HIGH-THROUGHPUT ANALYSIS OF THE INTESTINAL MICROBIOTA IN *CLOSTRIDIUM DIFFICILE*-ASSOCIATED DISEASE

ADAM LEWIS BERG

A thesis presented to the University of Leicester for the degree of Doctor of Philosophy

November 2012

For Mel, Will, and Hannah

Hippocrates: *"...death sits in the bowels..."*

Arabic proverb: *"The enemy of my enemy is my friend."*

ABSTRACT

The nosocomial pathogen *Clostridium difficile* is normally unable to thrive in the human gut due to colonisation resistance. The presence of the normal intestinal bacteria prevents its proliferation through competition for nutrients, or via other mechanisms as yet unknown. Disruption of the standard flora of an individual as a result of antibiotic administration attenuates this resistance such that colonisation can occur. The resultant infection, *Clostridium difficile*-associated disease (CDAD), is potentially life-threatening, and represents a considerable financial and logistical burden for healthcare institutions.

While exposure to the microbe is an absolute pre-requisite for development of the disease, other aspects of the epidemiology and pathogenesis of CDAD are as yet incompetely defined. However, the hypothesis at the core of the current research is that the composition of the microbiota may contribute to the multifactorial nature of the infection: certain individuals may have a microbiotic fingerprint which confers protection against the pathogen, while the flora of others means they are more susceptible to colonisation by *Clostridium difficile.*

The aim of the curent research was to investigate the microbiota of a number of individuals treated with antibiotics and subsequently falling into 3 distinct groups; those who contracted CDAD; those who developed diarrhoea not caused by *Clostridium difficile*; and those who displayed no evidence of intestinal disruption. To identify substantive differences between the groups it was essential to characterise the intestinal bacteria in a comprehensive manner, beyond the potential of existing techniques. Achievement of the research objective thus necessitated the development of novel array and sequencing approaches, along with complementary bioinformatic pipelines for analysis.

ACKNOWLEDGEMENTS

My gratitude and appreciation are due to everyone who has contributed to this course of research; it has been a challenging process, both from scientific and personal perspectives, and could not have been completed without the assistance of all concerned.

I would particularly like to thank my supervisors, Professors Mike Barer and Julian Ketley, both for granting the opportunity to undertake this project (their decision to offer the position must have required a significant leap of faith considering the initial interview), and their patience and understanding throughout the past 5 years, especially in view of the disappointing nature of some of the results. My appreciation for funding the research is extended to the Medical Research Council.

Thanks also to the post-docs who worked on the project, Dr Emma Beards and Dr Kelvin Lau, who have moved on to fresh pastures, hopefully not entirely due to their experiences at Leicester. Specific mentions for: Dr Robert Free, for his advice on coding; Reshma Bharkada, for her tireless efforts with 454 sequencing at Leicester and quiet optimism that problems could be resolved; and Dr Tim Gant, Dr Emma Marczylo and Kate Dudek for their suggestions and assitance with regards to the arcane world of microarrays.

In more general terms I would like to thank all the members of laboratory 121 in the Genetics department for creating such a congenial working environment, particularly Dr Richard Haigh, Dr Randeep Sandhu, Dr Claire Miller, Dr Ran Ren, Sue Hardy and Fadil Bidmos. Appreciation is also extended to the technical staff in the Genetics department, especially Keith Baker and Ian Townson. I would also like to express my gratitude to Dr Robert McKenzie, Dr Rueban Isaac and Dr Su-Min Lee, who helped primarily by just being there, and to Mick Batkin for helping me to listen and to see.

Last, but by no means least, thank you to all my family, for their unwavering support and encouragement, which I intend to return in kind in the course of their own continuing endeavours.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

A	Adenine
AAD	Antibiotic-associated diarrhoea
ADP	Adenosine diphosphate
AGE	Agarose gel electrophoresis
AIX	Ampicillin-IPTG-XGal
AMP	Antimicrobial peptide
ANOVA	Analysis of Variance
ATP	Adenosine triphosphate
BC	Bray Curtis
Bc	B lymphocyte
BHI	Brain heart infusion
BSA	Bovine serum albumin
C	Cytosine
C. difficile	*Clostridium difficile*
cAMP	Cyclic adenosine monophosphate
CCFA	Cycloserine, cefoxitin and fructose agar
CD	*Clostridium difficile*
CDAB	*Clostridium difficile* agar base
CDAD	*Clostridium difficile*-associated diarrhoea/disease
CDI	*Clostridium difficile* infection
CDMN	*Clostridium difficile* selective supplement
COPD	Chronic obstructive pulmonary disease
CR	Colonisation resistance
Cy3/ Cy5	Cyanine3/Cyanine5
ddH_2O	Double distilled water
ddNTPs	Dideoxynucleotide triphosphates
DGGE	Denaturing gradient gel electrophoresis
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotide triphosphates
FISH	Fluorescent In-Situ Hybridisation
G	Guanine

GALT	Gut-associated lymphoid tissue
GI	Gastro-intestinal
HMP/I	Human Microbiome Project/Initiative
I	Inosine
IBD	Inflammatory Bowel Disease
IEC	Intestinal epithelial cell
IgA	Immunoglobulin A
IgG	Immunoglobulin G
IL	Interleukin
JC	Jaccard's coefficient
K	Guanine or thymine (Keto)
LB	Luria-Bertani
LPS	Lipopolysaccharide
LSU	Large subunit
M	Adenine or cytosine (Amino)
MALT	Mucosa-associated lymphoid tissue
MC	M cell
MH	Morisita Horn
MID(s)	Multiplex Identifier (s), Tag(s), Barcode(s)
MLST	Multi-locus sequence typing
N	Any nucleotide
NA	Not applicable or Not available
NLR	NOD-like receptor
O/N	Overnight
OFARG	Oligonucleotide fingerprinting of arrayed ribosomal genes
OFRG	Oligonucleotide fingerprinting of ribosomal genes
OTU	Operational taxonomic unit
P.S.I	pounds per square inch
PBS	Phosphate-buffered saline
PC	Paneth cell
PCR	Polymerase Chain Reaction
Phylum *CFB*	*CytophagaFlavobacteriumBacteroides*
PMC	Pseudomembranous colitis
PMT	Photo-multipler tube

PRR	Pattern recognition receptor
PTP	Picotiter Plate
QIIME	Quantitative insights into Microbial Ecology
qPCR	Quantitative PCR
R	Adenine or guanine (Purine)
rDNA	Ribosomal deoxyribonucleic acid
RDP	Ribosomal database project
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
RT	Room temperature
SC	Sørensen's coefficient
SCFA	Short-chain fatty acid
SD	Standard deviation
SDS	Sodium dodecyl sulphate
SI	Simpson's Index
SNR	Signal-to-noise ratio
spp.	Species
SSC	Saline sodium citrate
SSU	Small subunit
T	Thymine
TAE	Tris-acetate EDTA
TGGE	Temperature-gradient gel electrophoresis
Th	T helper lymphocyte
TLR	Toll-like receptor
T_m	Melting temperature
T-RFLP	Terminal restriction fragment length polymorphism
U	Units
UOL	University of Leicester
Y	Cytosine or thymine (Pyrimidine)
ΔMS	Change in microbiotic structure

CHAPTER 1

INTRODUCTION

1.1 Prokaryotes and microbiology

1.1.1 Prokaryotes

The existence of infectious 'seeds' had been suspected for centuries before Girolamo Fracastoro's treatise on Syphilis in 1530, not least during the bubonic plague pandemics which decimated the populations of the known world in the 14th century, but probably first mooted by Lucretius in ancient Rome (Willey *et al*., 2008). However, the doorway to visualisation of these agents and appreciation of their nature was only opened in the late 1600s with the development of the microscope by Anton Van Leeuwenhoek (Lederberg, 2000); his description of 'animalcules' scraped from his own teeth represents the birth of microbiology.

Its transition to a science is attributed to the independent work of Pasteur and Koch (Lederberg, 2000), the latter's postulates forming the basis for definitive causal association of a microbial organism with a specific disease:

- The microorganism should be identified in every instance of the disease.
- Isolation and growth of the microorganism in pure *in vitro* cultures should be achieved from disease samples.
- Inoculation of a healthy host with the microorganism should induce the same disease

Koch's application of these founding principles (Willey *et al*., 2008) led to the identification of *Mycobacterium tuberculosis* in 1882, and a Nobel Prize for Koch in the early twentieth century. In addition, the painstaking work in the laboratory performed by Koch, his colleagues and other contemporaries provided many of the methodologies, techniques, growth media and apparatus which are the tools of today's microbiologists

and molecular biologists. However, there was early disregard for microbiology by the wider biological community, partly due to the inability of microscopes to resolve cellular structures of the bacteria. For instance, *Escherichia coli* has dimensions of approximately 1 μm by 4 μm, while a red blood cell is 7 μm in diameter (Willey *et al*., 2008). This led to the classification of prokaryotes (lacking a membrane-bound nucleus) as precellular, but studies by Griffith and, subsequently, the Avery-MacLeod-McCarty group on *Streptococcus pneumoniae* identified the 'transforming factor' of all life as DNA (Lederberg, 2000).

While the impetus for investigation of prokaryotes was their detrimental impact on the health of humans and agricultural resources, they are the most numerous and successful of the 'domains' of life on Earth with an estimated population range of 4-6 x 10^{30} cells (Whitman *et al*., 1998). Their importance to the planet's ecosystem can be assessed from conservative conversions of this cellular figure: they represent a carbon reservoir of 350-550 Pg of carbon (1Pg = 10^{15}g), almost equal to that estimated to reside in plants, while nitrogen levels may be 85-130Pg, or 10 times that extant in plant material (Whitman *et al*., 1998). An oft-quoted statement by Kluyver suggests that 'about one half of living protoplasm is microbial' (Kluyver and Van Niel, 1956), but this may well be an underestimate (Whitman *et al*., 1998).

The majority of prokaryotes are to be found in the soil, sedimentary layers and aquatic environments, where they participate in biogeochemical cycles such as the reconstitution of decaying material, and production of essential constituents of the atmosphere such as nitrogen and oxygen (Whitman *et al*., 1998). Metabolically, the prokaryotes are capable of utilising resources and exploiting niches which remain untapped by eukaryotic organisms, particularly due to their ability to dispense with oxygen as the terminal electron acceptor in respiration in favour of elements and compounds such as sulphur and nitrates, processes collectively termed lithotrophy

(Pace, 1997). This metabolic adaptability, encompassing the entire spectrum from organotrophy to autotrophy (fixation of CO_2 using energy provided by photosynthesis or lithotrophy), has enabled bacterial species to colonise the harshest of environments, from temperatures above 100°C and below 0°C, through extremes of acidity, alkalinity and pressure, to excesses of toxic heavy metals and ionizing radiation; *Thiobacillus* species, for instance, thrive at a pH below 2 (Truper, 1992). In addition, the 'domestication' and engineering of certain groups has allowed humans to exploit this metabolic diversity: lactobacilli have long been used for the production of cheeses and yoghurt, while the modern biotechnological industry relies heavily on bacteria for processes such as mass production of insulin (Johnson, 1983) and detoxification of waste materials (Lloyd and Lovley, 2001).

1.1.2 Diversity and traditional classification

Prokaryotes are thus numerous and ubiquitous, but such a description is uninformative with respect to their diversity and classification.

At the most basic level these can be assessed and described through morphological characteristics such as cell shape and conglomeration (clustering patterns), microscopically visible cell surface structures, motility, and cellular events such as sporulation and fission (Truper, 1992). Evidentally the microscope has been instrumental in such classification, supplemented (where gross visualisation is inadequate) by the use of staining, and antibodies directed against cell-surface structures for finer resolution (Willey *et al*., 2008). In addition, differentiation can be achieved on a metabolic basis, determination of the substrates utilised or biochemicals produced allowing for categorisation (Busse *et al*., 1995). Such techniques permitted an estimation of the numbers of bacteria in a given environment (total cell count) and a rudimentary view of the members of the community.

However, even with advances in microbiological techniques and concomitant increases in the number of identified bacterial species, it became clear that there was a disparity between the numbers extrapolated from microscopic examination and those resulting from *in vitro* cultivation, an observation which became known as the 'great plate count anomaly' (Staley, 2006). With hindsight this discrepancy is unsurprising, as the requirements of a given bacterial species for survival and proliferation may be intimately associated with its 'natural' environment, inclusive, from an ecological perspective, of other members of the community (Sachs and Hollowell, 2012). For instance, the cyanobacterium *Prochlorococcus* cannot be cultivated, even in media replicating the marine environment, in the absence of species which can degrade hydrogen peroxide through production of catalase (Morris *et al*., 2011). Thus, while traditional techniques remained the benchmark for characterisation of pure (homogeneous) cultures of bacteria in terms of phenotype (morphology, physiology and biochemistry), the differentiation of members of diverse (heterogeneous) populations represented a problematic undertaking.

However, development of reliable technologies enabling the determination of the sequences of entire genes (Sanger *et al*., 1977), coupled with work on the concept of an 'evolutionary clock' (Zuckerkandl and Pauling, 1962) based on neutral mutations in eukaryotes (Kimura, 1968) laid the foundations for a leap forward. The insights of Woese and contemporaries in the 1970s (Woese and Fox, 1977) would revolutionize assessment of diversity and classification of the entire bacterial domain, the basis for this paradigm shift being a constituent of the ribosome (Pace, 1997).

1.1.3 The ribosome

Ribosomes are the cellular structures governing the process of translation, whereby mRNA transcribed from genes directs the condensation reaction between amino acids leading to formation of proteins (Berg *et al*., 2007). They are found in all living

organisms and consist of a small subunit and a large subunit, the former mediating binding of mRNA and the correct choice of tRNA for the appropriate codon, while the latter directs translocation and peptide bond synthesis (Berg *et al*., 2007). In prokaryotic organisms the small subunit (SSU) has a Svedberg sedimentation coefficient of 30, while the large subunit (LSU) has a corresponding value of 50; the SSU is associated with the ribosomal RNA known as 16S (again due to the sedimentation coefficient) and 21 proteins, while the LSU comprises 23S rRNA, 5S rRNA and more than 30 proteins (Snyder and Champness, 2003).

1.1.4 Small subunit (16S) ribosomal RNA

The conservation of the function of ribosomes, and the near universality of the genetic code (Koonin and Novozhilov, 2009) for translation of the nucleotide sequence of the genome into the primary amino acid sequence of proteins, demand that its structural components be resistant to accumulation of mutations. Certainly, if deletions or substitutions were to occur in vital regions then the effect would be deleterious to the organism; considerable stretches of sequence, however, may undergo alteration without detrimental effect on function (Kitahara *et al*., 2012).

Woese and colleagues (Woese and Fox, 1977; Woese, 1987) recognised that the ubiqutous SSU rRNA with its slow rate of sequence change could form the basis of connecting all life, and, specifically for 16S rRNA, the foundation for determining the phylogenetic relationships between bacteria, thereby establishing a framework for classification (Hugenholtz, 2002). An early realisation of the approach was the adoption of the ‘domain’ (urkingdom) level of taxonomy, above that of kingdom, to differentiate between bacteria and archaea (Woese and Fox, 1977).

As analysis of 16S rRNA sequences became more widely accepted as a means of phylogenetic classification, the species concept for bacteria underwent refinement. From a total DNA homology of more than 70%, and a change in melting temperature

(ΔT_m) of no more than 5°C (Wayne *et al*., 1987), it became the more stringent requirement of greater than 95% homology of 16S rRNA (Truper, 1992), By the early years of this century the threshold had been raised to greater than 97% sequence identity of 16S rDNA (Stackebrandt and Goebel, 1994; Bäckhed *et al*., 2005), and most recently it has been asserted that bacteria sharing less then 98.7% 16S rDNA identity must be considered as belonging to different species (Stackebrandt and Ebers, 2006), although debate as to the exact nature and precise definition of a bacterial 'species' continues (Stackebrandt *et al*., 2002; de Queiroz, 2005; Doolittle and Papke, 2006).

1.2 Molecular classification approaches

Contemporaneous with (and partially responsible for) recognition that SSU RNA could provide the basis for phylogenetic classification was elucidation of the full sequences of a growing number of 16S rDNA genes, such as that for *Esherischia coli* (Brosius *et al*., 1978). With the increase in data came the possibility of constructing phylogenetic trees by pairwise alignment of the sequences with divergence representing arbitrary evolutionary distance (Pace, 1997). At least as important was the confirmation that regions of the 16S sequence are highly conserved across the bacterial domain (Baker *et al*., 2003), interspersing the nine hypervariable species-specific regions, V1-V9 (Chakravorty *et al*., 2007), as shown in figure 1.1.

Awareness of the limitations of culture as an approach to bacterial analysis, especially of complex communities, had become increasingly evident (Amann *et al*., 1995). This is founded in differential nutrient and atmospheric requirements, along with ignorance of the interplay between microbial species in their normal habitat (Van der Wielen *et al*., 2002); estimates of the cultivatable fraction of a bacterial community range from less than 1% in extreme biospheres (Sogin *et al*., 2006), to perhaps 20% of a population associated with the intestinal milieu of humans and other animals (Wilson and Blitchington, 1996; Zoetendal *et al*., 2004). Indeed, bacterial species amenable to *in vitro* elaboration have been described as 'the weeds' of the bacterial world (Hugenholtz, 2002). Although culture-based investigations had provided valuable insight into community structure (Moore and Holdeman, 1974), the potential to circumvent the inherent biases, particularly with regard to low G+C anaerobes, led to widespread adoption of the molecular approach to analysis of microbial populations (Hugenholtz, 2002).

Initially the culture-independent approach focused directly on the ribosomsal RNA (Lane *et al*., 1985), but development of the techniques to encompass analysis of the rDNA of an entire community soon followed. Such investigations led to the discovery of entire new lineages of bacteria, exemplified by the discovery of EM17 and EM19 in the extreme conditions of Octopus Spring pool of Yellowstone National Park (Reysenbach *et al*., 1994).

The numerous methodologies utilised at present follow a similar course: extraction of bacterial community genomic DNA, design of 'universal' primers with the potential to anneal to the conserved regions of the majority of members, amplification via PCR, and analysis of the sequences represented, the variable regions providing for taxonomic discrimination (Hugenholtz, 2002).

The predominant technique remains incorporation of the PCR products into a suitable vector for cloning, followed by traditional sequencing (Sanger *et al*., 1977), an approach used recently to study the microflora of the human intestine (Eckburg *et al*., 2005). The output of such an investigation allows for collation of the sequences in an appropriate catalogue, such as the Ribosomal Database Project (RDP; Cole *et al*., 2005), the data then feeding back into refinement of probes and primers for subsequent analyses using, for instance, (FISH) fluorescent-in-situ-hybridisation (Hugenholtz, 2002).

Other methods have been developed such as denaturing-gradient gel electrophoresis (DGGE), and a closely-related alternative, temperature-gradient gel electrophoresis (TGGE). Both rely on discrimination of 16S PCR amplicons through differential mobility under electrophoresis, caused by variations in dissociation rates of a DNA duplex dependent on oligonucleotide composition, those with a high G+C content being comparatively more stable (Muyzer, 1999). Terminal restriction fragment length polymorphism (T-RFLP) is another technique which relies on visualisation of PCR products. PCR is conducted with labelled primers before the products are digested with

a selection of restriction enzymes. Due to 16S sequence polymorphisms varying fragment lengths are created which can then be differentiated by gel electrophoresis (Moyer *et al*., 1994).

FIGURE 1.1: Variable regions of the 16S gene and classification accuracy

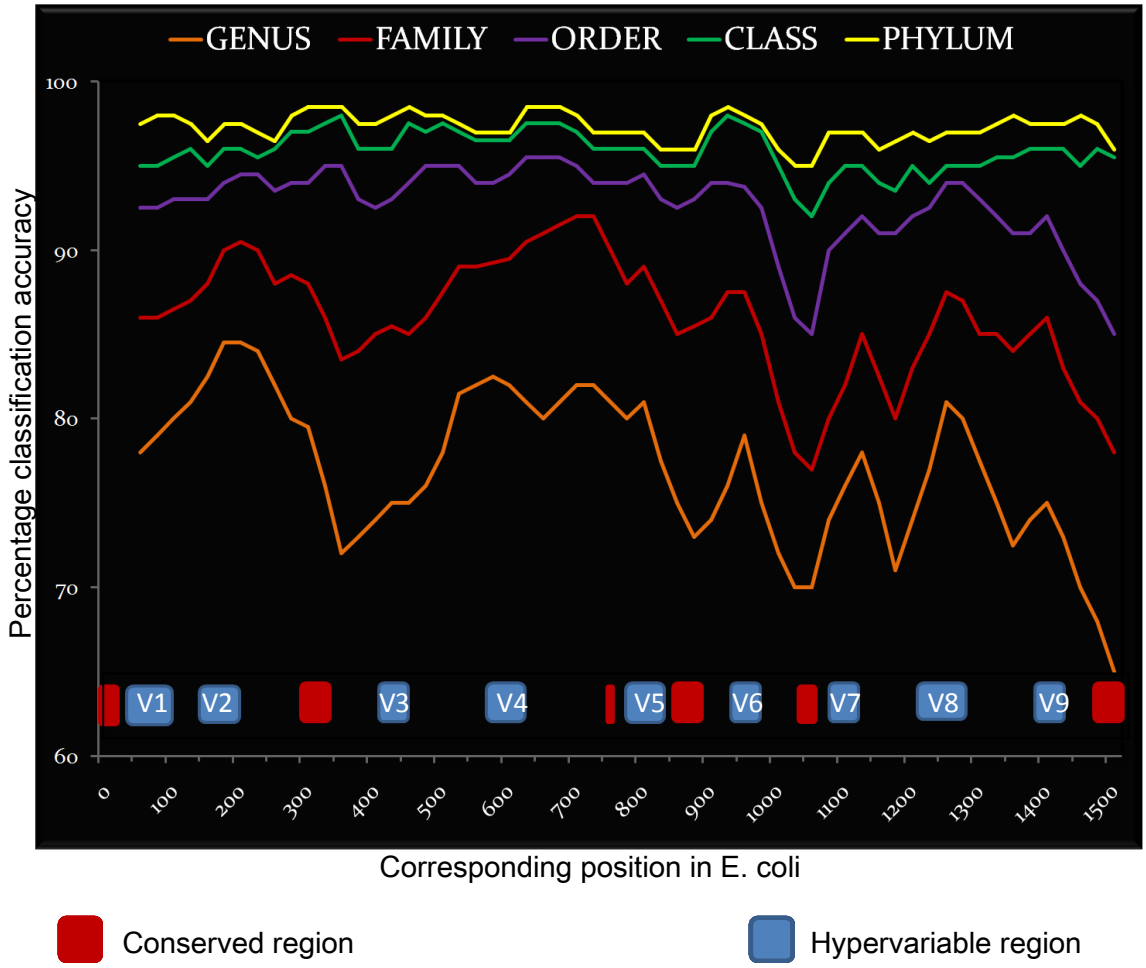


Figure 1.1 displays the approximate nucleotide positions of the nine hypervariable regions of the 16S rRNA gene, interspersed with highly conserved regions affording design of 'universal' primers for amplification. Classification accuracy on the y-axis is based on bootstrapping confidence levels using 100 bp fragments centred around the the 50 bp intervals on the x-axis and covers varying levels of taxonomy (adapted from Wang *et al*., 2007).

The advances catalysed by the supplementation of culture with SSU RNA analysis are manifold, even extending outside the microbial domain to the confirmation that mitochondria and chloroplasts are bacterial symbionts, descended respectively from proteobacteria and cyanobacteria (Pace, 1997). Analysis of phylotypes (phylogenetic identification independent of cultivation) has also revealed that bacterial diversity arose primarily as an ancient explosion of lineages rather than the incremental steps associated with eukaryotic evolution (Pace, 1997), while the paradigm shift of sequence-based approaches to classification has expanded the number of phyla from less than 10, to more than 40 (Hugenholtz, 2002), albeit with a reduction in 'shoehorning' of genera and species into inappropriate taxa, and a more systematic classification. Unexpected evolutionary nexi have also been illuminated, such as the close relationship between *Cytophaga, Flavobacterium* and *Bacteroides*, all now incorporated into the *Bacteroidetes* phylum (Hugenholtz, 2002).

It must be acknowledged, though, that such classification is not immutable: analyses of different genes such as *gyrB* and *polD* are capable of producing contrasting phylogenetic trees (Yamamoto and Harayama, 1998), and metabolic function is often diverse within phylogenies derived from SSU RNA investigation. Phylogenetic relationships are also complicated by the phenomenon of lateral gene transfer, whereby bacterial species obtain coding sequences from nominally unrelated bacteria in their environment through conjugation, transduction and transformation (Snyder and Champness 2003), even if informational genes are considered less susceptible (Hugenholtz, 2002).

In the context of community analysis, molecular techniques have illuminated features of the diversity and dynamics of microbial populations, from the human alimentary canal through cloning (Suau *et al*., 1999), and DGGE (Vanhoutte *et al*., 2004), to deep-sea hydrothermal vents via T-RFLP (Moyer *et al*., 1994). From a clinical perspective,

identification of the aetiological agent of Whipple's disease, *Tropheryma whippeli*, which had proven resistant to isolation, was achieved through analysis of 16S rDNA (Relman *et al*., 1992). Incidences of infectious diseases whose aetiological agent remains unculturable but has been causally associated through molecular means, have led to a relaxation in stringency of application of Koch's postulates, a situation resisted but foreseen by their author (Fredericks and Relman, 1996).

Molecular techniques do not stand alone, however; culture, morphological description, and metabolic analysis remain the benchmarks of microbiology, while genome sequencing, notation, and functional characterisation of genes are the standards required by genetics. Analysis of 16S rRNA has broadened the horizons of microbiology, but remains an (albeit powerful) adjunct to traditional approaches. It must be remembered that the molecular methodologies are associated with biases of their own, from variability caused by sampling and DNA extraction protocols to those inherently associated with attempts to perform PCR on complex samples (Wintzingerode *et al*., 1997).

Means of identification aside, it is humbling to acknowledge that there may be more than 10,000,000 species of bacteria, of which only about 5,000 have been fully categorised, with less than 150 being pathogenic to humans (Pace, 1997). The majority of those cultivated previously fall into just 4 of the 51 phyla: *Proteobacteria, Actinobacteria, Firmicutes,* and *Bacteroidetes*, while only 8 genera from these, including *Escherichia*, *Bacillus,* and *Staphylococcus* have been extensively studied (Hugenholtz 2002). The total diversity of the bacterial domain remains fundamentally unexplored.

1.3 The human intestine

The intestine is an anatomically distinct region of the alimentary canal which stretches from the pyloric sphincter to the anus. The small intestine comprises the duodenum, the jejunum and the ileum and is approximately 6 metres in length in an adult (Stevens and Lowe, 1997). In addition to villous projections which dramatically increase the surface area there are numerous glands and crypts which play a role in final digestion and absorption of the dietary intake. In this respect it is aided by secretions from the pancreas and liver. The large intestine commences at the ileocaecal valve and extends from the caecum through the ascending, transverse, descending and sigmoid portions of the colon to the rectum and anus. While only 1.5 m long it derives its name from the comparatively greater diameter of the lumen, and is concerned with reabsorption of water and electrolytes, and breakdown of waste materials (Clemente, 1987; Guyton, 1991).

At a finer level of resolution the tissues consist of: the mucosa containing epithelial cells and associated basement membrane; the *lamina propria* with associated blood vessels, nerve endings, lymphatics and *muscularis mucosae*, a thin layer of muscle; the submucosa, the support tissue layer; the *muscularis proper*, comprising the thick bands of circular and longitudinal muscle; and finally the adventitia which becomes continuous with the lining of the peritoneal cavity (Stevens and Lowe, 1992). The epithelial stem cells differentiate into a wide variety of histological types, from enterocytes concerned with absorption, to goblet cells (producing mucus) and M cells overlying lymphoid tissue and concerned with antigen presentation (Stevens and Lowe, 1992; O'Hara and Shanahan, 2006). Such aggregations of lymphocytes are known as gut-associated lymphoid tissue (GALT) and are concentrated in the submucosal layer, although they may extend into the lamina propria (Stevens and Lowe, 1992). In addition to these B and T cells which can be primed to produce immunoglobulin A (IgA) for

secretion into the lumen, there is a resident enteric population of monocytes and macrophages, distributed throughout the mucosa and capable of phagocytosing invaders and presenting their antigens to lymphocytes (Roitt, 1994), as well as dendritic cells which can sample the luminal interface directly through pattern recognition receptors (O'Hara and Shanahan, 2006). While the epithelium may be specialised to carry out a range of secretive and absorptive functions, its primary role remains a physical barrier for protection of the systemic milieu - from loss of vital nutrients such as water and invasion by potentially harmful agents such as bacteria (O'Hara, and Shanahan, 2006).

FIGURE 1.2: The human intestine and associated microbal density

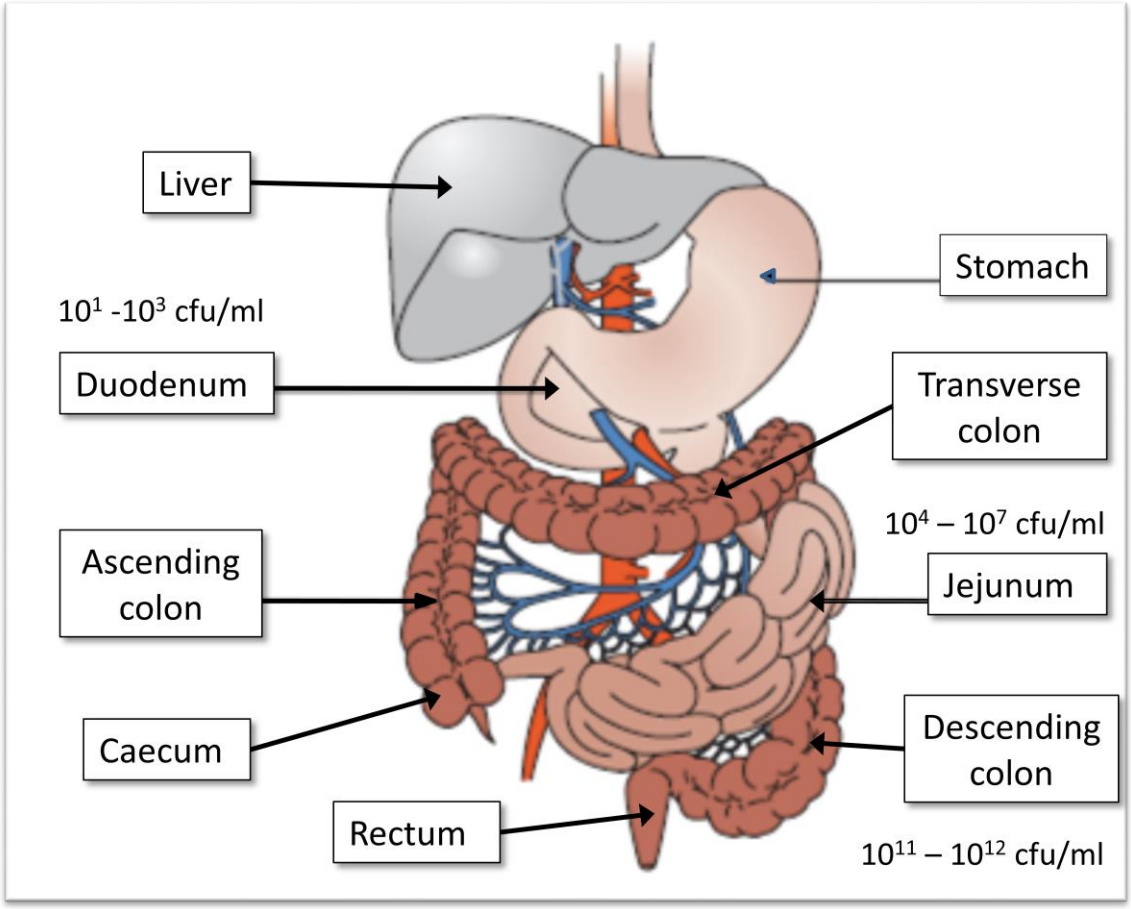


Figure 1.2 shows the human intestinal compartments and estimates of the number of bacteria per ml for each region.(Adapted from O'Hara and Shanahan, 2006).

1.4 Microbiota

1.4.1 Overview

The microorganisms resident in any system may collectively be described as the microbiota, an umbrella term which encompasses viruses, fungi, protozoa, archaea and bacteria (Hill and Artis, 2009), although it is often used generically, as here, to denote only the baterial proportion. An associated term is the 'microbiome', used to describe the total gene-pool of the microflora found in an environmental niche (Turnbaugh *et al*., 2007)

The microbiotic communities associated with animals, including humans, are some of the most densely concentrated as yet defined (Bäckhed *et al*., 2005). While numbers are relatively low on the skin, being approximately 10^3-10^4 bacteria per cm^2 except in the groin and axillary regions where density may be $10^6/cm^2$ (Whitman *et al*., 1998), the human intestine is thought to contain between 10^{13} and 10^{14} microorganisms (Gill *et al*., 2006; Ley at el., 2006), a figure more 'precisely' defined by Griff-Rhys Jones in the Yakult probiotic yoghurt advertisements as being 55 trillion. Such claims notwithstanding, it is thought that there may be as many as 10^{12} bacteria per gram of faeces (Tonna and Welsby, 2005), and more than 400 different resident intestinal species (McKenna *et al*., 2008). Even conservative estimates of the total population number mean that bacteria probably outnumber the somatic and germ cells of their host by a factor of 10 (Bäckhed *et al*., 2005) while for the microbiome in comparison to the human genome this factor may be more than 100 (Bäckhed *et al*., 2005).

1.4.2 Composition

Original investigations into the microbiota of the human intestine were conducted using traditional culture and microscopic techniques, but nonetheless revealed considerable diversity including *Fusobacterium prausnitzii*, Clostridium cluster IV

Ruminococcus spp., *Bacteroides fragilis*, and *Lactobacillus acidophilus* (Moore and Holdeman, 1974). More recent studies utilising molecular analyses have revealed the presence of some of these species and the overall dominance of the colonic intestinal environment by *Bacteroidetes* and *Firmicutes* (Suau *et al*., 1999), with bifidobacteria of the phylum *Actinobacteria* also commonly detected (Wang *et al*., 1996), and in some studies outnumbering the *Bacteroidetes* (Andersson et., 2008). Estimates of the proportion of unculturable bacteria from the colon vary between 40% (Wilson and Blitchington, 1996) and 80% (Suau *et al*., 1999), but it must be borne in mind that these figures are derived from disparities between microscopic observations and plate counts and thus do not represent the proportion of total phylotypes which may be resistant to *in vitro* cultivation (Zoetendal *et al*., 2004). In addition, the intestine is not homogenous; variations in oxygenation, pH and nutrient avaiability in the anatomical compartments from oral cavity to colon create environments which favour colonisation by certain species over others (Berg, 1996; Hayashi *et al*., 2005). For instance, the low pH in the gastric compartment prevents most bacteria from flourishing, apart from *Helicobacter pylori* and certain species of lactobacilli and streptococci which are unusually acid-tolerant (Laheij *et al*., 2003). Even within these gross anatomical regions there are also micro-environmental variations which may contribute to differences in autochthonous or indigenous populations (Mackie *et al*., 1999). Thus, in addition to the vertical stratification, horizontal variation can also be identified between luminal, epithelial and mucosal sites (Zoetendal *et al*., 2002), the latter also being sub-divided into a superficial and deep zone (Berg, 1996).

Despite these caveats and a modicum of discordance as to the precise composition and proportions at a more detailed level of classifiation, particularly with regard to minor representatives and allochthonous (transient) species, there is some consensus as to the major constituents of the human colonic intestinal microbiota. It is generally accepted

that the *Bacteroidetes* and *Firmicutes* constitute more than 90% of the colonic population (Turnbaugh *et al*., 2006), while obligate, as opposed to facultative, anaerobes predominate (Wang *et al*., 2005); this is evidenced by the domination of sequenced 16S library clones by the *Firmicutes* and *Bacteroidetes* (CFB) divisions/phyla (Gill *et al*., 2006; Eckburg *et al*., 2005), while other phyla encountered less frequently include the *Proteobacteria*, *Actinobacteria*, *Fusobacterium*, and *Verrucomicrobia* (Eckburg *et al*., 2005; Wang *et al*., 2003; Suau *et al*., 1999). Butyrate-forming clostridia tend to predominate amongst the *Firmicutes* (Andersson *et al*., 2008; Eckburg *et al*., 2005), with numerous sequences aligning to genera of clusters XIVa and IV (Suau *et al*., 1999; Wang *et al*., 1996; Hold *et al*., 2003), representing low G+C Gram-positives related to *Clostridium coccoides*, and Gram-positives related to *Clostridium leptum*, respectively (Hold *et al*., 2003). The genera identified most frequently are *Fusobacterium*, *Bifidobacterium*, *Bacteroides*, *Peptococcus*, *Ruminococcus*, *Clostridium*, *Eubacterium*, and *Peptostreptococcus* (Suau *et al*., 1999). Species of the phylum *Proteobacteria* are rarely identified, predictable from their generally facultative nature (Eckburg *et al*., 2005), although some such as *Esherischia coli* and *Campylobacter jejuni* are clearly capable of colonisation and pathogenesis. In view of the bacterial diversity of the human intestine, it is perhaps surprising that *Methanobrevibacter smithii* is the only species of the archaeal domain encountered with any frequency (Gill *et al*., 2006; Eckburg *et al*., 2005).

In many senses, though, comparisons between investigations are difficult to undertake. Sites and methods of sampling may differ, PCR conditions and efficacy could vary considerably, and definitions of OTUs (operational taxonomic units) span the range from 95-99% similarity (Zoetendal *et al*., 2004), while a multitude of factors may influence the composition of the intestinal microbiota at a given point in time.

1.4.3 Inter-individual variation

While it is possible to identify certain bacterial groups common to the majority of humans, especially at the higher levels of phylogenetic classification, there is also considerable inter-individual variation (Lay *et al*., 2005; Eckburg *et al*., 2005). This variability may arise due to a number of factors including genotype, gender, environment, diet, age, and accession (Turnbaugh *et al*., 2007; Dethlefsen *et al*., 2006).

The intestine of the foetus is free of microbes but the post-partum environment leads to rapid bacterial inoculation of the gut (Mackie *et al*., 1999), most commonly by *E. coli* and other enterobacteria and streptococci, with the vanguard affecting host gene expression such that colonisation by other bacteria is prevented while conditions for their own growth are enhanced (Xu and Gordon, 2003). In this respect there is potentially an element of succession in determination of the composition of the microbiota, whereby the first species to colonise influence the overall structure of the eventual commensal community (Eckburg *et al*., 2005; Turnbaugh *et al*., 2007); indeed, transplantation of the microbiota into gnotobiotic (germ-free) subjects from different species have shown that the resultant community is strongly dependent on the identity of the colonisers (Rawls *et al*., 2006). However, it seems that the prevalence of facultative anaerobes in neonates is not translated into long-term predominance, perhaps related to the dietary changes associated with weaning and progressive deoxygenation of the intestinal environment (Hooper, 2004). The latter effect commences relatively soon after birth such that the intestinal environment is sufficiently reduced for colonisation by bifidobacteria, bacteroides and clostridia within 4-7 days (Mackie *et al*., 1999). The process of succession continues over the next 12-24 months until the microbiota resembles that of an adult by about the second year (Stark and Lee, 1982), although the 'stable' climax community is not fully established until well into adolescence (Hopkins *et al*., 2001).

While there continue to be fluctuations in microbiotic compostition in response to other environmental factors throughout adulthood such that the term 'stable' may be a misnomer, a further distinct change in the microbiota is then evident as an individual becomes elderly (Claesson *et al*., 2010), perhaps partly in response to physiological changes such as reduced gastro-intestinal secretion and increased mucosal permeability (Woodmansey, 2007). Studies have identified a reduced number of bifidobacterial species (Hopkins *et al*., 2001) and attenuated diversity of *Bacteroides* species (Bartosch *et al*., 2004) along with changes in the relative dominance of clostridial species and a generalised reduction in the potential of the microbiota for degradation of complex polysaccharides to short chain fatty acids (Woodmansey *et al*., 2004).

Diet as a modulatory factor for composition of the microbiota has been mentioned in relation to the potential influence of weaning on changes associated with the progression from infancy to childhood (Hooper, 2004). At an even earlier stage of human development it has also been found that there are differences in the microbiota of infants who are breast-fed compared to those raised on formula substitutes, the latter group displaying a greater prevalence of staphylococci and clostridia compared to lactobacilli and streptococci in the former (Harmsen *et al*., 2000). Studies on adults, however, had pointed to a less dramatic influence of dietary changes (Finegold and Sutter, 1978), perhaps due to the capacity of many intestinal bacteria to metabolise a broad range of substrates (Flint, 2004), including the ubiquitous mucin (Cummings and Macfarlane 1991), and the inherent difficulties of colonisation for novel species when competing with well-established populations (Dethlefsen *et al*., 2006). More recent studies, though, have led to renewed interest in the diet as a driving force in dynamics of the intesinal mirobiota. One study found that there were significant differences in intestinal mirobial populations beween individuals on normal omnivorous, vegetarian, and vegan diets, particularly with regard to numbers of bifidobacteria and bacteroides

(Zimmer *et al*., 2011). Another investigation displayed a reduction in the number of *Roseburia* spp., and certain butyrate-producing members of clostridium cluster XIVa (*Lachnospiraceae*) consequent to a reduction in dietary carbohydrate intake (Duncan *et al*., 2007), while the abundance of erysipelotrichi was found to show a positive correlation with the level of dietary fat (Turnbaugh *et al*., 2009). It is possible then, that dietary influence is stronger than previously thought, but that the timescale of sampling, and/or approach to identification, were not adequate to identify these effects. In some sense diet can be regarded as one of a range of environmental factors which could influence the microbiotic membership or relative abundances of the indigenous taxa, but can be broadly described under the headings of geographical location (Mueller *et al*., 2006) and lifestyle (Dicksved *et al*., 2007).

The final significant determinant of inter-individual variation is that of the host genotype itself (Zoetendal *et al*., 2001). The host genome clearly has a direct bearing on aspects of the immune system (such as the *HLA* genes of the major histocompatibility complex responsible for antigen presentation to B cells) and the expression of cell-surface molecules with which commensal bacteria interact, while the metabolic phenotype of the host could also influence the availability of certain nutrients in the intestine (Spor *et al*., 2011). However, studies of twins have not displayed a profound similarity in microbiotic composition above that found in more distant family members (Zoetendal *et al*., 2001; Turnbaugh *et al*., 2009); indeed, the 'heritable' components may derive from the 'maternal effect', the successional and environmental factors mentioned earlier. Technological and analytical advances, though, have permitted the investigation of co-segregation of taxa and host genes in mice; one study found that an interleukin-22 gene (*Il22*), and a kinase gene (*Irak3*) which modulates a Toll-like receptor 2 (TLR2) pathway, partitioned with a reduced abundance of lactococci (Benson *et al*., 2010).

It has also been noted that a distinct *NOD2* genotype in humans was associated with a shift in relative proportions of certain bacterial taxa of the clostridial family *Lachnospiraceae* (Frank *et al*., 2011); while the correlation did not reach a level of statistical significance it is interesting that the *NOD2* gene is ubiquitously expressed in Paneth cells, which secrete defensins into the crypts of the distal intestine (Wehkamp *et al*., 2004). A further example of the potential for genotype to influence the microbiota comes from the effect of interferon (IFN) regulatory factor 9 deficiency in mice, whereby the stability of the microbiota over time is reduced (Thompson *et al*., 2010).

The inter-individual variation of the intestinal microbiota can thus be viewed, at least in part, as deterministic: a multitude of environmental and genotypic (host and microbial) factors are responsible, and, while the interplay may be complex, it should in theory be predictable. However, unknown stochastic factors can further complicate the landscape. For instance, in a study of the contribution of the leptin (*ob)* gene to the microbiota of mice, *ob/ob* (phenotypically obese) individuals were found to have similar shifts in the relative abundance of firmicutes and bacteroidetes in favour of greater numbers of the former (Ley *et al*., 2005); interestingly, though, this effect was superimposed on differences between siblings which arose despite the near identical nature of their genotypes and environmental exposures, the stochastically-derived variation was also then transferred to the subsequent generation (Ley *et al*., 2005).

1.4.4 Symbiosis

Comparative genomic analysis of gut-associated bacteria across a variety of intestinal niches highlights differences which point to long-term co-evolution of the alimentary microbiota and its host (Walter *et al*., 2010), while the potential to track human migratory patterns dating back tens of thousands of years via multi-locus sequence typing (MLST) of *H. pylori* strains (Linz *et al*., 2007) evinces an enduring symbiotic relationship between microbes and man (Bäckhed *et al*., 2005).

When viewed in this respect the microbiotic species of the mammalian intestine have most often been regarded as commensals (Yan and Polk, 2004), whereby bacteria benefit from a milieu whose temperature, pH and reductive potential are regulated by host homeostasis and nutrient supplies are replenished frequently (Guyton, 1991), while for the host the relationship would be deemed neutral (Hill and Artis, 2009). However, in many respects the relationship has evolved to be predominantly mutualistic (Xu and Gordon, 2003; Xu *et al*., 2003; Bäckhed *et al*., 2005; Ley *et al*., 2006) whereby both the microbiota and host gain from the association, although there is the potential even for true commensals and mutualists to become parasitic (Paulsen *et al*., 2003; Garrett *et al*., 2010; Hansson and Johansson, 2010).

That mammals derive distinct advantages from the presence of the microbiota is evidenced by studies of gnotobiotic (germ-free) animal models. These have revealed abnormalities of the enteric vascular system (Baez and Gordon, 1971) and cardiac musculature (Wostmann *et al*., 1982), while intestinal muscle is thinner and digestive enzyme efficacy reduced compared to colonised counterparts (Shanahan, 2002). In addition immunological function is impaired in that GALT (gut-associated lymphoid tissue) is deficient and immunoglobin levels and cytokines are diminished (Shanahan, 2002). However, introduction of *Bacteroides thetaiotaomicron* (Xu *et al*., 2003) into such a system has been shown to influence host gene expression to alter nutrient metabolism, enhance angiogenesis, and direct maturation of the immune system (Xu and Gordon, 2003).

The greater proportion of primary digestion and absorption has been completed once the remains of the gastric chyme pass through the ileocaecal valve into the colon, but the activities of the colonic flora lead to: metabolism of dietary carcinogens and xenobiotics (Sekirov *et al*., 2009); synthesis of nutrients such as Vitamin K, biotin and folate (Hill, 1997); Mg^{2+}/Ca^{2+}/Fe^{2+} absorption (Balzan *et al*., 2007) ; and fermentation

of complex carbohydrates 'inaccessible' to the host genome, with concomitant energy salvage (Flint, 2004; Gill *et al*., 2006; Li *et al*., 2008).

In this latter respect the microbiome is enriched for glycoside hydrolases and polysaccharide lysases (Ley *et al*., 2006; Gill *et al*., 2006), which catabolise biological polymers such as cellulose, pectin and starch into their constituent monosaccharides with subsequent production of short-chain fatty acids (SCFAs), such as lactate, acetate, and butyrate, through primary fermentation (Flint, 2004; Backhed *et al*., 2005; Ley *et al*., 2005). Prominent bacterial groups responsible for carbohydrate metabolism are the bacteroidetes, the bifidobacteria, members of clostridial cluster XIVa, and members of clostridial cluster IV (Flint, 2004). The SCFA butyrate can be absorbed by the host, and is metabolised preferentially by colonocytes, although some residual uptake of the resultant monosaccharides may also be possible (Gill *et al*., 2006).

Thus the microbiota assists the host indirectly through complex polysaccharide digestion, but the microbiota also appears to influence expression of genes such that the supplementary dietary intake is preferentially deposited in adipocytes (Turnbaugh *et al*., 2008). The microbiome also contributes to degradation of, for instance, phenols from plant material, harmful to human cells but broken down by β-glucosidases (Gill *et al* ., 2006). In addition, there are genes encoding enzymes involved in the pathways for biosynthesis of deoxyxylulose 5-phosphate (DXP), from which vitamin B_1 (thiamine) and vitamin B_6 (pyridoxine) may be derived (Gill *et al*., 2006). In the above respects it is notable that the intestinal microbiome is enriched in comparison to that of other environments (Gill *et al*., 2006). The mutualistic nature of the relationship is particularly evident from this metabolic perspective: symbionts provide the host with supplementary calorific extraction, while the intestine represents a controlled anoxic environment for the commensals, rich in glycans should dietary intake be reduced (Bäckhed *et al*., 2005).

With regard to the metabolic potential of the microflora, numerous studies have been conducted in mice suggesting a link between obesity and differential composition of the bacterial comunity (Ley *et al* 2005; Turnbaugh *et al*., 2006). Mutations in the leptin gene lead to increased food consumption and obesity; introduction of the microbiota from these individuals (mice) into gnotobiotic subjects causes a significantly greater increase in body weight than colonisation with the microbiome of lean individuals (Turnbaugh *et al*., 2006). The genetically obese individuals are found to have a 50% reduction in bacteroides numbers compared to the lean (Ley *et al*., 2005), with a concommitant increase in numbers of Firmicutes (Turnbaugh *et al*., 2006). Furthermore, diet-induced obesity in humans was found to encourage population growth in the Mollicutes class (Clostridium cluster XIV e.g. *Eubacterium dolichum*) of the Firmicutes phylum which could then be reversed by dietary restriction, and numbers of bacteroidetes are increased as weight loss occurs through dieting (Turnbaugh *et al*., 2008). A firmicutes-enriched microbiome has a greater capacity for fermentation which can be 'transplanted' to genetically 'lean' mice to cause increased adiposity, suggesting that the community is acting not only to alter the nutrients available to the host, but also to influence their metabolism once absorbed (Turnbaugh *et al*., 2008). It is interesting to speculate as to whether the changes are host-mediated in an effort to limit energy uptake (Ley *et al*., 2005), especially since *Bacteroides* spp. are normally thought to contribute to the majority of polysaccharide digestion in the colon (Van Tongeren *et al*., 2005). In addition to the intervention of the microbiota in nutritional pathways, host responses to pharmacological agents can also be influenced by intestinal commensals, both through provision of alternative routes for xenobiotic metabolism and release of metabolites capable of stimulating hepatic enzyme systems (Li *et al*., 2008; Sousa *et al*., 2008).

The microbiotic role in maintaining the competence of the immune system is perhaps even more extensive and profound than its contribution to host nutritional function.

Commensals are known to interact with intestinal epithelial cells (IECs) via a range of receptors on the host membrane (Hill and Artis, 2009). Toll-like receptors (TLRs) recognise LPS (Lipopolysaccharide), PSA (Polysaccharide A) and flagellin components, while bacterial peptidoglycans are ligands for NOD-like receptors (NLRs); in addition, G protein-coupled receptors (GPCRs) are activated by surface layer protein A (SlpA) and products of metabolism such as butyrate (Hill and Artis, 2009). Downstream effects of these receptor-ligand interactions are complex and multifarious, but absence of this commensal 'priming' of the immune system causes gnotobiotic animals to be deficient in both IgA-producing plasma cells and $CD4^+$ T cells in the intestinal *lamina propria* (Macpherson and Harris, 2004), while proximal and distal lymph nodes may be poorly developed (Bauer *et al*., 1963) and serum immunoglobulin levels are also reduced (Benveniste *et al*., 1971). Figures 1.3 and 1.4

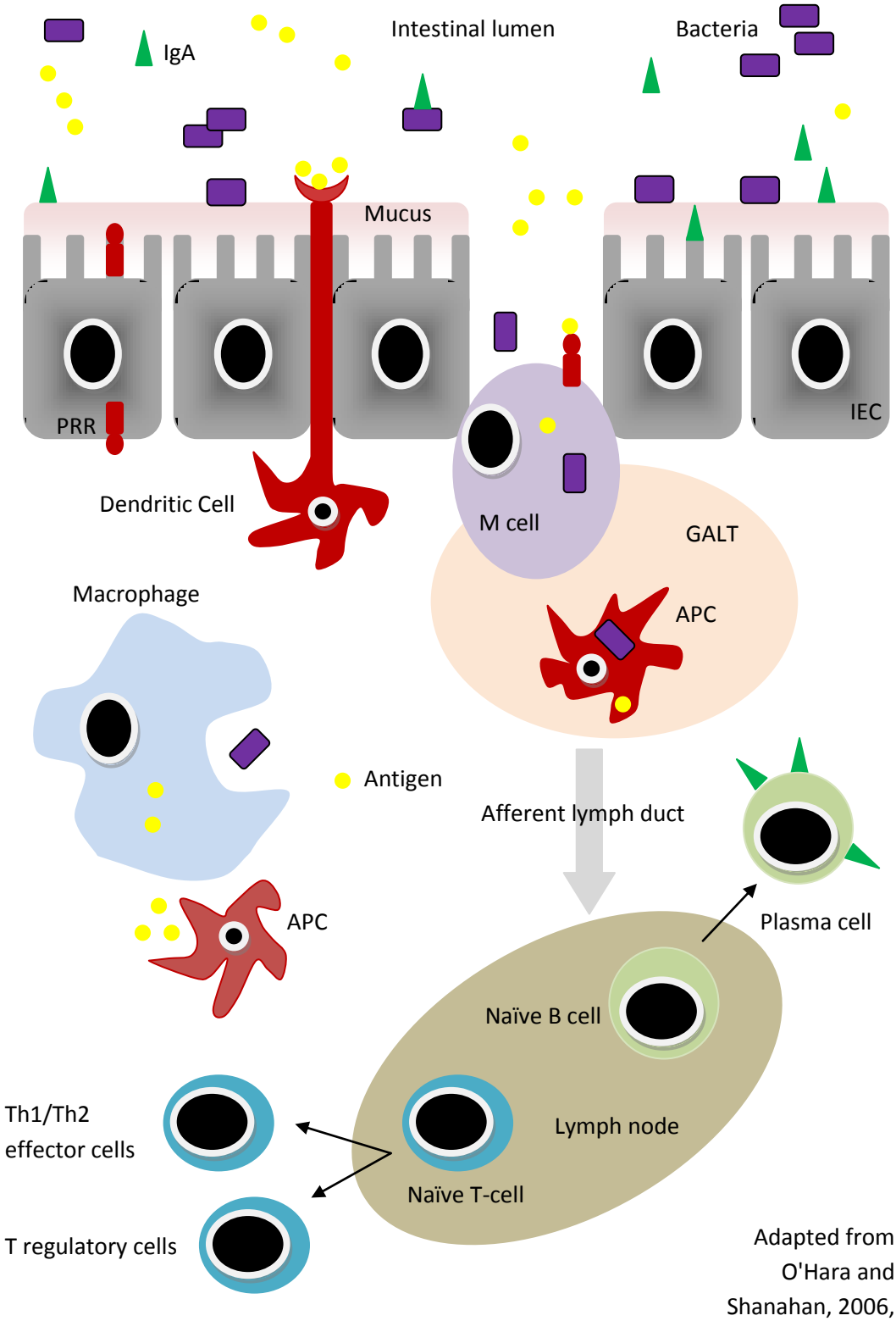
It appears that host pattern recognition receptor (PRR) systems such as the TLRs and NLRs associated with enterocytes and dendritic cells are unable to directly differentiate between commensal and pathogenic bacteria (Rakoff-Nahoum *et al*., 2004). One particular model for variable host responses suggests that pathogenic bacteria expressing virulence factors are more likely to breach the mucosal and epithelial barrier. Subsequently they would encounter macrophages as opposed to enterocytes and differential expression of cell-surface receptors on these populations of cells could account for the subdued response to commensal bacteria, the immune system effectively living in ignorance of their existence (Macpherson and Harris, 2004). However, certain commensals are known to release metabolites which exert direct anti-inflammatory effects via inhibition of TNF-α release induced by LPS from pathogens (Menard *et al*., 2004). In addition, increased activation of the pro-inflammatory transcription factor NF-κB by pathogenic bacteria (O'Hara and Shanahan, 2006) can be counteracted by *Bacteroides thetaiotaomicron* which enhances transport of the RelA subunit of NF-κB

from the nucleus via activation of the host nuclear PPARγ receptor (Kelly *et al*., 2005; Thomas and Versalovic, 2010), while other bacteria are capable of blocking the ubiquitination of IκB-α necessary for translocation into the nucleus (Neish et al, 2000). A different mechanism is responsible for suppression of damage induced by *Helicobacter hepaticus* which can cause colitis through increases in production of TNF-α and IL-17 (Mazmanian *et al*., 2008). In this instance, PSA from *Bacteroides fragilis* appears to neutralise this effect through induction of IL-10 production by $CD4^+$ T-cells (Mazmanian *et al*., 2008). Thus commensal symbionts appear capable of ameliorating the inflammatory damage cause by pathobionts.

