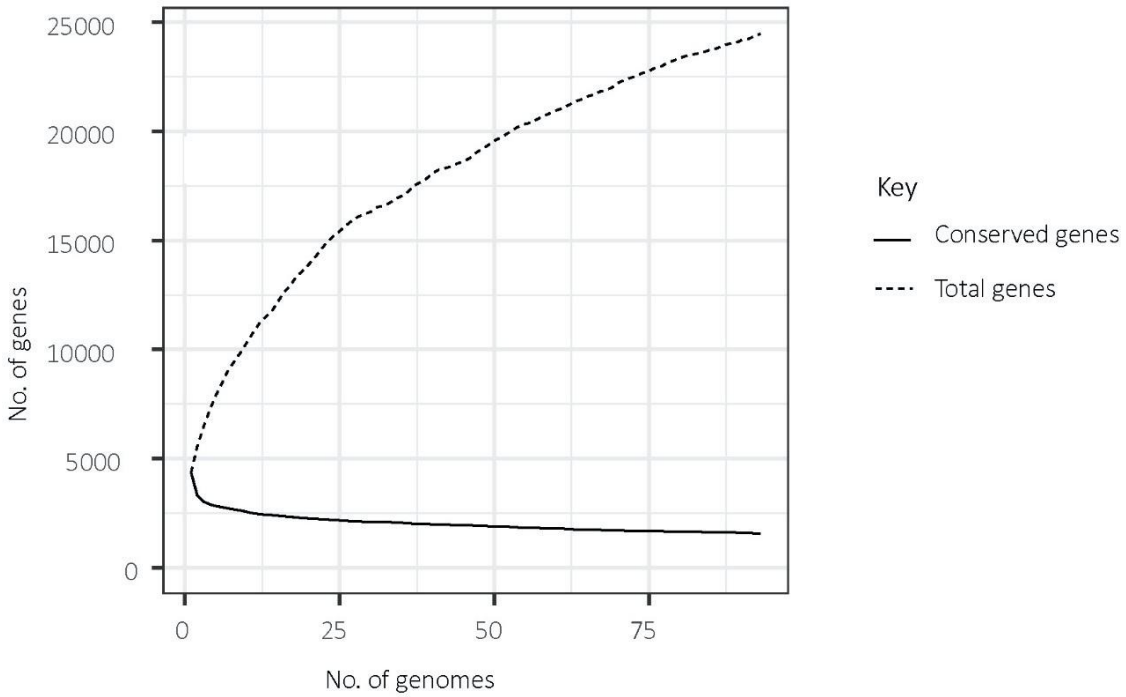

Figure 4.9: Number of conserved genes versus total genes in the *B. fragilis* pangenome
A comparison of conserved genes versus total genes was generated from the Roary output and used to determine the openness of the pangenome. Number of genomes displayed on the x axis and number of genes shown on the y axis.

PCoA was performed using the accessory genes (present in 5-95 % of isolates) to visualise any clusters. A total of 8,157 genes were present in 5-95 % of isolates, suggesting most genes in the pangenome were present in only a few genomes. Seventy-one isolates grouped together in the middle of the PCoA plot and little separation of any classification (i.e. clinical, enterotoxigenic, non-clinical) was observed. However, five outlying clusters and a main cluster were identified away from the main group (Figure 4.10).

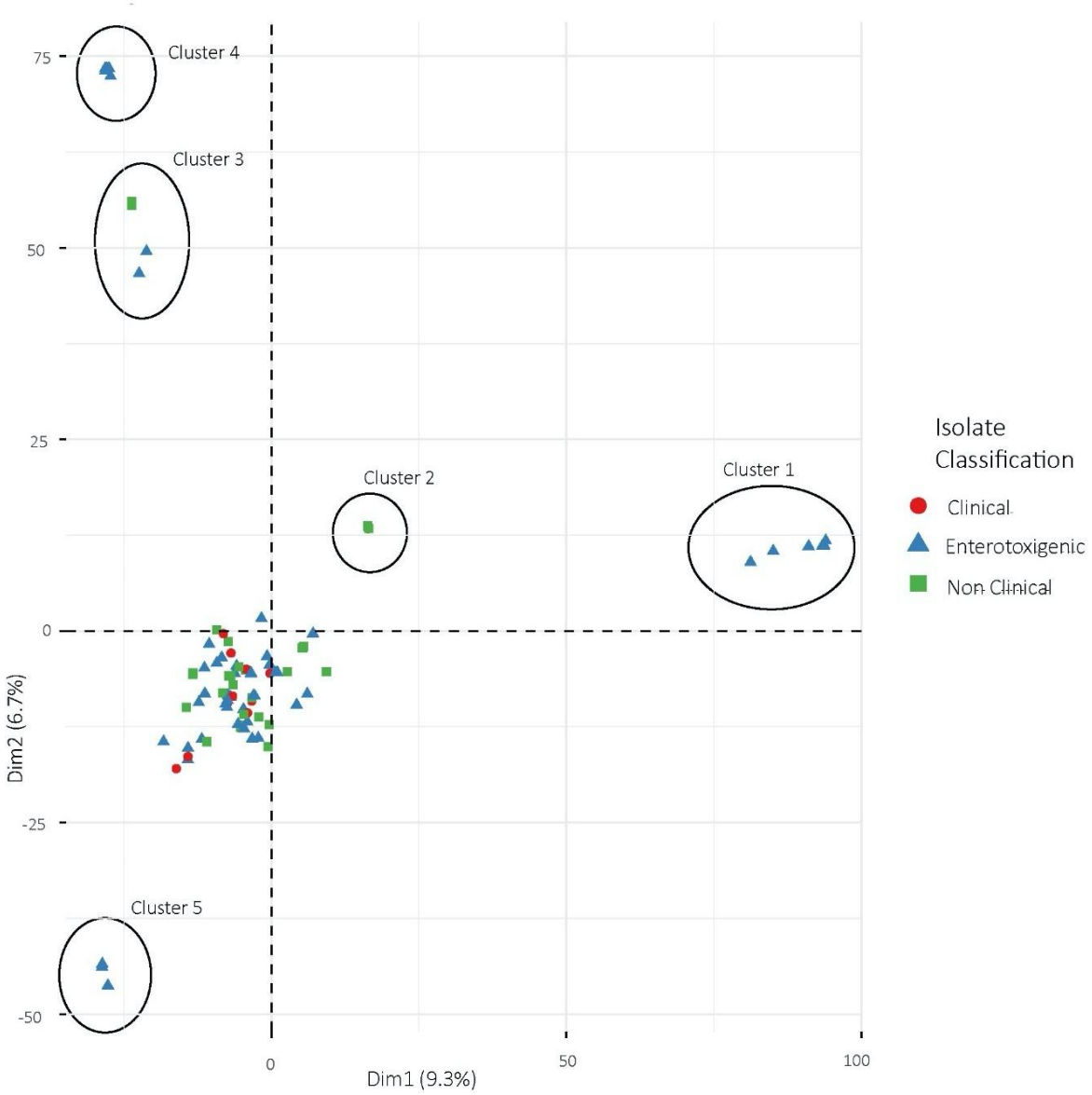

Figure 4.10: PCoA of accessory genome of the 93 *B. fragilis* genomes

A PCoA plot was generated from the number of accessory genes detected in each genome to examine the variation of each genome and determine if genomes clustered according to classification. Generated from the number of accessory genes detected in each genome. Dimension 1 (Dim1) explains 9.3 % of the variation within the dataset and Dimension 2 (Dim2) explains 6.7 % of the variation within the dataset. Each point represents a genome and shows different classifications:blue triangle, enterotoxigenic isolate; red circle, clinical isolate; green square, non-clinical isolate.Each of the five outlier Custers is circled and the main Cluster can be seen within the middle of the plot.

The main Cluster contains 72 isolates and contains a mix of enterotoxigenic, clinical and non-clinical genomes. This Cluster has a large spread and there does not appear to be any sub clustering according to classification within the main Cluster. Additionally, the clinical genomes are scattered within the Cluster and does not suggest these share significant virulence genes. Clusters 1, 4 and 5 contained enterotoxigenic isolates, Cluster 2 contained non-clinical/clinical isolates and Cluster 3 contained enterotoxigenic/non-clinical isolates. Cluster 1 contained six genomes from the same study and the isolates were recovered from human mucosal samples. Furthermore, Cluster 5 contained three genomes from the same study (Table 4.12).

The genomes belonging to Clusters 2 and 4 appeared to be grouped closer together than the other groups. The genomes in Cluster 3 all encoded *bft-1* and all isolates within cluster 5 encoded *cepA, Mef(En2)* and *tetQ*. No other Clusters contained genomes that exhibited a consistent *bft* or AMR gene profile.

A total of 1,645,251 SNPs were reported within the core genome according to snp-sites (Figure 4.11). A core SNP maximum likelihood phylogenetic tree was generated using IQTree and GTR+F+R7 model according to Bayesian information criterion. As seen with the accessory PCoA (Figure 4.10), the classifications did not cluster in the tree. This phylogenetic tree revealed two major clades, one containing 13 genomes. This smaller clade also contained Cluster 5 genomes, with all closely related. Additionally, the clades with multiple isolates showed a high level of relatedness due to the short node lengths. For example, the clade containing S12_NC, 3976T8, 3986NB22 and 3397N2 had similar node lengths. The same is seen with the clade containing Clusters 1 and 2. Clusters 3 and 4 shared a clade (they also clustered closely in the accessory gene PCoA shown in Figure 4.10). However, 1007-1-F#7 grouped with Cluster 4 and appeared to be more closely related to 3783N1-6 than the Cluster it was observed in the PCA. The low discrimination on Dim1 (9.3 %) and Dim2 (6.7 %) suggests that there is little genetic difference in genomes belonging to Cluster 3 and Cluster 4.

Table 4-12: Genomes belonging to each of the five outlying clusters identified in PCoA of accessory genes

Cluster	Isolate	Classification	Assembly accession
1	1007-1-F#10	Enterotoxigenic	GCA_000598685
	1007-1-F#4	Enterotoxigenic	GCA_000598545
	1007-1-F#5	Enterotoxigenic	GCA_000601035
	1007-1-F#6	Enterotoxigenic	GCA_000601095
	1007-1-F#8	Enterotoxigenic	GCA_000598265
	1007-1-F#9	Enterotoxigenic	GCA_000598885
	3397T14	Enterotoxigenic	GCA_000599165
2	CFPLTA004_1B	Non-clinical	GCA_007896595
	BFR_KZ03	Clinical	GCA_004798525
	TL139C_1B	Non-clinical	GCA_007896795
3	1007-1-F#7	Enterotoxigenic	GCA_000599145
	AD135F_2B	Non-clinical	GCA_009024655
	DS-166	Enterotoxigenic	GCA_00059824
	HAP130N_2B	Non-clinical	GCA_007896745
4	S6R8	Enterotoxigenic	GCA_000601075
	S6R6	Enterotoxigenic	GCA_00059924
	S6L5	Enterotoxigenic	GCA_000601015
	1007-1-F#3	Enterotoxigenic	GCA_000599265
5	2-F-2#4	Enterotoxigenic	GCA_000598825
	2-F-2#5	Enterotoxigenic	GCA_000598285
	2-F-2#7	Enterotoxigenic	GCA_000598145

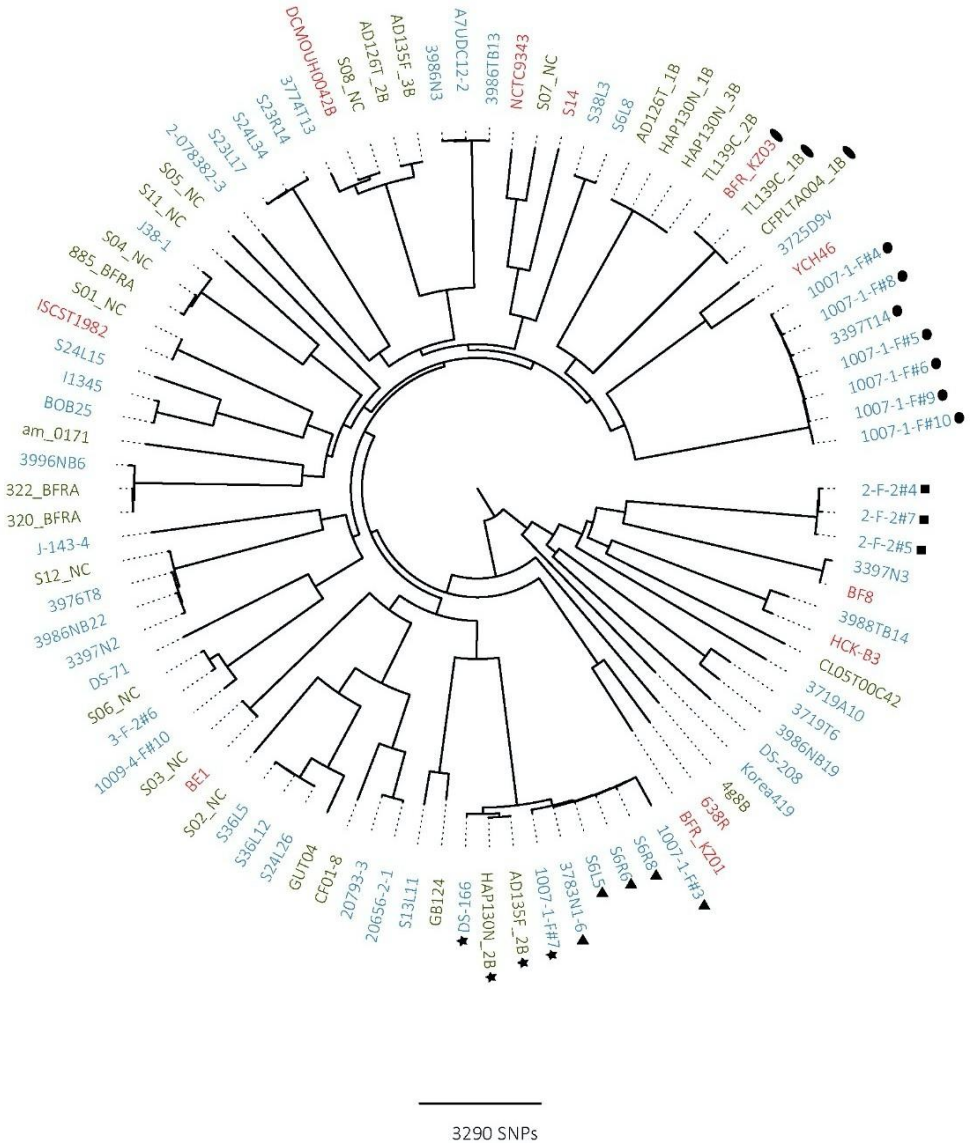

Figure 4.11: Maximum likelihood phylogenetic tree generated from core SNPs from the pangenome analysis

The core single nucleotide polymorphisms (SNPs) from each genome were used to generate a phylogenetic tree. The isolates are coloured according to classification: blue, enterotoxigenic; red, clinical; green, non-clinical. The genomes within each Cluster are shown by the following: Cluster 1, black circles;Cluster 2, black ovals; Cluster 3, black stars; Cluster 4, black triangles; Cluster 5, black diamonds. The tree was rooted at the midpoint and visualised using FigTree. Scale indicates number of SNPs.

4.3.6 Gene cluster analysis

4.3.6.1 Unique genes

Unique genes from each of the five outlying clusters and main cluster were determined and defined as any gene that was present across all isolates in one Cluster but not present in any other genomes. The clusters showed a wide range of unique genes; with Cluster 2 (167 genes) showing the highest number of unique genes and Cluster 3 showing the lowest (two genes). Cluster 1 had 163 unique genes, Cluster 4 had 25 genes and Cluster 5 had 137 unique genes. The main Cluster did not have any genes that were present consistently across all isolates and not present within the outlying clusters. Sixty-two genomes from the main Cluster shared a gene (group_2460). Additionally, 57 genomes from the main Cluster shared three genes with genomes from Cluster 5 and no other outlying Clusters (Table 4.13). It is possible that Cluster 3 and 4 share more unique genes due to the similarity observed in previous sections. However, this was not investigated.

Blastp was used to identify the unique genes within each cluster. Cluster 3 and 4 did not have any significant hits that met the threshold (percentage identity > 40 % and e value < 0.02). Among Clusters 1, 2 and 5, only 14 hits were considered significant (Table 4.14). The percentage identity of the blastp hits for the genes appeared to be relatively low (40 % - 69 %) and only one gene had a 100% percentage identity match (TetO gene; *B.fragilis*).

Table 4-13: Genes identified in the majority of isolates in the main Cluster

All originated from *B. fragilis*, except for 1009-4-F#10_00272 (*Bacteroides* spp.); all had an E value of 0.

Strain and gene location	Gene name	UniProtKB accession	Locus	Predicted product	Identity	Present†
1009-4-F#10_00365	group_2460	Q9XDJ0	F2Z25_15965	PepSY-like domain containing protein	460/460 (100%)	62
1009-4-F#10_00276	sigW_1*	A0A2K9H6L1	BUN20_03785	RNA polymerase sigma70 factor	180/180 (100%)	57
1009-4-F#10_00275	group_11926*	A0A015YH47	M076_0988	FecR family protein	382/382 (100%)	57
1009-4-F#10_00272	group_11924*	A0A372UVH7	DW640_10495	RagB/SusD family nutrient uptake outer membrane protein	614/614 (100%)	57

* Genes that were present in all isolates within outlying Cluster 5.

† Number of isolates within the main cluster the gene was present in.

Table 4-14: Unique genes identified in clusters 1,2 and 5 according to Blastp analysis

Cluster	Strain and gene location	Gene name	UniProtKB accession	Locus	Predicted product	Species	E Value	Identity (%)
1	1007-1-F#10_01722	ubiE_2	A6L3D5	MENG_BACV8	Demethylmenaquinone methyltransferase	*P. vulgatus* (strain ATCC 8482 / DSM 1447 / JCM 5826 / NBRC 14291 / NCTC 11154)	7.00E-69	96/238 (40%)
	1007-1-F#10_01244	group_10025	Q7A029	NREC_STAAW	Oxygen regulatory protein NreC	*Staphylococcus aureus* (strain MW2)	2.00E-09	25/61 (41%)
2	BFR_KZ03_02639	group_17607	P44189	Y1418_HAEIN	Uncharacterised protein HI_1418	*Haemophilus influenzae* (strain ATCC 51907 / DSM 11121 / KW20 / Rd)	3.00E-35	43/93 (46%)
	BFR_KZ03_02567	group_17563	P04043	MTD21_STREE	Modification methylase DpnIIA	*Streptococcus pneumoniae*	1.00E-10	27/50 (54%)
	BFR_KZ03_02554	group_17550	Q5WAX6	TOP3_BACSK	DNA topoisomerase 3	*Bacillus clausii* (strain KSM-K16)	1.00E-06	19/36 (53%)
	BFR_KZ03_01473	group_17496	Q32J95	YLBG_SHIDS	Uncharacterised protein YlbG	*Shigella dysenteriae* serotype 1 (strain Sd197)	3.00E-39	65/121 (54%)
5	2-F-2#4_04711	group_13007	P54992	YSNA_STRPR	Putative transposase in snaA-snaB intergenic region	*Streptomyces pristinaespiralis*	1.00E-181	157/373 (42%)
	2-F-2#4_04432	group_12926	B2RLI7	LPXE_PORG3	Lipid A 1-phosphatase	*Porphyromonas gingivalis* (strain ATCC 33277 / DSM 20709 / CIP 103683 / JCM 12257 / NCTC 11834 / 2561)	3.00E-10	32/77 (42%)
	2-F-2#4_04428	ltrA_1	P0A3U1	LTRA_LACLM	Group II intron-encoded protein LtrA	*Lactococcus lactis* subsp. *cremoris* (strain MG1363)	1.00E-145	246/569 (43%)
	2-F-2#4_04368	group_12898	Q6T1W6	FDTB_ANETH	dTDP-3-amino-3,6-dideoxy-alpha-D-galactopyranose transaminase	*Aneurinibacillus thermoaerophilus* (ATCC 700303)	4.00E-26	45/85 (53%)
	2-F-2#4_03529	fim1C_1	A7UZ95	FIM1C_BACUC	Putative fimbrium subunit Fim1C	*Bacteroides uniformis* (strain ATCC 8492 / DSM 6597 / CIP 103695 / JCM 5828 / NCTC 13054 / VPI 0061)	3.00E-108	169/303 (56%)
	2-F-2#4_02191	tetO	P70882	TETQ_BACFR	Tetracycline resistance protein TetO	*Bacteroides fragilis* (strain YCH46)	1.00E-112	158/158 (100%)
	2-F-2#4_01685	group_12548	P61888	RMLA2_SHIFL	Glucose-1-phosphate thymidylyltransferase 2	*Shigella flexneri*	4.00E-146	200/289 (69%)
	2-F-2#4_01098	group_12516	Q9JWD6	VSR_NEIMA	Putative very short patch repair endonuclease	*Neisseria meningitidis* serogroup A / serotype 4A (strain DSM 15465 / Z2491)	5.00E-38	63/121 (52%)

4.3.6.2 'Missing' genes

'Missing' genes from each Cluster were determined and defined as any gene not present in a Cluster but present in at least 50 % of all other isolates. These genes are referred to as 'missing' genes as they may be present, but the sequences differ enough to be classified as a different gene (according to Roary). Overall the clusters showed more 'missing' genes than unique genes; Cluster 1 had 135 'missing' genes, Cluster 2 had 144, Cluster 3 had 176, Cluster 4 had 169 and Cluster had 435 'missing' genes. No 'missing' genes were discovered within the main Cluster. To determine the identity of these genes, the nucleotide sequence was extracted from the pan-reference file and Blastx (default settings, v.2.10.0) used to determine the proposed identity. Hits were considered significant if the percentage identity was > 40 % and e-value < 0.02. As seen with the unique genes, very few of the genes produced significant hits; Cluster 1 had 22 'missing' genes with significant hits using Blastx, Cluster 2 had 21, Cluster 3 had 20, Cluster 4 had 40 and Cluster 5 had 43. The complete Blastx results for each Cluster can be found in Appendix 5.

Within Cluster 1, five of the 'missing' genes showed 99-100 % identity to genes from *B. fraglis* strain YCH46. These genes were chaperone protein DnaJ, CTP synthase, DNA mismatch repair protein MutL, polyribonucleotide nucleotidyltransferase *pnp* and diaminopimelate epimerase *dapF*. Cluster 2 also showed 99-100 % identity to genes identified in *B. fragilis* YCH46 and *B. fragilis* NCTC 9343^{T}. These genes included DnaJ, MutL, imidazole glycerol phosphate synthase subunit HisH, aspartate carbamoyltransferase *pyrB,* elongation factor *tufA* and putative membrane protein insertion efficiency factor. Additionally, 'missing' genes within Cluster 3 showed 99-100 % identity to *B. fragilis* YCH46 and *B. fragilis* NCTC 9343^{T}. These genes included *mutL, pnp, hisH*, riboflavin biosynthesis protein RibBA, glycine cleavage system H protein, glyceraldehyde-3-phosphate-dehydrogenase *gapA* and flavodoxin. There was also a 53 % identity hit to putative fimbrium anchoring subunit Fim4B from *Bacteroides ovatus* NCTC 11153^{T} and an 85 % identity hit to a probable butyrate kinase from *B. thetaiotaomicron* NCTC 10582^{T}. These two genes were also not present in Cluster 4, along with *mutL, pnp*, uracil-DNA glycosylase 1 *ung_1* and 7alpha-hydroxysteroid dehydrogenase from *B. fragilis* (YCH46/NCTC 9343^{T}). Cluster 5 'missing' genes showed significant hits to six *B. fragilis* genes including *dnaJ, dapF, mutL, pnp,* sialidase and CTP synthase. Additionally, a 40 % identity hit to TonB-dependent receptor SusC from *B. thetaiotaomicron* NCTC 10852^{T} was also noted.

It was noted that several of the clusters appeared to be missing similar genes (Table 4.15). For example, none of the five clusters possessed the same *mutL, rfbE* and *rfbG_2* genes that were present in > 50 % of the other isolates.

4.3.7 *rfb* gene analysis

It was noted in the previous section that all clusters did not possess the same *rfbE* and *rfbG_2* gene that is present in >50% of the other isolates. The amino acid sequences for all *rfb* genes were extracted from the isolates and identification predicted using a custom blastp created with *rfb* reference genes. The *rfb* profiles of each isolate was visualised with a heatmap showing the percentage identity to the top blastp hit (Figure 4.12). The heatmap revealed a high level of *rfb* gene variability within the isolates and no grouping according to isolate classification was noted. Interestingly, the previous clusters (1-5) observed in the PCA were also observed in the *rfb* heatmap; suggesting these genes contribute to the isolate clustering. Cluster 3 and 4 are also integrated on the heatmap, as seen previously with the accessory maximum likelihood tree.

All isolates showed a low percentage identity to *rfbJ* and *rfbX,* suggesting these genes are not present in *B. fragilis* or they no have homology to currently identified *rfbX and rfbJ* genes. All isolates possess the *rfbM* as there was high percentage identity to the *rfbM* gene from *B. fragilis* NCTC 9343^{T}. A total of 11 *rfbG* genes were identified across all isolates, with the majority possessing *rfbG_2*. Isolates within Cluster 1 did not appear to possess a *rfbG* gene and Cluster 3/4 isolates had a *rfbG_2* gene (rfbG_2_group_11059) not present in other isolates. As seen with the *rfbG* genes, the *rfbF* genes were highly variable across all samples. The majority of isolates possessed two *rfbF* genes; however, Cluster 1 did not have any *rfbF* genes. Similar to *rfbG*, isolates in Cluster 3 and 4 possessed an *rfbF* gene not seen in other isolates. However, this gene (rfbF_2_group_11061) has a low percentage identity compared to the *rfb* reference gene. A total of three *rfbE* genes were noted, with most isolates possessing *rfbE*. It was also noted that none of the isolates within Clusters 1-5 encoded an *rfbE* gene. Isolates BOB25 and I1345 encoded one copy of each *rfbE* gene. Nineteen *rfbC* genes were identified across all isolates and *rfbC_2* was present in all except Korea419 and 2-078382-3. Cluster 1 and 2 isolates had three versions of an *rfbC* gene: *rfbC_2, rfbC_2_group_3517* and *rfbC_2_10570.* Additionally, only isolates in Clusters 3 and 4 had *rfbC_1_group_10942.* Most isolates possessed an *rfbC_2* gene and few had an *rfbC_1* gene. Overall, few of the isolates had an *rfbB* gene and none of the isolates from Clusters 1-5 encoded an *rfbB* gene. A similar profile to the *rfbB* gene was noted for the *rfbA* gene. Due to the complexity of the *rfb* gene analysis, these genes were not extensively examined within the main Cluster. However, there appears to be a high level of diversity of *rfb* genes within the main cluster.

Table 4-15: Overview of 'missing' genes in each Cluster and the overlap between clusters

A 'missing' gene is defined as a gene not present in all isolates in a Cluster but present in > 50 % of all other isolates.

			Clusters				
Gene ID	**Blastx ID**	**Gene name**	**1**	**2**	**3**	**4**	**5**
nqrE	A5UFX2	Na(+)-translocating NADH-quinone reductase subunit E	X		X		X
capD_1	A8GRN9	UDP-glucose 4-epimerase	X				X
bioF_2	B0K590	8-amino-7-oxononanoate synthase					X
ribB	B1KNY2	3,4-dihydroxy-2-butanone 4-phosphate synthase	X			X	X
nth	O05956	Endonuclease III					X
mro_3	P05149	Aldose 1-epimerase	=				X
pabA	P06194	Aminodeoxychorismate synthase component 2					X
group_10882	P0A9L4	FKBP-type 22 kDa peptidyl-prolyl cis-trans isomerase					X
rfbE	P14169	CDP-paratose 2-epimerase	X	X	X	X	X
group_3316	P22036	Magnesium-transporting ATPase, P-type 1	X		X	X	X
asnB	P22106	Asparagine synthetase B [glutamine-hydrolysing]	X				X
group_873	P25906	Pyridoxine 4-dehydrogenase					X
rfbC_4	P26394	dTDP-4-dehydrorhamnose 3,5-epimerase	X	X			X
rfbG_2	P26397	CDP-glucose 4,6-dehydratase	X	X	X	X	X
spnQ	P26398	LPS biosynthesis protein RfbH	X				X
group_2916	P31206	Sialidase					X
group_2225	P33363	Periplasmic beta-glucosidase					X
group_3517	P37780	dTDP-4-dehydrorhamnose 3,5-epimerase	X				X
dbpA_2	P50729	Probable ATP-dependent DNA helicase RecS					X
rffH	P55255	Glucose-1-phosphate thymidylyltransferase	X	X			X
yhgF	P71353	Uncharacterised protein HI_0568					X
group_7966	P94519	Uncharacterised protein YsdA	X				X
atoC_2	Q06065	Regulatory protein AtoC					X
patB_2	Q08432	Cystathionine beta-lyase PatB		X	X	X	X
nqrB	Q1QX85	Na(+)-translocating NADH-quinone reductase subunit B					X
metH_2	Q24SP8	Corrinoid protein DSY3155					X
yqhD	Q46856	Alcohol dehydrogenase YqhD	X				X
bacC	Q56318	Uncharacterised oxidoreductase TM_0019					X
dnaJ	Q5LED4	Chaperone protein DnaJ	X	X			X
pnp	Q64N73	Polyribonucleotide nucleotidyltransferase	X		X	X	X
mutL	Q64NX1	DNA mismatch repair protein MutL	X	X	X	X	X
dapF	Q64SY7	Diaminopimelate epimerase	X				X
group_5173	Q64T27	CTP synthase	X				X
msbA	Q6AJW3	ATP-dependent lipid A-core flippase					X
apgM	Q74C57	Probable 2,3-bisphosphoglycerate-independent phosphoglycerate mutase					X
ppk_1	Q87S51	Polyphosphate kinase					X
group_7349	Q8A1G1	TonB-dependent receptor SusC					X
rfbF	Q8Z5I4	Glucose-1-phosphate cytidylyltransferase					X
rluA	Q8ZIK1	Dual-specificity RNA pseudouridine synthase RluA			X	X	X
yknY_4	Q92NU9	Macrolide export ATP-binding/permease protein MacB					X
glyD	Q9AEU2	Probable glycosyl transferase Gly					X
wbjC	Q9XC60	UDP-2-acetamido-2,6-beta-L-arabino-hexul-4-ose reductase					X
arnC_1	A0A0H2UR96	Glycosyltransferase GlyG		X			
group_3542	A0QV10	Uncharacterised oxidoreductase MSMEG_2408/MSMEI_2347		X			
group_6783	A7LXW1	Putative fimbrium anchoring subunit Fim4B			X	X	
ravA_2	B1LL73	ATPase RavA				X	
rnpA	B2RHI3	Ribonuclease P protein component		X			
group_1052	D5EV35	Acetylxylan esterase		X			
ald	E1V931	Alanine dehydrogenase	X			X	
yhgF_1	O31489	Uncharacterised protein YdcI			X		
yvgN	O32210	Glyoxal reductase		X			
rhaS_5	O34901	Uncharacterised HTH-type transcriptional regulator YobQ		X			
group_10884	P08696	Bacteriocin BCN5	X				
hup_3	P0A3H0	DNA-binding protein HU			X	X	
tufA	P33165	Elongation factor		X			
group_2802	P37515	Probable maltose O-acetyltransferase				X	

			Clusters				
Gene ID	**Blastx ID**	**Gene name**	**1**	**2**	**3**	**4**	**5**
group_10485	P40761	Uncharacterised protein YuxK				X	
group_1691	P71052	Probable polysaccharide biosynthesis protein EpsC			X	X	
group_10481	P9WNP3	3-hydroxyacyl-thioester dehydratase Z				X	
mdtB_1	Q48815	Protein HelA		X			
sufC	Q55791	Probable ATP-dependent transporter slr0075		X	X	X	
gapA	Q59199	Glyceraldehyde-3-phosphate dehydrogenase			X		
group_24443	Q5L7W4	Glycine cleavage system H protein			X		
group_321	Q5LA59	7alpha-hydroxysteroid dehydrogenase				X	
ung_1	Q5LA67	Uracil-DNA glycosylase 1				X	
btuD_3	Q5WNX0	Bacitracin transport ATP-binding protein BcrA		X			
group_24261	Q64N34	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (flavodoxin)			X		
hisH	Q64RT0	Imidazole glycerol phosphate synthase subunit HisH		X	X		
pyrB	Q64U74	Aspartate carbamoyltransferase		X			
ribBA	Q64YT3	Riboflavin biosynthesis protein RibBA			X		
group_24135	Q650K9	Putative membrane protein insertion efficiency factor		X			
group_23978	Q8A4P5	Probable butyrate kinase			X	X	
rfbE	Q8Z5I4	Glucose-1-phosphate cytidylytransferase	X				
group_23642	Q9WYS7	N5-carboxyaminoimidazole ribonucleotide mutase			X		
group_10498	Q9WZY4	O-acetyl-L-homoserine sulfhydrylase		X			

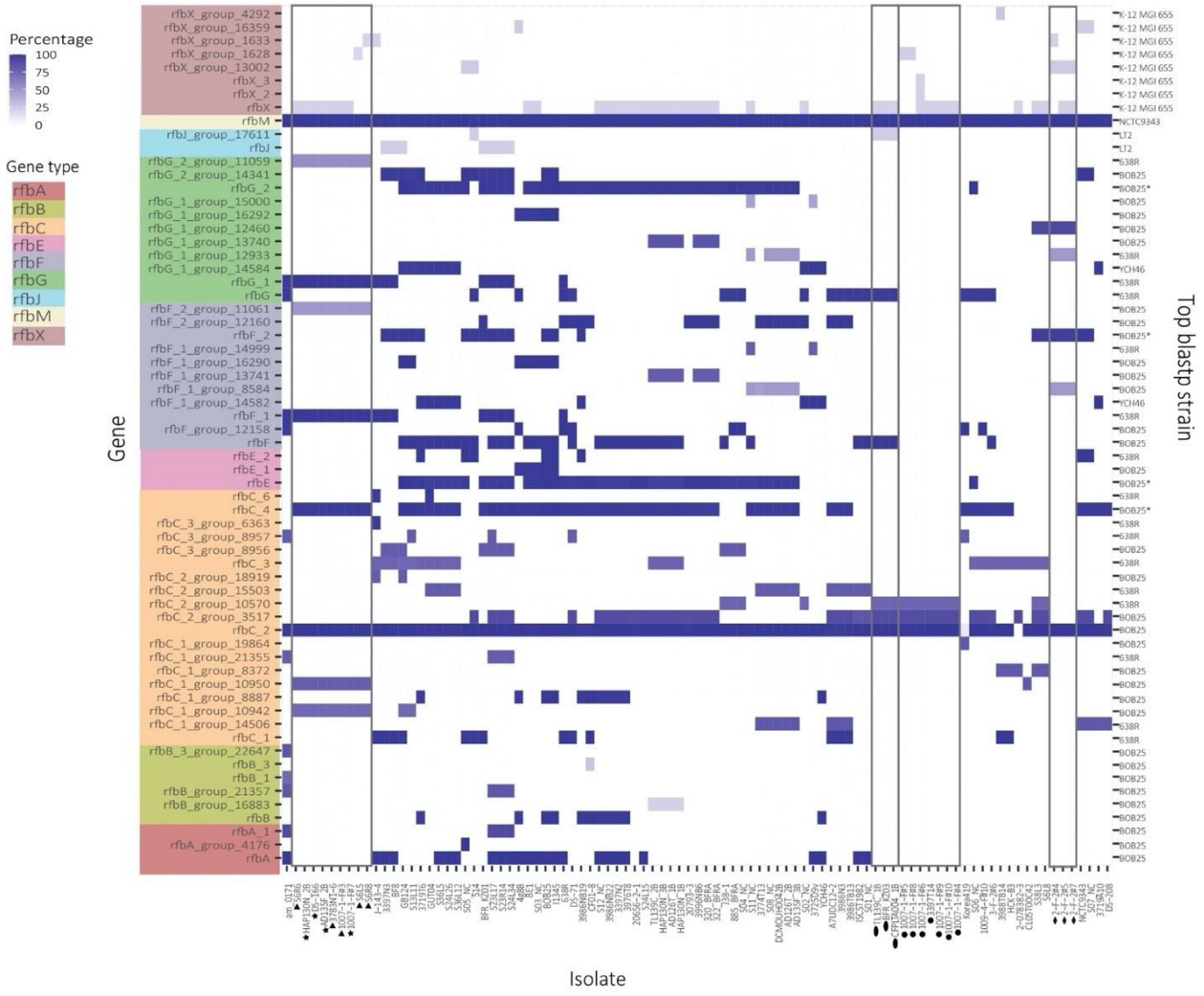

Figure 4.12: Heatmap of *rfb* genes present within each isolate according to identification with Blastp

The percentage identity of the isolate *rfb* genes to the Blastp hits were visulalised using a heatmap generated in R. The isolates are displayed along the x axis and *rfb* gene shown along the y axis. The presence of a *rfb* gene is represented by a purple square and percentage identity to *rfb* represented by darknessof the colour (See percentage legend). The *rfb* genes are coloured according to type (See Gene type legend). The right-hand axis shows thee top blastp hit for each *rfb* gene and corresponds to Table 4.2. The isolates within each Cluster are shown by the following: Cluster 1, black circles; Cluster 2, black ovals; Cluster 3, black stars; Cluster 4, black triangles; Cluster 5, black diamonds.

The *rfb* genes within each Cluster are also outlined in a grey rectangle to allow for easier interpretation of the heatmap.

The majority of isolate *rfb* genes consistently showed the highest percentage identity to those of strain BOB25. The top blast hit and associated KEGG ID can be found in Appendix 6. However, four isolate *rfb* genes contained top hits to multiple *rfb* reference genes (*rfbE, rfbF_2, rfbG_2* and *rfbC_4)*. For easier visualisation, one reference gene was chosen to be represented in the heatmap and the alternative reference gene/percentage identity are stored in Appendix 7. For example, the percentage identity of isolate *rfbG_2* to strain BOB25 (KEGG ID: VU15_16420) is represented on the heatmap. However, six isolates showed a higher percentage identity to strain 638R (KEGG ID: BF638R_3484).

It was determined through KEGG that *rfb* genes are involved in the O-antigen nucleotide sugar biosynthesis pathway (map00541). A map was generated in Adobe Illustrator to visualise the pathway and *rfb* genes involved (Figure 4.13).

rfb genes seem to be mainly involved in synthesis of dTDP-sugars and CDP-sugars via dTDP-glucose (dTDP-D-Glc; C00842) and D-Glucose alpha-1-phosphate (D-Glc-1P; C00103), respectively. The sugar residues created from the pathway form the repeating unit of the outermost and immunogenic domain of the LPS surrounding Gram-negative bacteria. *rfbM* is involved in the synthesis of GDP-sugars from beta-D-Fructose 6-phosphate (Fru-6P; C05345) via GDP-mannose (GDP-Man; C00096). It should be noted that the *rfb* genes are glycosyltransferases and can be involved in the synthesis of PSs other than the O-antigen[150].

The *rfb* genes within other Gram-negative bacteria, such as *Salmonella* and *Escherichia coli*, reside within a gene cluster. The gene arrangement and orientation of the *rfb* genes identified in each isolate were visualised with Geneious. All isolates showed some similarity in the gene organisation of *rfb* genes. For example, it was noted that a *rfbF* gene was followed by an *rfbF* gene. Additionally, the *rfb* genes were surrounded by other glycosyltransferase genes and upstream of the *rfb* genes were transcriptional regulation genes. Three isolates were chosen to show the variability and similarities between *rfb* gene clusters (DCMOUH0042B, BOB25 and AD135F_2B) and visualised in Adobe Illustrator (Figure 4.14). All isolates had *rfb* genes with *rmd* genes and *gtf* genes dispersed throughout the gene cluster. Additionally, all isolates had an LPS biosynthesis flippase protein gene upstream of *rfbG* and *rfbF.* DCMOUH0042B displayed an IS613 transposase in the middle of the cluster at a different orientation.

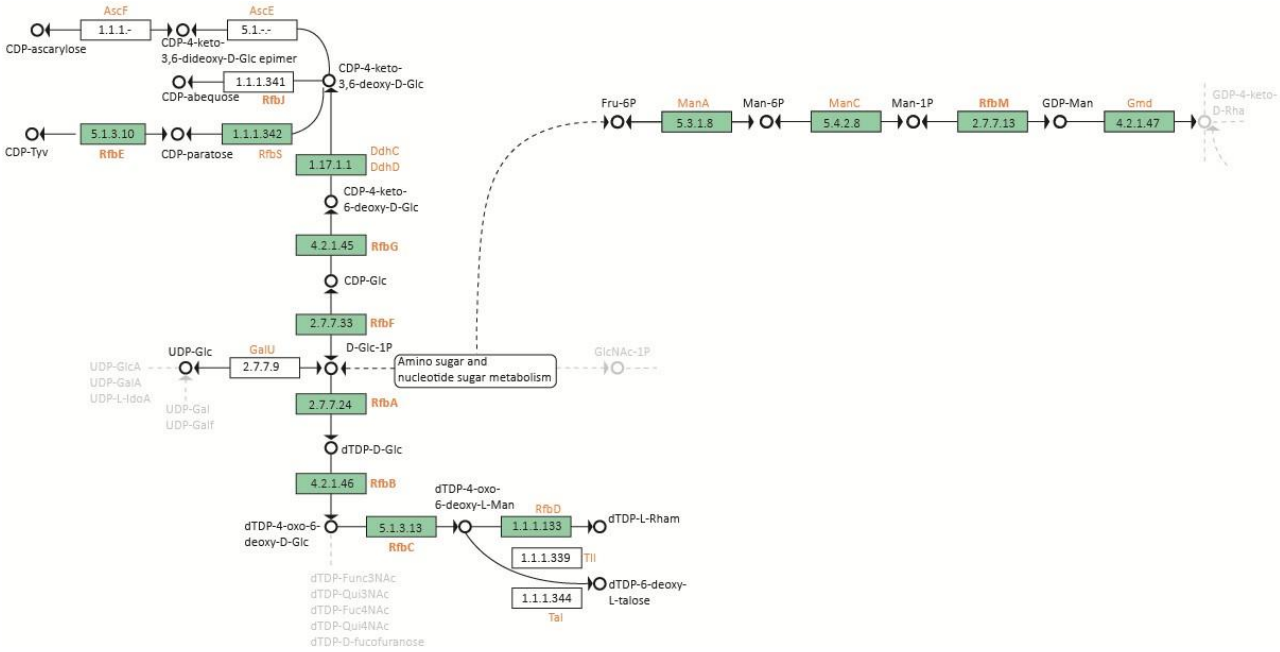

Figure 4.13: Annotated O-antigen nucleotide sugar biosynthesis KEGG pathway showing *rfb* gene involvement

Green rectangles represent genes present in the *B. fragilis* BOB25 KEGG O-antigen nucleotide sugar biosynthesis (PATHWAY: bfb00541) compared to the generic reference pathway. The numbers in the boxes correspond to Enzyme Commission identifiers, which classify enzymes according to the chemical reactions they catalyse. White circles represent the chemical compound (with name of the compound above) and arrows show the direction of the pathway. The orange text shows the gene that corresponds with the adjacent green rectangle. A dotted line represents a connection to another KEGG pathway. The grey text/lines show continuation of the pathway but was not relevant for this analysis.

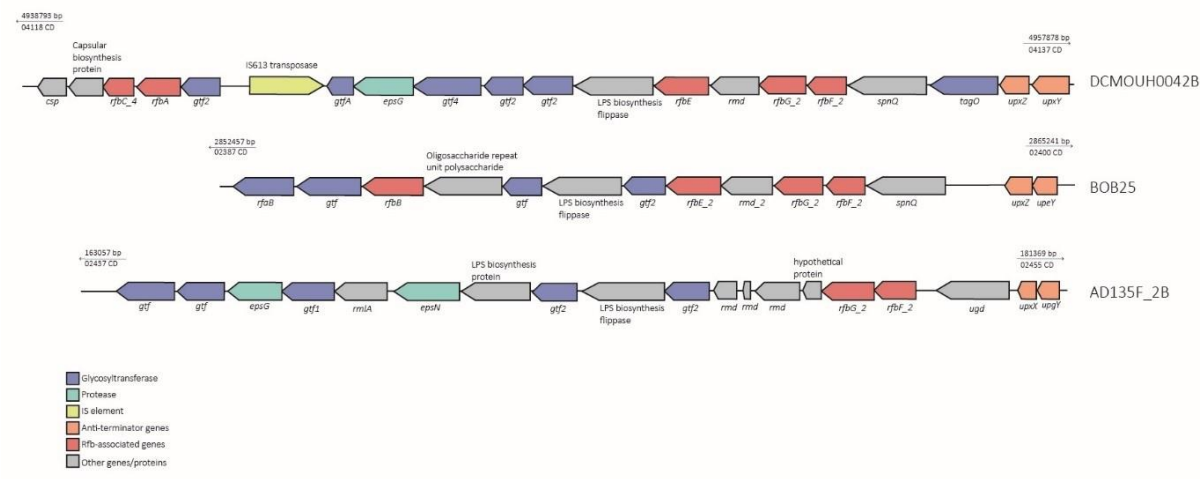

Figure 4.14: Genome arrangement of predicted PS loci of *B. fragilis* DCMOUH0042B, BOB25 and AD135F_2B

Three isolates were selected as examples to show the gene arrangement conservation between isolates. The coding region and base pair are show at either end of the DNA segment. The direction of the arrows represents orientation of the coding region and gene/protein name displayed. The colour of the arrow corresponds to the legend in the lower left-hand corner. The figure was generated in Adobe Illustrator.