It is perhaps worth noting at this point that designation of a bacterial species as commensal, mutualist, parasitic or pathogenic may be both circumstantial and temporal. Certain bacteria are regarded as inherently mutualistic, such as *Bacteroides thetaiotaomicron* (Xu *et al*., 2003), but the commensal population may contribute significantly to colitis induced by Enterobacteriaceae such as *Proteus mirabilis* (Garrett *et al*., 2010) while the change in the status of *Enterococcus faecalis* from commensal to opportunistic pathogen can be attributed to lateral (or horizontal) gene transfer of vancomycin resistance (Paulsen *et al*., 2003). Equally, suppression or deletion of virulence factors can lead to the 'domestication' of potential pathogens such as seen in *E.coli* 83972, which has lost several key pathogenicity islands along with the ability to mediate adhesion, and thus avoids eliciting an immune response (Klemm *et al*., 2007).

Further augmentation of the host defences against pathogens is achieved through fortification of the intestinal epithelial barrier. The microbiota is capable of stimulating epithelial cell regeneration via production of SCFAs (Shanahan, 2002), in addition to induction of IgA release (Macpherson *et al*., 2001), stimulation of anti-microbial peptide release by Paneth cells (Sekirov *et al*., 2009) and maintenance of tight junction integrity (Cario *et al*., 2007).

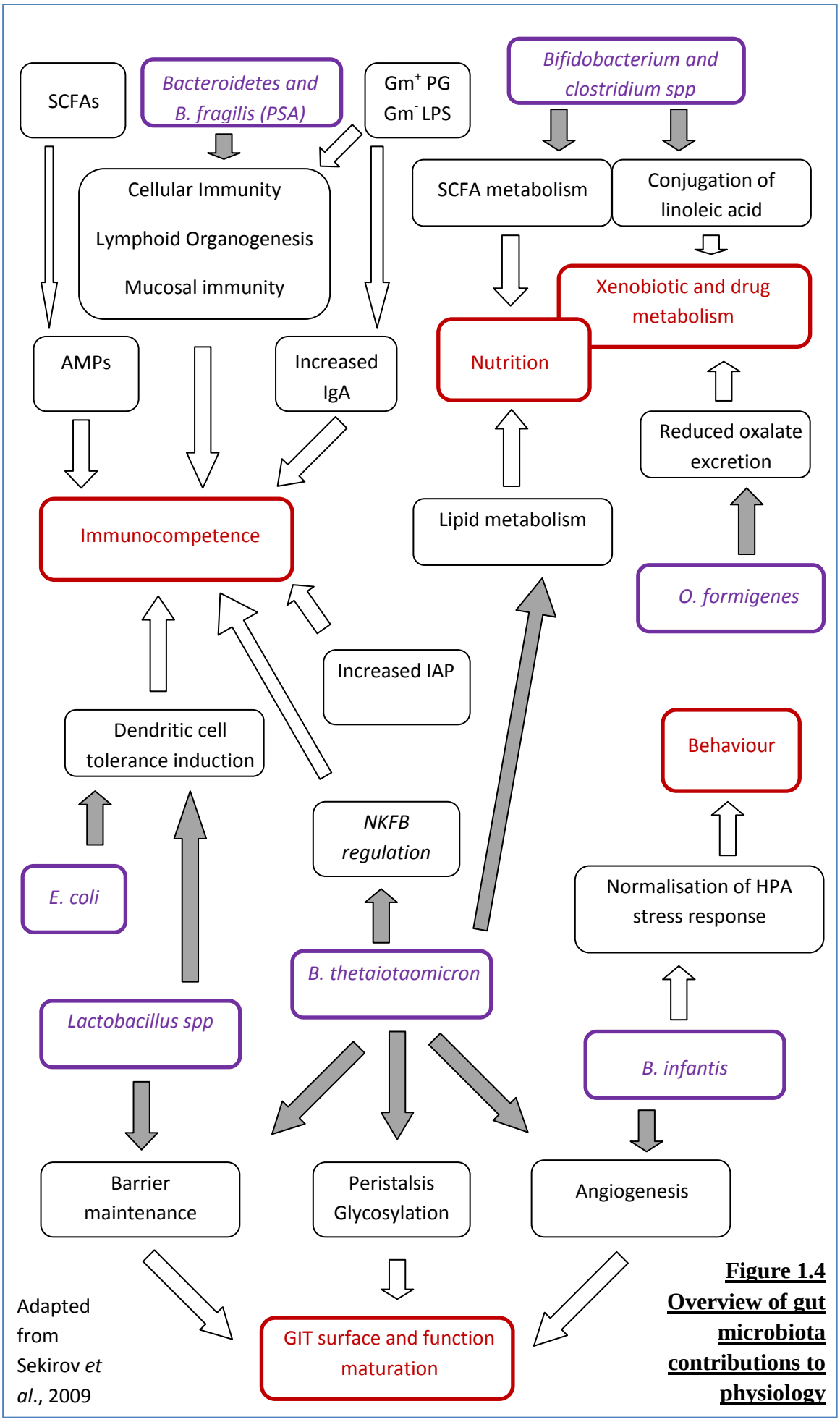
Figure 1.3 Diagrammatic representation of the mucosal immune system



Adapted from O'Hara and Shanahan, 2006, and Macpherson and Harris 2004

Figure 1.3 gives an overview of interactions of bacteria with the mucosal immune system. The mucus layer overlying the IECs (intestinal epithelial cells) is stratified such that the IEC apical layer is relatively free of bacteria while the luminal layer and the surface above it are heavily colonised (Macpherson and Harris, 2004). PRRs (pattern recognition receptors) such as TLRs and NLRs interact with commensal ligands to promote defensin (Ganz, 2003; Cario *et al*., 2007), IgA (Harrison and Maloy, 2011) and mucin release (Deplancke and Gaskins, 2001; Cario *et al*., 2007), while it is also likely that commensals have evolved to modulate certain immune interactions either at the effector stage or subsequent to the host response. Epithelial PRRs may also have a minor role in the process of antigen presentation to dendritic cells (Nagy-Szakal and Kellermayer, 2011). Intestinal bacteria interact frequently with M cells in mucosal associated lymphoid tissue, M cells expressing differing PRRs to other IECs and being capable of microbial product internalisation and whole bacterial translocation (Tyrer *et al*., 2006). Macrophages in the lamina propria phagocytose any bacteria which manage to penetrate the epithelial barrier via other routes (Macpherson and Harris, 2004). M cells and macrophages interact with antigen presenting cells (APCs) or dendritic cells (which may also sample the intestinal milieu directly) before these then prime T and B cells either in the MALT itself or after travel to proximal lymph nodes (Macpherson and Harris, 2004). The B cells differentiate into plasma cells for release of immunoglobulins, particularly IgA, while the T cells mature into Th1/Th2 effector cells or regulatory cells (O'Hara and Shanahan, 2006). Commensally-primed (tolerogenic), mucosal dendritic cells are suppressive of Th1 and Th17 cells (which release pro-inflammatory cytokines such as TNF-α, IFN-γ and IL-2) but stimulate Th2 and T regulatory cells to release cytokines such as IL-4, IL-5, IL-10 and IL-12 which suppress, or at least regulate, inflammation (Nagy-Szakal and Kellermayer, 2011).

Figure 1.4 gives an overview of some of the contributions by intestinal commensal bacteria to mammalian/ human physiology and immunology. The commensals shown in purple boxes are merely representative and should not be considered the sole mediators of the effects described by the grey block arrows. Boxes in red represent a phenotype. AMP = antimicrobial peptides e.g. defensins; DC = dendritic cells; Gm^- = Gram negative; HPA = Hypothalamus-pituitary-adrenal; IAP = Intestinal alkaline phosphatase; PG = peptidoglycan; PSA = Polysaccharide A. Certain pathways are not covered in the text, but the figure is primarily intended to display the range and complexity of mutualist interactions with host biology.



Adapted from Sekirov *et al.*, 2009

Figure 1.4 Overview of gut microbiota contributions to physiology

Perhaps the most vital of the mutualist roles played by the commensal microbiota is that of colonisation resistance (Van der Waaij *et al*., 1971) whereby proliferation of pathogens is restricted through occupation of potential colonisation sites, nutrient competition, and production of antimicrobials such as bacteriocins and lactic acid (Vollaard and Clasener, 1994).

Bacterial populations of the intestine may be planktonic (luminal) or sessile (adherent to the epithelum or overlying mucus layer) with the likelihood being that considerable numbers of the planktonic community arose from sloughing of more well-established sessile colonies or biofilms (Probert and Gibson, 2002). Indeed, it may be that there is a continuous cycle of this process as envisaged for the mutualist *Bacteroides thetaiotaomicron* (Xu *et al*., 2003), which appears to alternate between glycan-rich mucosal sites and luminal nutrient-particle platforms. (Bäckhed *et al*., 2005). In this respect, the interactions are mediated solely by glycan-specific outer membrane–binding proteins, since *B. thetaiotaomicron* lacks the genetic machinery to produce adhesive organelles (Sonnenburg *et al*., 2005). Its genome is, however, rich in glycoside hydrolases and polysaccharide lyases for metabolism of complex carbohydrates, and has an extensive repertoire of transposases and integrases for chromosomal rearrangement (Xu *et al*., 2003). This indicates that substrate promiscuity, to benefit from varying nutrient source availability, and genomic plasticity to effect immune system evasion are more important as adaptive strategies for a mutualist than mere adherence to the host. The existence of planktonic populations notwithstanding, bacterial adhesins are specific for a limited range of host extra-cellular carbohydate structures (Adlerberth *et al*., 2000), so occupancy of these sites by commensal bacteria may limit access for transient and/or pathogenic bacterial species, although studies with lactobacilli suggest that the blockade is due to general steric hindrance rather than competitive antagonism at receptor sites (Coconnier *et al*., 1993). Subsequent to inital association bacteria may

also produce enzymes capable of degrading carbohydrate structures or proteins in the local environment to which other bacteria or their toxins would bind, although such enzymatic activity may unmask other receptor sites (Kontani *et al*., 1996). Indeed , a model of colonisation resistance to *Clostridium difficile* involves the sequential activity of three *Clostridium* spp. on glycosaminoglycans of the extracellular matrix: *Clostridium cocleatum* is only able to colonise the intestine subsequent to *Clostridium indolis*, the former providing a niche for *Clostridium fusiformis* which then competes for sites or nutrients with *C. difficile* (Adlerberth *et al*., 2000; Liévin-Le Moal and Servin, 2006).

Competition for nutrients is another means by which colonisation resistance is effected, the commensal microbiota acting to deplete resources that might be available to invaders, and also enhance expression by the host such that their own proliferation is further augmented, as is the case for induction of fucosyltransferase activity by *B. thetaiotaomicron*, fucosylated glycoconjugates providing a further catabolic substrate for this mutualist (Adlerberth *et al*., 2000). The metabolic processes of the commensals may also alter the microenvironment of the intestine such that it is unfavourable for competitors (Fons *et al*., 2000), in much the same way that the initial bacterial colonists of the neonate progressively reduce the environment such that it becomes more hospitable for the obligate anaerobes (Mackie *et al*., 1999). Even in this respect physical location may be of importance i.e. oxygen concentrations in the intestine are highest at the epithelial surface so barring access to these sites would disadvantage facultative anaerobes (Fons *et al*., 2000). An example of this metabolic colonisation resistance is the inhibition of growth of species such as *C. difficile* by *Lactobacillus acidophilus* both through lowering of pH and the direct inhibitory effect of lactic acid (Fons *et al*., 2000).

The commensal bacteria are also able to compete with insurgents through production of antibiotics such as colicins and microcins to which they are themselves resistant

(Liévin-Le Moal and Servin, 2006), or through induction of Paneth cells and other epithelial cells to release their own stores of anti microbial peptides such as defensins and cathelicidin (Adlerberth *et al*., 2000; Liévin-Le Moal and Servin, 2006). A particular example is *Ruminococcus gnavus*, one of the Gram-positive anaerobic commensals, which produces ruminococcin A, an AMP with efficacy against numerous pathogenic clostridia which may have a role in colonisation resisitance against *C. difficile* (Dabard *et al*., 2001).

1.4.5 Dysbiosis and disease

The indigenous microbiota, once established, and in the absence of disease states or significant environmental changes, is considered to be a temporally stable and balanced community (Zoetendal *et al*., 1998; Rajilic-Stojanovic *et al*., 2007). Although the colonic community has been described as diverse, and at finer levels of phylogenetic classification this is indeed the case, at the level of phylum the presence of just seven divisions, and only 2 of these (*Bacteroidetes* and *Firmicutes*) in significant proportions, would suggest otherwise (Backhed *et al*., 2005). According to traditional theories of the relationship between diversity and stability this should make the intestinal microbiota susceptible to disruption (McCann, 2000), with an inherent vulnerability to loss of function (Yachi and Loreau, 1999). That the system is more resilient than theoretically envisaged is attributable to the genomic plasticity of bacteria such as *B. thetaiotaomicron* (Xu *et al*., 2003), which is thus capable of a wide variety of responses to changes in conditions.

Nevertheless, disruption of the *status quo* does occur, the combination of qualitative and quantitative imbalance that causes harm to the host (Holzapfel *et al*., 1998) being termed dysbiosis (Metchnikoff, 1907). The primary causes of dysbiosis are antibiotic usage, stress, dietary imbalance, infection and chronic pathological states (Sekirov *et al*., 2009), while a number of conditions have causal or correlative links with dysbiosis

such as inflammatory bowel disease (IBD), colorectal cancer, autism, obesity, allergies and Type II diabetes (Sekirov *et al*., 2009).

The differing aetiologies of the dysbiosis lead to divergence in terms of aberrant biochemistry, physiology or immunology, but aberrant bile metabolism and decreased production of SCFAs are common, particularly after administration of antibiotics (Högenauer *et al*., 1998). In addition to being the primary metabolic substrate for colonocytes, SCFAs have roles in reabsorption of water and electrolytes in the colon (Topping and Clifton, 2001), maintenance of intestinal mucosal integrity (Peng *et al*., 2009, Ferreira *et al*., 2012) and suppression of inflammation (Galvez *et al*., 2005; Maslowski *et al*., 2009; Vinolo *et al*., 2011) so their depletion can lead to diarrhoea, increased susceptibility to infection and exacerbation of inflammatory conditions.

The amelioration of colonisation resistance is another outcome of dysbiosis, the reduction in commensal populations providing opportunities for pathogens through increased availability of both adherence sites and nutrients (Fons *et al*., 2000; Adlerberth *et al*., 2000). In conjunction with other sequelae of commensal disruption, in particular the down-regulation of inflammatory suppression (Round and Mazmanian, 2009), the intestinal interface as a whole becomes severely compromised through the loss of homeostatic mechanisms (Sekirov *et al*., 2009). It is against this background of colonic dysfunction that pathogens such as *Clostridium difficile* are able to colonise and proliferate beyond the threshold required to cause disease (Denève *et al*., 2009).

1.5 Antimicrobials and AAD

Dysbiosis notwithstanding, the focus of previous sections can legitimately lead to the appraisal of the human intestinal microbiota as an accessory organ with considerable metabolic capacity and the potential for immunomodulation (Mazmanian *et al*., 2008), while humans can thus be viewed as 'superorganisms' (Mullard, 2008). However, despite a global reduction in the burden of infectious disease (ID) during the last decade of the 20th century (Lopez *et al*., 2006), the continuing prevalence of ID means bacteria may still be regarded as the enemies in a war of attrition, with considerable arsenals at the disposal of the combatants: antimicrobials for humans; sheer numbers and virulence factors, including potent toxins, for pathogens.

Bacteria may be described as pathogenic if they are capable of inflicting damage on a host, either directly or via activation of host inflammatory reponses, while virulence is the relative degree to which this occurs (Casadevall and Pirofski, 1999). Thus, *Clostridium difficile* is considered a pathogenic species but there are non-toxigenic avirulent strains such as M-1 (Borriello *et al*., 1988), a toxigenic strain with low virulence (BAT; Borriello *et al*., 1988) and hypervirulent strains such as 027 (Cookson, 2007). Shifts from commensal to pathogen may occur through acquisition of pathogenicity islands via horizontal gene transfer or even mutations in a solitary virulence factor, as described for the *fimH* gene of a uropathogenic *E. coli* strain (Sokurenko *et al*., 1998). A distinction should be drawn between the process of infection, which may require expression of virulence factors but can be limited to colonisation and contained proliferation of the microbe without significant damage to the host (asymptomatatic infection), and pathogenesis, which implies that the infection has passed a threshold of impairment at which symptoms or signs are manifest. Symptomatic infections generally arise when pathogenic bacteria have expressed a number of virulence factors such that adherence, invasion, interaction with innate host

defences, and release of exotoxin or endotoxin (LPS) have occurred (Wilson *et al*., 2002).

The immune systems, both innate (macrophages, complement, AMPs) and adaptive (T and B cells, cytokines, antibodies), have evolved to mount a comprehensive defence against bacterial invaders but medical science has appropriated environmental anti-microbials and developed suitable synthetic compounds to augment this process (Aminov, 2010).

Sterilisation techniques (e.g. autoclaving) remove all microorganisms (vegetative cells and spores) from a surface or medium (Rutala and Weber, 2004), while antiseptics (e.g. alcohol and anilides) and disinfectants (e.g. phenols), which together may be termed biocides, are not as efficacious, particularly against highly-resistant spores (McDonnell and Russell, 1999). Antiseptics and disinfectants differ in their mode of use, the latter not being suitable for living tissue, but have similar modes of action, often in causing damage to membranes or cross-linkage of microbial DNA (McDonnell and Russell, 1999). Overuse of biocides has been postulated to have contributed to the increasing incidence of atopic/allergic conditions in the 'western' world in recent years (Okada *et al*., 2010), the 'hygiene hypothesis' proposing that exposure to antigens early in life directs the immune system to respond appropriately via Th1 cells, while Th2-mediated allergies are promoted if the system is insufficiently stimulated at this stage (Folkerts *et al*., 2000); excessive exposure to disinfectants, which may be irritant chemicals and even 'poisons' such as formaldehyde or phenols, may also predispose to multiple chemical sensitivity (MCS), characterised by headaches, nausea, respiratory problems and disorientation in response to the odour of such commonplace products as aerosols and paints (Win, 2009).

Antiseptics, disinfectants and sterilisation are employed preventively to reduce the risk of infection or contamination, but once bacteria have colonised antibiotics are

utilised to contain their spread and limit proliferation such that pathogenesis is constrained and potentially fatal conditions such as septicaemia are avoided (or reversed) (Bochud *et al*., 2001).

Alexander Fleming first observed the lytic effects of penicillium mould on staphylococci in 1929 (Fleming, 1929), a discovery which led to the extension of millions of lives over subsequent decades. The family of compounds related to the bactericidal agent of penicillium mould (penicillin) are the β-lactams, which, like cephalosporins, act through inhibition of bacterial cell wall synthesis (Walker, 2007). Other groups of antibiotics and their sites of action include: polymyxins (cell membrane disruption); aminoglycosides such as gentamicin, and tetracyclines such as doxycycline, both groups being considered as bacteriostatic but may be bactericidal at higher concentrations through inhibition of protein synthesis through interaction with the 30S subunit of the ribosome; the macrolides such as erythromycin, which are often used as an alternative to the β-lactams against Gram-positive infections (inhibition of protein synthesis through interaction with the 50S subunit of the ribosome); clindamycin which has a similar mode of action to erythromycin, and is particularly effective against anaerobes; rifamycins such as rifampicin (inhibition of RNA synthesis); quinolones such as nalixidic acid and fluroquinolones such as ciprofloxacin (bactericidal through effects on DNA gyrase); and nitroimidazoles such as metronidazole (inhibition of DNA synthesis; Walker, 2007).

The rise of antibiotic resistance in groups of pathogenic bacteria such as methycillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococcus faecium* (VRE) has led to a resurgence in use of antibiotics such as the polymyxins, which had fallen into disfavour due to toxic side-effects, but against certain Gram-negatives there are often no alternatives (Arias and Murray, 2009). The reintroduction of such antibiotics, however, may prove to be a double-edged sword, potentially

precipitating loss of current microbotic members whose contribution to gut and systemic health is yet to be fully determined (Blaser and Falkow, 2009). This appears to be the case for *H. pylori*, whose association with human populations is on the wane just as its mediation of beneficial gastric effects, such as reduction in severity of acid reflux, has become evident (Atherton and Blaser, 2009).

Whether antibiotic administration is irrevocably altering the landscape of the intestinal microbiota remains to be seen, but it is clear that these antibacterial therapeutics can cause severe perturbation of the composition (Dethlefsen *et al*., 2008) which may last for many months after cessation of treatment (Jernberg *et al*., 2007). One of the consequences of such disruptions may be development of antibiotic-associated diarrhoea (AAD) which afflicts anywhere between 5% and 25% of those undergoing treatment (Bartlett, 2002; Beaugerie and Petit, 2004). The risk factor is greatest for broad-spectrum antibiotics to which *Clostridium difficile* is resistant, such as the cephalosporins, clindamycin and amoxicillin (Beaugerie and Petit, 2004; Talpaert *et al*., 2011), but tetracyclines, erythromycin, and quinolones (such as ciprofloxacin and levofloxacin), have all been implicated in the aetiology of AAD (Bartlett, 2002). Changes in the composition of the microflora are thought to contribute to AAD through disruption of both bile acid and carbohydrate metabolism in the intestine (Bartlett, 2002), but much of the pathology may be attributable to opportunistic colonisation by *Klebsiella* spp., *Clostridium perfringens* type A, *Staphylococcus aureus*, or species of Candida (Beaugerie and Petit, 2004; Song *et al*., 2008). Where there is a history of AAD and resolution is achieved through cessation of antibiotic administration these are strong candidates as contributory pathogens (Bartlett, 2002); where there is also evidence of colitis (inflammation of the colon), and in up to 60% of cases of AAD in a clinical setting, *C. difficile* is found to be the aetiological agent (Beaugerie and Petit, 2004; Bartlett, 2002).

1.6 *Clostridium difficile*

Clostridium difficile was first identified in the 1930s (Hall and O'Toole, 1935) from the stools of neonates, when it was designated *Bacillus difficilis* as a result of its resistance to isolation, the latter aspect of its nomenclature remaining well-deserved in the face of determined attempts to develop systems of genetic manipulation (Poxton, 2005). A member of the genus *Clostridium*, this Gram-positive pathogen has been linked with antibiotic-associated diarrhoea (AAD) since the 1970s, in which case the illness is termed (CDAD) *Clostridium-difficile*-associated diarrhoea (Noren, 2006; Noren, 2010).

1.6.1 The genus Clostridium

The genus *Clostridium* is a significant member of the *Firmicutes* phylum with approximately 100 constituent species (Collins *et al*., 1994), the majority of which are Gram-positive, spore-forming, anaerobic rods, in the region of a micrometre wide and up to 20 μm in length (Willey *et al*., 2008). Aerotolerance across the genus is actually variable, species such as *C. haemolyticum* being one of the strictest obligate anaerobes, while *C. histolyticum* thrives in aerobic environments, being merely capnophilic (Minton and Clarke, 1989). All, however, are unable to utilise oxygen as the terminal electron acceptor in respiration and are thus limited to fermentative pathways, generally adopting a saprophytic existence, subsisting on decaying organic material; they are characterised by inability to perform dissimilatory reduction of sulphates, and tend to be saccharolytic and proteolytic, though catalase negative (Minton and Clarke, 1989). The majority are motile with peritrichous flagellae and display pleomorphism, whereby appearance differs at various stages in the life cycle, particularly when forming endospores (Minton and Clarke, 1989). The genus contains a number of significant pathogens (Bruggemann, 2005), including *C. botulinum* (botulism), *C. tetani* (tetanus),

C. perfringens (gas gangrene), and *C. difficile* (enterocolitis). While the genus remains relatively stable, the order *Clostridiales* is distinctly heterophyletic, and may undergo considerable reclassification in the near future.

1.6.2 C. difficile: Phylogeny, metabolism and morphology

Within the clostridial genus, *Clostridium difficile* was grouped into Cluster XI with *C. sordellii* and *C. aminobutyricum*, although its closest relative appears to be the little-known *C. mangenotii*, a resident of soils and an occasional isolate from human faeces (Collins *et al*., 1994). The cluster corresponds to the rRNA homology group II-A (Johnson and Francis, 1975) and also contained species from the *Peptostreptococcus* and *Eubacterium* genera, both of which are non spore-forming, indicative of the phenotypic and taxonomic heterogeneity of clostridial clusters (Collins *et al*., 1994). Indeed, more recent classifications on the basis of 16S sequence place this pathogen in the family *Peptostreptococcaceae* and the genus *Peptostreptococcaceae Incertae Sedis* (Ludwig *et al*., 2005).

The primary environmental niche of *C. difficile* is the gastrointestinal tract of mammals, although strains have also been isolated from water and soils (Bongaerts and Lyerly, 1994). *C. difficile* can utilise available glucose, fructose, mannitol, mannose, xylitol and other monosaccharides but not disaccharides, oligosaccharides, or polysaccharides such as starch (Aktories and Wilkins, 2000). Through use of extracellular collagenases, proteases and mucopolysaccharide hydrolases such as hyaluronidase, heparinase and β-glucuronidase it can obtain substrates such as N-acetyl-glucosamine and N-acetyl-neuraminic acid from the mucins and proteoglycans of the mucosal extracellular surface, as well as from mucopolysaccharides and peptidoglycans of other bacteria (Bongaerts and Lyerly, 1997). It shares the ability to produce these enzymes with components of the native flora such as *Bacteroides* and *Ruminococcus* (Bongaerts and Lyerly, 1997).

FIGURE 1.5: *Clostridium difficile* colonies on CCFA agar with 5% horse blood

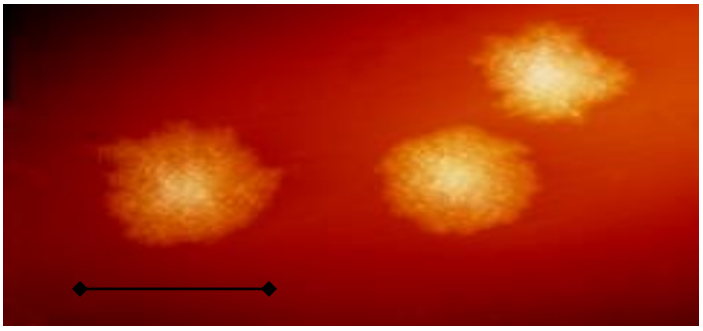


Figure 1.5 shows low resolution photograph of *C. difficile* colonies on CCFA agar (with 5% horse blood) after 48 hours of incubation. Figure displays the irregular outline of the colonies and absence of β-haemolysis with minimal α-haemolysis. Magnification is 5 X and scale bar has a length of 5 mm.

Under microscopic examination *C. difficile* is a comparatively sizeable bacterium, being up to 17 μm in length (normally about 5 μm), with a 'drumstick' morphology caused by terminal bulges (Aslam *et al*., 2005). While slow-growing, they can be cultivated *in vitro* on Brazier's CCFA medium, composed of cycloserine, cefoxitin and fructose agar with egg-yolk supplements (Brazier, 1993), in an anaerobic environment at between 35 and 37°C (Limaye *et al*., 2000). They form glossy, grey, circular colonies with rough edges which display chartreuse fluorescence under exposure to long-wave (365 nm) ultraviolet light; on blood-supplemented agar β-haemolysis is absent (see figure 1.5), as is lecithinase activity on standard CCFA, both of which form the basis of laboratory differentiation from closely-related species, while a layman's test is the characteristic 'farmyard smell', reminiscent of horse manure (Aktories and Wilkins, 2000).

These obligate anaerobes are prompted to sporulate when confronted with unfavourable environmental conditions, the spores being capable of withstanding extremes of temperature, dessication, and aerobic conditions for long periods of time

(Sebaihia *et al*., 2006), and even showing resistance to a range of disinfectants, although bleach-based cleaning agents are effective (Fawley *et al*., 2007). *C. difficile* also has a 'coat' around its external surface known as the S-layer, composed of two polypeptides which interact to form a consistent, crystalline boundary between the microbe and its environment (Poxton *et al*., 2001).

FIGURE 1.6: Scanning electron micrograph of Clostridium difficile

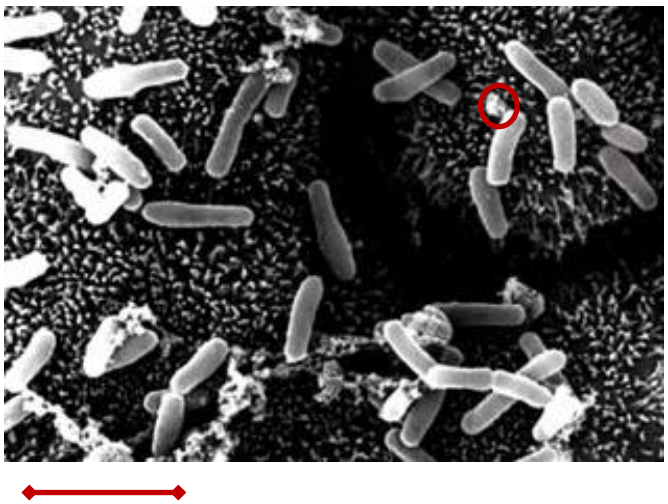


Figure 1.6 shows *Clostridium difficile* adhering to microvilli of intestinal epithelial cells, with spore visible in the upper right quadrant (circled). Magnification = 4000 X. Scale bar is equivalent to 5 µm.

1.6.3 Genetics of C. difficile

The genome sequence of the virulent *C. difficile* strain 630 was completed in 2006, revealing a single circular chromosome of 4.2 Mb and a 7.8 kb plasmid, p*CD*630 (Sebaihia *et al*., 2006). The genome is relatively low in G+C content (29%), with 3776 ORFs (Open Reading Frames), only 15% of which have been identified in the genomes of other sequenced *Clostridium*, and primarily representing essential functions (Sebaihia *et al*., 2006). A little more than 10% of the genome consists of mobile genetic elements

(MGEs), predominantly Tn5398 and 5397, encoding resistance to tetracycline and erythromycin, respectively (Sebaihia and Thomson, 2006). The high proportion of MGEs is of particular relevance as these are frequently involved in transfer of genes associated with virulence, surface structures, and interaction with the host (Wren, 2006).

The genome sequence also indicates that there are 11 rRNA operons and a similar number of clustered regularly interspersed palindromic repeats (CRISPRs), hypothesized to form a record of foreign DNA encountered (in the form of vestiges of the sequences), and possibly mediate a form of RNAi-based 'immune' response (Mojica *et al*., 2005).

There are many coding sequences associated with carbohydrate mobilisation and metabolism, potentially allowing for subsistence on a wide variety of nutrients (Wren, 2006); one particular cluster of 19 genes is concerned with ethanolamine metabolism, a potential advantage for *C. difficile* in the GI tract where such phospholipids abound as a subsidiary source of carbon and nitrogen (Sebaihia and Thomson, 2006). Another operon encodes enzymes which can decarboxylate the tyrosine degradation product *p*-hydroxyphenylacetate to produce *p*-cresol (Sebaihia *et al*., 2006), a bacteriostatic compound known to be elaborated by *C. difficile* (Sebaihia and Thomson, 2006). In addition, there are 426 sequences (11%) encoding transcriptional regulators including transcriptional antiterminators, signaling proteins and more than 40 two-component regulatory systems, all of which would permit the detection of, and response to, changes in environmental conditions (Sebaihia *et al*., 2006).

Overall, this pathogen appears highly suited to life in the intestinal tract with the ability to modify its metabolism supplemented by large numbers of phage and insertion sequence (IS) elements as well as conjugative transposons mediating antibiotic resistance (Wren, 2006), although its pathogenetic potential suggests it is not yet highly evolved in a mutualistic sense (Lederberg, 2000). In addition to this inherent

adaptability, the species also seems highly diverse, with only 40% of the genes being ubiquitous in a comparison of 8 strains (Sebaihia *et al*., 2006); for comparison, *Campylobacter jejuni* is considered to be a highly variable species with a core genome of approximately 60% (Sebaihia *et al*., 2006).

One portion of the chromosome of particular significance is the pathogenicity locus (PaLoc), a 19.6 kb region containing 5 genes which appears to be a site of considerable genomic rearrangement, although MGEs are absent from within its boundaries (Bruggemann, 2005). The genes are arranged as in Figure 1.7, *tcdA* encoding toxin A (enterotoxin), and toxin B (cytotoxin) being produced by *tcdB* (Rupnik *et al*., 2005). The *tcdR* gene was previously designated *tcdD* (Rupnik *et al*., 2005), and encodes an alternative sigma 70 factor (σ^{70}) which upregulates expression of toxins A and B (Bruggemann, 2005). The remaining genes in the locus, *tcdE* and *tcdC*, encode, respectively, a holin-like protein and a putative negative regulator of toxin expression (Rupnik *et al*, 2005). At a different chromosomal location of some strains are two further genes encoding components of a binary toxin, designated CDT but not to be confused with cytolethal distending toxin (Barth *et al*., 2004).

C. difficile is highly resistant to genetic manipulation. The Minton and Mullany laboratories are at the forefront of developing and refining techniques to create mutants for genetic analysis, a task which has so far eluded researchers, with the pathogen herein showing how richly it deserves its name (Wren, 2006).

1.6.4 Molecular toxicology

Toxins A and B (TcdA and TcdB) are part of a group of toxins known as large clostridial toxins or LCTs, the others being produced by the lesser-known species *Clostridium sordellii* and *Clostridium novyi* (Rupnik *et al*., 2005). Toxin A (308 kDa) consists of 2710 amino acids, while Toxin B (279 kDa) comprises 2366 amino acids,

with 49% identity between the two, indicating a gene duplication event (Moncrief *et al*., 1997); minimal identity is concentrated in the C-terminal region (Dillon *et al*., 1995). Each consists of an N-terminal glucosyl-transferase domain, and a C-terminal receptor-binding domain with approximately 40 repeat units (Bongaerts and Lyerly, 1994); the region sandwiched between these is a hydrophobic trans-membrane domain which may mediate translocation into the cytosol (Poxton *et al*., 2001).

FIGURE 1.7: The pathogenicity locus of *C. difficile*

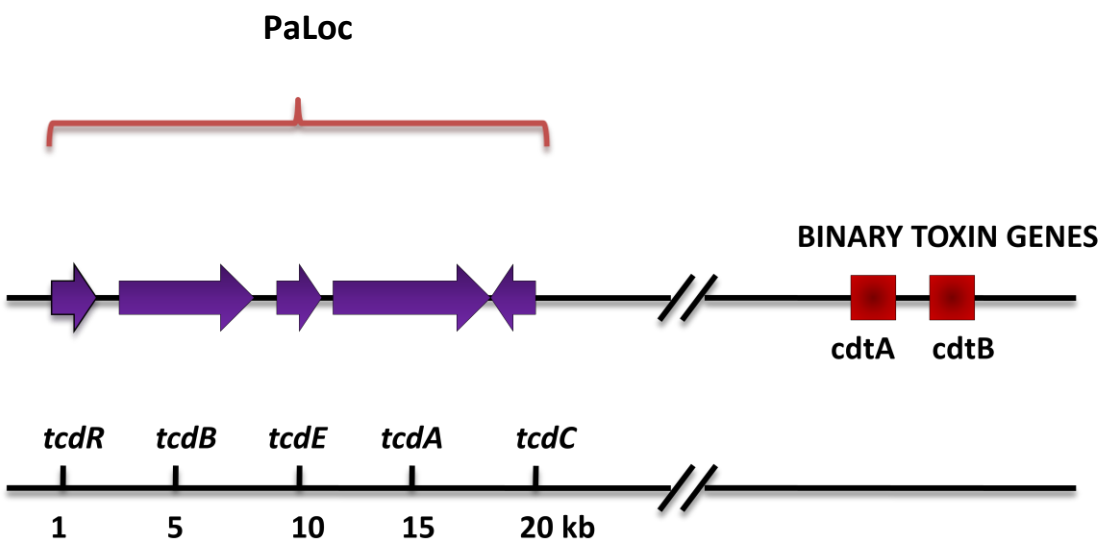


Figure 1.7 shows the 19.6 kb PaLoc including the *tcdB* and *tcdA* genes encoding the cytotoxin and enterotoxin respectively. The binary toxin genes are located elsewhere in the chromosome. The *tcdR* gene upregulates expression while *tcdE* encodes a holin-like protein and *tcdC* is a putative negative regulator (Adapted from Rupnik *et al*., 2005).

Toxin A binds to carbohydrates on the epithelial cell surface with a specific type-2 galactose moiety (Voth and Ballard, 2005), while toxin B is thought to bind to cells whose surfaces have a minimal glycocalyx (Poxton *et al*., 2001). The toxins are then endocytosed with subsequent activation in acidified endosomes before interacting with

GTPases such as Rho, Rac and CD42 (Voth and Ballard, 2005); this interaction takes the form of glycosylation of a specific threonine residue (Mylonakis *et al*., 2001), the sugar moiety being provided by UDP-glucose and resulting in transcriptional activation and condensation of actin from the filamentous F-form to the globular G-form (Dillon *et al*.,1995). The consequences of the covalent modifications are cell-rounding, disruption of tight junctions, and apoptosis (Voth and Ballard, 2005). Data from assays suggests that the cytotoxic potency of B is in the order of 1000 times that of A (Tonna and Welsby, 2005), while toxin A is considered to have greater pro-inflammatory efficacy and may be capable of triggering a neuroimmune response (Poxton *et al*., 2001).

FIGURE 1.8: Structural domains of *C. difficile* toxins

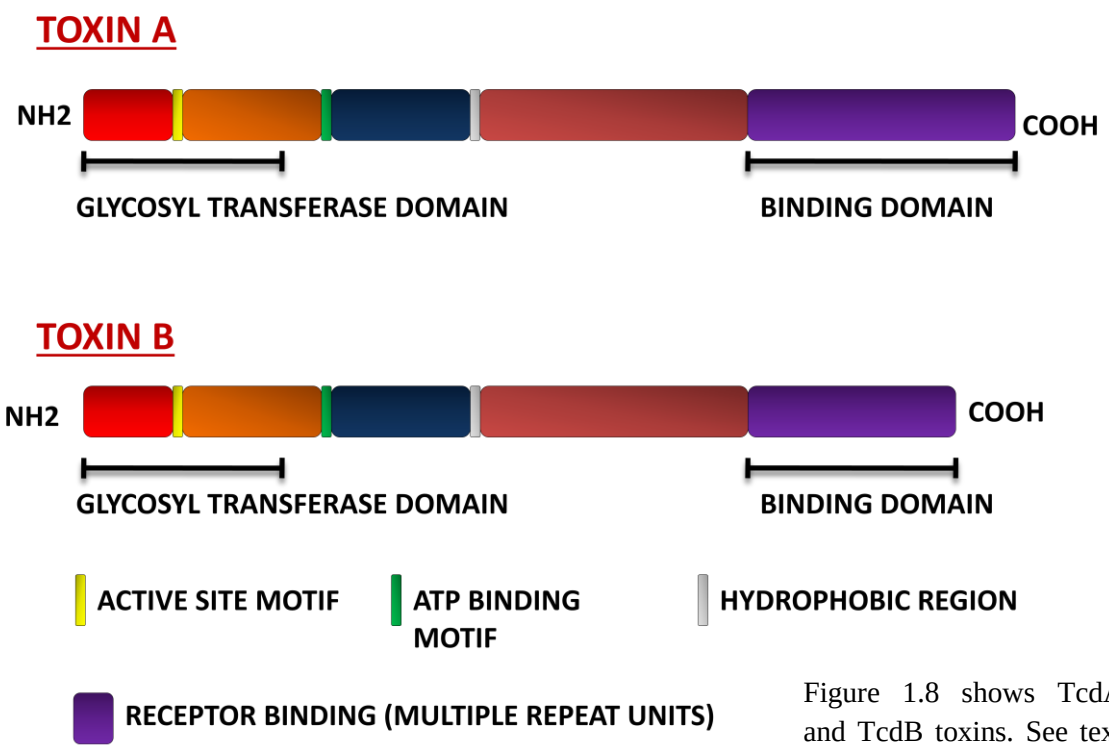
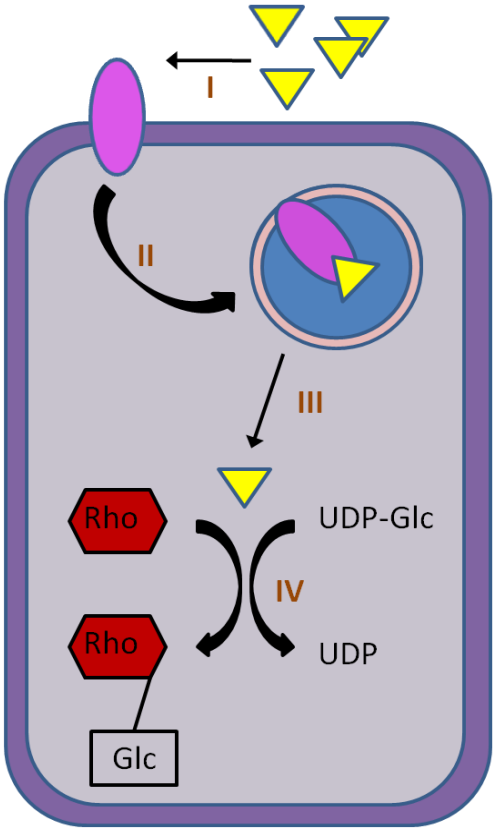


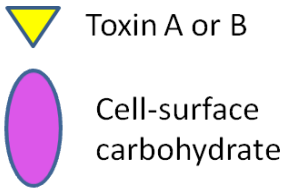
Figure 1.8 shows TcdA and TcdB toxins. See text for details.

The mode of action of CDT is bipartite in comparison to the synergy of the primary toxins (Geric *et al*., 2006). The CDTb units are synthesized as (99 kDa) precursors which are converted to active monomers (75 kDa) through proteolytic activity (Barth *et*

al., 2004). These 'activated' monomers aggregate into heptamers before binding to cell-surface components and mediating internalisation of the smaller (48 kDa) CDTa units (Barth *et al*., 2004). It is thought that the entire complex is translocated before CDTb perforates the endosomal membrane. The CDTa units then enter the cytosol where they mediate toxicity through ADP-ribosylation of G-actin, inducing cytoskeletal disorganisation and subsequent fluid loss (Geric *et al*., 2006). These effects are viewed as adjunctive to those of the primary toxins, since the binary toxin itself does not cause *Clostridium difficile*-associated disease (Barth *et al*., 2004).

FIGURE 1.9: Actions of *Clostridium difficile* toxins





I Toxins bind to carbohydrates on cell surface.

II Complex of toxin and carbohydrate is internalised via receptor-mediated endocytosis.

III Endosome is acidified through fusion with lysosome leading to activation and release of toxins into cytoplasm.

IV Toxin catalyses glycosylation of small GTP-ases leading to effects including breakdown of filamentous actin and subsequent cell death.

1.7 *Clostridium difficile*-associated disease

1.7.1 CDAD

Clostridium-difficile associated disease is essentially a spectrum of illnesses ranging from diarrhoea (CDAD), through colitis to (PMC) pseudomembranous colitis (Bongaerts and Lyerly 1997). The initial suspects in attempts to identify the cause of pseudomembranous enterocolitis were *Staphylococcus aureus* or *Candida albicans* (Baden, 1957), but by 1978 *Clostridium difficile* had been identified as the aetiological agent in the majority of cases (Bartlett *et al*., 1978).

Incidence of CDAD has risen dramatically in the past decade, and the cost to society is considerable, the disease causing significant morbidity resulting in increased duration of hospitalisation and, in some cases, mortality (Aslam *et al*., 2005). Aside from the immediate costs of treatment, difficulties in preventing spread of *C. difficile* lead to the additional economic burden of ward closures and patient isolation (Wren, 2006).

1.7.2 Epidemiology and clinical range

Clostridium difficile is the most common cause of nosocomial diarrhoea: in the UK in 2006 more than 50,000 cases were reported, representing an increase of more than 15% on the previous year (Wren, 2006), and double the figure from the turn of the millenium (Durai, 2007). In comparison to MRSA (methicillin-resistant *Staphylococcus aureus*), there are more than 3 times as many cases, and mortality is significantly higher (Wren, 2006).

Up to 30% of patients prescribed antibiotics in a clinical setting develop some form of diarrhoea (AAD or antibiotic-associated diarrhoea), of whom 10-25% will have *C. difficile* identified as the causative agent; however, 50-75% of cases of AA-colitis and more than 90% of those with AA-PMC are attributed to this pathogen (Aslam *et al*., 2005). Recent years have seen the emergence of hypervirulent strains such as

NAP1/027, responsible for the outbreak in Quebec in 2004 (Labbé *et al*., 2008), and capable of increased toxin production with concomitant higher mortality (Cloud and Kelly, 2007). In this case upregulation of toxin production may be attributable to a deletion in *tcdC* (Freeman *et al*., 2010), but pathogenic strains are being identified which produce shortened variants of the toxins or may not manufacture toxins at all (Wren, 2006).

FIGURE 1.10: Endoscopic image from colon of patient with early PMC

Figure 1.10 shows colonic mucosa in PMC with erythematous colitis and pseudomembranous foci. The initial exudate contains neutrophils and fibrin, with patches of epithelial necrosis developing into a more diffuse ulceration, possibly with an associated pseudomembrane surrounding an exudate of mucin, fibrin and leukocytes as well as the detritus of necrotised epithelial cells. Pseudomembranous plaques may be up to 2 cm in diameter and can conjoin to cover significant areas of the mucosa. (Adapted from Mylonakis *et al*., 2001)

Up to 5% of healthy adults (Mylonakis *et al*., 2001), 20% of those in hospital for 1 week, and 50% of those in hospital for more than 4 weeks are asymptomatic carriers of

C. difficile, these individuals representing a reservoir of infection (Tonna and Welsby, 2005); however, *C. difficile* should not be considered a commensal of the gastrointestinal tract (Johnson and Gerding, 1998). In instances where illness develops this will normally be limited to mild fluid-loss through diarrhoea, with possible oedema and hyperaemia of the rectum; this may be self-limiting, but resolution of infection can usually be achieved by cessation of treatment with the offending antibiotic (Tonna and Welsby, 2005). In an unspecified proportion of cases the condition will deteriorate to include erythematous colitis and a range of other symptoms including: fever, malaise, abdominal pain, bloody diarrhoea and leukocytosis (Hurley and Nguyen, 2002).

Approximately 10% of all cases of CDAD will progress to PMC, with formation of the characteristic yellowish plaques in the colon (Tonna and Welsby, 2005); a further danger in 3% of all incidences of CDAD is development of fulminant colitis, with associated high risks of toxic megacolon (dilation of the colon), intestinal perforation and death, especially if diarrhoea ceases due to ileus (Hurley and Nguyen, 2002). In 1.5% of all instances of CDAD the condition is fatal (Tonna and Welsby, 2005).

1.7.3 Multifactorial pathogenesis

A characteristic precedent of development of CDAD is antibiotic therapy, the supposition being that the normal microflora is disrupted thereby attenuating or abolishing colonisation resistance (Poxton *et al*., 2001; Cloud and Kelly, 2007). Antibiotics such as clindamycin, ampicillin, amoxicillin, and third-generation cephalosporins such as cefotaxime are most commonly implicated in creation of a permissive environment for *C. difficile* (Johnson and Gerding, 1998; Mylonakis *et al*., 2001; Stoddart and Wilcox, 2002), although anti-neoplastics such as methotrexate have also been associated with colonisation subsequent to microfloral perturbation (Riley, 1998). Recent years have also seen an increase in CDAD consequent to use of fluoroquinolones evidenced by the Quebec outbreak in 2004 where resistance was

widespread amongst the isolated strains (Cloud and Kelly, 2007). In addition, changes in antibiotic usage can precipitate CDAD outbreaks as evidenced by an increase in incidence when one hospital switched from levofloxacin to gatifloxacin, the latter being effective against an increased range of anaerobic bacteria; a switch back controlled the outbreak, although other measures including hygiene protocols also contributed (Cloud and Kelly, 2007). *C. difficile* is also frequently isolated from patients with inflammatory bowel disease and Crohn's disease; in this instance the disruption of the intestinal flora allows colonisation but it is not responsible for the underlying pathogenesis (Riley, 1998).

The cause of the disturbance notwithstanding, there are multiple favourable sequelae for the pathogen: not only do sites for colonisation become accessible through displacement of resident populations susceptible to antimicrobials, but nutrient availability is increased, especially with regard to monosaccharides and amino acids which are normally fermented by commensal species of the small intestine, while levels of bacteriocins or acids inhibitory to proliferation of *C. difficile* are diminished (Bongaerts and Lyerly, 1997).

There are contrasting views as to the timepoint of *C. difficile* contamination which allows exploitation of the intestinal conditions created by therapeutic intervention (Johnson and Gerding, 1998). The conventional view was that *C. difficile* spores are ubiquitous in a clinical setting, contamination occurring prior to administration of antibiotics, with subsequent germination of spores in the absence of the regulatory bacterial community leading to unchecked growth of *C. difficile* and disease; this implies that asymptomatic carriage is an intermediate stage in development of CDAD and that virulence of the strain is the sole determinant of progression (Johnson and Gerding, 1998). A contrasting perspective is that *C. difficile* is acquired subsequent to antimicrobial therapy with the clinical outcome somewhere on the spectrum of

asymptomatic carriage to PMC (Johnson and Gerding, 1998). Evidence that even the most virulent strains are as likely to cause asymptomatic carriage as CDAD suggests that the second model is more accurate (Johnson and Gerding, 1998), with establishment of the former generally protecting against pathogenesis (Shim *et al*., 1998). Thus, other extraneous factors must influence the clinical course (Hurley and Nguyen, 2002). In this respect, host susceptibility is a likely contributor, particularly febrility resulting from other conditions and immune status, while the extent, timing and specifics of intestinal commensal disruption may play an important role (Johnson and Gerding, 1998). However, exposure to the pathogen must occur, most commonly by ingestion of spores, with their subsequent passage through the acidic gastric environment and germination in the anaerobic environment of the distal large intestine (Mylonakis *et al*., 2001). Both toxinogenic and non-toxinogenic strains are capable of colonisation but only the former usually cause disease, the pathology normally being dependent on toxin elaboration, not the mere presence of the organism (Bongaerts and Lyerly, 1997).

Initial events in colonisation are undefined but probably involve interaction of the fimbriae with the mucosa and release of proteases such as chondroitin-4-sulfatase to degrade constituents of the extracellular matrix (Bongaerts and Lyerly, 1994); once established, production of substances such as *p*-cresol, ammonia, and volatile fatty acids (such as isocaproic acid), which are inhibitory for many other bacterial species, create an effective exclusion zone such that *C. difficile* becomes increasingly resistant to displacement by recolonisation of the normal microflora (Bongaerts and Lyerly, 1997).

Subsequent to colonisation *C. difficile* may begin to elaborate the toxins encoded on the PaLoc. Temperature and other aspects of the environmental milieu, such as the availability of specific nutrients, can all influence production of the *C. difficile* toxins e.g. *tCDR* is induced as cells enter the stationary phase, possibly by a CodY-like

regulator as found in *Bacillius subtilis* (Bruggemann, 2005). In addition, stress and limitation of biotin levels have been shown to increase expression (Voth and Ballard, 2005). Synthesis of the major toxins, though, is restricted to bacteria that have made the transition from exponential phase to stationary phase and sporulation (Poxton, 2005).

The majority of pathogenic *C. difficile* strains produce both of the toxins A and B, their actions being synergistic, with A mediating primary damage to mucosal tissue and B exacerbating the pathogenesis (Bongaerts and Lyerly, 1997), although pathogenic A^-/B^+ strains have been isolated (Limaye *et al*., 2000). The binary toxin is infrequently expressed and causes fluid loss but no apparent epithelial cell damage, and has not been shown to cause the disease in isolation (Cloud and Kelly, 2007). In addition to those previously mentioned, the total amount, and type, of toxin produced are further influences on the clinical outcome (Aslam *et al*., 2005).

It is of interest that up to 70% of neonates and 65% of children under 1 are asymptomatic carriers, the bacteria achieving colonisation due to the absence of a protective flora (Riley, 1998), with high levels of toxin being produced in the absence of progression to CDAD (Bongaerts and Lyerly, 1997). Some mechanism must therefore prevent pathogenesis. This may be the lower degree of mucin degradation in the infant colon (Bongaerts and Lyerly, 1997), but it is thought that the intestinal cell-surface structures of infants are represented by simple carbohydrates, in comparison to those of adults which consist of more complex multibranched structures; thus, receptors for toxin A, which binds exclusively to a glycoprotein with an α-linked galactose on the brush border of the epithelium (Mylonakis *et al*., 2001), may be scarce in the infant intestine (Bongaerts and Lyerly, 1997). Combined with relative immaturity of the associated intracellular G-protein systems, this may go some way to explaining comparative infant resistance (Bongaerts and Lyerly, 1994). Transitions in the resident microflora with increased age may lead to its subsequent exclusion, as *C. difficile* is

rarely detected in the absence of disease in adults (Bongaerts and Lyerly, 1994). The prevalent exposure during infancy may lead to acquisition of partial adaptive immunity, with more than 60% of adults carrying antibodies to *C.difficile* toxin A and low levels of serum IgG and mucosal IgA correlating with increased disease severity and increased probability of relapse (Mylonakis *et al*., 2001; Hurley and Nguyen 2002).