Transcription regulators *upxZ, upxY* and *upxX* were consistently found upstream of the *rfb genes,* confirming the *rfb* genes are within a PS locus. The 'x' within *upxZ, upxY* and *upxX* changes according to the PS locus the genes are located in (e.g. upaZ is found within PSA). Therefore, PS locus E of BOB25 and PS locus G of AD135F_2B is represented in Figure 4.14. The UpxZ proteins from a PS locus inhibit the action of UpxY anti-terminators from other PS loci. Although PS loci share a common genetic organisation, a high level of diversity was observed in PS loci between strains and agrees with previous studies. However, future work should investigate the PSs within the main Cluster to determine if there are any similarities.

4.3.8 Functional analyses of the pangenome

EggNOG-mapper was used to generate a COG annotation table from the Roary output. This revealed a marked difference between the proportion of annotated genes in each COG category (Figure 4.15).

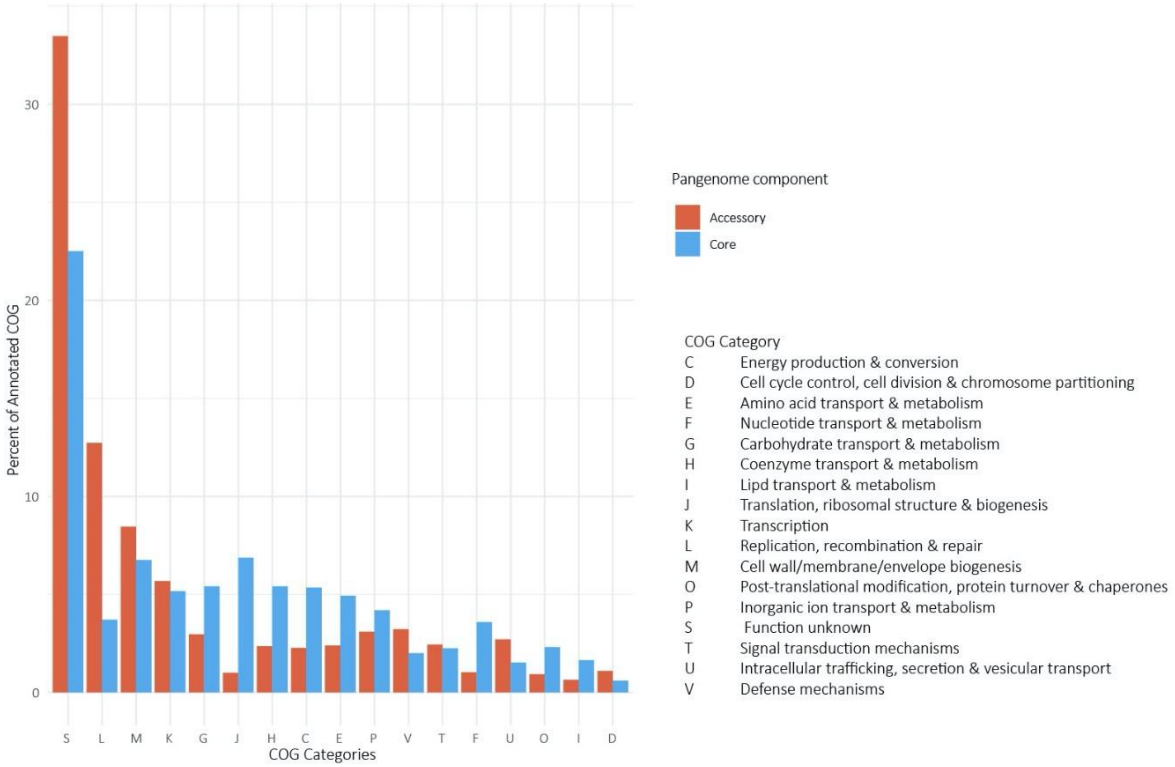

Figure 4.15: Proportion (%) of annotated COGs within the accessory and core genomes.
The percentage of annotated COGs in the accessory and core genome was determined from the KEGG output. The COG categories are shown on the x axis and percentage of annotated COGs in each pangenome component shown on the y axis. The accessory genome is represented by red bars and core genome represented by blue bars.

The largest COG category in the accessory and core genomes was S (Function unknown): 33.4 % and 22.5 %, respectively. COG category J (Translation, ribosomal structure and biogenesis) had the highest representation in the core genome (6.8 %). COG categories G (Carbohydrate transport and metabolism), H (Coenzyme transport and metabolism), C (energy production and conversion), E (Amino acid transport and metabolism) and I (Lipid transport and metabolism) were also higher in the core genome than in the accessory genome. The accessory genome (12.7%) showed a higher percentage of annotated genes in COG category L (Replication, recombination and repair) compared to the core genome (4 %). COG categories M (Cell wall, membrane and envelope biogenesis), K (Transcription), V (Defence mechanisms), U (Intracellular trafficking, secretion and vesicular transport) and D (Cell cycle control, cell division and chromosome portioning) were also higher in the accessory genome.

The COG categories for the unique genes in Clusters 1-5 (defined in the previous section) were identified (Figure 4.16).

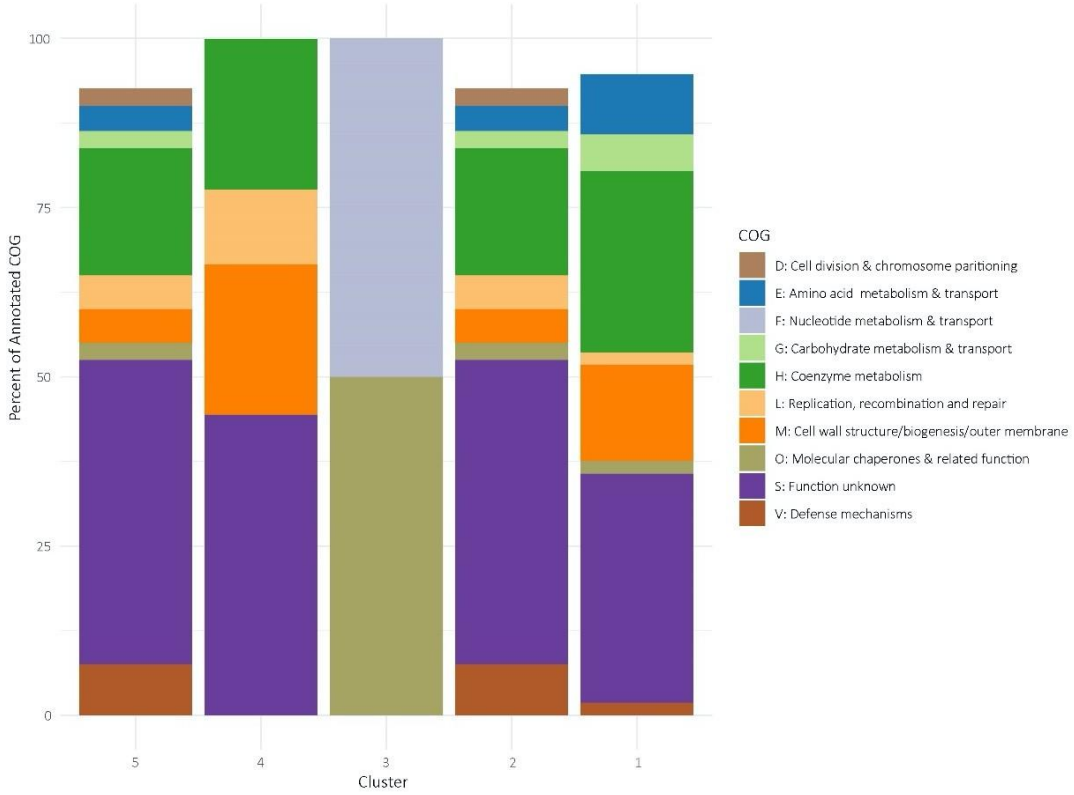

Figure 4.16: Stacked bar chart showing the proportion of annotated COGs of unique genes within each pangenome cluster and assigned COG category

The percentage of annotated COGs in the unique genes from the 5 clusters was determined from the KEGG output. Clusters 1-5 were defined according to the PCoA plot (Figure 4.10). The clusters are shown on the x axis and percentage of annotated COGs in each cluster shown on the y axis. The COG categories are shown by coloured sections and represented in the figure legend.

It should be noted that any COG categories under 1 % were not shown in the figure to allow for easier visualisation. In Clusters 1, 2, 4 and 5 the largest category was S (function unknown), followed by COG category H (Coenzyme metabolism) or M (Cell wall structure, biogenesis and outer membrane). Only two unique genes were found in Cluster 3 and each gene belonged to COG category F (Nucleotide metabolism and transport) or O (Molecular chaperones & related function), respectively.

The COG categories for the 'missing' genes in Clusters 1-5 were also identified (Figure 4.17). The clusters all showed a similar COG profile and this may be to due to the shared 'missing' genes between multiple clusters (as seen in Table 4.15). The COG category with the highest representation across all clusters was S (Function unknown), followed by COG category M (Cell wall, membrane and envelope biogenesis) or P (Inorganic ion transport and metabolism). Cluster 1 did not contain any 'missing' genes categorised as belonging to COG category J (Translation,

ribosomal structure and biogenesis) but did show the highest percentage of 'missing' genes belonging to T (Signal transduction mechanisms). All clusters showed 'missing' genes belonging toCOG category G (Carbohydrate transport/metabolism). This COG category contains the *rfb* genes analysed in the previous sections.

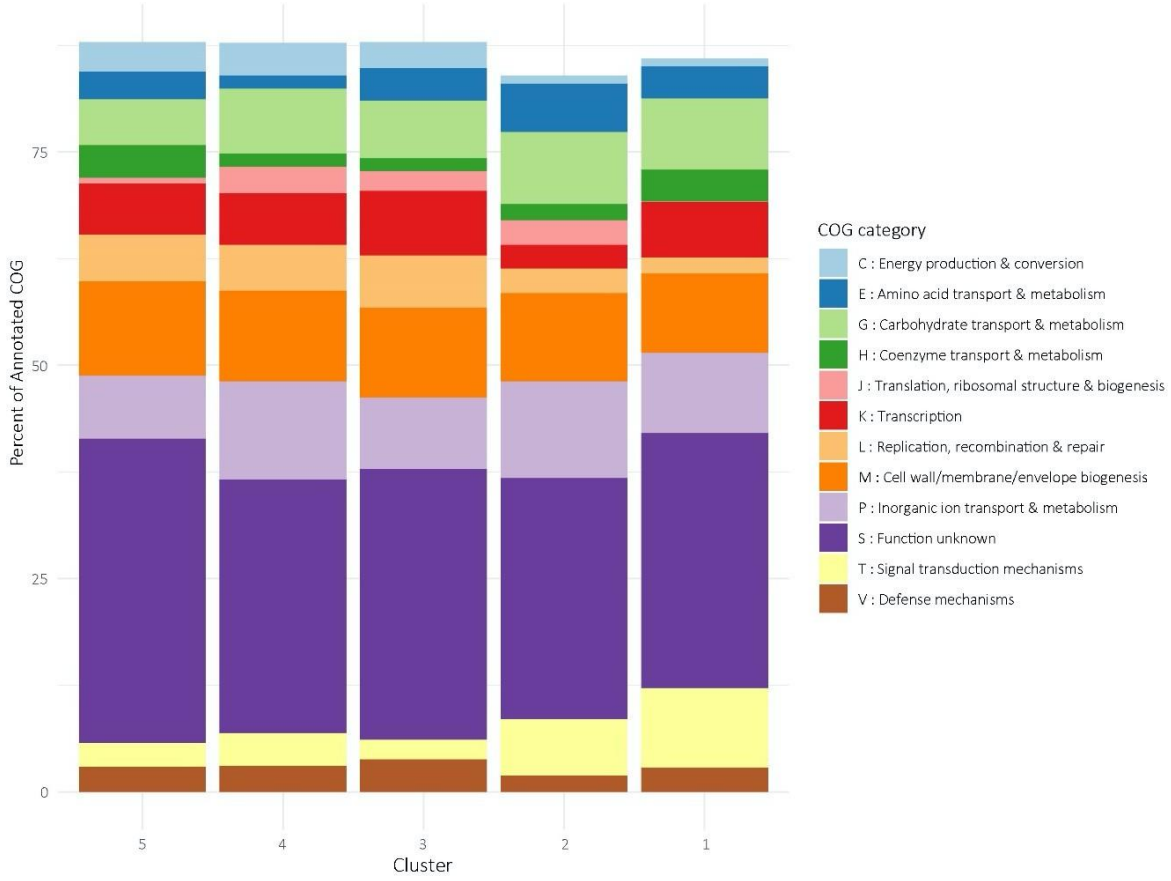

Figure 4.17: Stacked bar chart showing the proportion of annotated COGs of 'missing' genes within each pangenome Cluster and assigned COG category

The percentage of annotated COGs in the 'missing' genes from the 5 clusters was determined from the KEGG output. Clusters 1-5 were defined according to the PCoA plot (Figure 4.10). The clusters are shown on the x axis and percentage of annotated COGs in each cluster shown on the y axis. The COG categories are shown by coloured sections and represented in the figure legend.

4.3.9 Analysis of co-evolving genes

The presence of associating and dissociating genes was analysed by Dr Domingo-Sananes with Coinfinder. No significant associating genes were found within the pangenome dataset.

4.3.10 Identification of prophage within *B. fragilis* genomes

All genomes were screened for potential prophage using PhiSpy and the predicted prophage sequences manually searched for the presence of an integrase protein and structural proteins. Of the 93 isolates screened, 46 isolates encoded a prophage. A total of 78 prophage were predicted across the 46 isolates, with Korea 419 containing seven candidate prophage sequences (Table 4.16). However, only 67 of the 78-candidate prophages encoded an integrase protein and a further two prophages encoded one structural protein (Appendix 8). The presence of a capsid protein (major capsid protein) and a tail fibre protein (or tail spike protein) was needed to confidently assign a prophage sequence. No candidate prophage were taken forward for further analysis.

The structural proteins identified in 1007-1-F#10 and 1007-1-F#5 were characterised using Blastp and both showed 100 % identity to a phage tail tape measure protein from *B. fragilis* (WP_032533192.1).

4.4 Discussion

This Chapter reports the pangenome and comparative genomic analyses of 93 *B. fragilis* isolates from multiple isolation sites. No specific differences were noted between non-clinical, clinical and enterotoxigenic isolates; however, six pangenome Clusters were identified. There was a lack of clustering of *B. fragilis* isolates based on lifestyle or isolation site. Additionally, this study showed the fundamental need for accurate data curation and quality control on publicly downloaded genomes. A high level of diversity of the *rfb* gene between isolates was observed, except for those within specific Clusters. This is consistent with previous studies suggesting that the PS loci within *B. fragilis* are diverse[54,59]. The PS loci of *B.fragilis* share a common genetic structure and composed of diverse glycosyltransferases (e.g. *rfb*) , promoter region and other PS-associated genes. This analysis focused on *rfb* gene diversity within the isolates due lack of research regarding *rfb* gene diversity within isolates. Furthermore, these results suggest that the clustering of isolates may be in part due to similar *rfb* gene profiles and could explain why isolates of differing classifications grouped together.

Table 4-16: Overview of predicted prophage identified in isolates using PhiSpy

The predicted prophage region was accepted or rejected based on the presence of an integrase and presence of multiple structural proteins. All predicted prophage in the table below were rejected.

Strain	No. of prophage	Predicted prophage ID	Predicted prophage size (bp)	Integrase present?	No. of structural proteins
1007-1-F #10	2	pp1	8682	Yes	0
		pp1	38713	Yes	1
1007-1-F #4	1	pp1	20703	No	1
1007-1-F #5	1	pp1	39926	Yes	1
1007-1-F #6	1	pp1	21324	No	1
1007-1-F #9	1	pp1	26449	Yes	0
1009-4-F #10	1	pp1	30711	Yes	0
20793-3	1	pp1	25966	Yes	0
320_BFRA	1	pp1	10487	Yes	0
3397 T14	1	pp1	38714	Yes	0
3719 A10	2	pp1	30572	Yes	0
		pp2	20705	Yes	0
3725 D9(v)	3	pp1	66463	Yes	0
		pp2	2995	Yes	0
		pp3	25968	Yes	0
3774 T13	1	pp1	26447	Yes	0
3986 T(B)13	1	pp1	25436	Yes	0
4g8B	1	pp1	25968	Yes	0
638R	3	pp1	20605	Yes	0
		pp2	19824	No	0
		pp3	25969	Yes	0
AD126T_1B	2	pp1	57862	Yes	0
		pp2	10185	No	0
BE1	1	pp1	25968	Yes	0
BF8	1	pp1	25970	Yes	0
BFR_KZ01	1	pp1	9543	Yes	0
BFR_KZ03	1	pp1	37896	Yes	0
BOB25	2	pp1	30777	Yes	0
		pp2	25968	Yes	0
CFPLTA004_1B	2	pp1	5102	Yes	0
		pp2	37896	No	0
CL05T00C42	5	pp1	20777	Yes	0
		pp2	25294	Yes	0
		pp3	25967	Yes	0
		pp4	28352	Yes	0
		pp5	15493	Yes	0
DCMOUH0042B	1	pp1	36546	Yes	0
DS-166	2	pp1	30987	Yes	0
		pp2	25068	No	0

Strain	No. of prophage	Predicted prophage ID	Predicted prophage size (bp)	Integrase present?	No. of structural proteins
DS-71	1	pp1	42567	Yes	0
GUT04	1	pp1	25966	Yes	0
GB-124	2	pp1	25968	Yes	0
		pp2	18863	Yes	0
HAP130N_1B	2	pp1	11252	Yes	0
		pp2	57862	Yes	0
HAP130N_3B	2	pp1	10680	Yes	0
		pp2	57862	Yes	0
HCK-B3	2	pp1	35860	Yes	0
		pp2	14274	No	0
I1345	1	pp1	23802	No	0
ISCST1982	1	pp1	29370	Yes	0
J38-1	1	pp1	31198	Yes	0
Korea 419	7	pp1	15112	Yes	0
		pp2	9095	No	0
		pp3	18717	No	0
		pp4	15468	Yes	0
		pp5	33019	Yes	0
		pp6	8591	Yes	0
		pp7	25968	Yes	0
NCTC 9343	4	pp1	35889	Yes	0
		pp2	22082	Yes	0
		pp3	18554	Yes	0
		pp4	25967	Yes	0
S14	1	pp1	25967	Yes	0
S24L15	1	pp1	32223	Yes	0
S24L26	1	pp1	19890	Yes	0
S24L34	1	pp1	17484	Yes	0
S36L12	2	pp1	22748	Yes	0
		pp2	25967	Yes	0
S36L5	2	pp1	6855	No	0
		pp2	25967	Yes	0
S38L3	3	pp1	16759	Yes	0
		pp2	13806	Yes	0
		pp3	25969	Yes	0
S6L8	2	pp1	10192	Yes	0
		pp2	25969	Yes	0
S6R6	1	pp1	20599	Yes	0
S6R8	1	pp1	17318	Yes	0
1007-1-F #3	0	-	-	-	-
1007-1-F #7	0	-	-	-	-
1007-1-F #8	0	-	-	-	-
20656-2-1	0	-	-	-	-
2-078382-3	0	-	-	-	-
2-F-2 #4	0	-	-	-	-

Strain	No. of prophage	Predicted prophage ID	Predicted prophage size (bp)	Integrase present?	No. of structural proteins
2-F-2 #5	0	-	-	-	-
2-F-2 #7	0	-	-	-	-
322_BFRA	0	-	-	-	-
3397 N2	0	-	-	-	-
3397 N3	0	-	-	-	-
3719 T6	0	-	-	-	-
3783N1-6	0	-	-	-	-
3976T8	0	-	-	-	-
3986 N(B)19	0	-	-	-	-
3986 N(B)22	0	-	-	-	-
3986 N3	0	-	-	-	-
3988T(B)14	0	-	-	-	-
3996 N(B) 6	0	-	-	-	-
3-F-2 #6	0	-	-	-	-
885_BFRA	0	-	-	-	-
A7 (UDC12-2)	0	-	-	-	-
AD126T_2B	0	-	-	-	-
AD135F_2B	0	-	-	-	-
AD135F_3B	0	-	-	-	-
am_0171	0	-	-	-	-
CF01-8	0	-	-	-	-
HAP130N_2B	0	-	-	-	-
J-143-4	0	-	-	-	-
S13 L11	0	-	-	-	-
S23 R14	0	-	-	-	-
S23L17	0	-	-	-	-
S6L5	0	-	-	-	-
TL139C_1B	0	-	-	-	-
S12	0	-	-	-	-
S11	0	-	-	-	-
S06	0	-	-	-	-
S02	0	-	-	-	-
S08	0	-	-	-	-
S01	0	-	-	-	-
S04	0	-	-	-	-
S03	0	-	-	-	-
S07	0	-	-	-	-
S05	0	-	-	-	-

This study highlighted the need for quality control on assembled isolates from NCBI or publicly available databases. A total of 116 *B. fragilis* strains were downloaded from NCBI and 82 remained following the various quality control steps. Importantly, eight isolates were removed following ANI < 95 % when compared to *B. fragilis* NCTC 9343^{T}, including five multidrug-resistant strains and one isolate that was identified as *Parabacteroides distasonis*. Originally, *P. distasonis* was considered part of the *Bacteroides* genus but was reclassified in 2006[151]. Incorrect or poor-quality sequences could affect pangenome results with important inferences being missed or made incorrectly. Pangenome analysis relies upon gene clustering based on gene orthology; therefore, high-quality and taxonomically correct genomes are imperative for an accurate pangenome[89,90]. There is very little curation of bacterial genomes (draft or complete) available from publicly available databases, though NCBI has in the past 12 months started flagging ambiguously assigned assembled genomes in GenBank. Therefore, the responsibility of genome quality control lies with the user. It was also noted that many of the isolates on NCBI were enterotoxigenic and isolated from individuals with inflammatory diarrheal disease. However, these isolates lacked metadata and it was not possible to determine if ETBF has been confirmed as the causative agent or if isolation was a coincidence.

During curation of *B. fragilis* genomes from publicly available databases, a lack of genomes isolated from faecal samples of healthy individuals was noted. Of the 116 isolates collected from NCBI, only 25 were from faecal samples of individuals that did not have inflammatory diarrheal disease or labelled as enterotoxigenic. However, most of the non-clinical isolates collected from NCBI were not isolated from a traditionally healthy individual. For example, several isolates were sampled from ICU patients or cystic fibrosis-positive children. To gain a true picture of the *B. fragilis* population structure and an accurate pangenome, it was necessary to increase the proportion of isolates from healthy individuals in the work described herein. A literature search revealed a 2019 study that isolated 601 *B. fragilis* isolates from 12 individuals over a number of years[142]. The authors reported that each individual was dominated by a single *B. fragilis* lineage, which diversified over time to form coexisting sub lineages. The isolates collected were highly individualised and phylogenetically grouped according to subject. The introduction of all isolates from this study to the pangenome analysis could have skewed the results. Therefore, only one isolate from each lineage was chosen for the pangenome analysis. To achieve this, pangenome analysis using only the samples from the original study was undertaken. Interestingly, the core genome (27.5 %) was small compared to the core genome of other commensal/opportunistic pathogens, suggesting the pangenome is open. For example, the core genome of the commensal bacterium *Cutibacterium acnes* (formerly *Propionibacterium acnes*) was 88 % of the total gene

count[99]. These results are consistent with the original literature as intestinal *B. fragilis* is under constant selective pressure for long-term colonisation[142]. The accessory genome of all isolates was visualised with a PCoA plot and further confirmed that *B. fragilis* populations are highly individualised (Figure 4.4) and clustered according to the subject. A maximum likelihood tree generated from the core SNPs was also consistent with the original literature and the genomes grouped according to the subject (Figure 4.5). Isolates from subject 11 (S11) and subject 4 (S4) shared a subclade and isolates from subject 3 (S3) and subject 2 (S2) also shared a subclade. It should be noted that not all the isolate genomes generated from the original study were used due to technical difficulties encountered during genome assembly. One isolate from each lineage was selected for pangenome analysis with the isolates collected from NCBI.

The BFT metalloprotease is activated by the cysteine protease fragipain, inducing colonocyte E-cadherin cleavage and inflammatory cytokine secretion[9,29,30,152]. However, fragipain is found in all *B. fragilis* suggesting the protein has a role outside of BFT activation[29]. Additionally, *bft* is believed to play a role in extra-intestinal infection as isolates originating from blood samples are more likely to carry the *bft* gene[153,154]. A retrospective study in Kuwait screened 10-years of clinical *B. fragilis* isolates for the presence of *bft* and reported 49.9 % of extra-intestinal isolates were *bft*-positive[155]. This is considerably higher when compared to Poland (14.4 %), Japan (18.6 %), USA (6.2-38 %) and Hungary (13-25 %)[154,156-158]. The presence of BFT contributes to pathogenesis of anaerobic sepsis by weakening the intestinal epithelium and allowing bacteria to pass into the bloodstream[152].

The 93 high-quality genomes were screened for the presence of the BFT protein and fragipain to confirm the correct classification as 'enterotoxigenic'. Seventeen of the 93 isolates were found to encode a *bft* protein; with 13 of those classified as enterotoxigenic, three as nonclinical and one as clinical. However, 20656-2-1, 20793-3, BF8 and 3397N3 all contained *bft-2* and grouped in twos on the phylogenetic tree. A similar observation was noted with 3986NB22, 3397N2 and 3976T8. The prevalence of different *bft* isoforms is consistent with previous studies as most *bft*-positive isolates contain *bft-1*[33,159]. A Hungarian study reported 10 % of isolates encoded *bft-1* and 3 % encoded *bft-2*[156]. Three of the non-clinical were also *bft*-positive and this is in accordance with previous studies[7,31].

It is estimated up to 30 % of humans are asymptomatically colonised by ETBF.[31] This suggests that clinically significant disease depends on microbial virulence factors and host susceptibility factors. A 2017 study discovered a two-component system (RprXY) that suppresses *bft* expression to

maintain intestinal homeostasis and prevent lethal disease[160]. The authors determined that mucus-deficient mice had a higher susceptibility to ETBF colonisation if the regulatory system was disabled compared to mucus-deficient mice where the regulatory system was fully functional. This study further supports the theory that ETBF colonisation is dependent upon host mucosa integrity and homeostasis with ETBF can be achieved in healthy individuals[161]. Throughout this Chapter, the classification of isolates was confused by the lack of *bft*-positive isolates in the supposedly 'enterotoxigenic' isolates. The lack of metadata, as mentioned previously, confused the interpretation of these results. However, the classifications of the isolates remained as listed on NCBI. According to the literature, Korea-419 contains *bft-3*[149]. However, this was not discovered during this study. It is not known if this is due to an assembly or protein searching issue, even though the custom database contained all complete *bft* isoform sequences available. Additional protein searching tools should be used to confirm the presence/absence of *bft* genes within the isolates.

B. fragilis is able to develop resistance to several antimicrobials and the prevalence of resistance in clinical isolates has increased worldwide over the past decade. As noted in the introduction to this Chapter, resistance to tetracycline via *tetQ* and pencillin/cephalosporins other than cefoxitin via *cepA* is widespread in *B. fragilis*[15-17]. A total of 92 genomes encoded the *cepA* gene and 56 were positive for *tetQ.* Interestingly, seven genomes also encoded *cfxA* (CfxA3/CfxA2), which results in resistance to penicillin and cephalosporins including cefoxitin[20,21]. There did not appear to be a consistent AMR profile attributed to a specific classification (i.e. enterotoxigenic, clinical or non-clinical). However, there appeared to be more diversity in AMR genes present in enterotoxigenic isolates.

The mechanism of resistance to clindamycin in *Bacteroides* spp. is most commonly attributed to a mutation in the erythromycin resistance methylases (*erm*) genes, particularly *ermG, ermF* and *ermB*[162,163]*.* These genes were found in all genomes in all classifications; however, a high prevalence was noted in enterotoxigenic isolates. Linkage of *ermF* and *tetQ* on conjugative transposons has been described, and both genes are frequently found in clinical *Bacteroides* isolates[164]. Of the *ermF*-positive isolates, six encoded *tetQ*. Whereas two *ermF*-positive isolates did not encode *tetQ*. Additionally, 19 isolates were positive for *mef(En2).* This gene belongs to the major facilitator superfamily antibiotic efflux pump and confers resistance to macrolide antibiotics[165]. One *mef(En2)*-positive isolate was MDR DCMOUH0042B isolated from a clinical sample and encoding six other AMR genes. The *tetX* gene, encoded by four of the genomes, is associated with tetracycline resistance in the presence of oxygen via FAD- and NADPH-requiring

oxidoreductase. The presence of *tetX* infers resistance to multiple tetracycline derivatives[166]. It should be noted that the presence of AMR genes does not guarantee the phenotypic resistance to the antibiotics. Minimum inhibitory concentration testing should be completed to confirm the functioning of AMR genes within the isolates to complement the AMR genotype. Extensive analysis of AMR in *B. fragilis* is beyond the scope of this thesis, as it has been covered by another PhD student (English, Hoyles, Patrick and Grant, unpublished; L. Hoyles, personal communication).

The pangenome analysis described in this Chapter was undertaken with 93 isolates; 29 were non-clinical, 53 were enterotoxigenic and 11 were clinical. Of the 24,471 genes detected in the whole pangenome, only 1,571 genes made up the core genome (6.42 %). The majority of the genes were contained within relatively few isolates and this suggests *B. fragilis* has an open pangenome, as seen with the pangenome generated for the non-clinical isolates previously[99,102]. This is significantly smaller than the core genome of some pathogenic bacterial pangenomes[92,167,168]. For example, a recent study examined the pangenome of different bacterial species[169]. The lowest core genome percentage (53 %) was in 4401 *Escherichia coli* isolates and the entire pangenome contained 128,193 genes. The core genome of *Staphylococcus aureus* was 75 % of the total pangenome (22,133 total genes); this species also contained the smallest pangenome of the study. The authors proposed that new genes were less likely to be accepted and variations within the pangenome accumulated in the common region. The core genome of 190 *Bifidobacterium longum* strains also exhibited a small core genome (3.2%)[102]. The core genome size increased slightly when the authors only included *B. longum* subsp. *longum* in the pangenome (6 %).

The small core genome observed in this Chapter suggests that the core housekeeping genes necessary for basic survival are conserved among *B. fragilis* isolates, as noted with *B. longum*. Whereas the accessory genes are highly specialised and contribute towards long term persistence within the human host microbiota. A recent study revealed that constant adaptation of *B. fragilis* within the intestinal microbiome is a common feature of within-person evolution[142]. The authors revealed a rapid and continuous increase in the daily mutations of the *B. fragilis* isolates sampled from daily faecal samples from 12 individuals. It was estimated the frequency of mutations increased approximately 2 % daily. In contrast, within-person *E. coli* evolution is believed to be relatively low, particularly due to the low population size within the microbiome[170,171]. Analysis of the COG categories of the core and accessory genes revealed most of genes in the core genome belonged to category S (Function unknown; Figure 4.15). The majority of the genes within the core belonged to categories involved in translation, ribosomal structure, and biogenesis (J), coenzyme transport and metabolism (H), carbohydrate transport and metabolism (G) and amino

acid transport and metabolism (E). This further suggests that the genes within the core genome are basic housekeeping genes and the genes within the accessory genome are the consequence of constant adaptation.

A PCoA plot generated from the accessory genome of the isolates revealed there was little to no clustering of the isolates according to isolation site or pathogenesis (Figure 4.10). However, as mentioned previously, the interpretation of these results is clouded by the confusing metadata of the enterotoxigenic strains. Although there was no consistent clustering of the isolates according to isolation site, six clusters were evident (one main cluster and five outlying clusters). The outlying clusters were termed Clusters 1-5 and contained either non-clinical or enterotoxigenic isolates. The main cluster contained non-clinical, clinical and enterotoxigenic isolates. There was no grouping of specific classifications (i.e. non-clinical, clinical and enterotoxigenic) within the main cluster. This cluster appeared have a wider spread compared to the other clusters. The genomes belonging to Cluster 2 appeared to be most closely related to each other but closest to the main grouping. While Cluster 1 isolates were further spread out but furthest from the main grouping. Six of the seven isolates from Cluster 1 originated from the same study, according to NCBI. Due to the lack of metadata, it is unclear if these isolates were all from the same individual. However, the relatedness of the isolates could be explained if they were isolated from the same faecal sample or individual over time. A similar observation is noted for Cluster 5 as all its genomes appear to have similarly assigned isolate names. Cluster 3 and 4 originated from the same subclade and one isolate from Cluster 3 was positioned closer to Cluster 4. HAP130N_2B, AD135F_2B and DS-166 all encoded *bft-1* protein and this could explain why 1007-1-F#7 was not affiliated with Cluster 3. The population structure of the *B. fragilis* isolates was examined using a maximum likelihood phylogenetic tree generated from the core SNPs (Figure 4.11). Within the core genome, there was a high number of SNPs and this further highlighted the variability between the genomes. The same outlying clusters identified in the PCoA were also observed in the core SNPs phylogenetic tree. However, the relatedness of the isolates with the outlying Clusters was smaller suggesting these isolates have a similar SNP profile within the core genome. The isolates from the main cluster were distributed throughout the tree and did not show the same level of relatedness compared to the outlying Clusters, further suggesting the large spread within this Cluster.

A total of 494 unique genes was identified across all Clusters, with Cluster 3 only containing two unique genes (Table 4.12). Only 14 of the 494 unique genes were characterised using blastp. One gene from Cluster 1 showed 40 % percentage identity to demethylmenaquinone

methyltransferase (DMM) from *P. vulgatus* NCTC 11154. DMM is involved in the final step of menaquinone biosynthesis, in which it catalyses methylation of demthylmenaquinone using *S*-adenodylmethionine, resulting in the formation of menaquinone. Bacterially synthesised menaquinones form part of vitamin K[172]. Interestingly, a metaproteomic study analysing atopic dermatitis and the role of the infant gut microbiome suggested a potential important role of *Bacteroides*-synthesised DMM in metabolic alterations between healthy infants and infants with atopic dermatitis[173]. An additional potentially interesting gene within outlying Cluster 1 was oxygen regulatory protein NreC. The isolate protein sequence showed a 41 % percentage identity to the protein present in *Staphylococcus aureus.* This protein, along with NerB, is involved in reduction of nitrate/nitrite in the presence of oxygen, suggesting this could play a role in survival outside the anaerobic intestinal lumen[174]. All genomes belonging to Cluster 5 encoded an additional *tetO* gene. Several of the genes across all outlying Clusters were involved in bacterial structure (lipid A 1-phosphatase, putative fimbrium subunit Fim1C, glucose-1-phosphate thymidylyltrasnferase) or DNA regulation (DNA topoisomerase III, modification methylase DpnIIA, putative vert short patch repair endonuclease). No unique gene was discovered within the main cluster that was encoded in all genomes. However, one gene (PePSY-like domain containing protein) was present in 62 of the 72 isolates within the main cluster and not present in outlying Clusters. The PePSY domain is likely involved in regulation of peptidase activity; however, the role has not been studied in *Bacteroides* members[175]. A 2020 study investigated the role of PepSY domain-containing protein in *Francisella tularensis* pathogenicity[176]. The authors reported that deletion of the gene did not confer any obvious phenotypic changes and it was, therefore, unlikely to be essential for virulence.

Analysis of COGs associated with the unique genes within the outlying clusters showed that the majority of genes belonged to Category S (Function unknown), followed by Category H (Coenzyme metabolism) or Category M (Cell wall structure, biogenesis and outer membrane) (Figure 4.16). As Cluster 3 only had two unique genes, Figure 4.16 shows 50 % of genes belong to Category O (Molecular chaperones and related function) and 50 % to Category F (Nucleotide metabolism and transport). The predicted function of these genes could not be determined with Blastp. Additional tools, such as HMMER or DIAMOND, to identify the predicted function[177,178].

It is unlikely that the separation of the clusters from the main isolate group was due to unique genes. Therefore, the number of 'missing' genes was analysed. This was defined as a gene that was consistently not present in all genomes in a cluster but was present > 50 % of the remaining genomes. A total of 1,059 'missing' genes were identified across Clusters 1-5, with Cluster 5

showing the highest number of 'missing' genes (435). The main cluster did not contain any 'missing' genes, therefore the analysis below refers to outlying clusters only. Analysis of COG categories revealed a similar profile of 'missing' genes between clusters. The majority of genes belonged to Category S, as seen previously. Additionally, all clusters appeared to be 'missing' a similar proportion gene belonging to Category P (Inorganic ion transport and metabolism), M (Cell wall, membrane and envelope biogenesis) and G (Carbohydrate transport and metabolism) (Figure 4.17). Table 4.15 shows an overview of genes 'missing' from each cluster. As noted in the PCoA and phylogenetic tree, Clusters 3 and 4 appeared to be 'missing' similar genes that the other clusters were not; suggesting the genomes are closely related. The predicted products for these genes are butyrate kinase, PS biosynthesis protein EpsC, DNA binding protein HU and putative fimbrium anchoring subunit Fim4B[179].

All clusters showed a lack of a gene encoding for DNA mismatch repair protein (*mutL*)[180]. This gene has been located consistently upstream of the *ubb* region in 97 *B. fragilis* genomes[181]. The *ubb* region encodes for a eukaryotic-like ubiquitin protein (BfUbb) and shows toxicity against a subset of *B. fragilis* strains[182]. The authors reported that this region of *B. fragilis* is highly heterogenous and three genetic types exist[181]. However, all genetic types possess a similar *mutL* gene upstream of the *ubb* region. The variability within *ubb* regions was not explored within this Chapter. However, it would have been interesting to examine genomic diversity of the *ubb* region within these isolates.

It should be noted that the isolates containing 'missing' genes may contain a homologous gene that performs the same function. Therefore, the absence of the genes listed in Table 4.15 does not indicate that the isolate is unable to complete a function that gene would contribute to. The purpose of this analysis was to determine the gene diversity between the genomes within the Cluster and majority of genomes outside the Cluster to identify regions of genomic variability.

During analysis of the 'missing' genes, it was noted that several *rfb* genes were not present in multiple clusters. For example, CDP-paratose-2-epimerase (*rfbE*) and CDP-glucose 4,6-dehyratase (*rfbG_2*) were 'missing' from all clusters. As noted in the results, all isolates show a high diversity of *rfb* genes (Figure 4.12), with similarity noted between clusters. The main cluster showed a high level of diversity in *rfb* genes and little similarity between genomes was seen. The variability of *rfb* genes is associated with the variability of PS between isolates, as *rfb* genes are glycosyltransferases[183]. As noted in the introduction to this Chapter, the gene arrangement of PS loci between strains is conserved but genes within the loci are highly variable[46,53,54]. Visualisation

of three isolates confirmed that the PS loci regions were conserved as *upxX/upxZ* and *upxY* genes were upstream of the *rfb* genes. Despite numerous studies attempting to decipher the PS loci in *B. fragilis,* the true level of diversity remains unknown[46,53,54,61,65]. A 2018 study reported independent mutations in multiple *B. fragilis* isolates within the same individuals, with five of the 16 genes implicated in cell envelope biosynthesis[142]. In two separate individuals, multiple non-synonymous mutations within glycosyltransferase genes were noted. Additionally, isolates in four individuals showed mutations within the CPS biosynthesis protein (UngD2) over time. These results combined suggest the PS locus is under selective pressure to maintain colonisation and results in independent mutations in isolates within the same individual. Therefore, not only can *B. fragilis* isolates switch between expressed PS loci, diversity can also be introduced due to PS loci gene mutations. Evolutionary studies are needed to examine the mutations within isolate PS loci, the phenotypic presentation and influence on interactions with the microbiome, such as other bacteria, the human host and bacteriophage.

A transposase was noted within a PS locus in DCMOUH0042B, suggesting PS loci are transferable. A previous study reported the ability of large-scale chromosomal transfer between two *B. fragilis* isolates, HMW615 and 638R[184]. The authors reported a transfer of an entire PSA locus between isolates. An ICE (integrative and conjugative element) region downstream of the transferred PSA locus was also noted. This suggests that the diversity of PS loci within *B. fragilis* could also be attributed to HGT of the loci between isolates. This theory is further supported by the high level of similarity of *rfb* genes between isolates with the same cluster. It is possible that the PS locus is shared between these isolates. However, an extensive examination into the conservation of the genes within the loci in these clusters would be needed to confirm this.

Gene association and disassociation analyses were also undertaken using Coinfinder. However, this did not produce any significant results. It is unknown if this is due to the methodology or lack of gene association/disassociation between genomes. Future studies should use other methods to examine this, the alternatives to this are discussed in the Discussion Chapter. The presence of prophage was also investigated in this Chapter; however, no prophage regions were confidently assigned. PhiSpy predicted several prophage regions across the isolates (Table 4.16), but all were rejected following manual curation. It was noted that six isolates all had predicted prophage regions 25968 bp in length: 3725D9v pp3, 4g8b pp1, BE1 pp1, BOB25 pp2, GB124 pp2 and Korea 419 pp7. Additional prophage detection tools should be used or prophage should be induced to allow for their sequencing and characterisation.BV01, the only prophage identified within a *Bacteroides* species, is believed to be spontaneously induced[84]. Several attempts, including

antibiotic treatment, UV irradiation and co-culture with intestinal microbes, were made by authors to increase phage BV01 production but none were successful.

The genome size of *B. fragilis* is relatively large (~ 5.3 Mb) compared to other commensal bacteria such as *B. longum* (~ 2.2 Mb) and *S. epidermidis* (~ 2.4 Mb)[94,102,145]. This large genome size suggests that *B. fragilis* has a large repertoire of genes for adaptation and colonisation within differing ecological niches, such as intestinal to systemic. Further studies are needed to examine the virulence factors associated with extraintestinal colonisation. However, it could be possible that *B. fragilis* is a true opportunistic pathogen and the transfer of additional virulence genes is not needed for pathogenesis; hinting that infection is dependent on the human host's health (as suggested in the introduction to this Chapter). The results presented in this Chapter give a thorough overview of the *B. fragilis* pangenome; however, the data need to be examined further in extensive detail to determine the genetic differences between isolates. For example, the presence of mobile genetic elements was not investigated in this Chapter due to time constraints. Additionally, the *ubb* region within *B. fragilis* is an interesting research area due to the similarity to eukaryotic-like ubiquitin, diversity between strains and toxicity to a specific *B. fragilis* strains. Furthermore, further studies need to be undertaken to fully characterise the PS locus diversity between isolates.

4.5 References

1 Shen, Y. *et al.* Outer membrane vesicles of a human commensal mediate immune regulation and disease protection. *Cell Host Microbe* **12**, 509-520, doi:10.1016/j.chom.2012.08.004 (2012).

2 Wexler, H. M. Bacteroides: the good, the bad, and the nitty-gritty. *Clinical microbiology reviews* **20**, 593-621, doi:10.1128/CMR.00008-07 (2007).

3 Chung, L. *et al.* Bacteroides fragilis Toxin Coordinates a Pro-carcinogenic Inflammatory Cascade via Targeting of Colonic Epithelial Cells. *Cell Host Microbe* **23**, 203-214.e205, doi:https://doi.org/10.1016/j.chom.2018.01.007 (2018).

4 Namavar, F. *et al.* Epidemiology of the Bacteroides fragilis group in the colonic flora in 10 patients with colonic cancer. *J Med Microbiol* **29**, 171-176, doi:10.1099/00222615-29-3-171 (1989).

5 Nguyen, M. H. *et al.* Antimicrobial Resistance and Clinical Outcome of Bacteroides Bacteremia: Findings of a Multicenter Prospective Observational Trial. *Clinical Infectious Diseases* **30**, 870-876, doi:10.1086/313805 (2000).

6 Polk, B. F. & Kasper, D. L. Bacteroides fragilis subspecies in clinical isolates. *Ann Intern Med* **86**, 569-571, doi:10.7326/0003-4819-86-5-569 (1977).

7 Sears, C. L. *et al.* Association of enterotoxigenic Bacteroides fragilis infection with inflammatory diarrhea. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **47**, 797-803, doi:10.1086/591130 (2008).

8 Moore, W., Cato, E. & Holdeman, L. Some current concepts in intestinal bacteriology. *The American journal of clinical nutrition* **31**, S33-S42 (1978).

9 Shetab, R. *et al.* Detection of Bacteroides fragilisEnterotoxin Gene by PCR. *Journal of Clinical Microbiology* **36**, 1729-1732, doi:doi:10.1128/JCM.36.6.1729-1732.1998 (1998).