FIGURE 1.11: Effects of *C. difficile* toxins on intestinal epithelium

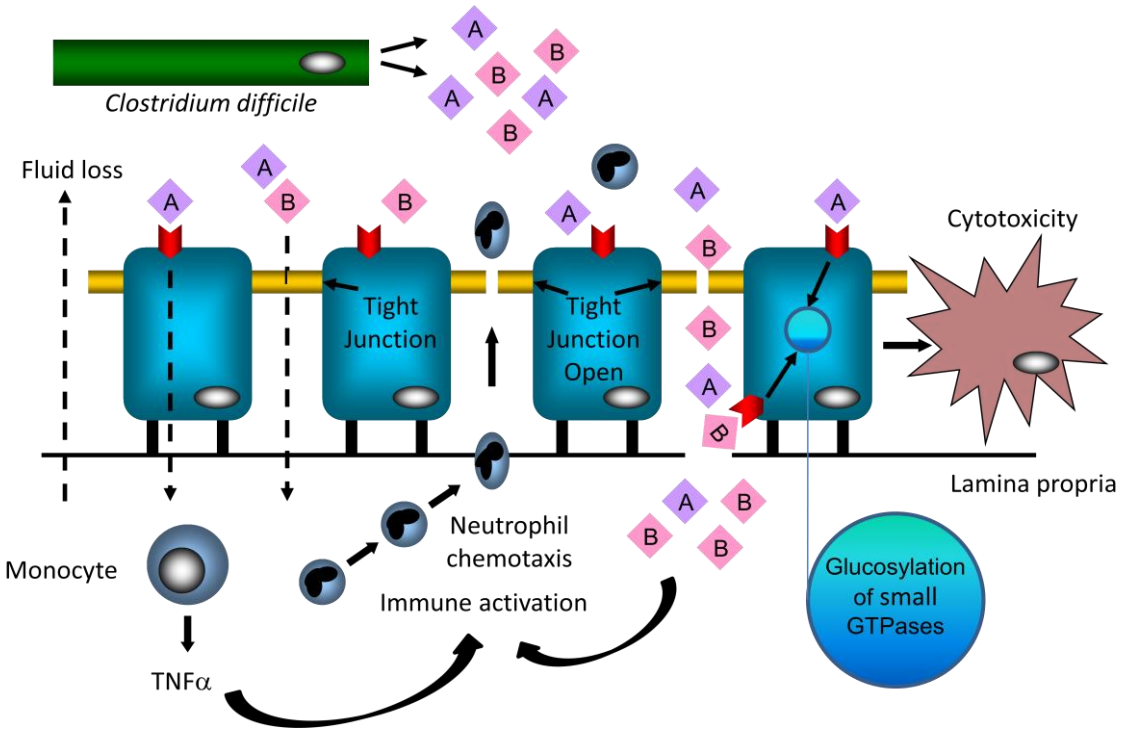


Figure shows schematic effects of Toxins A and B and exacerbation of fluid loss by excessive neutrophil activation and cytokine release. Note also putative basolateral entrance of toxin B subsequent to initial damage to tight junctions of the less potent toxin A. See figure 1.8 for more detail on glucosylation of small GTPases, and text for further details

The intracellular molecular toxicology of the toxins has been described previously (section 1.6.4), but pathogenesis of CDAD is not limited to the effects of glucosylation of low molecular weight Rho-type proteins of epithelial cells. This causes disruption of

protein synthesis and cytoskeletal organisation leading to loss of tight-junction integrity and death of cells. Effects of the toxins have primarily been studied on ileal tissue, where the microvillar brush border of the villous tips is denuded, probably due to cytoskeletal disruption, with eventual degradation of the tip itself (Bongaerts and Lyerly, 1994). Despite the lack of villi in the large intestine, invaginations caused by glandular crypts create a population of apical cells analogous to those of the small intestinal villi, with the microvilli characteristic of absorptive cells; adhesion of these cells to the basement membrane is reduced by the actions of toxin A in particular, thereby damaging the epithelial barrier (Tonna and Welsby, 2005). Toxin A also contributes to the viscous, bloody diarrhoea via effects on the endothelial cells of capillaries which initiate oedema, thereby enhancing the fluid loss resulting from tight-junction disintegration (Bongaerts and Lyerly, 1994). In essence, toxin A causes sloughing of epithelial cells and a generalised increase in intestinal permeability.

Both toxins are also highly pro-inflammatory, many of the symptoms of CDAD being attributable to excessive stimulation of the inflammatory cascade. They directly stimulate production of I-CAM 1, and interleukins such as IL-8, by intestinal monocytes in the lamina propria (Durai, 2007). It is of interest that a polymorphism in the IL-8 gene has recently been identified as a risk factor for development of CDAD (Cloud and Kelly, 2007). Toxin A is also known to cause direct activation of phospholipase A_2 through microfilament perturbation, leading to degradation of arachidonic acid to the eicosanoids (prostaglandins and leukotrienes) which can promote the release of cytokines including TNFα, a factor capable of inducing septic shock, (Bongaerts and Lyerly, 1994); Toxin B is even more potent in this respect, eliciting cytokine release comparable to that caused by LPS of Gram-negative bacteria (Bongaerts and Lyerly, 1994). These eicosanoids (particularly LTB_4), cytokines, and toxin A itself then act as chemoattractants for neutrophils leading to exacerbation of inflammation, primarily

through further disruption of tight junctions and capillary integrity caused by the invading leukocytes, and actions of reactive oxygen intermediates produced by the activated granulocytes (Bongaerts and Lyerly, 1994; Voth and Ballard, 2005).

The pathophysiology of colitis is thus characterised by the aggregation of large numbers of neutrophils in the colonic mucosa in response to release of cytokines induced by the toxins, with subsequent increases in fluid loss and necrosis (Poxton *et al*., 2001), supplemented by the direct cytotoxic effects of the toxins themselves (Voth and Ballard, 2005).

The extent of the inflammatory response initiated by the toxins of *C. difficile* essentially determines the progression and prognosis of the disease. Should pseudomembranous plaques develop then they are essentially explosions of fibrin, mucus, and leukocytes from the glandular crypts where the debris of inflammatory necrosis accumulates, numerous exudates coalescing to form the macroscopically visible structures which may overlie erosions and abcesses (Cotran *et al*., 1989).

1.7.4 Diagnosis and treatment

Diagnosis of CDAD is based on observation of the symptoms and detection of the pathogen, the latter attempted in cases of diarrhoea preceded by antibiotic usage in the preceding 2 months or admission to hospital within the past 72 hours. Diagnostic methods for *C. difficile* include culture, detection of organism-specific glutamate dehydrogenase (latex agglutination), the cytotoxicity assay, detection of A and/or B toxins by enzyme-linked immunoabsorbent assay (EIA), and, increasingly, PCR (Limaye *et al*., 2000; Durai, 2007). The cytotoxicity assay, whereby cultured fibroblasts are exposed to preparations of toxin from stools, is the most sensitive test but takes 48 hours to complete (Cloud and Kelly, 2007); the EIAs are more rapid and relatively inexpensive but sensitivity is about 10-fold less (Durai, 2007). Despite the move away from culture in recent years due to the availability of rapid toxin A+B kits, a return to

this practice has been advocated since it is possible for toxin-negative stools to occur in patients with CDAD, subsequent culture of *C. difficile* revealing that the isolates are actually toxin positive (Poxton, 2005).

Leukocytosis, an increase in the numbers of serum polymorphonuclear leukocytes, is often detected (as in many infective conditions) but may also be evident in faecal samples (Mylonakis *et al*., 2001). Complicated CDAD is associated with detection of high leukocyte counts (>20000) and high creatinine levels (Cloud and Kelly, 2007). Computerised tomography (CT) scans may show mural thickening in areas of oedema, while sigmoidoscopy reveals erythema and pseudomembranous plaques of between 1 and 2 cm in diameter in PMC (Durai, 2007).

Clinical laboratory differentiation of epidemic strains can be achieved through ribotyping, whereby variation in the intergenic space between the 16S and 23S rRNA genes is analysed through PCR (O'Neill *et al*., 1996), although the greatest discrimination is achieved by pulsed field gel electrophoresis (PFGE) of *SmaI*-digested DNA (Bidet *et al*., 2000). Interestingly, ribotyping has also revealed identical strains across mammalian species boundaries, indicating that zoonotic transmission may be possible (Arroyo *et al*., 2005; Wren, 2006), especially when one considers the habits of domestic animals, our close relationship, and the potential for *C. difficile* to persist in soil subsequent to sporulation (Riley, 1998).

Patients with suspected CDAD should be isolated and antibiotic treatment terminated, while stringent hygienic protocols should be implemented to prevent spread of infection (Aslam *et al*., 2005). Administration of fluids and electrolytes may be necessary to compensate for diarrhoea and such measures are effective in 25% of cases (Mylonakis *et al*., 2001).

Where required, the antibiotics of choice for treatment of CDAD are metronidazole or vancomycin hydrochloride (a glycopeptide antibiotic), the latter generally avoided due

to higher cost and the dangers of exerting selective pressure on vancomycin-resistant enterococcal bacteria (Noren et al., 2010). Vancomycin doses of 125-250 mg or metronidazole of 500 mg (both at 4 administrations/day) are effective in up to 90% of cases with a similar recurrence rate of approximately 15% (Aslam *et al*., 2005), the course of treatment generally being 14 days (Mylonakis *et al*., 2001). Vancomycin should be used in pregnant patients (Aslam *et al*., 2005), and is the drug of choice if metronidazole is ineffective within 48 hours, or if CDAD has progressed to septic shock or toxic megacolon (Cloud and Kelly, 2007). Bacitracin, teicoplanin and fusidic acid may also be effective and cholestyramine, a resin which binds *C. difficile* toxins, may be used as an adjunctive therapy (Mylonakis *et al*., 2001), although fusidic acid has been associated with an increased incidence of relapse/recurrence (Aslam *et al*., 2005). Other newer drugs include rifaximin and nitazoxanide, while rifalazil may soon be introduced (Cloud and Kelly, 2007). All anti-motility agents (such as loperamide and diphenoxylate) should be avoided and IV metronidazole (500mg/100ml saline) may be necessary in instances of toxic megacolon or ileus (Aslam *et al*., 2005). Severe complications such as toxic megacolon and sepsis may require colectomy (Durai, 2007).

Relapse or recurrence occurs because the antibiotics cannot eliminate spores, and may induce expression of virulence factors such that the pathogen is most toxinogenic towards the end of a course of antibiotics, proliferating significantly once treatment is stopped (Tonna and Welsby, 2005). Relapse may result from incomplete extirpation of the microorganism in the first instance (Johnson and Gerding, 1998) or from reinfection with a different strain (Riley, 1998), now thought to be the cause in 50% of cases (Aslam *et al*., 2005). A large response in terms of anti-toxin A IgG levels during the initial infection is associated with a lower risk of developing recurrent diarrhoea (Tonna and Welsby, 2005), while febrility, increased age and extended duration of hospitalisation increase the likelihood of relapse (Aslam *et al*., 2005).

Relapses can be treated with further courses of the original antibiotics, sometimes as tapered or interrupted regimes thus allowing spores to germinate with subsequent killing of vegetative cells, while pooled human IgG (200-500 mg/kg) can be used in refractory cases and immunocompromised patients even if commercial globulin preparations may not contain antibody to the toxin (i.e. the antibody to some other epitope is cross-reactive) (Aslam *et al*., 2005). Rectal infusions of faecal material subsequent to total gut lavage in attempts to restore the intestinal flora are another option with anecdoctal evidence of success (Johnson and Gerding, 1998), although such therapies could also transmit other undesirable organisms and are inherently unappealing (Aslam *et al*., 2005).

Another approach of particular interest is administration of probiotics containing organisms such as *Lactobacillus rhamnosus*, *Lactobacillus acidophilus*, *Saccharomyces boulardii*, and various bifidobacteria (Cloud and Kelly, 2007), the majority of which produce proteases capable of digesting the toxins of *C. difficile* and which can resist the low pH of the gastric environment (Tonna and Welsby, 2005). Use of *Saccharomyces boulardii*, however, may lead to *S. cerevisiae* fungaemia and is thus contraindicated in the immunocompromised (Cloud and Kelly, 2007). One study showed that growth of all *C. difficile* strains could be inhibited by certain lactobacillus species, particularly *L. paracasei* and *L. plantarum* species; the effect was generally strain specific with the more antibiotic-resistant and toxigenic strains found to be more susceptible to lactobacilli (Naaber *et al*., 2004). However, as yet, benefits from clinical trials have not proven statistically significant (Aslam *et al*., 2005).

1.7.5 Clostridium difficile and the microbiota

Exposure to *C. difficile* in nosocomial environments arises from the persistence of spores, which are highly refractory to many biocides and cleaning techniques (Fawley at al., 2007). The use of broad spectrum antibiotics such as cepahalosporins, to which *C.*

difficile is resistant, attenuates the ability of the intestinal microbiota to prevent colonisation and proliferation of the pathogen (Stoddart and Wilcox, 2002), thus fulfilling the second of the conditional prerequisites for development of CDAD.

Mechansims of colonisation resistance to *C. difficile* are incompletely established so it remains possible that this is a non-specific effect, mediated through occupancy of colonisation sites, or nutrient utilisation, by any of a broad number of bacterial groups. However, it is tempting to speculate that the process depends on antagonism of *C. difficile* by particular members of the standard intestinal microbiota.

Studies conducted into antagonism of *C. difficile* by total flora supplemented with antibiotics have suggested that certain groups are indeed responsible (Wilson and Freter, 1986) and that Gram-negative anaerobes, such as *Bacteroides* spp may be candidates (Wilson *et al*., 1981). In contrast, some evidence suggests that mere occupation of colonisation sites may be sufficient (Borriello and Barclay, 1985), while both production of lactic acid (Rolfe *et al*., 1981), and competition for metabolic substrates such as amino acids (Yamamoto-Osaki *et al*., 1994) or carbohydrates (Wilson and Perini, 1988; Borriello, 1990) have been postulated as mechanisms of CR to *C. difficile* as opposed to elaboration of a specific inhibitory substance (Borriello, 1990).

So the precise mechanism of CR may be unclear but these investigations serve to strengthen the concept of countering *C. difficile* infection on a probiotic level (Surawicz, 2003). The contribution of host defence (primarily in the form of anti-Tcd immunoglobulins) is significant in determining whether infection will be asymptomatic, or where on the spectrum of CDAD severity it will lie, but identification of the bacterial taxa effecting CR would be of assistance to clinicians with regard to both prophylaxis and treatment.

1.8 Metagenomics

1.8.1 Systems approaches

The past two decades have seen the growth of a ‘systems’ approach to molecular biological science, a move from the individual to the whole, fomented by technological advances and the burgeoning wealth of data available (Raes and Bork, 2008). This change of emphasis has given birth to the disciplines of genomics, metabolomics, microbiomics, metagenomics and even meta-transcriptomics. While gross characterisation of the microbiotic community in terms of phylogenetic clades or OTUs cannot quite be ascribed the '-omics' status, it remains an approach employed as a prelude to, or complementary to, more detailed microbiomics investigations. As such, application of techniques such as culture (Moore and Holdemann, 1972), cloning and sequencing (Eckburg *et al*., 2005), TGGE (Muyzer and Smalla, 1998), DGGE (Donskey *et al*., 2003), T-RFLP (Li *et al*., 2007), RISA (Kotlowski *et al*., 2007), FISH (Sekirov *et al*., 2008) and qPCR (Palmer *et al*., 2007) to description of the microbiota no longer provide either the level of resolution or the requisite throughput.

The Human Microbiome Project (HMP) is a multidisciplinary research undertaking to be conducted over the next few years at the cost of tens of millions of dollars (Turnbaugh *et al*., 2007). The objective is to characterise the microbiota associated with humans at all body sites and correlate this with the metagenome, metabolome, and metaproteome (Turnbaugh *et al*., 2007). In essence, it can be considered the next logical step in the human genome sequencing project (Venter *et al*., 2001), if we are, as suggested, 'superorganisms' (Mullard, 2008).

To achieve the aim of the project will require the deep sequencing of a multitude of anatomical sites from numerous subjects, as well as the subsidiary sequencing of a plethora of bacterial genomes (Turnbaugh *et al*., 2007). For this purpose the second-generation of sequencing technologies has been selected, although the multi-regional

nature of the undertaking means that other techniques, such as those listed above and specialised microarrays will be utilised to supplement the sequencing data (Turnbaugh *et al*., 2007).

1.9.2 Microarrays

Array technology is fundamentally based on the complementarity of the DNA duplex (Southern *et al*., 1999), but can be viewed as a development of observations made more than 40 years ago. DNA was found to associate with nitrocellulose membranes (Gillespie and Spiegelman, 1965), with this means of 'fixing' nucleic acids in an accessible fashion eventually evolving into 'blotting' procedures for analysis of heterogeneous populations of DNA (Southern, 1975). The subsequent progression of the technology was through nylon-filter-based screening of clone libraries, and gridded libraries on microtitre plates (Lander, 1999). By the late 1980s substrates such as glass were superceding membranes and gel pads as the immobilisation medium, allowing for eventual development of high density microarrays composed of thousands of individual features (Southern *et al*., 1999). The benefits of glass are manifold but include improved image acquisition and definition of feature location, along with the potential to mathematically model the kinetics of hybridisation (Southern *et al*., 1999). Traditional glass slides accommodated in the region of 20,000 features (Lander, 1999), with up to 0.1 pmol of DNA per square millimetre (Southern *et al*., 1999); modern robotics and inkjet technologies now allow for 'spotting' of DNA in 100-200 μm regions (with spacing being equivalent) such that 20,000 'features' can occupy an area of approximately 1 cm^2 (Jares, 2006), each feature containing in the region of 20 pg, the result of arraying approximately 1 nl (Cho and Tiedje, 2002). The most commonly used platform (substrates for immobilisation of oligonucleotide probes) remains glass, although others, such as microbead and plasmon resonance, have been investigated (Pozhitkov *et al*., 2007).

In general, a microarray is designed such that oligonucleotides are covalently bound to activated groups on the glass surface at the 5' end, spacers often being incorporated to reduce steric hindrance (Bodrossy and Sessitsch, 2004). The oligonucleotides are designed such that their perfect match (PM) duplex melting temperatures (T_ms) will coincide, either by adapting their lengths or through incorporation of tertiary amine salts into hybridisation buffers (Bodrossy and Sessitsch, 2004). Fluorescently-labelled targets from environmental/clinical samples are then used to challenge the array, ideally resulting in hybridisation of sequences which are perfectly complementary; to reduce mismatch (MM) binding, hybridisation is terminated by performing a succession of increasingly stringent washes in solutions of decreasing salt concentration (Bodrossy and Sessitsch, 2004).

Scanning equipment is then utilised to detect hybridisation of complementary interrogatory nucleic acids with intensity of output signal being considered dependent on concentration in the sample (Pozhitkov *et al*., 2007). Probe specificity and sensitivity are paramount considerations, with sensitivity being essential to elicit a signal-to-noise ratio which can be identified over background (Pozhitkov *et al*., 2007), while specificity is represented by stringency of binding and can be controlled through variations of hybridisation temperature and salt/buffer concentration, particularly during washing subsequent to hybridisation termination (Jares, 2006).

The array usually comprises the probes of known sequence, either cDNA from libraries or oligonucleotides synthesized *in situ* through the use of photolithography (Cho and Tiedje, 2002); the tester or target sample is commonly a population of cDNAs reverse transcribed from a population of mRNAs and incorporating dyes (Schena *et al*., 1995; Jares, 2006). The majority of array studies involve immobilisation of nucleic acids representative of genomes (Lander, 1999), such that interrogatory samples can be utilised to screen expression of genes, particularly with regard to modulations in

conditions such as cancer, or caused by toxic or pharmacological agents (Cho and Tiedje, 2002; Pozhitkov *et al*., 2007).

Although expression studies have been the predominant employment of the technology, arrays have proven adaptable to other forms of investigation including: analysis of sequence variations, particularly single nucleotide polymorphisms (SNPs); investigation of DNA-protein interactions through chromatin immuno precipitation (ChIP); comparisons of DNA copy number through genomic hybridisation; characterisation of DNA methylation patterns; and potential tumour gene identification through nonsense-mediated mRNA decay (NMD) inhibition (Jares, 2006).

Since 1998 the number of papers reporting on the use of microarrays for microbial identification has increased more than tenfold (Pozhitkov *et al*., 2007), the technology providing the possibility of high-throughput resolution of thousands of nucleic acid molecules in a single experiment, limitation in terms of microbial analysis being the specificity of appropriate probe sets (Bodrossy and Sessitsch, 2004; Harmsen *et al*., 2002). The approach involves creation of arrays with features representing the SSU rDNA of the majority of the bacterial domain (DeSantis *et al*., 2007; Palmer *et al*., 2006), or subsets of the population of particular interest, e.g sulphate-reducing prokaryotes (Loy *et al*., 2002).

One investigation employed *in situ* photolithographic synthesis which allows for greater density of features, each spot being in the region of 300 μm^2 and incorporating approximately 3×10^6 molecules (De Santis *et al*., 2007). The resultant 297851 25-mer probes on the array were specific to the SSU rRNA of more than 800 prokaryotic subfamilies, or represented paired mismatch probes for a subfamily which were devoid of the potential to hybridise specifically to other OTUs (De Santis *et al*., 2007). The targets were heterogeneous communities from a variety of environments, fragmented and labelled prior to hybridisation. The arrays incorporated a set of controls to assess

the fragmentation, labelling and hybridisation using known sequences 'spiked' into the samples at various points to which the controls were complementary. In addition, identical samples were analysed in parallel by a cloning and sequencing approach resulting in more than 1300 sequences. Their assessment was that 8% of the sequenced clones would not be identified by the array, despite the spotting of nearly 300,000 probes; conversely, phyla which were not identified through sequencing were detected by the array (De Santis *et al*., 2007). The 16S array approach allowed for greater overall coverage and higher throughput than traditional clone sequencing, which is time-consuming and costly, even for the analysis of 1000 sequences which is the typical number in a clone library (De Santis *et al*., 2007). However, it is clear that novel species cannot be identified in this manner, while quantification of the various community members is also precluded (De Santis *et al*., 2007).

Another study also utilised sets of probes specific to a particular phylogenetic group, although these were 40-mers with an additional 10 nucleotide poly-T linker to distance the binding sequence from the platform (Palmer *et al*., 2006). The approach investigated the microfloral composition of colonic tissue from human subjects and employed a comparative schema whereby tester samples incorporated the Cy5 dye while quantified reference samples were linked to Cy3 (Palmer *et al*., 2006). The relative abundance of species could then be calculated from the Cy5/Cy3 intensity ratio of each probe set (Palmer *et al*., 2006). The array approach allowed for detection of species not evidenced by parallel sequencing controls, although the results were generally in accordance with those of previous studies; the drawbacks, however, were identical to those of the 'universal' 16S array described above (Palmer *et al*., 2006; De Santis *et al*., 2007).

Two final microarray systems for characterisation of microbes are the Human Intestinal Tract (HITChip) Chip (Rajilić-Stojanović *et al*., 2009) and OFRG (Valinsky *et al*., 2002). The former utilised database sequences from the V1 and V6 regions of the

SSU rRNA to design a set of more than 4800 probes for immobilisation on a custom array, allowing for differentiation of more than 1100 phylotypes (Rajilic´-Stojanovic´ *et al*., 2009). OFRG (Oligonucleotide Fingerprinting of rRNA Genes) adopts an inverse hybridisation approach in that the sequences to be identified are immobilised to the solid support and are then interrogated with the probes of known sequence (Valinsky *et al*., 2002). This proceeds in an iterative fashion, the hybridised probe being stripped from the array surface using a heated buffer, before interrogation of the array with the subsequent probe in the set (Valinsky *et al*., 2002).

1.8.3 Second generation sequencing

Massively parallel sequencing technologies such as 454 (Roche), Solexa (Illumina) and SOLiD (ABI) have significant potential for the analysis of complex microbial communities, not least due to their economic advantage over Sanger sequencing (Sogin *et al*., 2006), but primarily due to the sheer volume of data they can provide and the high-throughput capability of the platforms (Ansorge, 2009)

The essence of 454 is the parallel analysis of tens of thousands of clonally amplified products, each of the 1.6 million wells of a sequencing plate containing a bead which has previously undergone emulsion-based PCR from a single nucleotide fragment such that there are approximately 10^7 copies per emulsion droplet (Margulies *et al*., 2005). Each fragment becomes attached to its bead via a system-specific oligonucleotide which can be incorporated into primers and then forms the basis for extension and calibration (Margulies *et al*., 2005). During the actual sequencing phase, nucleotides are allowed to flow sequentially across the system; incorporation of a nucleotide causes release of pyrophospate (PP_i) which is acted on by sulfurylase and luciferase to release light which is then detected by a scanner (Ronaghi *et al*., 2000). The integral software then converts

these signals into sequence data, eventually providing approximately 1 million reads of more than 150 bp in length per chip.

The technology has thus far been used primarily for metagenomics studies of environments such as the deep sea (Sogin *et al*., 2006), but is increasingly finding application in analyses of the intestinal microbiota (McKenna *et al*., 2008; Andersson *et al*., 2008).

The Solexa system provides a greater volume of data than 454 but at the cost of shorter read lengths, in the region of 50 bp; fragments of DNA are ligated to adapters and affixed to the solid support, before formation of 'bridges' and subsequent clonally amplified 'polonies' (Ansorge, 2009). Each 'polony' then forms the basis for the sequencing reaction, whereby terminator nucleotides labelled with dyes are incorporated in each round, and the dye is immediately detected. The end of a sequencing 'cycle' is determined by removal of both the terminator and the dye, thus priming the polony for subsequent incorporation (Ansorge, 2009). While the Solexa platform has been employed in a metagenomic study of the oral microbiome (Lazarevic *et al*., 2009) the shorter read lengths mean it is less suited to 16S sequencing than the Roche platform.

The third of the massively parallel sequencing technologies is the ABI SOLiD platform, which obtains sequence reads of around 35 bp in length through a system of clonal amplification, immobilisation of clonal populations on a solid support matrix, fluroescently-labelled octamer ligation, and cleavage (Ansorge, 2009). At the time of writing it is the least suitable of the three for microbiotic characterisation and has not found widespread application in the field.

1.9 Aims and Objectives

The primary aim of the research project was thus to investigate bacterial groups acting as markers for incidence of *Clostridium difficile*-associated disease in humans. The primary hypothesis was that these would be groups conferring resistance in the normal unperturbed microbiota, and extant even in individuals who had been treated with broad-spectrum antibiotics, such that colonisation and proliferation are limited to levels below a threshold for progression of infection. In essence, these bacterial groups maintain a state whereby *Clostridium difficile* does not flourish sufficiently to elaborate enough toxin to overwhelm mucosal (IgA) and serum (IgG) anti-toxin levels and initiate the positive-feedback inflammatory loop which is characteristic of acute infection. While the expectation was that these 'groups' or 'markers' would be inhibitory to *Clostridium difficile*, perhaps via basic colonisation resistance such as occupation of the intestinal niche favoured by *C. difficile* and competition for resources, it was deemed possible that the mechanism could be more complex e.g production and secretion of a bacteriocin antagonistic to the pathogen. Indeed, it was even postulated that those susceptible to infection subsequent to antibiotic treatment harbour a minor representative in their microbiota; proliferation of this taxon would be a pre-requisite for a bloom in CD sufficient to cause the disease state, perhaps by production of a metabolite which confers a competitive advantage for CD. Hypothesized mechanisms notwithstanding, the aim was to identify bacterial groups or operational taxonomic units (OTUs) whose presence or abundance changed significantly in a cohort of subjects with CDAD, compared to those with antibiotic-associated diarrhoea (AAD), or showing no symptoms after administration of antibiotics.

To achieve this would require characterisation of the microbiota of the distal intestine via classification of 16S rDNA, extracted from faecal samples and amplified using

'universal' primers. The depth of sequence assignment required to identify differences between groups of subjects necessitated the use of high-throughput techniques.

The second generation sequencing technology of 454 had already been utilised to investigate a number of microbiotic communities via analysis of community 16S sequences (section 1.8.3), so the availability of the platform at the University of Leicester (Genomics Services) permitted application of this approach. Subsidiary aims with regard to 454 were thus to establish an appropriate methodology, and optimise experimental protocols to obtain sufficient sequencing data for comprehensive analysis. In addition, while there were numerous extensive 16S databases, software and pipelines for processing large volumes of microbial ecology data were inchoate, so supplementary bioinformatic applications, algorithms, comparative indices, and statistics, would also require investigation and refinement.

Initial experiments to develop a novel array platform for analysis of microbial communities had been conducted at the University of Leicester by Dr Rob Free, Dr Colin Barker, and subsequently, Dr Brenda Kwambakana. Preliminary output suggested its suitability for further optimisation and application to the current research in collaboration with the microarray facility of the Leicester MRC Toxicology unit. The technique resembles OFRG (section 1.8.2) but extension to allow for significantly higher throughput necessitated development and optimisation of:

- a methodology for deposition of clone library representatives onto high-feature-density, glass microarrays
- techniques for hybridisation of probes to the custom arrays
- a novel, comprehensive probe set for interrogation of bacterial 16S rDNA libraries
- software for probe design and analysis of array output

The aim would then be to utilise both 454 and the array technique (OFARG) to characterise the microbiota of the same clinical samples. It was hoped that the differing approaches would provide complementary and synergistic profiles of the resident bacterial taxa such that markers for development of CDAD could be identified. Such an outcome would provide the foundation for subsequent more detailed and directed investigations, with a view to eventually enhancing therapeutic approaches to minimise the incidence of CDAD.

CHAPTER 2

MATERIALS AND METHODS

2.1 Bacterial strains and plasmids

A variety of *E.coli* strains and various other bacteria were utilised for the creation of reference clones from their 16S rDNA; these were kindly provided by Dr Richard Haigh (University of Leicester, UK). A list of species utilised for creation of reference clones can be found in Section 3.5. DNA from strains of *C. difficile* (CD 630 and NAP1/027) was kindly donated by members of Dr Martha Clokie's laboratory (University of Leicester, UK). *E. coli* DH5α and *E. coli* Top-10 (University of Leicester, Genetics lab-stock) were utilised as hosts for transformation of plasmid vectors. Long-term storage of bacteria was in 25% glycerol at -80°C.

The plasmid utilised as a vector for cloning of recombinant DNA was pGEM®-T Easy (Promega, USA), a linearised vector with single 3' thymidine overhangs. These allow for direct cloning of PCR products due to the extra deoxyadenosine residues added by certain thermostable polymerases to the 3' end during extension. The full sequence and map of the vector can be found in Appendix 1.

2.2 Bacterial culture and storage

2.2.1 Media and plate preparation for E.coli

Luria-Bertani (LB) medium (Roth, 1970) consists of 1% (w/v) bacto-tryptone, 0.5% (w/v) bacto-yeast extract and 0.5% (w/v) NaCl. LB broth for culture of *E. coli* was prepared by dissolving 10 g Bacto-tryptone, 5 g Bacto-yeast extract and 5 g NaCl in 950 ml ddH_2O followed by adjustment of pH to 7.0 with NaOH (5 M) and addition of water to 1 L. Solution was then sterilised by autoclaving at 121ºC, 15 p.s.i for 15 minutes prior to storage at room temperature until use.

To prepare Luria Bertani agar (LBA) 15 g of agar was added to 1 litre of LB broth (1.5% w/v) prior to heating of 400 ml in a microwave. Agar was then cooled to below

50˚C in a waterbath followed by addition of requisite supplements, mixture, and pouring into 90 mm diameter petri dishes. For subsequent blue/white selection procedures AIX (Ampicillin-IPTG-Xgal) plates were prepared by addition of 400 µl (1:1000 dilution) of the supplements described in Section 2.3 to 400 ml of liquefied LBA at 50˚C prior to pouring, such that final concentrations of supplements were 0.1 mg/ml ampicillin, 25 µg/ml X-Gal, and 0.1 mM IPTG. Plates were allowed to dry in the laminar flow prior to storage in the dark at 4˚C. LB-AIX 245 mm x 245 mm low-profile bioassay plates for large-scale colony picking with the QPix robot (Genetix, UK) were prepared in the same manner, 400 ml being sufficient for 2 plates. All plate preparation was conducted under aseptic conditions.

2.2.2 Cultivation and storage of E. coli

E. coli strains were routinely cultivated by streaking or spreading of confirmed glycerol stocks onto LBA plates without supplements under aseptic conditions, prior to incubation at 37˚C overnight. Liquid cultures were obtained by inoculation of 5 ml of LB broth, with subsequent incubation at 37˚C for 16 hours under agitation at 300 rpm. For indefinite maintenance of a strain liquid cultures were incubated as above for approximately 2 hours prior to assessment of OD_{600} using an Ultrospec10 Cell Density Meter (Amersham Biosciences, UK). Once the OD was between 0.4 and 0.5, indicating exponential-phase growth, the entire inoculate was centrifuged at 1200 x g for 10 minutes prior to resuspension of the bacterial cell pellet in 0.75 ml LB broth in a 2.0 ml cryotube. The resuspended pellet was then supplemented with 0.75 ml 50% (v/v) sterile glycerol prior to snap-freezing in dry ice and storage at -80 ˚C.

2.2.3 Media and plate preparation for C. difficile

Cycloserine cefoxitin fructose agar (CCFA; George *et al.*, 1979) contains proteose peptone (4% w/v), disodium hydrogen phosphate (0.5% w/v), potassium dihydrogen

phosphate (0.1% w/v), magnesium sulphate (0.01% w/v), sodium chloride (0.2% w/v), fructose (0.6% w/v) and agar (1.5% w/v). For preparation of plates, 34.55 g of CCFA (Oxoid) was suspended in 500 ml ddH_2O and gently boiled until completely dissolved, prior to sterilisation by autoclaving at 121°C and 15 p.s.i for 15 minutes. The solution was then allowed to cool to 50°C before aseptic addition of 1 vial of Oxoid Clostridium Difficile supplement together with 35 ml defibrinated horse blood (7% v/v). The supplement provides final concentrations of D-cycloserine and cefoxitin of 250 µg per ml and 8 µg per ml repectively. Mixture was then poured into sterile petri dishes and allowed to set before storage of plates at 4°C.

Brain heart infusion (BHI) broth for liquid culture of *Clostridium difficile* was prepared by dissolving brain infusion solids (1.25% w/v), beef heart infusion solids (0.5% w/v), proteose peptone (1% w/v), glucose (0.2% w/v), sodium chloride (0.5% w/v) and disodium phosphate (0.25% w/v) in 500 ml of distilled water prior to sterilisation by autoclaving and storage at 4°C.

2.2.4 Cultivation and storage of C. difficile

For culture of *C. difficile* from faecal specimens, faecal material was mixed in an approximate 1:1 ratio with industrial methylated spirit (IMS) and allowed to stand at RT for 30 mins. 'Alcohol shock' treatment acts as a pre-selection process, since non-sporing organisms should be eliminated by the addition of IMS (Borriello and Honour, 1981). Approximately 50 µl of the solution was then applied to the surface of CCFA plates, pre-equilibrated in anaerobic chambers with an atmosphere of 80% N_2, 10% H_2 and 10% CO_2. Subsequent to streaking to provide for single colonies plates were incubated anaerobically at 37°C for 48-72 hours. Colonies were selected based on morphology and yellow-green fluorescence under long wave (365 nm) ultraviolet light and transferred to Falcon™ tubes containing 10 ml BHI pre-equilibrated in anaerobic conditions. Inoculates were assessed after 12-16 hours, an OD_{550} of between 1.0 and 1.5

indicating exponential phase growth at which point 1 ml of the inoculate was transferred to a 2 ml cryotube containing 1 ml of cryofluid. Cryotubes were then stored at -80°C until required.

2.3 Chemicals, consumables, buffers and stock solutions

A comprehensive list of chemicals and consumables is provided in Appendix 2, while specifications of other laboratory equipment and details of manufacturer's and suppliers are included in Appendices 3 and 4 respectively.

2.3.1 Stock solutions

Unless otherwise stated, stock solutions, media and buffers were prepared using double-distilled water (ddH_2O), obtained from Elgastat Option2 Water Purifier (Elga, UK) or Ondeo Purite Select (Purite, UK) machines prior to autoclaving. Molecular biology grade water for dilution of probes and primers, and preparation of PCR, was from Invitrogen (UK). Where sterilisation of solutions was recommended as per established protocols (Sambrook and Russell, 2001), but autoclaving was inadvisable or impractical, this was achieved by passing the solution through 0.22 μm filter membranes. Small volumes were sterilised using acrodisc syringe filters (Pall Life Sciences) attached to 5 or 10 ml syringes (BD Plastipak); larger volumes required the use of Stericups (Millipore) and a vacuum pump (Fisher Scientific).

5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside (X-Gal) solution (25 mg/ml): X-Gal solution was prepared by dissolving 250 mg X-Gal in 10 ml N,N-dimethyl formamide (DMF) with subsequent storage at -20°C in the dark. One in one thousand dilutions of stock provided working concentrations of 25 μg/ml.

0.1 M Isopropyl-beta-D-thiogalactopyranoside (IPTG) solution: IPTG was prepared as 0.1 M solution by dissolving 238 mg IPTG in 10 ml ddH_2O prior to filter sterilisation and storage at -20°C. One in one thousand dilutions of stock provided working concentrations of 0.1 mM

Ampicillin solution was prepared as 100 mg/ml solution by dissolving 1000 mg of ampicillin sodium salt in 10 ml ddH_2O prior to filter sterilisation and storage at 4°C. One in one thousand dilutions of stock provided working concentrations of 100 µg/ml.

1 M Tris-Cl solution: Tris-Cl was prepared by dissolving 121 g Tris base in 900 ml H_2O, with adjustment of pH to 8.0 with concentrated HCl and supplementation to 1 L with ddH_2O.

10 M Sodium Hydroxide (NaOH) solution: NAOH solution was prepared by dissolving 400 g of tablets in 450 ml ddH_2O with addition of water to 1 L.

0.5 M Ethylenediamine tetra-acetic acid (EDTA): EDTA was prepared by dissolving 93.05 g $Na_2EDTA{\cdot}2H_2O$ in 450 ml ddH_2O concurrent with adjustment of pH to 8.0 using NaOH and supplementation with water to 500 ml.

50X Tris-Acetate-EDTA (TAE) solution: TAE was prepared by dissolving 242 g Tris-base in 850 ml of distilled water, followed by addition of 100 ml EDTA (0.5 M) and 57.1 ml glacial acetic acid. The solution's pH was then adjusted to 8.5 before addition of further ddH_2O up to a final volume of 1 L and sterilisation through autoclaving.

20X Sodium Saline Citrate (SSC) solution: SSC was prepared by dissolving 175.3 g NaCl and 88.2 g sodium citrate in 900 ml of distilled water, with adjustment of pH to 7.1 with concentrated HCl. Further ddH_2O was added up to a final volume of 1 L prior to autoclaving.

Phosphate-buffered saline (PBS): 10X PBS was prepared by dissolving 80 g NaCl, 2 g KCl, 11.5 g $Na_2HPO_4 \cdot 7H_2O$ and 2 g KH_2PO_4 in 950 ml ddH_2O with addition of further water to 1 L prior to autoclaving.

10% (w/v) Sodium Dodecyl Sulphate (SDS) stock solution: SDS was prepared by dissolving 50 g sodium dodecyl sulphate in 400 ml ddH_2O with subsequent addition of distilled water to 500 ml and purification by vacuum filtration.

10X Tris-EDTA (TE) buffer: TE was prepared by adding 100 ml Tris-Cl (1 M, pH 8) and 20 ml EDTA (0.5 M, pH 8) to 800 ml distilled water. Adjustment of pH to 8.0 was achieved via addition of concentrated HCl, prior to addition of ddH_2O to a final volume of 1 L and sterilisation by autoclaving.

Sodium acetate solution: 3 M solution was prepared by dissolving 40.82 g $CH_3COONa \cdot 3H_20$ in 90 ml ddH_2O with adjustment of pH to 5.2 with glacial acetic acid. Addition of further ddH_2O up to 100 ml was followed by autoclaving.

3-(*N*-morpholino) propanesulfonic acid (MOPS) buffer: MOPS buffer was prepared as 0.2 M MOPS, 0.5 M sodium acetate and 0.01 M EDTA by dissolving 20.93 g MOPS, 34.02 g sodium acetate and 3.72 g EDTA in 450 ml ddH_2O. Final addition of ddH_2O to 500 ml was followed by autoclaving to sterilise.

Salt-Optimised with Carbon (SOC) Medium: SOC medium was prepared by dissolving 20 g Bacto-tryptone and 5 g Bacto-yeast extract in 900 ml ddH_2O, with subsequent addition of 2.5 ml KCl (1 M) and 0.5 g NaCl. Adjustment of pH to 7.0 was via addition of NaOH with subsequent addition of ddH_2O to a final volume of 1 L and sterilisation by autoclaving. Prior to storage at -20°C, 20 ml of filter-sterilised glucose (1 M) was added to the medium.

Orange-G solution: 5X Orange-G loading buffer (0.3% w/v Orange G) was prepared by dissolving 0.03 g Orange-G loading buffer in 1ml 50X TAE, 3 ml 50% glycerol and 6 ml dd H_20. Addition of 1/5 volume per sample then allowed for electrophoresis.

Array spotting buffer: Spotting buffer was prepared by dissolving 40.548 g betaine inner salt monohydrate in 100 ml ddH_2O with addition of 30 ml 20X SSC, 60 ml ethylene glycol and further ddH_2O to 200 ml, providing a solution of 1.5 M betaine, 3X SSC and 30% (v/v) ethylene glycol. Solution was autoclaved to sterilise.

2.4 Samples and DNA extraction:

Sputum samples were obtained from patients with chronic pulmonary obstructive disease and stored at -80°C. Extraction of total bacterial community DNA was performed by Brenda Kwambana (University of Leicester) utilising a Gram-positive DNA extraction kit from Qiagen in accordance with the manufacturer's protocols.

Chicken faecal and caecal samples were obtained from Professor Tom Humphrey (Bristol Veterinary Research Centre) and transported in dry ice prior to storage at -20°C. Extraction of total bacterial DNA from 200 mg per sample of intestinal material was achieved using the Qiagen QIAamp® DNA Stool Mini Kit as per manufacturer's instructions. Optimal lysis of both Gram-positive and Gram-negative bacteria was through employment of an extended 95°C incubation in the associated lysis buffer (ASL) with removal of inhibitory substances present in faecal material via adsorption to a proprietary matrix. Proteins were degraded in the secondary lysis buffer (AL) containing proteinase K and chaotropic guanidine chloride, before adsorption of DNA to a column-based silica membrane, and washing with buffers containing ethanol. Volumes of DNA were then eluted in 200 µl of associated elution (AE) buffer as recommended and stored at -20°C until further analysis.

Faecal samples from hamsters were provided by Dr Gill Douce (University of Glasgow) and transported on dry ice. Storage of samples was at -20°C until extraction and elution with the QIAamp® DNA Stool Mini Kit as described above.

Human faecal samples were obtained from the Leicester Royal Infirmary in collaboration with Dr Martin Wiselka, Consultant in Infectious Diseases, as a preliminary to an ethically-approved clinical trial. Samples were stored at -20°C until collection and extraction was via the QIAamp® DNA Stool Mini Kit (as above) within 7 days of initial clinical microbiological analysis. Faecal samples were also obtained from volunteers at the University of Leicester on an *ad hoc* basis and treated as above.

2.5 Primers

All primers were obtained from Sigma-Aldrich (UK) and diluted to stock (100 μM) and working concentrations (40 μM) with purite ultrapure ddH_2O (18.2 MΩ) under sterile conditions, or for community 16S PCR, molecular biology grade water (Invitrogen). Primer sequences are displayed in Table 2.1.

Superscript [a] refers to the number of matches with the Ribosomal Database Project as of June 2011. Superscript [b] refers to the site on *E.coli* (or consensus sequence) 16S rDNA or pGEM-T® easy at which primer annealing occurs during PCR, the site given being the relative position of the 3' end of the primer. References for the sequence are detailed in the final column.

Table 2.1

NAME	SEQUENCE: 5'- 3'	RDP HITS [a]	SITE [b]	REFERENCE
Primers for amplification of *C. difficile* 16S rDNA				
CD-F	CTTGAATATCAAAGGTGAGCCA	113	198	Kikuchi *et al*., 2002
CD-R	CTACAATCCGAACTGAGAGTA	166	1244	Kikuchi *et al*., 2002
CD16S-F	TTGAGCGATTTACTTCGGTAAAGA	79	100	Rinttila *et al*., 2004
CD16S-R	CCATCCTGTACTGGCTCACCT	178	239	Rinttila *et al*., 2004
CXI-F1	ACGSTACTTGAGGAGGA	4370	431	Song *et al*., 2003
CXI-R2	GAGCCGTAGCCTTTCACT	1868	535	Song *et al*., 2003
Primers for amplification of inserts in pGEM-T® easy				
M13-F	CCCAGTCACGACGTTGTAAAACG	NA	2984	Promega (USA)
M13-R	AGCGGATAACAATTTCACACAGG	NA	190	Promega (USA)
pUC-F	GCCAGGGTTTTCCCAGTCACGA	NA	2972	Promega (USA)
pUC-R	TTGTGTGGAATTGTGAGCGGATAAC	NA	203	Promega (USA)
pGEM2166F	ATAAGGGCGACACGGAAATG	NA	2185	This study
pGEM500R	CCCTGATTCTGTGGATAACC	NA	481	This study

Primers for amplification of bacterial community 16S rDNA				
Bact-8F	AGAGTTTGATCCTGGCTCAG	172490	27	Baker *et al*., 2003; Weisburg *et al*., 1991
Bact-8FV	AGRGTTTGATCCTGGCTCAG	174684	27	Teske *et al*., 2002
Bact-27FV	AGAGTTTGATCMTGGCTCAG	213233	27	Ott *et al*., 2004
Bact-926R	CCGTCAATTCCTTTRAGTTT	805775	926	Baker *et al*., 2003; Reysenbach and Pace, 1995
Bact-1510R	GGTTACCTTGTTACGACTT	65865	1492	Baker *et al*., 2003; Turner *et al*., 1999
Bact-1492RVA	CGGCTACCTTGTTACGACTT	67233	1492	Teske *et al*., 2002; Sorensen *et al*., 2005
Bact-1492RVB	GGTTACCTTGTTACGACTT	65865	1492	Reysenbach *et al*., 1992; Leser *et al*., 2002
Bact-1407R	GACGGGCGGTGTGTRCA	395101	1391	Baker *et al*., 2003;
Bact-1391RVB	GACGGGCGGTGTGTRCA	395101	1391	Watanabe *et al*., 2004; Palmer *et al*., 2006
Bact-1391RVA	GACGGGCRGTGWGTRCA	427859	1391	Ashby *et al*., 2007
Bact-1541RV	AAGGAGGTGATCCANCCRCA	24128	1522	Suzuki and Giovanni, 1996
Bact-63FV	CAGGCCTAACACATGCAAGTC	87120	63	Fu *et al*., 2006; Park *et al*., 2005; Marchesi *et al*., 1998
Bact-334F	CCAGACTCCTACGGGAGGCAGC	931839	334	Baker *et al*.,2003; Rudi *et al*., 1997
Bact-334Fshort	CAGACTCCTACGGGAGG	948976	330	Rudi *et al*., 1997 (adapted this study)

Primers utilised for 454 run 1				
454-R1-357R	CTGCTGCCTYCCGTAG	1187980	341	Muyzer *et al*., 1993
454-R1-8F	AGAGTTTGATCCTGGCTCAG	172490	27	Baker *et al*., 2003; Weisburg *et al*., 1991
454-R1-1061R	CRRCACGAGCTGACGAC	926667	1061	Andersson *et al*., 2008
454-R1-784F	CAGGATTAGATACCCTGGTA	956644	803	Andersson *et al*., 2008
Primers utilised for 454 run 2				
454-R2-357F	CTACGGRAGGCAGCAG	1187980	357	Muyzer *et al*., 1993 (adapted this study)
454-R2-939Rs	GGGCCCCCGTCAATTC	568576	939	Rudi *et al*., 1997 (adapted this study)
454-R2-8F	AGAGTTTGATCCTGGCTCA	175365	27	Baker *et al*., 2003 (adapted this study)
454-R2-784R	ACTACCAGGGTATCTAATCC	956258	784	Baker *et al*., 2003 (adapted this study)
454-R2-1115R	TAAGGGTTGCGCTCGTTG	206023	1098	Baker *et al*., 2003; Reysenbach and Pace, 1995 (adapted here)
454-R2-515F	GCCAGCMGCCGCGG	1223592	529	Baker *et al*., 2003; Reysenbach *et al*., 1992 (adapted here)
Primers utilised for 454 runs 3-5				
926F	AAACTCAAAKGAATTGACGG	923955	926	Neufeld *et al*., 2008; Duncan *et al*., 2004; Muyzer *et al*., 1995
1391R	GACGGGCGGTGTGTRCA	395101	1391	Palmer *et al*., 2006; Watanabe *et al*., 2004; Lane *et al*., 1985

2.6 Standard DNA manipulation and analysis

2.6.1 Plasmid DNA extraction

Inoculates of selected *E. coli* colonies were incubated O/N at 37°C with agitation in 5-10 ml LBB with appropriate supplements as described previously (Section 2.2.2). Cells were harvested by centrifugation at 3000 x g for 10 minutes. Plasmid DNA was then extracted using the E.Z.N.A™ Plasmid Mini Kit 1 according to manufacturers' recommendations. Verification of plasmid presence was through agarose gel electrophoresis.

2.6.2 Agarose gel electrophoresis:

Gross visualisation of DNA fragments was achieved through electrophoresis of samples in 1-2% (w/v) agarose gels. Agarose was prepared by dissolving 4-8 g Seakem LE® agarose in 400 ml 1X TAE buffer (1 mM EDTA, 40 mM Tris-acetate) and heating on a medium microwave setting for 4-5 minutes. Subsequent to cooling to 50°C ethidium bromide was added to a final concentration of 0.5 µg/ml before casting in perspex trays using 12-22 prong combs to create wells. Samples of appropriate volume were mixed with 5X Orange-G loading buffer (e.g. 8 µl PCR product with 2 µl Orange-G) and electrophoresed at 5-8 V per cm of gel in 1X TAE buffer with appropriate markers (Hyperladder I, Bioline). Gels were then exposed to UV light and prints obtained using the Syngene Gene-Genius Bio-Imaging System (Synoptics, UK). Where necessary, extraction of DNA fragments from agarose gels was achieved using the Qiagen QIAquick spin columns in accordance with the manufacturer's instructions.

2.6.3 DNA quantification:

For preliminary DNA quantification an Eppendorf Biophotometer was employed, with standard 260/280 nm absorbance comparison for estimation of purity. A ratio of ~1.8 indicated good quality DNA while an A_{260} reading of 1 was considered equivalent

to 50 ng/µl in accordance with the Beer Lambert law (Ingle and Crouch, 1988). Lower 260/280 values indicate contamination with cellular debris, particularly proteins, and higher values suggest presence of residual RNA.. Secondary verification of spectrophotometer values was through comparison of the fluorescence of an electrophoresed sample under UV light with that of markers of known concentration. Where greater accuracy of quantification a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies, USA) was utilised as per manufacturer's guidelines.

2.6.4 PCR amplification

Quantities of DNA were amplified using the Polymerase Chain Reaction (PCR; Saiki *et al*., 1988) and Bio-X-Act^{TM} DNA polymerase (Bioline Reagents Ltd., UK), KAPA *Taq* DNA Polymerase (KAPA Biosystems, USA), or Phusion Hi-Fidelity DNA Polymerase (New England Biolabs Inc., USA) according to manufacturer's guidelines. Differing templates and primers necessitated tailored PCR conditions dependent on application; optimal constituents and cycling were determined through investigation of yield and specificity with primer annealment temperature gradients and titration of $MgCl_2$ and/or template dilutions. Negative controls to test for contamination of reaction mixtures were routinely included in all batches of PCRs. Below are details of colony PCR and use of primers (CD-F/R; CD16S-F/R; CXI-F1/CXI-R2) to test for the presence of *C. difficile*. Method-specific PCR conditions are detailed where appropriate.

2.6.4.1 Colony PCR

Colony PCR was performed by transferring a bacterial colony directly to 70 µl of sterile ddH_2O or 5 µl of inoculate with 65 µl sterile ddH_2O. Incubation at room temperature for 10 minutes preceded heating at 98˚C for 10 minutes in a thermal cycler. Subsequent centrifugation at 13000 x g for 5 minutes to pellet bacterial cell debris was followed by transfer of 5 µl of the supernatant as a template for a PCR reaction. In this

case, PCRs would be conducted utilising: 0.5 µl KAPA *Taq* polymerase (2.5 U/µl); 1 µl of the respective primers M13F and M13R (40 µM stocks; 0.8 µM final); 5 µl 10X Buffer A (dilution to 1X provides 1.5 mM Mg^{2+} concentration); 2 µl 25 mM $MgCl_2$ (2.5 mM Mg^{2+} final); 5 µl of 10 mM deoxynucleotide triphosphates (dNTPs), prepared by equimolar dilution of 100 mM stocks of dATP, dCTP, dGTP, and dTTP to 2.5 mM each; and PCR-quality ddH_2O to 50 µl. Cycling conditions were: primary denaturation for 5 minutes at 98˚C; followed by 25-30 cycles of denaturation for 30 seconds at 98˚C, primer annealing for 30 seconds at 53˚C, and extension at 72˚C for up to 2.5 minutes (extension rate = 1 Kb per minute); with a final extension step of 72˚C for 5 minutes.

2.6.4.2 PCR for detection of Clostridia

Reaction mixtures for gross detection of clostridia were routinely prepared in 50 µl as above. For primer pair CD-F and CD-R cycling conditions were as per section 2.6.3.1 but with annealing at 52°C and extension for 75 seconds. For primer pair CD16S-F and CD16S-R annealing was at 53°C with extension for 20 seconds. For primers CXI-F1 and CXI-R2 targetting the 16S rDNA of clostridial cluster XI (*Peptostreptococcaceae, Sporacetigenium* and *C. difficile*) annealing was at 49°C with extension for 20 seconds.

2.6.5 DNA purification:

Gross purification of PCR products to achieve removal of polymerase, buffer, primers and dNTPs was through use of the E.Z.N.A™ Cycle Pure Kit from Omega-BioTek or the Zymo DNA Clean and Concentrator™ Kit from Zymo Research, as per the respective manufacturer's instructions. Both kits employ a column-based silica membrane optimised to bind DNA while permitting removal of other reaction constituents. Elution of DNA was in volumes of between 30 and 50 µl ddH_2O.

Purification of DNA for electroporation was routinely achieved through addition of 10% (v/v) 3 M sodium acetate (pH 5.2), and 2.5X sample volume 100% ethanol prior to

thorough mixture by vortexing. Addition of 1 µl glycogen (20 mg/ml) was followed by further vortexing and incubation at 20 ˚C for 5 minutes with subsequent centrifugation at 13000 x g for 10 minutes. The supernatant was then discarded prior to washing with 1 ml 70% (v/v) ethanol to remove excess salts followed by a further centrifugation step of 13000 x g for 1 minute. Final removal of ethanol was through air-drying in a laminar flow for 1 hour followed by resuspension in 10 µl ddH_2O. Alternatively, samples were dialysed against ddH_2O on cellulose filter membranes (13 mm, 0.25 µm pore size; Millipore) for 40 minutes.

2.6.6 Ligation of DNA fragments

KAPA *Taq* and Bio-X-Act™ DNA polymerases catalyse the incorporation of single adenosine residues at the 3′ end of amplicons thereby allowing hydrogen-bonding with the thymidine residues of certain vectors and subsequent ligation. To ensure suitability of PCR products for ligation into the pGEM®-T easy vector an A-tailing procedure was performed where necessary. Addition of 1-7 µl of purified PCR product to KAPA *Taq* reaction buffer A (10X), 25 mM $MgCl_2$ to 2.5 mM final, dATP to 0.2 mM and 5 units of KAPA *Taq* polymerase preceded incubation for 30 minutes at 72°C.

Standard ligation reactions were performed at 4˚C for 16 hours (minimum) in 10 µl volumes containing 1 µl T4 DNA ligase (3units/µl), 5 µl 2X T4 DNA ligase Rapid Ligation Buffer, 3 µl of purified insert, and 1 µl pGEM®-T easy vector (50 ng/µl), with vector and insert in ratios of between 1:1 and 1:3. Subsequent purification of products was via ethanol precipitation as outlined above.

2.6.7 Restriction enzyme digestion

Restriction enzymes were utilised in accordance with manufacturer's instructions. Reactions were carried out with approximately 0.5 µg template plasmid DNA at 37˚C for 3 hours, while subsequent removal of enzymes and salts was via the Omega-Biotek

E.Z.N.A™ Cycle Pure Kit, as per the manufacturer's instructions. The specific sequence of the pGEM®-T easy vector system allowed for the use of the enzyme *NotI* to ensure complete excision of inserts. Results of restriction digestions were assessed via agarose gel electrophoresis.

2.6.8 DNA sequencing

Termination sequencing of DNA (Sanger *et al*., 1977) was through use of the BIG dye v 3.1 Cycle Sequencing Kit (Applied Biosciences) as per manufacturer's directions in 20 µl reaction volumes, inclusive of 1 µl Big Dye v 3.1, 3.5 µl 5X Sequencing buffer, between 200 and 300 ng of template plasmid DNA and 3µl of either pUCF (1 µM) or pUCR (1 µM). Cycling conditions were an initial denaturation step of 95°C for 5 minutes followed by 30 cycles of: denaturation at 96°C for 10 seconds, primer annealment at 50°C for 10 seconds and extension at 60°C for 4 minutes. Subsequent to thermal cycling, 2 µl 2.2% SDS was added to the samples to dissolve dye blobs, prior to boiling for 5 minutes at 98˚C. Samples were then further purified using Edge Bio Performa® DTR Gel Filtration Cartridges as per manufacturer's instructions and dispatched to the Protein and Nucleic Acid Chemistry Laboratory (PNACL, UOL) for capillary electrophoresis. Preliminary examination of sequence data for quality of traces was performed using Chromas Lite 2.01 (Technelysium Pty Ltd.). Analysis of data for confirmation of sequence identity and classification was performed using the BLAST website (http://www.ncbi.nlm.nih.gov/BLAST/ ; Altschul *et al*., 1990), or the Ribosomal Database Project (RDP) website release 10 (3/2011), inclusive of more than 1,600,000 bacterial and archaeal rRNA sequences (http://rdp.cme.msu.edu/ ; Cole *et al*., 2005). Further analysis procedures will be discussed as encountered in Chapters 3 and 4.

2.7 Creation of 16S rDNA clone libraries

2.7.1 Sample preparation

Total bacterial community DNA was extracted from intestinal contents as described previously; verification of purity and DNA concentration was then assessed using the Nanodrop ND-1000 spectrophotometer. As mentioned previously the 260/280 ratio is taken as an indication of sample DNA purity, with values of ~1.8 indicative of good quality DNA. However, contaminants such as protein and phenol also show significant absorbance at or near 280 nm while others such as carbohydrate and humic acids absorb in the 230 nm region (Nanodrop 2000 User Manual, Thermo Scientific). Therefore, when extracting from environmental samples such as faeces which are likely to include a range of potential contaminants, neither the 260/280 ratio nor the concentration as calculated from A_{260} values are absolute determinants of subsequent potential for PCR amplification. In fact, sample 260/280 ratios of 1.5-2.3 and concentrations as low as 1ng/µl were ultimately shown to provide good quality clone libraries. This caveat notwithstanding, attempts were made to standardise concentrations of samples to 10 ng/µl based on Nanodrop values.

2.7.2 Community 16S rDNA PCR

As discussed in Section 1.3, 16S rDNA contains regions of relatively conserved sequence in addition to the hypervariable regions. While the latter provide the means for taxonomic assignment, the former can be utilised to create amplicons representative of heterogeneous bacterial populations. Primers utilised in this study are listed in Table 2.1.

PCR components for 16S amplicon creation were routinely mixed in 50 µl volumes containing: 5 µl of DNA sample template (10 ng/ µl); 1 µl each of forward and reverse primer (40 µM); 5 µl dNTPs (10 mM); 5 µl 10X Optibuffer (Bioline); 2-2.5 µl 50 mM

$MgCl_2$ (2-2.5 mM Mg^{2+} final); 5 µl 10X Bovine Serum Albumin (BSA, added as per QIAamp® DNA Stool Mini Kit manufacturer's instructions for downstream PCR) to a reaction concentration of 0.1 µg/µl; 1 µl Bio-X-Act™ DNA polymerase (4 units per µl); and PCR-quality ddH_2O to 50 µl. Standard cycling conditions were as follows: initial denaturation at 98°C for 5 minutes; followed by 20-25 cycles of denaturation for 30 seconds at 98°C, annealing for 30 seconds at between 48 and 53°C dependent on primer pair (see below), and extension at 70°C for 2 minutes (extension rate = 1 kb per minute); and a final extension at 70°C for 7 minutes. The total number of cycles was minimised to avoid creation of extraneous amplicons or artefacts (Wintzingerode *et al*., 1997; Speksnijder *et al*., 2001;) and chimeras (Wintzingerode *et al*., 1997; Qiu *et al*., 2001), while also reducing the likelihood of PCR saturation, a state known to increase artificial equivalence of amplicon proportions as compared to their original parent templates (Suzuki and Giovannoni, 1996).

Annealing temperatures and Mg^{2+} concentrations for commonly utilised primer pairs are given in Table 2.2

Table 2.2 16S rDNA PCR annealing temperatures and magnesium concentration

Primer Pair	Temp	[Mg^{2+}]
Bact-8F with Bact-1391RVB	53°C	2.5 mM
Bact-8F with Bact-1492RVB	53°C	2 mM
Bact-8F with Bact-1541RV	55°C	2.5 mM
Bact-63FV with Bact-1391RVB	52°C	2.5 mM
Bact-63FV with Bact-1492RVB	53°C	2 mM
Bact-63FV with Bact-1541RV	54°C	2.5 mM

PCR products were analysed by gel electrophoresis to confirm successful amplification in terms of yield and presence of the desired amplicon. Removal of unwanted constituents was achieved through use of a commercially-available kit as outlined in Section 2.6.5.

2.7.3 Ligation

Samples were quantified utilising the NanoDrop™ before dilution with ddH_2O such that concentrations were standardised at 20 ng/µl. Ligation reactions were then prepared as described previously with 1 µl of pGEM®-T easy vector (50 ng/µl) and 3 µl of insert to provide a molecular vector to insert ratio of 1:3.

Where PCR samples were too dilute, ligation reaction volumes were adjusted accordingly (e.g. 6 µl of 10 ng/µl concentrations in 20 µl). Ligation reactions were allowed to proceed for at least 16 hours at 4 °C. DNA was then precipitated with ethanol as per section 2.6.5.

2.7.4 Transformation by electroporation:

Electrocompetent *E. coli* DH5α were prepared by first inoculating 10 ml LBB with cells from glycerol stocks, prior to incubation overnight at 37˚C with agitation at 300 rpm. The overnight culture then provided 1 ml which was utilised to inoculate 99 ml fresh LB broth in a conical flask. Incubation was then as above until an OD_{600} of between 0.4 and 0.5 could be recorded. Bacterial cells were then pelleted by centrifugation at 2500 x g for 10 minutes at 4°C, prior to resuspension in 2 ml of a solution containing 20% glycerol (v/v) and 1 mM MOPS, produced from autoclaved stocks of 200 mM MOPS, 50% glycerol and ddH_2O by mixing in the ratio 1:80:119. Further centrifugation as above was followed by resuspension in 1ml of 20% glycerol, 1 mM MOPS. A further 4 rounds of centrifugation at 13000 x g for 90 seconds at room temperature (RT) and washing were followed by final resuspension in 200 µl of 20%

glycerol, 1 mM MOPS. Aliquots of 50 µl were then flash-frozen over dry ice and stored indefinitely at -80˚C.

Transformation of precipitated recombinant plasmid vectors into bacterial cells employed a BioRad Gene Pulser set to 25 µFD capacitance and 1000 Ω resistance. Volumes of 2-10 µl purified DNA were added to 50 µl electrocompetent cells (prepared as per section 2.4) in pre-cooled 2 mm electroporation cuvettes. The resultant suspension was then chilled on ice for 10 minutes before being subjected to a single pulse of 1.5 kV. Time constants of 19-23 ms were taken as an indication that the procedure had been successful. The time constant is a product of the resistance and the capacitance of the pulse circuit.

Subsequent to electroporation, *E. coli* cells were resuspended in 1 ml SOC medium prior to incubation at 37˚C for 1 hour. Cells were then plated onto LBA-AIX plates prepared as described previously prior to O/N incubation at 37˚C. For small scale library preparation 100 µl was plated onto standard petri dishes prior to overnight incubation; for larger-scale libraries bioassay dishes (245 mm x 245 mm, low-profile) were utilised and the entire 1 ml of resuspended transformation was plated.

2.7.5 Identification and selection of desired recombinants

In conjunction with the host *E. coli* strains listed in section 2.1 the pGEM®-T Easy vector provides the means for identifying colonies harbouring the recombinant plasmids via ampicillin resistance and blue/white screening, white colonies being expected to contain recombinants due to interruption of the *lacZ* (beta-galactosidase) gene. The desired colonies were obtained primarily through selective antibiotic pressure prior to confirmation via plasmid extraction and DNA sequencing. The absence of blue product indicated successful transformation of bacterial cells with vector containing ligated insert. To create small-scale temporary libraries, colonies of interest were transferred to

patch plates using sterile toothpicks, for storage at 4°C after overnight incubation at 37°C.

For larger-scale libraries, LBA-AIX bioassay dishes were loaded onto the sub-illuminated picking tray of a Genetix QPix robotic system, linked to a PC with associated QPix picking software (version 3.6) installed; up to 15 96-well low-profile plates (containing 100 µl LBA with ampicillin per well) were accommodated in the destination plate holders. The three baths for washing of robotic head pins between rounds of selection and transfer were filled with 1% sodium hypochlorite solution, ddH2O and 80% ethanol respectively. The camera was calibrated (value > 0.94) to control for the characteristics (depth of agar, degree of illumination etc.) of each plate, and aligned based on identification of a single colony; adjustment of focus and contrast settings to provide the optimal image for picking was also performed for each individual plate. Subsequent standard steps in the picking procedure are detailed in the QPix Colony Picker Software Version 3.6 Guide and QPix Robot Manual (both produced by Genetix), but certain custom parameters can be chosen and incorporated into the picking script by the user as follows:

- Grey threshold was adapted manually such that all intensities between certain values are discerned as background i.e. agar. Values were generally between 120 and 180 (integral contrast units)

- Blue threshold was set such that colonies with a given intensity are rejected as non-white; threshold was generally between 185 and 205, while white colonies were found to show intensities of greater than 215 (integral contrast units)

- Diameter of colonies was set to between 10 and 100 µm, roundness* minimum was set at 0.6 and axis ratio* was set to 0.7. White colonies without these characteristics were not selected.

- Proximity detection* was enacted at a value of 8 and 'Check overlaps'* setting was activated with a value of 20. The former ensures that selected colonies are well separated while the latter prevents colonies being selected in a number of different frames since plates are divided into 30 different sectors or 'frames' which are visualised individually.

- The 'Delete 'Blobs'' setting causes other extraneous images, such as scratches and bubbles to be considered as background. Objects with a diameter less than 2 µm, an axis ratio less than 0.1 and an area greater than 500 μm^2 were thus excluded.

* Units for roundness, axis ratio, proximity detection and overlap check are not included as these are programme-defined variables; optimal settings for these values were obtained empirically.

Prior to initiation of the picking script, each frame was checked manually for anomalies such that suspect colonies chosen by the software for selection could be rejected. Plates were routinely able to provide between 700 and 1500 colonies of a satisfactory nature.

Sealed 96-well plates were then incubated at 37°C with agitation for 16 hours, before 50 µl of 60% glycerol with LBB was added to each well to give a final concentration of 20% glycerol. Plates were then sealed and stored at -80°C.

2.7.6 Classification of inserts

To determine the precise sequence of cloned 16S rDNA, sterile toothpicks were utilised to transfer a small amount of glycerol stock from a plate well into a 20 ml universal polystyrene container holding 5 ml LBB with ampicillin (0.1 mg/ml). Inoculates were incubated overnight at 37°C with agitation prior to extraction of plasmid DNA as described in Section 2.7.6. DNA was quantified prior to use of an

appropriate volume in sequencing reactions as described in Section 2.7.8. Where the sequence of the entire 1.5 kb 16S insert was required 2 reactions would be performed on each clone extract, the first with pUC-F and the second with pUC-R. Subsequent sequence processing and analysis was as per section 2.7.8, with additional use of the *revseq* and *merger* programs on the EMBOSS website (http://www.es.embnet.org/Services/MolBio/emboss-gui/) for creation of contiguous sequences representing the full insert.

2.8 Production of arrays

2.8.1 Preparation of PCR template from plates

To prepare templates for PCR from glycerol stocks of bacteria, 10 µl of bacterial suspension was transferred from each well to a fresh 96-well non-skirted PCR plate, each well having been pre-loaded with 40 µl of nanopure ddH_2O. Plates were then sealed and heated to 98°C for 10 minutes in a thermal cycler. Plates were centrifuged at 1200 x g for 10 minutes to sediment bacterial debris before 5 µl from the supernatant fraction of each well was removed for introduction into a 45 µl PCR mixture (section 2.8.2), previously deposited in each well of a fresh 96-well PCR plate.

2.8.2 PCR

PCR mixtures for amplification of 16S inserts in pGEM-T® easy were prepared in 45 µl volumes as follows: 5 µl KAPA *Taq* Buffer A (10X; dilution to working concentration provides 1.5 mM Mg^{2+}); 5 µl 10 mM dNTPs (1 mM final); 0.7 µl 40 µM M13F and 0.7 µl 40 µM M13R (0.8 µM final); 2 µl 25 mM $MgCl^2$ (2.5 mM Mg^{2+} final reaction concentration); 0.3 µl KAPA *Taq* DNA polymerase (2.5 U/ µl); and PCR-quality ddH_2O to 45 µl. For construction of arrays with pre-labelled spots (as described in section 3.5), M13R primer volume comprised 0.35 µl standard M13R and 0.35 µl Alexa-488 labelled M13R.

Cycling conditions were as follows: initial denaturation at 98°C for 5 minutes; followed by 28 cycles of denaturation for 30 seconds at 98°C, annealing for 30 seconds at 53°C, and extension at 72°C for 2.5 minutes (extension rate = 1 kb per minute); and a final extension at 72°C for 5 minutes. Due to the number of PCRs performed, each individual reaction was not assessed for successful creation of product, but 2 µl from each well of a randomly-selected column from each plate was routinely utilised for

preparation of mixtures for electrophoresis. Negative controls were also routinely included for each batch of reactions.

2.8.3 Isolation of PCR product

For purification and concentration of DNA, 10% (v/v) 3 M sodium acetate (pH 5.2) was added to each well containing 50 µl of PCR product, followed by addition of 55 µl isopropanol (propan-2-ol) with thorough mixture by pipette, before incubation at -20°C overnight as recommended by protocols for precipitation with isopropanol to increase yield.

Thawing of samples was followed by centrifugation at 1200 G for 1 hour. Removal of the supernatant preceded addition of 2X PCR sample volume (100 µl) of 70% ethanol prior to further centrifugation at 1200 G for a further 30 minutes. Subsequent to removal of ethanol the plates were dried in a laminar flow cabinet for 30-60 minutes. Precipitates were then resuspended in 15 µl of a solution comprising 1.5 M betaine, 3X SSC and 30% (v/v) ethylene glycol.

2.8.4 Control clones

Where perfect match clones for a probe could not be selected from screened libraries artificial matches were created. Probe sequences and their complements were ordered from Sigma-Aldrich inclusive of a single adenosine residue overhang at the 3' end. Equimolar concentrations of probe and complement were mixed and heated to 50°C for 10 minutes. Subsequent to cooling and appropriate dilution, double-stranded molecules were then ligated and transformed as per sections 2.7.3 and 2.7.4, with verification as per 2.7.6.

2.8.5 Slide and coverslip preparation

For manufacture of poly-L-lysine coated slides a saturated NaOH wash solution was first prepared by dissolving 100g NaOH pellets in 400ml ddH_2O, with addition of

further NaOH until a white precipitate had formed indicating saturation. Subsequent addition of 600 ml 95% EtOH was followed by further addition of ddH_20 until the precipitate was cleared. The solution was then vacuum-filtered using Millipore Stericup® and SteriTop® vacuum-filter accessories. Standard glass slides (25 mm x 75 mm x 1 mm) loaded in Shandon Lipshaw metal racks (30 per rack) were rinsed with ddH_20 before immersion in the above solution in glass staining-jars (135 mm x 135 mm x 135 mm); washing with agitation (Stuart See-Saw Rocker) then proceeded for 1 hour prior to rinsing with ddH_20 and immersion in fresh containers containing further ddH_20 where they were allowed to stand overnight. Poly-L-lysine solution was prepared by mixing 570 ml ddH_2O with 70 ml 10X phosphate-buffered saline (PBS) and 60 ml poly-L-lysine (Sigma-Aldrich) solution; slides were submerged in this solution for 1 hour, prior to a further wash with ddH_2O before initial drying by centrifugation at 1000 x g for 2 minutes. Drying was completed by incubation in a vacuum oven at 50°C for 30 minutes. Slides were stored in a sealed environment until spotting. Aldehyde and amino-silane coated Nexterion® glass slides (also 25 mm x 75 mm x 1 mm) were purchased from Schott as vacuum-sealed packs of 50 and required no subsequent processing.

Prior to usage, standard glass coverslips (44 mm x 22 mm) were washed in racks in 1% SDS for 30 minutes, then in 5 changes of ddH_2O for 5 minutes per wash. Subsequent to washing coverslips were dried by centrifugation at 2500 x g for 4 minutes and stored in a dust free environment. LifterSlips™ were purchased in sealed packaging from Thermo Scientific and required no pre-treatment.

2.8.6 Spotting of microbiomic library arrays

PCR products dissolved in betaine/SSC/ethylene-glycol spotting buffer were transferred from 96-well non-skirted PCR plates to low-profile Genetix 384-well plates using the Beckman Biomek 2000 Laboratory Automation Workstation (Beckman Coulter, Inc., UK).

Arrays were produced by the microarray facility in the Toxicology Unit at the University of Leicester using either an Ultra Marathon Arrayjet (Arrayjet, UK) or a Stanford spot printer, the latter having been custom-built by Dr Tim Gant (University of Leicester, UK) using components supplied by Western Technology Marketing (USA). Spotting of features onto slides by the Stanford printer is a contact process: steel pins are immersed in the 384-well plate and subsequently deposit approximately 1 nanolitre of sample per 10 μm spot; the Arrayjet, however, creates features using piezoelectric (inkjet) printheads, thus delivering the 100 picolitre sample to the array-surface without coming into contact with the substrate. Prior to hybridisation slides were stored in a dust-free environment away from exposure to ambient light.

During production a '.gal' file was also prepared allowing for correlation of PCR sample identification and slide feature location during downstream analysis utilising the GenePix®Pro (Version 6.1) Software package (Molecular Devices LLC., USA).

2.9 Hybridisation and detection

2.9.1 Hybridisation

Prior to hybridisation arrays were baked for 2 minutes at 100°C and 1 hour at 80°C to maximise fixation of DNA to the solid support matrix. Subsequently slides were washed twice for 2 minutes each in filtered 0.2% SDS, and then twice for two minutes each in ultrapure ddH_2O. Drying of slides was achieved through centrifugation at 2500 x g for 4 minutes.

Hybridisations were performed using the Genisphere 3DNA® Array50™ kit which provides for an initial hybridisation step followed by post-hybridisation labelling or capture. Probe design is discussed in detail in section 3.8, but probes were ordered as synthesized oligonucleotides incorporating the requisite 16S-specific sequence and proprietary capture sequences at the 5′ end. The capture sequence for Cy3 is 5′-TTCTCGTGTTCCGTTTGTACTCTAAGGTGGA-3′, while that for Cy5 is 5′-ATTGCCTTGTAAGCGATGTGATTCTATTGGA-3′. During the post-hybridisation labelling phase the trailing capture sequences of the bound probes allow for hybridisation with complementary sequences of the fluorescent dyes and subsequent detection.

Hybridisation mixtures were prepared with 7 µl of each 20 µM probe (140 pmol), 30 µl 2X Genisphere Enhanced Buffer and nuclease-free H_2O to a volume of 60 µl. Mixtures were heated to 80°C for 10 minutes followed by incubation at the hybridisation temperature until loading. Loading was by capillary action beneath a coverslip or as multiple droplets prior to application of the coverslip. Hybridisations were then allowed to proceed for between 16 and 24 hours in a sealed, humidified, 10-slide Genetix chamber inserted into a Techne Hybridisation HB-1D Oven set to the appropriate temperature.

2.9.2 Capture and detection

The arrays were subsequently washed in various buffers in glass staining jars to remove excess unbound probes and minimise non-specific binding. Buffers were prepared with dilutions of 10% SDS (w/v) and 20X SSC and vacuum-filtered prior to use. Wash Buffer 1 (WB1) was composed of 0.2% SDS and 2X SSC; Wash Buffer 2 (WB2) was 2X SSC; and Wash Buffer 3 (WB3) was 0.2X SSC. WB1 was pre-heated to the incubation temperature and hybridisations were terminated by inversion of arrays allowing coverslips to float off. Slides were then transferred to a rack in a fresh volume of pre-heated WB1 and washed with 50 rpm agitation on an HB-SHK1 orbital shaker for 2 minutes. Slide racks were then briefly rinsed with WB2 (RT) prior to immersion in a fresh volume of WB2 (RT) and washed with agitation for a further 2 minutes. Final transfer of slide racks was to a fresh volume of WB3 (RT) with subsequent washing for a further 6 minutes. Slides were then dried through centrifugation at 350 G for 4 minutes.

Capture mixtures were prepared in volumes of 44 µl containing 2.5 µl of the appropriate 3DNA® capture reagent (Genisphere, Inc), 22 µl of 2X Enhanced Buffer (Genisphere, Inc) and nuclease-free H_2O to the final volume. Mixtures were heated to 80°C for 10 minutes and then maintained at the incubation temperature until application. Loading of capture solutions and incubation were as described for hybridisations (section 2.9.1) but duration was 2-3 hours. Subsequent washes were exactly as for post-hybridisation but in the absence of light to minimise degradation of fluorescent reagents.

Slides were then analysed utilising the Axon 4200A 4-channel scanner from Molecular Devices Inc., with an excitation wavelength of 532 nm for Cy3 (appears green), and 635 nm for Cy5 (appears red). Where Alexa-488 labelled M13R had been utilised for amplification of clone inserts, scans were also conducted at an excitation

wavelength of 488 nm (appears blue). Since labelled M13R should have been incorporated into all PCR products scans at 488 nm allowed for estimation of spotting efficiency, and, to an extent, normalisation of signal in relation to product concentration. Settings for scanning were as described in the manual for GenePix®Pro, with laser power set to 95% and (PMT) Photo-Multiplier Tube or gain values of between 400 and 800 PMT (system-specific units) dependent on excitation wavelength, hybridisation quality and the heuristic rule that no more than 5% of features should show saturation intensity (appear white). Initial images were saved as '`.tif`' files with subsequent manual evaluation of all features. Where spots were considered inadequate for subsequent analysis due to inefficient printing, dust contamination, poor 3DNA reagent diffusion or high background fluorescence (*inter alia*), features were 'flagged' prior to saving of settings as '`.gps`' files and final conversion to '`.gpr`' files for export and further analysis in Excel.

2.10 Second generation sequencing

As an alternative to traditional Sanger sequencing or OPTIMarrays for analysis of complex microbiomic samples the 454 facility at University of Leicester, or the Centre for Genomic Research (CGR) at the University of Liverpool were utilised. Samples of bacterial DNA extracted as per section 2.4 were subjected to PCR amplification as described in section 4.2.3. Primers employed were designed to amplify hypervariable regions of the bacterial 16S rDNA gene. In addition to the template-specific sequence, the primers were designed to incorporate a short oligonucleotide (barcode or multiplex identifier - MID) for sample differentiation along with the primer A or primer B sequence at the 5′ end. The primer A/B sequences mediate binding of the ssDNA to the beads and, through complementarity to Roche 454 system primers, allow for extension of the fragment at both the emulsion PCR (emPCR) stage and during sequencing. Subsequent to PCR of bacterial DNA and AMPure purification, samples were quantified using the PicoGreen system, diluted to standard and equal concentrations, pooled, and dispatched for processing.

2.11 Supplementary methods and bioinformatics

Supplementary methods specific to the various phases of the investigations are included in short methodological sections in Chapters 3, 4 and 5. Bioinformatics platforms and procedures are also included in the relevant chapters as explanation of the approaches is more comprehensible in the context of the results.

A list of software and internet resources is included in Appendix 5, while scripts developed for analysis, bioinformatics workflows, statistical methods, and selected relevant ecological indices are detailed in Appendices 7 through 14.

CHAPTER 3

ARRAY METHOD DEVELOPMENT

3.1 Oligonucleotide Fingerprinting of Ribosomal Genes (OFRG)

In the current study all methodologies share a common protocol up to a point: collection of appropriate samples and extraction of community bacterial DNA, followed by PCR of the 16S rDNA genes with suitable 'universal' primers. At this point the samples could be prepared for analysis by massively parallel pyrosequencing such as 454 (Chapter 4) or further processed to form clone libraries by incorporation of the PCR products into a suitable vector and bacterial host. The clone libraries could then be directly sequenced to identify community members, or form the basis of a relatively novel approach known as oligonucleotide fingerprinting of rRNA genes (OFRG), borne of a technique utilised in a study on clones derived from human libraries (Drmanac *et al*., 1996)

OFRG was first described in relation to analysis of 18S rDNA clones of a fungal community, the PCR products from the clones fixed onto nylon membranes and interrogated with a set of 27 10-mer probes, designed to sort the clones into taxonomic clusters (Valinsky *et al*., 2002b). A similar approach was subsequently utilised to investigate a bacterial community associated with soil environments (Bent *et al*., 2006), and more recently for analysis of microbial communities in the intestines of turkeys (Scupham *et al*., 2007).

The fundamental difference between this approach and the array studies discussed previously (Section 1.9) is that the nucleic acids immobilised on the platform are those under investigation, while the oligonucleotides used to interrogate the 'arrays' are the 'probes' of known sequence; in this sense, the approach can be considered as inverse hybridisation, or as a return to a procedure more akin to 'Southern' blotting (Southern, 1975), as displayed in figure 3.1 (page 105).

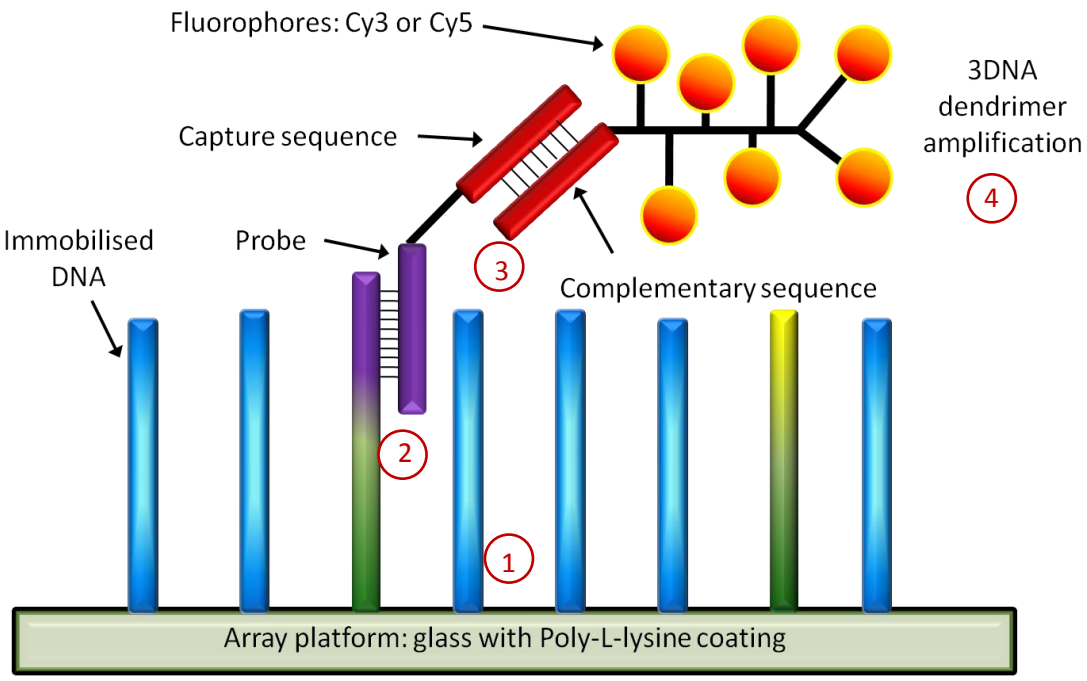
3.2 Oligonucleotide Fingerprinting of Arrayed Ribosomal Genes (OFARG)

One of the aims of the current study was to develop and modify the OFRG technique to create a methodology suitable for high-throughput investigation of faecal samples from human subjects. Studies utilising OFRG were limited to investigation of less than 5000 clones (Yin *et al*., 2005; Scupham *et al*., 2008; Valinsky *et al*., 2002a) due to the capacity of nylon membranes; the potential for use of glass as the substrate surface had also been explored, raising the limit to approximtely 10,000 (Bent *et al*., 2006), but with the improvement in spotting technologies this could be extended to around 30,000 clone products per slide. While probe sets for OFRG had been described (Bent *et al*., 2006), a secondary goal was to design a fresh library of oligonucleotides, optimised for differentiation of the maximum possible number of OTUs in the human intestinal milieu. Preliminary studies had also indicated that fluorescent signal strength (Rob Free and Colin Barker; unpublished data) with directly labelled-probes might prove limiting (Rob Free and Colin Barker; unpublished data); a labelling system for array hybridisations produced by Genisphere (Hatfield, PA, USA), incorporating 3DNA dendrimers (Stears *et al*., 2000; Mora *et al*., 2008) to provide amplification of fluorescent intensity, was thus chosen for integration into the methodology. To distinguish the method under development from that established previously it was designated OFARG: Oligonucleotide Fingerprinting of Arrayed Ribosomal Genes.

Glass slides coated with either poly-L-lysine, aldehyde or amino-silane constitute the array platform, while the immobilised nucleic acids are PCR products manufactured using plasmid-specific primers (pUC-F/R, M13F/R or pGEM2166F/pGEM500R) with clone libraries as the template (section 2.9); both PCR and print replicates can be included on an array to minimise the effects of variabilities induced by the spotting and hybridisation procedures such as black holes, doughnuts, coverslip scratching, and background fluorescence nebulae (Bowtell and Sambrook, 2003). Full hybridisation and

scanning methods are detailed in section 2.10 while the hybridisation approach is as outlined in figure 3.1 and the theoretical basis for probe design is described in figure 3.2.

Figure 3.1 OFARG overview



1 DNA from PCR of clone library is immobilised on coated glass slides. 2 Interrogation with probes in the first stage of hybridisation. Probe binds to complementary sequences on the array. Probes are designed to incorporate a fluorophore-specific capture sequence, either for Cy3 or Cy5. 3 Slides are washed to remove unbound reagents and non-specifically bound probe and challenged with the capture reagents. Capture reagents consist of the fluorophores complexed to an oligonucleotide sequence with complementarity to, and specificity for, the tag sequence on the probe. 4 Amplification of signal through 3DNA dendrimer technology allows for greater sensitivity of detection by the scanner (Stears *et al*., 2000)

Development of the technique can be partitioned into 3 phases. Phase I was focused on obtaining good quality images amenable to analysis with the GenePix software through optimisation of standard protocols. At this stage a minimal set of probes and a small pre-prepared library of 16S rDNA clones were adequate for purpose. Phase II saw an

attempt to refine the methodology through introduction of control and reference clones; arrayed libraries consisted of multiple replicates of fully-sequenced 16S clones and varied approaches to analysis were adopted. During phase II the novel probeset for intestinal investigations was also finalised. Phase III was then be implementation of OFARG for comparison of libraries derived from CDAD patients and subjects with AAD.

FIGURE 3.2 Rationale for fingerprint assignment with OFARG

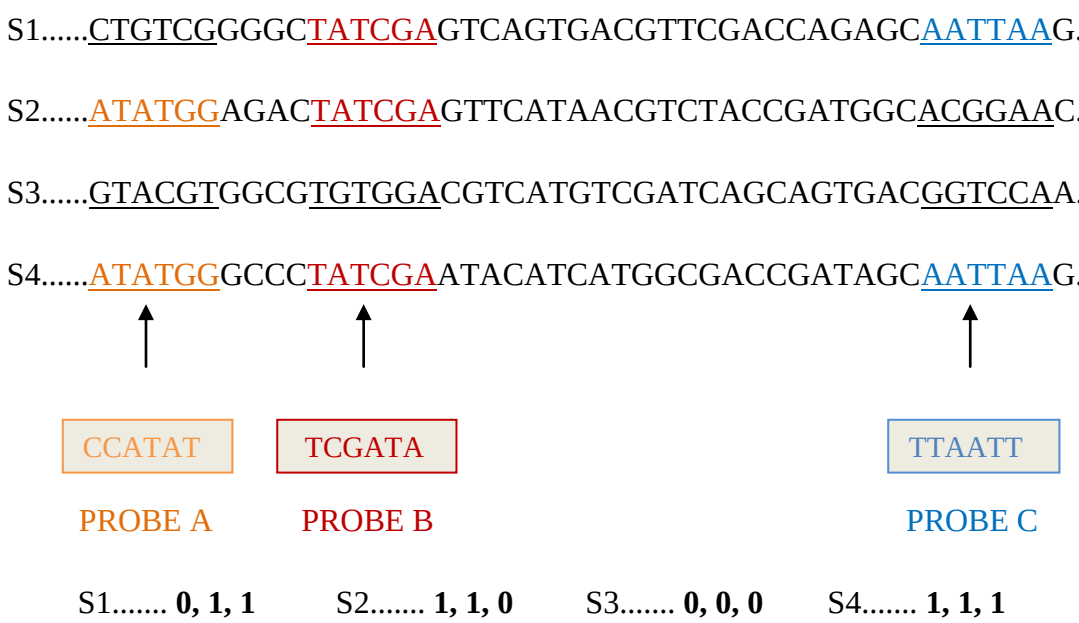


Figure 3.2 shows alignment of 4 sequences to identify regions of conservation. Sequence in orange is conserved between S2 and S4. Sequence in red is conserved between S1, S2 and S4. Sequence in blue is conserved between S1 and S4. Probes are 6-mers purely for representational purposes. Below the sequences is the expected hybridisation readout for challenge of a clone containing the sequence with probes A, B, and C respectively. The output is represented in binary form: 1 = hybridised, 0 = not hybridised. Note that differentiation of the 4 sequences could be achieved without challenge with probe B.

3.3 Phase I

The majority of hybridisations for initial optimisation of the array technique were performed on glass slides coated with poly-L-lysine and spotted with microbiotic 16S clone libraries from COPD samples, prepared by Brenda Kwambana (University of Leicester) as per section 2.9, with primers Bac-8F and Bac-1391R (Table 2.1). Each array comprised a total of 3456 spots. 2741 of these are the features, PCR products from the clone library, while a further 715 are empty spots or blanks containing only the printing buffer (betaine SSC).

Table 3.1 Probes used in OFARG Phase 1

Probe	Sequence	% Hits in RDP
338F	ACTCCTACGGGNGGCNGCA	69.2
574	GGGCGTAAAGCGTGCGCAGGCGG	4.2
755	CGAAGAACCTTACCAGGTCTTG	5.9
1239	GGGCTGCACACGTGCTACAATG	1.4
Cy3	TTCTCGTGTTCCGTTTGTACTCTAAGGTGGA	NA
Cy5	ATTGCCTTGTAAGCGATGTGATTCTATTGGA	NA

The initial probe set is shown in Table 3.1. Probes were designed and validated by R.C.Free (PhD thesis, 2005), except 338F, which is adapted from Baker *et al*., 2003. Probes were designed to incorporate the Cy3 or Cy5 capture sequences, property of Genisphere corporation (USA) for use in conjunction with the company's 3DNA signal amplification system. The table also shows the percentage of sequences in the Ribosomal Database Project (RDP) website database providing a positive match with each non-specific probe.

Initial attempts at hybridisation followed the manufacturer's protocol for the 3DNA Array Detection Array 50™ Kit (www.genisphere.com/pdf/Array50_Jan2011.pdf) which employs high hybridisation temperatures and long post-hybridisation washes, utilising either a formamide-based or SDS-based hybridisation buffer. However, the recommended temperatures proved excessive, contributing to substantial regions where no hybridisation can be detected ('black holes'), while the duration of washes increased background signal ('nebulae') rather than promoting attenuation. An example of an array scan obtained at this stage is shown in figure 3.3.

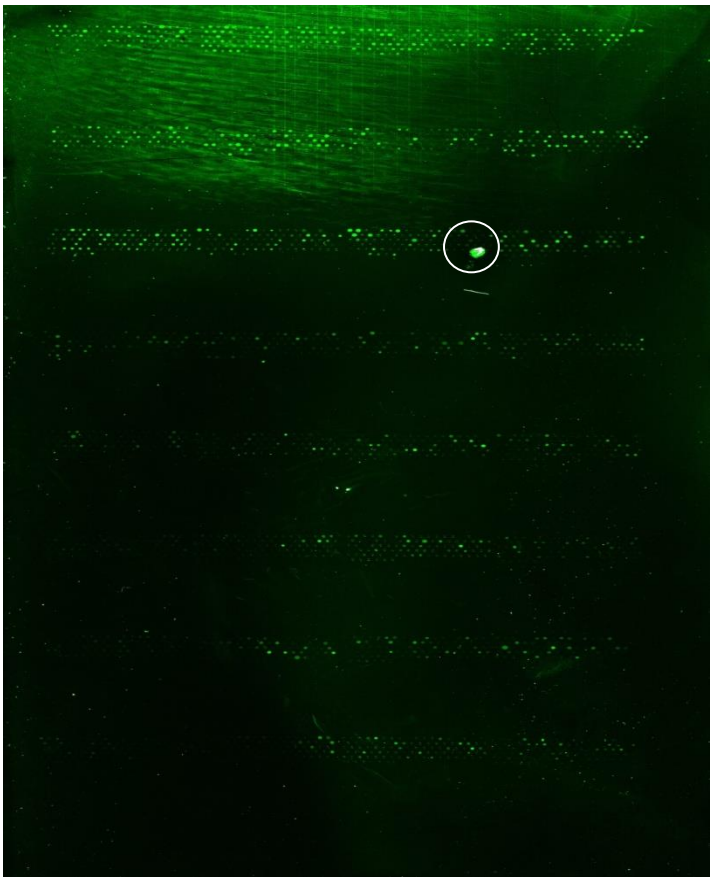


Figure 3.3 Phase I COPD array scan (early)

Figure 3.3 to the left displays pre-optimisation output with probe 338F and SDS-based buffer. Regions without signal (black holes; bottom left), scratches from coverslip removal (top left), nebulae (high background signal) due to inadequate dye removal (top left), and dust spots (circled) are all evident, while feature signal intensity is variable and poor. Problems were attributed to high temperatures, unsuitable buffer, and inexperience with the technique.

While reduction of the incubation temperatures from 60°C to 42°C improved hybridisation, the definitive adaptation for optimisation proved to be selection of an alternative incubation solution, both the SDS and formamide being inferior to a

proprietary buffer (Genisphere Inc.) known as 'Enhanced'; the constituents could not be ascertained but it is speculated that it provides a buffer similar to Denhardt's solution. And although arrays may be robust in certain respects a minimum level of expertise is required; this can be attained only through repeated implementation of the procedures.

The arrays utilised in Phase I do not contain the controls which would permit evaluation of non-specific binding and subsequent use of this value to assess positive hybridisation as described subsequently. However, once subjective consistency had been achieved the feature intensities were analysed on the basis of signal-to-noise ratio (SNR). For this calculation the background signal is assessed locally for each feature (average of pixels) and subtracted from the mean pixel intensity of the feature; this figure is then divided by the standard deviation of the background intensity, a SNR value of greater than 3 being taken as evidence of hybridisation (Manual for GenePix® Pro 6.0, Molecular Devices, 2005).

Table 3.2 Phase I analysis

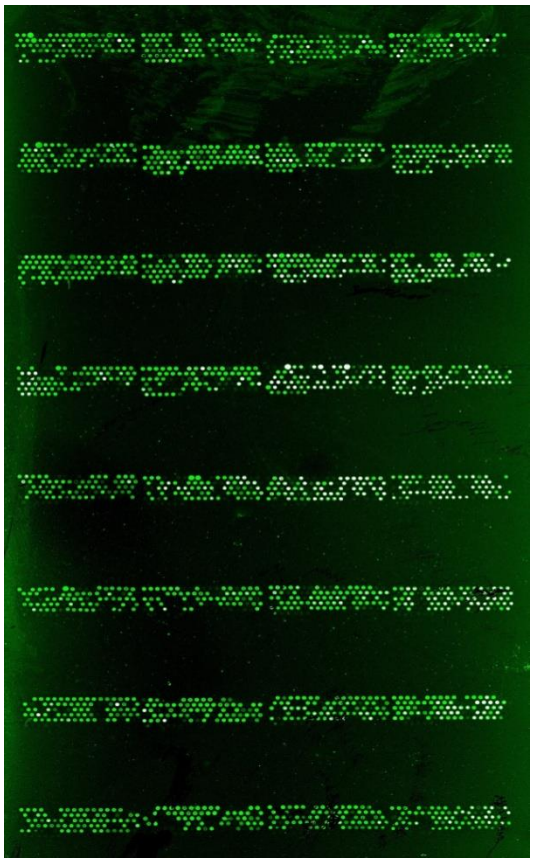
Category	338F	755
% Features Positive	74.84 (±1.04)	65.94 (±1.29)
% Blanks Positive	2.44 (±0.91)	1.31 (±0.88)
% Positives Unique	18.6	4.4
CoV	1.39	1.96

The above table displays the percentage of features found to be positive with each of probes 338F and 755, n = 6 (including dye-swaps where the probe is complexed with the alternative fluorophore) for hybridisations with a single probe and n = 2 for dual-probe hybridisations. The second row indicates the percentage of confirmed blanks

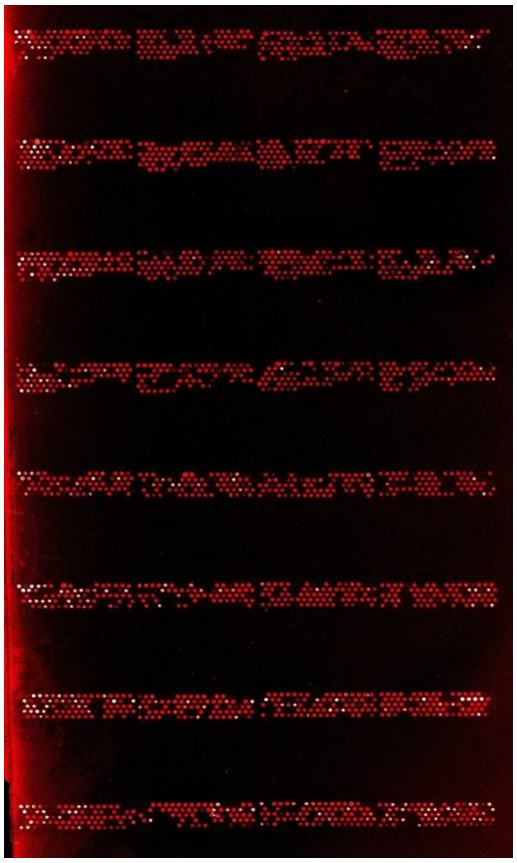
found to indicate hybridisation (false-positives), while the third row is applicable only to the dual probe hybridisation and is the percentage of hybridisation events unique to the respective probes. The coefficient of variation was calculated from the mean and standard error of the positive features across the replicates, values under 5% representing acceptable reproducibility. Importantly, neither dye-swaps nor dual-probe investigations introduced significant variability into the output. Few tests were undertaken with other probes, the focus of this phase being attainment of consistent signal; the close proximity of the T_m of 338F and 755 (55.4°C and 54.8°C respectively) led to their selection for this purpose.

It is clear that there is a considerable disparity between output from the array and the expected percentage of hits across the bacterial domain as predicted by the RDP. This is particularly evident for probe 755, expected to hybridise to less than 10%. However, the array is not representative of the entire bacterial domain but a microbiota sampled from patients with COPD. In this respect it would be surprising if the expected and observed figures showed correlation. However, the disparity could also be a result of the lower hybridisation temperatures utilised in attempts to attain a consistent level of signal intensity. In this sense, specificity of the probes had almost certainly been reduced, allowing hybridisation of the 755 probe with spotted clones which would not occur with increased stringency, although the post-hybridisation washes would be expected to reduce this mismatch hybridisation.

Concerns over specificity notwithstanding, the aims of Phase I had been achieved, but the lack of controls or reference features limited the scope of analytical approaches to the somewhat ‘fuzzy’ application of SNR. Subsequent phases would seek to rectify this deficiency and introduce greater statistical stringency to analysis.



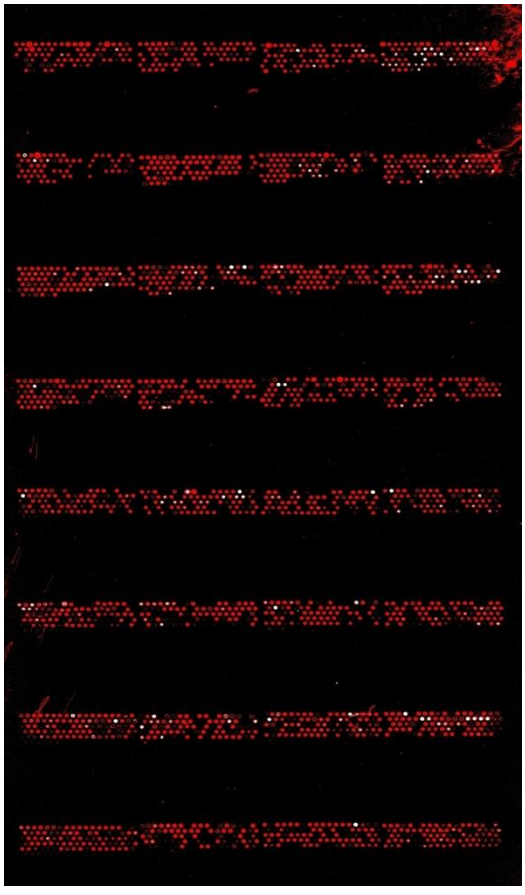
a) Cy3-338F



b) Cy5-338F

Figure 3.4 Phase I COPD arrays post-optimisation

c) Cy5-755



d) Cy3-338F and Cy5-755

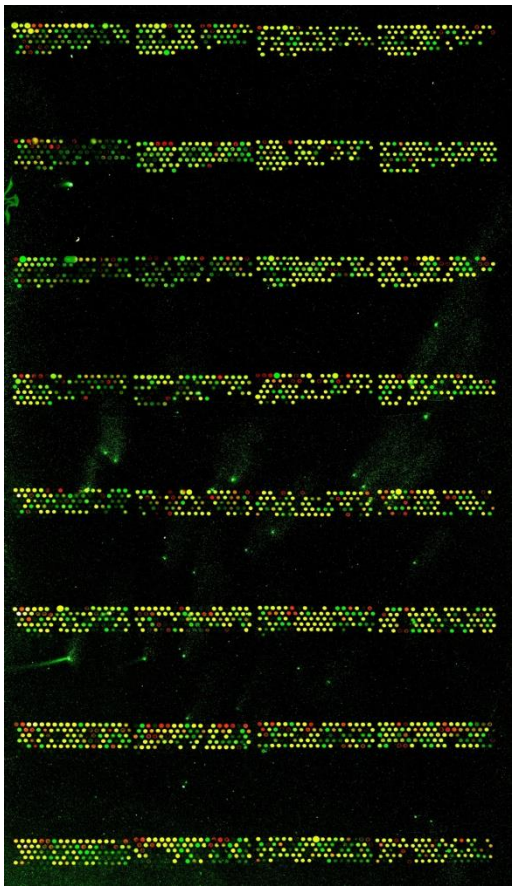


Figure 3.4 on the previous page shows single channel scans of a) Cy3-338F and b) Cy5-338F, the latter thus representing a 'dye-swap' for probe 338F. Figure c) shows a single channel scan of Cy5-755. Figure d) shows a dual-channel scan of the dual-probe hybridisation with Cy3-338F and Cy5-755, where yellow features are those spots with both probes hybridised to the target DNA.

3.4 Community PCR

Although not strictly associated with development of the OFARG methodology, PCR of heterogeneous bacterial DNA is an integral aspect of most approaches to microbial community analysis. Considerations as to minimisation of the production of extraneous artefactual bands (Wintzingerode *et al*., 1997; Acinas *et al*., 2005), maintenance of original species ratios (Wintzingerode *et al*., 1997; Suzuki and Giavannoni, 1996) and avoidance of chimera formation (Qiu *et al*., 2001) have led to a departure from standard PCR conditions in ecological studies. As such preliminary investigations were conducted to ascertain the optimal reaction parameters for preparation of amplicons.

3.4.1 Samples and extraction

Chicken faecal samples were obtained from Bristol Veterinary Research Centre. DNA was extracted from 200 mg of faecal material using methods as per section 2.6, obtaining concentrations of 50 to 250 ng/µl and absorbance values for A_{260}/A_{280} of between 1.6 and 2.0. DNA was eluted in ultrapure ddH_2O and stored at -20°C until use.

3.4.2 PCR

PCRs of the extracted DNA were performed under various conditions with a range of primers but generally according to the protocol described in section 2.8.2. Total amounts of template DNA ranging from 25-100 ng were utilised, along with total Mg^{2+} concentrations of between 1 and 4 mM. In addition, total cycle number varied between

10 and 30, while differing extension times (90-180 seconds) and primer annealing temperatures were investigated to maximise yield while maintaining specificity.

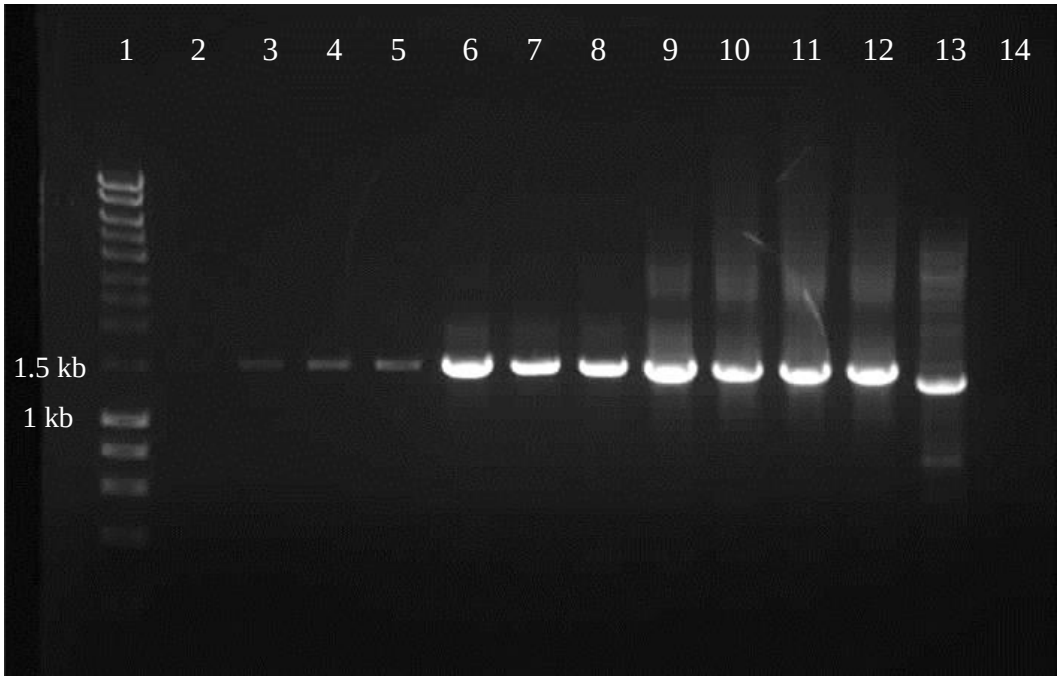
It was found that 25-50 ng of DNA and 20 cycles of extension were sufficient to provide quantities of DNA suitable for purification and ligation, while Mg^{2+} and annealment temperature were more specific to a primer pair. Figures 3.5 and 3.6 display results from a selection of optimisation PCRs.

Figure 3.5 shows the result of agarose gel electrophoresis (1.5% agarose, 100V) of a sample optimisation, revealing the presence of contaminating artefacts with increase in cycle number or template concentration, the former having the greater detrimental effect on the quality of PCRs. This observation is confirmed by the gels shown in figure 3.6 and subsequent community PCRs were conducted with the minimum number of cycles possible to obtain adequate yield for downstream procedures. Since total DNA in a sample does not necessarily correlate with the template available for 16S rDNA primer annealment, standardisation of input concentrations would only provide approximate equaliity for differing samples but was undertaken nonetheless. Cycle number, however, was subsequently maintained at 20 cycles unless samples proved resistant to amplification.