10 Bennion, R. S., Thompson, J. E., Baron, E. J. & Finegold, S. M. Gangrenous and perforated appendicitis with peritonitis: treatment and bacteriology. *Clin Ther* **12 Suppl C**, 31-44 (1990).

11 Osborne, N. The Role of Bacteroides fragilis in Pelvic Infections. *Journal of Gynecologic Surgery* **22**, 81-82, doi:10.1089/gyn.2006.22.81 (2006).

12 Elliott, D., Kufera, J. A. & Myers, R. A. The microbiology of necrotizing soft tissue infections. *Am J Surg* **179**, 361-366, doi:10.1016/s0002-9610(00)00360-3 (2000).

13 Farah, S., Alshehri, M. A., Alfawaz, T. S., Ahmad, M. & Alshahrani, D. A. Bacteroides fragilis meningitis in a Saudi infant: case report and literature review. *Int J Pediatr Adolesc Med* **5**, 122-126, doi:10.1016/j.ijpam.2018.05.003 (2018).

14 Hung, M. N. *et al.* Community-acquired anaerobic bacteremia in adults: one-year experience in a medical center. *J Microbiol Immunol Infect* **38**, 436-443 (2005).

15 Bryan, L. E., Kowand, S. K. & Van Den Elzen, H. M. Mechanism of aminoglycoside antibiotic resistance in anaerobic bacteria: Clostridium perfringens and Bacteroides fragilis. *Antimicrobial agents and chemotherapy* **15**, 7-13, doi:10.1128/AAC.15.1.7 (1979).

16 Shoemaker, N. B., Vlamakis, H., Hayes, K. & Salyers, A. A. Evidence for extensive resistance gene transfer among Bacteroides spp. and among Bacteroides and other genera in the human colon. *Appl Environ Microbiol* **67**, 561-568, doi:10.1128/aem.67.2.561-568.2001 (2001).

17 Veloo, A. C. M., Baas, W. H., Haan, F. J., Coco, J. & Rossen, J. W. Prevalence of antimicrobial resistance genes in Bacteroides spp. and Prevotella spp. Dutch clinical isolates. *Clinical Microbiology and Infection* **25**, 1156.e1159-1156.e1113, doi:https://doi.org/10.1016/j.cmi.2019.02.017 (2019).

18 Leng, Z., Riley, D. E., Berger, R. E., Krieger, J. N. & Roberts, M. C. Distribution and mobility of the tetracycline resistance determinant tetQ. *J Antimicrob Chemother* **40**, 551-559, doi:10.1093/jac/40.4.551 (1997).

19 Rogers, M. B., Parker, A. C. & Smith, C. J. Cloning and characterization of the endogenous cephalosporinase gene, cepA, from Bacteroides fragilis reveals a new subgroup of Ambler class A beta-lactamases. *Antimicrobial Agents and Chemotherapy* **37**, 2391-2400, doi:doi:10.1128/AAC.37.11.2391 (1993).

20 Pumbwe, L., Chang, A., Smith, R. L. & Wexler, H. M. Clinical significance of overexpression of multiple RND-family efflux pumps in Bacteroides fragilis isolates. *J Antimicrob Chemother* **58**, 543-548, doi:10.1093/jac/dkl278 (2006).

21 Pumbwe, L., Wareham, D. W., Aduse-Opoku, J., Brazier, J. S. & Wexler, H. M. Genetic analysis of mechanisms of multidrug resistance in a clinical isolate of Bacteroides fragilis. *Clinical Microbiology and Infection* **13**, 183-189, doi:https://doi.org/10.1111/j.1469-0691.2006.01620.x (2007).

22 Sóki, J. *et al.* Emergence and evolution of an international cluster of MDR Bacteroides fragilis isolates. *Journal of Antimicrobial Chemotherapy* **71**, 2441-2448, doi:10.1093/jac/dkw175 (2016).

23 Valguarnera, E. & Wardenburg, J. B. Good Gone Bad: One Toxin Away From Disease for Bacteroides fragilis. *J Mol Biol* **432**, 765-785, doi:10.1016/j.jmb.2019.12.003 (2020).

24 Round, J. L. *et al.* The Toll-like receptor 2 pathway establishes colonization by a commensal of the human microbiota. *Science* **332**, 974-977, doi:10.1126/science.1206095 (2011).

25 Wu, S., Lim, K.-C., Huang, J., Saidi, R. F. & Sears, C. L. Bacteroides fragilis enterotoxin cleaves the zonula adherens protein, E-cadherin. *Proceedings of the National Academy of Sciences* **95**, 14979-14984 (1998).

26 Haghi, F., Goli, E., Mirzaei, B. & Zeighami, H. The association between fecal enterotoxigenic B. fragilis with colorectal cancer. *BMC Cancer* **19**, 879, doi:10.1186/s12885-019-6115-1 (2019).

27 Boleij, A. *et al.* The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin Infect Dis* **60**, 208-215, doi:10.1093/cid/ciu787 (2015).

28 Casterline, B. W., Hecht, A. L., Choi, V. M. & Bubeck Wardenburg, J. The Bacteroides fragilis pathogenicity island links virulence and strain competition. *Gut microbes* **8**, 374-383, doi:10.1080/19490976.2017.1290758 (2017).

29 Franco, A. A. *et al.* Molecular evolution of the pathogenicity island of enterotoxigenic Bacteroides fragilis strains. *J Bacteriol* **181**, 6623-6633, doi:10.1128/jb.181.21.6623-6633.1999 (1999).

30 Pierce, J. V. & Bernstein, H. D. Genomic Diversity of Enterotoxigenic Strains of Bacteroides fragilis. *PLOS ONE* **11**, e0158171, doi:10.1371/journal.pone.0158171 (2016).

31 Sears, C. L. Enterotoxigenic Bacteroides fragilis: a rogue among symbiotes. *Clin Microbiol Rev* **22**, 349-369, Table of Contents, doi:10.1128/cmr.00053-08 (2009).

32 De Filippis, F. *et al.* Distinct Genetic and Functional Traits of Human Intestinal Prevotella copri Strains Are Associated with Different Habitual Diets. *Cell Host Microbe* **25**, 444-453.e443, doi:https://doi.org/10.1016/j.chom.2019.01.004 (2019).

33 Scotto d'Abusco, A. S., Del Grosso, M., Censini, S., Covacci, A. & Pantosti, A. The alleles of the bft gene are distributed differently among enterotoxigenic Bacteroides fragilis strains from human sources and can be present in double copies. *J Clin Microbiol* **38**, 607-612, doi:10.1128/jcm.38.2.607-612.2000 (2000).

34 Lobo, L. A., Jenkins, A. L., Jeffrey Smith, C. & Rocha, E. R. Expression of B acteroides fragilis hemolysins in vivo and role of HlyBA in an intra-abdominal infection model. *Microbiologyopen* **2**, 326-337 (2013).

35 Riepe, S. P., Goldstein, J. & Alpers, D. H. Effect of secreted Bacteroides proteases on human intestinal brush border hydrolases. *The Journal of clinical investigation* **66**, 314-322 (1980).

36 Reid, J. H. & Patrick, S. Phagocytic and serum killing of capsulate and non-capsulate Bacteroides fragilis. *J Med Microbiol* **17**, 247-257, doi:10.1099/00222615-17-3-247 (1984).

37 Galvão, B., Meggersee, R. & Abratt, V. Antibiotic resistance and adhesion potential of Bacteroides fragilis clinical isolates from Cape Town, South Africa. *Anaerobe* **17**, 142-146 (2011).

38 Reis, A. C. M., Silva, J. O., Laranjeira, B. J., Pinheiro, A. Q. & Carvalho, C. Virulence factors and biofilm production by isolates of Bacteroides fragilis recovered from dog intestinal tracts. *Brazilian Journal of Microbiology* **45**, 647-650 (2014).

39 Pruzzo, C., Dainelli, B. & Ricchetti, M. Piliated Bacteroides fragilis strains adhere to epithelial cells and are more sensitive to phagocytosis by human neutrophils than nonpiliated strains. *Infection and immunity* **43**, 189-194 (1984).

40 Ferreira, R. *et al.* Expression of Bacteroides fragilis virulence markers in vitro. *Journal of medical microbiology* **48**, 999-1004 (1999).

41 Potempa, J. & Pike, R. N. Corruption of innate immunity by bacterial proteases. *J Innate Immun* **1**, 70-87, doi:10.1159/000181144 (2009).

42 Thornton, R. F., Kagawa, T. F., O'Toole, P. W. & Cooney, J. C. The dissemination of C10 cysteine protease genes in Bacteroides fragilis by mobile genetic elements. *BMC Microbiol* **10**, 122, doi:10.1186/1471-2180-10-122 (2010).

43 Jotwani, R. & Gupta, U. Virulence factors in Bacteroides fragilis group. *Indian J Med Res* **93**, 232-235 (1991).

44 Zaleznik, D. F., Zhang, Z. L., Onderdonk, A. B. & Kasper, D. L. Effect of subinhibitory doses of clindamycin on the virulence of Bacteroides fragilis: role of lipopolysaccharide. *J Infect Dis* **154**, 40-46, doi:10.1093/infdis/154.1.40 (1986).

45 Onderdonk, A. B., Kasper, D. L., Cisneros, R. L. & Bartlett, J. G. The capsular polysaccharide of Bacteroides fragilis as a virulence factor: comparison of the pathogenic potential of encapsulated and unencapsulated strains. *J Infect Dis* **136**, 82-89, doi:10.1093/infdis/136.1.82 (1977).

46 Coyne, M. J. *et al.* Polysaccharide biosynthesis locus required for virulence of Bacteroides fragilis. *Infection and immunity* **69**, 4342-4350 (2001).

47 Tzianabos, A. O., Kasper, D. L., Cisneros, R. L., Smith, R. S. & Onderdonk, A. B. Polysaccharide-mediated protection against abscess formation in experimental intra-abdominal sepsis. *The Journal of clinical investigation* **96**, 2727-2731 (1995).

48 Los, F. C. O., Randis, T. M., Aroian, R. V. & Ratner, A. J. Role of Pore-Forming Toxins in Bacterial Infectious Diseases. *Microbiol Mol Biol R* **77**, 173-207, doi:doi:10.1128/MMBR.00052-12 (2013).

49 Braun, V. & Focareta, T. Pore-forming bacterial protein hemolysins (cytolysins). *Crit Rev Microbiol* **18**, 115-158, doi:10.3109/10408419109113511 (1991).

50 Robertson, K. P., Smith, C. J., Gough, A. M. & Rocha, E. R. Characterization of Bacteroides fragilis Hemolysins and Regulation and Synergistic Interactions of HlyA and HlyB. *Infection and Immunity* **74**, 2304-2316, doi:doi:10.1128/IAI.74.4.2304-2316.2006 (2006).

51 Lobo, L. A., Jenkins, A. L., Jeffrey Smith, C. & Rocha, E. R. Expression of Bacteroides fragilis hemolysins in vivo and role of HlyBA in an intra-abdominal infection model. *Microbiologyopen* **2**, 326-337, doi:10.1002/mbo3.76 (2013).

52 Round, J. L. & Mazmanian, S. K. Inducible Foxp3+ regulatory T-cell development by a commensal bacterium of the intestinal microbiota. *Proc Natl Acad Sci U S A* **107**, 12204-12209, doi:10.1073/pnas.0909122107 (2010).

53 Neff, C. P. *et al.* Diverse Intestinal Bacteria Contain Putative Zwitterionic Capsular Polysaccharides with Anti-inflammatory Properties. *Cell Host Microbe* **20**, 535-547, doi:10.1016/j.chom.2016.09.002 (2016).

54 Patrick, S. *et al.* Twenty-eight divergent polysaccharide loci specifying within- and amongst-strain capsule diversity in three strains of Bacteroides fragilis. *Microbiol-Sgm* **156**, 3255-3269 (2010).

55 Patrick, S. & Reid, J. H. Separation of capsulate and non-capsulate Bacteroides fragilis on a discontinuous density gradient. *J Med Microbiol* **16**, 239-241, doi:10.1099/00222615-16-2-239 (1983).

56 Reid, J. H., Patrick, S. & Tabaqchali, S. Immunochemical characterization of a polysaccharide antigen of Bacteroides fragilis with an IgM monoclonal antibody. *J Gen Microbiol* **133**, 171-179, doi:10.1099/00221287-133-1-171 (1987).

57 Patrick, S., Reid, J. H. & Coffey, A. Capsulation of in vitro and in vivo grown Bacteroides species. *J Gen Microbiol* **132**, 1099-1109, doi:10.1099/00221287-132-4-1099 (1986).

58 Patrick, S., Houston, S., Thacker, Z. & Blakely, G. W. Mutational analysis of genes implicated in LPS and capsular polysaccharide biosynthesis in the opportunistic pathogen Bacteroides fragilis. *Microbiology (Reading)* **155**, 1039-1049, doi:10.1099/mic.0.025361-0 (2009).

59 Patrick, S., Gilpin, D. & Stevenson, L. Detection of intrastrain antigenic variation of Bacteroides fragilis surface polysaccharides by monoclonal antibody labelling. *Infection and immunity* **67**, 4346-4351, doi:10.1128/IAI.67.9.4346-4351.1999 (1999).

60 Lutton, D. A. *et al.* Flow cytometric analysis of within-strain variation in polysaccharide expression by Bacteroides fragilis by use of murine monoclonal antibodies. *J Med Microbiol* **35**, 229-237, doi:10.1099/00222615-35-4-229 (1991).

61 Bayley, D. P., Rocha, E. R. & Smith, C. J. Analysis of cepA and other Bacteroides fragilis genes reveals a unique promoter structure. *FEMS Microbiology Letters* **193**, 149-154, doi:10.1111/j.1574-6968.2000.tb09417.x (2000).

62 Patrick, S. *et al.* Multiple inverted DNA repeats of Bacteroides fragilis that control polysaccharide antigenic variation are similar to the hin region inverted repeats of Salmonella typhimurium. *Microbiology (Reading)* **149**, 915-924, doi:10.1099/mic.0.26166-0 (2003).

63 Woude, M. W. v. d. & Bäumler, A. J. Phase and Antigenic Variation in Bacteria. *Clinical Microbiology Reviews* **17**, 581-611, doi:doi:10.1128/CMR.17.3.581-611.2004 (2004).

64 Chatzidaki-Livanis, M., Coyne, M. J. & Comstock, L. E. A Family of Transcriptional Antitermination Factors Necessary for Synthesis of the Capsular Polysaccharides of Bacteroides fragilis. *Journal of Bacteriology* **191**, 7288-7295, doi:doi:10.1128/JB.00500-09 (2009).

65 Wang, Y., Kalka-Moll, W. M., Roehrl, M. H. & Kasper, D. L. Structural basis of the abscess-modulating polysaccharide A2 from Bacteroides fragilis. *Proceedings of the National Academy of Sciences* **97**, 13478-13483, doi:10.1073/pnas.97.25.13478 (2000).

66 Silhavy, T. J., Kahne, D. & Walker, S. The bacterial cell envelope. *Cold Spring Harbor perspectives in biology* **2**, a000414 (2010).

67 Raetz, C. R. & Whitfield, C. Lipopolysaccharide endotoxins. *Annual review of biochemistry* **71**, 635-700 (2002).

68 Hansen, G. H., Rasmussen, K., Niels-Christiansen, L. L. & Danielsen, E. M. Lipopolysaccharide-binding protein: localization in secretory granules of Paneth cells in the mouse small intestine. *Histochem Cell Biol* **131**, 727-732, doi:10.1007/s00418-009-0572-6 (2009).

69 Seydel, U., Oikawa, M., Fukase, K., Kusumoto, S. & Brandenburg, K. Intrinsic conformation of lipid A is responsible for agonistic and antagonistic activity. *Eur J Biochem* **267**, 3032-3039, doi:10.1046/j.1432-1033.2000.01326.x (2000).

70 Cigana, C. *et al.* Pseudomonas aeruginosa Exploits Lipid A and Muropeptides Modification as a Strategy to Lower Innate Immunity during Cystic Fibrosis Lung Infection. *PLOS ONE* **4**, e8439, doi:10.1371/journal.pone.0008439 (2009).

71 Smith, E. E. *et al.* Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients. *Proc Natl Acad Sci U S A* **103**, 8487-8492, doi:10.1073/pnas.0602138103 (2006).

72 Weintraub, A., Larsson, B. E. & Lindberg, A. A. Chemical and immunochemical analyses of Bacteroides fragilis lipopolysaccharides. *Infect Immun* **49**, 197-201, doi:10.1128/iai.49.1.197-201.1985 (1985).

73 Poxton, I. R. & Edmond, D. M. Biological activity of Bacteroides lipopolysaccharide--reappraisal. *Clin Infect Dis* **20 Suppl 2**, S149-153, doi:10.1093/clinids/20.supplement_2.s149 (1995).

74 Berezow, A. B. *et al.* The structurally similar, penta-acylated lipopolysaccharides of Porphyromonas gingivalis and Bacteroides elicit strikingly different innate immune responses. *Microb Pathog* **47**, 68-77, doi:10.1016/j.micpath.2009.04.015 (2009).

75 Meredith, T. C., Aggarwal, P., Mamat, U., Lindner, B. & Woodard, R. W. Redefining the Requisite Lipopolysaccharide Structure in Escherichia coli. *ACS Chemical Biology* **1**, 33-42, doi:10.1021/cb0500015 (2006).

76 Lindberg, A. A., Weintraub, A., Zähringer, U. & Rietschel, E. T. Structure-activity relationships in lipopolysaccharides of Bacteroides fragilis. *Rev Infect Dis* **12 Suppl 2**, S133-141, doi:10.1093/clinids/12.supplement_2.s133 (1990).

77 Maskell, J. P. Electrophoretic analysis of the lipopolysaccharides of Bacteroides spp. *Antonie Van Leeuwenhoek* **65**, 155-161, doi:10.1007/bf00871756 (1994).

78 Maskell, J. P. The resolution of bacteroides lipopolysaccharides by polyacrylamide gel electrophoresis. *J Med Microbiol* **34**, 253-257, doi:10.1099/00222615-34-5-253 (1991).

79 Jacobson, A. N. *et al.* The Biosynthesis of Lipooligosaccharide from Bacteroides thetaiotaomicron. *mBio* **9**, e02289-02217, doi:doi:10.1128/mBio.02289-17 (2018).

80 Chu, H. *et al.* Gene-microbiota interactions contribute to the pathogenesis of inflammatory bowel disease. *Science* **352**, 1116-1120, doi:10.1126/science.aad9948 (2016).

81 Tan, H., Zhao, J., Zhang, H., Zhai, Q. & Chen, W. Novel strains of Bacteroides fragilis and Bacteroides ovatus alleviate the LPS-induced inflammation in mice. *Appl Microbiol Biotechnol* **103**, 2353-2365, doi:10.1007/s00253-019-09617-1 (2019).

82 Nuding, S. *et al.* Antibacterial activity of human defensins on anaerobic intestinal bacterial species: a major role of HBD-3. *Microbes Infect* **11**, 384-393, doi:10.1016/j.micinf.2009.01.001 (2009).

83 Troy, E. B. & Kasper, D. L. Beneficial effects of Bacteroides fragilis polysaccharides on the immune system. *Front Biosci (Landmark Ed)* **15**, 25-34, doi:10.2741/3603 (2010).

84 Campbell, D. E. *et al.* Infection with Bacteroides Phage BV01 Alters the Host Transcriptome and Bile Acid Metabolism in a Common Human Gut Microbe. *Cell Reports* **32**, 108142, doi:https://doi.org/10.1016/j.celrep.2020.108142 (2020).

85 Yao, L. *et al.* A selective gut bacterial bile salt hydrolase alters host metabolism. *Elife* **7**, e37182 (2018).

86 Joyce, S. A. *et al.* Regulation of host weight gain and lipid metabolism by bacterial bile acid modification in the gut. *Proceedings of the National Academy of Sciences* **111**, 7421-7426 (2014).

87 Shkoporov, A. N. *et al.* Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. *BMC Biology* **19**, 163, doi:10.1186/s12915-021-01084-3 (2021).

88 Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* **102**, 13950-13955 (2005).

89 Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Current opinion in genetics & development* **15**, 589-594 (2005).

90 Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* **11**, 472-477 (2008).

91 Rouli, L., Merhej, V., Fournier, P. E. & Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect* **7**, 72-85, doi:10.1016/j.nmni.2015.06.005 (2015).

92 Holt, K. E. *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proceedings of the National Academy of Sciences* **112**, E3574-E3581, doi:10.1073/pnas.1501049112 (2015).

93 Kiu, R., Caim, S., Alexander, S., Pachori, P. & Hall, L. J. Probing genomic aspects of the multi-host pathogen Clostridium perfringens reveals significant pangenome diversity, and a diverse array of virulence factors. *Frontiers in microbiology* **8**, 2485 (2017).

94 Conlan, S. *et al.* Staphylococcus epidermidis pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. *Genome Biology* **13**, R64, doi:10.1186/gb-2012-13-7-r64 (2012).

95 Otto, M. Staphylococcus epidermidis—the'accidental'pathogen. *Nature reviews microbiology* **7**, 555-567 (2009).

96 Wisplinghoff, H. *et al.* Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. *Clinical infectious diseases* **39**, 309-317 (2004).

97 Fitz-Gibbon, S. *et al.* Propionibacterium acnes strain populations in the human skin microbiome associated with acne. *Journal of investigative dermatology* **133**, 2152-2160 (2013).

98 Grice, E. A. & Segre, J. A. The skin microbiome. *Nature reviews microbiology* **9**, 244-253 (2011).

99 Tomida, S. *et al.* Pan-Genome and Comparative Genome Analyses of Propionibacterium acnes Reveal Its Genomic Diversity in the Healthy and Diseased Human Skin Microbiome. *mBio* **4**, e00003-00013, doi:doi:10.1128/mBio.00003-13 (2013).

100 Odamaki, T. *et al.* Genomic diversity and distribution of Bifidobacterium longum subsp. longum across the human lifespan. *Scientific Reports* **8**, 85, doi:10.1038/s41598-017-18391-x (2018).

101 Yang, S. *et al.* Selective Isolation of Bifidobacterium From Human Faeces Using Pangenomics, Metagenomics, and Enzymology. *Frontiers in Microbiology* **12**, doi:10.3389/fmicb.2021.649698 (2021).

102 Albert, K., Rani, A. & Sela, D. A. Comparative Pangenomics of the Mammalian Gut Commensal Bifidobacterium longum. *Microorganisms* **8**, 7, doi:10.3390/microorganisms8010007 (2019).

103 O'Callaghan, A., Bottacini, F., O'Connell Motherway, M. & van Sinderen, D. Pangenome analysis of Bifidobacterium longum and site-directed mutagenesis through by-pass of restriction-modification systems. *BMC Genomics* **16**, 832, doi:10.1186/s12864-015-1968-4 (2015).

104 Arboleya, S. *et al.* Gene-trait matching across the Bifidobacterium longum pan-genome reveals considerable diversity in carbohydrate catabolism among human infant strains. *BMC Genomics* **19**, 33, doi:10.1186/s12864-017-4388-9 (2018).

105 Joshi, N. & Fass, J. (2011).

106 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

107 Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology* **13**, e1005595 (2017).

108 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

109 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**, 1043-1055, doi:10.1101/gr.186072.114 (2015).

110 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**, 5114, doi:10.1038/s41467-018-07641-9 (2018).

111 Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350-3352, doi:10.1093/bioinformatics/btv383 (2015).

112 Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* **67**, 2640-2644, doi:10.1093/jac/dks261 (2012).

113 Feldgarden, M. *et al.* Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrobial agents and chemotherapy* **63**, e00483-00419, doi:10.1128/AAC.00483-19 (2019).

114 Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**, D32-36, doi:10.1093/nar/gkj014 (2006).

115 Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics (Oxford, England)* **28**, 464-469, doi:10.1093/bioinformatics/btr703 (2012).

116 Galata, V., Fehlmann, T., Backes, C. & Keller, A. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res* **47**, D195-d202, doi:10.1093/nar/gky1050 (2019).

117 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/s0022-2836(05)80360-2 (1990).

118 Chen, L. *et al.* VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* **33**, D325-328, doi:10.1093/nar/gki008 (2005).

119 Zhao, S. *et al.* Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host Microbe* **25**, 656-667.e658, doi:10.1016/j.chom.2019.03.007 (2019).

120 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).

121 Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075, doi:10.1093/bioinformatics/btt086 (2013).

122 Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* **21**, 180, doi:10.1186/s13059-020-02090-4 (2020).

123 Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **48**, D9-d16, doi:10.1093/nar/gkz899 (2020).

124 Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* **48**, D517-d525, doi:10.1093/nar/gkz935 (2020).

125 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).

126 Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693, doi:10.1093/bioinformatics/btv421 (2015).

127 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534, doi:10.1093/molbev/msaa015 (2020).

128 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518-522, doi:10.1093/molbev/msx281 (2017).

129 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587-589, doi:10.1038/nmeth.4285 (2017).

130 Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056, doi:10.1099/mgen.0.000056 (2016).

131 Consortium, T. U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480-D489, doi:10.1093/nar/gkaa1100 (2020).

132 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457-D462, doi:10.1093/nar/gkv1070 (2015).

133 Liu, D., Cole, R. A. & Reeves, P. R. An O-antigen processing function for Wzx (RfbX): a promising candidate for O-unit flippase. *Journal of Bacteriology* **178**, 2102-2107, doi:doi:10.1128/jb.178.7.2102-2107.1996 (1996).

134 Jiang, X.-M. *et al.* Structure and sequence of the rfb (O antigen) gene cluster of Salmonella serovar typhimurium (strain LT2). *Molecular Microbiology* **5**, 695-713, doi:https://doi.org/10.1111/j.1365-2958.1991.tb00741.x (1991).

135 Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* **34**, 2115-2122, doi:10.1093/molbev/msx148 (2017).

136 Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026-1028, doi:10.1038/nbt.3988 (2017).

137 Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366-368, doi:10.1038/s41592-021-01101-x (2021).

138 Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

139 Whelan, F. J., Rusilowicz, M. & McInerney, J. O. Coinfinder: detecting significant associations and dissociations in pangenomes. *Microb Genom* **6**, doi:10.1099/mgen.0.000338 (2020).

140 Crispim, J. S. *et al.* Screening and characterization of prophages in Desulfovibrio genomes. *Scientific Reports* **8**, 9273, doi:10.1038/s41598-018-27423-z (2018).

141 Akhter, S., Aziz, R. K. & Edwards, R. A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic acids research* **40**, e126-e126, doi:10.1093/nar/gks406 (2012).

142 Zhao, S. *et al.* Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host Microbe* **25**, 656-667 e658, doi:10.1016/j.chom.2019.03.007 (2019).

143 Roach, D. J. *et al.* A Year of Infection in the Intensive Care Unit: Prospective Whole Genome Sequencing of Bacterial Clinical Isolates Reveals Cryptic Transmissions and Novel Microbiota. *PLOS Genetics* **11**, e1005413, doi:10.1371/journal.pgen.1005413 (2015).

144 Jiang, X. *et al.* Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science* **363**, 181-187, doi:10.1126/science.aau5238 (2019).

145 Nikitina, A. S. *et al.* Complete Genome Sequence of an Enterotoxigenic Bacteroides fragilis Clinical Isolate. *Genome Announc* **3**, e00450-00415, doi:10.1128/genomeA.00450-15 (2015).

146 Tariq, M. A. *et al.* Genome Characterization of a Novel Wastewater Bacteroides fragilis Bacteriophage (vB_BfrS_23) and its Host GB124. *Frontiers in Microbiology* **11**, doi:10.3389/fmicb.2020.583378 (2020).

147 Kuwahara, T. *et al.* Genomic analysis of Bacteroides fragilis reveals extensive DNA inversions regulating cell surface adaptation. *P Natl Acad Sci USA* **101**, 14919-14924, doi:10.1073/pnas.0404172101 (2004).

148 Rogers, M. B., Parker, A. C. & Smith, C. J. Cloning and characterization of the endogenous cephalosporinase gene, cepA, from Bacteroides fragilis reveals a new subgroup of Ambler class A beta-lactamases. *Antimicrob Agents Chemother* **37**, 2391-2400, doi:10.1128/aac.37.11.2391 (1993).

149 Chung, G. T. *et al.* Identification of a third metalloprotease toxin gene in extraintestinal isolates of Bacteroides fragilis. *Infect Immun* **67**, 4945-4949, doi:10.1128/iai.67.9.4945-4949.1999 (1999).

150 Davies, M. R., Broadbent, S. E., Harris, S. R., Thomson, N. R. & van der Woude, M. W. Horizontally Acquired Glycosyltransferase Operons Drive Salmonellae Lipopolysaccharide Diversity. *PLOS Genetics* **9**, e1003568, doi:10.1371/journal.pgen.1003568 (2013).

151 Sakamoto, M. & Benno, Y. Reclassification of Bacteroides distasonis, Bacteroides goldsteinii and Bacteroides merdae as Parabacteroides distasonis gen. nov., comb. nov., Parabacteroides goldsteinii comb. nov. and Parabacteroides merdae comb. nov. *Int J Syst Evol Microbiol* **56**, 1599-1605, doi:10.1099/ijs.0.64192-0 (2006).

152 Wu, S. *et al.* The Bacteroides fragilis toxin binds to a specific intestinal epithelial cell receptor. *Infect Immun* **74**, 5382-5390, doi:10.1128/iai.00060-06 (2006).

153 Claros, M. *et al.* Characterization of the Bacteroides fragilis pathogenicity island in human blood culture isolates. *Anaerobe* **12**, 17-22 (2006).

154 Kato, N., Kato, H., Watanabe, K. & Ueno, K. Association of enterotoxigenic Bacteroides fragilis with bacteremia. *Clinical infectious diseases* **23**, S83-S86 (1996).

155 Jamal, W. *et al.* Prevalence and antimicrobial susceptibility of enterotoxigenic extra-intestinal Bacteroides fragilis among 13-year collection of isolates in Kuwait. *BMC Microbiol* **20**, 14, doi:10.1186/s12866-020-1703-4 (2020).

156 Sárvári, K. P. *et al.* Detection of enterotoxin and protease genes among Hungarian clinical Bacteroides fragilis isolates. *Anaerobe* **48**, 98-102, doi:10.1016/j.anaerobe.2017.07.005 (2017).

157 Claros, M. C. *et al.* Occurrence of Bacteroides fragilis enterotoxin gene-carrying strains in Germany and the United States. *Journal of clinical microbiology* **38**, 1996-1997, doi:10.1128/JCM.38.5.1996-1997.2000 (2000).

158 Kierzkowska, M. *et al.* The presence of antibiotic resistance genes and bft genes as well as antibiotic susceptibility testing of Bacteroides fragilis strains isolated from inpatients of the Infant Jesus Teaching Hospital, Warsaw during 2007-2012. *Anaerobe* **56**, 109-115, doi:10.1016/j.anaerobe.2019.03.003 (2019).

159 Ulger Toprak, N. *et al.* The distribution of the bft alleles among enterotoxigenic Bacteroides fragilis strains from stool specimens and extraintestinal sites. *Anaerobe* **12**, 71-74, doi:10.1016/j.anaerobe.2005.11.001 (2006).

160 Hecht, A. L., Casterline, B. W., Choi, V. M. & Bubeck Wardenburg, J. A Two-Component System Regulates Bacteroides fragilis Toxin to Maintain Intestinal Homeostasis and Prevent Lethal Disease. *Cell Host Microbe* **22**, 443-448.e445, doi:10.1016/j.chom.2017.08.007 (2017).

161 Casadevall, A. & Pirofski, L.-a. The damage-response framework of microbial pathogenesis. *Nature Reviews Microbiology* **1**, 17-24, doi:10.1038/nrmicro732 (2003).

162 Eitel, Z., Soki, J., Urban, E., Nagy, E. & Infection, E. S. G. o. A. The prevalence of antibiotic resistance genes in Bacteroides fragilis group strains isolated in different European countries. *Anaerobe* **21**, 43-49, doi:10.1016/j.anaerobe.2013.03.001 (2013).

163 Johnsen, B. O. *et al.* erm gene distribution among Norwegian Bacteroides isolates and evaluation of phenotypic tests to detect inducible clindamycin resistance in Bacteroides species. *Anaerobe* **47**, 226-232, doi:https://doi.org/10.1016/j.anaerobe.2017.06.004 (2017).

164 Waters, J. L. & Salyers, A. A. Regulation of CTnDOT conjugative transfer is a complex and highly coordinated series of events. *MBio* **4**, e00569-00513 (2013).

165 Wang, J., Shoemaker, N. B., Wang, G. R. & Salyers, A. A. Characterization of a Bacteroides mobilizable transposon, NBU2, which carries a functional lincomycin resistance gene. *J Bacteriol* **182**, 3559-3571, doi:10.1128/jb.182.12.3559-3571.2000 (2000).

166 Yang, W. *et al.* TetX Is a Flavin-dependent Monooxygenase Conferring Resistance to Tetracycline Antibiotics*. *Journal of Biological Chemistry* **279**, 52346-52352, doi:https://doi.org/10.1074/jbc.M409573200 (2004).

167 Salipante, S. J. *et al.* Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. *Genome research* **25**, 119-128, doi:10.1101/gr.180190.114 (2015).

168 Deng, X., Phillippy, A. M., Li, Z., Salzberg, S. L. & Zhang, W. Probing the pan-genome of Listeria monocytogenes: new insights into intraspecific niche expansion and genomic diversification. *BMC genomics* **11**, 500-500, doi:10.1186/1471-2164-11-500 (2010).

169 Park, S.-C., Lee, K., Kim, Y. O., Won, S. & Chun, J. Large-Scale Genomics Reveals the Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for Estimating Pan-Genome Size. *Frontiers in Microbiology* **10**, doi:10.3389/fmicb.2019.00834 (2019).

170 Ghalayini, M. *et al.* Evolution of a Dominant Natural Isolate of Escherichia coli in the Human Gut over the Course of a Year Suggests a Neutral Evolution with Reduced Effective Population Size. *Appl Environ Microbiol* **84**, doi:10.1128/aem.02377-17 (2018).

171 Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61-66, doi:10.1038/nature23889 (2017).

172 Ramotar, K., Conly, J. M., Chubb, H. & Louie, T. J. Production of menaquinones by intestinal anaerobes. *J Infect Dis* **150**, 213-218, doi:10.1093/infdis/150.2.213 (1984).

173 Kingkaw, A. *et al.* Analysis of the infant gut microbiome reveals metabolic functional roles associated with healthy infants and infants with atopic dermatitis using metaproteomics. *PeerJ* **8**, e9988, doi:10.7717/peerj.9988 (2020).

174 Fedtke, I., Kamps, A., Krismer, B. & Gotz, F. The nitrate reductase and nitrite reductase operons and the narT gene of Staphylococcus carnosus are positively controlled by the novel two-component system NreBC. *Journal of bacteriology* **184**, 6624-6634 (2002).

175 Yeats, C., Rawlings, N. D. & Bateman, A. The PepSY domain: a regulator of peptidase activity in the microbial environment? *Trends in Biochemical Sciences* **29**, 169-172, doi:https://doi.org/10.1016/j.tibs.2004.02.004 (2004).

176 Kopeckova, M., Pavkova, I., Link, M., Rehulka, P. & Stulik, J. Identification of Bacterial Protein Interaction Partners Points to New Intracellular Functions of Francisella tularensis Glyceraldehyde-3-Phosphate Dehydrogenase. *Frontiers in microbiology* **11**, 576618-576618, doi:10.3389/fmicb.2020.576618 (2020).

177 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

178 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic acids research* **39**, W29-W37, doi:10.1093/nar/gkr367 (2011).

179 Vital, M., Howe, A. C., Tiedje, J. M. & Moran, M. A. Revealing the Bacterial Butyrate Synthesis Pathways by Analyzing (Meta)genomic Data. *mBio* **5**, e00889-00814, doi:doi:10.1128/mBio.00889-14 (2014).

180 Jun, S. H., Kim, T. G. & Ban, C. DNA mismatch repair system. Classical and fresh roles. *Febs j* **273**, 1609-1619, doi:10.1111/j.1742-4658.2006.05190.x (2006).

181 Chatzidaki-Livanis, M. *et al.* Gut Symbiont Bacteroides fragilis Secretes a Eukaryotic-Like Ubiquitin Protein That Mediates Intraspecies Antagonism. *mBio* **8**, e01902-01917, doi:doi:10.1128/mBio.01902-17 (2017).

182 Patrick, S. *et al.* A unique homologue of the eukaryotic protein-modifier ubiquitin present in the bacterium Bacteroides fragilis, a predominant resident of the human gastrointestinal tract. *Microbiology (Reading)* **157**, 3071-3078, doi:10.1099/mic.0.049940-0 (2011).

183 Comstock, L. E. *et al.* Analysis of a capsular polysaccharide biosynthesis locus of Bacteroides fragilis. *Infect Immun* **67**, 3525-3532, doi:10.1128/iai.67.7.3525-3532.1999 (1999).

184 Husain, F. *et al.* Novel large-scale chromosomal transfer in Bacteroides fragilis contributes to its pan-genome and rapid environmental adaptation. *Microbial genomics* **3**, e000136, doi:10.1099/mgen.0.000136 (2017).

Chapter 5 : General discussion

5.1 Summary

This Thesis presented results from investigations of the human intestinal microbiota, including an overview of the ME/CFS microbiota (Chapter 2), characterization of an intestinal-associated phage (Chapter 3) and a pangenome analysis of *Bacteroides fragilis* (Chapter 4), an important intestinal opportunistic pathogen. Firstly, the analysis of the ME/CFS microbiota highlighted the heterogenous nature of the disease and the importance of study design in microbiome studies. Second, the first in-depth exploration of *B. fragilis* phage diversity using a curated database of known/unknown phage revealed a novel *B. fragilis* phage family. Finally, the pangenome of 93 *B. fragilis* genomes was investigated and potential genetic differences between different classifications (with respect to whether isolates were enterotoxigenic, clinical or non-clinical) examined.

The metagenomic analysis of the intestinal (faecal) microbiota of 14 ME/CFS patients and five controls did not reveal any specific microbial signatures associated with the disease, which was inconsistent with previous studies (Table 5.1)[1-6].

The patient group in the study described in Chapter 2 exhibited a more diverse microbiota, in composition and predicted function, compared to the controls. However, it is difficult to compare different ME/CFS studies due to differing patient recruitment criteria and disease stratification. For example, Chapter 2 describes the only study to date that has focussed solely on severe and very severe ME/CFS patients. Previous studies, as outlined in Table 5.1, have recruited far more patients than in this study. Additionally, the majority of the studies used 16S rRNA gene sequencing, instead of metagenomic sequencing used in Chapter 2. Additionally, none of the previous studies recruited participants from the United Kingdom and geographical variation in the intestinal microbiota is well documented[8]. This could further explain why the results in Chapter 2 are not consistent with previous studies.

Table 5-1: Overview of number of participants, diagnostic criteria, type of study and recruitment location for ME/CFS studies where the information is available

Study	Participants	Diagnostic Criteria	Type of study	Recruitment location	Reference
Chapter 2	14 severe ME/CFS and 5 healthy controls	Fukuda, Canadian Criteria and Oxford Criteria	Shotgun sequencing	United Kingdom	N/A
Lupo (2021)	35 ME/CFS and 70 healthy controls	Fukuda	16S rRNA gene sequencing & metabolomics	Italy	[1]
Armstrong (2017)	34 ME/CFS and 25 healthy controls	Canadian Criteria	Culture & metabolomics	America	[7]
Nagy-Szakal (2017)	50 ME/CFS and 50 healthy controls	Fukuda and/or Canadian Criteria	Shotgun sequencing	America	[3]
Giloteaux (2016)	Monozygotic twin pair (1 ME/CFS and 1 control)	Fukuda	16S rRNA gene sequencing	America	[5]
Giloteaux (2016)	49 ME/CFS and 39 healthy controls	Fukuda	16S rRNA gene sequencing	America	[4]
Fremont (2013)	43 ME/CFS and 36 healthy controls	Fukuda	16S rRNA gene sequencing	Norway & Belgium	[6]

Sample size in microbiota studies has been strongly associated with beta diversity measures and a larger sample size normally decreases heterogeneity within cohorts[9]. Power calculations are used to determine numbers of samples required in studies to allow robust data to be generated to detect a statistically relevant difference between patients and controls. Calculations can only be done if those designing studies can specify *"(i) the smallest relevant deviation from the null hypothesis that is to be detected at some specified significance level, and (ii) a realistic guess of the variability in the sample*"[10]. In a microbiota study comparing patients and controls, the null hypothesis is that there are no differences in the microbiota composition of the two groups. Mattiello et al. (2016) proposed a means of generating power calculations for microbiota studies, modelling abundance data using a Dirichlet-Multinomial distribution[10]. Inputting the number of controls (*n*=5) and patients (*n*=14) from my study and carrying out an analysis based on non-stratification of samples, I found that under all criteria tested, the sample sizes used in Chapter 2 do not provide sufficient power to draw meaningful conclusions from microbiota-based study data (Figure 5.1).

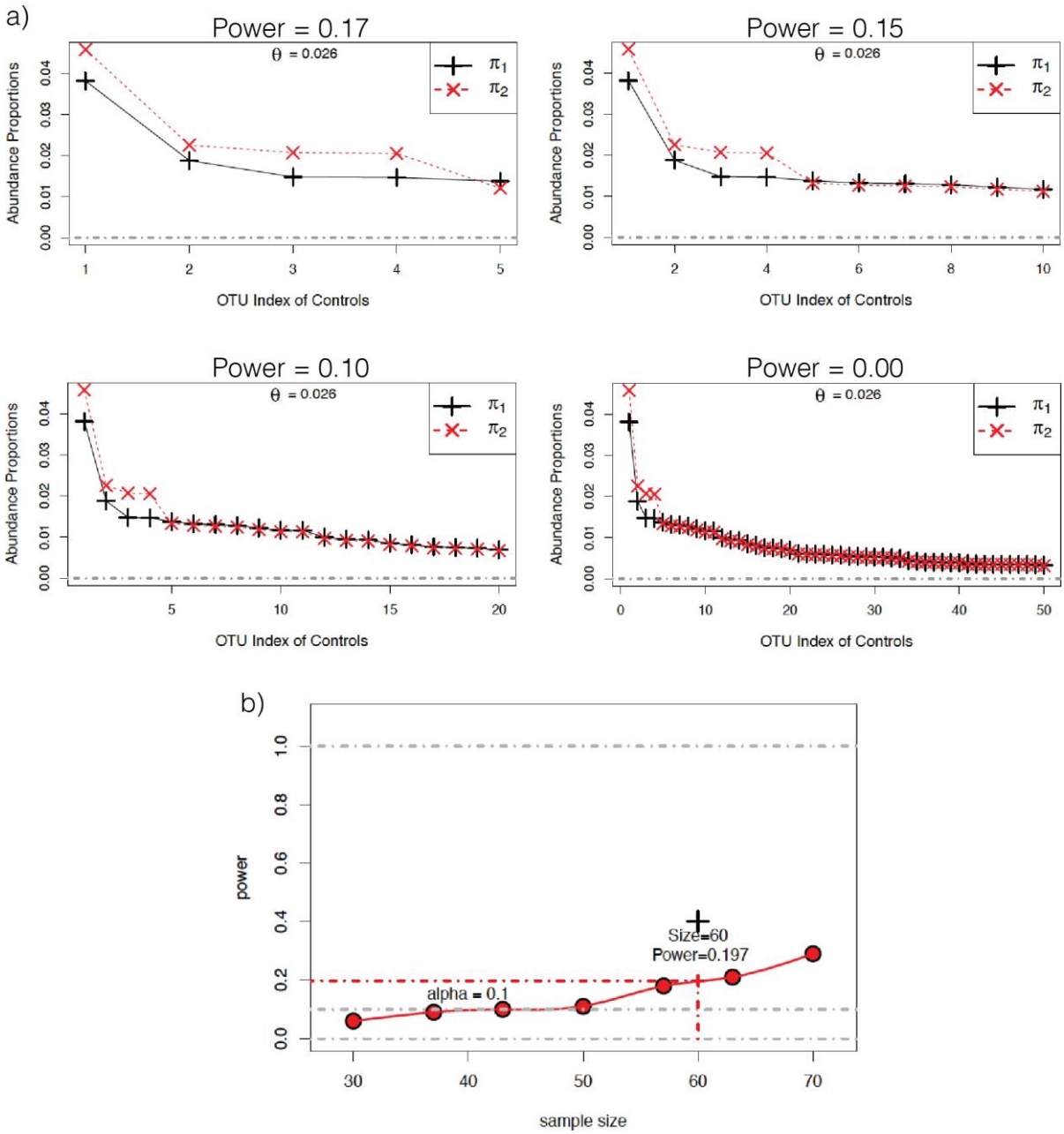

Figure 5.1: Power calculations determined for number of patients (*n*=5, ν1) and controls (*n*=14, ν2) included in the ME/CFS study

a) Power calculations were done by means of a Monte Carlo approach (100 replications) in which, for the given sample sizes, data of *k* operational taxonomic units (5, 10, 20 and 50 in the examples shown) were randomly generated from a Dirichlet-Multinomial distribution using the default stool model (derived from Human Microbiome Project data). θ, Within-sample excess of variability with respect to a multinomial distribution. (b) A minimum of 60 patients and 60 controls would need to be included in a study comparing the two groups (patient and control, without stratification) with an alpha value of 0.1 and power of 0.2. All analyses shown in a) produced the same outcome for b).