3.4.3 Libraries

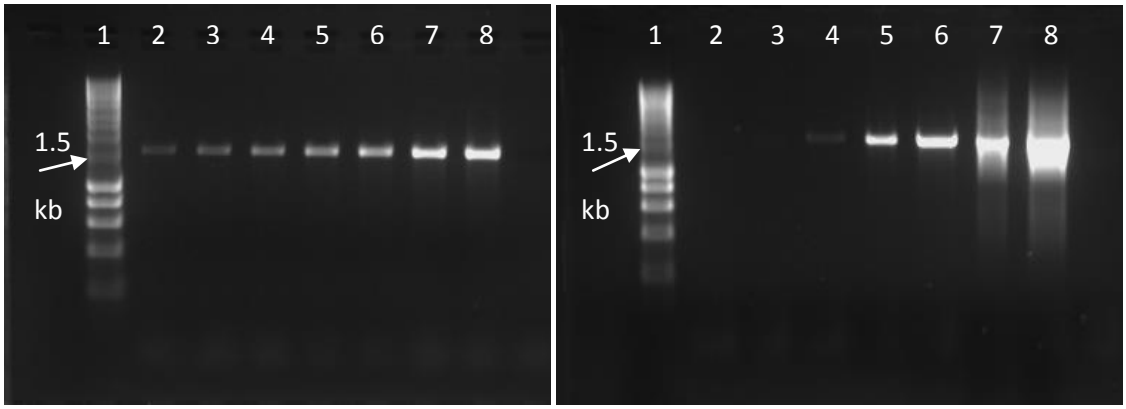
A number of small-scale tester libraries were created as per sections 2.8.2 through 2.8.5 using primer pair Bact-8F and Bact-1492RVA (table 2.1). Five white colonies were selected from each of 5 libraries and prepared for sequencing as per section 2.7.6, sequencing reactions being completed using primers pUC-F and pUC-R (table 2.1) as per section 2.7.8. Contigs for the sequenced clone inserts were created using EMBOSS (http://bips.u-strasbg.fr/EMBOSS/) and classified using the RDP. All clones were shown to contain bacterial 16S inserts and classified as such with greater than 95% confidence by the RDP.

Figure 3.5 Visualisation of optimisation PCR products on 1.5% agarose



Lane 1 = 5 µl Hyperladder I (Bioline). Lane 2 = 50 ng and 10 cycle PCR. Lanes 3, 4 & 5 = 15 cycle PCR of 25 ng, 50 ng and 100 ng respectively. Lanes 6, 7 & 8 = 50 ng and 20 cycles of PCR with 4, 3, 2 mM Mg^{2+} respectively. Lane 9 = 25 cycle PCR of 25 ng and 1.5 mM Mg^{2+}. Lanes 10, 11 & 12 = 30 cycle PCR of 25 ng with 1 mM Mg^{2+}. Lane 13 = 30 cycle PCR of 25 ng with 1 mM Mg^{2+}. Lane 14: negative control. Product is generally of 1.5 Kb in length, except lane 13 where a different primer pair was used, leading to a product of approximately 1.35 Kb.

Figure 3.6 Agarose gel (1.5%) visualisation of community PCR optimisations



Gel on the left shows: Lane 1 = Hyperladder I ; Lanes 2-8 = 5, 10, 15, 30, 50, 75 and 100 ng of template DNA for 20 cycles. Gel on the right shows: Lane 1 = Hyperladder I; Lanes 2-8 = 30 ng template DNA for 5, 10, 15, 20, 25, 30 and 35 cycles respectively. Cycle number has a stronger influence on quality of the 1.5 kb amplicon.

3.5 Improved array design rationale

Arrays utilised in the first phase of OFARG development had been produced from 16S libraries of COPD samples and were invaluable for preliminary optimisation of hybridisation protocols. However, evaluation of the hybridisation status of a spot was constrained to calculation of the individual SNR, with the threshold for 'positive' being set at three, based on the recommendation of the GenePix ® Pro Version 6.0 User's Guide (Molecular Devices).

An improved rationale for conversion of signal intensities to a binary hybridisation fingerprint was devised through incorporation of additional elements into the OFRG analysis approach. To assign fingerprints using OFRG the background intensity is first subtracted from the spot intensity before the adjusted value for an experimental probe is divided by that for a reference probe, for instance 338F, which is expected to hybridise to the majority of 16S sequences. Each experimental probe also has complementary (positive) control spots to which it is expected to hybridise, along with controls with which there should be no interaction (negative controls). The lowest value obtained for a probe with its positive controls is the threshold above which an intensity represents a hybridisation event; conversely the highest value for the probe with its negative controls establishes the negative threshold below which spots are considered as un-hybridised (Valinsky *et al*., 2002a). Intensities falling between these thresholds lead to assignment of a clone as 'unclassified' or 'N' (Valinsky *et al*., 2002a). The drawback of this approach is that many probes may then be uninformative for a given clone, limiting the potential for OTU binning, while OFARG in conjunction with the Array50 kits opens up multiple hybridisation possibilities not exploited previously.

The proposed approach would also commence with subtraction of the background intensity for each feature, but from both that of the reference and informative probe independently since the background intensity for a particular dye is found to vary.

Given the fact that intensities may also vary across the slide due to differential quantities of PCR product and stochastic experimental variability these figures can then be normalised by comparison with background-subtracted feature intensities from the 'neutral' control spots, designed to hybridise to the reference probe and whose concentration has been precisely controlled and evaluated. The average of hybridisation intensities for these features with the reference probe provides a value to which each spot intensity can be scaled. This then allows concomitant adjustment of the informative probe intensity for each feature by the same scalar factor, thus effectively normalising for concentration variation. Additional control spots would correlate with the positives from OFRG, showing complementarity to the experimental probe set; the average of these values per probe, minus 2SD (standard deviation), would be expected (assuming normal distribution) to encompass 95% of positives and would thus represent the positive threshold. The final control group are any spots not expected to hybridise to a given probe, these negatives being utilised to establish a lower threshold by adding 2SD to the average intensity of this group. The primary advantage of the adapted approach is the potential for two-tailed T-test assessment of a feature set against both the mean of the positive controls and that of the negative controls. Only if the mean is considered significantly different from both thresholds would the feature prove uninformative, the likelihood of either positive (1) or negative (0) classification being greater than that in previous OFRG studies which suffered from numerous 'N' classifications (unassigned). Standard assessment of a clone's assignment, though, would be whether the adjusted intensity lies above the positive threshold ('1') or below the negative threshold ('0').

To assess intensities reliably, three major factors must be taken into consideration. Firstly, the number of amplicons per feature may vary and thus intensity from a given feature requires standardisation or normalisation. For OFARG, adjusting for this variation is expected through scaling of feature intensities to the average of the

reference probe intensities for ‘neutral’ control spots, whose concentration is standardised. Secondly, it is estimated that mismatched probes will hybridise to clones; the extent to which this occurs will require empirical evaluation, and controls with a defined number of mismatches to each probe should eventually be included in the arrayed set. Lastly, signal output across an array is not entirely consistent for stochastic and technical reasons, so to prevent an array being uninformative for a given clone, replicates must be spotted in different regions of the slide, and, in any event, a single data point is statistically insignificant.

With the above in mind, subsequent arrays were designed as follows:

- features constituting the library of the sample microbiota with at least 3 print replicates of each PCR for each clone
- multiple neutral controls of verified concentration designed to hybridise only to the reference probe (338)
- sets of positive controls designed to hybridise to each of the individual probes
- sets of negative controls expected not to hybridise to the probes in question

The latter 2 groups are overlapping in that the positive control for probe X also serves as a negative control for probe Y, although the negative controls should also include entirely ‘empty’ spots (betaine/SSC only) and a number of PCR products of similar length but with no homology to 16S rDNA. The following sections outline the process for creation of arrays suitable for such analysis, and implementation of tests of the system in Phase II.

3.6 Control clones

OFARG system control features for phase II of development were created for the probe set detailed in table 3.3. These include a number previously utilised in phase I and additional oligos identified as potentially informative in both testing and final implementation. The BAC and FIR probes are relatively specific for the phyla *Bacteroidetes* and *Firmicutes* respectively, while the LAC probe targets lactobacilli and various members of other firmicute families. The CDProbe and CDSpec probes target *Clostridium difficile* and were included for their relevance to the research aim and with a view to testing the specificity of the system.

Table 3.3 Probes used in OFARG phase II

Probe	Sequence	Ref
338F	ACTCCTACGGGAGGCAGCA	**1**
755	CGAAGAACCTTACCAGGTCTTG	**2**
BAC	ATGGCACTTAAGCCGACACC	**3,5**
LAC	CAGCCTACAATCCGAACTGAGA	**3,6**
FIR	CCGAAGATTCCCTACTGCTG	**3,7**
CDProbe	CTCTTGAAACTGGGAGACTTGA	**4**
CDSpec	GGGAACTCTCCGATTAAGGAG	**3**

1 = Baker *et al*., 2003; 2 = Rob Free, PhD Thesis, 2005; 3 = Probebase (Loy *et al*., 2007); 4 = Gumerlock *et al*., 1991; 5 = Weller *et al*., 2000; 6 = Daly *et al*., 2003; 7 = Meier *et al*., 1999

The choice of probes then dictated design of the positive control features, which were created as described in section 2.8.4 through the simple expedient of ordering each oligo and its complement with an extra adenine residue at the 3' end; mixture of the two then

created a double-stranded complex which could be ligated into pGEM-T easy and transformed into *E. coli* DH5α. For addition of the 338F sequence into these clones the plasmids were isolated as per section 2.6.1, prior to digestion as per 2.6.7 with *Sac*I and *Spe*I, the sites for which both lie upstream of the T-cloning site (Appendix 1) and thus cause linearisation and excision without removal of the control sequence. The larger band was then extracted from agarose gels as per section 2.6.2, addition of the 338 sequence being through ligation of duplexes ordered as complementary oligos with tails to correspond to the *Sac*I and *Spe*I sites. All clones were verified as containing the desired inserts through sequencing.

Additional control clones of identified sequence and classification to genus level were obtained from test preparations of clone libraries with random selection or amplification of reference strains obtained through Dr Richard Haigh (University of Leicester). All procedures for clone production and verification were as per section 2.7. While all of the clones selected did not provide perfect matches to a probe they allowed for empirical evaluation of the degree of mismatch binding, and thus the specificity of the system; inclusion of CD-derived reference clones and utilisation of the CDProbe (table 3.3; specific for *Clostridium difficile* 16S) would further investigate this characteristic. In addition, reproducibility could be investigated through feature replicates and multiple hybridisations, along with variations in hybridisation conditions to assess optimal annealment temperatures; sensitivity could be evaluated through variation of control spot and probe concentrations.

The final set of controls were the negative group, these being either ‘empty’ spots or PCR products of a similar length to the 16S inserts, the latter produced through amplification of approximately 1.5 Kb of the *SFU1* gene from *Candida albicans*, primers and template DNA provided by Dr Alex Woodacre (University of Leicester). All control features were arrayed as multiple replicates.

Figure 3.7 Control clones inserted into the RDP classification tree

domain Bacteria

phylum "Firmicutes"

class "Bacilli"

Clone 17 *Bacillus spp.*

Bacillus cereus

Geobacillus stearothermophilus

Staphylococcus aureus

Clone 26 *Lactobacillus mucosae*

Enterococcus faecalis

class "Clostridia"

Clostridium difficile

Clostridium perfringens

Clone 2 "Ruminococcaceae"

Clone 6 "Ruminococcaceae"

Clone 8 "Ruminococcaceae"

Clone 9 "Ruminococcaceae"

Clone 20 "Ruminococcaceae"

Clone 21 "Ruminococcaceae"

Clone 23 "Ruminococcaceae"

Clone 29 "Ruminococcaceae"

Clone 31 "Ruminococcaceae"

Clone 32 "Ruminococcaceae"

Clone 10 Faecalibacterium sp

Clone 22 Faecalibacterium sp

Clone 24 Faecalibacterium prausnitzii

Clone 16 Subdoligranulum sp

Clone 19 Subdoligranulum sp

Clone 25 Papillibacter sp

Clone 4 "Lachnospiraceae"

Clone 7 "Lachnospiraceae"

Clone 14 "Lachnospiraceae"

Clone 27 "Lachnospiraceae"

Clone 28 "Lachnospiraceae"

class "Cocci"

Streptococcus pneumoniae

class "Erysipelotrichi"

Clone 11 *Erysipelotrichaceae* sp.

phylum "Bacteroidetes"

class "Bacteroidia"

Clone 1 *Bacteroides (fragilis)*

Clone 3 Bacteroides (fragilis)

Clone 15 Bacteroides (fragilis)

Clone 18 Bacteroides (fragilis)

Clone 13 Bacteroides spp.

Clone 5 Alistipes spp.
Bacteroides spp
Bacteroides fragilis
class "Flavobacteria"
class "Sphingobacteria"
phylum "Actinobacteria"
Mycobacterium tuberculosis CDC
Micrococcus luteus
Rhodococcus spp
phylum "Fibrobacteres"
phylum "Fusobacteria"
phylum "Proteobacteria"
class "Alphaproteobacteria"
Rhodospirillum rubrum
Ochrobactrum spp
class "Betaproteobacteria"
Neisseria meningitidis
class "Deltaproteobacteria"
class "Epsilonproteobacteria"
Campylobacter jejuni
class "Gammaproteobacteria"
Serratia spp
Enterobacter aerogenes
Haemophilus influenzae
Pseudomonas aeruginosa
Morganella morganii
Proteus mirabilis
Clone 30 Shigella/Escherichia coli
Clone 12 Escherichia coli
Escherichia coli
Enterobacter cloacae
Pseudomonas aeruginosa
Klebsiella pneumoniae
phylum "Spirochaetes"
phylum "Synergistetes"
phylum "Tenericutes"
phylum "Verrucomicrobia"

Figure 3.7 shows approximate coverage of the bacterial domain by control clones through insertion into the RDP classification hierarchy. Other phyla and certain classes shown are those encountered in a selection of studies of the human intestinal microbiota.

3.7 Phase II

Arrays for phase II of OFARG development were designed as described in the previous section, the primary set consisting of 7680 features per slide with multiple replicates of the clones listed in table 3.4. The focus of this stage was assessment of the analytical approach described previously, but while multiple hybridisations were conducted, relatively few slides were deemed suitable in terms of quality. Indeed, approximately 40% of arrays provided no viable scanned images, in part due to excessive evaporation of hybridisation buffer which manifests as dye hotspots extending into the feature region from the periphery (Bowtell and Sambrook, 2003).

The final hybridisations selected on the basis of subjective quality of scanned images were conducted at 53ºC using Cy3-338/Cy5-755 (Slide A) and Cy3-BAC/Cy5-338 (Slide B). A section of the 338 features had been designated as reference spots for the purposes of normalisation; concentration of these features had been standardised for array construction via spectrophotometric measurement and appropriate dilution prior to addition of spotting solution. Intensities from these features were utilised to establish a standardised figure for the 338 probe (subsequent to local background 'noise' subtraction) and a scalar factor was calculated for each included feature – certain features were flagged as 'bad' (excluded) in the initial stage of analysis where feature circumference, location and quality are manually assessed and, where necessary, adjusted. The scalar factor required to adjust 338 intensities to the normalised value was then applied to the backgound-subtracted value for the informative probe, in this case either 755 or BAC. Scaled intensities for the replicate probe control features were then averaged, with the standard deviation being subtracted from the mean to establish the threshold intensity for evaluation of hybridisation. Replicate intensities for each of the 'test' clones (clones with verified sequence) were used to calculate the mean array value for a clone for comparison to this threshold. Results are shown in tables 3.4 and 3.5.

Table 3.4 Intensities, % contiguity and hybridisation status for Slide A (755)

	Average	%Cont	Hyb		Average	%Cont	Hyb
Clone 1	810.45	45.45	N	**Clone19**	735.39	100.00	N
Clone 2	2431.93	100.00	N	**Clone20**	1151.64	77.27	N
Clone 3	-218.25	45.45	N	**Clone21**	2885.48	100.00	Y/N
Clone 4	4719.37	68.18	Y/N	**Clone22**	4329.08	68.18	Y/N
Clone 5	1077.09	45.45	N	**Clone23**	1453.86	77.27	N
Clone 6	2860.92	100.00	Y/N	**Clone24**	827.52	68.18	N
Clone 7	2450.20	63.64	N	**Clone25**	5446.87	77.27	Y
Clone 8	1571.29	77.27	N	**Clone26**	1949.02	100.00	N
Clone 9	629.26	63.64	N	**Clone27**	494.41	68.18	N
Clone 10	2466.63	68.18	N	**Clone28**	496.80	63.64	N
Clone 11	130.52	63.64	N	**Clone29**	3286.63	77.27	Y/N
Clone 12	683.02	63.64	N	**Clone30**	1909.82	63.64	N
Clone 13	583.08	45.45	N	**Clone31**	250.80	77.27	N
Clone 14	5772.98	68.18	Y	**Clone32**	-181.92	77.27	N
Clone 15	819.67	45.45	N	***E. coli***	2689.18	63.64	N
Clone 16	401.12	100.00	N	***C. difficile***	372.92	63.64	N
Clone 17	1486.79	45.45	N	***M. morganii***	383.17	63.64	N
Clone 18	500.46	100.00	N	***P. mirabilis***	1658.53	63.64	N

Table 3.4 shows the average feature intensity (Average; in scanner specific units), the maximum percentage contiguity (% Cont) with 755, and the estimated hybridisation status (Hyb) for each clone. Calculated mean value and standard deviation of the mean for the 755 probe were 7848.94 and 2572.41 respectively giving a threshold value of 2704.12 when 2SD was utilised for the calculation. Multiple clones with low contiguity were categorised as hybridised with this threshold, so 1SD was also subtracted from the

mean giving a more stringent threshold of 5276.53. However this eliminated clone 6 (100% contiguity for probe 755) leaving only clone 14 (maximum of 68% contiguity) in the 'hybridised' category. Indeed, of the 7 clones verified as containing a region of sequence with 100% identity to the 755 probe only two (clone 6 and clone 21) would be considered as hybridised even with the more lenient of the threshold values.

Table 3.5 Intensities, % contiguity and hybridisation status for Slide B (BAC)

	Ave	%Cont	Hyb		Ave	%Cont	Hyb
Clone1	3244.29	100.00	N/Y	Clone19	674.28	0.00	N
Clone2	765.14	0.00	N	Clone20	2144.01	0.00	N
Clone3	3214.96	100.00	N/Y	Clone21	1338.15	0.00	N
Clone4	2286.72	0.00	N/Y	Clone22	463.74	0.00	N
Clone5	1533.39	40.00	N	Clone23	322.81	0.00	N
Clone6	-402.74	0.00	N	Clone24	2003.48	0.00	N
Clone7	469.17	0.00	N	Clone25	874.61	0.00	N
Clone8	1551.56	0.00	N	Clone26	563.87	0.00	N
Clone9	712.96	55.00	N	Clone27	719.93	45.00	N
Clone10	892.27	0.00	N	Clone28	1768.48	0.00	N
Clone11	259.32	0.00	N	Clone29	638.14	0.00	N
Clone12	1848.89	0.00	N	Clone30	214.05	0.00	N
Clone13	4126.99	100.00	N/Y	Clone31	146.06	0.00	N
Clone14	414.95	0.00	N	Clone32	1613.32	0.00	N
Clone15	1065.56	100.00	N	*E. coli*	228.42	0.00	N
Clone16	1687.59	0.00	N	*C. difficile*	390.80	0.00	N
Clone17	1570.44	100.00	N	*M. morganii*	544.19	0.00	N
Clone18	3019.68	0.00	N/Y	*P. mirabilis*	418.65	0.00	N

Table 3.5 shows the same schema for the BAC probe, the mean being 7250.51 with a standard deviation of 2543.49. This provides thresholds of 4707.01 and 2163.52 using

1SD and 2SD respectively. Three of the 5 clones with regions of sequence displaying 100% contiguity to BAC are categorised as hybridised if the less stringent (2SD) threshold is employed. However, this also necessitates inclusion in this category of clone 18, which displays no sequence homology to the BAC probe whatsoever, as can be seen from table 3.6. The average intensity for the features representing clone 18 is close to that for clones 1 and 3 (100% contiguity for BAC), and 3 times that for clone 15 (also 100% BAC contiguity), so no simple means of setting a threshold that could be described as both sensitive (all expected positives are positive) and specific (all expected negatives are negative) presents itself. Attempts were made to calculate the threshold in a different manner, i.e. by dividing the intensity with the interrogative probe by that with the reference probe for each feature individually. However, this led to greater variance of the mean and apparent decreased specificity of the system.

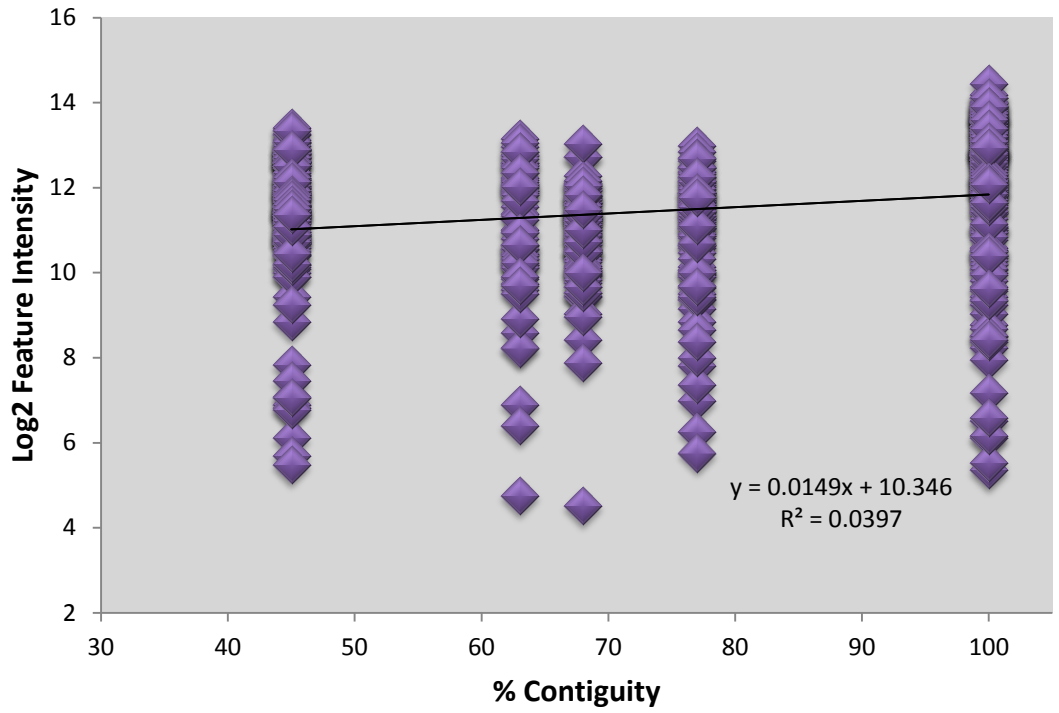
It is worth noting that each clone was represented by at least 80 replicates on the slide, derived from a minimum of 5 individual PCRs, so the discrepancy in average feature intensity for certain clones cannot easily be attributed to stochastic variation across the slide or technical errors. In addition, preliminary optimisations for experimental conditions in phase II had been conducted, with the finding that hybridisation was effectively abolished above 60ºC while quality of scanned images was poor (numerous nebulae and hot-spots) if hybridisation was conducted below 35ºC. However, within this temperature range variations in hybridisation, as assessed by reference probe (338) binding, were not evident. Hence, the temperature at which this set of hybridisations was conducted does not account for the variability.

In an attempt to discern whether there was a correlation between percentage identity (or maximum percentage contiguity) of the probe with a clone and feature intensity values from hybridisations were plotted as shown in figures 3.8 (for probe 755) and 3.9 (for BAC).

Table 3.6 % Contiguity (left) and % Match (right) of clones with phase II

	755	BAC	338	CDProbe	755	BAC	338	CDProbe
Clone1	45.45	100.00	100.00	0.00	81.82	100.00	100.00	0.00
Clone2	100.00	0.00	100.00	40.91	100.00	0.00	100.00	77.27
Clone3	45.45	100.00	100.00	0.00	81.82	100.00	100.00	0.00
Clone4	68.18	0.00	100.00	0.00	95.45	0.00	100.00	0.00
Clone5	45.45	40.00	100.00	0.00	86.36	90.00	100.00	0.00
Clone6	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00
Clone7	63.64	0.00	100.00	0.00	95.45	0.00	100.00	0.00
Clone8	77.27	0.00	100.00	40.91	95.45	0.00	100.00	40.91
Clone9	63.64	55.00	100.00	0.00	63.64	55.00	100.00	0.00
Clone10	68.18	0.00	100.00	0.00	95.45	0.00	100.00	0.00
Clone11	63.64	0.00	68.42	0.00	63.64	0.00	94.74	0.00
Clone12	63.64	0.00	100.00	0.00	95.45	0.00	100.00	0.00
Clone13	45.45	100.00	100.00	45.45	81.82	100.00	100.00	77.27
Clone14	68.18	0.00	100.00	0.00	90.91	0.00	100.00	0.00
Clone15	45.45	100.00	100.00	0.00	81.82	100.00	100.00	0.00
Clone16	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00
Clone17	45.45	100.00	100.00	0.00	81.82	100.00	100.00	0.00
Clone18	100.00	0.00	100.00	40.91	100.00	0.00	100.00	72.73
Clone19	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00
Clone20	77.27	0.00	100.00	0.00	95.45	0.00	100.00	0.00
Clone21	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00
Clone22	68.18	0.00	100.00	0.00	95.45	0.00	100.00	0.00
Clone23	77.27	0.00	100.00	40.91	95.45	0.00	100.00	40.91
Clone24	68.18	0.00	100.00	0.00	95.45	0.00	100.00	0.00
Clone25	77.27	0.00	100.00	0.00	95.45	0.00	100.00	0.00
Clone26	100.00	0.00	100.00	36.36	100.00	0.00	100.00	68.18
Clone27	68.18	45.00	100.00	0.00	95.45	65.00	100.00	0.00
Clone28	63.64	0.00	100.00	0.00	63.64	0.00	100.00	0.00
Clone29	77.27	0.00	100.00	0.00	95.45	0.00	100.00	0.00
Clone30	63.64	0.00	100.00	0.00	95.45	0.00	100.00	0.00
Clone31	77.27	0.00	100.00	0.00	95.45	0.00	100.00	0.00
Clone32	77.27	0.00	100.00	40.91	95.45	0.00	100.00	40.91
E. coli	63.64	0.00	100.00	100.00	63.64	0.00	100.00	100.00
C. difficile	63.64	0.00	100.00	0.00	95.45	0.00	100.00	0.00
M.. morganii	63.64	0.00	100.00	0.00	63.64	0.00	100.00	0.00
P. mirabilis	63.64	0.00	100.00	40.91	63.64	0.00	100.00	40.91

Figure 3.8a Log2 feature intensity vs % contiguity for 755



Figures show log2 of feature intensity plotted against maximum percentage contiguity of probe 755 with clones (a) or percentage match (b). Lines of best fit and R^2 correlation coefficient values are indicated on the figures.

Figure 3.8b Log2 feature intensity vs % match for 755

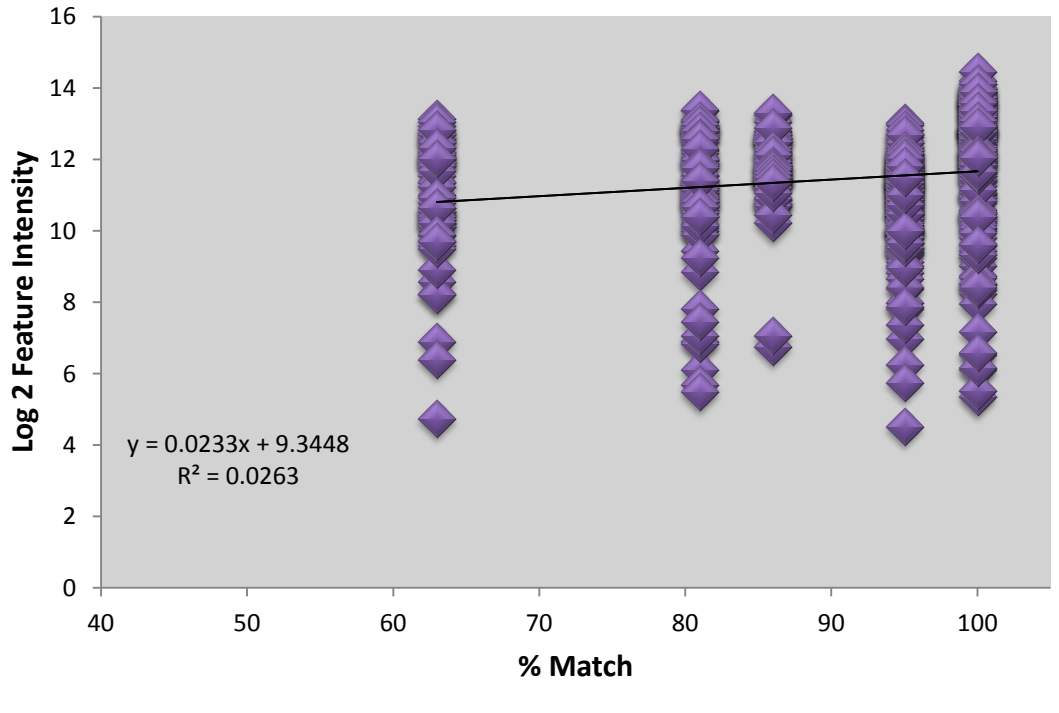


Figure 3.9 Log2 feature intensity vs % match for BAC

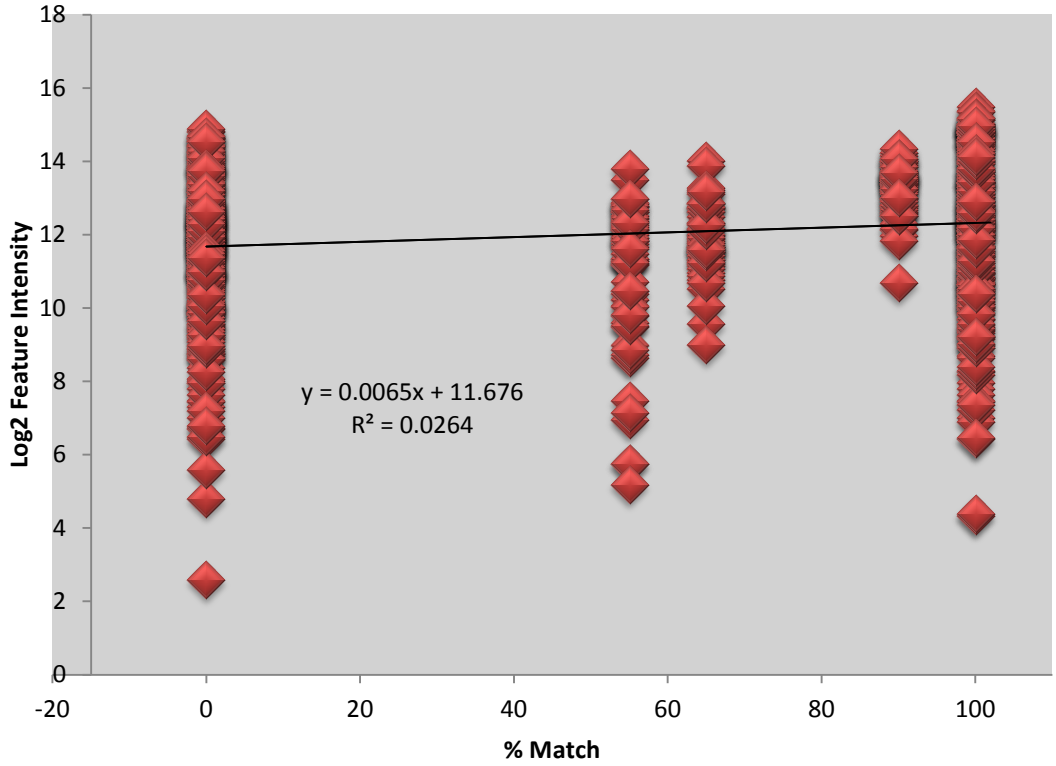


Figure 3.9 shows log2 of feature intensity plotted against percentage match for the BAC probe. A plot for contiguity is not included due to the close relationship of the two for the BAC probe. Plot also shows trendline and correlation coefficient R^2.

R^2 values indicate a very weak correlation between contiguity or percentage match and feature intensity. Since this should be the dominant factor in determination of hybridisation kinetics it is clear that significant non-specific binding of probe had occurred. With temperatures at the higher end of the viable range to minimise extraneous binding (Wetmur, 1991; Beattie *et al*., 1995), and only a single hybridisation buffer available, options for remediation were limited. However, there was the suspicion that concentration of DNA in the spots varied considerably from feature to feature which may then have been contributing to skewed intensities (Peterson *et al*., 2001). Use of the 338 probe as a 'reference' probe for standardisation of intensities had been intended to correct for this, but this introduces further potentially variable

hybridisation kinetics (Dai *et al*., 2002) into assessment of informative probe duplex formation: in effect, non-specific binding of 338 had the potential to amplify intensity variability caused by concentration of immobilised DNA.

Although the Genisphere 3DNA system is limited to two fluorophores, others including Alexa-488 are available, and it was possible to incorporate this into the M13R primer to prepare pre-labelled PCR products for array creation (section 2.8.2). Assessment of the required ratio of labelled to unlabelled primer for adequate detection without saturation was via titration of primer concentration and printing of a small scale array. An equimolar ratio of the two was deemed to be optimal based on subjective appraisal of the scanned image. Subsequently arrays were prepared such that all features were pre-labelled with Alexa-488 using the Ultra Marathon Arrayjet for slide printing. All clones for the fresh prints were as previously described but total feature number was reduced to 3872.

Arrays were pre-scanned at 488 nm excitation wavelength to assess suitability for hybridisation. Arrays for analysis were of hybridisations conducted at 53ºC utilising probes 338 and CDprobe. Slide C was challenged with Cy3-338 and Cy5-CDprobe while Slide D was the dye swap experiment for this combination where Cy3-CDProbe and Cy5-338 were employed. Hybridisation conditions and washes were as described in section 2.9. Subsequent to hybridisation slides were scanned at 488nm (Alexa 488), 532 nm (Cy3) and 635 nm (Cy5). Slide scans were assessed manually for quality of image and poor quality features were removed from further analysis. Final intensities for each feature with 338 and CDprobe were calculated by dividing the background-subtracted value for the probe by the equivalent value at 488 nm.

Figures 3.10 and 3.11 display scans of slides C and D respectively. I shows the single channel Cy3 scan at 532 nm, while II shows the dual-channel scan for both probes. III shows the scan at 488 nm to identify the labelled product spotted on the features and IV shows the single channel Cy5 scan at 635 nm. Slides C and D represent a dye-swap experiment for 338 and CDProbe.

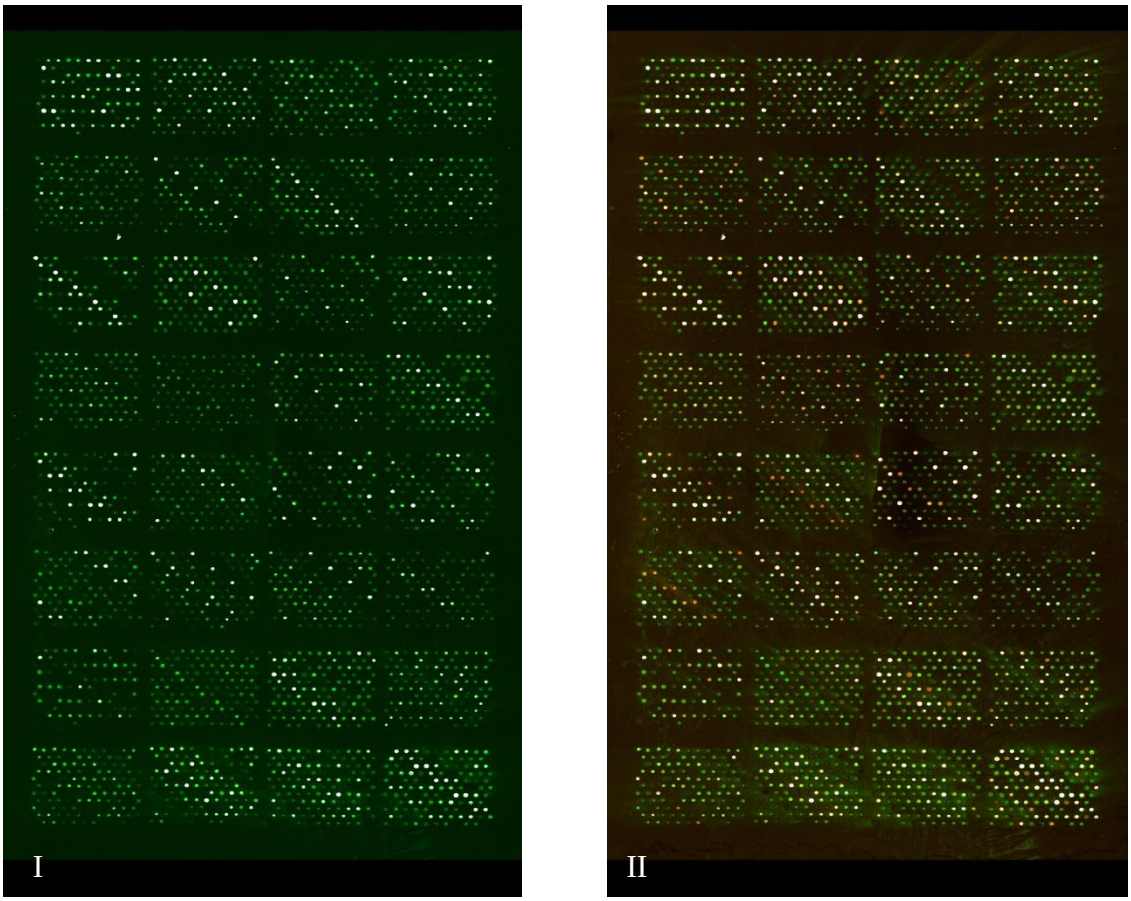
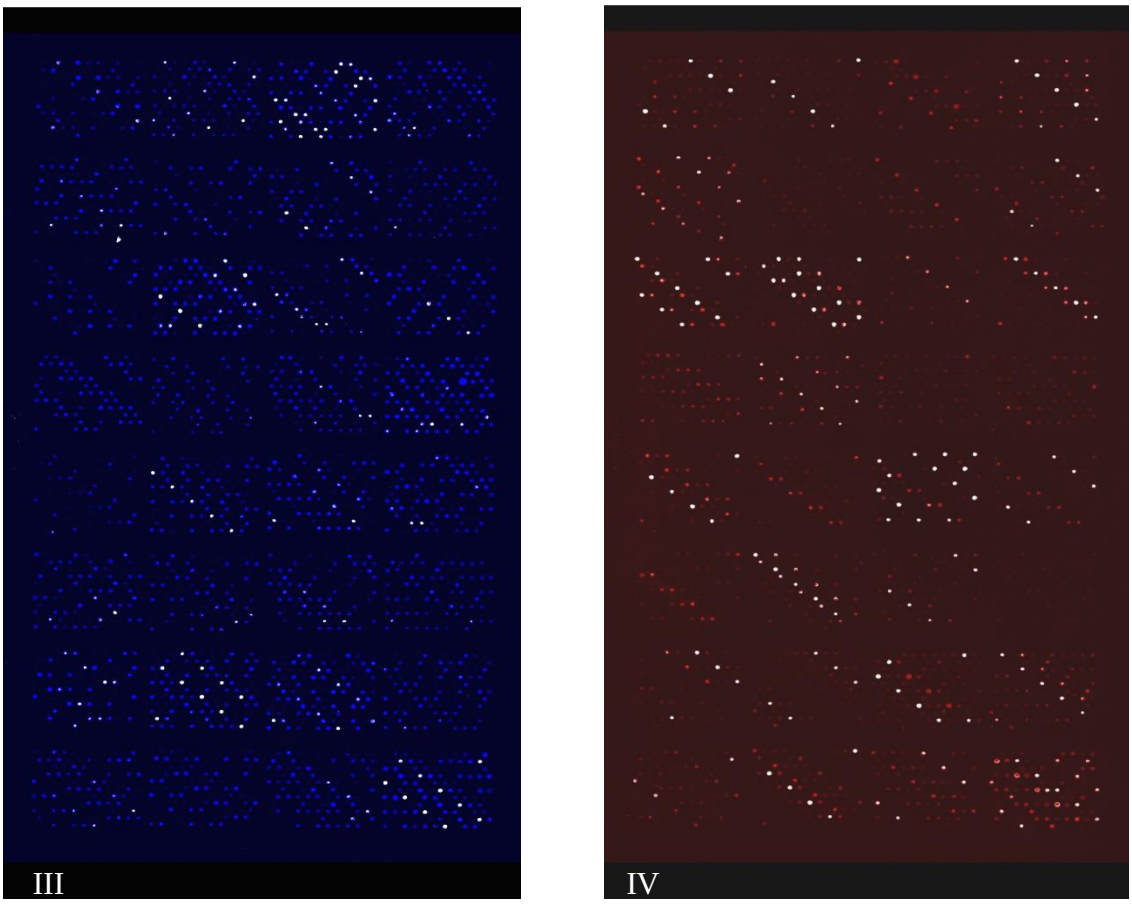


Figure 3.10 Scans of Slide C at 488, 532, and 635 nm



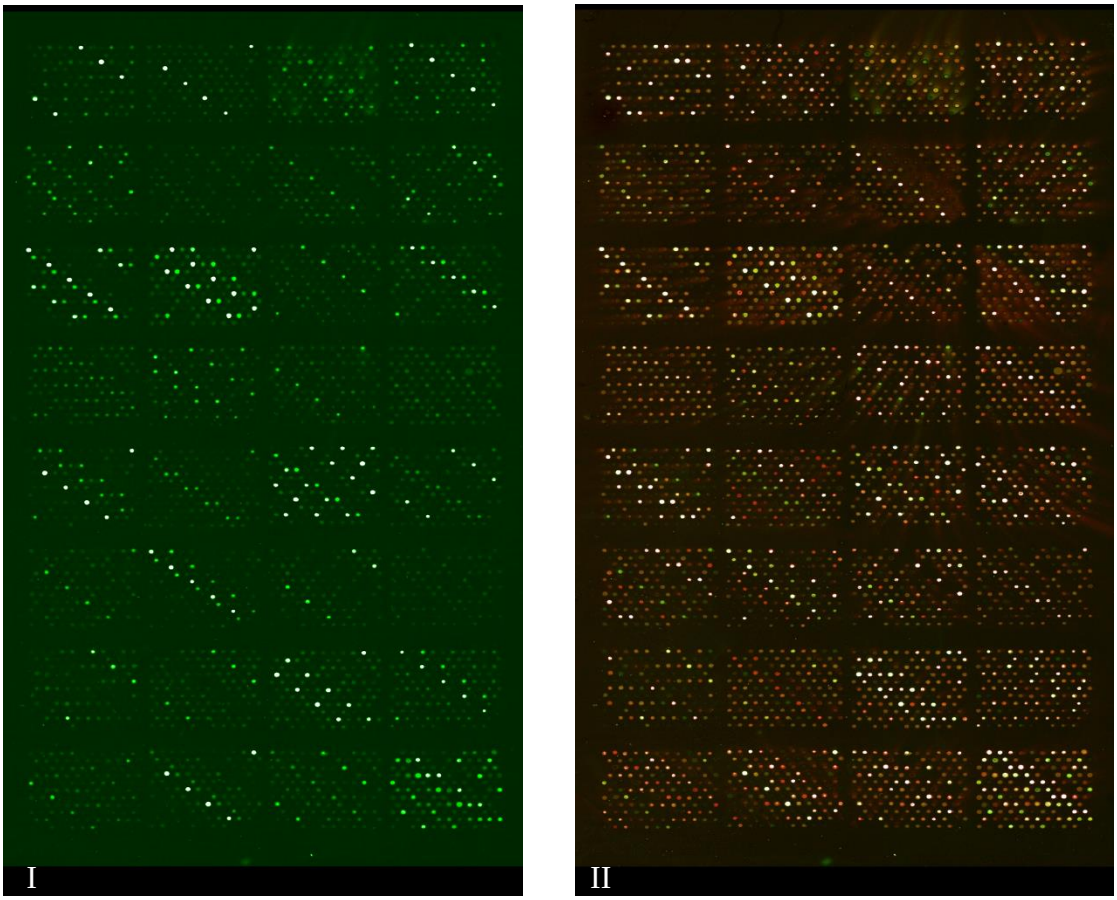


Figure 3.11 Scans of Slide D at 448, 532 and 635 nm

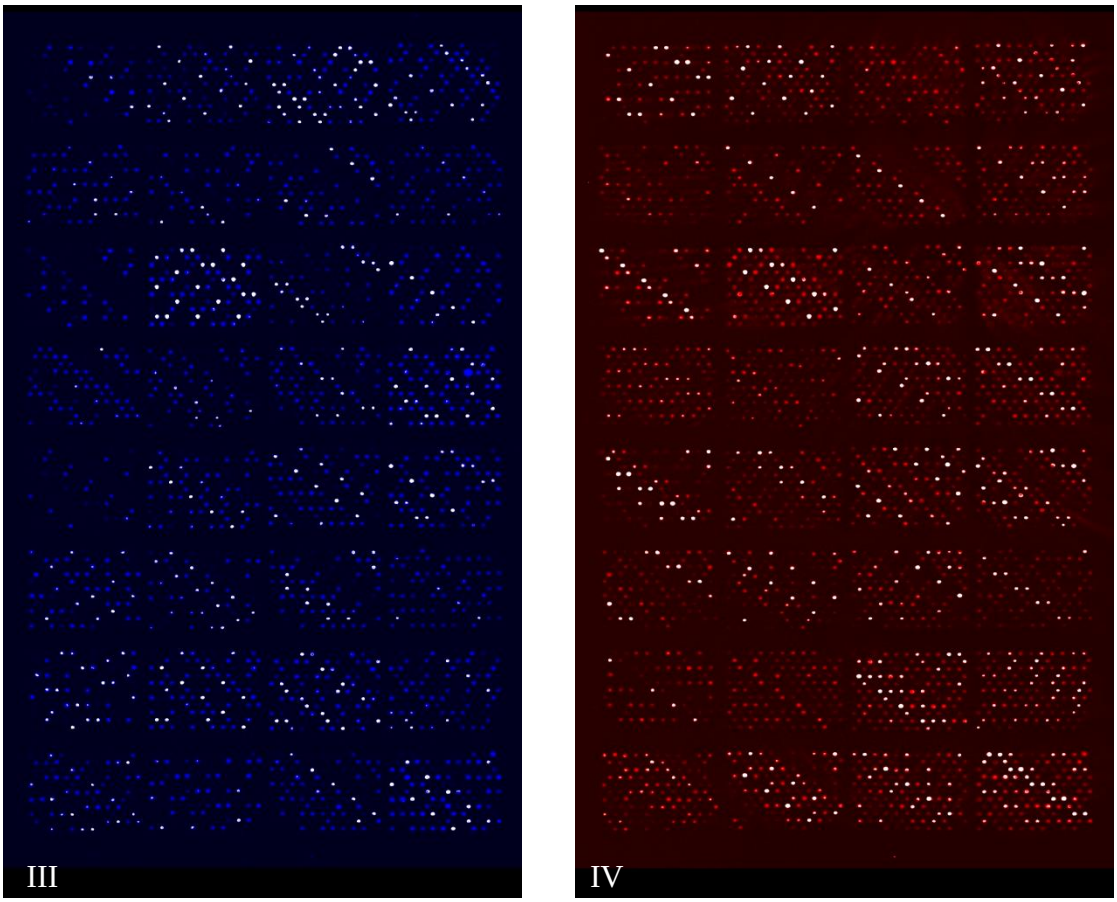


Figure 3.12 Slide C scaled intensity v % match for CDprobe

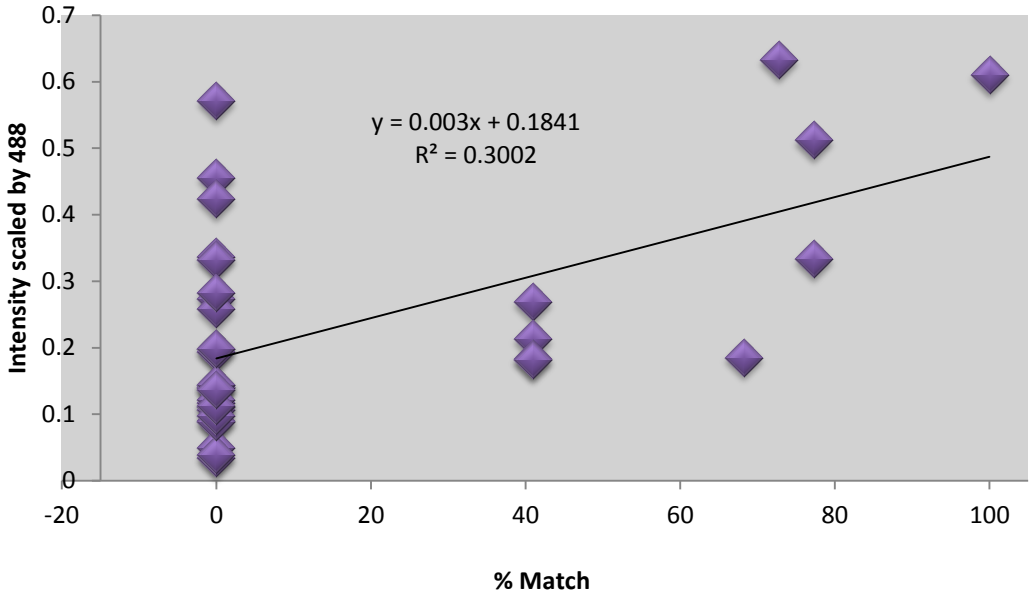


Figure 3.12 displays the mean of 488-adjusted intensities for each of the 36 clone replicate groups plotted against the percentage sequence match of that group with the CDprobe. Values in figure 3.12 are derived from slide C with CDprobe in the red (635 nm) channel. Figure 3.13 represents the same family of values but for slide D where the CDprobe was complexed to Cy3 and therefore fluorescing at 532 nm. Slide D is therefore the 'dye-swap' for C. Figures also show the line of best fit and the correlation coefficient (R^2).

Figure 3.13 Slide D scaled intensity v % match for CDprobe

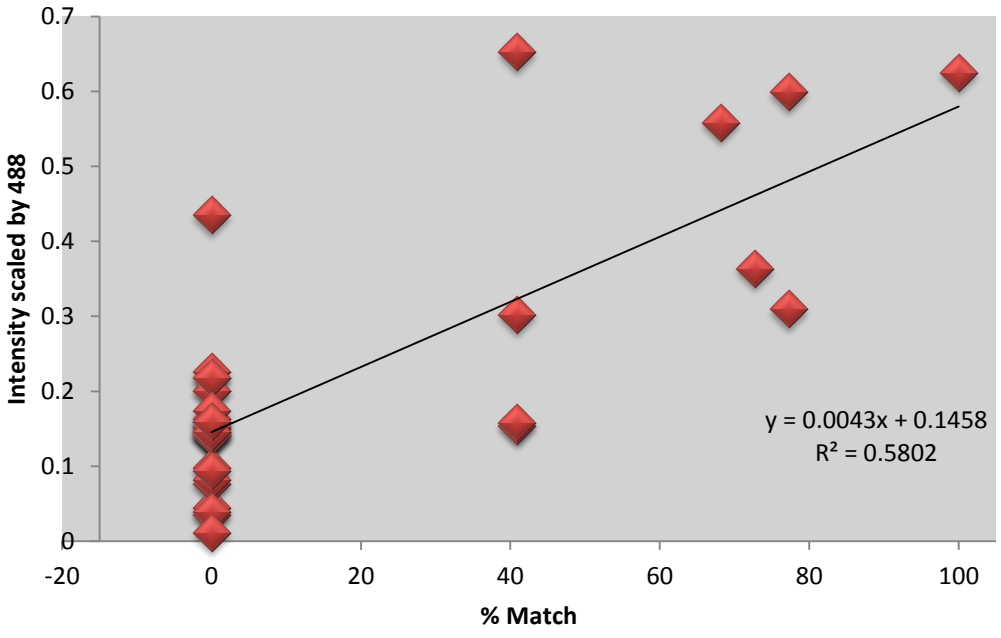


Table 3.7 Calculated hybridisation status and % match for clones with CDprobe

Clone	Match %	Mean C	Hyb C	Mean D	Hyb D
Clone1	0.00	0.2743	0	0.1418	0
Clone2	77.27	0.3335	0	0.3092	0
Clone3	0.00	0.259	0	0.2009	0
Clone4	0.00	0.3367	0	0.1431	0
Clone5	0.00	0.0341	0	0.2262	0
Clone6	0.00	0.1408	0	0.1495	0
Clone7	0.00	0.0494	0	0.1636	0
Clone8	40.91	0.2692	1/0	0.6524	1
Clone9	0.00	0.4561	1/0	0.1598	0
Clone10	0.00	0.1399	0	0.1528	0
Clone11	0.00	0.0901	0	0.0348	0
Clone12	0.00	0.0381	0	0.0441	0
Clone13	77.27	0.5118	1/0	0.599	1/0
Clone14	0.00	0.4232	1/0	0.201	0
Clone15	0.00	0.5714	1/0	0.2177	0
Clone16	0.00	0.3327	0	0.2175	0
Clone17	0.00	0.1444	0	0.0981	0
Clone18	**72.73**	**0.6334**	**1**	**0.3635**	**0**
Clone19	0.00	0.0882	0	0.0768	0
Clone20	0.00	0.1067	0	0.1414	0
Clone21	0.00	0.2818	0	0.1734	0
Clone22	0.00	0.1177	0	0.1402	0
Clone23	40.91	0.2132	0	0.1546	0
Clone24	0.00	0.1993	0	0.0812	0
Clone25	0.00	0.1116	0	0.1464	0
Clone26	68.18	0.1854	0	0.5582	1/0
Clone27	0.00	0.1932	0	0.1593	0
Clone28	0.00	0.1216	0	0.0942	0
Clone29	0.00	0.0981	0	0.1454	0
Clone30	0.00	0.1441	0	0.1587	0
Clone31	0.00	0.1981	0	0.0393	0
Clone32	40.91	0.1838	0	0.3021	0
C. difficile	100.00	0.6108	1	0.6247	1
E. coli	0.00	0.1122	0	0.0118	0
M. morganii	0.00	0.1364	0	0.4352	0
P. mirabilis	40.91	0.1819	0	0.1575	0

Subsequent to background subtraction (local setting, applied to individual features), and scaling of the CDprobe intensity by the 488 nm value for the feature, the mean and standard deviation for the CDprobe control spots were calculated:

- Slide C: Mean = 0.7916; SD = 0.1931; 1SD Threshold = 0.5985; 2SD Threshold = 0.4054
- Slide D: Mean = 0.7519; SD = 0.1373; 1SD Threshold = 0.6146; 2SD Threshold = 0.4773

The mean was also calculated for each clone replicate group and compared to both the relaxed (2SD) and stringent (1SD) thresholds. Results are displayed in table 3.7 with relevant percentage match of the clone to CDprobe.

The only confirmed positive from this experiment was the clone created from the SSU rDNA of *Clostridium difficile*, and the majority of those clones with no match to the probe were confirmed as negative. In addition, the SD values for the probe control replicates are smaller than previously encountered, indicating reduced variability in intensities and the improvement effected through scaling by a factor not subject to hybridisation kinetics.

However, there are some concerns in terms of reproducibility and specificity with regard to a number of feature groups, in particular clone 18, which has a region of sequence with 72.73% identity with CDprobe. Whether this clone would provide evidence of a hybridisation event was uncertain *a priori*, the identity being on the boundary of values (75-80% identity) at which cross-hybridisation may be considerable (Kane *et al*., 2000; Evertsz *et al*., 2001; Tomiuk and Hofmann,2001; Dai *et al*., 2009). Categorisation of the clone as either hybridised or not hybridised would have allowed incorporation of percentage identity into the establishment of thresholds and OTUs for the OFARG schema; however, a contradictory (mutually exclusive) outcome for the dye-swap experiments raised concerns about the reliability and potential of the

methodology, especially in view of the small number of clones being investigated, all of which were of verified sequence and with mutiple (>50) technical replicates per array.

Although focus was drawn to clone 18 due to the polarised nature of the result, figures 3.12 and 3.13 display that the correlation coefficients differ for the two dyes, while none of clones 8, 26, 13, 14, 15 and 9 provided a definitive hybridisation status despite zero identity with CDprobe for the clones 14, 15, and 9. As a whole the results suggested that the OFARG system was far from optimised, with each probe's hybridisation dynamics likely to require further extensive empirical modelling; potentially, each probe would eventually demand differing hybridisation conditions such as constitution of buffers and temperature. In addition, the causes of extensive non-specific binding and dye-dependent variability would need to be established and remediation effected.

Time and funding constraints (Genisphere 3DNA kits for 25 arrays cost £350 for instance) dictated that such an undertaking would not be possible, and further experiments utilising the OFARG system were not conducted.

Despite the discontinuation of the OFARG approach, the probe set theoretically represented a valuable tool for future investigations, especially given enough resources to optimise the hybridisation conditions for each individual probe. As such, the development of the set, and the final list, are described briefly in the subsequent section.

3.8 Probe design

One of the challenges which presented itself in the current undertaking was the evaluation of probes which maintain specificity under comparable hybridisation conditions. OFRG studies utilised a simulated annealing algorithm for assessment of probe hybridisation conditions (Borneman *et al*., 2001), but successive hybridisations often demanded significantly different compositions of hybridisation buffers and wash times. The aim here was to develop a probeset capable of maximal 16S differentiation under equivalent hybridisation conditions, thus minimising the number of arrays required to assign an OTU fingerprint to a clone. It was originally intended that the entire set for the human intestinal microbiota would comprise approximately 20 oligos.

The initial catalogue of probes was compiled from three sources: probeBase (www.microbial-ecology.net/probebase/; Loy *et al*., 2007); literature related to 16S primers and probes such as Baker *et al*., 2003; and the output of probe-design software devised and coded by Dr Rob Free (University of Leicester, unpublished work). The last of these can be used to align multiple 16S sequences and create probes from a degenerate consensus sequence based on the frequencies of bases at aligned positions. As such its output is confined to certain limited regions and probes may both overlap and display a high degree of similarity, with perhaps only two or three variant positions in a 20-mer oligonucleotide. Nonetheless, the combined total of some 2,500 oligonucleotide sequences provided an extensive initial database for *in silico* evaluation. In addition to the 'probe' sequences a further reference database of 16S sequences was compiled against which the oligos could be challenged to assess hybridisation potential. Based on a review of relevant literature, (Eckburg *et al*., 2005; Wilson, 2009; *inter alia)* a total of 264 full-length 16S sequences were selected and downloaded from the RDP to represent the spectrum of the human intestinal microbiota.

The probes were then queried against the reference sequence database using a downloaded version of the Blast software (http://www.ncbi.nlm.nih.gov/blast; blast-2.2.23+; Altschul et al., 1990) to provide a textual output of matches and maximum contiguity. This file was then utilised as input for the 'ProbeScript' (Appendices 6 and 7) coded in Perl for this project. The 2 versions of this program both initially create a binary output database with all sequences in rows and probes in columns. Subsequent code examines the columns to identify probes with identical outputs, replicates being removed so that no two probes have the same fingerprint for the reference sequences. Combined with manual examination of the remaining probes to retain only those with 19 ± 5 nucleotides, or melting temperatures of 50 ± 12°C, the probe set was reduced to 847 through this first iteration. The probescript also provides an output identifying probes which are of limited discriminatory value i.e. hybridising to a high/low proportion of the sequences, or providing a similar profile to other probes. These were then filtered separately over the course of multiple iterations, eventually providing a list of 113 probes which could differentiate the entire reference set. Despite randomisations of the dataset this could not be reduced further so the reference sequence list was edited to include fewer members of the same genus. The resultant 214 species are detailed in Appendix 5 and formed the basis for a final round of iterations in which sequences were also 'binned' in OTUs; the ultimate 141 unique 'fingerprints' for the perfect match script were distingushed with 26 probes, although 5 of these were included to identify particular species considered important (*C. difficile* and *Faecalibacterium prausnitzii*) or differentiate particular groups (e.g the class *Beta-proteobacteria*), while the last was the reference probe from earlier phases, 338F. Use of the contiguity script with filter set for 70% results in discrimination of the RDP reference set with the creation of 179 OTUs. Table 3.8 describes the final probe set including melting temperatures for final pairing, 16S binding region and RDP hits with zero, one, and two mismatches.

Table 3.8 Final probelist, showing sequence, source, annealing site on 16S, melting temperature, and RDP hits with mismatches

ID for this study	Sequence	Reference	Site	Tm	RDP 0	RDP 1	RDP 2
pB01001	ACTGCTTGTGCGGGCTCC	ProbeBase- Loy *et al*., 2007; Daly *et al*., 2006	926-943	54.9	781	39443	721318
pB00718	AAACCACACGCTCCGCT	ProbeBase- Loy *et al*., 2007; Gich *et al*., 2001	958-941	49.5	8926	37053	60593
pB01041	TTCTTCCTAATCTCTACGCA	ProbeBase- Loy *et al*., 2007; Kusel *et al*., 1999	864-883	47.7	10334	32547	167192
pB00300	CGGCGTCGCTGCGTCAGG	ProbeBase- Loy *et al*., 2007; Amann *et al*., 1990	385-402	59.4	26167	274924	553241
pB00543	ACCGCTTGTGCGGGCC	ProbeBase- Loy *et al*., 2007; Liu *et al*., 2008	927-942	52.6	376614	851213	956728
pB00428	ACGGGCGGTGTGTACAAG	ProbeBase- Loy *et al*., 2007; Loy *et al*., 2002	1389-1406	52.6	367664	430849	464745
Robs 26	TGCATGGCTGTCGTCAGCTCGTG	This study	1050-1072	60.6	455409	927733	951601
Robs 138	GGCTAACTCCGTGCCAGCAGC	This study	501-521	60.2	300796	1106135	1223265

Robs 495	CCTGGGGAGTACGACCGCAAGG	This study	855-876	62.3	166963	552192	629554
Robs17	TACGGGAGGCAGCAGTGGGG	This study	343-362	60	764993	1175512	1223745
338	ACTCCTACGGGAGGCAGCA	ProbeBase- Loy *et al*., 2007; Amann *et al*., 1990	338-357	55.4	1166134	1223596	1249252
755	CGAAGAACCTTACCAGGTCTTG	This study	968-946	54.8	103793	271080	554598
BAC (CFB1082)	ATGGCACTTAAGCCGACACC	ProbeBase- Loy *et al*., 2007; Weller *et al*., 2000	1081-1100	53.8	44717	47920	56745
LAC	CAGCCTACAATCCGAACTGAGA	ProbeBase- Loy *et al*., 2007; Daly *et al*., 2006	1295-1317	54.8	51296	201548	318952
CDProbe	CTCTTGAAACTGGGAGACTTGA	Gumerlock *et al*., 1991	692-713	53	319	786	3240
Gamma-proteobacteria	GAAGCCACGCCTCAAGGGCACAA	Shen *et al*., 2010	856-834	60.6	19100	20858	27300
Faecalibacterium	CCTCTGCACTACTCAAGAAAAAC	Suau *et al*., 2001	645-667	53.5	11238	11713	12763

Clep1240	CGTTTTGTCAACGGCAGTC	ProbeBase- Loy *et al*., 2007; Sghir *et al*., 2000	1240-1257	51.1	28832	31594	290477
Erec482	GCTTCTTAGTCAGGTACCG	ProbeBase- Loy *et al*., 2007; Franks *et al*., 1998	482-500	51.1	61684	76745	116665
Beta 1	CCCATTGTCCAAAATTCCCC	Ashelford *et al*., 2002	359-378	51.8	97220	273354	669726
Bifido	CCACCGTTACACCGGGAA	Langendijk *et al*., 1995	662-679	52.6	2050	20857	103694
IV815	CCCACACCTAGTAATCATCGTT	ProbeBase- Loy *et al*., 2007; Daly *et al*., 2006	815-836	53	8546	38354	106860
FIR	CCGAAGATTCCCTACTGCTG	ProbeBase- Loy *et al*., 2007; Meier *et al*., 1999	354-371	53.8	48317	178627	266194
574	GGGCGTAAAGCGTGCGCA	This study	574-591	54.9	86365	257937	555358
1239	GGGCTGCACACGTGCTACAAT	This study	1239-1260	56.3	16631	515509	723367

3.8 Summary

Despite considerable efforts with the OFARG system, including the creation and hybridisation of more than 200 arrays, the approach could not be optimised to fulfill its hypothetical promise. It was visualised as a relatively cheap and high-throughput alternative to the sequencing of hundreds of clones, which could be applied to identification of the dynamics of the intestinal microbiota under differing conditions. In this sense, absolute phylogenetic characterisation of clones was neither essential nor intended; consistency of assignment of a given clone (or species) to an OFARG OTU was, however, fundamental.

While initial technical issues were overcome, and approximately 3 of 4 hybridisations were eventually providing scanned images suitable for analysis, the expectation that a set of hybridsation conditions could be devised for universal application with regard to the probe set may have been unrealistic. The widely accepted paradigm states that temperature and buffers are significant in determining the specificity of probe behaviour on microarrays (Miller *et al*., 2002; Relógio *et al*., 2002), although some groups have found that non-specific binding is not as intimately related to reaction conditions as believed (Evertsz *et al*., 2001). In truth, many factors may influence specificity of probe duplex formation (Relógio *et al*., 2002) and it is likely that the relative contribution of any one of these is system-specific. Thus, continuation of OFARG development necessitated time and resources, both of which were eventually in short supply.

In closing, though, it is perhaps noteworthy that the most recent appearance of OFRG in the scientific literature ,in relation to characterisation of microbial communities, was more than 5 years ago (Bent *et al*., 2006); in addition, the technique at the time involved sequential use of the probes comprising the set, each with its own demands in terms of experimental conditions such as buffers and temperature (Bent *et al*., 2006).

CHAPTER 4

PRELIMINARY 454

4.1 Introduction

The massively parallel sequencing (MPS or second-generation) technologies briefly mentioned in Chapter 1 have significant potential for the analysis of complex microbial communities, not least due to their economic advantage over Sanger sequencing (Schuster, 2008); indeed, the Human Microbiome Project (HMP), in which millions of dollars has been invested in order to characterise the microbial communities associated with humans, drew heavily on their availability (Turnbaugh *et al*., 2007). While platforms such as 454 (Roche Life Sciences), Solexa/Hi-Seq (Illumina Inc.), SOLiD™ 4 (Applied Biosystems) and others utilise differing chemistries and approaches (Metzker, 2009), all rely on the parallel examination of multiple, clonally-amplified populations of the DNA under investigation, allowing for higher throughput than is possible with capillary sequencers (Mardis, 2008).

The choice of platform is clearly informed by the aims of the research: the Solexa and SOLiD systems provide a large number of sequence reads and data volumes of around 30 Gb (Metzker, 2009) but at the cost of shorter reads, approximately 35 to 100 bases (Mardis, 2008; Pallen *et al*., 2010); 454 allows for sequence output of approximately 400 bases per read but a total of only 0.5 Gb of sequence data (Metzker, 2009). While there is considerable overlap on their application in microbiology the former provide greater coverage and are ideally suited to elucidation of variants via genome resequencing, while the longer reads of 454 are more suited to *de novo* genome assembly (Metzker, 2009). Such a complementary application was employed by one group using 454 (and Sanger) to sequence the genomes of six *Clostridium difficile* strains before examining SNP variation between more than 20 isolates on the Illumina platform, the resultant data leading to the suggestion that the pathogenicity of this species may have developed independently in multiple lineages (He *et al*., 2010).

In the fields of metagenomics and microbial ecology the 454 system has generally been favoured as the longer read lengths permit differentiation of sequences at lower taxonomic levels approaching that of species. Studies have been conducted on microbial niches ranging from deep-sea environments (Sogin *et al*., 2006) to the intestinal milieus of macacque monkeys (McKenna *et al*., 2008) and humans (Andersson *et al*., 2008); one study even analysed the species present in material impacted on the windscreen of a car over the course of a 300-mile journey (Pond *et al*., 2009). More recently, though, improvements in the read-length available with the Illumina platform have led to an increase in its application to microbiomic studies of, for instance, the human gut (Qin *et al*., 2010) and oral flora (Lazarevic *et al*., 2009).

At the inception of this course of research the choice of 454 was influenced primarily by its predominance in the literature for microbiomic investigations in comparison to other NGS (Next-Generation Sequencing) technologies. Further support for use of the 454 platform was provided by the decision to use 16S rDNA as the target gene for classification based on the availability of multiple comprehensive databases (RDP, Greengenes and ARB-Silva). In addition, the purchase of 454 sequencing equipment by the University of Leicester (Genomics services, Department of Genetics) allowed for on-site optimisation of the procedures and protocols.

The essence of 454 is the parallel analysis of tens of thousands of clonally amplified products, each of the 1.6 million wells of the plate containing a bead which has previously undergone emulsion-based PCR from a single nucleotide fragment (or 'denatured' amplicon) such that there are eventually in the region of 10^7 copies per emulsion droplet (Margulies *et al*., 2005). Each fragment becomes attached to its bead via a system-specific oligonucleotide which can be incorporated into primers and then forms the basis for extension and calibration (Margulies *et al*., 2005). Nucleotides are

allowed to flow sequentially across the system, incorporation causing release of pyrophospate (PP_i), which is then acted on by sulfurylase and luciferase to release light (Ronaghi *et al*., 2000). The chemistry of the 454 sequencing reaction is displayed in Figure 4.1, while the entire procedure is represented digrammatically in Figure 4.2.

An integral aspect of the parallel approach, and the basis for the high throughput that can be attained, lies in the proprietary sequences which mediate attachment of target DNA to beads for deposition in wells, and prime both emPCR and sequencing. For genomic applications these are known as adapters and are ligated to regions of sequence subsequent to fragmentation; for methodologies which require an initial amplification step they are described as 'fusion' sequences, and can be incorporated at the 5′ end of the sequence-specific region of oligonucleotide PCR primers during synthesis. In addition, though, this allows for multiplexing of sample libraries for pooling on regions of a chip, each sample having a unique 'barcode' identifier between the sequence-specific region of the primer and the 454-fusion sequence (figure 4.3), allowing for deconvolution of data at the post-processing stage (Andersson *et al*., 2008).

Figure 4.1 Chemistry of the 454 sequencing reaction

$$(DNA)_n + dNTP \xrightarrow{DNA\ polymerase} (DNA)_{n+1} + PP_i$$

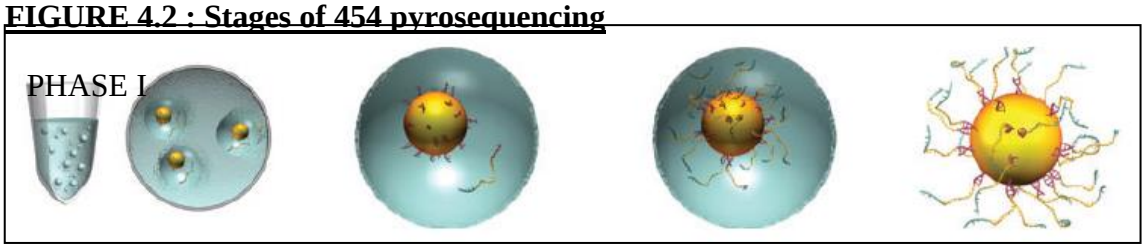
$$PP_i + \text{adenosine 5′ phosphosulphate} \xrightarrow{ATP\ sulfurylase} ATP + SO_4^{2-}$$

$$ATP + \text{luciferin} + O_2 \xrightarrow{luciferase} AMP + PP_i + CO_2 + \text{oxy-luciferin} + \text{protons}$$

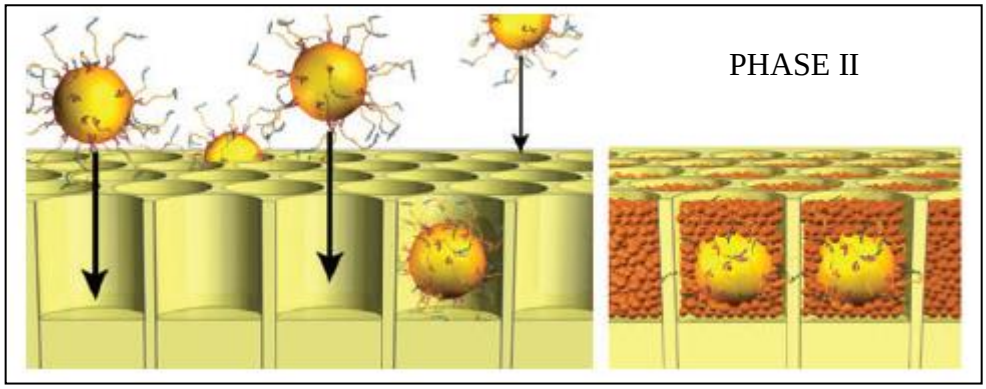
$$ATP \text{ and } dNTP \xrightarrow{apyrase} AMP \text{ and } dXMP + 4P_i$$

Figure 4.1 shows the sequential reactions of the 454 sequencing process leading to the release of protons which are detected to signify the incorporation of nucleotides (Ronaghi et al., 2000)

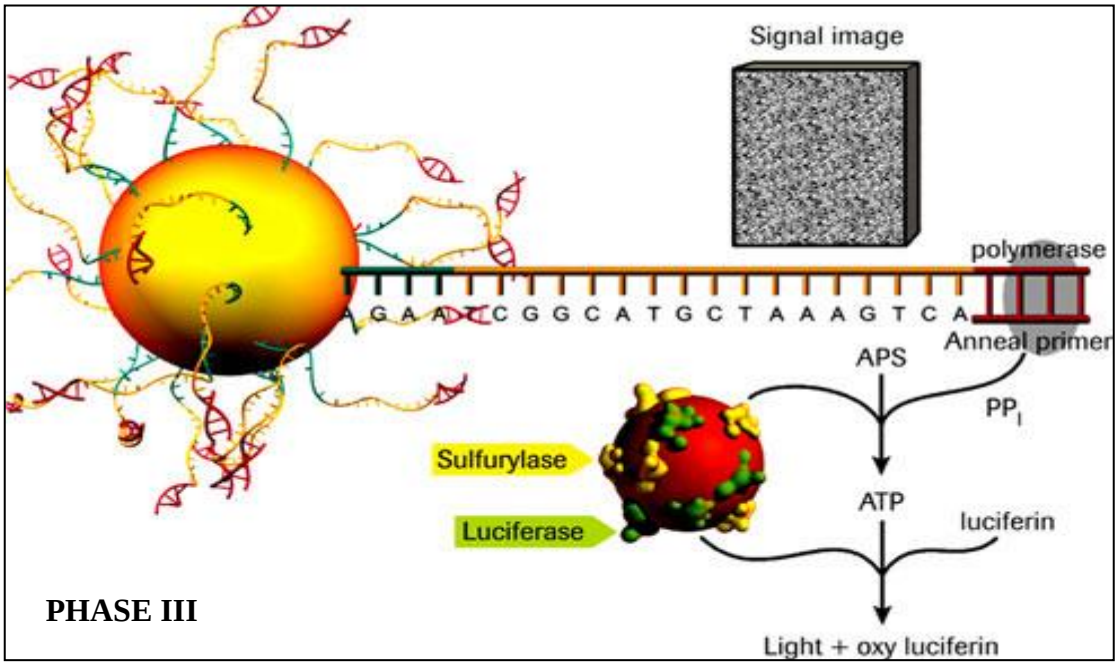
FIGURE 4.2 : Stages of 454 pyrosequencing



Phase I shows the emulsification of beads and single-stranded DNA (in oil and water) to create PCR 'micro-reactors', leading to a population of beads with multiple copies of a single template. Assessment of the correct ratio of concentration of sample to number of beads, (such that the majority of beads are associated with a single molecule prior to emPCR) is performed by titration. An enrichment step selects for beads with attached DNA.



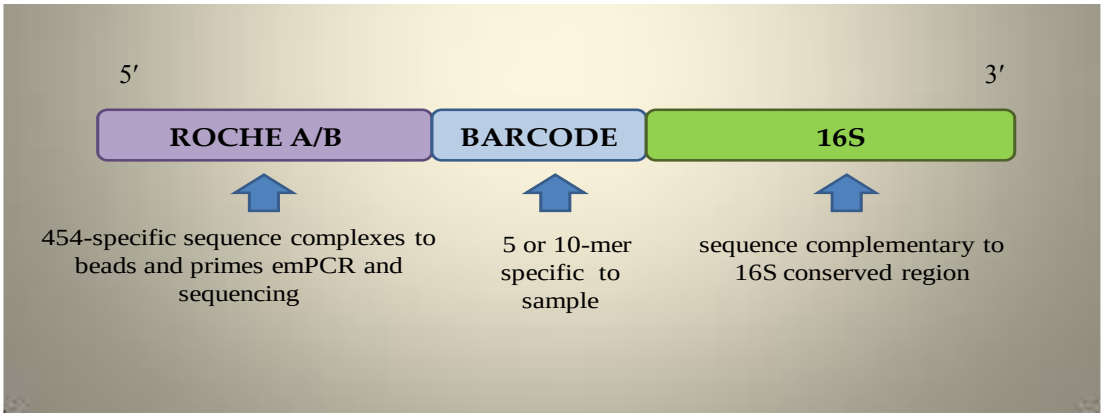
In Phase II the enriched 'beads' are loaded onto the PTP (Pico Titer Plate) where the diameter of the wells (40 µm) ensures that only one bead may occupy a single site; wells are packed with the enzymes and other reactants necessary for the sequencing process. Phase III shows the reaction itself whereby incorporation of a nucleotide causes emission of light and detection by the scanning equipment.



Reproduced and adapted with permission from Roche (McLeod, 2009).

Figure 4.3 Primer design incorporating barcodes for multiplexed 454

Figure 4.3 shows 16S-specific region at the 3′ end of the complete oligonucleotide with barcode (or tag/MID) between this and the 454-specific



Once primers for amplicon production have been selected preparation of samples for submission is relatively straightforward. Steps from pooling of samples onwards (i.e. emPCR and sequencing) are generally undertaken by the sequencing centres, with eventual return of flowgrams and/or sequence read files; the challenge, as in metagenomics studies conducted using any platform, is presented by structured and meaningful analysis of the GBs of output data.

The eventual objective was to utilise 454 sequencing for comparison of the intestinal microbiota of numerous CDAD patients with subjects diagnostically free of infection subsequent to antibiotic treatment, either with or without diarrhoea. However, the relative novelty of the technology and the potential complexity of analysis, coupled with the expected difficulty in obtaining an appropriate set of samples, informed the requirement for preliminary projects using a range of subjects. Each of these would provide the opportunity for refinement of the methodological workflow and empirical assessment of such key factors as primer selection, while addressing subsidiary biological questions relating to the dynamics of diverse intestinal microbiota.

4.2 Methods

The following section briefly describes additional methods utilised to process samples for 454 sequencing. Since approaches to analysis evolved over the course of time these are detailed in the relevant results sections.

4.2.1 Samples

On collection, all samples (intestinal contents or faecal matter) were stored at -20°C until further processing. Total bacterial community DNA was extracted from 200 mg of sample using the QIAamp DNA Stool Mini Kit (Qiagen, UK) and associated standard protocol. The initial lysis incubation step was performed at 95°C and a negative extraction control (tap water with no faecal material) was included for each subject set. Purity and quantity of DNA were assessed with the NanoDrop™ ND2000c spectrophotometer (Thermo Fisher Scientific, Wilmington, USA), 260/280 absorbance values between 1.75 and 2.10 being considered acceptable for downstream amplification.

4.2.2 Primers

PCR primers were selected to target the 16S rDNA of bacterial genomes. Choice of primers and modifications were intended to maximize coverage across phyla previously encountered in the intestinal milieu (RDP: http://pyro.cme.msu.edu) whilst minimising creation of extraneous products. Unless otherwise stated, universal 16S primers 926F (Table 2.1; Muyzer *et al*., 1995) and 1391R (Table 2.1; Lane *et al*., 1985), were utilised for amplification of hypervariable regions V6-V8. To allow for parallel analysis of multiple samples on a single region of a sequencing chip, ten-nucleotide sequences were included between the adapter sequence 'A' and the 16S-specific region of the primers; sequences utilised were those recommended as MIDs (Multiplex Identifiers) by the

manufacturer of the 454 Genome Sequencer FLX platform (Roche Diagnostics). All primers were supplied by Sigma-Aldrich (Dorset, UK).

4.2.3 Amplicon production

PCR reactions were prepared in 50 µl containing 1 x High-Fidelity PCR Buffer, 250 µM of each deoxynucleotide triphosphate (Promega, WI, USA), 0.8 M Betaine HCl (Sigma-Aldrich, Dorset, UK), 4 µg BSA (NEB ,UK), 2.3 mM $MgCl_2$, 0.8 µM each primer and 1 U Phusion High-Fidelity DNA Polymerase (Finnzymes, Finland), optimal constituent ratios being ascertained through gradient PCR. Extracted DNA samples were diluted with molecular biology grade H_2O prior to addition to the appropriate PCR mix. PCR was performed in Eppendorf Mastercyclers with the following conditions: 98°C for 5 min followed by 30 cycles of 98°C for 40 s, 58°C for 30 s and 72°C for 20 s, with a final extension at 72°C for 4 min. Optimal conditions and constituents had been established via gradient PCR, the temperature during the primer annealing phase of the PCR cycle being somewhat higher than expected for the T_m of the primers due to the nature of the proprietary buffer; the total number of cycles is higher than that utilised for standard community PCR since yield was found to be low, possibly due to an inhibitory effect of the fusion sequences and MIDs incorporated into primers.

Initial visualisation and estimation of purity was via electrophoresis in 1.5% agarose/TAE gels containing ethidium bromide. Negative controls were performed with the amplification of each set of samples.

4.2.4 Purification and quantification

Reactions were cleaned of PCR constituents and primer dimers using the Agencourt® AMPure® XP magnetic bead purification system (Beckman Coulter, USA) according to manufacturer's instructions. Subsequent quantification was via the Quant-iT™ PicoGreen ® (Molecular Probes Inc., Invitrogen, USA) assay technique as per the

manufacturer's instructions. Fluorescence of samples was assessed in duplicate, with excitation at 480 nm and emission detection at 520 nm, using a FluoStar Omega Spectrophotometer (BMG Labtech, UK), concentration of dsDNA in samples being calculated from the standard curve. Secondary verification of DNA concentration ratios between samples intended for pooling was performed using the NanoDrop™ ND2000c spectrophotometer. Each sample was then diluted with 1 x TE buffer such that the concentrations were standardised at 10^9 molecules/µl. Each sample then contributed 5 µl to a pool, the purity of which was determined using the Agilent 2100 Bioanalyzer (Agilent Technologies, UK) high-sensitivity dsDNA kit. Pools were dispatched for sequencing utilising the genome Sequencer FLX Instrument (454 Life Sciences, Roche Diagnostics, UK). In the latter stages with the use of Titanium kit chemistry expected output under optimal conditions was approximately 600,000 reads (in excess of 400 bp) per chip.

4.2.5 Data analysis

RDP: The raw output of a 454 run is in the form of either fasta (.fna) and quality (.qual) files or proprietary standard flowgram format (.sff) files. For early runs fasta and quality files were uploaded to the RDP pyrosequencing pipeline (Cole *et al*., 2008; http://rdp.cme.msu.edu/) for initial deconvolution and quality screening. At this stage the reads are allocated to their respective samples based on their barcodes, and sequences falling short of specified thresholds are rejected, i.e.sequences less than 100 bp in length and those containing ambiguous bases or base-calls with quality scores below 20. Sequences from the individual samples were then aligned using the fast Infernal aligner (Nawrocki and Eddy, 2007) and clustered into OTUs using a complete-linkage clustering algorithm, akin to the furthest-neighbour method whereby no member of an OTU is more than the cutoff distance away from every other constituent. The

pipeline then allows for creation of distance matrices and associated 'groups' files which can be utilised as input files for Mothur.

The RDP was also employed as the primary means of assigning taxonomic classification to the individual reads. Preliminary comparisons between the RDP and other sites such as Greengenes (DeSantis *et al*., 2006a; DeSantis *et al*., 2006b; NAST aligner; http://greengenes.lbl.gov/cgi-bin/nph-index.cgi) and ARB-Silva (Pruesse *et al*., 2007; SINA aligner; http://www.arb-silva.de) displayed greater than 95% correlation in classification assignment at the genus level, but the RDP provides for a greater capacity in terms of data upload and is therefore better suited to analysis of pyrosequencing output. The classifier implements a naïve Bayesian approach based on a word-size of 8 and returns the highest-scoring taxonomy along with a bootstrapped confidence value (Wang *et al*., 2007). There is also facility for comparison of libraries based on a standard T-test of each of the taxa (Wang *et al*., 2007), or the Library Compare Tool (adapted from an algorithm for comparison of expression levels in 'Northern' blots) if membership of the taxa is small (Audic and Claverie, 1997).

MOTHUR: Mothur (http://www.mothur.org/) is an open-source platform for analysis of microbial communities (Schloss *et al*., 2009). This software brings a variety of tools such as SONS (Schloss and Handelsmann, 2006a), DOTUR (Schloss and Handelsmann, 2005), Tree-Climber (Schloss and Handelsmann, 2006b), Libshuff (Schloss *et al*., 2004), Metastats (White *et al*., 2009), and UniFrac (Lozupone and Knight, 2005) together under one umbrella, with added functionality for processing of pyrosequencing data. Mothur will accept distance matrices from the RDP as input (as utilised in Run1), but can also perform similar deconvolution functions to the RDP and thus process the original fasta and quality files. Files displaying the commands utilised are included in Appendices 11 and 14.

The Mothur approach can be partitioned into 3 phases. The first phase is sequence processing, including: deconvolution, filtering of poor reads and denoising, de-replication, alignment (integral NAST-like aligner), chimera removal, distance matrix creation and production of files for the next phase. The second phase is initiated with clustering by the integral furthest neighbour method, before representatives of each OTU are chosen for assignment of a taxonomic classification to an OTU if desired. Mothur subsequently allows for a number of alpha-diversity (within sample diversity) and beta-diversity (between sample and group diversity) measures to be calculated, to describe and contrast community structure (OTUs present) and community membership (relative abundance of members of OTUs). These can then be represented in numerous formats such as dendrograms, tables, collectors and rarefaction curves, the latter being a process whereby comparisons can be made between samples at equivalent numbers of sequence reads (Brewer and Williamson, 1994). The final phase is implementation of statistical methods to test whether communities differ including: parsimony, Unifrac, libshuff, amova and homova. In addition, the means to display clustering of the samples in the form of PCA, PcoA and NMDS is provided.

QIIME: Towards the end of 2010 a further platform for analysis of complex microbial communities became available in the form of QIIME (Quantitative Insights Into Microbial Ecology; http://qiime.org/; Caporaso *et al*., 2010). QIIME incorporates many of the utilities and tools found in Mothur, but with superior built-in graphical capabilities, and the potential for faster processing of computationally-demanding steps, such as denoising and chimera-checking, via a cloud-based version.

Qiime, like Mothur, is able to accept the raw standard flowgram format files from the 454 platform. Processing begins with an integrated denoising, deconvolution and quality-screening phase, resulting in acceptable, trimmed sequence reads being partitioned into their respective samples. Denoising is achieved through implementation

of an integrated algorithm, with a stringency such that many (but not all) unique sequences are identified as erroneous artefacts of the pyrosequencing procedure (Reeder and Knight, 2010). At this point QIIME also accommodates the fusion of outputs from various regions of a chip, although only if the 16S gene amplicons can be aligned. Subsequent to initial processing the reads are binned at a 97% (default; approximately species level) sequence similarity using uclust (Edgar, 2010); this can be performed against a reference set to determine the 'seeds' of the clusters (as oppposed to *de novo* clustering) thereby standardising the clustering process across multiple samples and studies. Sequences are then aligned (PyNAST; Caporaso *et al*., 2010a) against the Greengenes core reference alignment (DeSantis *et al*., 2006b) prior to chimera-checking and removal (Chimeraslayer; Haas *et al*., 2011), phylogenetic tree development (FastTree 2; Price *et al*., 2010) and OTU table creation, the latter representing the input for downstream applications. Heatmaps, network views and abundance plots can be visualised, before estimation of multiple alpha and beta diversity calculators, the majority of which are shared by Mothur. These calculators can themselves form the basis of 2-D and 3-D PcoA plots, while UPGMA trees derived from distance metrics can be subjected to jack-knifing and bootstrapping to estimate support for clustering of samples across experimental groups. In addition, ANOVA, Pearson correlation, and a paired T-test are available to identify OTUs with varying membership across experimental treatments.

R: Certain evaluations were also carried out using 'R', a statistical computing language. The benefit of the use of 'R' lay in the flexibility of custom scripts, but the evolution of Mothur and emergence of Qiime led to the discontinuation of this approach, apart from custom visualisation of OTU heatmaps and PCoA plots. Other software and resources are described where utilised.

4.3 Run 1: CDAD and AAD

The following section details experiments undertaken to provide preliminary data on the intestinal microbiota of patients with CDAD as compared to that of subjects with AAD, based on analysis of bacterial DNA extracted from stool samples and sequenced using the 454 platform. In addition to examining taxonomic differences between the subject groups, subsidiary aims were to investigate experimentally-induced differences in composition engendered by choice of primer pairs, and the development of the bioinformatics pipeline.

4.3.1 Samples and extraction

Human faecal samples were donated by Dr Martha Clokie (University of Leicester). Anonymised samples had been acquired from the Leicester Royal Infirmary as part of ongoing investigations into *C. difficile* and its bacteriophages; samples were those submitted to clinical microbiology for diagnostic ELISA testing for CD toxin A, and had been stored at 4°C. A number of samples had been found to be negative for *C. difficile*. Two CD-positive (CD7 and CD8) and 2 CD-negative samples (CD18 and CD19) were selected for investigation. 200 mg of each faecal sample was subjected to extraction of total community DNA as per section 4.2.1. Extracts were assessed for purity and concentration using an Eppendorf BioPhotometer, providing values of 1.85 ± 0.14 and 40.8 ± 20.2 ng/µl respectively. The latter is lower than expected for extraction with the Qiagen protocol but not outside the acceptable range, and test PCRs with generic 'universal' community 16S primers (Bact-8F and Bact-1541RV; table 2.1) provided a band of the expected size.

4.3.2 Amplicons

Primer pairs for this phase of investigations were 454-R1-8F with 454-R1-357R and 454-R1-784F with 454-R1-1061R (Table 2.1). Primers were designed to incorporate the

454-adapter sequences (Roche Diagnostics Ltd, UK) at the 5′ end. Reverse (R) primers were synthesisized with fusion sequence 'A' (GCCTCCCTCGCGCCATCAG) while forward (F) primers incorporated the 'B' sequence (GCCTTGCCAGCCCGCTCAG). Pentamer oligonucleotides were incorporated into R primers between the fusion and 16S-specific regions such that sequencing from 'A' towards 'B' (mediating attachment to beads) would provide a sample-specific barcode (Andersson *et al*., 2008) for each sequence read allowing for subsequent deconvolution of pooled amplicon data.

The rationale for selection of primer pairs at this stage was governed by the available literature and system-specific considerations. Returning to figure 1.2 it is clear that maximal classification accuracy is achieved via investigation of multiple hypervariable regions (Wang *et al*., 2007), ideally in a single amplicon (i.e. bi-directional Sanger sequencing of a 1.5 kb cloned insert), but the 454 system was limited to amplicon lengths of 400 bp (GS FLX Amplicon DNA Library Preparation Method Manual 2008, Roche). A recent investigation (Andersson *et al*., 2008) had successfully utilised 784F/1061R (table 2.1) and coverage of this region (V5/V6) was also suggested by the RDP, so this amplicon was felt to be sufficiently supported. Early empirical evidence had also evinced the utility of primers annealing at positions towards the start of the 16S gene (Baker *et al*., 2003). These allowed amplicon coverage of V1, V2 or V3, thus offering the potential for high classification accuracy and confidence of taxonomic assignment (Wang *et al*., 2007). The constraints mentioned above and limited availability of conserved regions for primer annealing within this region (Baker *et al*., 2003) led to the choice of the 8F/357R pair and amplicon production across V1/V2.

Subsequent to optimisation of conditions with the chosen primer pairs, PCR reactions were performed as per section 4.2.3 with visualisation of products on 1.5% agarose gels as shown in Figure 4.3. Presence of a smaller band could not be entirely eradicated

through alteration of cycling conditions (reduced [Mg^{2+}] and total cycle number, and increased annealing temperature) so bands were excised as per section 2.7.1.

Figure 4.4 PCR amplicons for 454 Run1

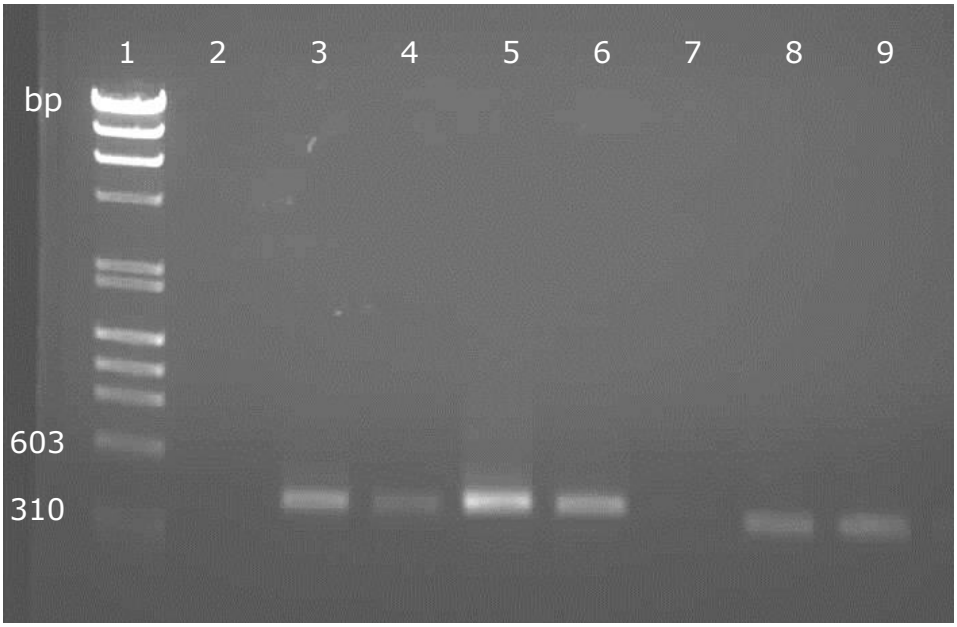


Figure 4.4 shows 1.5% agarose gel visualisation of PCR products for 454 subsequent to gel extraction. Lane 1 shows λ/φ marker. Lanes 2 and 7 show negative controls. Lanes 3-6 show samples CD7, CD8, CD18 and CD19 with primer pair 8F and 357R. Lanes 8 and 9 show PCR products for CD7 and CD8 with 784F and 1061R. Amplicon sizes are 380 bp and 310 bp respectively.

4.3.3 Purification, quantification and pooling

Gel extraction to remove artefacts at approximately 270 bp obviated the need to perform the AMPure clean-up procedure to remove primer dimers which were evident in all 454 preparation PCRs, possibly due to the high concentration of primers required to achieve adequate yield. However, manufacturer's recommendations (GS FLX Amplicon DNA Library Preparation Method Manual 2008, Roche) stress the importance of the procedure, so all samples were subjected to AMPure purification as per 4.2.4, followed by quantification using the PicoGreen system detailed in the same

section. Concentrations of samples were then standardised and diluted to the required level for emPCR procedures (1 x 10^7 molecules/µl) before pooling and submission.

4.3.4 Sample names

The naming convention adopted for the samples utilised in Run 1 is detailed in Table 4.1. The faecal sample extracts were amplified with both of the primer pairs to create 2 separate amplicon libraries for each with coverage of hypervariable regions as previously detailed.

Table 4.1 Run 1 samples

Original name	Primer pair 8F/357R	Primer pair 784F/1061R
CD 7 (CD +ve)	Sample7 (S7)	Sample11 (S11)
CD 8 (CD +ve)	Sample8 (S8)	Sample12 (S12)
CD 18 (CD -ve)	Sample9 (S9)	Sample13 (S13)
CD 19 (CD -ve)	Sample10 (S10)	Sample14 (S14)
MK23B (respiratory)	Sample6 (S6)	N/A

Sample6 represents bacterial DNA extracted from a COPD patient sputum sample and amplified as per the faecal samples with primer pair 1. Six respiratory samples derived from a parallel study were included in the run and S6 was chosen to represent an 'outgroup' for ecological analysis.

4.3.5 Pyrosequencing output and data-processing

Samples S6 (respiratory) and S7-S14 were sequenced on ¼ of a PTP device using the Roche 454 GS-FLX sequencer and GS-LR70 sequencing kit. Expected output was in

the region of 70,000 reads of 250 bp length for a total of 17.5 Mbp; raw read number for the entire region was 77,452. Fasta and quality files were uploaded to the RDP along with primer sequences and barcodes for deconvolution; initial processing removed reads of less than 100 bp in length and those sequences containing ambiguous bases, providing 54,312 reads across 14 samples, with a further 1,954 assigned to 'no-tag', indicating at least one nuceotide error in the barcode region. Average number of reads per sample was 3,880, (range 2960-6584) with average length of 224 bp.

Samples S6-S10 and S11-S14 were aligned as 2 groups with the bacterial model, the lack of overlap between sequenced regions preventing global alignment. The groups were then clustered at distances from zero (identical) to 0.3 (70% identity) in 0.03 distance increments. The OTU distances thus approximate to the various levels of distance conventionally associated with species, genus, family and phylum (0.03, 0.05, 0.10 and 0.20 respectively; Schloss and Handelsmann, 2004) although it must be remembered that these figures apply to distance comparisons over the full length (1.5 Kb) of the 16S gene (Schloss and Handelsmann, 2004). Grouped alignment files were also utilised to create uncorrected lower-triangular distance matrices for export to Mothur. In addition, deconvolution of samples provides individual fasta files which were utilised for classification and input for the RDP Library Compare facility.

4.3.6 Classification

Assignment of taxonomic classification to sequence reads was performed using the RDP classifier function (Wang *et al*., 2007) and 80% confidence threshold. The classifier provides a confidence level (based on percentage of bootstrap trials returning the output) for the most common match at each taxonomic level. The RDP recommends a 50% threshold for short reads but this allowed the inclusion of certain sequences which could not be assigned to the bacterial domain; these were found to correspond to archaea, host or intestinal material via implementation of global Blast searches

(Altschul *et al*., 1990) on the NCBI website (http://blast.ncbi.nlm.nih.gov/Blast.cgi; Benson *et al*., 2009). A threshold of 80% excludes these sequences while still allowing for tentative assignment at lower taxonomic levels.

Figure 4.4 displays the total composition of the samples by percentage at the level of phylum (A) and family (B) respectively. Those phyla or families which contributed at least 0.5% of the total reads in at least 3 of the samples are included in the legends for the figure; the remainder are grouped into the 'Other' category which also includes those bacterial sequences unclassified at the family level with 80% confidence.

Although the number of samples does not easily lend itself to statistical comparison of the 2 groups a number of observations are germane to both technical and biological aspects of the study.

At the phylum level the composition is broadly as expected for faecal samples, the phyla *Firmicutes*, *Bacteroidetes* and *Proteobacteria* dominating the 'fingerprint' with a lower incidence of *Actinobacteria* and *Fusobacterium* (Eckburg *et al*., 2004; Andersson *et al*., 2008); levels of *Proteobacteria* are comparatively high, most likely due to antibiotics administered to target the Gram-positives. While there appears to be little correlation between the samples within the CD-positive and CD-negative groups (S7 and S10 being the most similar), the repeats of the respective samples with a different primer pair show a reasonable degree of correlation on preliminary examination, all but one of the composition values being within 10% of its 'replicate'. These observations are also applicable to classifications at the family level, particularly for samples S7/S11 and S10/S14. It is noteworthy that *C. difficile* (*Peptostreptococcaceae*) was detected only in the CD group, albeit at low levels in S7/S11. This latter finding is promising from the perspective of detection of relatively rare groups, numbers of *C. difficile* present in the CDAD state estimated as 10^5-10^8 per gram of faeces (Bartlett *et al*., 1980) compared to 10^{12} total bacteria (Donskey, 2004).

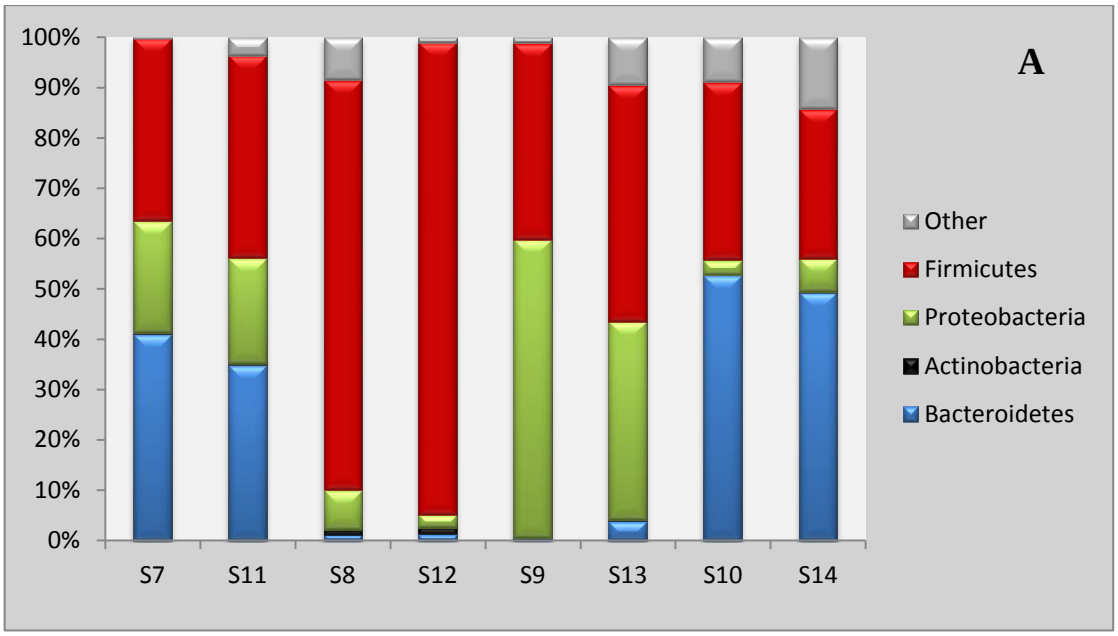


Figure 4.5 Percentage composition of samples S7-S14 at phylum (A) and family (B) level

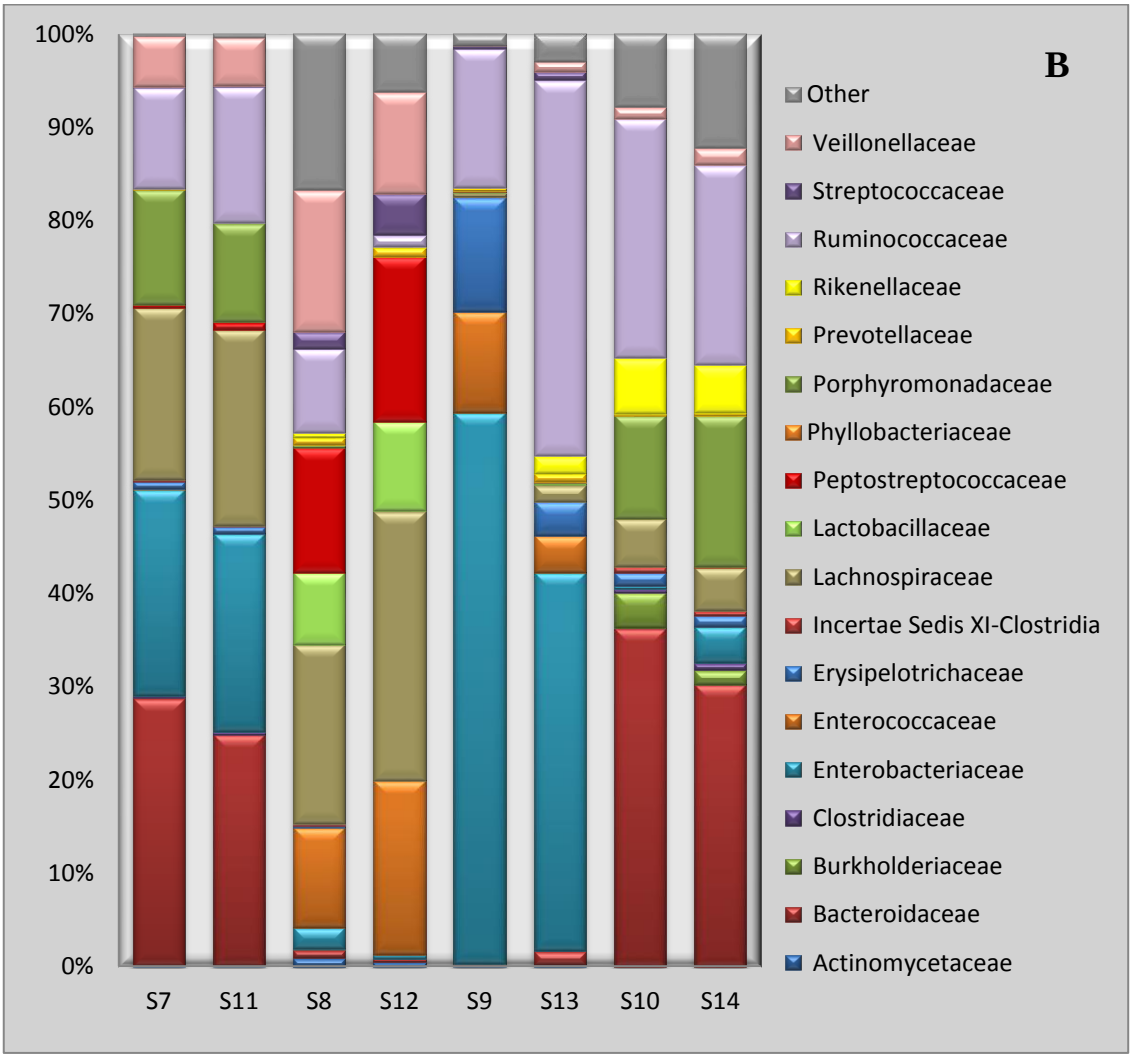


Figure 4.5 displays phylum (A) and family (B) level classification for samples S7-S14. Identical samples with the differing primer pairs are shown side by side. S7/S11 and S8/S12 are the CD +ve samples while S9/S13 and S10/S14 are CD -ve. Legends to the right display colour association for the predominant phyla and families.

From a biological perspective, the family *Ruminococcaceae* (previously clostridial cluster XIVa, *Clostridium coccoides* type) is more prevalent in the non-CDAD group while both the *Lachnospiraceae* (clostridial cluster IV, *Clostridium leptum* type) and the *Veillonellaceae* (clostridial cluster IX) are comparatively under-represented in this group.

4.3.7 RDP LibCompare and Spade

The RDP provides a means to test the similarity of classified communities on the basis of each taxon encountered in one or other of the libraries. The statistics utilised are described in Appendix 8. Each of the samples was compared in this fashion to all of the others, a total of 36 tests. Multiple test correction is unavailable for RDP output so tests are likely to indicate an artificially high level of difference between samples, but the number of p-values less than 0.05 was calculated as a percentage of total tests in each pair-wise comparison. The average of percentage significant differences between primer repeat samples (total = 4) was then compared to the average between all other samples (total = 32); 35.88% compared to 54.46%, while the range of the latter was 43.2% to 70.77%. Although this is an unsatisfactory approach to determining the absolute bias introduced by primer pair choice, not least since it is unclear if the composition revealed by either pair is truly representative of the community, it does provide a rudimentary measure of reproducibilty. While all values for 'sister' (identical source but differing primer pair) samples lie outside the range for unrelated comparisons it is clear that primer selection has at least some impact upon apparent microbiotic composition.

The RDP also provides an option to format output for use of Spade (Species Prediction And Diversity Estimation; Chao and Shen, 2010) which, like Mothur, is capable of calculating a variety of diversity indices and similarity coefficients, but is more suited to species abundance data than distance matrices. Table 4.2 displays the Morisita-Horn similarity coefficient as calculated with Spade (Appendix 7) for each of the pairwise library comparisons. S6, the respiratory sample, is included as an 'outgroup'. Comparisons between 'sister' samples are displayed in red font.

Table 4.2 Morisita-Horn similarity values for family level comparisons in SPADE

	S6	**S7**	**S11**	**S8**	**S12**	**S9**	**S13**	**S10**	**S14**
S6	1	0.018	0.017	0.078	0.026	0.016	0.015	0.005	0.007
S7		1	**0.992**	0.415	0.348	0.503	0.553	0.773	0.803
S11			1	0.462	0.372	0.54	0.613	0.739	0.789
S8				1	**0.923**	0.152	0.198	0.182	0.208
S12					1	0.084	0.068	0.105	0.085
S9						1	**0.948**	0.093	0.204
S13							1	0.211	0.348
S10								1	**0.965**
S14									1

Four measures of similarity were calculated in Spade: the Jaccard (Jaccard, 1901), Sørensen (Sørensen, 1948), Bray-Curtis (Bray and Curtis, 1957), and Morisita-Horn (Horn, 1966) similarity coefficients. These represent a subset of those available (Wolda, 1981) which examine both community membership (OTUs present or absent; Jaccard and Sorensen) and structure (abundances within OTUs; Bray-Curtis and Morisita-Horn). The Morisita-Horn coefficient has been chosen as representative from those

calculated as this estimator is the least susceptible to differences in reads per sample while still accounting for relative abundance (Chao *et al*., 2006). All calculators displayed the same trend in that values for replicate samples were close to maximal, while the lowest values were recorded for comparisons of the intestinal extracts against the respiratory sample. In addition, an indirect estimation of the bias introduced by choice of primer pair is the correlation between values of a replicate pair with any other sample e.g. S7 and S11 v S6. These values can be seen to diverge, particularly for S8 and S12, but indicate the same relative similarity: e.g CD7 is most similar to CD19 independent of primer pair utilised for analysis, even though the former is CD +ve while the latter is CD -ve

4.3.8 Mothur

Where less than 50,000 sequences are to be analysed the RDP can be utilised to create distance matrices in either column or lower-triangular (Phylip) format for export to software such as Mothur. While it is now possible to utilise Mothur for preliminary processing of sequences this option was not available in the initial stages, so in this instance use of Mothur commenced at the second phase (as per section 4.2.5) with clustering of sequences via the furthest neighbour method and a cutoff of 0.10. All commands can be found in Appendix 4 but for this run a truncated analysis was undertaken to provide alpha and beta diversity indices/coefficients for each of the samples S6-S14.