A minimum of 60 patient and 60 controls would need to be included to undertake a properly powered study (Figure 5.1b). Patients with severe and very severe disease were included in the study, but were not stratified by disease severity in the analyses undertaken in Chapter 2. As ME/CFS is a heterogeneous disease, stratification of patients by disease severity would need to be considered in power calculations, as would confounders[9,10]. It was difficult to recruit patients with severe ME/CFS during this study, which introduced logistical challenges for sample collection. Recruitment of additional participants to a sufficiently powered study would also be required to account for potential drop-outs[10]. This highlights the necessity for careful study design, especially in multi-factorial complex diseases such as ME/CFS. Furthermore, no ME/CFS studies to date (Table 5.1) have recruited sufficient patients and controls to meet adequate power based on the microbiota-based analysis done here (Figure 5.1).

A recent study used activity bracelets, cardiopulmonary exercise testing and a validated activity questionnaire to confirm the severity grading self-reported by ME/CFS patients[11]. These techniques could be used in future studies to group patients according to quantifiable severity measures, instead of self-reporting. In addition to disease severity, onset event and symptom presentation can vary significantly between patients[12]. Therefore, future studies should also attempt to collect metadata regarding disease onset and symptom types. Additionally, as shown by previous studies, the microbiota in ME/CFS patients with and without irritable bowel syndrome (IBS) show marked differences[3,13,14]. It is not known if the patients from this study were co-morbid for IBS.

As mentioned in Chapter 1 and Appendix 1, patient symptom type and severity can vary daily or weekly depending on patient levels of physical and/or mental exertion[15]. Microbiota alterations should be examined longitudinally, preferably with a symptom diary, instead of single 'snap-shot' samples to gain a true picture of the ME/CFS microbiota. During this study household controls were recruited (normally female first-relatives); however, this may not have been the best course of action. The occurrence of ME/CFS within female relatives of the same family is well documented[16]. Therefore, the use of female first relatives for controls may introduce an unknown confounding factor and cloud interpretation of the data. A 2021 microbiota study recruited external controls unrelated to the patient and familial controls[1]. This allowed the authors to show that first relatives of ME/CFS patients shared a closer microbiota to patients than external controls. It also allowed the authors to negate several confounding factors, as the familial control and ME/CFS patient most likely have similar genetics, diet and living environment. Previous microbiota studies, such as a recent autism spectrum disorder (ASD) study, have used first-degree

relatives to mitigate any genetic confounding factors. This allowed authors to determine that the autism-gut microbiome associations noted in ASD are due to autism-related dietary preferences rather than the disorder itself[17]. Therefore, future studies should aim to recruit external and relative controls or collect sufficient metadata (e.g. food diary) to account for several confounding factors. The effect of host variables on microbiota disease analysis is well known[9]. A 2021 study examined specific covariates and the effect on microbiota study outcome[9]. The authors used faecal metagenomes available through American Gut to determine the robustness of microbiota associations when host variables are accounted for. By matching subjects and controls according to confounding variables, the authors were able to determine the magnitude of microbial differences, compared to non-matching subject and control. Studies in patients with type II diabetes (T2DM) showed the largest decrease in microbial association following covariate matching, and the authors attributed this to differences in alcohol intake and bowel movement quality. Additionally, inflammatory bowel disease was shown to have the greatest microbial alterations between co-variate matched cases and controls. This study highlighted the importance of careful and meticulous subject selection and control matching to reduce false-positive microbiota disease associations.

Given the complexity of ME/CFS and the involvement of multiple body systems, additional 'omics techniques should be used to examine the disease. Lupo (2021) used metabolomics to complement metagenomics[1]. Once a complete picture of the microbiota in ME/CFS has been established, strain-level analysis should be carried out. For example, a recent study reported a decrease in the butyrate-producing bacteria *Faecalibacterium*, *Roseburia*, and *Eubacterium* in subjects with ME/CFS[18]. Alterations in these microbes have been associated with various metabolic diseases, such as obesity, T2DM and liver disease, raising the question of if ME/CFS is a metabolic disease or a microbiota-related disease?[19] Reduced metabolic potential has been suggested to explain the aetiology of ME/CFS and it has been proposed that ME/CFS patients exhibit extensive metabolic dysregulation via a defect in the tricarboxylic acid cycle in the mitochondria. The resulting decreased production of adenosine triphosphate and excessive lactate production upon exertion could explain the variety of symptoms, especially delayed fatigue onset[20,21]. StrainPhlAn and PanPhlAn, mentioned in Chapter 1, could be used to determine if the population structure of these strains was similar between patients and if there were any genomic differences present that could also account for the reduced SCFA levels also reported in the faecal metabolome[18,22,23]. Serum and urine metabolomics may also allow identification of host-associated metabolic pathways perturbed by ME/CFS, as these may be more relevant to disease progression and/or maintenance than the intestinal microbiome.

In Chapter 3 I curated a custom dataset of 2,636 *Bacteroides* phage from three databases: NCBI Virus, IMGVR and GPD. This revealed an extensive amount of unexplored diversity of *Bacteroides* phage. Gene-sharing network analysis showed the 2,636 *Bacteroides* phage grouped into 100 viral clusters (VCs), representing potentially 100 uncharacterised phage taxonomic groups. Furthermore, two of these VCs were unrelated to any phage currently characterised and present an interesting avenue for future research. Investigating one of the VCs revealed a previously undescribed potential *B. fragilis* phage family and contained all phage currently isolated with *B. fragilis.* A wide range of intestinal metagenome studies should be screened to examine the geographical and age distribution of this phage family. As mentioned in the Discussion of Chapter 3, additional novel phage taxonomic groups could have been determined if additional phage genomes were added to the gene-sharing network analysis, such as the database carefully curated by MillardLab.Due to the low sequence similarity noted between structurally related phage, it is imperative thatnovel phage are classified according to genetic similarity[24]. A 2021 publication by Turner et al. proposed the abolishment of the order *Caudovirales* and families *Myoviridae, Podoviridae* and *Siphoviridae* to allow for reclassification of phage families that are based on evolutionary relationship[25].

A disadvantage of metagenomic phage characterisation is that the physical phage are not available for phenotypic and bacteria-host interaction assays. Four of the phage within the novel *B. fragilis* phage family described in Chapter 3 have been physically isolated; two with isolate GB-124, one with isolate HSP-40 and one unknown isolate. Therefore, a combination of physical phage isolation and metagenomic phage discovery techniques should be used to elucidate phage lifestyle and interaction with bacterial host. Due to the narrow host range of known *B. fragilis* phage, stability of *B. fragilis* strains within a host and individualization of *B. fragilis* strains, isolation of *B. fragilis* phage with a *B. fragilis* strain from the same individual may provide a better insight in the interplay between phage and host. Furthermore, annotation of the phage genomes could have been improved by using the Prokaryotic Virus Remote Homologous Groups database (PHROGs)[26]. This database contains 38,880 protein orthologous groups from ~ 15,000 phage genomes (including prophages).

A total of 93 *B. fragilis* genomes were collected for the pangenome analysis described in Chapter 4. The completion of this Chapter was hindered greatly by the low quality of publicly available *B. fragilis* genomes. Additionally, the lack of metadata associated with the genomes limited any potential conclusions that could be drawn from the analysis. For example, the majority of the 'enterotoxigenic' isolates did not contain a detectable *bft* gene so it was unknown if these isolates

were truly causing inflammatory diarrheal disease or isolation was coincidence. Furthermore, very few of the isolates on NCBI had a published article attached. The characterization of the remaining isolates was only determined by the minimal information provided by the depositing scientist, such as host disease. Of the four ETBF isolates that did contain a referenced article on the NCBI profile, the presence of the BFT was confirmed in only three isolates (via Western blot)[27]. It is also unknown if the depositing researchers performed a PCR to confirm the presence of a *bft* gene. This has highlighted to me the need for data curation prior to pangenome analysis. During genome selection, it became evident that the number of *B. fragilis* isolates recovered and characterise from healthy subjects was very low. Given the pivotal role *Bacteroides* spp. play in the maintenance of health, efforts should be made to increase the number of isolates from healthy individuals within publicly available databases.

The pangenome analysis could have been improved by the addition of phenotypic assays, such as antimicrobial resistance profiling, secretion of *bft,* oxygen tolerance assays and other assays related to virulence. With phenotypic data, genome wide-association studies (GWAS) could have been undertaken, instead of grouping the isolates according to apparent arbitrary classifications (non-clinical, clinical, 'enterotoxigenic').

This study confirmed the complexity and diversity of outer polysaccharides of *B. fragilis* seen in previous studies[28-31]. The genomes collected during this study appeared to cluster partly due to the *rfb* gene diversity, and potentially polysaccharide (PS) diversity. The PS loci within *B. fragilis* comprise various regulatory genes, diverse glycosyltransferases and an intergenic promoter region[32,33]. However, it is not possible to examine intergenic regions from the standard output of a pangenome analysis. The approach used by Roary excludes non-coding protein regions, such as intergenic regions, that can account for up to 15 % of the genome[34,35]. A recently developed bioinformatic tool, Piggy, identifies core and accessory intergenic regions with the standard Roary output[34]. This would allow detection of "switched" intergenic regions within the *B. fragilis* pangenome and the downstream genes affected. Determination of these "switched" intergenic regions may explain the explain the ability of non-clinical faecal *B. fragilis* isolates to thrive outside the intestinal environment and cause infection, as has been suggested with *Pseudomonas aeruginosa.* During infection in cystic fibrosis patients, intergenic changes within *P. aeruginosa* are strongly positively selected for and may play a pivotal in persistence of infection[36]. Furthermore, bioinformatic analysis of the PS regions in *B. fragilis* should be complemented with phenotypic assays, such as evolutionary studies to determine the factors for PS switching (e.g. phage challenge or co-culture with other microbes). As noted previously, a single population can express

multiple PSs and this variation would not be detected by short-read whole genome sequencing[30]. Therefore, the use of transcriptomics and hybrid genome assemblies should be explored to determine the proportion of different PSs expressed in a single *B. fragilis* population.

Gene association and disassociation studies were attempted in Chapter 4; however, no gene associations were found. This could be because no gene dis-/associations exist, or the small number of genomes analysed. The authors of Coinfinder examined the effect of sample size on Coinfinder's ability to discover gene-gene associations by subsetting 534 *Streptococcus pneumoniae* genomes into datasets of between 400 and 50 genomes[37]. They reported that as sample size decreased, the ability of Coinfinder to confidently discover gene-gene associations decreased substantially. Furthermore, no associations were detected with a 50-genome dataset. Increasing the number of *B. fragilis* genomes or including additional members of the genus *Bacteroides* could increase the likelihood of detecting gene dis-/associations.

This Thesis presents the intestinal microbiome of severe ME/CFS patients compared to controls, characterized a novel *Bacteroides fragilis* phage isolated from sewage water and explored the pangenome of phenotypically distinct *Bacteroides fragilis* strains. Investigation of the intestinal microbiome of severe ME/CFS patients revealed a high level of compositional heterogeneity that has not been widely reported previously. Furthermore, a novel *B.fragilis* phage was discovered from sewage effluent and the analysis of all *Bacteroides* metagenome-assembled phage genomes revealed a novel genus. The phage within the novel genera showed high similarity at genomic and protein level. Additionally, the genomic differences of 93 *Bacteroides fragilis* strains of intestinal or systemic origin were explored. This Chapter was the first known in-depth study of the *B.fragilis* pangenome and revealed no significant genetic differences of strains with differing lifestyles. However, the results suggested polysaccharide capsule loci contribute to genomic diversity and forms a good basis for further study.

5.2 References

1 Lupo, G. F. D. *et al.* Potential role of microbiome in Chronic Fatigue Syndrome/Myalgic Encephalomyelits (CFS/ME). *Sci Rep* **11**, 7043, doi:10.1038/s41598-021-86425-6 (2021).

2 Armstrong, C. W., McGregor, N. R., Lewis, D. P., Butt, H. L. & Gooley, P. R. The association of fecal microbiota and fecal, blood serum and urine metabolites in myalgic encephalomyelitis/chronic fatigue syndrome. *Metabolomics* **13**, 8, doi:10.1007/s11306-016-1145-z (2016).

3 Nagy-Szakal, D. *et al.* Fecal metagenomic profiles in subgroups of patients with myalgic encephalomyelitis/chronic fatiguesyndrome. *Microbiome* **5**, 44, doi:10.1186/s40168-017-0261-y (2017).

4 Giloteaux, L. *et al.* Reduced diversity and altered composition of the gut microbiome in individuals with myalgicencephalomyelitis/chronic fatigue syndrome. *Microbiome* **4**, 30, doi:10.1186/s40168-016-0171-4 (2016).

5 Giloteaux, L., Hanson, M. R. & Keller, B. A. A Pair of Identical Twins Discordant for Myalgic Encephalomyelitis/ChronicFatigue Syndrome Differ in Physiological Parameters and Gut Microbiome Composition. *Am J Case Rep* **17**, 720-729, doi:10.12659/ajcr.900314 (2016).

6 Fremont, M., Coomans, D., Massart, S. & De Meirleir, K. High-throughput 16S rRNA gene sequencing reveals alterations ofintestinal microbiota in myalgic encephalomyelitis/chronic fatigue syndrome patients. *Anaerobe* **22**, 50-56, doi:10.1016/j.anaerobe.2013.06.002 (2013).

7 Alfayyadh, M. *et al.* Recurrence of Clostridium difficile infection in the Western Australian population. *Epidemiol Infect* **147**,e153, doi:10.1017/s0950268819000499 (2019).

8 Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e620, doi:https://doi.org/10.1016/j.cell.2019.01.001 (2019).

9 Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448-454,doi:10.1038/s41586-020-2881-9 (2020).

10 Mattiello, F. *et al.* A web application for sample size and power calculation in case-control microbiome studies. *Bioinformatics* **32**, 2038-2040, doi:10.1093/bioinformatics/btw099 (2016).

11 van Campen, C., Rowe, P. C. & Visser, F. C. Validation of the Severity of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome by Other Measures than History: Activity Bracelet, Cardiopulmonary Exercise Testing and a Validated Activity Questionnaire: SF-36. *Healthcare (Basel)* **8**, doi:10.3390/healthcare8030273 (2020).

12 Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/Chronic Fatigue, S., Board on the Health of Select, P. & Institute of, M. in *Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Redefining an Illness* (National Academies Press (US) Copyright 2015 by the National Academy of Sciences. All rights reserved., 2015).

13 Aaron, L. A., Burke, M. M. & Buchwald, D. Overlapping conditions among patients with chronic fatigue syndrome, fibromyalgia, and temporomandibular disorder. *Arch Intern Med* **160**, 221-227, doi:10.1001/archinte.160.2.221 (2000).

14 Quigley, E. M. The enteric microbiota in the pathogenesis and management of constipation. *Best Pract Res Clin Gastroenterol* **25**, 119-126, doi:10.1016/j.bpg.2011.01.003 (2011).

15 Cortes Rivera, M., Mastronardi, C., Silva-Aldana, C. T., Arcos-Burgos, M. & Lidbury, B. A. Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: A Comprehensive Review. *Diagnostics (Basel)* **9**, doi:10.3390/diagnostics9030091 (2019).

16 Valdez, A. R. *et al.* Estimating Prevalence, Demographics, and Costs of ME/CFS Using Large Scale Medical Claims Data and Machine Learning. *Front Pediatr* **6**, 412, doi:10.3389/fped.2018.00412 (2018).

17 Yap, C. X. *et al.* Autism-related dietary preferences mediate autism-gut microbiome associations. *Cell* **184**, 5916-5931.e5917, doi:https://doi.org/10.1016/j.cell.2021.10.015 (2021).

18 Guo, C. *et al.* Deficient butyrate-producing capacity in the gut microbiome of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome patients is associated with fatigue symptoms. *medRxiv*, 2021.2010.2027.21265575, doi:10.1101/2021.10.27.21265575 (2021).

19 Wang, P.-X., Deng, X.-R., Zhang, C.-H. & Yuan, H.-J. Gut microbiota and metabolic syndrome. *Chin Med J (Engl)* **133**, 808-816, doi:10.1097/CM9.0000000000000696 (2020).

20 Naviaux, R. K. *et al.* Metabolic features of chronic fatigue syndrome. *Proc Natl Acad Sci U S A* **113**, E5472-5480, doi:10.1073/pnas.1607571113 (2016).

21 Fluge, Ø. *et al.* Metabolic profiling indicates impaired pyruvate dehydrogenase function in myalgic encephalopathy/chronic fatigue syndrome. *JCI Insight* **1**, e89376, doi:10.1172/jci.insight.89376 (2016).

22 Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, doi:10.7554/eLife.65088 (2021).

23 Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res* **27**, 626-638, doi:10.1101/gr.216242.116 (2017).

24 Dion, M. B., Oechslin, F. & Moineau, S. Phage diversity, genomics and phylogeny. *Nature Reviews Microbiology* **18**, 125-138 (2020).

25 Turner, D., Kropinski, A. M. & Adriaenssens, E. M. A Roadmap for Genome-Based Phage Taxonomy. *Viruses* **13**, 506, doi:10.3390/v13030506 (2021).

26 Terzian, P. *et al.* PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics and Bioinformatics* **3**, doi:10.1093/nargab/lqab067 (2021).

27 Pierce, J. V. & Bernstein, H. D. Genomic Diversity of Enterotoxigenic Strains of Bacteroides fragilis. *PLOS ONE* **11**, e0158171, doi:10.1371/journal.pone.0158171 (2016).

28 Comstock, L. E. *et al.* Analysis of a capsular polysaccharide biosynthesis locus of Bacteroides fragilis. *Infect Immun* **67**, 3525-3532, doi:10.1128/iai.67.7.3525-3532.1999 (1999).

29 Patrick, S., Gilpin, D. & Stevenson, L. Detection of intrastrain antigenic variation of Bacteroides fragilis surface polysaccharides by monoclonal antibody labelling. *Infection and immunity* **67**, 4346-4351, doi:10.1128/IAI.67.9.4346-4351.1999 (1999).

30 Lutton, D. A. *et al.* Flow cytometric analysis of within-strain variation in polysaccharide expression by Bacteroides fragilis by use of murine monoclonal antibodies. *J Med Microbiol* **35**, 229-237, doi:10.1099/00222615-35-4-229 (1991).

31 Patrick, S. *et al.* Twenty-eight divergent polysaccharide loci specifying within- and amongst-strain capsule diversity in three strains of Bacteroides fragilis. *Microbiol-Sgm* **156**, 3255-3269 (2010).

32 Bayley, D. P., Rocha, E. R. & Smith, C. J. Analysis of cepA and other Bacteroides fragilis genes reveals a unique promoter structure. *FEMS Microbiology Letters* **193**, 149-154, doi:10.1111/j.1574-6968.2000.tb09417.x (2000).

33 Jiang, X. *et al.* Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science* **363**, 181-187, doi:10.1126/science.aau5238 (2019).

34 Thorpe, H. A., Bayliss, S. C., Sheppard, S. K. & Feil, E. J. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *GigaScience* **7**, doi:10.1093/gigascience/giy015 (2018).

35 McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology* **10**, 13-26 (2012).

36 Khademi, S. H. & Jelsbak, L. Host adaptation mediated by intergenic evolution in a bacterial pathogen. *bioRxiv*, 236000 (2017).

37 Whelan, F. J., Rusilowicz, M. & McInerney, J. O. Coinfinder: detecting significant associations and dissociations in pangenomes. *Microb Genom* **6**, doi:10.1099/mgen.0.000338 (2020).

Appendix 1

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

PORTLAND PRESS

Review Article

Does the microbiome and virome contribute to myalgic encephalomyelitis/chronic fatigue syndrome?

Fiona Newberry[1,2], Shen-Yuan Hsieh[1,2], Tom Wileman[1,2] and Simon R. Carding[1,2]

[1]Norwich Medical School, University of East Anglia, Norwich NR4 7TJ, U.K.; [2]The Gut Health and Food Safety Research Programme, The Quadram Institute, Norwich Research Park, Norwich, U.K.

Correspondence: Fiona Newberry (fiona.newberry@quadram.ac.uk)

OPEN ACCESS

Myalgic encephalomyelitis (ME)/chronic fatigue syndrome (CFS) (ME/CFS) is a disabling and debilitating disease of unknown aetiology. It is a heterogeneous disease characterized by various inflammatory, immune, viral, neurological and endocrine symptoms. Several microbiome studies have described alterations in the bacterial component of the microbiome (dysbiosis) consistent with a possible role in disease development. However, in focusing on the bacterial components of the microbiome, these studies have neglected the viral constituent known as the virome. Viruses, particularly those infecting bacteria (bacteriophages), have the potential to alter the function and structure of the microbiome via gene transfer and host lysis. Viral-induced microbiome changes can directly and indirectly influence host health and disease. The contribution of viruses towards disease pathogenesis is therefore an important area for research in ME/CFS. Recent advancements in sequencing technology and bioinformatics now allow more comprehensive and inclusive investigations of human microbiomes. However, as the number of microbiome studies increases, the need for greater consistency in study design and analysis also increases. Comparisons between different ME/CFS microbiome studies are difficult because of differences in patient selection and diagnosis criteria, sample processing, genome sequencing and downstream bioinformatics analysis. It is therefore important that microbiome studies adopt robust, reproducible and consistent study design to enable more reliable and valid comparisons and conclusions to be made between studies. This article provides a comprehensive review of the current evidence supporting microbiome alterations in ME/CFS patients. Additionally, the pitfalls and challenges associated with microbiome studies are discussed.

What is the microbiome?

Virtually every surface of the human body is colonized by vast populations of microbes, including prokaryotes, archaea, viruses, fungi and unicellular eukaryotes [1-3]. Bacteria of the phyla Bacteriodetes and Firmicutes dominate the diverse and complex intestinal bacteriome of most animals [4]. Microbial colonization begins rapidly at birth when the infant is first exposed to microbes in its immediate environment. The microbiome increases in diversity during the first 2–4 years of life in response to various hosts (i.e. genetics), and environmental factors including diet, lifestyle and behaviour [5-7]. It is believed that the early colonizers of the infant intestine play a key role in laying the foundations for the development of the complex and diverse adult microbiome and lifelong health [8]. In recent years, the role of microbiome in health of the host and its contribution to disease development has emerged [9-11]. It contributes to various body systems including immunity, metabolism, neurological signalling and homeostasis [12,13].

Received: 08 December 2017
Revised: 11 January 2018
Accepted: 16 January 2018

Version of Record published:
09 March 2018

PORTLAND PRESS

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

Table 1 Overview of important faecal virome studies in health and disease

Year	Study participants	Comments	Reference
2003	Healthy adults	First virome metagenomics study	[168]
2006	Healthy adults	Plant RNA viruses contribute towards virome	[169]
2008	Infants	Virome establishment begins within 1 week of birth	[21]
2011	Healthy adults	Diversity and abundance of ssDNA viruses	[170]
2011	Monozygotic twins and mothers	Virome is individualized and highly stable	[22]
2011	Healthy adults	Virome is influenced by diet	[157]
2012	Healthy adults	Hypervariation driven by unique reverse transcriptase based mechanism	[171]
2013	Healthy adult	Virome is relatively stable; 80% of virome remained through 2.5-year study	[172]
2013	Pediatric CD patients	CD patients exhibited higher bacteriophage levels than controls	[49]
2013	CD patients	Similar results as above; results depend on interpretation of data	[173]
2015	Infants	Longitudinal study of virome establishment in infant twins	[174]
2015	Malnourished Malawian twins	Virome establishment affected by severe malnourishment	[178]
2015	IBD patients	Virome in IBD patients	[51]
2015	IBD patients	Increase in phage-richness abundance compared with healthy controls	[175]
2015	CD patients	Alterations in virome according to disease status and therapy	[58]

Abbreviations: CD, Crohn's disease; IBD; inflammatory bowel disease.

Describing the microbiome in detail is beyond the scope of this article, however several excellent review papers have been published recently [7,14-17].

The neglected virome

The vast majority of microbiome research has to date focused on its bacterial component, largely neglecting the other organisms. However, the influence of these lesser studied organisms, such as viruses are just beginning to be understood, thanks primarily to recent advancements in sequencing technology and bioinformatics capability (Table 1) [18].

It is estimated that there are 10^{31} different DNA and RNA viruses on the planet; many of which remain undiscovered [19]. This collection of viruses (dsDNA, ssDNA, dsRNA and ssRNA) within an ecosystem is defined as the virome [20]. Similar to the bacteriome, the intestinal virome is established from birth and increases in diversity/complexity with age [21]. A large proportion of this complex environment consists of prokaryotic viruses (bacteriophage); with archaea-, human-, plant- and amoeba-infecting viruses found at lower frequencies [20]. The tailed, dsDNA viruses of the Order Caudovirales (Siphoviridae, Myoviridae, Podoviridae) dominate the bacteriophage portion of the virome (Figure 1) [22].

The microbiota in health and disease

The importance of the intestinal microbiome in maintaining health is an emerging research topic with advances in high-throughput sequencing technology allowing the identification and characterization of microbes that contribute to host health [10,11]. The microbiota has been implicated in immunomodulation, pathogen resistance, maintenance of intestine structure/function and nutrition and host metabolism [12,13]. It provides the host with a physical barrier to pathogen invasion and infection by, for example competitive exclusion and competing for nutrients, occupation of attachment sites and production of antimicrobial proteins [23-26]. Importantly, various microbiome members have been found to contribute to the intestinal metabolome, through for example vitamin synthesis, bile salt metabolism and xenobiotic degradation [27]. There is bidirectional communication between the microbiome and the local host immune system [28]. The immune system influences the composition of the microbiota and gut microbes and their products (e.g. metabolites and microbe associated molecular pattern (MAMPs) molecules) direct immune maturation and the development and possibly maintenance of immune (microbial) tolerance and homoeostasis [29,30].

There is increasing evidence that an imbalance of the intestinal microbiota (dysbiosis) may contribute to the pathogenesis of diseases affecting the gastrointestinal (GI) tract and other organ systems. Dysbiosis is characterized by a detrimental alteration of intestine microbial populations and ecology that can result in the growth and expansion of pathogenic microbes (pathobionts) and the production of factors toxic or harmful to host cells. These alterations

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

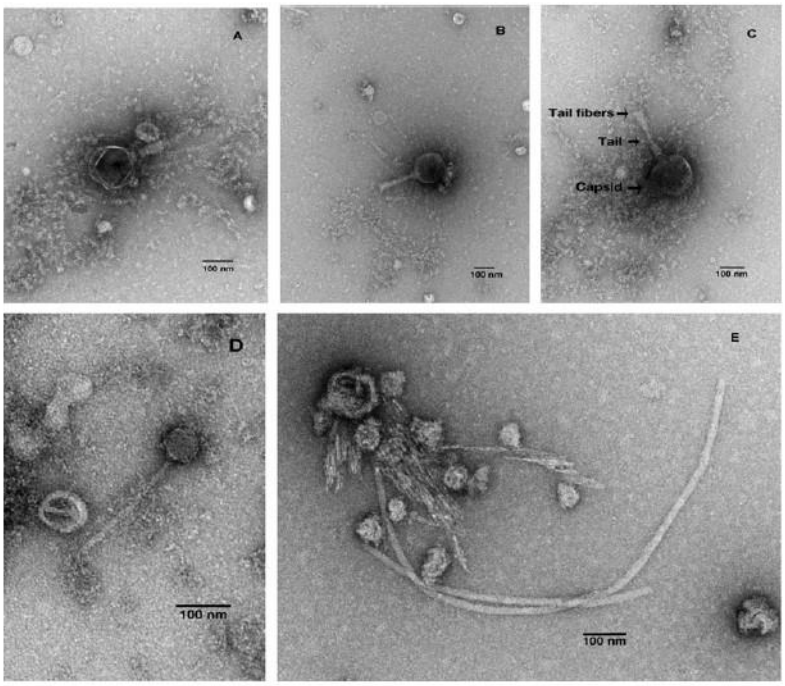

Figure 1. TEM images of Caudovirales from faecal water
(**A**–**C**) Myoviridae and (**D**,**E**) Siphoviridae. Imaging completed by S.-Y.H. and K.C.

are normally held in check by an intact microbiome but dysbiosis can result in the development and/or maintenance of chronic inflammatory infections caused by *Clostridium difficile* and *Helicobacter pylori*, metabolic syndrome and obesity, colorectal cancer, irritable bowel syndrome (IBS) and inflammatory bowel disease (IBD) [14,31-34]. The stability of the microbiome is largely influenced by age, behaviour and lifestyle [12,35,36]. It has been hypothesized that intestinal microbial dysbiosis can lead to an imbalance in the immune system, resulting in diseases such as IBD, common variable immunodeficiency and rheumatoid arthritis [37,38]. However, to understand the significance of dysbiosis, it is necessary to establish if microbiome alterations cause, follow, precede or simply correlate with disease onset. Dysbiosis can be precipitated by drugs and medications (i.e. antibiotics), immune dysregulation, age-associated reduction in microbiota diversity, colonization by pathogenic microbes, stress and changes in diet [36,38-41]. The precise trigger or cause of dysbiosis in any disease has yet to be established but is likely to be multifactorial.

The virome in health

Viruses utilize a lytic or lysogenic life cycle. In the lytic life cycle, infected host cells are destroyed during viral replication whereas in the lysogenic life cycle the virus integrates into the host chromosome as a prophage. Lytic phages can have both narrow or broad host ranges, and lysogenic phages can be converted into a lytic cycle in response to environmental stressors such as antibiotics [20]. Lytic phages can alter the microbiome by killing bacterial hosts, providing a competitive growth advantage to bacteria resistant to phages. Prophages encode mobile genetic elements which contribute to horizontal gene transfer between bacteria altering antibiotic resistance, virulence or metabolic pathways [42]. This can provide a competitive advantage by allowing bacteria to metabolize new nutrient sources or acquire antibiotic resistance [43-45]. Temperate phages, able to perform lysogenic or lytic cycle, have been shown to influence

PORTLAND PRESS

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

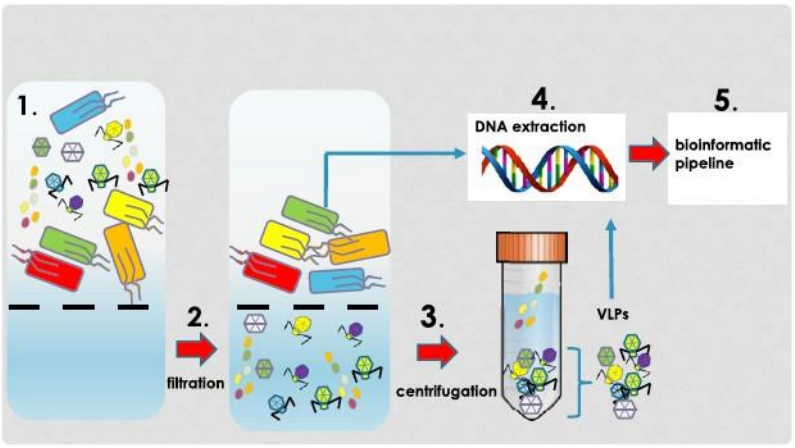

Figure 2. VLP isolation protocol

Overview of VLP isolation protocol involving filtration (**1**,**2**) and centrifugation (**3**). Following isolation of concentration VLPs, DNA is extracted (**4**), sequencing and (**5**) bioinformatic tools applied to determine virome community composition.

the dynamics of biofilms and dispersal by a number of important pathogens such as *Pseudomonas aeruginosa* (opportunistic pathogen), *Streptococcus pneumoniae* (e.g. pneumonia) and *Bacillus anthracis* (e.g. anthrax) [46-48]. For example, the presence of lysogenic phages Bcp1, Wip1, Wip4 and Frp2 in *B. anthracis* results in the formation of a durable, complex and viable biofilm; allowing prolonged survival of the bacterium [48]. There is a constant shift of phages between lytic and lygosenic forms that is presumed to contribute to microbiome homoeostasis and that a differential spatial distribution of phages is correlated with health.

The virome in disease

Alterations in the virome have been implicated as sources of intestinal microbial (prokaryotic) dysbiosis for several different diseases [49-52]. Prophage induction in response to various environmental stressors can induce 'community shuffling' which alters the ratio of symbionts to pathobionts creating an imbalance within microbial communities that can lead to occupation of symbiont niches by pathobionts [42]. These events provide an explanation for the raised number of virus-like particles (VLPs, see Figure 2) as well as microbial population shifts in patients with GI-related disorders. Of note, an experimental model of *Salmonella typhimurium* diarrhoea has shown that inflammation increases lysogenic conversion of prophages [53].

The involvement of lytic phages in disease pathogenesis has been demonstrated through studying prophages that encode virulence factors (e.g. Shiga toxin) [54]. In the healthy intestine, toxin gene expression is silent in lysogenic phages that infect *Escherichia coli*. However, dysbiosis is accompanied by induction of prophages and activation of Shiga toxin genes resulting in release of the toxin into the intestine [55]. Additionally, *in vitro* experiments have shown phages can transmigrate across epithelial barrier cells. A recent study suggests that intestinal phages can interact directly with eukaryotic cells outside the GI tract; likely contributing to human health and immunity [56].

Although the virome is suspected to play a role in disease, relatively few studies have been undertaken with most studies focusing on IBD and HIV [49,50,52,57]. In one study, Crohn's disease (CD) patients were shown to have a higher number of VLPs, which provide a crude estimate of phage numbers, in colonic biopsies compared with healthy controls. Patients with ulcerated mucosa had significantly fewer VLPs than non-ulcerated mucosa. The authors hypothesized that viruses had higher survival rates or a higher frequency of viruses in non-ulcerated areas. Based upon these findings, it was proposed that phages play an indirect role in the immune dysregulation evident in CD patients through microbiome alterations [57]. A later study also described differences in the virome in CD patients according to their disease status (newly diagnosed, active onset, active presurgery) and therapy. Newly diagnosed patients had a higher viral diversity in faecal and biopsy samples than those with active disease. Additionally, there were significant

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

differences in virome diversity between patients on immunosuppressive therapy, steroids, combination therapy or no therapy. The clinical relevance of these results is unknown although more detailed studies of the alteration in virome composition in IBD patients is warranted [58]. Alterations in the enteric virome have also been reported in children susceptible to developing type 1 diabetes (T1D), prior to disease onset. Importantly, a disease-specific sequencing 'fingerprint' was identified in children susceptible to T1D that went on to develop the disease, compared with children that did not develop the disease. The present study reported that specific components of the virome were both directly and inversely associated with the development of T1D in these patients [59].

Additionally, a bidirectional communication network between the intestinal and central nervous system (the gut–brain–axis) is gaining research focus [60]. Its' role is to monitor and integrate intestine functions as well as linking emotional and cognitive centres of the brain with peripheral intestinal functions and mechanisms such as immune system activation, intestinal permeability, enteric reflex, pain perception and enteroendocrine signalling [61]. Both clinical and experimental evidence suggest that the enteric microbiota has an important impact on communication pathways between the intestine and brain, also known as the gut–microbiome–brain axis. The microbiome can interact locally with intestinal cells and enteric nervous system, but also have indirect interactions with the CNS through neuroendocrine and metabolic pathways. Therefore, significant alterations in the resident microbiota or their metabolites might have a direct effect on the host nervous system and lead to neurological pathologies [62]. For example, changes in the microbiome have been associated with autism, depression, schizophrenia, Alzheimer's disease and Parkinson's disease [63-67].

Our own research is focused on developing a mechanistic understanding of the intestine–microbiome–brain axis and the GI tract microbiome in the pathogenesis of the neurological disorder, myalgic encephalomyelitis (ME)/chronic fatigue syndrome (CFS) (ME/CFS).

ME/CFS

Historical perspective

The causative factor(s) of ME/CFS remain elusive resulting in a lack of robust diagnostics and effective treatments [68,69]. The disease onset and progression varies from patient to patient with the onset normally associated with an acute flu-like viral infection, which is either gradual or rapid [68]. Approximately 25% of patients become house- or bed-bound with less than 10% returning to predisease levels of function [70]. The socioeconomic burden of ME/CFS is significant and estimated to be between $17 and $24 billion per annum. This considerable cost is due to direct and indirect effects of the illness, such as healthcare and loss of work for patient and/or family carers [71].

The heterogenic nature of ME/CFS suggests a multifactorial and self-sustaining disorder [72]. Several theories have been proposed including mitochondrial dysfunction, viral infection and autoimmunity [68,73,74]. Important clues for the involvement of (viral) infections in the aetiology of ME/CFS can be obtained from historical reports of epidemic or sporadic outbreaks of cases; the first of which was reported in 1934 in a suspected epidemic of poliomyelitis in Los Angeles, California [75,76]. The inconsistent disease pattern observed in patients led doctors to classify this epidemic as atypical; differing from polio cases endemic at the time by the lack of flaccid paralysis, which normally defines poliomyelitis [77]. Additionally, the affected cases were mainly older children and young adults compared with polio which affected infants and children of less than 5 years of age [78]. The disease at the onset consisted of an acute upper respiratory tract infection accompanied by muscle weakness, fever, pain, malaise and photophobia. The patients reported recurrence of fever and other symptoms during recovery, which were at a greater incidence than those in typical epidemic poliomyelitis [76].

A similar apparent epidemic of poliomyelitis appeared in Akureyri, Iceland between 1948 and 1949. There were striking similarities between this outbreak of atypical poliomyelitis and the one recorded in Los Angeles in 1934, including both overlapping symptoms and occurrence of relapse. This disease was named Iceland (or Akureyri) disease [79]. Sixty-one other outbreaks of a similar disease were reported worldwide between 1934 and 1990 [75]. The most significant outbreak was in 1955 at the Royal Free Hospital in London, where 292 hospital staff were affected by the illness. The disease when fully developed showed features of a generalized infection with involvement of the lymphoreticular system, and widespread involvement of the central nervous system. The mysterious polio-like illness (including the disease at the Royal Free Hospital) was renamed ME and later extended to CFS (ME/CFS) to include a seemingly identical disease [80,81].

What is ME/CFS?

In both historical and current cases of ME/CFS persistent fatigue is the dominant and defining symptom, which is accompanied by a range of heterogeneous symptoms that are universally present in all the patients. It is classified by

PORTLAND PRESS

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

Table 2 ME/CFS microbiome articles selected following literature review

Number	Year	Author	Title	Area of study
1	2017	Armstrong [93]	The association of faecal microbiota and faecal, blood, serum and urine metabolites in ME/CFS	Microbiome and metabolites
2	2017	Nagy-Szakal [90]	Faecal metagenomic profiles in subgroups of patients with ME/CFS	Microbiome
3	2016	Giloteaux [91]	Reduced diversity and altered composition of the gut microbiota in individuals with ME/CFS	Microbiome
4	2016	Giloteaux [107]	A pair of identical twins discordant for ME/CFS differ in physiological parameters and gut microbiome composition	Microbiome and virome
5	2013	Fremont [92]	High-throughput 16S rRNA gene sequencing reveals alterations of intestinal microbiota in ME/CFS patients	Microbiome
6	2009	Sheedy [94]	Increased d-lactic acid intestinal bacteria in patients with CFS	Microbiome and metabolites
7	2009	Fremont [109]	Detection of herpes virus and parvovirus B19 in gastric and intestinal mucosa of CFS patients	Virome
8	2008	Chia [108]	CFS is associated with chronic enterovirus infection of the stomach	Virome
9	2007	Evengård [176]	Patients with CFS have higher numbers of anaerobic bacteria in the intestine compared with healthy subjects	Microbiome
10	2001	Butt [177]	Bacterial colonosis' in patients with persistent fatigue	Microbiome
11	1998	Butt [98]	Faecal microbial growth inhibition in chronic fatigue/pain patients	Microbiome and metabolites

Abbreviations: CFS, chronic fatigue syndrome; ME, myalgic encephalomyelitis

the World Healthy Organization International Classification of Diseases (ICD-10) as a neurological disorder (WHO Reference 93.3). Patients often report delayed exacerbation of symptoms following mental or physical exertion and daily or weekly variations in symptom severity that have a significant impact on day-to-day living [69,82]. A standardized criterion for ME/CFS is urgently needed, with diagnosis relying heavily upon clinical observations and by exclusion of other disorders. This situation is further complicated by the use of different diagnostic criteria within the same country and between different countries. As a result, it is can take several years for sufferers to receive a diagnosis [83-85]. To date, an effective treatment for ME/CFS does not exist, with current treatments aimed at alleviating symptoms [86].

An intestinal origin for ME/CFS

The co-morbidity of ME/CFS and GI symptoms is well documented, with one study reporting 92% of patients exhibiting IBS [87]. Additional studies have reported increased mucosal and systemic levels of pro-inflammatory cytokines such as IL-6, IL-8, IL-1β and TNFα in patients with coexistent IBS [88,89]. The significant GI symptoms often experienced by ME/CFS patients has led researchers, including ourselves, to investigate the microbiome in these patients. Several studies have reported significant changes in microbiota composition of ME/CFS patients compared with controls [90-92]. However, ME/CFS microbiome studies to date have largely focused on alterations in bacterial populations. The advancement in sequencing technology and emerging influence of the virome on human health has enabled studies of the virome of ME/CFS [18].

Intestinal microbiome and ME/CFS

We completed a literature search to determine the extent of microbiome research in ME/CFS using the following search terms: 'Myalgic encephalomyelitis', 'Chronic Fatigue Syndrome', 'CFS/ME', 'ME/CFS' in combination with 'virome', 'microbiome', 'metabolome', 'metagenomics', 'viromics' and 'metabolomics'. The resulting papers were screened according to abstract contents. Articles were excluded if an intervention was used (e.g. probiotics) and measurements not reported prior to the invention. This resulted in 11 papers that had examined the microbiome and/or intestinal metabolome of ME/CFS patients, dating from 1998 to 2017 (Table 2). Due to inconsistencies in study design including small sample sizes, different sequencing platforms and bioinformatics software analyses, microbial sequencing depth and a single time point'snapshot' of sampling and analysis; it was not possible to compare the studies statistically. However, from examining the articles individually there is sufficient evidence to support the claim of an altered intestinal microbiome in ME/CFS patients.

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

Table 3 Overview of articles selected studying the microbiome in ME/CFS

Author	Number of patients	Number of controls	Studying	Study design
Armstrong [93]	34	25	Microbiome and metabolites	Culture + MS
Nagy-Szakal [90]	50	50	Microbiome	Metagenomics
Giloteaux [91]	48	39	Microbiome	16s rRNA gene sequencing
Giloteaux [107]	1	1	Microbiome and virome	16s rRNA gene sequencing
Frémont [92]	43	36	Microbiome	16s rRNA gene sequencing
Sheedy [94]	108	177	Microbiome and metabolites	Culture
Evengard [176]	10	10	Microbiome	Culture
Butt [177]	1390	-	Microbiome	Culture
Butt [98]	27	4	Microbiome and metabolites	Culture

Abbreviation: MS, mass spectrometry

The literature search revealed nine articles that examined the microbiome in ME/CFS patients, with three articles also examining intestinal metabolites (Table 3); it is challenging to compare these studies because of different diagnostic criteria, patient selection, use or non-use of appropriately matched control subjects and microbial identification techniques. Of the nine articles, four used sequencing technologies, but different platforms, with five using culture-based techniques. One study (2017) performed metagenomic sequencing on 50 patient samples and was able to determine species-level differences compared with samples from control subjects [90]. A simple comparison between studies revealed eight similar results and seven conflicting results (Table 4). For example, while Giloteaux et al. [91] and Armstrong et al. [93] reported a general decrease in bacterial abundance, Sheedy et al. [94] reported an increase. These studies utilized different microbial identification techniques, which might account for the conflicting results. The lack of statistical analysis of the datasets constrains direct cross-study comparisons. From the limited cross-study analysis shown in Table 4, one finding of note was the decrease in *Faecalibacterium* seen in three studies. A reduction in butyrate-producing genus, which includes *Faecalibacteria* has been associated with dysbiosis in CD patients [95]. Butyrate has several protective properties, including improving the mucosal barrier, and immunomodulatory and anti-inflammatory effects by down-regulating pro-inflammatory cytokines [96]. However, decrease in the relative abundance of *Faecalibacterium* are associated with several other disorders and is not therefore specific for ME/CFS [97]. Interestingly, there was an increase in Enterobacteriaceae in two studies [91,98]. However, this may result from ME/CFS symptoms instead of a disease-specific microbial alteration. Enterobacteriaceae are dominant in the upper GI tract and are present in at low levels in the faeces of healthy individuals [99]. These taxa likely become enriched with faster stool transit time (i.e. signature of diarrhoea). The notable increase in this family would be consistent with increased transit time and reported in IBS-like symptoms in patients [100]. A depletion in the butyrate-producing family Ruminococcaceae was recorded in two studies and is also associated with diarrhoea [101]. Interestingly, *Bacteroides* spp. known for producing butyrate were reduced in several studies [97]. It would be beneficial to perform longitudinal studies of the microbiome in ME/CFS patients throughout the duration of the illness. This may provide more insightful clues as to the significance of any microbiome compositional changes with disease progression and severity within an individual patient. Several studies have been published on the longitudinal evaluation of ME/CFS patients; however, these focus on immune aspects, rehabilitative treatments and employment status [102-105].