4.3.8.1 Alpha diversity

Alpha diversity indices are quantitative descriptors of communities based on the number of OTUs encountered and the relative distribution of individuals within these taxonomic groups (Whittaker, 1972). While the diversity per se is not of fundamental importance to the current research these were viewed as a means of indirectly

comparing the sister samples, the assumption being that values for diversity should be independent of primer pair utilised and thus equivalent. However, correlation between diversity and pathology would be of some interest in that reduced diversity is suggested to correlate with reduced 'health' of an ecosystem (Chapin *et al*., 1998) and increased potential for infection (Keesing *et al*., 2006), while reduced diversity of the human intestinal microbiome has been implicated in recurrence of CDAD (Chang *et al*., 2008) Although Mothur provides for calculation of a plethora of indices, for current purposes only 4 were utilised: Chao1 (Chao, 1984), Berger-Parker Dominance (Berger and Parker, 1970), Simpson (Simpson, 1949), and Shannon-Wiener (Shannon, 1948; Wiener, 1949), the formulae for which can be found in Appendix 8. In addition, Good's coverage (Good, 1953), a measure of the adequacy of sampling, was also calculated. These indices were thought to provide sufficient overall measures of community 'diversity' despite caveats for individual application (Magurran, 2004).

Table 4.3 Observed OTU's, diversity indices and coverage for S7-S14 at 0.03 cutoff

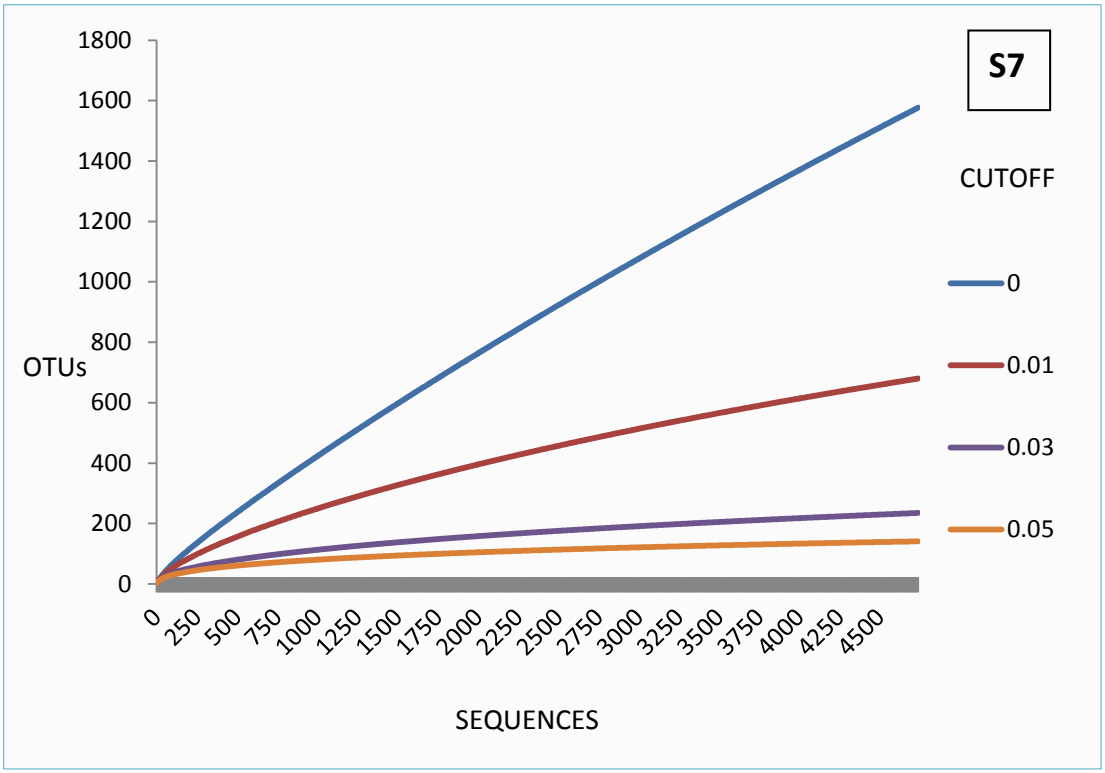
	Observed	Chao1	LCI*	UCI*	Shannon	Simpson	BPD	Good's
S7	587	1005.1	971.3	1064.4	4.629	0.025	0.091	72.600
S11	501	970.4	936.6	1010.7	4.670	0.025	0.090	67.200
S8	1233	2838.2	2669.6	3327.4	5.485	0.016	0.097	66.800
S12	486	774.6	734.1	828.9	4.141	0.053	0.145	78.100
S9	230	369.2	345.5	401.3	3.838	0.044	0.254	79.700
S13	455	737.4	697.2	796.1	3.928	0.091	0.243	75.600
S10	466	659.0	635.0	691.6	3.818	0.075	0.118	91.200
S14	753	1687.3	1629.7	1743.6	5.010	0.036	0.138	59.300

Table 4.3 displays the values calculated for all samples with a cluster cutoff of 0.03, equivalent to the members of each cluster showing 97% identity. Lower (LCI*) and upper (UCI*) 95% confidence intervals for the Chao1 index are shown indicating the similarity of S7 and S11; confidence intervals for the Shannon index are not shown but indicate the similarity of both the S7/S11 and S9/S13 pairs. This approach to compare 'sister' samples thus indicates that use of differing primer pairs may introduce bias, but the small number of samples and lack of technical replicates prevent any definitive conclusions. In addition, the indices may be inherently unsuited to such utilisation: diversity values can vary with relatively minor changes in community composition while conversely being identical for communities with significant structural dissimilarity. Such considerations notwithstanding, the diversity of the two groups was not found to differ for the presence or absence of *C. difficile* infection, p-values for Chao1, Shannon and Simpson being 0.232, 0.175 and 0.098 respectively. Values of the indices also indicate that the communities sampled are diverse (Shannon approaching or exceeding 4, and Simpson approaching zero) and evenly partitioned (BPD approaching zero).

Good's coverage values are expressed as a percentage and, although only an estimate, indicate that a high proportion of the OTUs have been encountered. Such a conclusion could also be inferred from rarefaction curves plotted on the basis of Mothur output, rarefaction being calculated by performing repeated sampling of the total pool at various levels from zero to the maximum and taking the average number of OTUs encountered at each sampling depth to produce a smooth curve. In addition to allowing for contrast of communities at equivalent levels of sampling, if the curve is seen to become asymptotic sampling can be considered to have been sufficiently comprehensive to encounter the majority of species in an environment.

.

Figure 4.6 Rarefaction curves for OTUs observed in S7 and S11



Figures show rarefaction curves for observed OTUs for S7 (above) and S11 (below) at various cutoffs from identical to 0.05, approximately corresponding to species (0.03) and genus level (0.05) for the length of amplicon examined. Curves can be seen to be approaching asymptote for species level at 2500 sequence reads.

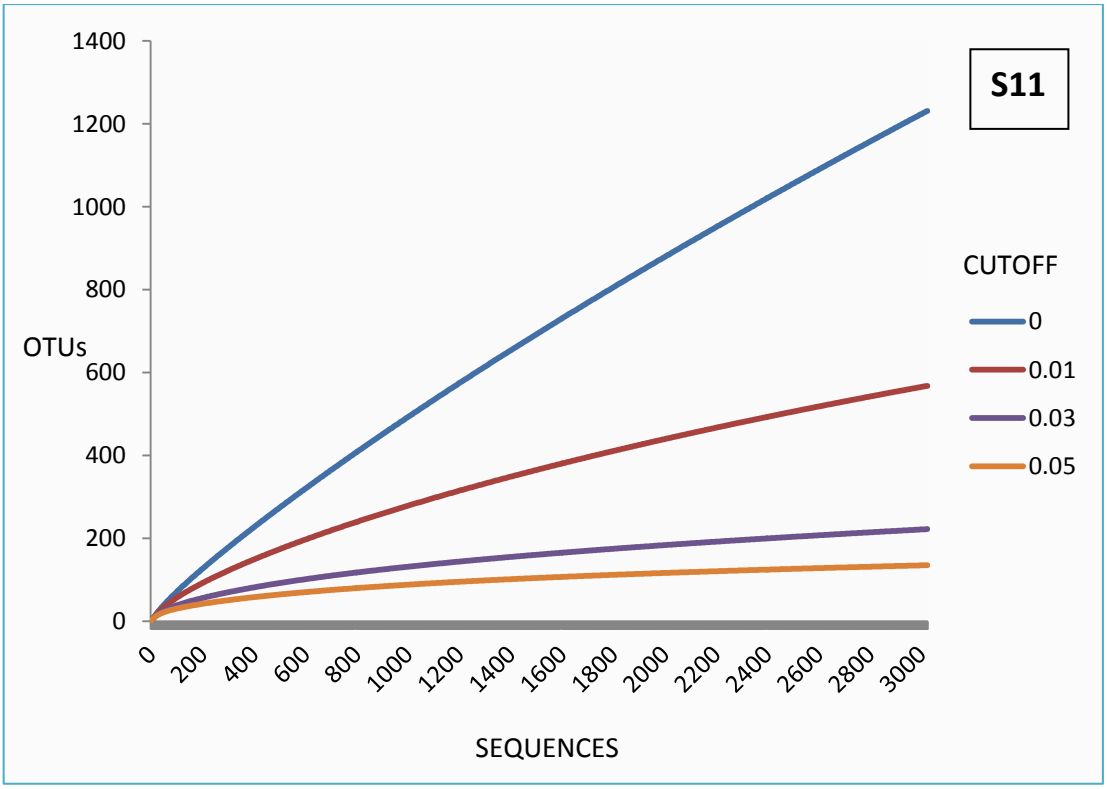


Figure 4.6 shows rarefaction curves for S7 and S11, although those derived from all samples were similar; these indicate that a minimum of 2500 sequences per sample are desirable.

As a whole, investigation of alpha diversity proved unsatisfactory for comparison of samples at this stage due to the small number of samples, an experimental design flaw which could be rectified in future phases. Alpha diversity indices also tend to ignore abundances and are therefore best employed as a preliminary measure to assess communities. This is particularly applicable with regard to rarefaction, which is an essential aspect of processing to effectively 'normalise' sequence reads per sample

4.3.8.2 Beta Diversity

In addition to alpha diversity to assess the richness and eveness of individual samples, beta diversity values can be calculated in an attempt to quantify the similarity (or otherwise) of samples. Broadly, beta diversity coefficients can be compartmentalised into 2 groups: those based on community membership or presence/absence of OTUs; and those which factor in abundance within OTUs and therefore provide a measure of community structure as a whole. Once again, Mothur provides for derivation of a mulititude of these values including the following: the Sørensen, Jaccard and Kulczynski-Cody similarity coefficients (Chao et al. 2006), all 3 of which are incidence-based; and the abundance-adjusted Jaccard, abundance-adjusted Sørensen, Morisita-Horn and Bray-Curtis similarity coefficients (Chao et al. 2006), the Smith theta similarity coefficient (Smith *et al*., 1996), the Yue and Clayton theta similarity coefficient (Yue and Clayton, 2005) and the Euclidean measure of similarity (Faith *et al*., 1987) all of which examine communities from the structural or abundance perspective. Formulae for all the above are included in Appendix 9, while results are displayed in Table 4.4.

The values indicate that the CD and AAD groups are not differentiated by the beta-coefficients, the most similar samples being the S7/S9 pair or the corresponding sister pair S11/S13. It is clear though, that both the choice of co-efficient and the level of comparison can influence the conclusion; values for the adjusted Sorensen and adjusted Jaccard indicate that S8/S9 or S12/S13 are the most similar, while the Morisita-Horn coefficients for comparison at class level (Table 4.2) indicate the highest level of similarity is that between S7/S10 or S11/S14. The values for S7 against S6 are the comparison of an intestinal community with a respiratory, included to provide a scalar reference point for the superficially arbitrary coefficients.

Table 4.4a Beta-diversity value Mothur output for Run 1 (S6-S10)

comparison	1	2	3	4	5	6	7	8	9	10
S7 v S8	48	0.027	0.052	0.060	0.212	0.3498	0.032	0.029	0.016	0.145
S7 v S9	**127**	**0.137**	**0.241**	**0.244**	0.214	0.3527	**0.453**	**0.201**	**0.293**	**0.201**
S7 v S10	36	0.046	0.088	0.108	0.197	0.3301	0.033	0.042	0.016	0.140
S8 v S9	59	0.035	0.069	0.087	**0.351**	**0.5200**	0.024	0.025	0.012	0.192
S8 v S10	21	0.014	0.028	0.054	0.113	0.2039	0.004	0.012	0.002	0.063
S9 v S10	12	0.017	0.034	0.038	0.133	0.2360	0.013	0.007	0.006	0.103
S6 v S7	**9**	**0.007**	**0.015**	**0.015**	**0.008**	**0.0163**	**0.0004**	**0.002**	**0.0002**	**0.006**

Key to the respective coefficients in column 1 of each tables 4.4a and b: 1 = Observed shared OTUs; 2 = Sorensen; 3 = Jaccard; 4 = Kulczynski-Cody; 5 = abundance-adjusted Jaccard; 6 = abundance-adjusted Sorensen; 7 = Morisita-Horn; 8 = Bray-Curtis; 9 = Yue and Clayton; 10 = Smith theta. Highest values for each index are highlighted in blue.

Table 4.4b Beta-diversity value Mothur output for Run 1 (S11-S14)

comparison	1	2	3	4	5	6	7	8	9	10
S11 v S12	33	0.066	0.034	0.066	0.193	0.324	0.018	0.026	0.009	0.143
S11 v S13	**82**	**0.165**	**0.090**	**0.165**	0.177	0.301	**0.369**	**0.174**	**0.226**	0.167
S11 v S14	79	0.125	0.067	0.131	0.260	0.413	0.121	0.078	0.064	0.184
S12 v S13	68	0.144	0.077	0.144	**0.414**	**0.586**	0.007	0.040	0.003	**0.237**
S12 v S14	25	0.040	0.020	0.042	0.312	0.476	0.005	0.011	0.002	0.129
S13 v S14	36	0.059	0.030	0.063	0.169	0.290	0.046	0.028	0.023	0.140

4.3.9 Run 1 Summary

The first attempt at investigation of human intestinal communities thus achieved certain objectives whilst illuminating areas requiring further refinement.

Broadly, the bacterial composition was found to correlate subjectively with that established by previous investigations (Eckburg *et al*., 2005, ; Andersson *et al*., 2008), the relative prevalence of the phylum *Proteobacteria* and scarcity of *Firmicutes* attributable to the disease state and administration of antibiotics (Young and Schmidt, 2004; Dethlefsen *et al*., 2008; Croswell *et al*., 2009). Of particular interest from a biological perspective was the relatively low incidence of the family *Veillonellaceae* in the non-CDAD samples with either primer pair. The veillonella are notable for their fermentation of lactate to acetate and propionate (Duncan *et al*., 2004); if such a differential were to be confirmed by further investigation a simple protective mechanism would potentially present itself through prevention of *C. difficile* overgrowth by lactate levels, although any causal relationship between bacterial groups

and mechanism of colonisation resistance is likely to be multi-factorial and metabolically complex.

Results also indicate that bias introduced by choice of primer pair may not be as significant as expected (Schmalenberger *et al*., 2001), both taxonomic proportions and ecological indices suggesting subjective correlation of 'sister' samples. However, the small sample size restricted the statistical tests applicable such that conclusions were relatively superficial observations; an increase in the number of replicates would be utilised to rectify this where significant numbers of individual samples could not be procured.

In addition, this preliminary investigation provided an introduction to the resources available for analysis of community data, both with regard to the bioinformatics applications (RDP, Mothur, Spade) and the ecological indices (alpha and beta diversity metrics) they employ. In particular, the indices are easily applied erroneously (Hill *et al*., 2006), while even the terminology is an area of continuing debate (Jurasinski and Koch, 2011). Awareness of some of the limitations, though, does permit the use of both alpha and beta indices/coefficients as a starting point for analysis, while rarefaction of data from individual samples to allow for comparison at equivalent sampling levels is essential.

4.4 Run 2: Effects of diet and *C. jejuni* infection on the chicken caecal microbiota

As part of ongoing research into *C. jejuni* infection in chickens a sample set of caecal contents was made available. In addition to those birds innoculated with *C. jejuni*, intestinal contents from birds of a similar age and raised under equivalent conditions (without innoculation) were obtained, while a further subset were raised on a contrasting diet. While the host and bacterial species differ from the primary aims of this research these provided the opportunity to assess the impact on a microbiota of introduction of a new species and changes in nutrient intake.

Additional aims of this phase were to further investigate primer bias and develop the analysis pipeline, specifically with regard to the introduction of 'R', a powerful computer language for statistical and graphical applications. Investigations were conducted subsequent to introduction of the Titanium chemistry for the 454 platform, allowing for read lengths in the region of 400 bp; it was thus also utilised as an opportunity to ascertain how processing would differ from that for the FLX standard.

4.4.1 Samples

Samples were the caecal contents of chickens culled at between 3 and 6 weeks of age, with immediate removal of the caecum and extrusion of the contents prior to storage in airtight containers at -20°C. Samples were obtained from the School of Clinical Veterinary Science at the University of Bristol, transported on dry ice and stored at -20°C until extraction. The 12 samples were in four triplets: 3 innoculated with *Campylobacter jejuni* and culled 3 days later; 3 innoculated with *Campylobacter jejuni* and culled 3 weeks later having been fed on a standard diet in the intervening time period; 3 additional chickens were raised on a standard diet and a further 3 separately on an organic diet. All chickens were culled at 6 weeks.

Table 4.5 Annotation of Samples for Run2

Sample set	Primers	Annotation
C. jejuni innoculated; early cull	357F/939Rs	Cj1, Cj2, Cj3
C.jejuni innoculated; late cull	357F/939Rs	Cj8, Cj9, Cj10
Standard diet	357F/939Rs	S4, S5, S7,
Organic (Hubbard) diet	357F/939Rs	H4, H5, H7
S5 repeat	357F/939Rs	S5rep1, S5rep2, S5rep3
S5 repeat	8F/784R	S5rep4, S5rep5, S5rep6
S5 repeat	515F/1115R	S5rep7, S5rep8, S5rep9

4.4.2 Preparation

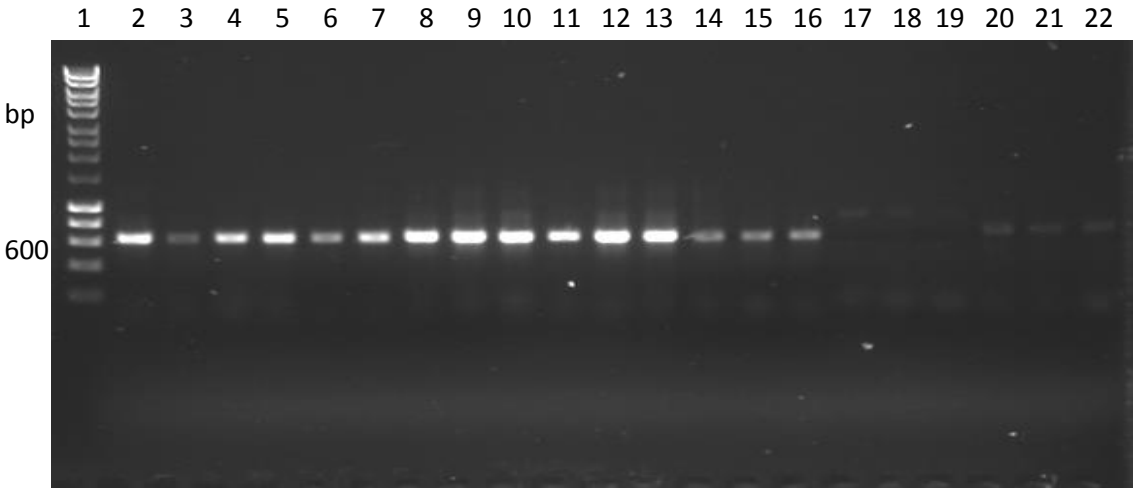
Community bacterial DNA was extracted from samples as described previously (Section 4.2.1), values for purity and yield being within the recommended range for downstream amplification. Standardised concentrations of sample DNA were then amplified with one of 3 pairs of primers, synthesized with either fusion sequence B (CTATGCGCCTTGCCAGCCCGCTCAG) or fusion sequence A (CGTATCGCCTCCCTCGCGCCATCAG) and a MID at the 5' end. Rationale for primer pair selection was again dictated by manufacturers recommendations for the 454 platform (Amplicon Library Preparation Method Manual for GS FLX Titanium Series 2009, Roche; 454 Sequencing System Guidelines for Amplicon Experimental Design

2011, Roche) and available conserved regions for primer annealing within the 16S gene. Successful amplification in the previous run informed the use of 3 of the 4 primers, albeit in different combinations, with others selected from a variety of sources in the literature (see Table 2.1) and adapted where necessary such that the T_m of each member of a primer pair was matched to within 5°C. Those selected were 357F with 939Rs (pair A), 8F with 784R (pair B), and 1115R with 515F (pair C), the second member of each pair being complexed with fusion sequence A and thus the MID. Primer pairs covered hypervariable regions V5/V4, V4/V3 and V4/V5/V6 respectively, the overlap intended to reduce classification bias based on region but unfortunately not extensive enough to allow alignment of reads from all samples. Sample annotation is displayed in Table 4.5.

Constituents of PCR reaction mixtures were as per section 4.2.3 and PCR was performed in Eppendorf Mastercyclers with the following conditions: 98°C for 5 min followed by 30 cycles of 98°C for 30 s, annelaing at 60°C (for pair A) or 58°C (for pair B) or 62°C (for pair C) for 40 s and 72°C for 25 s, with a final extension at 72°C for 4 min. Amplicons were subsequently visualised via agarose gel electrophoresis as detailed previously. Results are displayed in figures 4.5 and 4.6.

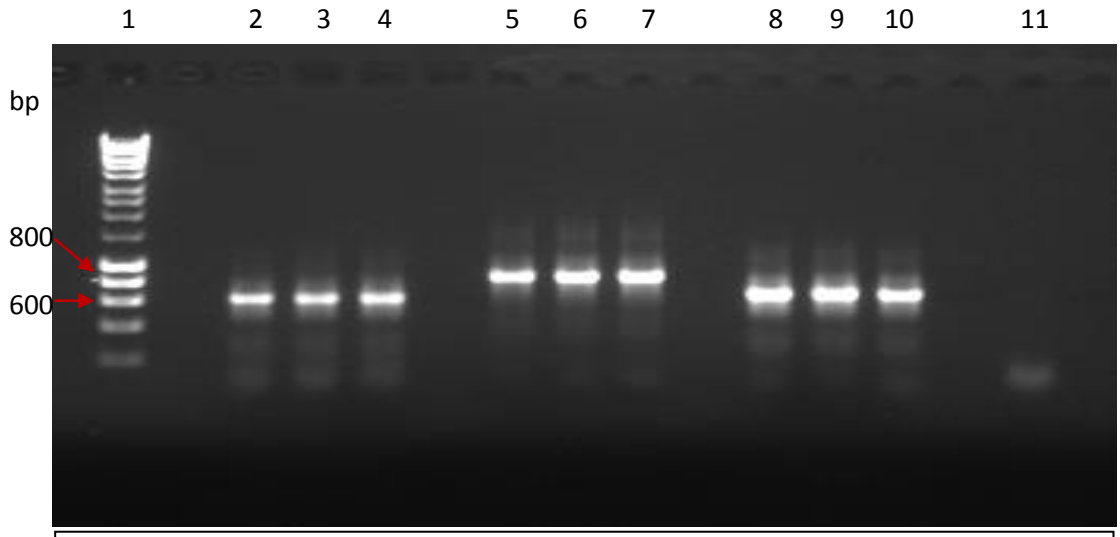
All samples were then subjected to AMPure purification as per 4.2.4, followed by quantification using the PicoGreen system detailed in the same section. Concentrations of samples were then standardised at equimolar levels (1×10^9 molecules/µl) prior to further dilution, pooling and submission. EmPCR and sequencing were performed by Reshma Bharkada at the Genomic Services Facility, University of Leicester.

FIGURE 4.7 AGE visualisation of amplicons from Cj, S, H and S5rep series



2 µl of PCR product with 2 µl OG (5x) and 6 µl H_2O. Lane 1 = HyperladderI; Lanes 2-4 = S4-S7; Lanes 5-7 = H4-H7; Lanes 8-10 = Cj1-Cj3; Lanes 11-13 = Cj8-Cj10; Lanes 14-16 = S5rep1-S5rep3; Lanes 17-19 = S5rep4-S5rep6; Lanes 20-22 = S5rep7-S5rep9. Negative controls were included but not shown and S5rep1-S5rep9 are repeated below due to low yield.

Figure 4.8 AGE visualisation of amplicons for S5rep series



Loading volume per well as for figure 4.5. Lane 1 = HyperladderI; Lanes 2-4 = S5rep1-S5rep3; Lanes 5-7 = S5rep4-S5rep6; Lanes 8-10 = S5rep7-S5rep9; Lane 11 = negative control. Temperature for annealment cycle of PCR had been reduced to improve yield, so low-intensity artefactual bands can be visualised. Negative control in lane 11 shows primer dimers.

4.4.3 Output and initial processing

Samples were sequenced as one pool on one quarter of a picotitre plate (PTP), expected to provide in the region of 150,000 reads of up to 400 bp in length. However, the initial upload to the RDP for deconvolution and classification consisted of only 26,650 reads, less than 20% of that predicted, with some samples providing as few as 200 reads (S5rep6). Investigations into the cause of the shortfall via run diagnostics provided no definitive answer, but problems with subsequent runs suggest that initial processing of the raw image intensity data was too stringent with the default settings. The introduction of the titanium chemistry was not accompanied by guidelines as to thresholds for the settings of the filtering algorithms, and the 'tweaks' necessary were only fully implemented some 18 months later as a result of empirical findings by the sequencing centres.

Despite the reduced read number pre-processing was performed via the RDP site and provided a total of 22,744 reads after filtering for length (bp > 150), ambiguous bases (N=0), overall quality and sample assignment based on barcodes/MIDs; average length of the reads was 350.45 bp. Sequences were assigned a taxonomic classification with figure 4.8 displaying the proportional representation of the 13 most prevalent families. Based on this evaluation of the sample composition a number of T-tests were conducted to compare: the standard diet and organic diet groups; non-inoculated with inoculated; and early-cull inoculated with late-cull inoculated. Of these tests only Bacteroidaceae, Prevotellaceae and Bifidobacteriaceae between non-inoculated and inoculated, Bacteroidaceae between organic and standard, and Lachnospiraceae between early and late-cull inoculated were significant at the $\alpha = 0.05$ level ($p = 0.037, 0.032, 0.04, 0.048, 0.021$ repectively). Implementation of the (conservative) Bonferroni correction for multiple tests subsequently brought all values above the $\alpha = 0.05$ threshold.

Figure 4.9 Run 2 family level classification and percentage representation

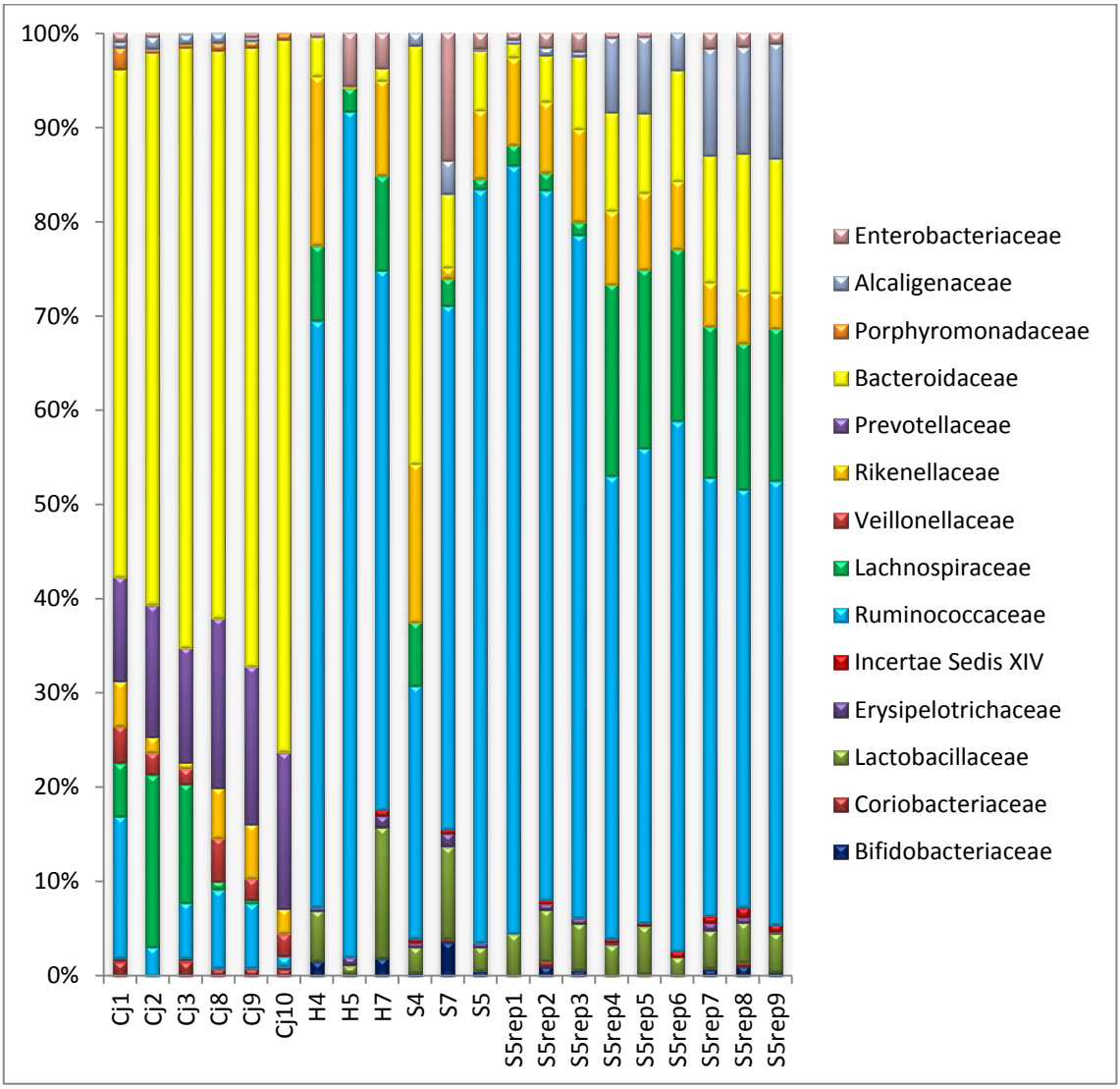


Figure 4.9 shows classification by family for Run 2. Legend shows the 13 most prevalent families; for inclusion a family must have accounted for at least 2% of the total in at least 1 of the samples. Cj1-Cj3 were inoculated with C. jejuni and culled early; Cj8-10 were inoculated with C. jejuni and culled late; H4, H5 and H7 were not inoculated and fed on an organic diet; S4, S5 and S7 were not inoculated and fed on a standard diet

As a whole the composition of the samples is in broad agreement with previous studies of the caecal microflora of chickens, with domination by the family *Ruminococcaceae* (Zhu *et al.*, 2002; Zhu and Joerger, 2003), previously clostridial cluster IV,

Lachnospiraceae (Wise and Siragusa, 2005), formerly designated clostridial cluster XIV, and *Bacteroidaceae* (Lu *et al*., 2007). This is evident primarily for the non-inoculated birds, those infected with *C. jejuni* displaying a greater prevalence of bacteroidaceae and prevotellaceae; interestingly, the inoculated groups distinguished only by time of cull show a decreased number of lachnospiraceae as duration of infection increases.

11 of the families displayed in figure 4.8 were encountered in the S5rep series and values were utilised for one-way ANOVA on the 3 primer groups to evaluate primer-derived bias. P-values for each of the tests are shown in table 4.4 along with Bonferroni-corrected values to adjust for multiple testing. Values in red indicate a corrected p-value below the alpha threshold and therefore significant differences between the replicate groups.

Table 4.6 P-values from family level ANOVA for S5rep series at alpha = 0.05

Family	p-value	Bonferroni-corrected
Bifidobacteriaceae	0.105656	1
Coriobacteriaceae	0.782215	1
Lactobacillaceae	0.225516	1
Erysipelotrichaceae	0.27349	1
Incertae Sedis XIV	0.006278	0.069062
Ruminococcaceae	9.76E-05	0.001074
Lachnospiraceae	1.79E-07	1.97×10^{-6}
Rikenellaceae	0.003331	0.036646
Bacteroidaceae	0.004098	0.045079
Alcaligenaceae	0.000225	0.002472
Enterobacteriaceae	0.024393	0.268321

As mentioned, read numbers were lower than expected, and replicate numbers were not sufficient to provide significant calculation power but table 4.4 suggests that choice of primers can significantly affect the composition estimates from a quantitative structural perspective, five of 11 comparisons being significant at $\alpha = 0.05$. However, of 27 families detected in the S5rep series only 5 were missing from one of the other groups, and none were peculiar to one group alone, so the low output and skewed inter-sample distribution (minimum reads per sample = 197, maximum = 2858) are likely to have contributed significantly to the ANOVA calculation. In addition, despite superficial similarity, the coefficients of variation (COV) within groups per family were in the range 4-71%, possibly since COV is susceptible to value ranges close to zero where the mean is low in relation to the variance.

Despite the caveats, the data tentatively support the established paradigm for the optimal approach to PCR-based microbial community studies: amplicon pools should be produced from a variety of oligonucleotide pairs to minimise primer-induced bias (Schmalenberger *et al*., 2001). However, for the purposes of the current research it was considered that this would introduce undesirable practical complexity in that a greater number of tagged primers would be required and intra-sample rarefaction would also become essential. The resultant compromise was the use of a single primer pair for the remainder of the experiments; 926F and 1391R were chosen for this purpose by the research group as a whole, based on recommendations from the HMP and a fresh review of the literature. Early PCR tests and sequencing of cloned inserts from small-scale libraries displayed that the pair would be suitable for further experimentation.

4.4.4 Analysis with 'R'

A subsidiary aim of the second 454 run was to investigate the suitability of 'R' for analysis of pyrosequencing data to supplement the capabilities provided by the RDP and Mothur. In particular, the former provides only rudimentary statistical options while the

latter is restrictive in terms of presentation of data. The 'R' script utilised in this instance is detailed in the appendix 12 and provides for processing of files produced by the RDP, filtering of sequences of archaeal and eukaryotic origin, derivation of OTU heatmaps and PCA (principal component analysis).

4.4.4.1 Input

Input files for 'R' were created through deconvolution of fasta files in the RDP with subsequent use of the classifier function. The assignment files for each of the following samples were utilised: Cj1, S4 and S5rep1 to S5rep9. Initial processing examines the bootstrap confidence levels of assignments and resulted in the removal of any sequences assigned to the kingdom *Archaea* (857 sequences), or classified as bacteria with less than 95% confidence (821 sequences); 99.04% of these sequences were from the S5rep7 to S5rep9 series indicating the unsuitability of 1115R and 515F for future investigations. Filtering in this manner introduces flexibility into the classification process since a defined cutoff is required by the RDP. If this value is too low sequences such as archaea are included, while a higher threshold causes the process to be too stringent leading to an excessive proportion of sequences remaining unclassified at lower taxonomic levels.

4.4.4.2 OTU heatmaps

Filtered sequence files from above were passed to subsequent stages of the script for creation of OTU heatmaps. While this function is available in Mothur the output files cannot be manipulated such that figures produced are relatively uninformative. OTU heatmaps for family and genus were derived with logarithmic and root transformations in addition to the basic proportional data to identify the best fit for multi-variate community data. Heatmaps are shown in figures 4.10 and 4.11 with details in the text.

Figure 4.10 Genus and family level OTU heatmaps for S5rep series with raw abundance

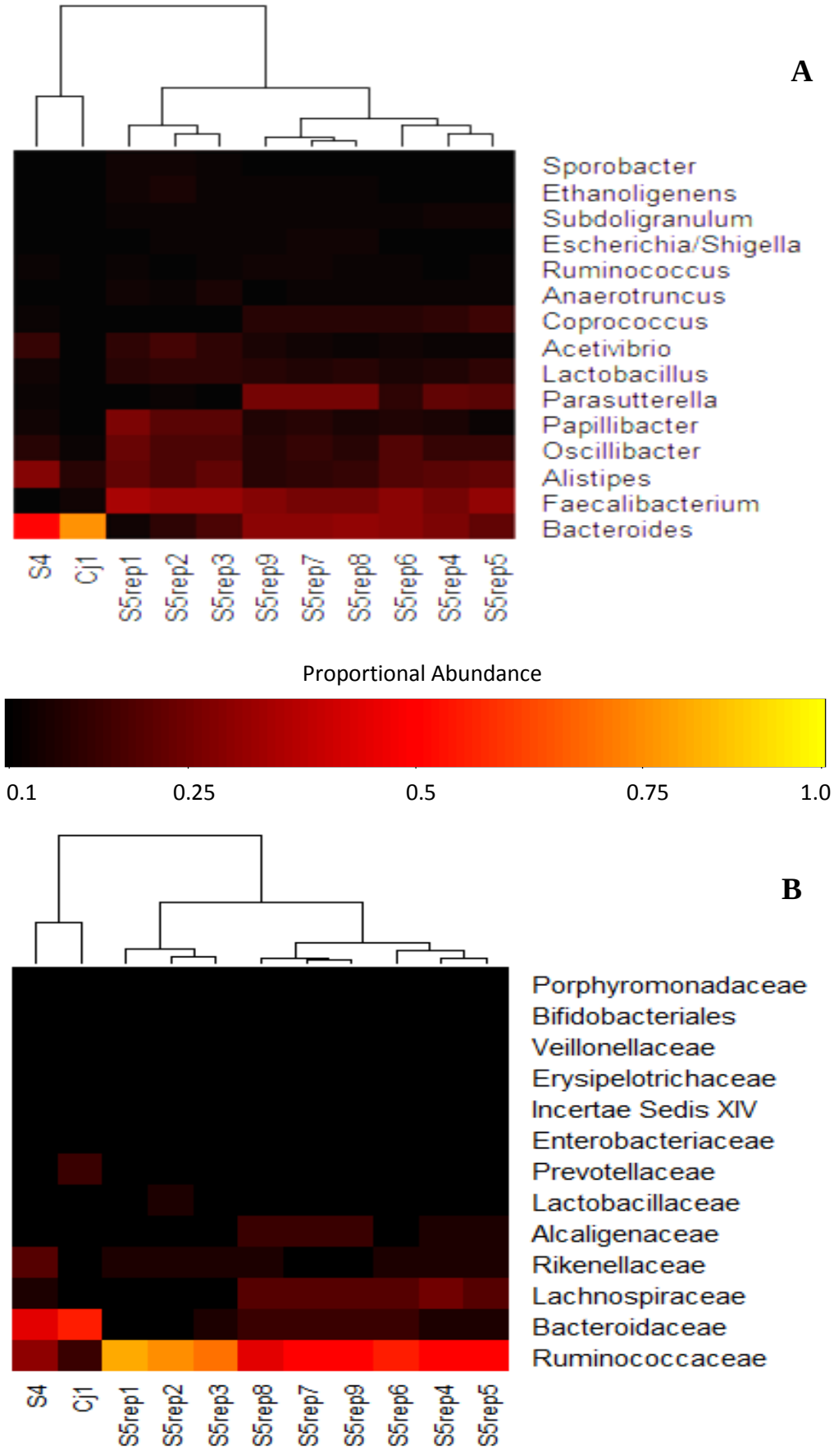


Figure 4.10 displays genus (A) and family (B) level heatmaps for the S5rep series of samples with S4 and Cj1 as outgroups. Each of the heatmaps includes a clustering dendrogram calculated by 'R' as part of the heatmap function. Heatmaps in the figure are based on simple proportional abundance without transformation. Primer repeat samples cluster together at both family and genus level.

Figures 4.10 and 4.11 show the OTU heatmap data for the S5rep series of samples inclusive of data for Cj1 and S4 as 'outgroups'. The 15 most abundant genera and 13 most abundant families are shown, selected by the 'R' script from the original OTU tables on the basis of representing at least 2% of the total count of at least one sample. For each figure, 'A' is the genus heatmap and 'B' is the family heatmap. Taxonomic association is represented on the left of each heatmap while samples are represented along the bottom edge ordered according to integral 'R' clustering as shown by the dendrograms along the top edge. The figures display differing transformations of the data: figure 4.10 shows raw abundance while 4.11 shows root-transformed abundance.

Despite the transformations both heatmaps were unsatisfactory to a degree, in that discrimination is limited for those OTUs with lower abundance; unfortunately these account for a considerable proportion of microbial communities and the heatmaps are actually representative of the data in that the samples are dominated by a relatively small number of genera and families e.g *Faecalibacterium*, *Bacteroides*, *Ruminococcaceae*, *Lachnospiraceae* and *Bacteroidaceae*.

Figure 4.11 displays genus (A) and family (B) level heatmaps for the S5rep series of samples with S4 and Cj1 as outgroups. Each of the heatmaps includes a clustering dendrogram calculated by 'R' as part of the heatmap function. Heatmaps in the figure are based on root-transformed abundance without transformation. Primer repeat samples cluster together at both genus and family level.

Figure 4.11 Genus and family level OTU heatmaps for S5rep series with root-transformed abundance

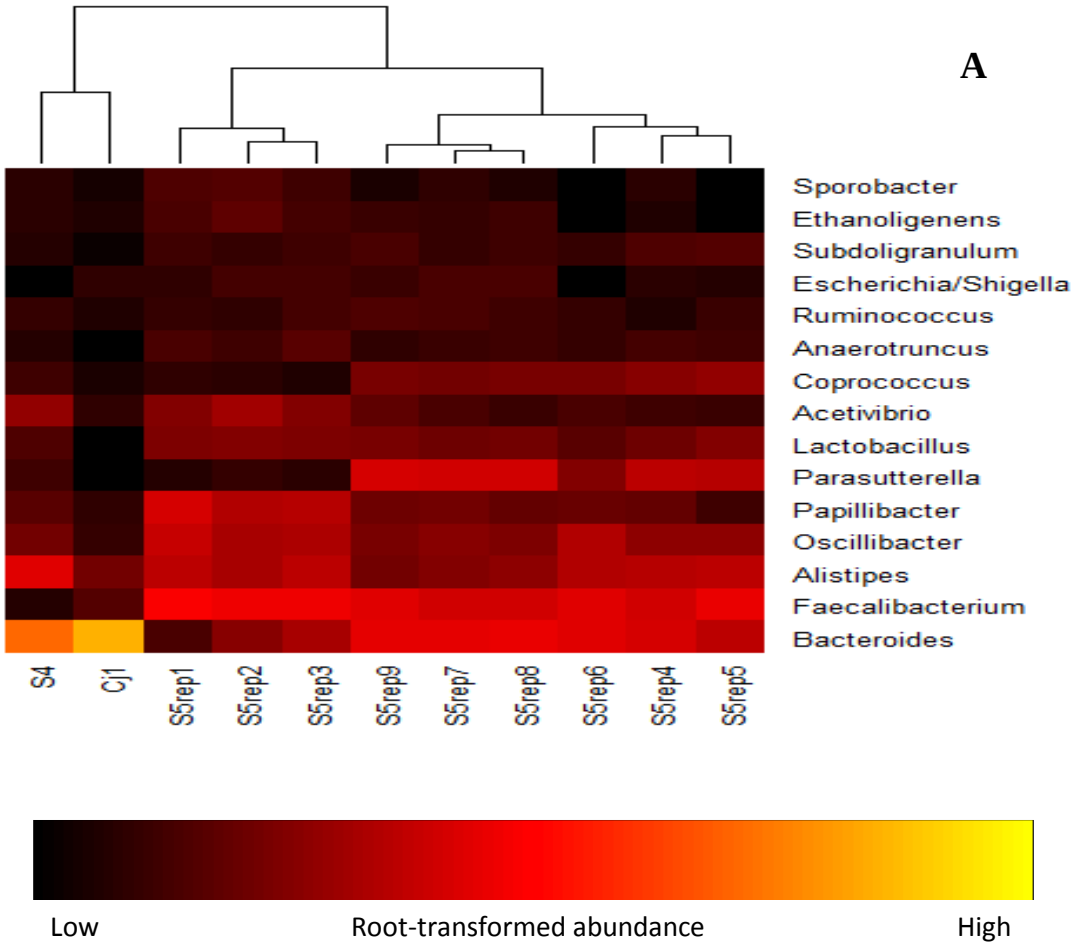


Figure 4.12 PCA plot of PC1 vs PC2 for Run2 at genus level

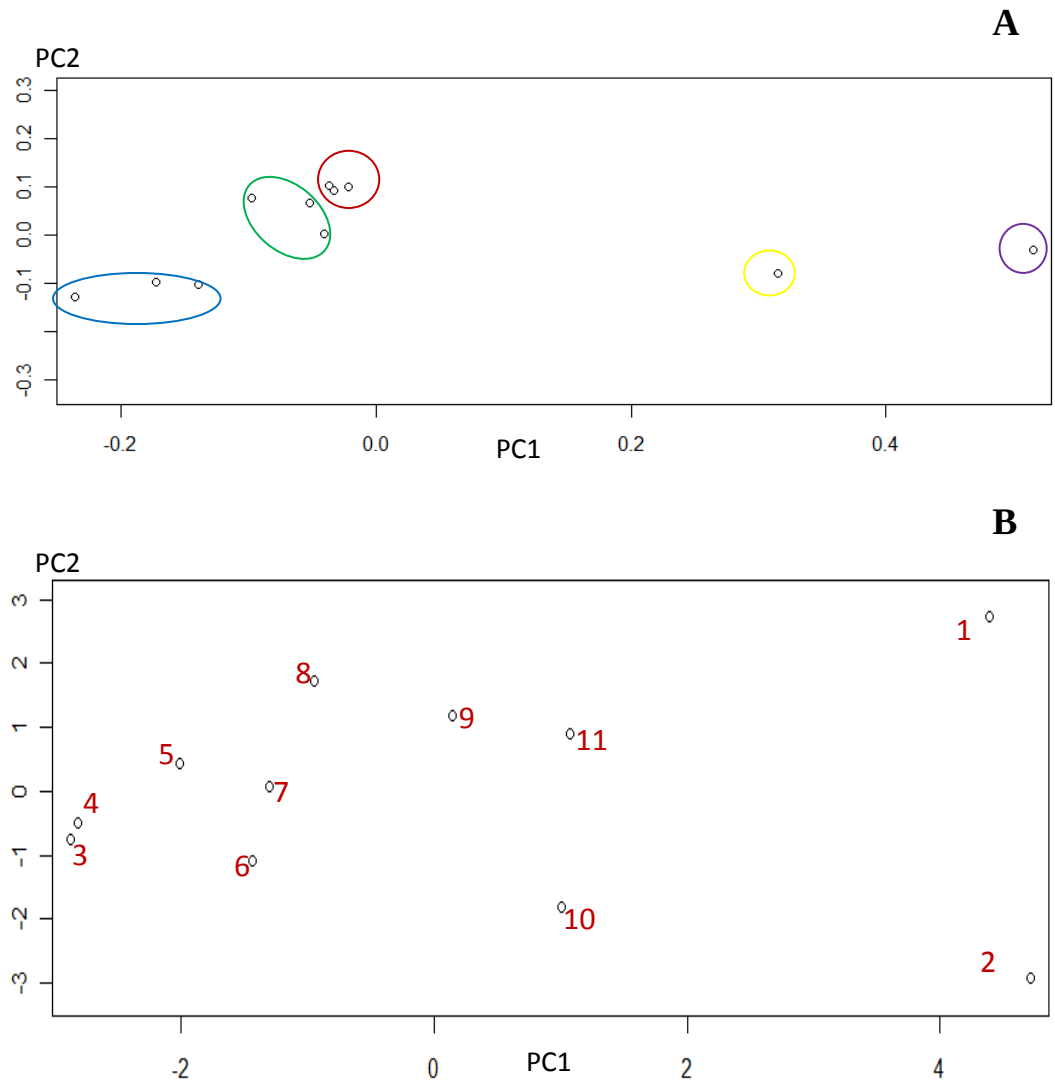


Figure 4.12 displays PCA plots for S5rep samples, Cj1 and S4 (A) with the corresponding loading plot for the contributory genera (B). The X-axis shows the value for PC1 and y-axis shows the value for PC2. Blue circle encloses S5rep1-3; green circle encloses S5rep4-6; red circle encloses S5rep7-9; yellow circle is S4; purple circle is Cj1. OTUs: 1 = Faecalibacterium; 2 = Bacteroides; 3 = Ruminococcus; 4 = Anaerotruncus; 5 = Acetivibrio; 6 = Coprococcus; 7 = Lactobacillus; 8 = Papillibacter; 9 = Oscillibacter; 10 = Parasuterella; 11 = Alistipes

4.4.4.3 PCA

Heatmap dendrograms displayed a number of potential clustering patterns based purely on the parameters utilised by the heatmap function of 'R', not necessarily corresponding to the true relatedness of samples. Clustering is best visualised in this

respect through ordination techniques such as principal component analysis (PCA) whereby the variance between samples across many characteristics (in this case, OTUs) is converted into eigenvalues which can be plotted two-dimensionally. Figure 4.12 displays the PCA genus abundance plot created in 'R' for the Srep series with Cj1 and S4 included as 'outgroups'. Total variation in the samples is effectively explained by 11 principal components, with PC1 and PC2 accounting for 93.8% of this. Also included in the figure is the contribution (or loading) made by the OTUs to the first 2 principal components plotted. While the S5rep series cluster together they display particular association within the respective primer pairs, some separation being evident between the groups primarily due to differing abundances of the genera *Bacteroides*, *Parasutterella* and *Oscillibacter*.

The various analyses of Run 2 with 'R' provided greater insight into how the data for repeat samples with differing primer pairs varied and an opportunity to test approaches to presentation of data. The heatmaps and PCA plot have also confirmed that while the primer repeats cluster together there remain compositional differences which prevent them from being considered as true replicates.

4.4.5 Metastats

A further resource for analysis of microbial community data is the Metastats application (White *et al*., 2009), available online at http://metastats.cbcb.umd.edu/. Metastats takes OTU abundance tables and, like the RDP libcompare function, calculates p-values for the differences between communities through a variant of the T-test. The advantages of Metastats, though, are the ability to compare groups of libraries and application of multiple test corrections.

OTU tables were manually compiled from RDP hierarchy files for samples Cj1 and Srep1 to 9, the purpose of Cj1 being to serve as an outgroup. Comparisons were then undertaken between Cj1 and each of the primer groups individually, and then between

the respective primer groups. The average number of OTUs showing a significant difference ($p < 0.05$) between the outgroup and the Srep series was 8.1%; the average of the inter-group comparisons was 6.2%. A T-test for these values confirmed that there was no significant difference ($p = 0.1353$ at $\alpha = 0.05$), indicating that the primer repeats were as dissimilar to each other as an unrelated outgroup sample. Clearly, though, choice of outgroup sample informs this result, and other chicken caecal samples are not truly unrelated.

4.4.6 Run 2 summary

Overall, the changes in diet and state of infection imposed on the chickens did not appear to have caused the expected alteration in the caecal microbiota. Although read numbers were low, the superficial similarity of samples within groups and for the replicates of S5 with the differing primer pairs suggest that this was not a function of lack of coverage. However, the relatively small sample size may have impacted on the potential to observe significant differences. It is noteworthy that a number of taxonomic groups do show partitioning between the sample sets and the disruption caused by infection is clearly visible, particularly in terms of numbers of the 2 main families of firmicutes (*Ruminococcaceae* and *Lachnospiraceae*) and the family *Bacteroidaceae*. It is also of interest that lachnospiraceae are diminished over time after inoculation.

Heatmaps and PCA displayed that S5 replicates with differing primer pairs clustered together and distinct from included 'outgroups', but analysis of the groups via Metastats suggested that the repeats should be considered as showing the expected experimentally-induced variability (Schmalenberger *et al*., 2001).

4.5 Run 3: CDAD and normal control

The third preliminary 454 run was undertaken while samples for the final phase were being identified and collected. The aim was to utilise QIIME (Quantitative Insights into Microbial Ecology), a platform developed specifically for analysis of 454 data (Caporaso *et al*., 2010b), with a view to its integration into the final processing pipeline. In addition, the potential variability introduced by the extraction process and PCR were to be investigated using human stool samples from subjects with CDAD and a 'normal' volunteer

4.5.1 Samples

Nine faecal samples were obtained from 3 human subjects, two with CDAD and the third a normal volunteer with no known infections or pathophysiological conditions. Eight of the samples were collected over the course of 90 days by the staff of the Clinical Infectious Diseases Unit of the LRI with the permission of Dr Martin Wisielka and the respective patients and stored at 4°C for a maximum of 7 days until further processing. Seven of the samples represent a longitudinal time-series from the same patient at various time-points during the course of the infection (Table 4.5), the eighth from an unrelated CDAD subject. The final sample was obtained from a volunteer with their permission as to its use in the investigation and extracted immediately. Samples were designated as follows: CDLady1-CDLady7 for the longitudinal series; CD1A, CD1B and CD1C for repeat extractions of the individual CDAD subject; and VOL1.1, VOL1.2, VOL1.3 and VOL1.4 for repeat amplification reactions of the identical extract from the volunteer subject. Table 4.7 displays the limited clinical metadata available for the CDLady series of samples, with timepoint in days, result of tests for *C. difficile* toxin, and clinical notes inclusive of treatment regime.