As with any microbiome study, it is difficult to determine if the alterations observed cause, precede or correlate with disease. The microbiome of a patient would exhibit disease-specific microbial signatures and general microbial changes due to an unbalanced microbiome [106]. It is important, therefore, to separate microbial alterations associated with an unbalanced microbiome from those associated with a specific disease (microbiome disease biomarkers).

Virome and ME/CFS

Of the 11 articles selected in our literature search, only 3 examined the intestinal virome of CFS/ME patients (Table 5). Of these, two articles used direct virus detection (e.g. PCR or immunostaining) and one article used a high-throughput sequencing technique (Illumina MiSeq). An increase in bacteriophage richness, particularly Siphoviridae and Myoviridae, in patients was noted in the Giloteaux et al. study [107]. However, this study is statistically underpowered due to its small sample size. Chia and Chia [108] and Frémont et al. [109] used virus detection techniques to examine the presence of eukaryotic viruses within the gastric/intestinal mucosa. These studies reported an increase in parvovirus B19, enteroviral RNA and viral capsid protein 1 in patients. Also of note, Nagy-Szakal et al. [90] used

PORTLAND PRESS

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

Table 4 Basic comparison of microbiome composition alterations noted in articles selected for review

Micro-biome comparison	Article details								
	Armstrong (2017) [93]	Nagy-Szakal (2017) [90]	Giloteaux (2016) [91]	Giloteaux (2016) [107]	Fremont (2013) [92]	Sheedy (2009) [94]	Evangård (2007) [176]	Butt (2001) [177]	Butt (1998) [98]
Overall abundance	↓		↓			↑			
Phylum Firmicutes			↓	↑					
Phylum Proteobacteria			↑	↓					
Family Bacteroidaceae			↓	↑					
Family Enterobacteriaceae			↑						↑
Family Prevotellaceae			↑	↓					
Family Rickenellaceae			↓	↓					
Family Ruminococcaceae			↓	↓					
Genus *Bacteroides*	↓								↓
Genus *Bifidobacterium*			↓	↓			↑	↓	↓
Genus *Clostridium*		↑	↓						
Genus *Coprobacillus*		↑	↑						
Genus *Faecalibacterium*		↓	↓	↓					
Genus *Haemophilus*		↓	↓						
Genus *Ruminococcus*			↓		↓				
Species *Enterococcus faecalis*						↑			↑
Species *E. coli*								↓	↓

Seventeen criteria were either similar or conflicting between studies (microbiome composition). The down arrows represent a decrease in patients and up arrows represent an increase in patients.

Table 5 Overview of articles selected studying the virome in ME/CFS

Author	Number of patients	Number of controls	Study design
Giloteaux [107]	1	1	Viral metagenomics
Fremont [109]	48	35	PCR detection
Chia [108]	165	34	PCR detection and immunoperoxidase staning

metagenomic based approach on a large (n=50) cohort of patients although they did not perform virome analysis on the dataset. The authors reported significant changes in the bacterial components of the microbiome in ME/CFS patients compared with controls [90]. Virome analysis could be performed with the available date to determine if significant changes are observed in the viral components of ME/CFS patients.

Metabolomics studies

The identification of the bacterial and viral components of the microbiome is an important step forward, as is understanding how the use of nutrients by these microorganisms influences the overall metabolism within the gut. Metabolomics can be used to identify metabolites within the microbiome [110]. Only a handful of studies have

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

Table 6 Overview of articles selected for studying the metabolome in ME/CFS

Author	Number of patients	Number of controls	Study design
Armstrong [93]	34	25	NMR spectroscopy
Sheedy [94]	108	177	C^{13}-labelled bacteria/metabolites for HPLC and NMR
Butt [98]	27	4	Specific metabolites

Abbreviations: NMR, nuclear magnetic resonance; HPLC, high performance liquid chromatography

attempted to characterize faecal metabolites in ME/CFS patients despite its potential for deciphering microbiome function (Table 6) [93,94,98]. There are significant challenges associated with identifying faecal metabolites due to differing metabolite properties and range of metabolite concentrations in samples [162,163]. A major challenge is not only to identify all metabolites (insufficient reference libraries available) but also to produce metadata (i.e. sample origin, tissue, experimental conditions) in a format that is easily interpreted [166]. The biological interpretation of metabolites as potential disease-associated biomarkers is often challenging as it requires data analysis and integration [167] and targeted and non-targeted metabolomics to dissect the metabolic pathway(s) and origin of metabolite(s) of interest [164]. Currently, ^{1}H NMR is the most used analytical technique for metabolite profiling and is routinely used in clinical or pharmaceutical research and applications [165].

Armstrong et al. [93] quantified metabolites using high-throughput ^{1}H NMR spectroscopy from ME patient faecal filtrates. This technique provides a non-targeted metabolic profile that measures all high concentration metabolites with non-exchangeable protons [111]. In addition to faecal metabolomics, the authors performed urine and blood serum metabolite analysis. The present study presented a robust metabolome workflow and eluded to the relationship of faecal metabolites and microbes with host blood serum and urine metabolites [93]. Two older studies used selective culture based systems to examine the metabolic output of specific bacteria [94,98]. It is difficult to draw conclusions from these older studies because of the culture-based techniques used. It is possible that the isolation of bacteria from the complex intestinal environment alters the excreted/secreted metabolites, resulting in metabolites specific to the artificial *in vitro* culture environment. To obtain a true picture of the faecal metabolome, samples should be prepared directly from the faecal sample.

Unfortunately, it is not possible to compare these three metabolomics studies directly because different metabolites were studied. However, it is possible to make general comparisons between metabolites and microbes. For example, Sheedy et al. [94] reported an increase in lactic acid and an increase in *Enterococcus faecalis*, a lactic acid producing bacteria. Interestingly, Armstrong et al. [93] reported a general increase in the short chain fatty acids (SCFAs) butyrate, isovalerate and valerate. This contradicts the microbiome studies as known SCFA-producing bacteria (*Faecalibacterium*, *Eubacteria*, *Roseburia* and *Ruminococcus*) were consistently decreased across multiple studies [90-92,107]. A decrease in lactate was also reported in this study [93]. Several bacterial members of the microbiota produce lactate, which is the most common short chain hydroxyl-fatty acid in the intestinal lumen [97,112]. It can be converted into other SCFAs by a subgroup of lactate-fermenting bacterial species. Changes in these lactate-fermenting bacterial species were not noted in the current microbiome studies. Future studies will need to examine microbiome and metabolome alterations in tandem and then integrate the data to reveal a truer picture of microbiome metabolism.

Studying the microbiome: techniques and challenges

Within recent years, the increased interest in trying to understand the effect of the microbiome on health and disease has resulted in significant advancements in techniques to characterize it [9-11]. In particular, metagenomics is increasingly popular and favoured over sequencing bacterial 16s rRNA due to increased taxonomic sensitivity and potential for functional interpretation. Additionally, established techniques are being applied to microbiome research, such as metabolomics [11,113]. The research at the Quadram Institute Bioscience is focused on optimizing protocols, standardizing microbiome studies and applying this to ME/CFS. There are several pitfalls and challenges associated with microbiome studies, which need to be addressed prior to patient recruitment and sample collection [114]. The considerations that need to be made in designing microbiome studies in ME/CFS and some recommendations are outlined in Figure 3. Below we describe in some detail the particular constraints on microbiome and virome studies in ME/CFS and the approaches that can be taken to mitigate against or overcome them.

PORTLAND PRESS

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

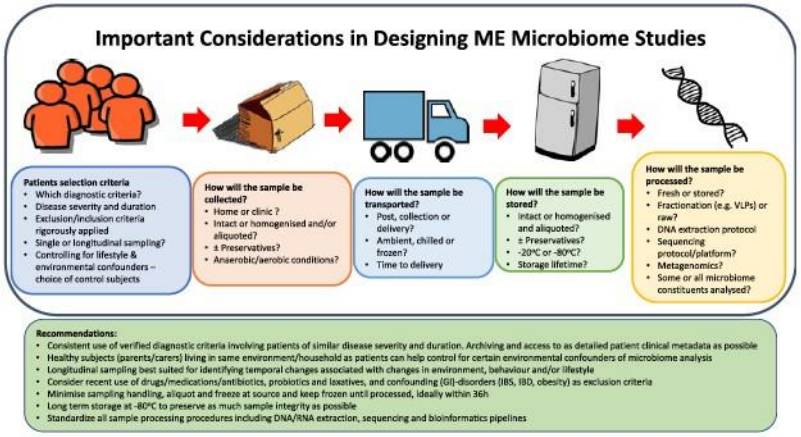

Figure 3. Important considerations in designing ME microbiome studies
Recommendations for designing a microbiome study and important questions to consider

Patient recruitment

A standardized criterion for ME/CFS diagnosis is lacking, with diagnosis relying heavily upon clinical observations and exclusion [83,115]. Multiple diagnosis checklists have been created, with each checklist differing slightly on symptom emphasis and severity [69,83,84]. The International ME criteria and Canadian criteria place greater emphasis on the delayed exacerbation of physical and mental symptoms following exertion. However, it does not exclude psychiatric illness such as depression or anxiety [116,117]. Therefore, it is difficult to determine how many patients recruited have an accurate diagnosis of ME/CFS and how many have been misdiagnosed using inadequate criteria. Several studies have attempted to address this by using multiple diagnostic checklists, with the majority using the 1994 Fukuda diagnostic scale. The severity of ME/CFS is ranked according to impact of illness upon daily life and ranges from mild to very severe. The severity grade given to a patient is subjective and generally given by the diagnosing clinician. Several studies have used patient questionnaires to assess the level of illness [90,91,94,107]. However, four different patient questionnaires were used by four different studies (Short Form 36 Healthy Survey, Multidimensional Fatigue Inventory, Bell's Disability Scale and McGregor 1995 questionnaire). Due to the multifactorial nature of the disease, standardization in diagnosis and disease severity are imperatives. These are a basic requirement to produce robust and reproducible microbiome studies. It is very difficult to determine if any microbiome differences are due to a true ME/CFS signature or complexities of patient recruitment. Future studies should aim to stratify patients according to disease duration and onset (sudden or gradual).

As microbiome research has increased, the need for properly matched controls has become apparent. The complexity of the GI microbiome and potential role within healthy/diseased states produce confounding factors [118]. As many of these confounding factors need to be taken into consideration in a microbiome study, and as many as possible should be accounted for or eliminated. Age, lifestyle, medications and drug use, geography and diet have all been reported to influence microbiome function and composition [5-7]. The effect of antibiotics on the microbiome is well documented [119]. However, other prescription and recreational drugs can affect microbiome analyses [120]. For example, decreasing stomach acidity with proton pump inhibitors allows upper GI microbes to move down into the intestine more readily, altering the composition of the lower GI microbiota and increasing risk of *C. difficile* infections [121].

Diet also influences the microbiome. Long-term dietary patterns have been linked to faecal microbiomes dominated by certain genera [118,122]. High protein/animal fat diets are associated with the prevalence of *Bacteriodes*, whereas diets high in carbohydrates are associated with high *Prevotella* [123]. To account for this, details of food consumption at least 48 h prior to sample collection should be obtained. Moreover, healthy household controls could be used to identify and exclude environmental confounders (e.g. diet, living environment); increasing the likelihood

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

PORTLAND PRESS

of identifying disease-specific microbiome signatures [118]. Additionally, the microbiome changes during aging and declines in diversity in the elderly [124]. Therefore, the use of age-matched controls would be beneficial to account for this important variable.

Recently, the influence of gender on the microbiome (termed as 'microgenderome') has become evident [125-128]. The intestine and its' microbiome serves as a virtual endocrine organ due to the metabolites and neurotransmitters and hormones it can produce [129]. For example, early microbial exposure increases testosterone levels in male mice, leading to a protective effect against T1D [128]. Additionally, microbiome alterations are observed in pre- and post-menopausal women; highlighting hormonal cross-talk within the microbiome [130]. Certain microbes have also been discovered to be a source of hormones and neurotransmitters. Experimental models have confirmed the bidirectional relationship between the intestinal microbiota, sex hormones and the immune system and provided an explanation for sexual dimorphism in T1D [128,131]. The results of these studies revealed evidence of sex-specific microbial communities, sex-specific responses to the same microbial communities, the role of sexual maturation impacting on changes on microbial communities and that microbial communities can play a protective and therapeutic role by influencing hormonal, metabolic and immune pathways [125-128]. A 2015 study compared the microbiome of male and female patients with ME/CFS revealing significant sex-specific interactions between Firmicutes (*Clostridium*, *Streptococcus*, *Lactobacillus and Enterococcus*) and symptoms, regardless of compositional similarity of microbial levels across the sexes [132]. This study highlights the need for gender-matched controls to account for any gender bias from future microbiome studies.

Although it is often impractical and perhaps impossible to control for all confounding factors within a microbiome study, efforts should be made to account for as many as possible.

Sample collection, storage and processing

As the number of microbiome studies has increased, the need for consistency in sampling techniques and standard operating procedures (SOPs) has also increased. An excellent review of the critical factors for sample collection, storage, transport and 'gold standard' techniques for longitudinal microbiome studies in human populations was recently published [114]. The most important considerations for storing microbiome samples are to reduce changes in the original microbiota from sample collection to processing and to keep storage conditions consistent for all samples in a study [133,134]. Sample storage conditions are not always consistent due to study or research group-specific downstream applications and resource limitations. Additionally, considerations are not always taken for preserving anaerobic bacteria within an anaerobic environment. Different studies often store samples at differing temperature (e.g. 4 to −80°C), affecting the long-term preservation of certain bacteria [114]. Additionally, the length of time for which the sample is stored and frequency of freeze/thaw cycles can significantly affect the microbiome composition. For example, *Bacteroides* is sensitive to freezing and should be processed within 6 weeks of storage (at −80°C) to avoid bacterial degradation [135,136]. The microbiome and ME/CFS studies reviewed here used different sample collection and storage techniques; including storage at <12°C, immediate processing, −20 and −80°C. For logistical reasons, it can be difficult to standardize this across all studies. However, it is important to be aware of these limitations.

The microbe composition changes laterally and longitudinally along the GI tract, therefore it has been suggested that there is significant variation within a single faecal sample. A 2015 study reported a reduction in intrasample variation following homogenization of the whole faecal sample [137]. However, several studies use a random section of the faeces without homogenization.

Additionally, different DNA extraction techniques have been used as a prelude to sequencing (MoBio PowerSoil DNA isolation kit, QIAmp DNA Stool Mini Kit and DNeasy Blood and Tissue kit). These kits differ in protocol, bead size, reagents used and are likely to introduce unnecessary bias [138].

Identification of prokaryotes

The development of sequencing, characterization of the bacterial component of the faecal microbiome relied on culture-based techniques that allow the identification of anaerobic and aerobic bacteria using selective or non-selective culture conditions and media; albeit taxonomic resolution and sensitivity is relatively low [12]. However, this approach does inform the cultured organism's growth requirements and substrate utilization and other physiological parameters, which cannot be obtained from sequence-based approaches [12]. Next-generation sequencing technology now makes it possible to characterize the bacterial microbiome using the 16S rRNA gene 'fingerprint' for identification and as an indicator of genetic diversity [4]. The 16S rRNA gene was chosen because of its relatively small

PORTLAND PRESS

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

size (~1.5 kb) and harbouring enough variation to distinguish between different species, yet enough similarity to assign members belonging to the same larger phylogenetic group (e.g. order, family or phylum) [5,139]. However, this approach has its limitations. It only detects and analyses a short, specific genomic region and taxonomic resolution or functional inference is therefore limited [11]. For example, this assay cannot recognize the different serovars within *Salmonella enterica* or detect toxin genes that could distinguish pathogenic *C. difficile* or distinguish pathogenic *Escherichia* strains from non-pathogenic strains [140]. This is particularly problematic in comparative studies of the microbiome in healthy and diseased states. It also provides no insight into functionality of the bacteriome [11].

Metagenomic sequencing is increasingly being chosen over 16S rRNA sequencing due to its higher taxonomic resolution and ability to infer functional potential [140]. It provides sequence information from the collective genomes of the microbiota, which in turn can be used to infer or predict functional contributions and biological roles of this complex community in human health and disease [11,139]. However, the absence of whole genome sequences in public databases limits the ability to identify gene function based on known sequence information. In comparison with 16S-based sequencing approaches, whole community metagenomics with an appropriate sequencing depth and coverage can be used to identify other microbes (i.e. archaea and viruses) within the microbiome [140,141]. Although it is possible to infer functional potential from metagenomic analysis through gene presence/abundance, the presence of a gene does not necessarily infer function; it is possible for the gene to be present but not transcribed [113]. Therefore, careful consideration needs to be taken when inferring functional potential from metagenomic sequences and, if possible, the predicted function should be examined using laboratory based techniques (e.g. antibiotic resistance), assuming the candidate microbe(s) can be cultured in isolation.

To date, only one study has utilized metagenomics in ME/CFS microbiome studies [90]. However, the analysis was incomplete and did not fully exploit the data produced. Whenever possible, metagenomics should be applied to microbiome studies in ME/CFS in order to achieve the required taxonomic resolution to fully examine the bacteriome and virome.

Identification of viruses

Virus genomes do not encode universally conserved genes such as the 16S or 18S genes of prokaryotes and eukaryotes respectively, and they are genetically highly diverse [142]. Consequently, it is not possible to use metataxonomic approaches such as 16S rRNA gene sequencing to characterize VLPs within an ecosystem [20]. Traditionally, classical approaches of microscopy and cultivation have been used to characterize VLPs isolated from faecal samples originating in the human intestine [143,144]. The only reliable molecular method currently available for surveying the human virome is metagenomics. However, to achieve an adequate sequencing depth, lytic VLPs need to be separated from the faecal material [145,146]. An excellent review describing the human virome and its characterization was recently published [147].

The efficient isolation of VLPs is an essential step in viral metagenomics in order to obtain an accurate picture and profile of the virome [148,149]. The workflow for sequencing the nucleic acid in VLPs (Figure 2) from faecal material begins with homogenization of faeces in buffer, centrifugation to remove cell debris followed by filtration to remove bacteria [150]. Ultracentrifugation can be used to separate the sample into differing densities and a specific density containing VLPs can be selected for downstream processing. Within the intestinal microbiome community, viral genomes represent a small proportion of the total DNA compared with bacterial genomes [149,151]. It is important therefore to use a reliable, robust and efficient VLP isolation protocol with as few manipulations of the sample as possible to minimize loss of VLPs. Various VLP protocols have been published that differ in details such as filter pore size, centrifugation speed and the inclusion/omission of ultracentrifugation [148-150,152,153]. Importantly, these protocols have yet to be directly compared. It is highly desirable therefore that standardized faecal VLP isolation and DNA extraction techniques are adopted to enable direct comparisons of datasets from different virome studies.

Viral metagenomic data are analysed in a manner similar to bacterial metagenomic data [154,155]. High-quality reads are aligned to reference databases, assembly is then attempted with non-aligned reads and functional characteristics inferred [147]. However, the lack of conserved genes, high genetic variation and under-representation of virus genomes within reference databases means a minority of the reads can be taxonomically assigned. It is predicted that less than 0.001% of the predicted phage diversity is represented in global sequence databanks [156]. One virome study has reported that 98% of the generated reads did not significantly match to an identified sequence within a database [157]. Therefore, a majority of sequencing reads are unassigned to any known genomes and are considered 'viral dark matter'. Additionally, assembly of sequencing reads is made difficult due to their short-read lengths [158,159]. Several research groups are now investigating the possibility of long-read sequencing to characterize the virome. The release of the Oxford Nanopore MinION has drastically reduced the cost of long-read sequencing (compared with PacBio

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

sequencing). These have the potential to provide complete or near complete phage genomes without the need for alignment or representation in databases [160].

Perhaps the biggest challenge in studying the intestinal virome is the lack of bioinformatics tools for the analysis of sequence data [146,147,161]. To date, there is no easy-to-use pipeline that uses raw reads, can remove host DNA, can search for bacterial contaminants and assign taxonomy and functionality to viruses within the sample. However, efforts are being made to generate such tools. In addition to isolation and sequencing of VLPs, it is possible to identify prophages and the bacterial host(s) from metagenomic sequencing. To accurately study the virome, both techniques should be utilized to study the lytic and lysogenic phages [154,155,157]. The Norwich U.K., ME/CFS research group is currently optimizing and standardizing VLP isolation and DNA/RNA sequencing protocols in addition to developing fit-for purpose viromics pipelines to comprehensively analyse the virome in ME/CFS patients.

Concluding remarks

Several microbiome studies have been performed on ME/CFS patients in the hope of identifying disease-specific signatures. This disease should be viewed as multifactorial and that the alteration of one body system (e.g. microbiome) may not be the exclusive cause. The dysregulation of the microbiome may be variously placed in a disease progression pathway interfacing with other systems (immune, neuroendocrine and mitochondrial), tipping the body into persistent imbalance. Although studies to date often report conflicting results, microbiome dysbiosis in ME/CFS patients is evident. However, in order to discover disease-specific microbe alterations, future studies need to adopt standardized techniques and analyses. The recent advancements in sequencing technology allows the characterization of the previously neglected virome. As virome research increases, it is becoming clear that the virome can directly and indirectly affect host health, and may play a role in the pathogenesis of ME/CFS. Confirmation of such a role will be largely dependent on the adoption of robust patient selection, reproducible study design and appropriate data analyses by different research groups investigating the microbiome/virome in complex diseases such as ME/CFS.

Clinical perspectives

- Several studies have reported alterations in the intestinal microbiome of ME/CFS patients, suggesting the involvement of microbial dysbiosis.
- Study design needs to be consistent to allow statistical comparison between different microbiome studies.
- Future microbiome studies should take account of the virome.

Acknowledgements

We thank Navena Navaneetharaja, Nadine Davies, Daniel Vipond and Lesley Hoyles for their contribution towards this research; additionally, Kathryn Cross for the TEM imaging of intestinal viruses. ME/CFS research was carried out in the laboratories of S.R.C. and T.W.

Competing interests

The authors declare that there are no competing interests associated with the manuscript.

Funding

This work was supported in part by Invest in ME to S.R.C. and T.W.; and the Ph.D. funding [grant number R205102 to F.N.)].

Author contribution

F.N. wrote the manuscript with support from S.R.C. and T.W. T.W. also designed the graphical illustrations. S.-Y.H. prepared samples for TEM imaging.

Abbreviations

CD, Crohn's disease; CFS, chronic fatigue syndrome; GI, gastrointestinal; IBD, inflammatory bowel disease; IBS, irritable bowel syndrome; IL, Interleukin; ME, myalgic encephalomyelitis; ME/CFS, myalgic encephalomyelitis/chronic fatigue syndrome; SCFA,

PORTLAND PRESS

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

short chain fatty acid; T1D, type 1 diabetes; TNFα, Tumor necrosis factor alpha; VLP, virus-like particle; WHO, World Health Organization.

References

1 Kunz, C., Kuntz, S. and Rudloff, S. (2009) Intestinal flora. *Adv. Exp. Med. Biol.* **639**, 67–79, https://doi.org/10.1007/978-1-4020-8749-3_6
2 Morelli, L. (2008) Postnatal development of intestinal microflora as influenced by infant nutrition. *J. Nutr.* **138**, 1791S–1795S, https://doi.org/10.1093/jn/138.9.1791S
3 Neish, A.S. (2009) Microbes in gastrointestinal health and disease. *Gastroenterology* **136**, 65–80, https://doi.org/10.1053/j.gastro.2008.10.080
4 Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M. et al. (2005) Diversity of the human intestinal microbial flora. *Science* **308**, 1635–1638, https://doi.org/10.1126/science.1110591
5 Palmer, C., Bik, E.M., DiGiulio, D.B., Relman, D.A. and Brown, P.O. (2007) Development of the human infant intestinal microbiota. *PLoS Biol.* **5**, 1556–1573, https://doi.org/10.1371/journal.pbio.0050177
6 Koenig, J.E., Spor, A., Scalfone, N., Fricker, A.D., Stombaugh, J., Knight, R. et al. (2011) Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4578–4585, https://doi.org/10.1073/pnas.1000081107
7 Tamburini, S., Shen, N., Wu, H.C. and Clemente, J.C. (2016) The microbiome in early life: implications for health outcomes. *Nat. Med.* **22**, 713–717, https://doi.org/10.1038/nm.4142
8 Mackie, R.I., Sghir, A. and Gaskins, H.R. (1999) Developmental microbial ecology of the neonatal gastrointestinal tract. *Am. J. Clin. Nutr.* **69**, 1035S–1045S
9 Robinson, C.J., Bohannan, B.J.M. and Young, V.B. (2010) From structure to function: the ecology of host-associated microbial communities. *Microbiol. Mol. Biol. Rev.* **74**, 453–476, https://doi.org/10.1128/MMBR.00014-10
10 Bassis, C., Young, V. and Schmidt, T. (2013) Methods for characterizing microbial communities associated with the human body. *The Human Microbiota*, pp. 51–74, John Wiley & Sons, Inc., Hoboken, NJ, U.S.A.
11 Di Bella, J.M., Bao, Y., Gloor, G.B., Burton, J.P. and Reid, G. (2013) High throughput sequencing methods and analysis for microbiome research. *J. Microbiol. Methods* **95**, 401–414, https://doi.org/10.1016/j.mimet.2013.08.011
12 Sekirov, I., Russell, S. and Antunes, L. (2010) Gut microbiota in health and disease. *Physiol. Rev.* **90**, 859–904, https://doi.org/10.1152/physrev.00045.2009
13 O'Hara, A.M. and Shanahan, F. (2006) The gut flora as a forgotten organ. *EMBO Rep.* **7**, 688–693, https://doi.org/10.1038/sj.embor.7400731
14 Shreiner, A.B., Kao, J.Y. and Young, V.B. (2016) The gut microbiome in health and in disease. *Curr. Opin. Gastroenterol.* **31**, 69–75, https://doi.org/10.1097/MOG.0000000000000139
15 Pfeiffer, J.K. and Virgin, H.W. (2016) Viral immunity. Transkingdom control of viral infection and immunity in the mammalian intestine. *Science* **351**, aad5872, https://doi.org/10.1126/science.aad5872
16 Lim, E.S., Wang, D. and Holtz, L.R. (2016) The bacterial microbiome and virome milestones of infant development. *Trends Microbiol.* **24**, 801–810, https://doi.org/10.1016/j.tim.2016.06.001
17 Columpsi, P., Sacchi, P., Zuccaro, V., Cima, S., Sarda, C., Mariani, M. et al. (2016) Beyond the gut bacterial microbiota: The gut virome. *J. Med. Virol.* **88**, 1467–1472, https://doi.org/10.1002/jmv.24508
18 Haynes, M. and Rohwer, F. (2011) The human virome. *Metagenomics of the Human Body*, pp. 63–77, Springer New York, New York
19 Breitbart, M. and Rohwer, F. (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13**, 278–284, https://doi.org/10.1016/j.tim.2005.04.003
20 Virgin, H.W. (2014) The virome in mammalian physiology and disease. *Cell* **157**, 142–150
21 Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R.A., Felts, B. et al. (2008) Viral diversity and dynamics in an infant gut. *Res. Microbiol.* **159**, 367–373, https://doi.org/10.1016/j.resmic.2008.04.006
22 Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F. et al. (2011) Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338, https://doi.org/10.1038/nature09199
23 Cario, E., Gerken, G. and Podolsky, D.K. (2007) Toll-like receptor 2 controls mucosal inflammation by regulating epithelial barrier function. *Gastroenterology* **132**, 1359–1374, https://doi.org/10.1053/j.gastro.2007.02.056
24 Hooper, L.V., Wong, M.H., Thelin, A., Hansson, L., Falk, P.G. and Gordon, J.I. (2001) Molecular analysis of commensal host-microbial relationships in the intestine. *Science* **291**, 881–884, https://doi.org/10.1126/science.291.5505.881
25 Lutgendorff, F., Akkermans, L.M.A. and Söderholm, J.D. (2008) The role of microbiota and probiotics in stress-induced gastro-intestinal damage. *Curr. Mol. Med.* **8**, 282–298, https://doi.org/10.2174/156652408784533779
26 Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S. and Medzhitov, R. (2004) Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell* **118**, 229–241, https://doi.org/10.1016/j.cell.2004.07.002
27 Hooper, L.V., Bry, L., Falk, P.G. and Gordon, J.I. (1998) Host-microbial symbiosis in the mammalian intestine: exploring an internal ecosystem. *Bioessays* **20**, 336–343, https://doi.org/10.1002/(SICI)1521-1878(199804)20:4%3c336::AID-BIES10%3e3.0.CO;2-3
28 Rooks, M.G. and Garrett, W.S. (2016) Gut microbiota, metabolites and host immunity. *Nat. Rev. Immunol.* **16**, 341–352, https://doi.org/10.1038/nri.2016.42
29 Hooper, L.V. and Macpherson, A.J. (2010) Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nat. Rev. Immunol.* **10**, 159–169, https://doi.org/10.1038/nri2710
30 Kabat, A.M., Srinivasan, N. and Maloy, K.J. (2014) Modulation of immune development and function by intestinal microbiota. *Trends Immunol.* **35**, 507–517, https://doi.org/10.1016/j.it.2014.07.010

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

31 Kostic, A.D., Chun, E., Robertson, L., Glickman, J.N., Gallini, C.A., Michaud, M. et al. (2013) *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* **14**, 207–215, https://doi.org/10.1016/j.chom.2013.07.007
32 Kostic, A.D., Xavier, R.J. and Gevers, D. (2014) The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology* **146**, 1489–1499, https://doi.org/10.1053/j.gastro.2014.02.009
33 Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G. et al. (2013) Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546, https://doi.org/10.1038/nature12506
34 Distrutti, E., Monaldi, L., Ricci, P. and Fiorucci, S. (2016) Gut microbiota role in irritable bowel syndrome: New therapeutic strategies. *World J. Gastroenterol.* **22**, 2219–2241, https://doi.org/10.3748/wjg.v22.i7.2219
35 Rajilić-Stojanović, M., Heilig, H.G.H.J., Tims, S., Zoetendal, E.G. and de Vos, W.M. (2013) Long-term monitoring of the human intestinal microbiota composition. *Environ. Microbiol.* **15**, 1146–1159, https://doi.org/10.1111/1462-2920.12023
36 David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E. et al. (2014) Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563, https://doi.org/10.1038/nature12820
37 Scher, J.U., Littman, D.R. and Abramson, S.B. (2016) Microbiome in inflammatory arthritis and human rheumatic diseases. *Arthritis Rheumatol.* **68**, 35–45, https://doi.org/10.1002/art.39259
38 Berbers, R.M., Nierkens, S., van Laar, J.M., Bogaert, D. and Leavis, H.L. (2017) Microbial dysbiosis in common variable immune deficiencies: evidence, causes, and consequences. *Trends Immunol.* **38**, 206–216, https://doi.org/10.1016/j.it.2016.11.008
39 Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M. et al. (2012) Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227
40 Foster, J.A., Rinaman, L. and Cryan, J.F. (2017) Stress & the gut-brain axis: regulation by the microbiome. *Neurobiol. Stress* **7**, 124–136, https://doi.org/10.1016/j.ynstr.2017.03.001
41 Dethlefsen, L., Huse, S., Sogin, M.L. and Relman, D.A. (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rRNA sequencing. *PLoS Biol.* **6**, e280
42 De Paepe, M., Leclerc, M., Tinsley, C.R. and Petit, M.-A. (2014) Bacteriophages: an underestimated role in human and animal health? *Front. Cell Infect. Microbiol.* **4**, 39, https://doi.org/10.3389/fcimb.2014.00039
43 Modi, S.R., Lee, H.H., Spina, C.S. and Collins, J.J. (2013) Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**, 219–222, https://doi.org/10.1038/nature12212
44 Meessen-Pinard, M., Sekulovic, O. and Fortier, L.C. (2012) Evidence of *in vivo* prophage induction during *Clostridium difficile* infection. *Appl. Environ. Microbiol.* **78**, 7662–7670, https://doi.org/10.1128/AEM.02275-12
45 Matos, R.C., Lapaque, N., Rigottier-Gois, L., Debarbieux, L., Meylheuc, T., Gonzalez-Zorn, B. et al. (2013) *Enterococcus faecalis* prophage dynamics and contributions to pathogenic traits. *PLoS Genet.* **9**, e1003539, https://doi.org/10.1371/journal.pgen.1003539
46 Webb, J.S., Lau, M. and Kjelleberg, S. (2004) Bacteriophage and phenotypic variation in *Pseudomonas aeruginosa* biofilm development. *J. Bacteriol.* **186**, 8066–8073, https://doi.org/10.1128/JB.186.23.8066-8073.2004
47 Carrolo, M., Frias, M.J., Pinto, F.R., Melo-Cristino, J. and Ramirez, M. (2010) Prophage spontaneous activation promotes DNA release enhancing biofilm formation in *Streptococcus pneumoniae*. *PLoS ONE* **5**, e15678, https://doi.org/10.1371/journal.pone.0015678
48 Schuch, R. and Fischetti, V.A. (2009) The secret life of the anthrax agent *Bacillus anthracis*: Bacteriophage-mediated ecological adaptations. *PLoS ONE* **4**, e6532, https://doi.org/10.1371/journal.pone.0006532
49 Wagner, J., Maksimovic, J., Farries, G., Sim, W.H., Bishop, R.F., Cameron, D.J. et al. (2013) Bacteriophages in gut samples from pediatric Crohn's disease patients: metagenomic analysis using 454 pyrosequencing. *Inflamm. Bowel Dis.* **19**, 1598–1608, https://doi.org/10.1097/MIB.0b013e318292477c
50 Pérez-Brocal, V., García-López, R., Vázquez-Castellanos, J.F., Nos, P., Beltrán, B., Latorre, A. et al. (2013) Study of the viral and microbial communities associated with Crohn's disease: a metagenomic approach. *Clin. Transl. Gastroenterol.* **4**, e36, https://doi.org/10.1038/ctg.2013.9
51 Wang, W., Jovel, J., Halloran, B., Wine, E., Patterson, J., Ford, G. et al. (2015) Metagenomic analysis of microbiome in colon tissue from subjects with inflammatory bowel diseases reveals interplay of viruses and bacteria. *Inflamm. Bowel Dis.* **21**, 1419–1427
52 Monaco, C.L., Gootenberg, D.B., Zhao, G., Handley, S.A., Ghebremichael, M.S., Lim, E.S. et al. (2016) Altered virome and bacterial microbiome in human immunodeficiency virus-associated acquired immunodeficiency syndrome. *Cell Host Microbe* **19**, 311–322, https://doi.org/10.1016/j.chom.2016.02.011
53 Diard, M., Bakkeren, E., Cornuault, J.K., Moor, K., Hausmann, A., Sellin, M.E. et al. (2017) Inflammation boosts bacteriophage transfer between *Salmonella* spp. *Science* **355**, 1211–1215, https://doi.org/10.1126/science.aaf8451
54 Boyd, E.F. (2012) Bacteriophage-encoded bacterial virulence factors and phage-pathogenicity island interactions. *Adv. Virus Res.* **82**, 91–118, https://doi.org/10.1016/B978-0-12-394621-8.00014-5
55 Gamage, S.D., Patton, A.K., Hanson, J.F. and Weiss, A.A. (2004) Diversity and host range of Shiga toxin-encoding phage. *Infect. Immun.* **72**, 7131–7139, https://doi.org/10.1128/IAI.72.12.7131-7139.2004
56 Nguyen, S., Baker, K., Padman, B.S., Patwa, R., Dunstan, R.A., Weston, T.A. et al. (2017) Bacteriophage transcytosis provides a mechanism to cross epithelial cell layers. *MBio* **8**, e01874–17, https://doi.org/10.1128/mBio.01874-17
57 Lepage, P., Colombet, J., Marteau, P., Sime-Ngando, T., Doré, J. and Leclerc, M. (2008) Dysbiosis in inflammatory bowel disease: a role for bacteriophages? *Gut.* **57**, 424–425, https://doi.org/10.1136/gut.2007.134668
58 Pérez-Brocal, V., García-López, R., Nos, P., Beltrán, B., Moret, I. and Moya, A. (2015) Metagenomic analysis of Crohn's disease patients identifies changes in the virome and microbiome related to disease status and therapy, and detects potential interactions and biomarkers. *Inflamm. Bowel Dis.* **21**, 2515–2532, https://doi.org/10.1097/MIB.0000000000000549

PORTLAND PRESS

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

59 Kramná, L., Kolářová, K., Oikarinen, S., Pursiheimo, J.P., Ilonen, J., Simell, O. et al. (2015) Gut virome sequencing in children with early islet autoimmunity. *Diabetes Care* **38**, 930–933, https://doi.org/10.2337/dc14-2490
60 Rhee, S.H., Pothoulakis, C. and Mayer, E.A. (2009) Principles and clinical implications of the brain-gut-enteric microbiota axis. *Nat. Rev. Gastroenterol. Hepatology* **6**, 306–314, https://doi.org/10.1038/nrgastro.2009.35
61 Carabotti, M., Scirocco, A., Maselli, M.A. and Severi, C. (2015) The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. *Ann. Gastroenterol.* **28**, 203–209
62 Grenham, S., Clarke, G., Cryan, J.F. and Dinan, T.G. (2011) Brain-gut-microbe communication in health and disease. *Front. Physiol.* **2**, 94, https://doi.org/10.3389/fphys.2011.00094
63 Mulle, J.G., Sharp, W.G. and Cubells, J.F. (2013) The gut microbiome: a new frontier in autism research. *Curr. Psychiatry Rep.* **15**, 337, https://doi.org/10.1007/s11920-012-0337-0
64 Evrensel, A. and Ceylan, M.E. (2015) The gut-brain axis: the missing link in depression. *Clin. Psychopharmacol. Neurosci.* **13**, 239–244, https://doi.org/10.9758/cpn.2015.13.3.239
65 Dickerson, F., Severance, E. and Yolken, R. (2017) The microbiome, immunity, and schizophrenia and bipolar disorder. *Brain Behav. Immun.* **62**, 46–52, https://doi.org/10.1016/j.bbi.2016.12.010
66 Vogt, N.M., Kerby, R.L., Dill-McFarland, K.A., Harding, S.J., Merluzzi, A.P., Johnson, S.C. et al. (2017) Gut microbiome alterations in Alzheimer's disease. *Sci. Rep.* **7**, 13537, https://doi.org/10.1038/s41598-017-13601-y
67 Sampson, T.R., Debelius, J.W., Thron, T., Janssen, S., Shastri, G.G., Ilhan, Z.E. et al. (2016) Gut microbiota regulate motor deficits and neuroinflammation in a model of Parkinson's disease. *Cell* **167**, 1469.e12–1480.e12, https://doi.org/10.1016/j.cell.2016.11.018
68 Bansal, A.S., Bradley, A.S., Bishop, K.N., Kiani-Alikhan, S. and Ford, B. (2012) Chronic fatigue syndrome, the immune system and viral infection. *Brain Behav. Immun.* **26**, 24–31, https://doi.org/10.1016/j.bbi.2011.06.016
69 Bansal, A.S. (2016) Investigating unexplained fatigue in general practice with a particular focus on CFS/ME. *BMC Fam. Pract.* **17**, 81, https://doi.org/10.1186/s12875-016-0493-0
70 Kerr, J.R., Cunniffe, V.S., Kelleher, P., Bernstein, R.M. and Bruce, I.N. (2003) Successful intravenous immunoglobulin therapy in 3 cases of parvovirus B19-associated chronic fatigue syndrome. *Clin. Infect. Dis.* **36**, e100–6, https://doi.org/10.1086/374666
71 Jason, L.A., Benton, M.C., Valentine, L., Johnson, A. and Torres-Harding, S. (2008) The economic impact of ME/CFS: individual and societal costs. *Dyn. Med.* **7**, 6, https://doi.org/10.1186/1476-5918-7-6
72 Hickie, I., Lloyd, A., Hadzi-Pavlovic, D., Parker, G., Bird, K. and Wakefield, D. (1995) Can the chronic fatigue syndrome be defined by distinct clinical features? *Psychol. Med.* **25**, 925–935, https://doi.org/10.1017/S0033291700037417
73 Sanders, P. and Korf, J. (2008) Neuroaetiology of chronic fatigue syndrome: an overview. *World J. Biol. Psychiatry* **9**, 165–171, https://doi.org/10.1080/15622970701310971
74 Morris, G., Berk, M., Galecki, P. and Maes, M. (2014) The emerging role of autoimmunity in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). *Mol. Neurobiol.* **49**, 741–756, https://doi.org/10.1007/s12035-013-8553-0
75 Underhill, R.A. (2015) Myalgic encephalomyelitis, chronic fatigue syndrome: an infectious disease. *Med. Hypotheses.* **85**, 765–773, https://doi.org/10.1016/j.mehy.2015.10.011
76 Meals, R.W., Hauser, V.F. and Bower, A.G. (1935) Poliomyelitis-The Los Angeles Epidemic of 1934: part I. *Cal. West. Med.* **43**, 123–125
77 Mateen, F.J. and Black, R.E. (2013) Expansion of acute flaccid paralysis surveillance: beyond poliomyelitis. *Trop. Med. Int. Health* **18**, 1421–1422, https://doi.org/10.1111/tmi.12181
78 Hall, W.J., Nathanson, N. and Langmuir, A.D. (1957) The age distribution of poliomyelitis in the United States in 1955. *Am. J. Hyg.* **66**, 214–234
79 Sigurdsson, B., Sigurjónsson, J., Sigurdsson, J.H.J., Thorkelsson, J. and Gudmundsson, K.R. (1950) A disease epidemic in Iceland simulating poliomyelitis. *Am. J. Epidemiol.* **52**, 222–238, https://doi.org/10.1093/oxfordjournals.aje.a119421
80 Staff, M. (1957) An outbreak of encephalomyelitis in the royal free hospital group, London, in 1955. *Br. Med. J.* **2**, 895–904, https://doi.org/10.1136/bmj.2.5050.895
81 Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome; Board on the Health of Select Populations and Institute of Medicine (2015) *Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Redefining an Illness*, National Academies Press, Washington (DC)
82 Meeus, M., Van Eupen, I., Van Baarle, E., De Boeck, V., Luyckx, A., Kos, D. et al. (2011) Symptom fluctuations and daily physical activity in patients with chronic fatigue syndrome: a case-control study. *Arch. Phys. Med. Rehabil.* **92**, 1820–1826, https://doi.org/10.1016/j.apmr.2011.06.023
83 Twisk, F.N. (2015) Accurate diagnosis of myalgic encephalomyelitis and chronic fatigue syndrome based upon objective test methods for characteristic symptoms. *World J. Methodol.* **5**, 68–87, https://doi.org/10.5662/wjm.v5.i2.68
84 Fukuda, K., Straus, S.E., Hickie, I., Sharpe, M.C., Dobbins, J.G. and Komaroff, A. (1994) The chronic fatigue syndrome: a comprehensive approach to its definition and study. International Chronic Fatigue Syndrome Study Group. *Ann. Intern. Med.* **121**, 953–959
85 Capelli, E., Zola, R., Lorusso, L., Venturini, L., Sardi, F. and Ricevuti, G. (2010) Chronic fatigue syndrome/myalgic encephalomyelitis: an update. *Int. J. Immunopathol. Pharmacol.* **23**, 981–989, https://doi.org/10.1177/039463201002300402
86 Whiting, P., Bagnall, A.-M., Sowden, A.J., Cornell, J.E., Mulrow, C.D. and Ramírez, G. (2001) Interventions for the treatment and management of chronic fatigue syndrome. *JAMA* **286**, 1360, https://doi.org/10.1001/jama.286.11.1360
87 Aaron, L.A., Burke, M.M. and Buchwald, D. (2000) Overlapping conditions among patients with chronic fatigue syndrome, fibromyalgia, and temporomandibular disorder. *Arch. Intern. Med.* **160**, 221, https://doi.org/10.1001/archinte.160.2.221
88 Quigley, E.M.M. (2011) The enteric microbiota in the pathogenesis and management of constipation. *Best Pract. Res. Clin. Gastroenterol.* **25**, 119–126, https://doi.org/10.1016/j.bpg.2011.01.003
89 Quigley, E.M. (2011) Gut microbiota and the role of probiotics in therapy. *Curr. Opin. Pharmacol.* **11**, 593–603, https://doi.org/10.1016/j.coph.2011.09.010

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

90 Nagy-Szakal, D., Williams, B.L., Mishra, N., Che, X., Lee, B., Bateman, L. et al. (2017) Fecal metagenomic profiles in subgroups of patients with myalgic encephalomyelitis/ chronic fatigue syndrome. *Microbiome* **5**, 44
91 Giloteaux, L., Goodrich, J.K., Walters, W.A., Levine, S.M., Ley, R.E., Hanson, M.R. et al. (2016) Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome* **4**, 953–959, https://doi.org/10.1186/s40168-016-0171-4
92 Fremont, M., Coomans, D., Massart, S. and De Meirleir, K. (2013) High-throughput 16S rRNA gene sequencing reveals alterations of intestinal microbiota in myalgic encephalomyelitis/chronic fatigue syndrome patients. *Anaerobe* **22**, 50–56, https://doi.org/10.1016/j.anaerobe.2013.06.002
93 Armstrong, C.W., McGregor, N.R., Lewis, D.P., Butt, H.L. and Gooley, P.R. (2017) The association of fecal microbiota and fecal, blood serum and urine metabolites in myalgic encephalomyelitis/chronic fatigue syndrome. *Metabolomics* **13**, 8, https://doi.org/10.1007/s11306-016-1145-z
94 Sheedy, J.R., Wettenhall, R.E.H., Scanlon, D., Gooley, P.R., Lewis, D.P., Mcgregor, N. et al. (2009) Increased d-lactic acid intestinal bacteria in patients with chronic fatigue syndrome. *In Vivo* **23**, 621–628
95 Wright, E.K., Kamm, M.A., Teo, S.M., Inouye, M., Wagner, J. and Kirkwood, C.D. (2015) Recent advances in characterizing the gastrointestinal microbiome in Crohn's disease: a systematic review. *Inflamm. Bowel Dis.* **21**, 1219–1228
96 Machiels, K., Joossens, M., Sabino, J., De Preter, V., Arijs, I., Eeckhaut, V. et al. (2014) A decrease of the butyrate-producing species *Roseburia hominis* and *Faecalibacterium prausnitzii* defines dysbiosis in patients with ulcerative colitis. *Gut* **63**, 1275–1283, https://doi.org/10.1136/gutjnl-2013-304833
97 Ríos-Covián, D., Ruas-Madiedo, P., Margolles, A., Gueimonde, M., De los Reyes-Gavilán, C.G. and Salazar, N. (2016) Intestinal short chain fatty acids and their link with diet and human health. *Front. Microbiol.* **7**, 185
98 Butt, H., Dunstan, R., McGregor, N. and Roberts, T. (1998) Faecal microbial growth inhibition in chronic fatigue/pain patients. *Proceedings of the AHMF International Clinical and Scientific Conference*, Newcastle, Australia
99 Gall, A., Fero, J., McCoy, C., Claywell, B.C., Sanchez, C.A., Blount, P.L. et al. (2015) Bacterial composition of the human upper gastrointestinal tract microbiome is dynamic and associated with genomic instability in a Barrett's esophagus cohort. *PLoS ONE* **10**, e0129055, https://doi.org/10.1371/journal.pone.0129055
100 Carroll, I.M., Ringel-Kulka, T., Siddle, J.P. and Ringel, Y. (2012) Alterations in composition and diversity of the intestinal microbiota in patients with diarrhea-predominant irritable bowel syndrome. *Neurogastroenterol. Motil.* **24**, 521–530, e248, https://doi.org/10.1111/j.1365-2982.2012.01891.x
101 Vandeputte, D., Falony, G., Vieira-Silva, S., Tito, R.Y., Joossens, M. and Raes, J. (2016) Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65**, 57–62, https://doi.org/10.1136/gutjnl-2015-309618
102 Goldsmith, K., Chalder, T., White, P., Sharpe, M. and Pickles, A. (2015) Longitudinal mediation in the PACE randomised clinical trial of rehabilitative treatments for chronic fatigue syndrome: modelling and design considerations. *Trials* **16** (Suppl. 2), 043, https://doi.org/10.1186/1745-6215-16-S2-043
103 Hardcastle, S.L., Brenu, E.W., Johnston, S., Nguyen, T., Huth, T., Ramos, S. et al. (2015) Longitudinal analysis of immune abnormalities in varying severities of Chronic Fatigue Syndrome/Myalgic Encephalomyelitis patients. *J. Transl. Med.* **13**, 299, https://doi.org/10.1186/s12967-015-0653-3
104 Nyland, M., Naess, H., Birkeland, J.S. and Nyland, H. (2014) Longitudinal follow-up of employment status in patients with chronic fatigue syndrome after mononucleosis. *BMJ Open* **4**, e005798, https://doi.org/10.1136/bmjopen-2014-005798
105 Brenu, E.W., van Driel, M.L., Staines, D.R., Ashton, K.J., Hardcastle, S.L., Keane, J. et al. (2012) Longitudinal investigation of natural killer cells and cytokines in chronic fatigue syndrome/myalgic encephalomyelitis. *J. Transl. Med.* **10**, 88, https://doi.org/10.1186/1479-5876-10-88
106 Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A. and Alm, E.J. (2017) Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**, 1784, https://doi.org/10.1038/s41467-017-01973-8
107 Giloteaux, L., Hanson, M.R. and Keller, B.A. (2016) A pair of identical twins discordant for myalgic encephalomyelitis/chronic fatigue syndrome differ in physiological parameters and gut microbiome composition. *Am. J. Case Rep.* **17**, 720–729, https://doi.org/10.12659/AJCR.900314
108 Chia, J.K.S. and Chia, A.Y. (2007) Chronic fatigue syndrome is associated with chronic enterovirus infection of the stomach. *J. Clin. Pathol.* **61**, 43–48, https://doi.org/10.1136/jcp.2007.050054
109 Frémont, M., Metzger, K., Rady, H., Hulstaert, J. and De Meirleir, K. (2009) Detection of herpesviruses and parvovirus B19 in gastric and intestinal mucosa of chronic fatigue syndrome patients. *In Vivo* **23**, 209–213
110 Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K. and Narasimhan, G. (2016) Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis. *Evol. Bioinform. Online* **12** (Suppl. 1), 5–16
111 Matysik, S., Le Roy, C.I., Liebisch, G. and Claus, S.P. (2016) Metabolomics of fecal samples: a practical consideration. *Trends Food Sci. Technol.* **57**, 244–255, https://doi.org/10.1016/j.tifs.2016.05.011
112 Duncan, S.H., Louis, P. and Flint, H.J. (2004) Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. *Appl. Environ. Microbiol.* **70**, 5810–5817, https://doi.org/10.1128/AEM.70.10.5810-5817.2004
113 Morgan, X.C. and Huttenhower, C. (2014) Meta'omic analytic techniques for studying the intestinal microbiome. *Gastroenterology* **146**, 1437.e1–1448.e1, https://doi.org/10.1053/j.gastro.2014.01.049
114 Vandeputte, D., Tito, R.Y., Vanleeuwen, R., Falony, G. and Raes, J. (2017) Practical considerations for large-scale gut microbiome studies. *FEMS Microbiol. Rev.* **41** (Supp_1), S154–S167
115 Attree, E.A., Arroll, M.A., Dancey, C.P., Griffith, C. and Bansal, A.S. (2014) Psychosocial factors involved in memory and cognitive failures in people with myalgic encephalomyelitis/chronic fatigue syndrome. *Psychol. Res. Behav. Manag.* **7**, 67–76
116 Carruthers, B.M. (2007) Definitions and aetiology of myalgic encephalomyelitis: how the Canadian consensus clinical definition of myalgic encephalomyelitis works. *J. Clin. Pathol.* **60**, 117–119
117 Carruthers, B.M., Van de Sande, M.I., De Meirleir, K.L., Klimas, N.G., Broderick, G., Mitchell, T. et al. (2011) Myalgic encephalomyelitis: International Consensus Criteria. *J. Intern. Med.* **270**, 327–338, https://doi.org/10.1111/j.1365-2796.2011.02428.x

PORTLAND PRESS

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

118 Kim, D., Hofstaedter, C.E., Zhao, C., Mattei, L., Tanes, C., Clarke, E. et al. (2017) Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* **5**, 52
119 Langdon, A., Crook, N. and Dantas, G. (2016) The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Med.* **8**, 39, https://doi.org/10.1186/s13073-016-0294-z
120 Wilson, I.D. and Nicholson, J.K. (2017) Gut microbiome interactions with drug metabolism, efficacy, and toxicity. *Transl. Res.* **179**, 204–222, https://doi.org/10.1016/j.trsl.2016.08.002
121 Minalyan, A., Gabrielyan, L., Scott, D., Jacobs, J. and Pisegna, J.R. (2017) The gastric and intestinal microbiome: role of proton pump inhibitors. *Curr. Gastroenterol. Rep.* **19**, 42, https://doi.org/10.1007/s11894-017-0577-6
122 Singh, R.K., Chang, H.-W., Yan, D., Lee, K.M., Ucmak, D., Wong, K. et al. (2017) Influence of diet on the gut microbiome and implications for human health. *J. Transl. Med.* **15**, 73, https://doi.org/10.1186/s12967-017-1175-y
123 Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.Y., Keilbaugh, S.A. et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–108, https://doi.org/10.1126/science.1208344
124 Claesson, M.J., Cusack, S., O'Sullivan, O., Greene-Diniz, R., de Weerd, H., Flannery, E. et al. (2011) Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4586–4591, https://doi.org/10.1073/pnas.1000097107
125 Mueller, S., Saunier, K., Hanisch, C., Norin, E., Alm, L., Midtvedt, T. et al. (2006) Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study. *Appl. Environ. Microbiol.* **72**, 1027–1033, https://doi.org/10.1128/AEM.72.2.1027-1033.2006
126 Haro, C., Rangel-Zúñiga, O.A., Alcalá-Díaz, J.F., Gómez-Delgado, F., Pérez-Martínez, P., Delgado-Lista, J. et al. (2016) Intestinal microbiota is influenced by gender and body mass index. *PLoS ONE* **11**, e0154090, https://doi.org/10.1371/journal.pone.0154090
127 Yurkovetskiy, L., Burrows, M., Khan, A.A., Graham, L., Volchkov, P., Becker, L. et al. (2013) Gender bias in autoimmunity is influenced by microbiota. *Immunity* **39**, 400–412, https://doi.org/10.1016/j.immuni.2013.08.013
128 Markle, J.G.M., Frank, D.N., Mortin-Toth, S., Robertson, C.E., Feazel, L.M., Rolle-Kampczyk, U. et al. (2013) Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science* **339**, 1084–1088, https://doi.org/10.1126/science.1233521
129 Sudo, N. (2014) Microbiome, HPA axis and production of endocrine hormones in the gut. *Adv. Exp. Med. Biol.* **817**, 177–194, https://doi.org/10.1007/978-1-4939-0897-4˙8
130 Baker, J.M., Al-Nakkash, L. and Herbst-Kralovetz, M.M. (2017) Estrogen-gut microbiome axis: Physiological and clinical implications. *Maturitas* **103**, 45–53, https://doi.org/10.1016/j.maturitas.2017.06.025
131 Fransen, F., van Beek, A.A., Borghuis, T., Meijer, B., Hugenholtz, F., van der Gaast-de Jongh, C. et al. (2017) The impact of gut microbiota on gender-specific differences in immunity. *Front. Immunol.* **8**, 754, https://doi.org/10.3389/fimmu.2017.00754
132 Wallis, A., Butt, H., Ball, M., Lewis, D.P. and Bruck, D. (2016) Support for the microgenderome: associations in a human clinical population. *Sci. Rep.* **6**, 19171, https://doi.org/10.1038/srep19171
133 Choo, J.M., Leong, L.E. and Rogers, G.B. (2015) Sample storage conditions significantly influence faecal microbiome profiles. *Sci. Rep.* **5**, 16350, https://doi.org/10.1038/srep16350
134 Tedjo, D.I., Jonkers, D.M.A.E., Savelkoul, P.H., Masclee, A.A., Van Best, N., Pierik, M.J. et al. (2015) The effect of sampling and storage on the fecal microbiota composition in healthy and diseased subjects. *PLoS ONE* **10**, e0126685, https://doi.org/10.1371/journal.pone.0126685
135 Rochet, V., Rigottier-Gois, L., Rabot, S. and Doré, J. (2004) Validation of fluorescent in situ hybridization combined with flow cytometry for assessing interindividual variation in the composition of human fecal microflora during long-term storage of samples. *J. Microbiol. Methods* **59**, 263–270, https://doi.org/10.1016/j.mimet.2004.07.012
136 Jia, J., Frantz, N., Khoo, C., Gibson, G.R., Rastall, R.A. and McCartney, A.L. (2010) Investigation of the faecal microbiota associated with canine chronic diarrhea. *FEMS Microbiol. Ecol.* **71**, 304–312, https://doi.org/10.1111/j.1574-6941.2009.00812.x
137 Gorzelak, M.A., Gill, S.K., Tasnim, N., Ahmadi-Vand, Z., Jay, M. and Gibson, D.L. (2015) Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. *PLoS ONE* **10**, e0134802, https://doi.org/10.1371/journal.pone.0134802
138 Bag, S., Saha, B., Mehta, O., Anbumani, D., Kumar, N., Dayal, M. et al. (2016) An improved method for high quality metagenomics DNA extraction from human and environmental samples. *Sci. Rep.* **6**, 26775, https://doi.org/10.1038/srep26775
139 Peterson, D.A., Frank, D.N., Pace, N.R. and Gordon, J.I. (2008) Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases. *Cell Host Microbe* **3**, 417–427, https://doi.org/10.1016/j.chom.2008.05.001
140 Ranjan, R., Rani, A., Metwally, A., McGee, H.S. and Perkins, D.L. (2016) Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **469**, 967–977, https://doi.org/10.1016/j.bbrc.2015.12.083
141 Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T. et al. (2016) Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front. Microbiol.* **7**, 459, https://doi.org/10.3389/fmicb.2016.00459
142 Woolhouse, ME., Howey, R., Gaunt, E., Reilly, L., Chase-Topping, M. and Savill, N. (2008) Temporal trends in the discovery of human viruses. *Proc. Biol. Sci.* **275**, 2111–2115, https://doi.org/10.1098/rspb.2008.0294
143 Flewett, T.H., Bryden, A.S. and Davies, H. (1974) Diagnostic electron microscopy of faeces: I The viral flora of the faeces as seen by electron microscopy. *J. Clin. Pathol.* **27**, 603–608, https://doi.org/10.1136/jcp.27.8.603
144 Mitgutsch, K., Schirra, S. and Verrilli, S. (2013) Movers and shakers. *CHI'13 Ext. Abstr. Hum. Factors Comput. Syst. CHI EA'13* **976**, 715
145 Lecuit, M. and Eloit, M. (2013) The human virome: new tools and concepts. *Trends Microbiol.* **21**, 510–515, https://doi.org/10.1016/j.tim.2013.07.001
146 Wylie, K.M., Weinstock, G.M. and Storch, G.A. (2013) Virome genomics: a tool for defining the human virome. *Curr. Opin. Microbiol.* **16**, 479–484, https://doi.org/10.1016/j.mib.2013.04.006

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

147 Carding, S.R., Davis, N. and Hoyles, L. (2017) Review article: the human intestinal virome in health and disease. *Aliment. Pharmacol. Ther.* **46**, 800–815, https://doi.org/10.1111/apt.14280
148 Castro-Mejía, J.L., Muhammed, M.K., Kot, W., Neve, H., Franz, C.M.A.P., Hansen, L.H. et al. (2015) Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome* **3**, 64, https://doi.org/10.1186/s40168-015-0131-4
149 Kleiner, M., Hooper, L.V. and Duerkop, B.A. (2015) Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* **16**, 7, https://doi.org/10.1186/s12864-014-1207-4
150 Hoyles, L., McCartney, A.L., Neve, H., Gibson, G.R., Sanderson, J.D., Heller, K.J. et al. (2014) Characterization of virus-like particles associated with the human faecal and caecal microbiota. *Res. Microbiol.* **165**, 803–812, https://doi.org/10.1016/j.resmic.2014.10.006
151 Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F. and Gordon, J.I. (2012) Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* **10**, 607–617, https://doi.org/10.1038/nrmicro2853
152 Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W. et al. (2015) Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci. Rep.* **5**, 16532, https://doi.org/10.1038/srep16532
153 Hall, R.J., Wang, J., Todd, A.K., Bissielo, A.B., Yen, S., Strydom, H. et al. (2014) Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J. Virol. Methods* **195**, 194–204, https://doi.org/10.1016/j.jviromet.2013.08.035
154 Bikel, S., Valdez-Lara, A., Cornejo-Granados, F., Rico, K., Canizales-Quinteros, S., Soberón, X. et al. (2015) Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: Towards a systems-level understanding of human microbiome. *Comput. Struct. Biotechnol. J.* **13**, 390–401, https://doi.org/10.1016/j.csbj.2015.06.001
155 Hurwitz, B.L., U'Ren, J.M. and Youens-Clark, K. (2016) Computational prospecting the great viral unknown. *FEMS Microbiol. Lett.* **363**, https://doi.org/10.1093/femsle/fnw077
156 Ogilvie, L.A. and Jones, B.V. (2015) The human gut virome: a multifaceted majority. *Front. Microbiol.* **6**, https://doi.org/10.3389/fmicb.2015.00918
157 Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D. et al. (2011) The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625, https://doi.org/10.1101/gr.122705.111
158 Brum, J.R. and Sullivan, M.B. (2015) Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13**, 147–159, https://doi.org/10.1038/nrmicro3404
159 Mizuno, C.M., Rodriguez-Valera, F., Kimes, N.E. and Ghai, R. (2013) Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9**, e1003987, https://doi.org/10.1371/journal.pgen.1003987
160 Lu, H., Giordano, F. and Ning, Z. (2016) Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* **14**, 265–279, https://doi.org/10.1016/j.gpb.2016.05.004
161 Wylie, K.M., Weinstock, G.M. and Storch, G.A. (2012) Emerging view of the human virome. *Transl. Res.* **160**, 283–290, https://doi.org/10.1016/j.trsl.2012.03.006
162 De Preter, V. and Verbeke, K. (2013) Metabolomics as a diagnostic tool in gastroenterology. *World J. Gastrointest. Pharmacol. Ther.* **4**, 97, https://doi.org/10.4292/wjgpt.v4.i4.97
163 Smirnov, K.S., Maier, T.V., Walker, A., Heinzmann, S.S., Forcisi, S., Martinez, I. et al. (2016) Challenges of metabolomics in human gut microbiota research. *Int. J. Med. Microbiol.* **306**, 266–279, https://doi.org/10.1016/j.ijmm.2016.03.006
164 Vernocchi, P., Del Chierico, F. and Putignani, L. (2016) Gut microbiota profiling: metabolomics based approach to unravel compounds affecting human health. *Front. Microbiol.* **7**, https://doi.org/10.3389/fmicb.2016.01144
165 Zhang, C., Yin, A., Li, H., Wang, R., Wu, G., Shen, J. et al. (2015) Dietary modulation of gut microbiota contributes to alleviation of both genetic and simple obesity in children. *EBio Med.* **2**, 968–984
166 Weckwerth, W. and Morgenthal, K. (2005) Metabolomics: from pattern recognition to biological interpretation. *Drug Discov. Today* **10**, 1551–1558, https://doi.org/10.1016/S1359-6446(05)03609-3
167 Nassar, A.-E.F. and Talaat, R.E. (2004) Strategies for dealing with metabolite elucidation in drug discovery and development. *Drug Discov. Today* **9**, 317–327, https://doi.org/10.1016/S1359-6446(03)03018-6
168 Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., Salamon, P. et al. (2003) Metagenomic analyses of an uncultured viral community from human feces metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220–6223, https://doi.org/10.1128/JB.185.20.6220-6223.2003
169 Zhang, T., Breitbart, M., Lee, W.H., Run, J.Q., Wei, C.L., Soh, S.W.L. et al. (2006) RNA viral community in human feces: Prevalence of plant pathogenic viruses. *PLoS Biol.* **4**, 0108–0118, https://doi.org/10.1371/journal.pbio.0040003
170 Kim, M.-S., Park, E.-J., Roh, S.W. and Bae, J.-W. (2011) Diversity and abundance of single-stranded DNA viruses in human feces. *Appl. Environ. Microbiol.* **77**, 8062–8070, https://doi.org/10.1128/AEM.06331-11
171 Minot, S., Grunberg, S., Wu, G.D., Lewis, J.D. and Bushman, F.D. (2012) Hypervariable loci in the human gut virome. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 3962–3966, https://doi.org/10.1073/pnas.1119061109
172 Minot, S. and Bryson, A. (2013) Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 12450–12455, https://doi.org/10.1073/pnas.1300833110
173 Pérez-Brocal, V., García-López, R., Vázquez-Castellanos, J.F., Nos, P., Beltrán, B., Latorre, A. et al. (2013) Study of the viral and microbial communities associated with Crohn's disease: a metagenomic approach. *Clin. Transl. Gastroenterol.* **4**, e36, https://doi.org/10.1038/ctg.2013.9
174 Lim, E.S., Zhou, Y., Zhao, G., Bauer, I.K., Droit, L., Ndao, I.M. et al. (2015) Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **21**, 1228–1234, https://doi.org/10.1038/nm.3950
175 Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., Keller, B.C. et al. (2015) Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460, https://doi.org/10.1016/j.cell.2015.01.002

PORTLAND PRESS

Clinical Science (2018) **132** 523–542
https://doi.org/10.1042/CS20171330

176 Evengård, B., Nord, C.E. and Sullivan, Å (2007) P1239 Patients with chronic fatigue syndrome have higher numbers of anaerobic bacteria in the intestine compared to healthy subjects. *Int. J. Antimicrob. Agents* **29**, S340, https://doi.org/10.1016/S0924-8579(07)71079-8

177 Butt, H., Dunstan, R., McGregor, N. and Roberts, T. (2001) Bacterial colonosis in patients with persistent fatigue. *Proceedings of the AHMF International Clinical and Scientific Conference*, Sydney, Australia

178 Reyes, A., Blanton, L., Cao, S., Zhao, G., Manary, M., Trehan, I. et al. (2015) Gut viromes of Malawian twins discordant for severe acute malnutrition. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11941–11946, https://doi.org/10.1073/pnas.1514285112

Appendix 2

frontiers in Microbiology

ORIGINAL RESEARCH
published: 23 October 2020
doi: 10.3389/fmicb.2020.583378

Genome Characterization of a Novel Wastewater *Bacteroides fragilis* Bacteriophage (vB_BfrS_23) and its Host GB124

Mohammad A. Tariq[1†], Fiona Newberry[1,2†], Rik Haagmans[1,2], Catherine Booth[1], Tom Wileman[1,2], Lesley Hoyles[3], Martha R. J. Clokie[4], James Ebdon[5] and Simon R. Carding[1,2]*

[1] Gut Microbes and Health Research Programme, Quadram Institute Biosciences, Norwich Research Park, Norwich, United Kingdom, [2] Norwich Medical School, University of East Anglia, Norwich, United Kingdom, [3] Department of Biosciences, Nottingham Trent University, Nottingham, United Kingdom, [4] Department of Genetics and Genome Biology, Leicester University, Leicester, United Kingdom, [5] Environment and Public Health Research Group, School of Environment and Technology, University of Brighton, Brighton, United Kingdom

OPEN ACCESS

Edited by:
Andrew S. Lang, Memorial University of Newfoundland, Canada

Reviewed by:
Alexander P. Hynes, McMaster University, Canada
Ananda Shankar Bhattacharjee, Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana–Champaign, United States

****Correspondence:***
Simon R. Carding
Simon.Carding@quadram.ac.uk

†These authors share first authorship

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *14 July 2020*
Accepted: *05 October 2020*
Published: *23 October 2020*

Citation:
Tariq MA, Newberry F, Haagmans R, Booth C, Wileman T, Hoyles L, Clokie MRJ, Ebdon J and Carding SR (2020) Genome Characterization of a Novel Wastewater Bacteroides fragilis Bacteriophage (vB_BfrS_23) and its Host GB124. Front. Microbiol. 11:583378. doi: 10.3389/fmicb.2020.583378

Bacteroides spp. are part of the human intestinal microbiota but can under some circumstances become clinical pathogens. Phages are a potentially valuable therapeutic treatment option for many pathogens, but phage therapy for pathogenic *Bacteroides* spp. including *Bacteroides fragilis* is currently limited to three genome-sequenced phages. Here we describe the isolation from sewage wastewater and genome of a lytic phage, vB_BfrS_23, that infects and kills *B. fragilis* strain GB124. Transmission electron microscopy identified this phage as a member of the *Siphoviridae* family. The phage is stable when held at temperatures of 4 and 60°C for 1 h. It has a very narrow host range, only infecting one host from a panel of *B. fragilis* strains ($n = 8$). Whole-genome sequence analyses of vB_BfrS_23 determined it is double-stranded DNA phage and is circularly permuted, with a genome of 48,011 bp. The genome encodes 73 putative open reading frames. We also sequenced the host bacterium, *B. fragilis* GB124 (5.1 Mb), which has two plasmids of 43,923 and 4,138 bp. Although this phage is host specific, its isolation together with the detailed characterization of the host *B. fragilis* GB124 featured in this study represent a useful starting point from which to facilitate the future development of highly specific therapeutic agents. Furthermore, the phage could be a novel tool in determining water (and water reuse) treatment efficacy, and for identifying human fecal transmission pathways within contaminated environmental waters and foodstuffs.

Keywords: bacteriophage, *Bacteroides fragilis*, genomes, waste water, human

INTRODUCTION

Bacteriophages (phage) are viruses that infect bacterial cells and as a result can influence their growth, fitness and response to stress (Casjens and Hendrix, 2015; Davies et al., 2016). They are estimated to be numerically the most abundant biological entity on earth numbering at least 10^{31} (Hendrix et al., 1999; Bar-On et al., 2018). Phages are also a major constituent of the human

Abbreviations: BPRM, *Bacteroides* phage recovery medium; CDS, coding sequence; ORF, open reading frame.

microbiome and in particular, the intestinal microbiota where they can outnumber bacterial cells and human cells by up to 10:1 (Sender et al., 2016).

Bacteroides spp. are a dominant component of the intestinal bacteriome, accounting for between 5 and 40% of all anaerobes (Gorvitovskaia et al., 2016). In a recent study looking at 98 gut samples the relative abundance of *Bacteroides* spp. ranged from 0.37 to 98.82% (King et al., 2019). Although *Bacteroides fragilis* represents a smaller fraction of this group, it was present in all samples (King et al., 2019). *Bacteroides* spp. confer significant health benefits to their host including the digestion, processing and extraction of nutrients from complex plant-based polysaccharides, promoting colonic motility and angiogenesis, and the development of the gut-associated immune system (Hooper et al., 2003; Xu et al., 2003; Mazmanian et al., 2005).

Bacteroides spp. are also important clinical pathogens and can contribute to anaerobic infections (Perez-Brocal et al., 2015). *B. fragilis* is one of the most commonly isolated anaerobic pathogens from soft tissue infections and bacteremia (Shenoy et al., 2017). The capsular polysaccharide complex of *B. fragilis* consisting of two distinct polysaccharides is the primary mediator of intra-abdominal abscess formation (Tzianabos et al., 1993). Enterotoxigenic *B. fragilis* also produces metalloprotease toxins (fragilysin), which unlike pore-forming toxins, breakdown connective tissue through proinflammatory cytokine signaling leading to an increase in the permeability of the epithelial barrier, causing diarrheal diseases and acute inflammation (Wu et al., 2004; Kim et al., 2006).

Anaerobic bacteria and their phages have been proposed as candidates for indicators of fecal pollution as they do not replicate in estuarine waters and the phages are resistant to chlorine inactivation (Booth et al., 1979; Tartera and Jofre, 1987; McMinn et al., 2014, 2017; Dias et al., 2018). Despite the clinical importance of *B. fragilis*, and a study reporting up to 3×10^4 PFU/100 mL of *B. fragilis* phages in sewage influent (Sun et al., 1997), only three complete genomes of *B. fragilis* phages have been described to date, two of which have been published with the third deposited under accession MN078104; all are virulent phages (Puig and Girones, 1999; Hawkins et al., 2008; Ogilvie et al., 2012). Here we describe the isolation of a new *B. fragilis* phage, vB_BfrS_23 from municipal wastewater and detail its genome characteristics. We also sequenced the host bacterium *B. fragilis* strain GB124 isolated from a United Kingdom municipal wastewater sample (Payan et al., 2005).

MATERIALS AND METHODS

Bacterial Culture and Growth Conditions

Bacteroides fragilis GB124 was used as the host reference strain for phage isolation and has been used to detect human fecal contamination in water sources (Payan et al., 2005; Ebdon et al., 2007), and to test the treatment efficacy of water reuse technologies (Purnell et al., 2015, 2016; Dias et al., 2018). *Bacteroides* phage recovery medium (BPRM) was used to cultivate host GB124 and propagate the phage (**Supplementary Table 1**).

Phage Isolation and Purification

Phage vB_BfrS_23 was isolated from 100 mL raw (untreated) municipal wastewater from a United Kingdom-based treatment plant. Wastewater was filtered with a 0.45 μm PES membrane syringe filter (Sartorius UK Ltd.) and concentrated by centrifugation for 15 min at 5,000 *g* using Amicon Ultra-15 10K centrifugal filter units. One milliliter of this concentrated sewage filtrate was mixed with 1 mL of mid-exponential growth phase (OD_{620nm} 0.3–0.4) *B. fragilis* GB124 allowing 5 min for adsorption and was then added to semi-soft BPRM agar (0.35%) and poured on BPRM agar (1.5%) (**Supplementary Table 1**; Ebdon et al., 2007). After 18 h anaerobic incubation (5% CO_2, 5% H_2 and 90% N at 37°C and ∼25 psi pressure) the plates were screened for plaques. A single plaque was picked using a sterile pipette and resuspended in 10 mL BPRM medium containing sub-cultured host (OD_{620nm} 0.3–0.4). The suspension was incubated for 18 h to allow further propagation of the phages. The sample was filtered through a 0.22 μM PES membrane filter (Sartorius UK Ltd.). The procedure was repeated a further three times to obtain a pure phage stock. This stock was used to further propagate and increase the phage titer. Fifty microliter was used to perform serial dilutions and was added to semi-soft BPRM agar (0.35%) with 200 μL of mid-log phase (OD_{620nm} 0.3–0.4) bacterial host. The plates were incubated for 16 h in an anaerobic cabinet (5% CO_2, 5% H_2 and 90% N at 37°C and ∼25 psi pressure). Five milliliter SM buffer (100 mM NaCl, 8.1 mM $MgSO_4.7H_2O$ and 50 mM Tris.HCl pH 7.4) was added to a plate of complete cell lysis and left on a mini gyratory shaker SSM3 (Stuart, United Kingdom) for 1 h. The top agar was harvested along with the buffer and transferred to a 50 mL tube (Corning, United Kingdom), after a brief vortex the tube was centrifuged at 3,000 × *g* for 10 min and the supernatant filtered through a PES membrane bottle top vacuum filter using ∼100 psi pressure (Millipore Millivac, Merck UK). The titer was evaluated using dilutions 10^{-1} to 10^{-9} and the titer adjusted to 1×10^8 PFU/mL for temperature assays and was stored at 4°C for the duration of the experiments.

Transmission Electron Microscopy

Briefly, a small drop of phage suspension containing ∼1×10^7 PFU/mL was applied to a formvar/carbon coated copper transmission electron microscopy (TEM) grid (Agar Scientific, Stansted, United Kingdom) and left for 1 min. Excess liquid was removed with filter paper. A small drop of 2% uranyl acetate (BDH 10288) was applied to the grid surface and left for a further 1 min after which it was removed with filter paper. Grids were left to thoroughly dry before viewing and imaging using a Talos F200c TEM with Gatan Oneview digital camera.

Host Range Assay

In total, eight *B. fragilis* strains (**Supplementary Table 2**) were used to determine the host range and specificity of vB_BfrS_23. Bacterial strains were cultured in BPRM broth to exponential

phase (OD_{620} 0.3–0.33) prior to incorporation into double BPRM agar overlays (Ebdon et al., 2007). Dilutions of vB_BfrS_23 were then spotted onto the double agar overlay and observed for plaques following 16 h in an anaerobic cabinet (5% CO_2, 5% H_2 and 90% N at 37°C and ~25 psi pressure).

One-Step Growth Curve and Eclipse Period

To determine the burst size and latency period of vB_BfrS_23, a one-step growth curve was carried out (Kropinski, 2018). Initially, 9.9 mL of *B. fragilis* GB124 was grown anaerobically (5% CO_2, 5% H_2 and 90% N at 37°C and ~25 psi pressure) to mid-exponential phase and 0.5 OD_{620}. One-hundred µL of 1×10^7 PFU/mL phage was then added for 5 min to allow phage adsorption. A 0.1 mL aliquot was then used to make ten-fold serial dilutions to a final dilution of 1×10^1. As an adsorption control, a 1 mL aliquot from the 1×10^3 dilution flask aliquot was added to 50 µL of $CHCl_3$ and kept on ice for the duration of the experiment (less than 4 h). At various time points 0.1 mL was taken from each dilution and mixed with 200 µL of bacterial host suspension (in BPRM broth) and plated using 0.35% (w/v) BPRM agar. The data were normalized by multiplying the adsorption control and the value obtained from 1×10^3 PFU/mL flask by x10, 1×10^2 PFU/mL flask by x100 and 1×10^1 PFU/mL flask by x1000. The burst size was determined as previously described (Kropinski, 2018).

At each sampling point 475 µL was taken to determine the eclipse period. The sample was added to 25 µL of chloroform (5%v/v), vortexed for 10 s and kept on ice until the end of the experiment to allow the chloroform to settle. One hundred microliter was taken from each timepoint sample and added to 200 µL of bacterial host and plated using 0.35% (w/v) BPRM agar. The plates were incubated for 16 h in an anaerobic cabinet (5% CO2, 5% H2 and 90% N at 37°C and ~25 psi pressure). Each one-step growth and eclipse experiment were repeated to give three biological replicates.

Thermal Assay

The viability of vB_BfrS_23 at different temperatures was assessed by incubating phage preparations of known titers at 4, 24, 30, 37, 40, 45, 60, 70, or 80°C for 15, 30, or 60 min out of direct sunlight. Serial dilutions of the phage stocks were incubated with 200 µL of bacterial host culture in BPRM broth for 15 min at 37°C prior to mixing with 5 mL BPRM semi-soft agar (0.35%, w/v) and pouring onto BPRM agar plates and incubated for 18 h at 37°C in anaerobic cabinet (5% CO_2, 5% H_2 and 90% N at 37°C and ~25 psi pressure). For accuracy, plaques were counted on plates containing between 30 and 300 plaques.

DNA Extraction

For Illumina sequencing, phage preparations (~10^9 PFU/mL) were incubated with RNase A (100 U Ambion) and Turbo DNase (2U Invitrogen) at 37°C for 30 min to remove bacterial chromosomal DNA. Nucleases were heat-inactivated at 65°C in the presence of 15 mM EDTA for 10min. The Norgen Phage DNA isolation kit (Geneflow Limited, Lichfield, United Kingdom) was used to extract the phage DNA. For Nanopore sequencing, phage was PEG-precipitated (10% (w/v) PEG 8000 and 6% (w/v) NaCl), treated with DNase (4 U Turbo DNase; Invitrogen) and RNase A (100 U; Ambion) followed by treatment with SDS (0.5%, w/v) and 4 µL (80 µg of proteinase K 20 mg/mL, Ambion) treatment at 55°C for 1 h and heat inactivation at 75°C for 15 min. Lipids and proteins were removed by mixing the sample 1:1 with chloroform and vigorous shaking for a few seconds followed by centrifugation at 15,000 *g* at 20°C for 5 min. The upper aqueous phase was treated with NaCl (0.2 M final concentration) prior to mixing 1:1 with isopropanol and left in -20°C for 16 h. The sample was centrifuged at 13,000 *g* at 20°C for 1 h followed by two washes with 70% ethanol prior to resuspending the pellet in nuclease-free water (Invitrogen). Bacterial DNA was extracted from an overnight culture of GB124 grown in BPRM broth, the sample was centrifuged at 3,000 *g* for 20 min, the pellet was resuspended using 300 µL of TE buffer in accordance to the Promega Maxwell® RSC Cultured Cell DNA Kit (AS1620) protocol (FB211) and run on the Promega Maxwell® RSC Instrument (AS4500).

DNA Sequencing

Phage and bacterial genomic DNA were sequenced using Illumina and MinION ONT sequencing platforms. For MinION sequencing, the standard ONT protocol and native barcoding kit EXP-NBD104 with the ligation sequencing kit SQK-LSK109 were used. In brief, 1 µg of high quality vB_BfrS_23 and *B. fragilis* GB124 DNA was end-repaired and dA-tailed using the NEBNext FFPE Repair Mix (M6630) and NEBNext End Repair/dA-tailing (E7546). The native barcode (EXP-NBD104 kit) was used to barcode the samples and they were ligated using NEB Blunt/TA Ligase Master Mix (M0367). The sequence adapters were ligated with NEBNext Quick Ligation Module (E6056) and the samples were primed and loaded using the Flow Cell Priming Kit (EXP-FLP001) on the MinION R9 4.1 FLO-MIN106. Samples were run for 72 h, and the raw reads were base-called using Guppy v3.5.1.[1] Adapters were removed using Porechop v0.2.3 (rrwick/Porechop, 2020).[2] Genomic bacterial DNA was also sequenced using the Illumina MiSeq system. Briefly, the Illumina Nextera XT (Illumina, Saffron Walden, United Kingdom) library preparation kit was used to prepare sequencing libraries prior to running on Illumina MiSeq 2 × 150-cycle v2 chemistry. Paired-end sequencing reads were provided as FASTQ files with the raw reads having their adapters removed using Trimmomatic (Bolger et al., 2014) prior to quality trimming using Sickle at –q 30 and –l 15 (Joshi and Fass, 2011).

Phage Genome Assembly and Annotation

Illumina MiSeq and MinION generated sequences were assembled using Unicycler v0.4.8 (Wick et al., 2017) resulting in a single contiguous circular sequence of 48,011 bp. The genome was annotated using RAST (Aziz et al., 2008; Overbeek et al., 2014; Brettin et al., 2015). The putative functions of the coding

[1]https://nanoporetech.com
[2]https://github.com/rrwick/Porechop

regions (CDS) were predicted using NCBI-nr (June 15, 2020) and Conserved Domain Database (CDD) (June 15, 2020) searches using BlastP and tBlastn. For Blastp and tBlastn, hits were considered significant if the e-values were lower than 1 e^{-5} at ≥60% protein identity (Altschul et al., 1990). For CDD searches, only hits with an e-value of 0.01 or lower were considered significant (Marchler-Bauer et al., 2015).

GB124 Genome Assembly, Quality Checks and Annotation

The genome was assembled using Illumina MiSeq and ONT MinION reads, using Unicycler (Wick et al., 2017). Following host assembly, the contig was annotated using Prokka v.1.14.6 (Seemann, 2014). Antimicrobial resistance genes were investigated with ABRicate v.0.9.8 (tseemann/abricate, 2020)[3] using Resfinder v3.2 (database version September 10, 2019) (Zankari et al., 2012), NCBI and AMRFinderPlus v3.8 (database version 2020-05-04.1) (Feldgarden et al., 2020). Insertion elements were predicted using ISfinder (Siguier et al., 2006).[4] Significant hits (Score > 100 and e value <4e-11) in ISfinder were examined in the Prokka GenBank file and protein sequence submitted to BlastP. Suspected insertion sequence (IS) elements were visualized in Artemis 18.1.0 (Carver et al., 2012) and investigated for downstream Anti-Microbial Resistance (antimicrobial resistance) genes. ABRicate hits were considered significant if the coverage and identity were >90%.

Plasmids were identified using the PLSDB web server[5] (data v2020_03_04) and coding regions found using Prokka v1.14.6 (Seemann, 2014; Galata et al., 2019). The putative functions assigned by Prokka were checked using BlastP according to NCBI-nr (July 1, 2020) and Conserved Domain Database (July 1, 2020) (Altschul et al., 1990; Marchler-Bauer et al., 2015). Hits were considered significant if the e-values were lower than $1e^{-}5$ at ≥ 80% protein identity. Plasmids were visualized using Brig v0.95 (Alikhan et al., 2011). The plasmids were screened for antimicrobial resistance and virulence genes using ABRicate v.0.9.8 and Resfinder, NCBI AMRFinderPlus and VFDB (Zankari et al., 2012; Chen et al., 2016; Feldgarden et al., 2020). The completeness and contamination of the seven-contig assembly were assessed using CheckM v1.0.18 (Parks et al., 2015). Average nucleotide identity with the type strain of *B. fragilis* was assessed using fastANI v1.2 (Jain et al., 2018) and *B. fragilis* CCUG 4856^{T} (RefSeq assembly accession GCF_005706655).

vB_BfrS_23 Phage Genome Comparison of Large Terminase Subunit and Tail Fiber

The coding region for the genes were tblastx searched using default parameter and amino acid sequences sharing identity to the large terminase subunit and the tail fiber sequences were aligned using MAFFT v7 (Katoh and Standley, 2013). The L-INS-i algorithm with default parameters was used to improve accuracy. The alignment file was used to create a p-distance analysis in MEGA7 (Kumar et al., 2016) following construction of a neighbor-joining tree on p-distance using 1,000 bootstrap analyses using default parameters (Saitou and Nei, 1987).

vB_BfrS_23 and ϕB124-14 Linear Genome Comparison Alignment

A detailed comparison of the vB_BfrS_23 with ϕ124-14 (Ogilvie et al., 2012), B40-8 and *Bacteroides* phage Barc2635 was performed using tBLASTx in Easyfig (Sullivan et al., 2011). The annotated GenBank file of vB_BfrS_23 was compared with the GenBank file for ϕB124-14, B40-8 and *Bacteroides* phage Barc2635.

RESULTS

Phage Isolation and Phenotypic Characterization

Bacteroides fragilis GB124 was used as a host for phage discovery and the starting point for the screening of a filtered raw wastewater sample. We identified and isolated a virulent phage capable of infecting GB124, that generated plaques that ranged in size between 0.5 and 2 mm (**Figure 1A**). TEM images revealed the presence of a non-contractile long tail ∼150 nm and an icosahedral head ∼50 nm in size consistent with vB_BfrS_23 belonging to the *Siphoviridae* family of the order *Caudovirales* (**Figure 1B**).

vB_BfrS_23 Phage Characteristics

Phage vB_BfrS_23 was seen to infect and lyse only one of the eight *B. fragilis* strains tested, GB124, which was the host strain used for isolating vB_BfrS_23 (**Supplementary Table 2**).

The one-step growth curve experiment (**Figure 1C**) showed that the phage had a burst size of ∼44 phage/cell (mean of three independent experiments) and latency period of ∼37 min. The eclipse period was determined to be ∼23 min ($n = 3$). It was also stable at temperatures between 4°C and 60°C (**Figure 1D**) with viability decreasing at 60°C with a more rapid inactivation seen at 70°C. No plaques were seen at 80°C. At 37°C plaques were of sizes up to 2 mm (**Supplementary Figure 7**). Interestingly, a slight increase in PFU/mL was seen between 40°C and 45°C, with the plaques being more uniform and smaller (0.5 mm) at 45°C (**Supplementary Figure 7**).

Phage Genome and Phylogeny

vB_BfrS_23 is a double-stranded DNA phage of 48,011 bp with a GC content of 38.6%, containing 73 putative CDS (**Figure 2** and **Supplementary Table 3**). It was most similar to the virulent phage ϕB124-14 (86% query coverage) followed by *Bacteroides* phage Barc2635 (85% query coverage) and then ϕB40-8 (73% query coverage). The linear genome comparison of the vB_BfrS_23, Barc2635, ϕB124-14 and ϕB40-8 phages is illustrated in **Figure 3**.

The terminase large subunit and the tail fiber proteins were used to generate a phylogenetic tree (**Figure 4**). Both the tail fiber (**Figure 4A**) and terminase large subunit (**Figure 4B**)

[3]https://github.com/tseemann/abricate
[4]http://www-is.biotoul.fr
[5]https://ccb-microbe.cs.uni-saarland.de/plsdb/

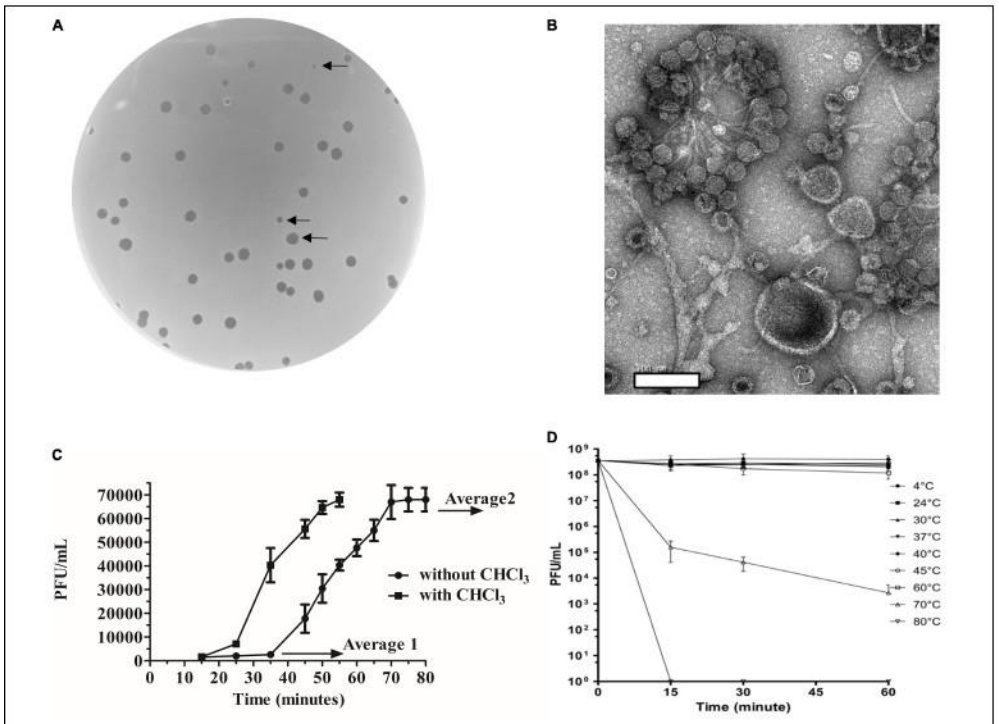

FIGURE 1 | Morphological and biological characteristics of vB_BfrS_23. **(A)** Plaques of vB_BfrS_23D seen on a lawn of *B. fragilis* GB124. Arrows identify plaques of different sizes (2mm to 0.5mm). **(B)** Negatively stained TEM images of vB_Bfr_23D. Scale bar = 200nm. **(C)** One-step growth curve: The latent and rise periods were determined to be approximately 37 and 30 min, respectively, with a rise period and a burst size per cell of approximately 44 phage/cell (mean of 3 replicates). The Average 1 value is the mean PFU/mL of three time points before the rise period with the Average 2 value being the mean PFU/mL of three time point after the one step burst. The error bars depict SEM values (n = 3 biological replicates). The eclipse period was determined to be about 23 min. **(D)** Temperature assay: The vB_BfrS_23 phage was stable up to 60°C, with declining viability at higher temperatures after 1 h incubation. The error bars depict SEM values (*n* = 3 biological replicates).

were shown to be most similar to ϕB124-14. BlastP[6] revealed 13 of the CDS had a putative function and 8 CDS contained conserved domain signatures. Most of the CDS were assignable to genome structure and replication/regulation, with the remainder associated with lysis and DNA structure. Putative CDS of similar function clustered together to form modules. However, 6 putative proteins identified in phage ϕB124-14 were not found within the vB_BfrS_23 genome (Ogilvie et al., 2012). Ten putative CDS showed no homology to any protein within the database, with 27 sharing the highest sequence similarity to genes in ϕB124-14, 27 to Barc2635 and 8 to ϕB40-8 (the following to prophage regions) and 1 to a *Bacteroides ovatus* phage (**Supplementary Table 3**).

Like ϕB124-14, Barc2635 and ϕB40-8, vB_BfrS_23 lacked an obvious virulent genome module and only contained 1 putative protein that alluded to a strictly lytic life cycle (CD 18). CD18 exhibited closest homology to a putative peptidase protein identified within ϕB124-14 and contained a peptidase superfamily domain. The peptidase protein appeared to reside within a cluster of unassigned proteins, suggesting it may be a putative virulent module.

Five CDs were assigned a predicted function relating to virus replication and regulation. CD11 encoded a putative thymidylate synthase, which is a key enzyme in the synthesis of 2′-deoxythymidine-5′-monophosphate, an essential precursor for DNA replication. A conserved domain region identified within the protein suggested it encodes a ThyA-like enzyme as reported in the ϕB124-14 genome (Ogilvie et al., 2012). CD7 (recombination protein) and CD70 (anti-repressor) were also encoded within the replication and regulation genome module, promoting transcription of phage genes (Lemire et al., 2011).

[6]https://blast.ncbi.nlm.nih.gov/Blast.cgi

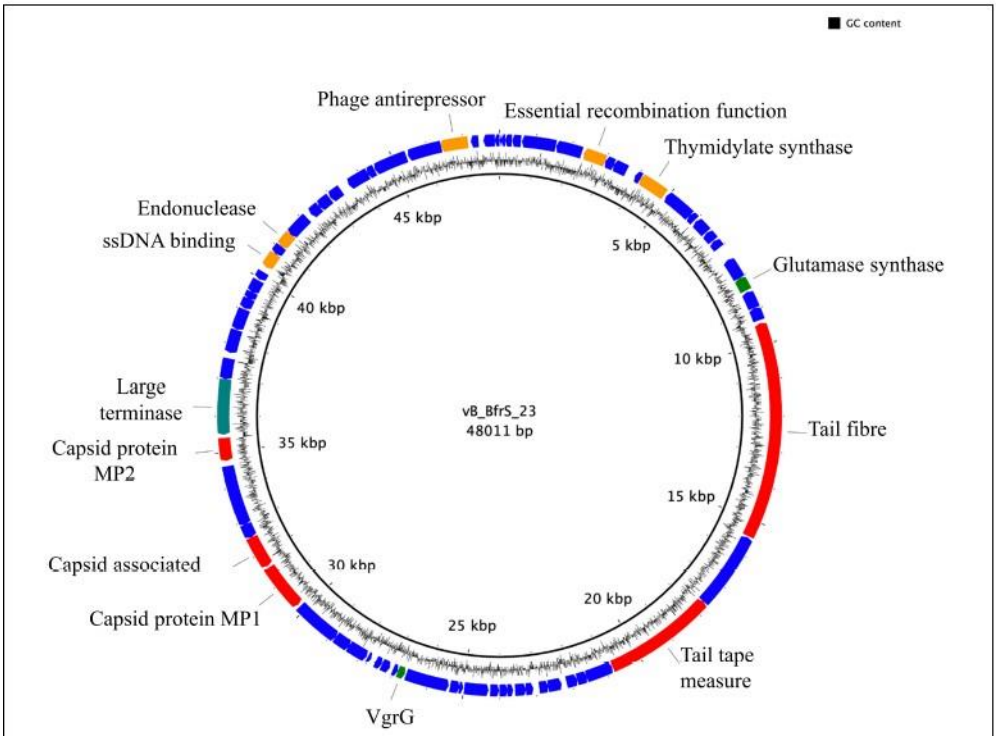

FIGURE 2 | Genome map of vB_BfrS_23. The prediction and direction of the coding regions are indicated by arrow heads. The blue coding regions are hypothetical proteins with no known function. The 13 putatively known functions of the phage are annotated. Also shown are genes with predicted roles in replication and regulation (orange), DNA packaging (green), and structure (red).

B. fragilis GB124 Genome Assembly and Annotation

The genome was assembled into 7 contigs of >100 bp (N50 4,986,460 bp). CheckM analysis showed the genome to be 99.26% complete with no contamination. It shared 99.03% average nucleotide identity with *B. fragilis* CCUG 4856^T, confirming GB124 as an authentic strain of *B. fragilis* (Chun et al., 2018). Two contigs were complete assemblies for plasmids which were identified using PLSDB and were named PBf1 and PBf2. PBf1 consisted of 4148 bp, was circular and contained eight predicted open reading frames (ORFs; **Figures 5A–C** and **Supplementary Figure 4**). PLSDB revealed an exact match to *B. xylanisolvens* strain H207 plasmid unnamed2 (NZ_CP041232.1). A further Blastn search showed a 100% identity and query cover match to one other plasmid, *B. ovatus* strain 3725 D1 iv plasmid unnamed3. PBf1 contained a YoeB toxin, a toxic component of a type II toxin-antitoxin system that helps to maintain plasmid stability by post-segregation killing or genetic addition (**Figure 5B** and **Supplementary Figure 5**) (Gerdes et al., 1986; Yarmolinsky, 1995). PBf2 consisted of 43,923 bp, was circular and contained 57 coding regions (5 domains of unknown function, 30 hypothetical proteins and 22 of putative function) (**Figure 5C**). A PLSDB search revealed 2 hits: *B. ovatus* strain 3725 D1 iv plasmid unnamed2 (NZ_CP041397.1) and *B. thetaiotaomicron* F9-2 plasmid p1-F9 DNA (AP022661.1). Interestingly, an additional blastn search reported a 97% query cover and 98.2% percentage identify match to *B. xylanisolvens* strain H207 plasmid unnamed1. No virulence or antibiotic resistance genes were identified on either plasmid.

The remaining five contigs were identified as belonging to the GB124 genome of 5,093,249 bp with a GC content of 43.87%. A total of 4266 ORFs were predicted of which 72 were tRNA, 1 tmRNA and 2 CRISPR repeat regions (**Supplementary Table 6**). Two incomplete prophage regions were identified using PHASTER (Arndt et al., 2016). They were 42.5 and 13.2 kbp in size, all features are depicted in **Figure 5A**. A total of 5 IS elements were identified. However, investigation of the GenBank file and

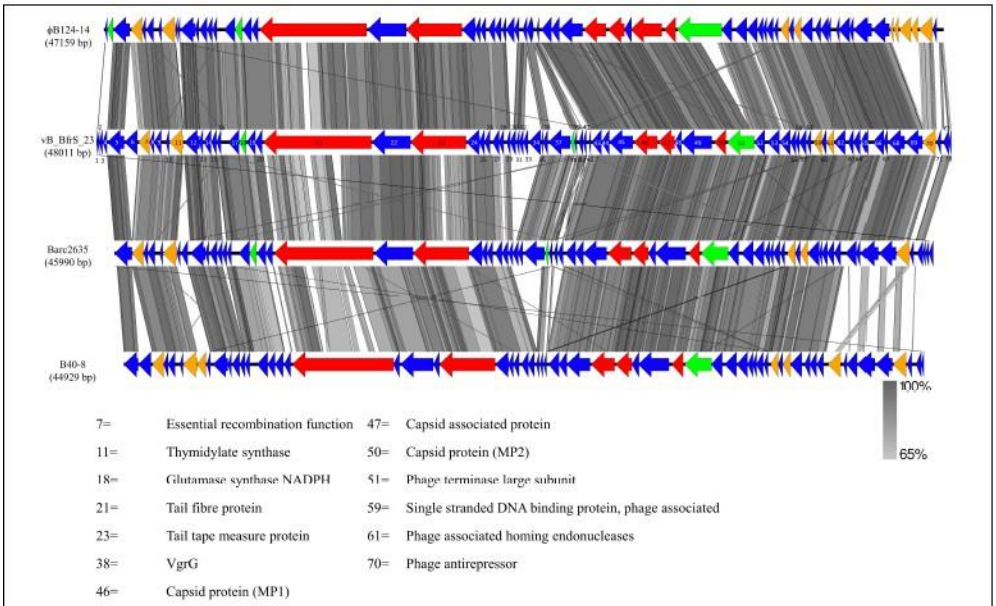

FIGURE 3 | Genome comparison of vB_BfrS_23, Barc2635, ϕB124-14, and B40-8. Sequence similarity is represented by the gray scale bar. Blue colored regions are hypothetical proteins with no known function. Genes with predicted role in replication and regulation (orange), DNA packaging terminase (green) and structural (red) are also shown. The coding sequences (CDS) for vB_BfrS_23 are numbered and represented by directional arrows. CDS with putative function are listed and the numbers are linked to the CDS of vB_BfrS_23.

protein sequences revealed only 1 IS element, IS1182 family ISBf3. Genes flanking the IS element had no known function. No *B. fragilis* virulence factors were identified. Resfinder, NCBI and AMRFinderPlus databases revealed two antimicrobial resistance genes (*cepA* and *tetQ*). The *cepA* gene, observed in >90% of *B. fragilis* isolates, encodes the β-lactamase protein and confers resistance to cephalosporins (except cefoxitin) and penicillin (Rogers et al., 1993; Mastrantonio et al., 1996). *tetQ* gene-related resistance is common among *B. fragilis* isolates and encodes a protein that protects the bacterial ribosomes from tetracycline (Rasmussen et al., 1993; Roberts, 1996). Bacterial contigs >200 bp were submitted to GenBank thus omitting 162bp contig from the assembly.

DISCUSSION

The escape of *B. fragilis* from the gut environment into other parts of the body can result in major pathology, including bacteremia and abscess formation in various tissues. Although phages are a therapeutic option to treat and kill pathogenic *B. fragilis* strains, only three to date had been described and their genomes sequenced. Here, we identify a new highly specific virulent phage that is only able to infect a single host among a panel of *B. fragilis* strains tested. This supports similar finding of ϕ B124-14 which infected 5 out of 15 *B. fragilis* spp. tested (Ogilvie et al., 2012). This narrow host range may reflect extreme niche specialization exhibited by close phylogenetic and evolutionary relationships to gut bacteria (Zaneveld et al., 2010; Ogilvie et al., 2012). The morphological features of the phage identify it as *Siphoviridae*. The absence of any identifiable genes essential for the lysogenic life cycle is consistent with vB_BfrS_23 being a virulent phage. Despite identifying recombinase (CD7) and anti-repressor genes (CD70) which are associated with temperate life cycles, no integrase or excision genes that are essential for lysogenic life cycle were identified. Recombinase and anti-repressor genes have also been identified in ϕB124-14, B40-8 and Barc2635 (**Figure 3**). The investigators that initially described ϕB124-14 concluded that it was a virulent phage based upon a deviation in GC content between the phage and host, as we have seen between vB_BfrS_23 (38.6% GC content) and *B. fragilis* GB124 (43.87% GC content) (Deschavanne et al., 2010; Ogilvie et al., 2012). Thus, we assume that vB_BfrS23 is a virulent phage that may be a model candidate for human-specific microbial source tracking in contaminated surface and groundwater.

The phage vB_BfrS_23 contains a putative peptidase protein, but lacks any homology to known holin proteins (small

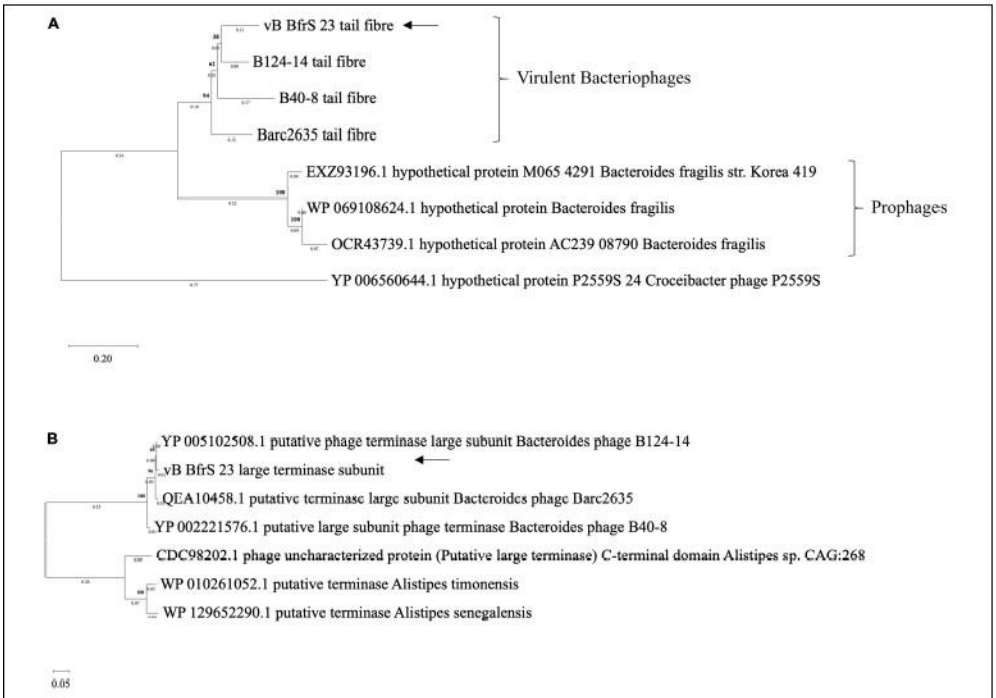

FIGURE 4 | Phylogenetic comparison of vB_BfrS_23. Amino acid sequences of tail fiber proteins **(A)** and large terminase subunit proteins **(B)** were aligned using MAFFT v7 the L-INS-i algorithm with default parameters used to improve accuracy with the p-distance used to generate a neighbor joining tree with 1000 bootstrap replicate. The arrows depict vB_BfrS_23 location in the tree. The bold number above the node show the percentage bootstrap. The sum of the tree length in panel **(A)** is 2.17 and 0.87 in panel **(B)**.

membrane proteins) which is not unusual for phages belonging to the *Siphoviridae* family (Hawkins et al., 2008; Duhaime et al., 2011). Double-stranded DNA phages typically lyse host cells using a holin-endolysin system. Active degradation of bacterial peptidoglycan is achieved with a muralytic enzyme or endolysin (Young, 1992; Young and Blasi, 1995). Endolysins accumulate in an active state in the cytosol, the holin proteins bind to the membrane, and the membrane is permeabilized to the endolysin. This leads to the breakdown of murein's resulting in the cell bursting. All this is time dependent and is programmed into the holin gene (Wang et al., 2000). It appears that the putative peptidase protein resides within an undefined lytic life cycle module in which there may be a holin-endolysin system. Interestingly, a putative thymidylate synthase was identified (CDS11) within the replication and regulation gene module. It is highly conserved across bacterial and mammalian species and shares remarkable structural and functional similarities (Carreras and Santi, 1995; Escartin et al., 2008). The exact function of ThyA within the phage genome is unknown but its additional copies may be of importance for survival of its host by enhancing growth (Stern et al., 2010). No tRNA genes were identified. The genome map highlights only 13 of the 73 predicted coding regions with a putative function, emphasizing the fact that phages are under-characterized.

In comparing the genome of vB_BfrS_23 with that of ϕB124-14, the former is 852 bp larger. Both genomes have genes unique to them that are primarily located around the same gene module and near the cos site, possibly due to recombination events. The Barc2635 genome is 2,021 bp smaller than vB_BfrS_23. There are distinct putative genes present in vB_BfrS_23 that are missing in Barc2635 including CDS 6, 31-34 and 71 hypothetical proteins. The tail fiber protein (CDS 21) is also smaller in Barc2635, B40-8 and ϕB124-14 compared to vB_BfrS_23 by 40 – 483bp (**Figure 3**). The relatively large tail fiber is consistent with that is seen for *Bacteroides fragilis* phages.

The large terminase subunit and tail fiber phylogenetic comparison shows these genes share homology with other known *B. fragilis* phages. Furthermore, the large terminase subunit is smaller in vB_BfrS_23 (CDS 51) and Barc2635 compared to ϕB124-14. Interestingly the lytic tail fiber genes share higher

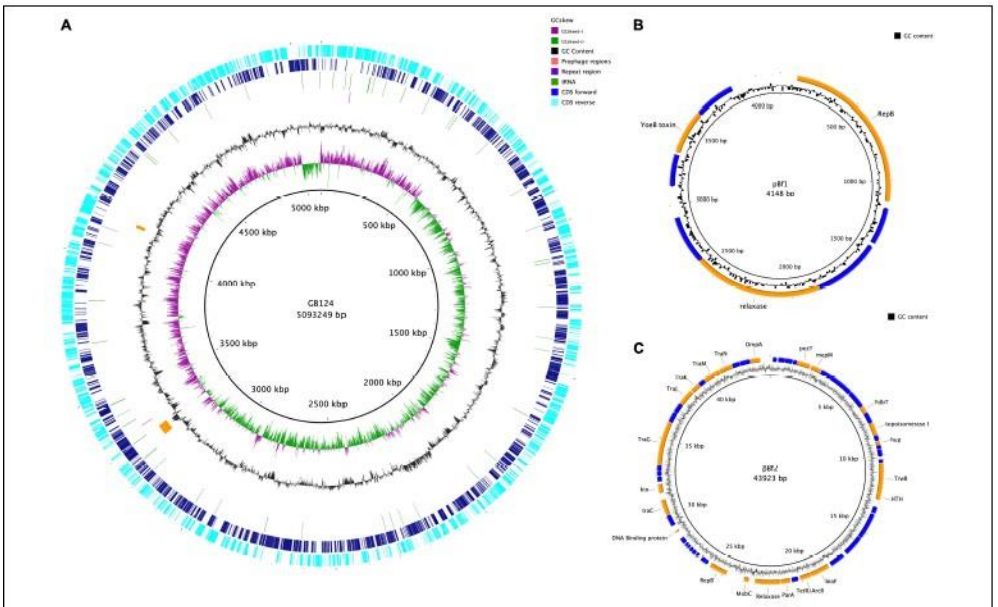

FIGURE 5 | Genome map of *B. fragilis* GB124 and its plasmids. **(A)** The outer most ring depicts the coding sequences (CDS) in both anti-clockwise (aqua) and clockwise (blue) orientation. tRNAs are depicted as green arcs with the purple arcs depicting 2 CRISPR repeat regions. The orange arcs depict two prophage regions with by GC content shown as black and GC skewing by green and purple. **(B)** Circular map of plasmid 1 (S4). The blue coding regions are hypothetical proteins with no known function. The orange coding regions highlights a putative function. **(C)** Circular map of plasmid 2 (S5) using the same color coding as in panel **(B)**.

nucleotide identity with other virulent *B.fragilis* phages compared to the prophage tail fibers, as they appear on another clade. This may be due to the life cycle of the phages and may be an indicator of differences in host range of virulent versus temperate phages. The large terminase subunit shares more identity to ϕB124-14, Barc2635 and B40-8 *B. fragilis* phages (**Figure 4A**).

The thermal assays provided useful information about vB_BfrS_23. The phage was resilient to the lower temperatures tested (4, 24, 30, 37°C) and at 40 and 45°C an increase in the number of plaques was observed. A key observation is that the phage forms aggregates (**Figure 1B**) and at higher temperatures (≤ 40 and 45°C) the plaque size was consistently 0.5 mm (**Supplementary Figure 7**). This characteristic of *B. fragilis* phages showing different plaques sizes and the phage agglomeration or clumping has been reported previously (Keller and Traub, 1974) for *B. fragilis* phages isolated from animal sera (Keller and Traub, 1974). It is possible therefore that the temperature kinetics impacts phage clustering by influencing the binding between phages, potentially mediated by Ig-like or carbohydrate adherence domains (CAD; Sathaliyawala et al., 2010; Barr et al., 2013; Tariq et al., 2015) enabling a more accurate determination of phage numbers in the sample. Alternatively, this could be an artifact of the experimental procedure used to generate samples for EM imaging including the duration of vortexing, the media and need for high phage titers. The vB_BfrS_23 thermal stability characteristics supports work done previously with naturally occurring GB124 phages (Bertrand et al., 2012; McMinn et al., 2014). The first of these studies showed the phages persisted longer at 5°C compared to higher temperatures (20 and 35°C), where few plaques were observed after 7 days (McMinn et al., 2014). Similarly, a meta-analysis of virus inactivation found inactivation occurred faster at temperatures ≥50°C than at <50°C. Adaptation to temperature is important in viral ecology as it influences phage infection, propagation and importantly, viability (Meschke and Sobsey, 2003; Jacquet et al., 2005; Ebdon et al., 2007; Jonczyk et al., 2011).

The exact origin of *B. fragilis* strain GB124 is unknown. It was isolated from untreated wastewater from a treatment plant in South East England and wastewater and impacted surface waters in the United Kingdom, United States, Brazil, and India (Payan et al., 2005; Prado et al., 2018; Wadhwa et al., 2018). Coupled with previously reported existence of a clear human gut-associated eco-genomic signature within the *Bacteroides* phage genomes (Ogilvie et al., 2012, 2013, 2018), it is assumed to be a human gut commensal.

Bacteroides fragilis phages when compared with other *Bacteroides* spp. phages have been shown to share little homology (Gilbert et al., 2017). A study looking at ϕBrb01

and ϕBrb02 phages, that are capable of infecting a *Bacteroides* isolate, has shown them to be phylogenetically distant to both ϕB124-14 and ϕB40-8 based on the comparative analysis of the large terminase subunit gene (Gilbert et al., 2017). Similarly, vB_BfrS_23 shares little or no sequence identity to 27 recently published *B. thetaiotaomicron* phages (Hryckowian et al., 2020). Although some CrAssphages such as ϕCrAss001 infect *Bacteroides intestinalis* (Shkoporov et al., 2018), they share little sequence identity to ϕB124-14 and ϕB40-8 (Garcia-Aljaro et al., 2017). Considering that the *B. fragilis* phage presented in this study shares a high sequence homology to other *B. fragilis* phages including ϕB124-14 and ϕB40-8 and thus we can infer that vB_BfrS_23 is also unrelated to other *Bacteroides* spp phages and CrAssphages.

CONCLUSION

The isolation and characterization of phage vB_BfrS_23 not only adds to and builds on fledgling phage databases, but it should also facilitate the detection other *Bacteroides* phages in human fecal metagenomes. As both the bacterial host and the new phage reported here are sourced from municipal wastewater they have considerable potential as, (1) highly specific novel therapeutic agents, (2) as tools for testing the efficacy of water and wastewater reuse technologies (spiking studies), and (3) as molecular or metagenome-based Microbial Source Tracking genetic marker for identifying human fecal transmission pathways in contaminated water and food (McMinn et al., 2014; Dias et al., 2018).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, SAMN14 843706 SRA; https://www.ncbi.nlm.nih.gov/genbank/, SRX8283257; https://www.ncbi.nlm.nih.gov/genbank/, SRX SRX828326; https://www.ncbi.nlm.nih.gov/genbank/, SRX8275163; https://www.ncbi.nlm.nih.gov/genbank/, SRX8275162.

AUTHOR CONTRIBUTIONS

SC and MT conceived and designed the experiments. MT, FN, JE, LH, and SC wrote the manuscript. SC supervised the research. MT, FN, RH, CB, and LH executed the experimental work. MT, FN, LH, MC, TW, JE, and SC carried out the data interpretation. All authors revised, read, and approved the final manuscript.

FUNDING

This work was supported in part by the UK Biotechnology and Biological Sciences Research Council (BBSRC) and BBSRC Institute Strategic Programme grant BB/R012490/1 to the Gut Microbes and Health programme and its constituent project(s) BBS/E/F/000PR10353 and BBS/E/F/000PR10356 (SRC), Invest in ME Research and UEA-Faculty of Health co-funded Ph.D. studentships (FN and RH), and award R42894 from The Solve ME/CFS Initiative (FN).

ACKNOWLEDGMENTS

We thank Dr. Lesley Ogilvie (Max Planck Institute for Molecular Genetics, Berlin, Germany) and Dr. Brian Jones (Dept of Biology and Biochemistry, University of Bath, BA2 7AY, UK) for useful discussions during design and analysis of this study. Additionally, we thank the late Dr. Ella Bond (Norwich, UK) for providing *Bacteroides fragilis* strains. We would like to thank Shen-Yuan Hsieh (Gut Microbes and Health Research Programme, Quadram Institute Biosciences, Norwich Research Park, Norwich, UK and Norwich medical School, University of East Anglia, Norwich NR4 7TJ, UK) for useful laboratory support. We thank the JIC Bioimaging facility and staff for enabling TEM phage analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.583378/full#supplementary-material

REFERENCES

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. doi: 10.1186/1471-2164-12-402

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y. J., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21. doi: 10.1093/nar/gkw387

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Bar-On, Y. M., Phillips, R., and Milo, R. (2018). The biomass distribution on Earth. *Proc. Natl. Acad. Sci. U.S.A.* 115, 6506–6511. doi: 10.1073/pnas.1711842115

Barr, J. J., Auro, R., Furlan, M., Whiteson, K. L., Erb, M. L., Pogliano, J., et al. (2013). Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10771–10776. doi: 10.1073/pnas.1305923110

Bertrand, I., Schijven, J. F., Sanchez, G., Wyn-Jones, P., Ottoson, J., Morin, T., et al. (2012). The impact of temperature on the inactivation of enteric viruses in food and water: a review. *J. Appl. Microbiol.* 112, 1059–1074. doi: 10.1111/j.1365-2672.2012.05267.x

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Booth, S. J., Vantassell, R. L., Johnson, J. L., and Wilkins, T. D. (1979). Bacteriophages of *Bacteroides*. *Rev. Infect. Dis.* 1, 325–336.

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., et al. (2015). RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* 5:8365. doi: 10.1038/srep08365

Carreras, C. W., and Santi, D. V. (1995). The catalytic mechanism and structure of thymidylate synthase. *Ann. Rev. Biochem.* 64, 721–762. doi: 10.1146/annurev.bi.64.070195.003445

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., and McQuillan, J. A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 28, 464–469. doi: 10.1093/bioinformatics/btr703

Casjens, S. R., and Hendrix, R. W. (2015). Bacteriophage lambda: early pioneer and still relevant. *Virology* 479, 310–330. doi: 10.1016/j.virol.2015.02.010

Chen, L. H., Zheng, D. D., Liu, B., Yang, J., and Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis-10 years on. *Nucleic Acids Res.* 44, D694–D697. doi: 10.1093/nar/gkv1239

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D. R., da Costa, M. S., et al. (2018). Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.* 68, 461–466. doi: 10.1099/ijsem.0.002516

Davies, E. V., Winstanley, C., Fothergill, J. L., and James, C. E. (2016). The role of temperate bacteriophages in bacterial infection. *FEMS Microbiol. Lett.* 363:fnw015. doi: 10.1093/femsle/fnw015

Deschavanne, P., Dubow, M. S., and Regeard, C. (2010). The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. *Virol. J.* 7:163. doi: 10.1186/1743-422x-7-163

Dias, E., Ebdon, J., and Taylor, H. (2018). The application of bacteriophages as novel indicators of viral pathogens in wastewater treatment systems. *Water Res.* 129, 172–179. doi: 10.1016/j.watres.2017.11.022

Duhaime, M. B., Wichels, A., Waldmann, J., Teeling, H., and Glockner, F. O. (2011). Ecogenomics and genome landscapes of marine *Pseudoalteromonas* phage H105/1. *ISME J.* 5, 107–121. doi: 10.1038/ismej.2010.94

Ebdon, J., Maite, M., and Taylor, H. (2007). The application of a recently isolated strain of *Bacteroides* (GB-124) to identify human sources of faecal pollution in a temperate river catchment. *Water Res.* 41, 3683–3690. doi: 10.1016/j.watres.2006.12.020

Escartin, F., Skouloubris, S., Liebl, U., and Myllykallio, H. (2008). Flavin-dependent thymidylate synthase X limits chromosomal DNA replication. *Proc. Natl. Acad. Sci. U.S.A.* 105, 9948–9952. doi: 10.1073/pnas.0801356105

Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., et al. (2020). Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* 64:e00361-20. doi: 10.1128/AAC.00361-20

Galata, V., Fehlmann, T., Backes, C., and Keller, A. (2019). PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.* 47, D195–D202. doi: 10.1093/nar/gky1050

Garcia-Aljaro, C., Balleste, E., Muniesa, M., and Jofre, J. (2017). Determination of crAssphage in water samples and applicability for tracking human faecal pollution. *Microb. Biotechnol.* 10, 1775–1780. doi: 10.1111/1751-7915.12841

Gerdes, K., Rasmussen, P. B., and Molin, S. (1986). Unique type of plasmid maintenance function - postsegregational killing of plasmid-free cells. *Proc. Natl. Acad. Sci. U.S.A.* 83, 3116–3120.

Gilbert, R. A., Kelly, W. J., Altermann, E., Leahy, S. C., Minchin, C., Ouwerkerk, D., et al. (2017). Toward understanding phage: host interactions in the rumen; complete genome sequences of lytic phages infecting rumen bacteria. *Front. Microbiol.* 8:2340. doi: 10.3389/fmicb.2017.02340

Gorvitovskaia, A., Holmes, S. P., and Huse, S. M. (2016). Interpreting *Prevotella* and *Bacteroides* as biomarkers of diet and lifestyle. *Microbiome* 4:15. doi: 10.1186/s40168-016-0160-7

Hawkins, S. A., Layton, A. C., Ripp, S., Williams, D., and Sayler, G. S. (2008). Genome sequence of the *Bacteroides fragilis* phage ATCC 51477-B1. *Virol. J.* 5:97. doi: 10.1186/1743-422x-5-97

Hendrix, R. W., Smith, M. C. M., Burns, R. N., Ford, M. E., and Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl. Acad. Sci. U.S.A.* 96, 2192–2197. doi: 10.1073/pnas.96.5.2192

Hooper, L. V., Stappenbeck, T. S., Hong, C. V., and Gordon, J. I. (2003). Angiogenins: a new class of microbicidal proteins involved in innate immunity. *Nat. Immunol.* 4, 269–273. doi: 10.1038/ni888

Hryckowian, A. J., Merrill, B. D., Porter, N. T., Van Treuren, W., Nelson, E. J., Garlena, R. A., et al. (2020). *Bacteroides thetaiotaomicron*-infecting bacteriophage isolates inform sequence-based host range predictions. *Cell Host Microbe* 28, 371–379.e5. doi: 10.1016/j.chom.2020.06.011

Jacquet, S., Domaizon, I., Personnic, S., Ram, A. S. P., Hedal, M., Duhamal, S., et al. (2005). Estimates of protozoan- and viral-mediated mortality of bacterioplankton in Lake Bourget (France). *Freshwater Biol.* 50, 627–645. doi: 10.1111/j.1365-2427.2005.01349.x

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9:5114. doi: 10.1038/s41467-018-07641-9

Jonczyk, E., Klak, M., Miedzybrodzki, R., and Gorski, A. (2011). The influence of external factors on bacteriophages-review. *Folia Microbiol.* 56, 191–200. doi: 10.1007/s12223-011-0039-8

Joshi, N. A., and Fass, J. N. (2011). *Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files [Online]*. Available online at: https://github.com/najoshi/sickle (accessed July 24, 2019).

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Keller, R., and Traub, N. (1974). Characterization of *Bacteroides-fragilis* bacteriophage recovered from animal sera - observations on nature of *Bacteroides* phage carrier cultures. *J. Gen. Virol.* 24, 179–189. doi: 10.1099/0022-1317-24-1-179

Kim, J. M., Lee, J. Y., Yoon, Y. M., Oh, Y. K., Kang, J. S., Kim, Y. J., et al. (2006). *Bacteroides fragilis* enterotoxin induces cyclooxygenase-2 and fluid secretion in intestinal epithelial cells through NF-kappa B activation. *Eur. J. Immunol.* 36, 2446–2456. doi: 10.1002/eji.200535808

King, C. H., Desai, H., Sylvetsky, A. C., LoTempio, J., Ayanyan, S., Carrie, J., et al. (2019). Baseline human gut microbiota profile in healthy people and standard reporting template. *PLoS One* 14:e0206484. doi: 10.1371/journal.pone.0206484

Kropinski, A. M. (2018). Practical advice on the one-step growth curve. *Methods Mol. Biol.* 1681, 41–47. doi: 10.1007/978-1-4939-7343-9_3

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Lemire, S., Figueroa-Bossi, N., and Bossi, L. (2011). Bacteriophage crosstalk: coordination of prophage induction by trans-acting antirepressors. *PLoS Genet.* 7:e1002149. doi: 10.1371/journal.pgen.1002149

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S. N., Chitsaz, F., Geer, L. Y., et al. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 43, D222–D226. doi: 10.1093/nar/gku1221

Mastrantonio, P., Cardines, R., and Spigaglia, P. (1996). Oligonucleotide probes for detection of cephalosporinases among *Bacteroides* strains. *Antimicrob. Agents Chemother.* 40, 1014–1016. doi: 10.1128/Aac.40.4.1014

Mazmanian, S. K., Liu, C. H., Tzianabos, A. O., and Kasper, D. L. (2005). An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* 122, 107–118. doi: 10.1016/j.cell.2005.05.007

McMinn, B. R., Ashbolt, N. J., and Korajkic, A. (2017). Bacteriophages as indicators of faecal pollution and enteric virus removal. *Lett. Appl. Microbiol.* 65, 11–26. doi: 10.1111/lam.12736

McMinn, B. R., Korajkic, A., and Ashbolt, N. J. (2014). Evaluation of *Bacteroides fragilis* GB-124 bacteriophages as novel human-associated faecal indicators in the United States. *Lett. Appl. Microbiol.* 59, 115–121. doi: 10.1111/lam.12252

Meschke, J. S., and Sobsey, M. D. (2003). Comparative reduction of Norwalk virus, poliovirus type 1, F+ RNA coliphage MS2 and *Escherichia coli* in miniature soil columns. *Water Sci. Technol.* 47, 85–90.

Ogilvie, L. A., Bowler, L. D., Caplin, J., Dedi, C., Diston, D., Cheek, E., et al. (2013). Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences. *Nat. Commun.* 4:2420. doi: 10.1038/ncomms3420

Ogilvie, L. A., Caplin, J., Dedi, C., Diston, D., Cheek, E., Bowler, L., et al. (2012). Comparative (meta)genomic analysis and ecological profiling of human gut-specific bacteriophage phi B124-14. *PLoS One* 7:e35053. doi: 10.1371/journal.pone.0035053

Ogilvie, L. A., Nzakizwanayo, J., Guppy, F. M., Dedi, C., Diston, D., Taylor, H., et al. (2018). Resolution of habitat-associated ecogenomic signatures in bacteriophage genomes and application to microbial source tracking. *ISME J.* 12, 942–958. doi: 10.1038/s41396-017-0015-7

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* 42, D206–D214. doi: 10.1093/nar/gkt1226

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Payan, A., Ebdon, J., Taylor, H., Gantzer, C., Ottoson, J., Papageorgiou, G. T., et al. (2005). Method for isolation of *Bacteroides* bacteriophage host strains suitable for tracking sources of fecal pollution in water. *Appl. Environ. Microbiol.* 71, 5659–5662. doi: 10.1128/Aem.71.9.5659-5662.2005

Perez-Brocal, V., Garcia-Lopez, R., Nos, P., Beltran, B., Moret, I., and Moya, A. (2015). Metagenomic analysis of crohn's disease patients identifies changes in the virome and microbiome related to disease status and therapy, and detects potential interactions and biomarkers. *Inflamm. Bowel Dis.* 21, 2515–2532. doi: 10.1097/Mib.0000000000000549

Prado, T., Bruni, A. D., Barbosa, M. R. F., Bonanno, V. M. S., Garcia, S. C., and Sato, M. I. Z. (2018). Distribution of human fecal marker GB-124 bacteriophages in urban sewage and reclaimed water of Sao Paulo city, Brazil. *J. Water Health* 16, 289–299. doi: 10.2166/wh.2017.011

Puig, M., and Girones, R. (1999). Genomic structure of phage B40-8 of *Bacteroides fragilis*. *Microbiol Sgm* 145, 1661–1670. doi: 10.1099/13500872-145-7-1661

Purnell, S., Ebdon, J., Buck, A., Tupper, M., and Taylor, H. (2015). Bacteriophage removal in a full-scale membrane bioreactor (MBR)–implications for wastewater reuse. *Water Res.* 73, 109–117. doi: 10.1016/j.watres.2015.01.019

Purnell, S., Ebdon, J., Buck, A., Tupper, M., and Taylor, H. (2016). Removal of phages and viral pathogens in a full-scale MBR: implications for wastewater reuse and potable water. *Water Res.* 100, 20–27. doi: 10.1016/j.watres.2016.05.013

Rasmussen, B. A., Bush, K., and Tally, F. P. (1993). Antimicrobial resistance in *Bacteroides*. *Clin. Infect. Dis.* 16, S390–S400. doi: 10.1093/clinids/16.Supplement_4.S390

Roberts, M. C. (1996). Tetracycline resistance determinants: mechanisms of action, regulation of expression, genetic mobility, and distribution. *FEMS Microbiol. Rev.* 19, 1–24. doi: 10.1111/j.1574-6976.1996.tb00251.x

Rogers, M. B., Parker, A. C., and Smith, C. J. (1993). Cloning and characterization of the endogenous cephalosporinase gene, cepa, from *Bacteroides fragilis* reveals a new subgroup of ambler class-a beta-lactamases. *Antimicrob. Agents. Chemother.* 37, 2391–2400. doi: 10.1128/Aac.37.11.2391

rrwick/Porechop (2020). *Porechop [Online]*. Available online at: https://github.com/rrwick/Porechop (accessed Jun 10, 2020).

Saitou, N., and Nei, M. (1987). The neighbor-joining method–a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Sathaliyawala, T., Islam, M. Z., Li, Q., Fokine, A., Rossmann, M. G., and Rao, V. B. (2010). Functional analysis of the highly antigenic outer capsid protein, Hoc, a virus decoration protein from T4-like bacteriophages. *Mol. Microbiol.* 77, 444–455. doi: 10.1111/j.1365-2958.2010.07219.x

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol* 14:e1002533. doi: 10.1371/journal.pbio.1002533

Shenoy, P. A., Vishwanath, S., Gawda, A., Shetty, S., Anegundi, R., Varma, M., et al. (2017). Anaerobic bacteria in clinical specimens–frequent, but a neglected lot: a five year experience at a tertiary care hospital. *J. Clin. Diagn. Res.* 11, DC44–DC48. doi: 10.7860/JCDR/2017/26009.10311

Shkoporov, A. N., Khokhlova, E. V., Fitzgerald, C. B., Stockdale, S. R., Draper, L. A., Ross, R. P., et al. (2018). Phi CrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. *Nat. Commun.* 9:4781. doi: 10.1038/s41467-018-07225-7

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* 34, D32–D36. doi: 10.1093/nar/gkj014

Stern, A., Mayrose, I., Penn, O., Shaul, S., Gophna, U., and Pupko, T. (2010). An evolutionary analysis of lateral gene transfer in thymidylate synthase enzymes. *Syst. Biol.* 59, 212–225. doi: 10.1093/sysbio/syp104

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039

Sun, Z. P., Levi, Y., Kiene, L., Dumoutier, N., and Lucena, F. (1997). Quantification of bacteriophages of *Bacteroides fragilis* in environmental water samples of Seine river. *Water Air Soil Poll.* 96, 175–183.

Tariq, M. A., Everest, F. L. C., Cowley, L. A., De Soyza, A., Holt, G. S., Bridge, S. H., et al. (2015). A metagenomic approach to characterize temperate bacteriophage populations from cystic fibrosis and non-cystic fibrosis bronchiectasis patients. *Front. Microbiol.* 6:97. doi: 10.3389/fmicb.2015.00097

Tartera, C., and Jofre, J. (1987). Bacteriophages active against *Bacteroides fragilis* in sewage-polluted waters. *Appl. Environ. Microb.* 53, 1632–1637. doi: 10.1128/Aem.53.7.1632-1637.1987

tseemann/abricate (2020). *Abricate [Online]*. Available online at: https://github.com/tseemann/abricate (accessed Jun 10, 2020).

Tzianabos, A. O., Onderdonk, A. B., Rosner, B., Cisneros, R. L., and Kasper, D. L. (1993). Structural features of polysaccharides that induce intraabdominal abscesses. *Science* 262, 416–419. doi: 10.1126/science.8211161

Wadhwa, A., Dutta, S., Ebdon, J., Chowdhary, G., Kapoor, R., Wang, Y. K., et al. (2018). Successful application of microbial source tracking using Gb-124 bacteriophage as an indicator of human fecal contamination in environmental samples in Kolkata, India. *Am. J. Trop. Med. Hyg.* 99, 639–639.

Wang, I. N., Smith, D. L., and Young, R. (2000). Holins: the protein clocks of bacteriophage infections. *Annu. Rev. Microbiol.* 54, 799–825. doi: 10.1146/annurev.micro.54.1.799

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* 13:e1005595. doi: 10.1371/journal.pcbi.1005595

Wu, S. G., Powell, J., Mathioudakis, N., Kane, S., Fernandez, E., and Sears, C. L. (2004). *Bacteroides fragilis* enterotoxin induces intestinal epithelial cell secretion of interleukin-8 through mitogen-activated protein kinases and a tyrosine kinase-regulated nuclear factor-kappa B pathway. *Infect. Immun.* 72, 5832–5839. doi: 10.1128/Iai.72.10.5832-5839.2004

Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., et al. (2003). A genomic view of the human *Bacteroides thetaiotaomicron* symbiosis. *Science* 299, 2074–2076. doi: 10.1126/science.1080029

Yarmolinsky, M. B. (1995). Programmed cell-death in bacterial populations. *Science* 267, 836–837. doi: 10.1126/science.7846528

Young, R., and Blasi, U. (1995). Holins - form and function in bacteriophage lysis. *FEMS Microbiol. Rev.* 17, 191–205. doi: 10.1016/0168-6445(94)00079-4

Young, R. Y. (1992). Bacteriophage lysis - mechanism and regulation. *Microbiol. Rev.* 56, 430–481. doi: 10.1128/Mmbr.56.3.430-481.1992

Zaneveld, J. R., Lozupone, C., Gordon, J. I., and Knight, R. (2010). Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res.* 38, 3869–3879. doi: 10.1093/nar/gkq066

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* 67, 2640–2644. doi: 10.1093/jac/dks261

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix 3

Coding regions of *Bacteroides fragilis* GB-124 plasmid PBf1

This table is reproduced from Tariq *et al.*, 2018 (Appendix 2) under terms of the Creative Commons Attribution License (CC BY) of *Frontiers in Microbiology*.

CD	Start	End	Size (aa)	Putative product	E value	Identity (aa)
1	106	1131	341	Initiator RepB protein [*Bacteroides* sp. D22]; Rep 3 (Pfam 01051)	0	341/341 (100 %)
2	1418	1870	150	MULTISPECIES: hypothetical protein [Bacteria]	1.00E-07	150/150 (100 %)
3	1827	2621	264	Relaxase/mobilization nuclease domain protein [*Bacteroides cellulosilyticus* DSM 14838]; Relaxase (Pfam 03432)	0	264/264 (100 %)
4	2618	2917	99	MULTISPECIES: hypothetical protein [Bacteria]	5.00E-63	99/99 (100 %)
5	3110	3307	65	MULTISPECIES: hypothetical protein [Bacteria]	2.00E-39	65/65 (100 %)
6	3318	3590	90	Type II toxin-antitoxin system mRNA interferase toxin, RelE/StbE family [*Bacteroides xylanisolvens];* YoeB toxin (Pfam 06769)	2.00E-58	90/90 (100 %)
7	3590	3841	83	MULTISPECIES: hypothetical protein [Bacteria]	8.00E-53	83/83 (100 %)

Appendix 4

Coding regions of *Bacteroides fragilis* GB-124 plasmid PBf2

This table is reproduced from Tariq *et al.*, 2018 (Appendix 2) under terms of the Creative Commons Attribution License (CC BY) of *Frontiers in Microbiology*.

CD	Start	End	Size (aa)	Putative product	E value	Identity (aa)
1	161	430	89	MULTISPECIES: hypothetical protein [Bacteroides]	7.00E-54	89/89 (100 %)
2	502	951	149	MULTISPECIES: DUF3791 domain-containing protein [Bacteroidales]; DUF3791 (Pfam 12668)	5.00E-104	149/149 (100 %)
3	948	1427	159	MULTISPECIES: DUF3990 domain-containing protein [Bacteroidales]; DUF3990 (Pfam 13151)	3.00E-114	159/159 (100 %)
4	1449	1709	86	MULTISPECIES: hypothetical protein [Bacteria]	2.00E-53	86/86 (100 %)
5	1696	2562	288	MULTISPECIES: AAA family ATPase [Bacteria]; Zeta toxin (Pfam 06414)	0	288/288 (100 %)
6	2609	3274	221	MULTISPECIES: M23 family metallopeptidase [Bacteroidales]; Peptidase M23 (Pfam 01551)	2.00E-163	221/221 (100 %)
7	3315	4034	239	Hypothetical protein M097_4003 [Bacteroides vulgatus str.3775 SL(B) 19 (iv)]; DUF3945 (Pfam 13101)	3.00E-169	238/239 (99 %)
8	4036	5595	519	MULTISPECIES: DUF3945 domain-containing protein [Bacteroidales]; DUF3945 (Pfam 13101)	0	519/519 (100 %)
9	5600	6709	369	MULTISPECIES: DUF3991 domain-containing protein [Bacteroidales]; Toprim-like (Pfam 13155)	0	369/369 (100 %)
10	6706	7209	167	MULTISPECIES: PH domain-containing protein [Bacteria]; YdbT (COG3428)	0	167/167 (100 %)
11	7190	7897	235	MUTLISPECIES: hypothetical protein [Bacteria]	0	235/235 (100 %)
12	7898	8734	278	MULTISPECIES: DNA topoisomerase I [Bacteroidales]; Toprim_Crpt (Pfam 13342)	0	278/278 (100 %)
13	8731	9084	117	MULTISPECIES: hypothetical protein [Bacteroidales]	6.00E-136	117/117 (100 %)
14	9096	9374	92	MULTISPECIES: HU family DNA-binding protein [Bacteroidales]; Bacterial DNA-binding protein (Pfam 00216)	3.00E-99	92/92 (100 %)
15	9365	9637	90	MULTISPECIES: hypothetical protein [Bacteroidales]	2.00E-99	90/90 (100 %)
16	9648	10091	147	MULTISPECIES: hypothetical protein [Bacteroidales]	6.00E-164	147/147 (100 %)
17	10266	10490	74	MULTISPECIES: hypothetical protein [Bacteroidales]; COG1107	3.00E-80	74/74 (100 %)
18	10494	12449	651	MULTISPECIES: type IV secretory system conjugative DNA transfer family protein [Bacteroidales]; TrwB AAD bind (Pfam 10412)	0	651/651 (100 %)
19	12442	12828	128	MUTLISPECIES: helix-turn-helix transcriptional regulator [Bacteria]; HTH 31 (Pfam 13560)	1.00E-140	128/128 (100 %)
20	13263	13691	142	Hypothetical protein BFAG_03571 [Bacteroides fragilis 3_1_12]	4.00E-163	142/142 (100 %)
21	13735	15414	559	MULTISPECIES: hypothetical protein [Bacteria]	0	559/559 (100 %)
22	15424	16746	440	MULTISPECIES: hypothetical protein [Bacteria]	0	440/440 (100 %)
23	16958	17956	332	MULTISPECIES: hypothetical protein [Bacteroidales]	0	332/332 (100 %)
24	18125	19360	410	MUTLISPECIES: aminotransferase class I/II-fold pyridoxal phosphate-dependent enzyme [Bacteroidales]; KBL like (cd06454)	0	409/410 (99 %)
25	19363	20085	240	TetR/AcrR family transcriptional regulator [Bacteroides intestinalis]; AcrR (COG1309)	0	240/240 (100 %)
26	20174	20608	144	MULTISPECIES: hypothetical protein [Bacteria]	3.00E-158	144/144 (100 %)
27	20667	21290	207	MULTISPECIES: ParA family protein [Bacteroidales]; ParAB family (cd02042)	0	207/207 (100 %)

CD	Start	End	Size (aa)	Putative product	E value	Identity (aa)
28	21307	22926	539	Putative mobilization protein [Bacteroides caccae]; Relaxase (Pfam 03432)	0	539/539 (100 %)
29	23303	23632	109	Hypothetical protein BSBG_04822 [Bacteroides sp. 9_1_42FAA]; MobC (Pfam 05713)	3.00E-121	107/109 (98 %)
30	24763	25920	385	MULTISPECIES: replication initiation protein [Bacteria]; Rep 3 (Pfam 01051)	0	385/385 (100 %)
31	26114	26293	59	Hypothetical protein [Bacteroides dorei]	9.00E-47	49/49 (100 %)
32	26283	26618	111	Hypothetical protein M082_5909 [Bacteroides fragilis str. 3725 D9 ii]	6.00E-125	111/111 (100 %)
33	26827	27060	77	Hypothetical protein [Bacteroides ovatus]	1.00E-62	62/62 (100 %)
34	27100	27348	82	Hypothetical protein [Bacteroides sp. HMSC068A09]	9.00E-82	77/78 (99 %)
35	27360	27584	74	MULTISPECIES: hypothetical protein [Bacteria]	2.00E-65	65/65 (100 %)
36	27626	27787	53	Hypothetical protein [Bacteroides eggerthii]	7.00E-43	45/45 (100 %)
37	27804	27983	59	Hypothetical protein [Parabacteroides distasonis]	4.00E-61	59/59 (100 %)
38	28030	28419	129	MULTISPECIES: hypothetical protein [Bacteria]	7.00E-147	129/129 (100 %)
39	28862	28990	42	DNA-binding protein [Campylobacter jejuni]	3.00E-13	21/21 (100 %)
40	29366	29926	186	MULTISPECIES: hypothetical protein [Bacteroidales]	0	184/186 (99 %)
41	29877	30101	74	MULTISPECIES: hypothetical protein [Bacteroidales]	2.00E-76	74/74 (100 %)
42	30098	31108	335	MULTISPECIES: DUF1738 domain-containing protein [Bacteroidales]	0	335/335 (100 %)
43	31533	32126	197	DNA invertase Pin-like site-specific DNA recombinase [Butyricimonas paravirosa]; Sertine recombinase family (cd 03768)	0	197/197 (100 %)
44	32272	32613	113	MULTISPECIES: hypothetical protein [Bacteroidales]]	1.00E-125	113/113 (100 %)
45	32633	32947	104	MULTISPECIES: DUF4134 domain-containing protein [Bacteroidales]; DUF4134 (Pfam 13572)	1.00E-111	104/104 (100 %)
46	32949	33248	99	MULTISPECIES: hypothetical protein [Bacteroidales]	2.00E-106	99/99 (100 %)
47	33254	36094	946	MULTISPECIES: hypothetical protein [Bacteroidales]; Bacteroides conjugation system ATPase TraG family (TIGR 03783)	0	946/946 (100 %)
48	36105	36779	224	MULTISPECIES: hypothetical protein [Bacteroidales]	1.00E-161	224/224 (100 %)
49	36781	37437	218	MULTISPECIES: hypothetical protein [Bacteria]	3.00E-158	218/218 (100 %)
50	37430	38453	340	MULTISPECIES: plasmid transfer protein [Bacteroidales]; Bacteroides conjugative transposon TraJ protein (TIGR 03782)	0	340/340 (100 %)
51	38485	39099	204	MULTISPECIES: conjugative transposon protein TraK [Bacteroidales]; Bacteroides conjugative transposon TraK protein (TIGR 03781)	1.00E-147	204/204 (100 %)
52	39099	39530	143	MULTISPECIES: hypothetical protein [Bacteroidales]	5.00E-99	143/143 (100 %)
53	39534	40643	369	MULTISPECIES: conjugative transposon protein TraM [Bacteroidales]; Bacteroides conjugative transposon TraM protein (TIGR 03779)	0	369/369 (100 %)
54	40645	41484	279	MULTISPECIES: conjugative transposon protein TraN [Bacteroidales]; Bacteroides conjugative transposon TraN protein (TIGR 037800	0	279/279 (100 %)
55	41497	42000	167	MULTISPECIES: hypothetical protein [Bacteroidales]	2.00E-120	167/167 (100 %)
56	42003	42662	219	MULTISPECIES: hypothetical protein [Bacteroidales]	8.00E-158	219/219 (100 %)
57	42665	43306	213	MULTISPECIES: OmpA family protein [Bacteroidales]; OmpA family (Pfam 00691)	7.00E-152	213/213 (100 %)

Appendix 5

Top blastx hit of 'missing' genes within the 5 outlying clusters

Cluster	Strain and gene location	Gene name	UniProtKB accession	Locus	Predicted product	Species	E value	Identity
1	1007-1-F#3_02999	group_7966	P94519	YSDA_BACSU	Uncharacterized protein YsdA	Bacillus subtilis (strain 168)	3.00E-13	33/66 (50 %)
	1007-1-F#7_03731	dnaJ	Q5LED4	DNAJ_BACFN	Chaperone protein DnaJ	Bacteroides fragilis (strain ATCC 25285/ DSM 2151/ JCM 11019/ NCTC 9343)	0.00E+00	394/394 (100 %)
	20656-2-1_02655	group_5173	Q64T27	PYRG_BACFR	CTP synthase	Bacteroides fragilis (strain YCH46)	0.00E+00	532/533 (99 %)
	1007-1-F#3_04196	mutL	Q64NX1	MUTL_BACFR	DNA mismatch repair protein MutL	Bacteroides fragilis (strain YCH46)	0.00E+00	624/625 (99 %)
	1007-1-F#3_04408	pnp	Q64N73	PNP_BACFR	Polyribonucleotide nucleotidyltransferase	Bacteroides fragilis (strain YCH46)	0.00E+00	708/708 (100 %)
	1007-1-F#3_02558	dapF	Q64SY7	DAPF_BACFR	Diaminopimelate epimerase	Bacteroides fragilis (strain YCH46)	0.00E+00	267/269 (99 %)
	1007-1-F#3_00723	group_10884	P08696	BCN5_CLOPF	Bacteriocin BCN5	Clostridium perfringens	2.00E-02	30/55(55 %)
	1007-1-F#3_02563	yqhD	Q46856	YQHD_ECOLI	Alchohol dehydrogenase YqhD	Escherichia coli (strain K12)	4.00E-129	193/386 (50 %)
	1007-1-F#3_02554	asnB	P22106	ASNB_ECOLI	Asparagine synthetase B [glutamine-hydrolysing]	Escherichia coli (strain K12)	0.00E+00	357/558 (64 %)
	1007-1-F#3_01790	nqrE	A5UFX2	NQRE_HAEIG	Na(+)-translocating NADH-quinone reductase subunit E	Haemophilus influenzae (strain PittGG)	2.00E-86	130/208 (63 %)
	1007-1-F#3_01388	group_4904	E1V931	DHA_HALED	Alanine dehydrogenase	Halomonas elongata (strain ATCC 33173/ DSM 2581/ NBRC 15536/ NCIMB 2198/ 1H9)	4.00E-129	207/3636 (57 %)
	1007-1-F#3_03556	rffH_3	P55255	RMLA_NEIMB	Glucose-1-phophase thymidylyltransferase	Neisseria meningitidis serogroup B (strain MC58)	7.00E-147	192/291 (66 %)
	1007-1-F#7_01584	rffH_1	P55255	RMLA_NEIMB	Glucose-1-phosphate thymidylytransferase	Neisseria meningitidis serogroup B (strain MC58)	1.00E-143	191/289 (66 %)
	3397N2_01535	capD_1	A8GRN9	CAPD_RICRS	UDP-glucose 4-epimerase	Rickettsia rickettsii (strain Sheila Smith)	1.00E-168	221/339 (65 %)
	20656-2-1_03572	rfbE	Q8Z5I4	RFBF_SALTI	Glucose-1-phosphate cytidylytransferase	Salmonella typhi	3.00E-102	147/259 (57 %)

Cluster	Strain and gene location	Gene name	UniProtKB accession	Locus	Predicted product	Species	E value	Identity
	20656-2-1_03569	rfbE	P14169	RFBE_SALTI	CDP-paratose 2-epimerase	Salmonella typhi	5.00E-141	209/337 (62 %)
	20656-2-1_03571	rfbG_2	P26397	RFBG_SALTY	CDP-glucose 4,6-dehydratase	Salmonella typhimurium (strain LT2/ SGSC1412/ ATCC 700720)	1.00E-129	178/345 (52 %)
	1009-4-F#7_03292	spnQ	P26398	RFBH_SALTY	Lipopolysaccharide biosynthesis protein RfbH	Salmonella typhimurium (strain LT2/ SGSC1412/ ATCC 700720)	5.00E-168	234/419 (56 %)
	1007-1-F#3_03555	rfbC_4	P26394	RMLC_SALTY	dTDP-4-dehydrorhamnose 3,5-epimerase	Salmonella typhimurium (strain LT2/ SGSC1412/ATCC 700720)	3.00E-64	103/179 (58 %)
	1007-1-F#7_00207	group_3316	P22036	ATMB_SALTY	Magnesium-transporting ATPase, P-type 1	Salmonella typhimurium strain (strain Lt2/ SGSC1412/ ATCC 700720)	0.00E+00	480/902 (53 %)
	1007-1-F#3_01267	ribB	B1KNY2	RIBB_SHEWAM	3,4-dihydroxy-2-butanone 4-phosphate synthase	Shewanella woodyi (strain ATCC 51908/ MS32)	1.00E-94	133/205 (65 %)
	1007-1-F#7_01585	group_3517	P37780	RMLC_SHIFL	dDTP-4-dehydrohamnose 3,5-epimerase	Shigella flexneri	5.00E-63	101/186 (54 %)
2	1007-1-F#10_02412	rhaS_5	O34901	YOBQ_BACSU	Uncharacterized HTH-type transcriptional regulator YobQ	Bacillus subtilis (strain 168)	2.00E-08	28/63 (44 %)
	1007-1-F#10_01834	patB_2	Q08432	CBL_BACSU	Cystathionine beta-lyase PatB	Bacillus subtilis (strain 168)	1.00E-39	67/144 (47 %)
	1007-1-F#10_01831	yvgN	O32210	GR_BACSU	Glyoxal reductase	Bacillus subtilis (strain 168)	6.00E-37	76/171 (44 %)
	1007-1-F#10_03731	dnaJ	Q5LED4	DNAJ_BACFN	Chaperone protein DnaJ	Bacteroides fragilis (strain ATCC 25285 / DSM 2151 / JCM 11019 / NCTC 9343)	0.00E+00	394/394 (100 %)
	1007-1-F#10_02377	hisH	Q64RT0	HIS5_BACFR	Imidazole glycerol phosphate synthase subunit HisH	Bacteroides fragilis (strain YCH46)	1.00E-145	195/196 (99 %)
	1007-1-F#10_01955	pyrB	Q64U74	PYRB_BACFR	Aspartate carbamoyltransferase	Bacteroides fragilis (strain YCH46)	0.00E+00	308/308 (100 %)
	1007-1-F#3_04241	tufA	P33165	EFTU_BACFR	Elongation factor	Bacteroides fragilis (strain YCH46)	0.00E+00	394/394 (100 %)
	2-F-2#4_03751	group_3938	Q64NX1	MUTL_BACFR	DNA mismatch repair protein MutL	Bacteroides fragilis (strain YCH46)	0	624/625 (99 %)
	1007-1-F#10_00116	group_24135	Q650K9	YIDD_BACFR	Putative membrane protein insertion efficiency factor	Bacteroides fragilis (strain YCH46)	8.00E-49	73/73 (100 %)
	1007-1-F#10_04249	btuD_3	Q5WNX0	BCRA_ENTFL	Bacitracin transport ATP-binding protein BcrA	Enterococcus faecalis	7.00E-51	86/216 (40 %)

Cluster	Strain and gene location	Gene name	UniProtKB accession	Locus	Predicted product	Species	E value	Identity
	1007-1-F#3_03865	mdtB_1	Q48815	HELA_LEGPN	Protein HelA	Legionella pneumophila	0.00E+00	458/1021 (45 %)
	1007-1-F#10_01830	group_3542	A0QV10	Y2408_MYCS2	Uncharacterized oxidoreductase MSMEG_2408/MSMEI_2347	Mycolicibacterium smegmatis (strain ATCC 700084 / mc(2)155)	8.00E-82	123/261 (47 %)
	1007-1-F#3_03556	rffH_3	P55255	RMLA_NEIMB	Glucose-1-phosphate thymidylyltransferase	Neisseria meningitidis serogroup B (strain MC58)	7.00E-147	192/291 (66 %)
	1007-1-F#10_00115	rnpA	B2RHI3	RNPA_PORG3	Ribonuclease P protein component	Porphyromonas gingivalis (strain ATCC 33277 / DSM 20709 / CIP 103683 / JCM 12257 / NCTC 11834 / 2561)	3.00E-27	57/119 (48 %)
	1007-1-F#10_04364	group_1052	D5EV35	AXEA1_PRER2	Acetylxylan esterase	Prevotella ruminicola (strain ATCC 19189 / JCM 8958 / 23)	1.00E-50	97/206 (47 %)
	20656-2-1_03569	rfbE	P14169	RFBE_SALTI	CDP-paratose 2-epimerase	Salmonella typhi	5.00E-141	209/337 (62 %)
	1007-1-F#3_03555	rfbC_4	P26394	RMLC_SALTY	dTDP-4-dehydrorhamnose 3,5-epimerase	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	3.00E-64	103/179 (58 %)
	20656-2-1_03571	rfbG_2	P26397	RFBG_SALTY	CDP-glucose 4,6-dehydratase	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	1.00E-129	178/345 (52 %)
	1007-1-F#10_00224	arnC_1	A0A0H2UR96	GLYG_STRPN	Glycosyltransferase GlyG	Streptococcus pneumoniae serotype 4 (strain ATCC BAA-334 / TIGR4)	1.00E-46	92/225 (41 %)
	1009-4-F#10_00267	sufC	Q55791	Y075_SYNY3	Probable ATP-dependent transporter slr0075	Synechocystis sp. (strain PCC 6803 / Kazusa)	3.00E-108	152/248 (61 %)
	1007-1-F#10_03303	group_10498	Q9WZY4	METY_THEMA	O-acetyl-L-homoserine sulfhydrylase	Thermotoga maritima (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)	7.00E-142	223/426 (52 %)
3	1007-1-F#10_03773	group_1691	P71052	EPSC_BACSU	Probable polysaccharide biosynthesis protein EpsC	Bacillus subtilis (strain 168)	5.00E-112	188/458 (41 %)
	1007-1-F#10_01834	patB_2	Q08432	CBL_BACSU	Cystathionine beta-lyase PatB	Bacillus subtilis (strain 168)	1.00E-39	67/144 (47 %)
	1007-1-F#10_03712	yhgF_1	O31489	YDCI_BACSU	Uncharacterized protein YdcI	Bacillus subtilis (strain 168)	4.00E-59	99/196 (51 %)

Cluster	Strain and gene location	Gene name	UniProtKB accession	Locus	Predicted product	Species	E value	Identity
	1007-1-F#3_04345	group_24443	Q5L7W4	GCSH_BACFN	Glycine cleavage system H protein	Bacteroides fragilis (strain ATCC 25285 / DSM 2151 / JCM 11019 / NCTC 9343)	4.00E-77	126/126 (100 %)
	1007-1-F#10_00972	gapA	Q59199	G3P_BACFR	Glyceraldehyde-3-phosphate dehydrogenase	Bacteroides fragilis (strain YCH46)	0.00E+00	333/333 (100 %)
	1007-1-F#3_04347	group_24261	Q64N34	ISPG_BACFR	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (flavodoxin)	Bacteroides fragilis (strain YCH46)	0.00E+00	623/626 (99 %)
	2-F-2#4_03751	group_3938	Q64NX1	MUTL_BACFR	DNA mismatch repair protein MutL	Bacteroides fragilis (strain YCH46)	0	624/625 (99 %)
	1007-1-F#10_02377	hisH	Q64RT0	HIS5_BACFR	Imidazole glycerol phosphate synthase subunit HisH	Bacteroides fragilis (strain YCH46)	1.00E-45	195/196 (99 %)
	1007-1-F#10_04408	pnp	Q64N73	PNP_BACFR	Polyribonucleotide nucleotidyltransferase	Bacteroides fragilis (strain YCH46)	0.00E+00	708/708 (100 %)
	1007-1-F#10_00630	ribBA	Q64YT3	RIBBA_BACFR	Riboflavin biosynthesis protein RibBA	Bacteroides fragilis (strain YCH46)	0.00E+00	404/404 (100 %)
	1007-1-F#10_00030	group_6783	A7LXW1	FIMB_BACO1	Putative fimbrium anchoring subunit Fim4B	Bacteroides ovatus (strain ATCC 8483 / DSM 1896 / JCM 5824 / NCTC 11153)	2.00E-87	134/252 (53 %)
	20656-2-1_04230	group_23978	Q8A4P5	BUK_BACTN	Probable butyrate kinase	Bacteroides thetaiotaomicron (strain ATCC 29148 / DSM 2079 / NCTC 10582 / E50/ VPI-5482)	0	298/352 (85 %)
	1007-1-F#10_04548	hup_3	P0A3H0	DBH_GEOSE	DNA-binding protein HU	Geobacillus stearothermophilus	9.00E-22	41/82 (50 %)
	1009-4-F#10_02791	group_4967	A5UFX2	NQRE_HAEIG	Na(+)-translocating NADH-quinone reductase subunit E	Haemophilus influenzae (strain PittGG)	2.00E-86	130/208 (63 %)
	20656-2-1_03569	rfbE	P14169	RFBE_SALTI	CDP-paratose 2-epimerase	Salmonella typhi	5.00E-141	209/337 (62 %)
	1007-1-F#10_00207	group_3316	P22036	ATMB_SALTY	Magnesium-transporting ATPase, P-type 1	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	0.00E+00	480/902 (53 %)
	20656-2-1_03571	rfbG_2	P26397	RFBG_SALTY	CDP-glucose 4,6-dehydratase	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	1.00E-129	178/345 (52 %)
	1009-4-F#10_00267	ufC	Q55791	Y075_SYNY3	Probable ATP-dependent transporter slr0075	Synechocystis sp. (strain PCC 6803 / Kazusa)	3.00E-108	152/248 (61 %)

Cluster	Strain and gene location	Gene name	UniProtKB accession	Locus	Predicted product	Species	E value	Identity
	1007-1-F#3_04346	group_23642	Q9WYS7	PURE_THEMA	N5-carboxyaminoimidazole ribonucleotide mutase	Thermotoga maritima (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)	1.00E-55	88/167 (53 %)
	1007-1-F#10_01272	rluA	Q8ZIK1	RLUA_YERPE	Dual-specificity RNA pseudouridine synthase RluA	Yersinia pestis	9.00E-46	92/210 (44 %)
4	1007-1-F#10_03273	group_10485	P40761	YUXK_BACSU	Uncharacterized protein YuxK	Bacillus subtilis (strain 168)	2.00E-32	51/118 (43 %)
	1007-1-F#10_03773	group_1691	P71052	EPSC_BACSU	Probable polysaccharide biosynthesis protein EpsC	Bacillus subtilis (strain 168)	5.00E-112	188/458 (41 %)
	1007-1-F#10_03252	group_2802	P37515	MAA_BACSU	Probable maltose O-acetyltransferase	Bacillus subtilis (strain 168)	2.00E-46	81/190 (43 %)
	1007-1-F#10_01834	patB_2	Q08432	CBL_BACSU	Cystathionine beta-lyase PatB	Bacillus subtilis (strain 168)	1.00E-39	67/144 (47 %)
	1007-1-F#10_03275	group_321	Q5LA59	HDHA_BACFN	7alpha-hydroxysteroid dehydrogenase	Bacteroides fragilis (strain ATCC 25285 / DSM 2151 / JCM 11019 / NCTC 9343)	7.00E-177	256/259 (99 %)
	1007-1-F#10_03267	ung_1	Q5LA67	UNG1_BACFN	Uracil-DNA glycosylase 1	Bacteroides fragilis (strain ATCC 25285 / DSM 2151 / JCM 11019 / NCTC 9343)	1.00E-166	220/220 (100 %)
	2-F-2#4_03751	group_3938	Q64NX1	MUTL_BACFR	DNA mismatch repair protein MutL	Bacteroides fragilis (strain YCH46)	0	624/625 (99 %)
	1007-1-F#10_04408	pnp	Q64N73	PNP_BACFR	Polyribonucleotide nucleotidyltransferase	Bacteroides fragilis (strain YCH46)	0.00E+00	708/708 (100 %)
	1007-1-F#10_00030	group_6783	A7LXW1	FIMB_BACO1	Putative fimbrium anchoring subunit Fim4B	Bacteroides ovatus (strain ATCC 8483 / DSM 1896 / JCM 5824 / NCTC 11153)	2.00E-87	134/252 (53 %)
	20656-2-1_04230	group_23978	Q8A4P5	BUK_BACTN	Probable butyrate kinase	Bacteroides thetaiotaomicron (strain ATCC 29148 / DSM 2079 / NCTC 10582 / E50 / VPI-5482)	0	298/352 (85 %)
	1007-1-F#10_01890	ravA_2	B1LL73	RAVA_ECOSM	ATPase RavA	Escherichia coli (strain SMS-3-5 / SECEC)	2.00E-95	145/296 (49 %)
	1007-1-F#10_04548	hup_3	P0A3H0	DBH_GEOSE	DNA-binding protein HU	Geobacillus stearothermophilus	9.00E-22	41/82 (50 %)
	1007-1-F#10_01528	ald	E1V931	DHA_HALED	Alanine dehydrogenase	Halomonas elongata (strain ATCC 33173 / DSM 2581 / NBRC 15536 / NCIMB 2198 / 1H9)	5.00E-128	206/363 (57 %)

Cluster	Strain and gene location	Gene name	UniProtKB accession	Locus	Predicted product	Species	E value	Identity
	1007-1-F#10_03272	group_10481	P9WNP3	HTDZ_MYCTU	3-hydroxyacyl-thioester dehydratase Z	Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)	1.00E-36	60/145 (41 %)
	20656-2-1_03569	rfbE	P14169	RFBE_SALTI	CDP-paratose 2-epimerase	Salmonella typhi	5.00E-141	209/337 (62 %)
	1007-1-F#10_00207	group_3316	P22036	ATMB_SALTY	Magnesium-transporting ATPase, P-type 1	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	0.00E+00	480/902 (53 %)
	20656-2-1_03571	rfbG_2	P26397	RFBG_SALTY	CDP-glucose 4,6-dehydratase	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	1.00E-129	178/345 (52 %)
	1007-1-F#10_01267	ribB	B1KNY2	RIBB_SHEWM	3,4-dihydroxy-2-butanone 4-phosphate synthase	Shewanella woodyi (strain ATCC 51908 / MS32)	1.00E-94	133/205 (65 %)
	1009-4-F#10_00267	sufC	Q55791	Y075_SYNY3	Probable ATP-dependent transporter slr0075	Synechocystis sp. (strain PCC 6803 / Kazusa)	3.00E-108	152/248 (61 %)
	1007-1-F#10_01272	rluA	Q8ZIK1	RLUA_YERPE	Dual-specificity RNA pseudouridine synthase RluA	Yersinia pestis	9.00E-46	92/210 (44 %)
5	1007-1-F#10_04505	mro_3	P05149	GALM_ACICA	Aldose 1-epimerase	Acinetobacter calcoaceticus	7.00E-100	158/365 (43 %)
	1007-1-F#10_04054	dbpA_2	P50729	RECS_BACSU	Probable ATP-dependent DNA helicase RecS	Bacillus subtilis (strain 168)	5.00E-81	145/335 (43 %)
	1007-1-F#3_02999	group_7966	P94519	YSDA_BACSU	Uncharacterized protein YsdA	Bacillus subtilis (strain 168)	3.00E-13	33/66 (50 %)
	1007-1-F#10_01834	patB_2	Q08432	CBL_BACSU	Cystathionine beta-lyase PatB	Bacillus subtilis (strain 168)	1.00E-39	67/144 (47 %)
	1007-1-F#10_03731	dnaJ	Q5LED4	DNAJ_BACFN	Chaperone protein DnaJ	Bacteroides fragilis (strain ATCC 25285 / DSM 2151 / JCM 11019 / NCTC 9343)	0.00E+00	394/394 (100 %)
	1007-1-F#10_02558	dapF	Q64SY7	DAPF_BACFR	Diaminopimelate epimerase	Bacteroides fragilis (strain YCH46)	0.00E+00	267/269 (99 %)
	1007-1-F#10_04057	group_2916	P31206	NANH_BACFR	Sialidase	Bacteroides fragilis (strain YCH46)	3.00E-109	179/444 (40 %)
	20656-2-1_02655	group_5173	Q64T27	PYRG_BACFR	CTP synthase	Bacteroides fragilis (strain YCH46)	0.00E+00	532/533 (99 %)
	1007-1-F#10_04196	mutL	Q64NX1	MUTL_BACFR	DNA mismatch repair protein MutL	Bacteroides fragilis (strain YCH46)	0.00E+00	624/625 (99 %)
	1007-1-F#10_04408	pnp	Q64N73	PNP_BACFR	Polyribonucleotide nucleotidyltransferase	Bacteroides fragilis (strain YCH46)	0.00E+00	708/708 (100 %)
	1007-1-F#10_02803	group_7349	Q8A1G1	SUSC_BACTN	TonB-dependent receptor SusC	Bacteroides thetaiotaomicron (strain ATCC 29148 / DSM 2079 / NCTC 10582 / E50 / VPI-5482)	7.00E-05	29/72 (40 %)

Cluster	Strain and gene location	Gene name	UniProtKB accession	Locus	Predicted product	Species	E value	Identity
	1007-1-F#10_01787	nqrB	Q1QX85	NQRB_CHRSD	Na(+)-translocating NADH-quinone reductase subunit B	Chromohalobacter salexigens (strain ATCC BAA-138 / DSM 3043 / CIP 106854 / NCIMB 13768 / 1H11)	1.00E-108	190/406 (47 %)
	1007-1-F#10_02392	metH_2	Q24SP8	MTGC_DESHY	Corrinoid protein DSY3155	Desulfitobacterium hafniense (strain Y51)	3.00E-47	85/213 (40 %)
	1007-1-F#3_00316	msbA	Q6AJW3	MSBA_DESPS	ATP-dependent lipid A-core flippase	Desulfotalea psychrophila (strain LSv54 / DSM 12343)	7.00E-130	207/504 (41 %)
	1007-1-F#10_02554	asnB	P22106	ASNB_ECOLI	Asparagine synthetase B [glutamine-hydrolysing]	Escherichia coli (strain K12)	0.00E+00	357/558 (64 %)
	1007-1-F#3_01424	atoC_2	Q06065	ATOC_ECOLI	Regulatory protein AtoC	Escherichia coli (strain K12)	4.00E-109	182/454 (40 %)
	1007-1-F#10_04416	group_2225	P33363	BGLX_ECOLI	Periplasmic beta-glucosidase	Escherichia coli (strain K12)	0.00E+00	351/745 (47 %)
	1007-1-F#10_01836	group_873	P25906	PDXI_ECOLI	Pyridoxine 4-dehydrogenase	Escherichia coli (strain K12)	1.00E-06	21/43 (49 %)
	1007-1-F#10_02563	yqhD	Q46856	YQHD_ECOLI	Alcohol dehydrogenase YqhD	Escherichia coli (strain K12)	4.00E-129	193/386 (50 %)
	1007-1-F#3_00672	apgM	Q74C57	APGM_GEOSL	Probable 2,3-bisphosphoglycerate-independent phosphoglycerate mutase	Geobacter sulfurreducens (strain ATCC 51573 / DSM 12127 / PCA)	6.00E-112	194/407 (48 %)
	1007-1-F#10_03675	yhgF	P71353	Y568_HAEIN	Uncharacterized protein HI_0568	Haemophilus influenzae (strain ATCC 51907 / DSM 11121 / KW20 / Rd)	0.00E+00	347/717 (48 %)
	1007-1-F#10_01790	nqrE	A5UFX2	NQRE_HAEIG	Na(+)-translocating NADH-quinone reductase subunit E	Haemophilus influenzae (strain PittGG)	2.00E-86	130/208 (63 %)
	1007-1-F#10_02569	pabA	P06194	PABA_KLEAE	Aminodeoxychorismate synthase component 2	Klebsiella aerogenes	2.00E-57	89/187 (48 %)
	1007-1-F#10_01584	rffH_1	P55255	RMLA_NEIMB	Glucose-1-phosphate thymidylyltransferase	Neisseria meningitidis serogroup B (strain MC58)	1.00E-143	191/289 (66 %)
	1007-1-F#3_03556	rffH_3	P55255	RMLA_NEIMB	Glucose-1-phosphate thymidylyltransferase	Neisseria meningitidis serogroup B (strain MC58)	7.00E-147	192/291 (66 %)
	3397N2_01536	wbjC	Q9XC60	WBJC_PSEA1	UDP-2-acetamido-2,6-beta-L-arabino-hexul-4-ose reductase	Pseudomonas aeruginosa (strain ATCC 29260 / BCRC 12902 / CIP 102967 / NCIMB 11965 / PA103)	2.00E-110	165/379 (44 %)

Cluster	Strain and gene location	Gene name	UniProtKB accession	Locus	Predicted product	Species	E value	Identity
	1007-1-F#10_03874	yknY_4	Q92NU9	MACB_RHIME	Macrolide export ATP-binding/permease protein MacB	Rhizobium meliloti (strain 1021)	4.00E-70	108/221 (49 %)
	1007-1-F#10_02960	nth	O05956	END3_RICPR	Endonuclease III	Rickettsia prowazekii (strain Madrid E)	6.00E-62	91/205 (44 %)
	3397N2_01535	capD_1	A8GRN9	CAPD_RICRS	UDP-glucose 4-epimerase	Rickettsia rickettsii (strain Sheila Smith)	1.00E-168	221/339 (65 %)
	20656-2-1_03572	rfbF	Q8Z5I4	RFBF_SALTI	Glucose-1-phosphate cytidylyltransferase	Salmonella typhi	3.00E-102	147/259 (57 %)
	20656-2-1_03569	rfbE	P14169	RFBE_SALTI	CDP-paratose 2-epimerase	Salmonella typhi	5.00E-141	209/337 (62 %)
	1007-1-F#10_00207	group_3316	P22036	ATMB_SALTY	Magnesium-transporting ATPase, P-type 1	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	0.00E+00	480/902 (53 %)
	1007-1-F#3_03555	rfbC_4	P26394	RMLC_SALTY	dTDP-4-dehydrorhamnose 3,5-epimerase	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	3.00E-64	103/179 (58 %)
	20656-2-1_03571	rfbG_2	P26397	RFBG_SALTY	CDP-glucose 4,6-dehydratase	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	1.00E-129	178/345 (52 %)
	1009-4-F#10_03292	spnQ	P26398	RFBH_SALTY	Lipopolysaccharide biosynthesis protein RfbH	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	5.00E-168	234/419 (56 %)
	1007-1-F#10_01267	ribB	B1KNY2	RIBB_SHEWM	3,4-dihydroxy-2-butanone 4-phosphate synthase	Shewanella woodyi (strain ATCC 51908 / MS32)	1.00E-94	133/205 (65 %)
	1007-1-F#3_00490	group_10882	P0A9L4	FKBB_SHIFL	FKBP-type 22 kDa peptidyl-prolyl cis-trans isomerase	Shigella flexneri	3.00E-16	46/111 (41 %)
	1007-1-F#10_01585	group_3517	P37780	RMLC_SHIFL	dTDP-4-dehydrorhamnose 3,5-epimerase	Shigella flexneri	5.00E-63	101/186 (54 %)
	1007-1-F#10_01773	glyD	Q9AEU2	GLY_STRGN	Probable glycosyl transferase Gly	Streptococcus gordonii	8.00E-57	109/268 (41 %)
	1007-1-F#3_02004	bioF_2	B0K590	BIOF_THEPX	8-amino-7-oxononanoate synthase	Thermoanaerobacter sp. (strain X514)	2.00E-87	145/356 (41 %)
	1007-1-F#10_01276	bacC	Q56318	Y019_THEMA	Uncharacterized oxidoreductase TM_0019	Thermotoga maritima (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)	5.00E-48	95/238 (40 %)
	1007-1-F#10_02561	ppk_1	Q87S51	PPK1_VIBPA	Polyphosphate kinase	Vibrio parahaemolyticus serotype O3:K6 (strain RIMD 2210633)	0.00E+00	273/663 (41 %)
	1007-1-F#10_01272	rluA	Q8ZIK1	RLUA_YERPE	Dual-specificity RNA pseudouridine synthase RluA	Yersinia pestis	9.00E-46	92/210 (44 %)

Appendix 6

Top *rfb* gene hits for each gene type and KEGG ID

This table shows the top blastp hit for each *rfb* gene across all isolates and relates to *rfb* heatmap (Figure 4.12). The KEGG ID relates to KEGG ID found in Table 4.2.

rfb gene	**Gene type**	**KEGG ID**
rfbA	rfbA	VU15_04710
	rfbA_group_4176	VU15_16360
	rfbA_1	VU15_03380
rfbB	rfbB	VU15_03390
	rfbB_group_16883	VU15_03390
	rfbB_group_21357	VU15_03390
	rfbB_1	VU15_03390
	rfbB_3	VU15_03390
	rfbB_3_group_22647	VU15_03390
rfbC	rfbC_1	BF638R_1545
	rfbC_1_group_14506	BF638R_3473
	rfbC_1_group_10942	VU15_03385
	rfbC_1_group_8887	VU15_03385
	rfbC_1_group_10950	VU15_16355
	rfbC_1_group_8372	VU15_16355
	rfbC_1_group_21355	BF638R_1545
	rfbC_1_group_19864	VU15_16355
	rfbC_2	VU15_09970
	rfbC_2_group_3517	VU15_03385
	rfbC_2_group_10570	BF638R_3473
	rfbC_2_group_15503	BF638R_3473
	rfbC_2_group_18919	VU15_16355
	rfbC_3	BF638R_3473
	rfbC_3_group_8956	VU15_16355
	rfbC_3_group_8957	BF638R_3473
	rfbC_3_group_6363	BF638R_3473
	rfbC_4	VU15_16355
	rfbC_6	BF638R_3473
rfbE	rfbE	VU15_16410
	rfbE_1	VU15_06465
	rfbE_2	BF9343_2519
rfbF	rfbF	VU15_16425
	rfbF_group_12158	VU15_16425
	rfbF_1	BF638R_0779
	rfbF_1_group_14582	BF1534
	rfbF_1_group_8584	VU15_11525
	rfbF_1_group_13741	VU15_11525
	rfbF_1_group_16290	VU15_06445

rfb gene	**Gene type**	**KEGG ID**
	rfbF_1_group_14999	BF638R_0779
	rfbF_2	VU15_11525
	rfbF_2_group_12160	VU15_16425
	rfbF_2_group_11061	VU15_16425
rfbG	rfbG	BF638R_2596
	rfbG_1	BF638R_0780
	rfbG_1_group_14584	BF1536
	rfbG_1_group_12933	BF638R_0780
	rfbG_1_group_13740	VU15_16420
	rfbG_1_group_12460	VU15_11520
	rfbG_1_group_16292	VU15_06455
	rfbG_1_group_15000	VU15_16420
	rfbG_2	VU15_16420
	rfbG_2_group_14341	VU15_11520
	rfbG_2_group_11059	BF638R_0780
rfbJ	rfbJ	STM2089
	rfbJ_group_17611	STM2089
rfbM	rfbM	BF9343_4017
rfbX	rfbX	B2037
	rfbX_2	B2037
	rfbX_3	B2037
	rfbX_group_13002	B2037
	rfbX_group_1628	B2037
	rfbX_group_1633	B2037
	rfbX_group_16359	B2037
	rfbX_group_4292	B2037

Appendix 7

Top alternative *rfb* gene hits for each gene type and KEGG ID

This table shows the *rfb* genes that had multiple KEGG IDs for top blastp hits. It also shows the isolates that had an alternative top hit and the gene location. It relates to the *rfb* heatmap (Figure 4.12) and KEGG ID to those found in Table 4.2.

***rfb* gene**	**KEGG ID**	**Isolate and gene location**	**Identity (%)**
rfbE	BF638R_3482	DS-71_03753	98
		GUT04_02820	98
		HAP130N_1B_04417	98
		HAP130N_3B_04417	98
		J38-1_02424	98
		S04_NC_02762	98
		S08_NC_01817	99
		S11_NC_00910	98
		S12_NC_00769	98
		S23L17_03356	100
		S23R14_03138	100
		S24L26_03347	98
		S24L34_03415	100
		S36L12_03442	98
		S36L5_02212	98
		TL139C_2B_03183	98
		3397N2_03388	98
		3774T13_03702	99
		3976T8_03377	98
		3986N22_03494	98
		885_BFRA_01743	98
		AD126T_1B_02663	98
		AD126T_2B_03187	99
		638R_03448	100
		AD135F_3B_03480	99
		BFR_KZ01_02897	100
		DCMOUCH0042B_04130	99
rfbF_2	BF9343_2522	3397N3_02656	100
		3719T6_02504	100
		3986NB19_02018	100
		BF8_02559	100
		BFR_KZ01_04093	100
		GB124_02918	100
		NCTC9343_02571	100
		S05_NC_04194	100
		S07_NC_01728	100
		S13L11_02187	100

rfb gene	KEGG ID	Isolate and gene location	Identity (%)
		S14_01816	100
		S23L17_02358	100
		S23R14_02288	100
		S24L34_02468	100
rfbG_2	BF638R_3484	3719T6_03493	99
		3986NB19_02859	99
		638R_03450	100
		BE1_03320	99
		BFR_KZ01_02899	100
		S03_NC_01943	99
rfbC_4	BF638R_3473	1007-1-F#3_03555	100
		1007-1-F_7_04015	100
		1009-4-F_10_02520	100
		20656-2-1_03558	99
		20793-3_01313	99
		20793-3_01313	99
		320_BFRA_04385	98
		322_BFRA_00205	98
		3397N2_03377	100
		3719A10_03576	100
		3719T6_03482	100
		3774T13_03699	100
		3783N1-6_03531	100
		3976T8_03366	100
		3986N3_03346	100
		3986NB19_02853	100
		3986NB22_03483	100
		3986TB13_03375	100
		3988TB14_03587	100
		3996NB6_03942	99
		4g8B_00703	100
		638R_03439	100
		A7UDC12-2_03370	100
		AD126T_1B_02652	100
		AD126T_2B_03178	100
		AD135F_2B_02912	100
		AD135F_3B_03471	100
		BE1_03309	100
		BFR_KZ01_02888	100
		CF01-8_02233	99
		DCMOUH0042B_04120	100
		DS-166_03581	100
		DS-208_03119	100
		DS-71_03742	100
		GUT04_02809	100
		HAP130N_1B_04406	100

rfb gene	KEGG ID	Isolate and gene location	Identity (%)
		HAP130N_2B_03108	100
		HAP130N_3B_04406	100
		HCK-B3_00360	100
		Korea419_03739	100
		NCTC9343_03439	100
		S03_NC_01932	100
		S06_NC_01739	100
		S07_NC_00291	100
		S08_NC_01826	100
		S08_NC_01826	100
		S11_NC_00899	100
		S12_NC_00758	100
		S24L15_03328	100
		S24L26_03336	100
		S36L12_03431	100
		S36L5_03169	100
		S6L5_03498	100
		S6R6_03660	100
		S6R8_03544	100
		TL139C_2B_03194	100

Appendix 8

Top alternative *rfb* gene hits for each gene type and KEGG ID

This table shows the *rfb* genes that had multiple KEGG IDs for top blastp hits. It also shows the isolates that had an alternative top hit and the gene location. It relates to the *rfb* heatmap (Figure 4.12) and KEGG ID to those found in Table 4.2.

Isolate and gene location	Length (aa)	Accession	Predicted product	Species	Identity (%)	E value
1007-1-F#10_01987	1433	WP_032533192.1	Phage tail tape measure protein	*B. fragilis*	100	0
1007-1-F#5_02130	1433	WP_032533192.1	Phage tail tape measure protein	*B. fragilis*	100	0