Table 4.7 Clinical metadata for CDL series of samples

Sample	Timepoint	CDT	Clinical notes
CDLady1	0	+ve	Diagnostic; diarrhoea for 3+ days
CDLady2	+3	+ve	2nd diagnostic
CDLady3	+50	=ve	Recurrence diagnostic (trimethoprim)
CDLady4	+55	+ve	Colitis (vancomycin, metronidazole)
CDLady5	+64	-ve	No diarrhoea (imipimem, vancomycin)
CDLady6	+76	+ve	Diarrhoea (4 days post vancomycin cessation)
CDLady7	+89	-ve	No diarrhoea (vancomycin)

4.5.2 Preparation of amplicons

Community bacterial DNA was extracted from samples within 7 days of collection as described previously in section 4.2.1, the solitary sample from one of the CDAD subjects being processed in triplicate to provide 3 separate substrate mixtures for amplification. Normalised concentrations of sample DNA were amplified using the 454 primer pair 926F and 1391R (table 2.1). These had been synthesized with Roche fusion adapter sequences A (CCATCTCATCCCTGCGTGTCTCCGACTCAG) and B (CCTATCCCCTGTGTGCCTTGGCAGTCTCAG) complexed to their respective 5′ ends while the forward primer also included the sample-specific decamer MIDs between adapter and 16S-specific sequence. Constituents of reaction mixtures were as detailed in section 4.2.3 and PCR was performed in Eppendorf Mastercyclers with the

following conditions: 98°C for 5 min followed by 30 cycles of 98°C for 30 s, 60°C for 40 s and 72°C for 25 s, with a final extension at 72°C for 4 min. Visualisation of amplicons to verify products of the expected size and identify samples with artefactual bands (thus requiring repeat PCR) was through electrophoresis on 1.5% agarose gels; resultant gels are shown in figure 4.13.

Sample amplicons were then purified, quantified, diluted and pooled as detailed in section 4.2.4 prior to submission to the Genomic Services Facility, University of Leicester for emPCR and sequencing, occuping one half of a PTP chip.

Figure 4.13 AGE visualisation of amplicons for Run 3

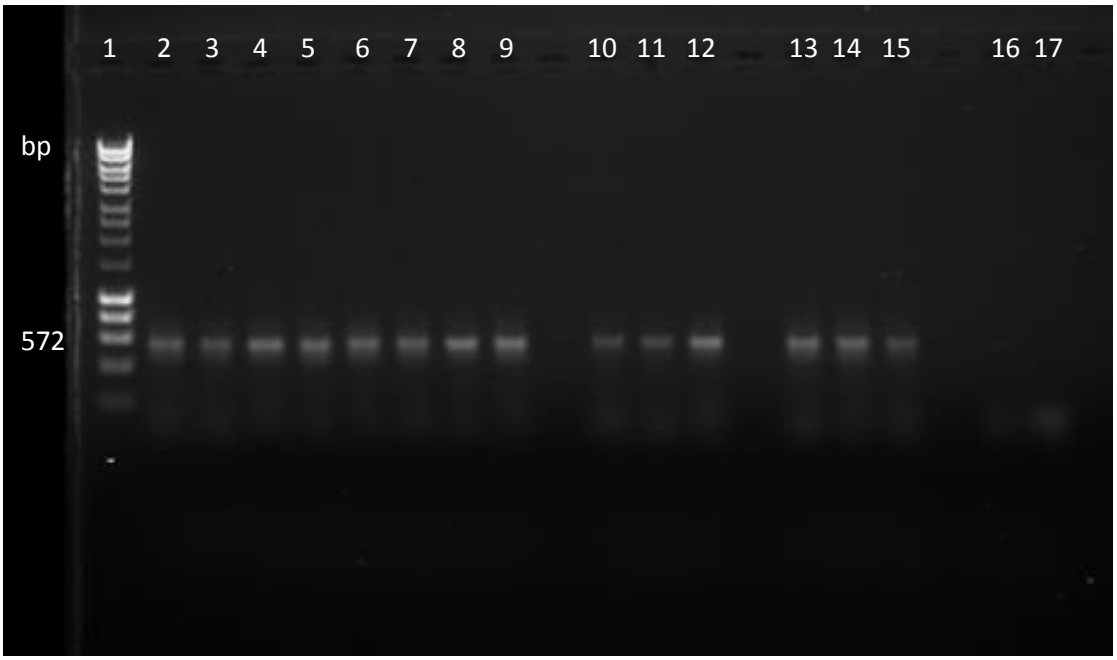


Figure 4.13 shows PCR products on 1.5% agarose gel. Lane 1 = Hyperladder 1; Lanes 2-8 = CDLady1-7; Lane 9 = CD1; Lanes 10-12 = CD1A, CD1B and CD1C (PCRs of repeat extractions); Lanes 13-15 = VOL1.1, VOL1.2, VOL1.3; Lane 16 = PCR of blank extract; Lane 17 = PCR negative control. Expected product is 572 bp in length.

4.5.3 Run overview

Initial output for Run 3 amounted to 232,744 reads which were uploaded to the RDP for deconvolution, filtering and trimming, with parameters set to 150 nucelotides (minimum length) and zero ambiguous bases permitted. Resultant sample files contained a total of 148,016 reads with an average length of 252 nucleotides, the majority of reads being removed on the basis of quality and length. Range of reads per sample was between 7,605 and 25,635 with a further 2,497 not being allocated to any sample due to errors in the barcodes. Figure 4.14 is a histogram of the number of reads over the range of sequence lengths subsequent to filtering with the RDP.

Figure 4.14 Read length histogram for Run 3

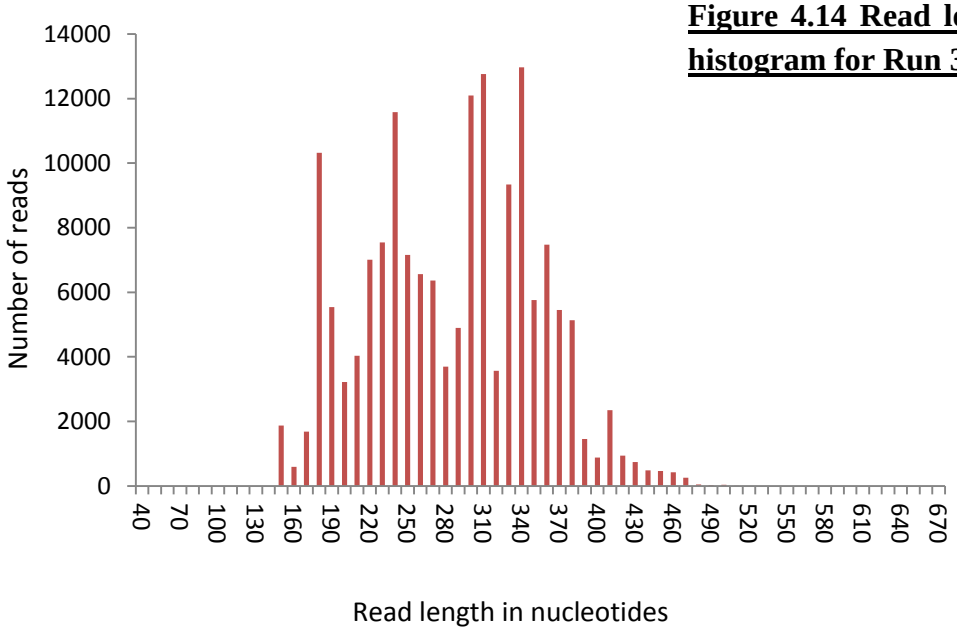


Figure 4.14 displays the read length distribution for 454 Run 3 subsequent to initial filtering stages which remove sequences below 150 bp and reads with poor quality

4.5.4 QIIME

Sequences from the RDP were subsequently exported to the QIIME platform (Caporaso *et al*., 2010b), introduced in section 4.2.5. The version utilised here was Qiime 1.3.0 on the Virtual Box and the commands employed are detailed in Appendix 13. In addition to the read files, a mapping file relating reads to samples and including metadata is also required. Initial processing in QIIME included clustering (uclust) at 0.03 cutoff, selection of representative sequences, alignment, chimera-checking, phylogenetic tree derivation, and ultimately, creation of OTU tables which form the basis for all subsequent analyses. Recommended even sampling depth for subsequent rarefaction analysis was calculated by QIIME as 5,000 sequences per sample.

4.5.4.1 Classification

QIIME classifies reads by comparison of a reference sequence from each OTU with the Greengenes dataset, although still allowing for use of the RDP nomenclature. Figure 4.15 displays the composition of each sample by percentage at the class level, VOL1.1 being absent due to the return of less than 200 reads for this sample. It is evident that the repeat samples superficially show a strong degree of correlation and that the various CD samples are distinguished by a high proportion of the class *Gamma-proteobacteria* compared to the 'normal' volunteer, where the classes *Clostridium* and *Bacteroidia* account for more than 90% of the reads. At lower levels of taxonomic assignment the genera *Enterococcus*, *Escherichia*, *Kluyvera*, *Klebsiella*, and *Salmonella* were the most prevalent across the CDAD samples with the VOL1 samples displaying greater levels of Bacteroides and Parabacteroides along with numerous indeterminate genera from the *Clostridiales* order. The seventh of the longitudinal samples (CDLady7) also displays a drastic reduction in the number of Proteobacteria detected compared to preceding samples; in terms of a clinical correlation, this was one of only 2 samples to test negative for *Clostridium difficile*.

Figure 4.15 Percentage classification at class level for Run 3

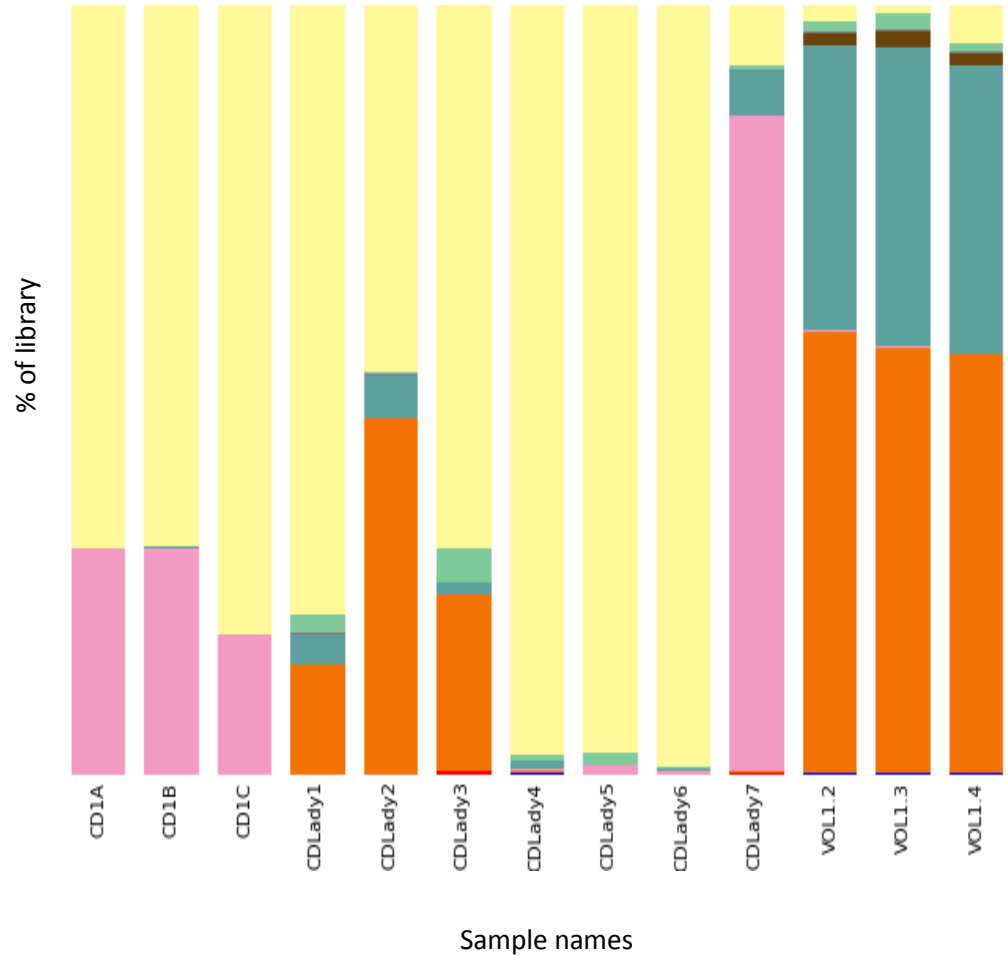


Figure 4.15 shows Run3 Qiime classification comparison for the six most prevalent classes. Other classes accounted for less than 5% of each sample and were excluded from the figure, although not from further analysis. CD1A-CD1C are repeat extractions of aa faecal sample from a patient with CDAD; CDLady1-7 are longitudinal samples from a further patient with CDAD; VOL1.2-1.4 are repeat PCRs of a 'normal' volunteer

4.5.4.2 Alpha diversity

Subsequent to alignment, clustering, and creation of OTU tables, data was rarefied at a level of 5000 sequences per sample allowing for comparisons at equivalent levels of sequencing depth. Calculation of diversity indices at intervals of 500 sequences up to this point indicated that asymptote had been reached. A total of 9 indices were calculated in Qiime including the Berger-Parker index of dominance and the Simpson index as shown in Table 4.8. Patterns in the data were independent of index chosen to represent diversity, partially negating the concerns about their relative suitability and correct application (Magurran, 2004; Hill *et al*., 2003; Jost, 2007).

Table 4.8 Alpha diversity values for Run 3

SAMPLE	Berger-Parker	Simpson
CD1A	0.3024	0.77686
CD1B	0.3254	0.765785
CD1C	0.3736	0.758645
CDLady1	0.7108	0.47767
CDLady2	0.4084	0.671853
CDLady3	0.3384	0.796743
CDLady4	0.4018	0.74006
CDLady5	0.8086	0.333367
CDLady6	0.43	0.703118
CDLady7	0.5862	0.626698
VOL1.2	0.3322	0.835275
VOL1.3	0.3246	0.841096
VOL1.4	0.3188	0.859675

Values approaching 1 for the Simpson index indicate greater diversity, while those approaching the same value for the Berger-Parker index indicate that a high proportion of sequences are assigned to a relatively small number of OTUs. While the dominance values do not differ signifiacntly between the groups (CDAD against normal) a T-test of the Simpson's values displays a clear partition, the normal samples being more diverse than those from the CDAD patients ($P = 0.00208$; $\alpha = 0.05$). However, the small sample size and nature of the normal/control group (repeat PCRs of the same extraction) mean the result must be treated with caution.

4.5.4.3 Beta diversity and Principal Co-ordinate Analysis

In addition to those indices which represent the alpha diversity of samples, a variety of ecological metrics can be applied to estimate the similarity of the samples. Many of these were previously encountered in Mothur, such as the Morisita-Horn, Bray-Curtis and Kulcyzcynski-Cody similarity coefficients and the Euclidean measure of similarity. While all were utilised to verify consistency of clustering, values for the MH were prefered for final preparation of plots for reasons outlined previously. Prior to calculation of similarity coefficients all read sets had been rarefied at an even sampling depth of 5000 sequences, while 3D-PCoA biplots were visualised using Kinemage (http://kinemage.biochem.duke.edu/).

Figure 4.16 displays PCoA data for Run 3 samples, with clear partitioning of the groups apart from the final longitudinal (recovery/ CD -ve) sample, which clusters with the repeat extraction samples. The inclusion of weighted taxon information allows confirmation of the earlier observation that the VOL1 samples are distinguished by the relative prevalence of *Clostridiales* and *Bacteroides* while separation of the extraction and longitudinal (CD1 v CDL) samples can be attributed to *Enterococcus* and *Salmonella* in the CD1 series, and *Kluyvera* and, to a lesser extent, *Proteus* in the CDL

series. Clustering of samples in this manner, CDL7 associating with the CD1 group, was observed in all PCoA plots constructed from the various similarity coefficients, the basis for this being the high proportion of *Bacillus* in this sample. It is not a simple matter to visually interpret the 3D-biplot without the interactive facility for rotation of the axes but CDL1,CDL2 and CDL 3 also show displacement along the PC3 axis, whereby the contribution of *Bacteroides* is dominant and causes a degree of correlation between these 3 samples and those of VOL1. The close relationship of the repeat samples is encouraging and suggests that PCR and extraction procedures do not have a significant effect on the snapshot of the community.

Figure 4.16 Morisita-Horn 3-D PCoA bi-plot for Run 3

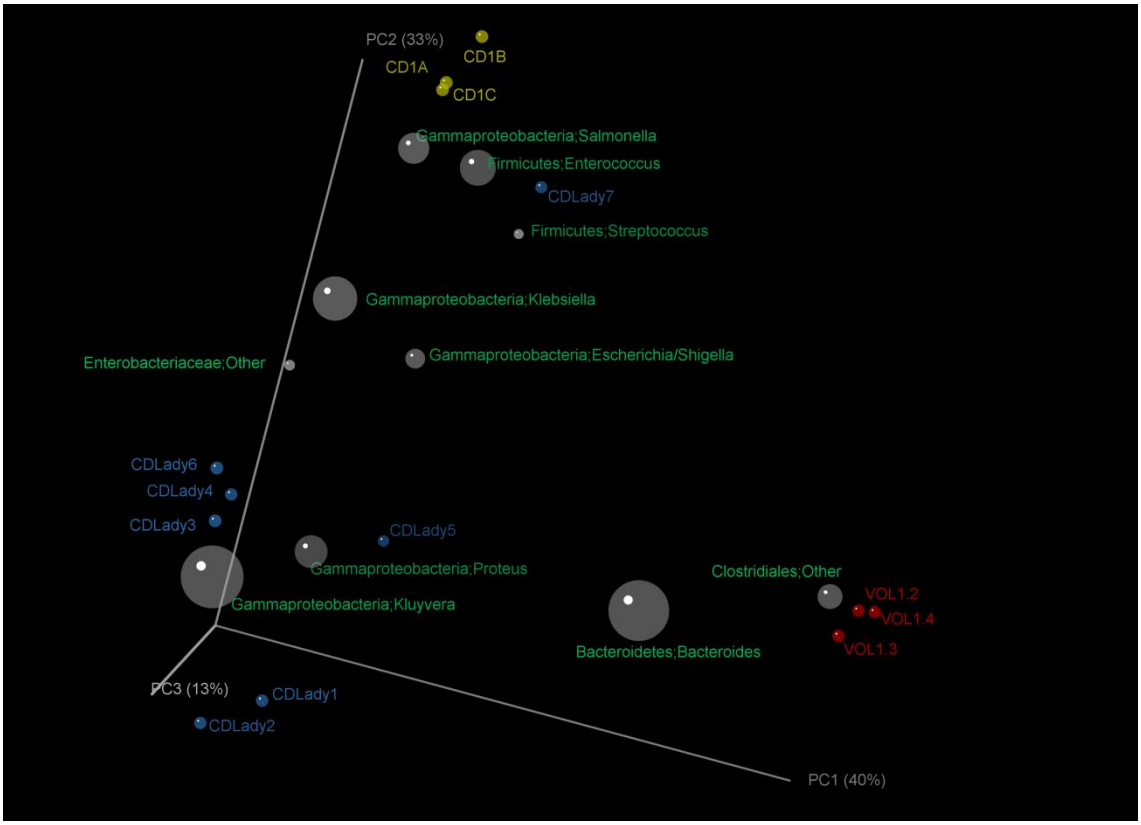


Figure 4.16 displays 3D-PCoA MH biplot for Run3. CD1 samples are yellow, CDLady are blue and VOL1 are red. The first three PCs are shown and account for 86% of the variation. Also shown are grey spheres with green taxonomic labels representing relative contributions of OTUs to the dispersion and clustering of groups, the size of the sphere indicating the weighting.

Integrated into Qiime is the software Cytoscape (Shannon *et al*., 2003; Smoot *et al*., 2011; http://www.cytoscape.org/), a tool for visualisation of biological networks which takes data from Qiime as input. The OTUs and samples are assigned node status with 'edges' connecting samples to OTUs in which they are represented, the 'weight' of the edge being proportional to the membership. Edge weights then contribute to clustering in that an OTU node is 'pulled' to its connecting sample; where multiple samples contribute to an OTU the algorithm acts to position the nodes such that the overall tension in the network is minimised, thus causing clustering both of OTUs and related samples. Such a network for the Run 3 samples is displayed in Figure 4.17.

Figure 4.17 Cytoscape network image of samples from Run3

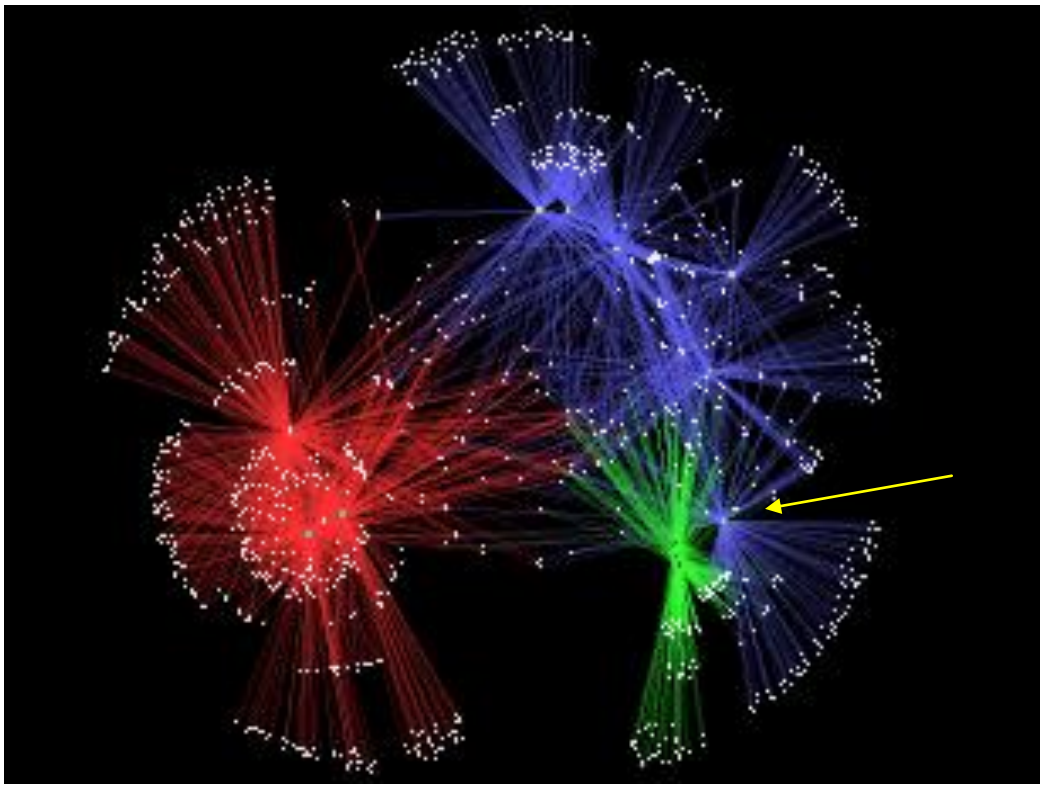


Figure 4.17 displays association of samples and OTUs for Run 3. Sample nodes are not easily discerned but can be seen at the centre of radiations, purple for CD1, yellow for CDL and aqua for VOL1. OTU nodes are white. Edge colours are green for CD1, blue for CDL and red for VOL1. Yellow arrow points to the CDLady7 node, once again clustering with CD1 samples.

The Cytoscape figure provides further confirmation of the clustering pattern in that the distances between repeat samples are minimal while the CD1 and CDL groups are more closely associated with each other than with VOL1.

4.5.4.4 UPGMA clustering

UPGMA (Unweighted Pair Group Method with Arithmetic mean) clustering is a further approach to grouping of samples using average linkage, and in Qiime can be combined with jackknifing (repeated sub-sampling of the sets) to estimate the validity of the tree created.

Figure 4.18 UPGMA clustering for Run3

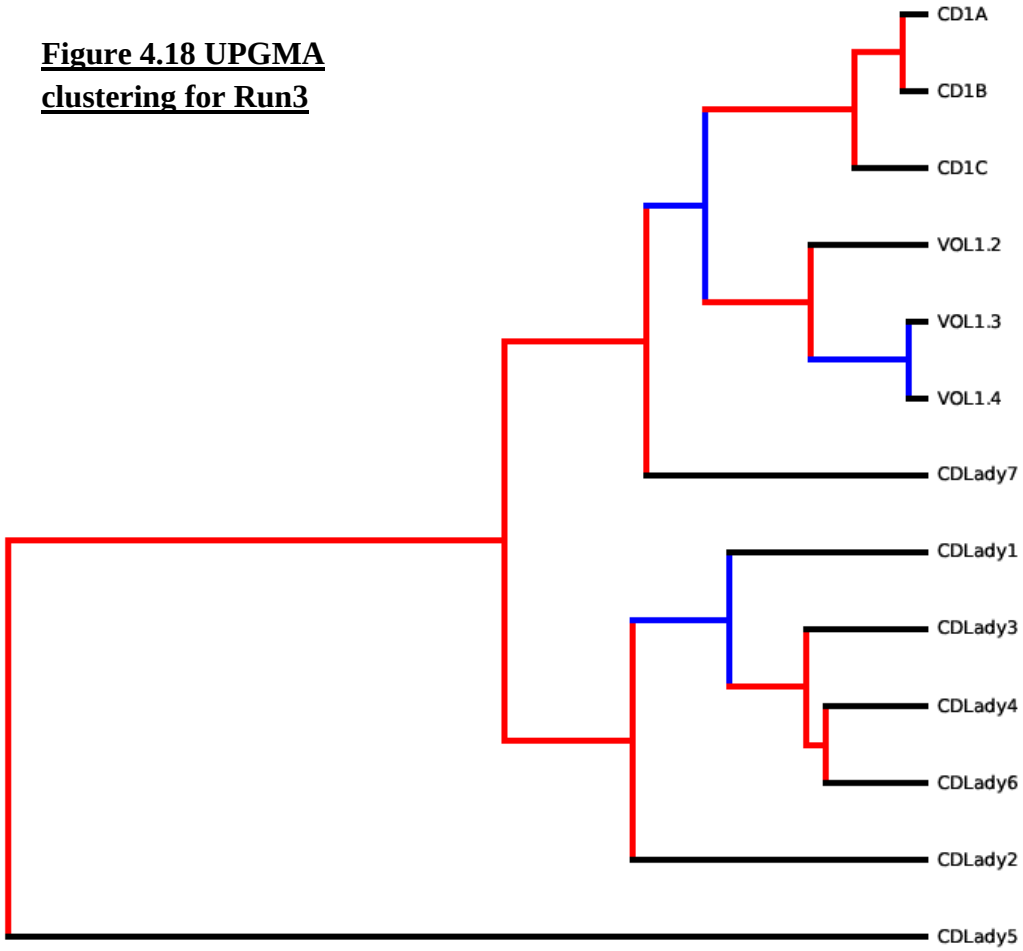


Figure 4.18 shows clustering of samples from Run 3, with jackknifed support and using Euclidean measure of similarity. Red colouring of nodes indicates > 75% support while blue colouring indicates < 25% support.

Figure 4.18 displays the jackknifed UPGMA results for Run 3 using the Euclidean measure of similarity on OTU tables rarefied at 500 read intervals from 500 up to 5000 with 10 iterations at each level, all of which are then analysed via tree_compare.py and make_bootstrapped_tree.py to create the master tree.

Node colouring of the tree indicates that this derivation is strongly supported and generally reflects the clustering seen with other approaches, with only one of the CDLady internal nodes and separation of the CD1 and VOL1 samples providing low confidence values. However, once again, the CDLady7 sample clusters away from other longitudinal samples while all those from repeat groups cluster together.

4.5.4.5 Statistical approaches

Qiime provides for two groups of statistical tests the first of which implement the UniFrac (Lozupone *et al*., 2006) algorithms for comparison of trees, resulting in either the weighted UniFrac significance output or the P-test value (Martin, 2002). In both cases the OTU table and the master tree for the entire run are the input, the script then creating sub-trees for each sample before comparison of their structure. Since there are multiple comparisons the values are subsequently scaled using the Bonferroni correction, an conservative implementation for reduction of type-I errors (false positives). The P-test values for inter-sample comparisons were all less than 0.02 apart from the following: CD1A v CD1B; CD1A v CD1C; CD1B v CD1C; VOL1.2 v VOL1.3; VOL1.2 v VOL1.4; and VOL1.3 v VOL1.4, all of which returned corrected values of greater than 0.05. The P-test values thus provide statistical support at the 95% confidence level for the clustering displayed in earlier figures and for the reproducibility of output with repeat PCRs and extractions.

In addition to gross comparisons of libraries the compositional differences between samples and groups can be assessed through the use of ANOVA, the G-test of independence, and Pearson's correlation on longitudinal samples or for additional

variables. The CDL series of samples can be broadly described in terms of taxonomic presence by a sustained dominance of Proteobacteria with a reduction in Bacteroidetes numbers, until the final sample where Bacilli become prevalent at the expense of all other taxa. Without a baseline sample prior to infection and incomplete clinical information Pearson's correlation was performed only to verify that no definitive pattern could be discerned, the expectation being that the statistic would be more usefully applied for the final set of samples. While 'r' values for the genera *Bacteroides*, *Lactonifactor* (of the family *Ruminococcaceae*) and *Roseburia* (of the family *Lachnospiraceae*) were close to -1 (negative correlation), and those for a number of proteobacterial genera (*Klebsiella*, *Kluyvera*, *Proteus* and *Escherichia*) approached zero (indicating little relative change), FDR-corrected p-values were all in excess of 0.5 so no firm conclusions about OTUs associated with the course of infection can be drawn.

The difference between the G-test and ANOVA is analogous to that between indices of community membership and community structure; the former based on the presence or absence of taxa and the latter incorporating abundance within OTUs. ANOVA tests were performed with the groups CD1, CDL and VOL1 to identify OTU categories showing differential abundance; although the experimental design was not directed towards this estimation (so the calculation of ANOVA is somewhat artificial) it was hoped that preliminary findings might guide the final phase of investigations. All OTU tables were rarefied to even sampling depth (5000 reads) prior to calculation of p-values with Bonferroni and FDR corrections. Of 1129 OTUs, 242 were found to be significant for ANOVA (α=0.05), 98 of these displaying reduced prevalence in comparisons between VOL1 and both of the other 2 groups. The majority of these were from the phyla *Firmicutes* and *Bacteroidetes*, only the genera *Collinsella*, *Atopobium* (both phylum *Actinobacteria*) and *Sutterella* (of the class *Betaproteobacteria*) originating from other phyla. Interestingly, 3 separate OTUs classified as *Sutterella* appear in this

group (indicative of different species at 97% cluster cutoff) with one of these providing a value from the G-test of 0.21, the lowest overall for FDR-corrected values with this statistic. Whilst this value is not considered significant the genus *Sutterella* may prove of interest in later trials. The G-test generally provided higher values than ANOVA, possibly due to a greater effect on the power of the test when sample numbers are low.

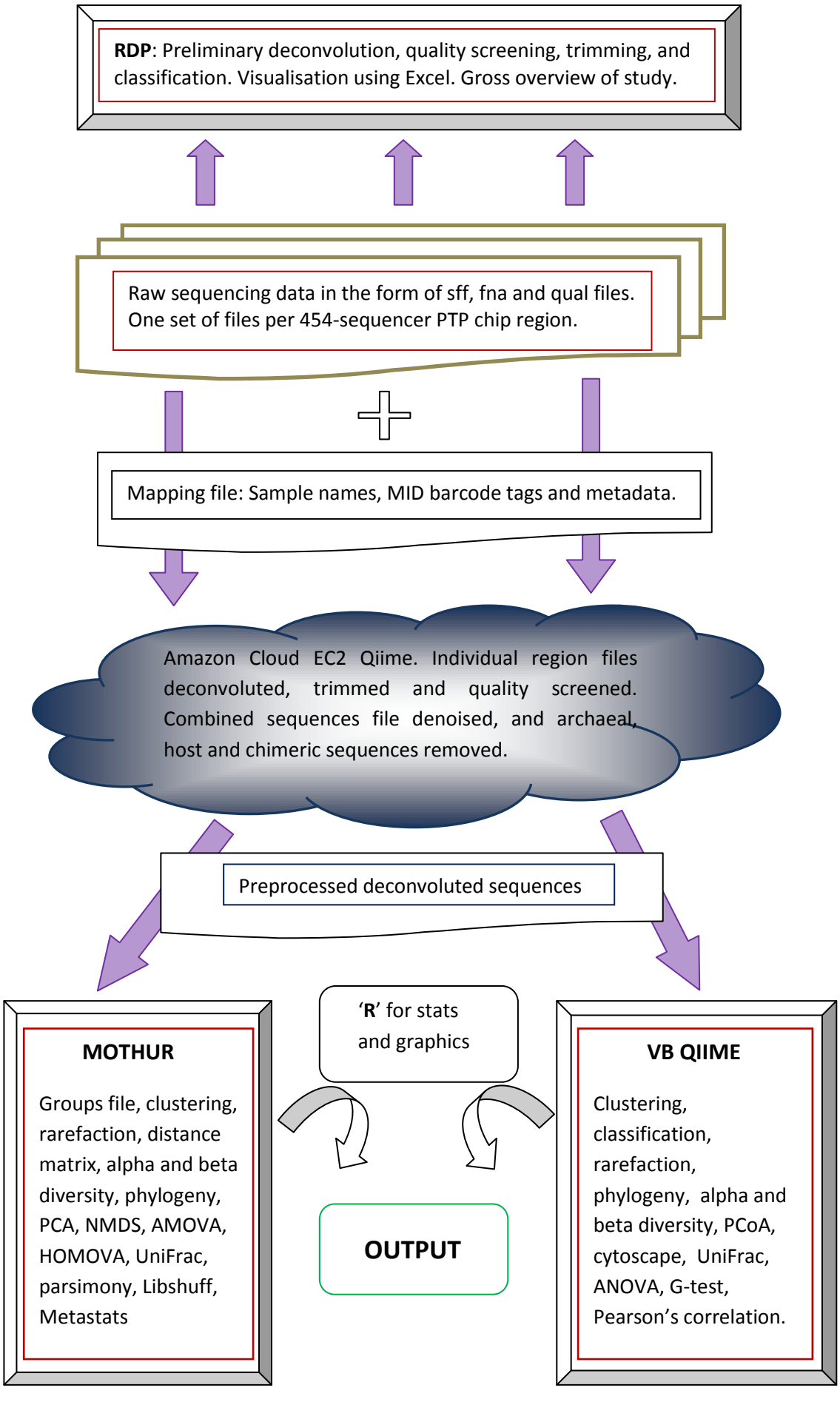
4.6 Summary and final 454 analysis pipeline

The investigations described in the preceding sections had shed little biological light on bacterial taxa which may contribute to colonisation resistance, or provide markers for susceptibility to CDAD, although studies had not been conducted specifically with this aim. However, data from the experiments detailed in section 4.5 had highlighted a reduced diversity in CDAD samples in comparison to samples derived from a normal subject, a finding which would be further examined in the final phase.

Where the preliminary experiments had proved vital was in their contribution to development and optimisation of practical procedures and the analysis pipeline. In particular, the applications available for data processing were novel, while additional features were provided with each version especially in the early stages of development. The final analysis pipeline is an amalgamation of the various platforms utilised in the 3 preliminary 454 runs, incorporating the RDP, Mothur, 'R' and Qiime, and is displayed in Figure 4.19, a detailed list of commands being described in Appendix 13.

A number of the investigations to this point had focussed on variability introduced by methodological approaches. Results from repeat PCRs and extractions displayed overall reproducibility as evidenced by consistency of OTU tables, alpha diversity values (Simpson's index, Fisher's alpha index, Berger-Parker dominance index), beta diversity values (Morisita-Horn index, Bray-Curtis index, euclidean measure of similarity and Jaccard/Sørensen coefficients), PCA plots and UPGMA clustering. Where these provided a value, differences between repeats were no greater than 5%, or a single standard deviation from the mean. Pooling of material from repeat extractions and amplicons from replicate PCRs was therefore added to the methodology for the final phase.

Figure 4.19 454 Data Analysis Workflow



Choice of primer pair for the investigations, however, was found to affect the apparent composition of the community, although this was evident more from the perspective of structure (abundance per OTU) than membership (presence/absence of OTUs). To ameliorate this effect the ideal approach is a mixture of primer pairs (Schmalenberger *et al*., 2001), each barcoded to provide a link to a sample for deconvolution. However, this was deemed impractical for current purposes due to the extra cost (primers and other PCR constituents), and complexities introduced at the analytical stage where each primer pair must be subjected to denoising, chimera-checking and alignment independently. In lieu of an approach involving multiple primer pairs, consistent application of an empirically-approved pair with broad coverage of the bacterial domain was deemed necessary. Since 926F and 1391R had been tested on human samples and provide the requisite coverage this primer pair was selected for use in the final phase of experiments. As with any investigation, consistency in application of general methodology, nomenclature, statistical tests and ecological indices is also of paramount importance; rarefaction, such that the number of reads per sample is standarised to a comparable level, can be viewed as falling within this category.

In addition, results indicated that longitudinal samples would prove vital to meaningful analysis since baseline inter-individual variation could be considerable (Eckburg *et al*., 2005; Gill *et al*., 2006), while improved metadata for the samples (provenance, clinical status, antibiotics administered) was also deemed beneficial if correlation between clinical status and microbiotic compositionwas to be examined. The primary limitation of investigations to date had been the relatively low number of available samples which had restricted both the implementation of statistics and the power of the resultant values. While this could not be fully rectified for the final phase due to ethical considerations, it was hoped that the sample set would at least contain multiple representatives for the each of the desired groups.

CHAPTER 5

CDAD vs AAD

5.1 Introduction

In the final phase of the study the aim was to apply the protocols and workflows described previously to investigation of the intestinal microbiota in both patients with *Clostridium difficile*-associated diarrhoea and subjects with antibiotic-associated diarrhoea. In addition, 'normal' volunteers who had undergone antibiotic treatment without experiencing subsequent bowel dysfunction were selected as controls, inclusive of a pre-trial sample representing their basal intestinal microbiota.

The objective, as previously stated, was to identify bacterial taxa acting as markers for a predisposition towards, or protection against, infection with *Clostridium difficile*. In particular, distinctive characteristics of an AAD-microbiota in contrast to that found in CDAD patients would be indicative of a protective mechanism, this cohort being most closely matched for age, infirmity, antibiotic treatment and the potential for exposure to *Clostridium difficile*.

It is worth noting, at this stage, the difficulties encountered both with acquisition of samples, and accumulation of sufficient data. The intention had been that ethical approval for a preliminary study into CDAD would be obtained relatively promptly but, for various reasons, this was not the case, and samples became available in an unsatisfactory piecemeal fashion. In addition, for reasons discussed in section 6.4, the volume of sequence reads provided by 454 was not commensurate with expectations. Final read numbers were low despite stringent quality control at the amplicon preparation stage and 2 sequencing attempts on the second pool submitted to the CGR.

5.2 Samples

Accumulation of samples suitable for analysis took place over the course of two years, with collection, storage, and subsequent extraction of community DNA as per section 4.2.1. All CDAD and AAD samples were acquired from the LRI with the permission of Dr Martin Wiszielka and the patients themselves. Samples were categorised as either AAD or CDAD based on clinical observations (loose stools for a period of at least 3 days subsequent to antibiotic treatment in the past 6 weeks) and laboratory diagnostic tests (absence or presence of CD toxin A or B as assessed by ELISA of stool samples); a total of 5 CDAD and 5 AAD faecal samples were eventually included in subsequent steps. Samples from normal volunteers were donated by various members of Lab 121 of the Genetics Department (University of Leicester); community bacterial DNA was then extracted immediately and archived for future potential utilisation at -20°C. Where one of the volunteers subsequently underwent treatment with antibiotics a further sample was also acquired. All subjects were aware of the purpose of sample collection and all data was fully anonymised.

Samples were allocated a tag based on their category, the five CDAD being CD 1-5, the AAD being AAD 1-5, and those from the 'normal' control volunteers being annotated as V1-5, where V1.1 is the basal sample and V1.2 is the sample collected subsequent to antibiotic treatment. Volunteer 5 provided only a pre-antibiotic sample such that a total of 19 samples were utilised for amplicon production and analysis.

5.3 Amplicon Preparation

Amplicons suitable for submission to the CGR (Centre for Genomic Research) sequencing centre at the University of Liverpool for 454 processing were prepared as described previously in section 4.5.2 with primers 926F and 1391R (table 2.1). Primers were purchased synthesized to include the Roche-454 fusion tags and sample-specific MIDs between the 16S region and the A-adapter sequence of the 926F oligonucleotide. PCR constituents and cycling conditions (sections 4.2.3 and 4.5.2 respectively) were optimised such that all 19 samples could be processed simultaneously along with negative PCR and blank extraction controls. Final PCR products were visualised via agarose gel electrophoresis as shown in figure 5.1 and, where possible, multiple PCR reactions for each sample were mixed prior to subsequent steps.

PCR products were then purified, quantified, diluted and pooled as per section 4.2.4, with final verification of amplicon presence, size and quantity, in addition to overall quality of samples, using the Agilent BioAnalyser system as described in the same section. Where possible, designated PCR labs and/or cabinets were utilised for reaction preparation to avoid contamination of amplicons. In addition, quality control involved incorporation of negative controls at every stage of the procedure, along with comparison of values and figures with those obtained perviously from successful sequencing runs.

Samples were packed into dry ice containers and transported to the CGR for sample quality verification, emPCR, enrichment, and sequencing. The final dataset was derived from 2 demi-regions and a quarter-region of PTP devices (a total of 1.25 chips) with data being provided in the form of .sff (simple flowgram format) files of between 800 MB and 1.6 GB.

Figure 5.1 AGE of amplicons for final 454 run before (B) and after (A) AMPure

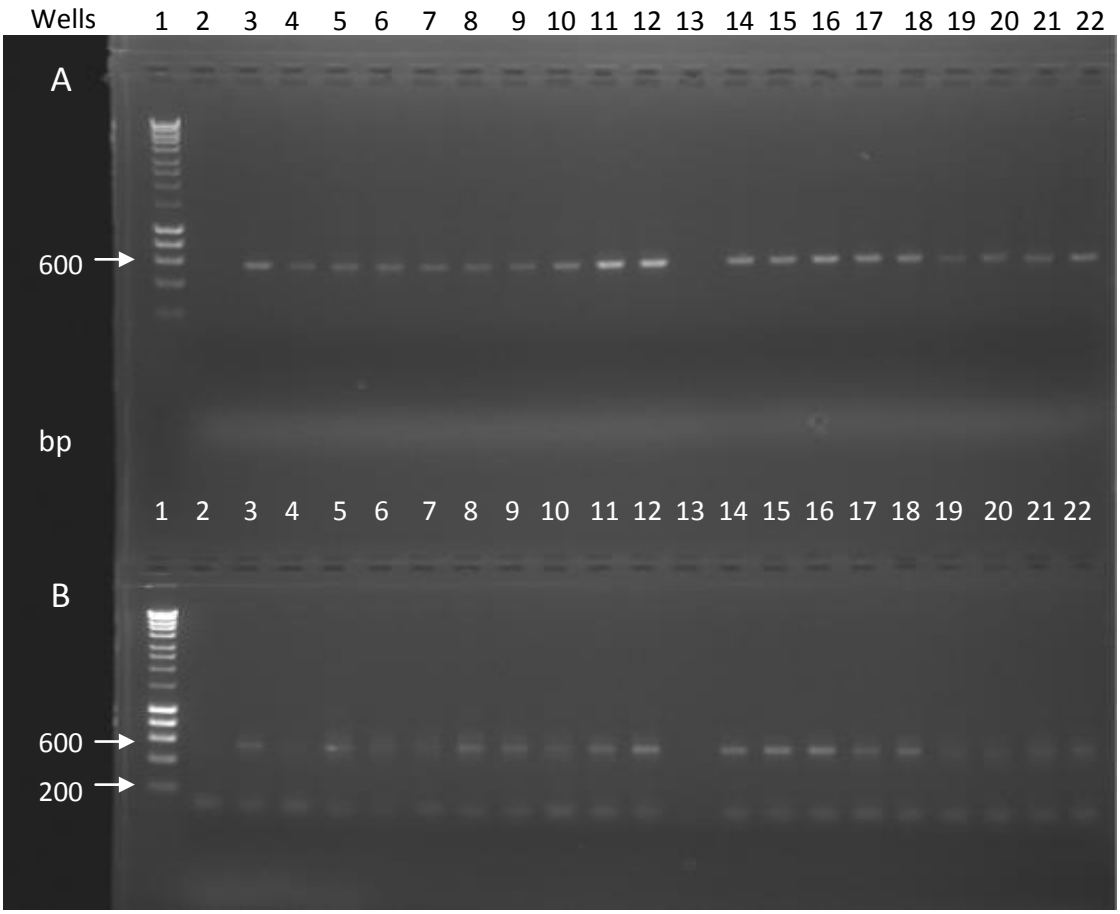


Figure 5.1 shows 1.5% agarose gel electrophoresis images showing 572 bp amplicons prepared for final 454 sequencing run. Lane 1 - Hyperladder I DNA marker; Lane 2 - Negative PCR control; Lanes 3 to 7 - CDAD; Lanes 8 to 12 - AAD; Lane 13 – blank extraction control; Lanes 14 to 21 – Pre and post antibiotic samples from normal volunteers; Lane 22 – Single sample from normal volunteer. Set (B) shows the PCR products prior to AMPure clean-up. The band at approximately 140 bp represents primer dimers and is absent after AMPure.

5.4 Output and initial processing

For initial processing of data inclusive of deconvolution of reads into respective samples according to barcode, filtering of reads for quality (average score across 50 bp windows > 20) and length (between 100 and 450 bp), denoising (Reeder and Knight, 2010) and chimera-checking (Chimeraslayer; Haas *et al*., 2011), Qiime version 1.4.0 was employed (Caporaso *et al*., 2010a). Since these steps are computationally intensive, this required use of the Amazon Cloud (EC2) Qiime V1.4.0 image (ami-458d5b2c) with n3phele (http://www.n3phele.com/) as a gateway to facilitate flexibility. Prior to denoising and chimera-checking to provide the final output, the integrated PYNAST aligner (Caporaso *et al*., 2010b), uclust OTU picker (Edgar, 2010), and Greengenes core reference alignment (DeSantis *et al*., 2006a) were utilised. All steps for bioinformatic analysis are described in Appendix 14.

Figure 5.2 Number of reads per sample subsequent to deconvolution, quality filtering, and chimera-checking

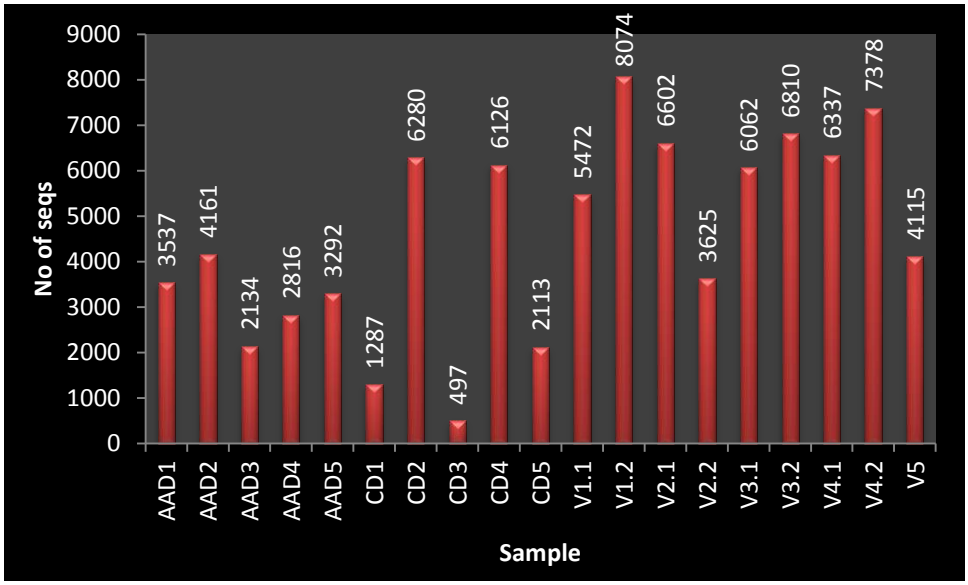


Figure 5.2 shows final number of reads per sample subsequent to deconvolution according to decamer barcodes (MIDs or tags), quality filtering (number of unassigned nucleotides or low quality reads) and chimera-checking. CD1 and CD3 were removed from certain stages of analysis requiring a minimum of 2000 reads per sample.

The numbers of raw sequence reads obtained for the three regions were: 168,494 (A, ½ chip), 95,157 (B, ¼ chip), and 225,696 (C, ½ chip), for a total of 489,347 reads. Sequence numbers subsequent to initial filtering and quality checking amounted to : 30,934 (A), 36,522 (B), and 40,953 (C), for a total of 108,301. This represents a loss of nearly 80% of the data, possible reasons for which are discussed in section 6.5.

Figure 5.3 Length distribution histograms for final combined 454 output

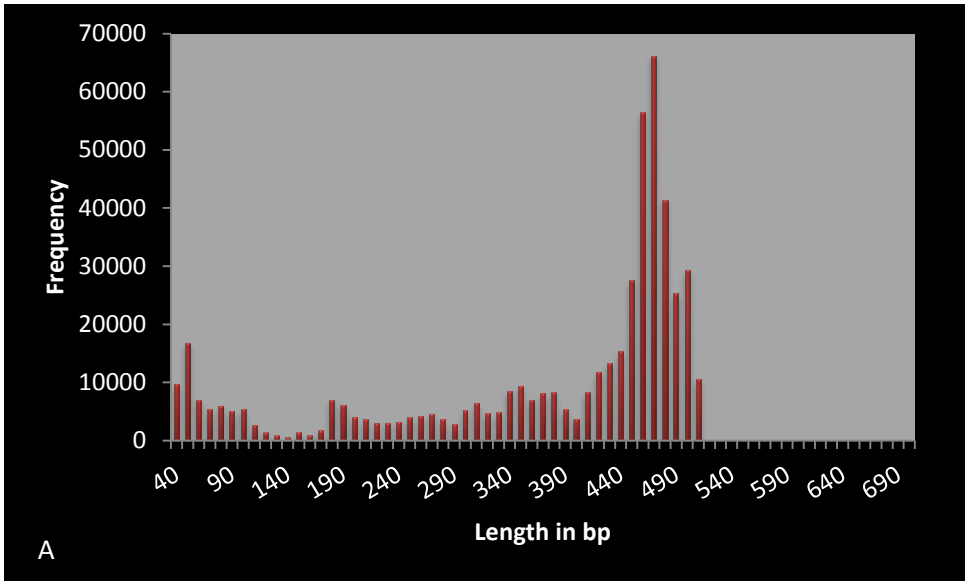
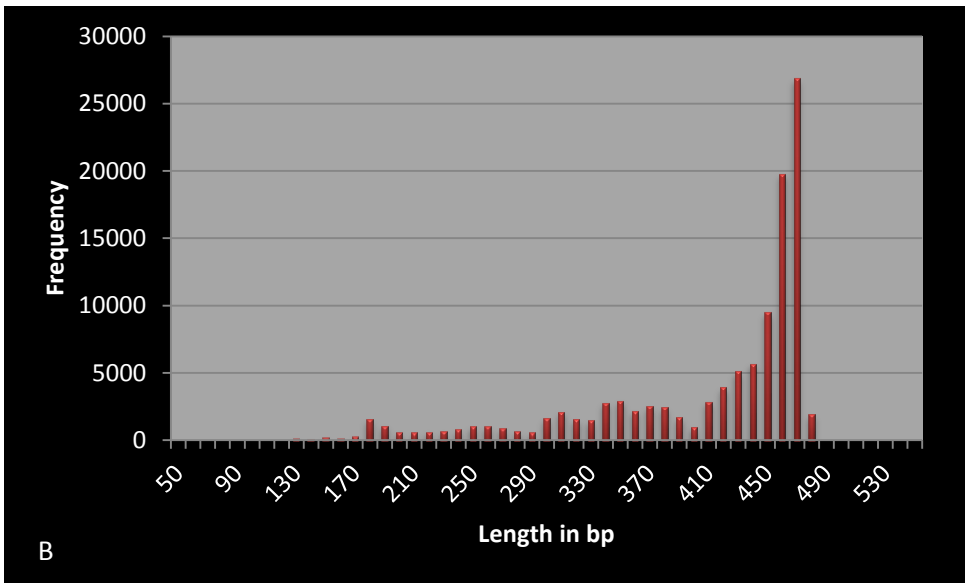


Figure 5.3 shows length distribution histograms for the combined output of the 3 half chips. Figure A is the output prior to filtering while figure B shows numbers subsequent to filtering for poor quality and length.



A further 21,583 reads were discarded by Chimeraslayer to leave a total of 86,718 for the 19 samples, an average of 4,564 per sample with a mean length of 411 bp. Distribution of reads between the samples is shown in figure 5.2, while figure 5.3 displays the range of lengths of reads rejected at the initial filtering stage.

Time and funding constraints prevented resequencing of the samples or preparation of further amplicon pools for submission – regions A and B represent attempts to perform such remedial action – so despite particularly low numbers for certain samples such as CD1 and CD3, the decision was made to proceed with analysis and accept that observations would lack statistical power.

In addition to Qiime version 1.4.0, running locally on VirtualBox-4.1.2-73507-Win, certain steps in analysis required the use of:

- Mothur version 1.21.1 (Schloss *et al*., 2009)
- 'R' version 2.13.0 (R Development Core Team, 2008)
- Ribosomal database project release 10 (Olsen *et al*., 1992; Wang *et al*., 2007; Cole *et al*., 2009)

Prior to chimera-checking, sequences from the 3 regions were combined into a single fasta file. Once divested of chimeras, this file was utilised as input for Mothur and, subsequent to deconvolution into the individual sample files, for classification of reads using the RDP.

5.5 Classification

Sequence files for the individual samples were uploaded to the RDP and classified with a bootstrap cutoff of 80%. All but 40 of the sequences were classified at least to the level of phylum and no archaeal sequences were identified.

Phyla encountered were *Actinobacteria*, *Bacteroidetes*, *Proteobacteria*, *Firmicutes*, *Synergistetes* and *Verrucomicrobia*, although the latter two only in 3 samples and extremely low numbers. T-tests were conducted on percentage abundance of the first four phyla between the AAD, CDAD, and control groups. No distinction was drawn between the samples acquired from the 'normal' volunteers since it became apparent that all but one had waited less than 24 hours after inception of the course of antibiotics before providing the sample. In addition, little variation could be discerned from superficial pairwise comparison of percentage abundance at the higher phylogenetic levels.

P-values for T-tests at the level of phylum displayed significance for the following at $\alpha = 0.05$:

- AAD v Control, Bacteroidetes = 0.02304
- CDAD v Control, Bacteroidetes = 0.00009032
- CDAD v Control, Proteobacteria = 0.0156

Figure 5.4 displays the total composition of the samples by percentage at the level of class. A total of 13 classes were encountered, *Bacteroidia* and *Clostridium* dominating in the control group but diminished or absent in the CDAD group, while the AAD samples are more variable in this respect. The relative lack of variation between the pre- and post-antibiotic samples for the 'normal' cohort is evident, only V3.1 and V3.2 displaying a single noticeable difference in the abundance of *Betaproteobacteria*. The comparative prevalence of *Gammaproteobacteria* and *Bacillus* outside of the control group is also apparent.

Figure 5.4 Class level proportional abundance derived from RDP

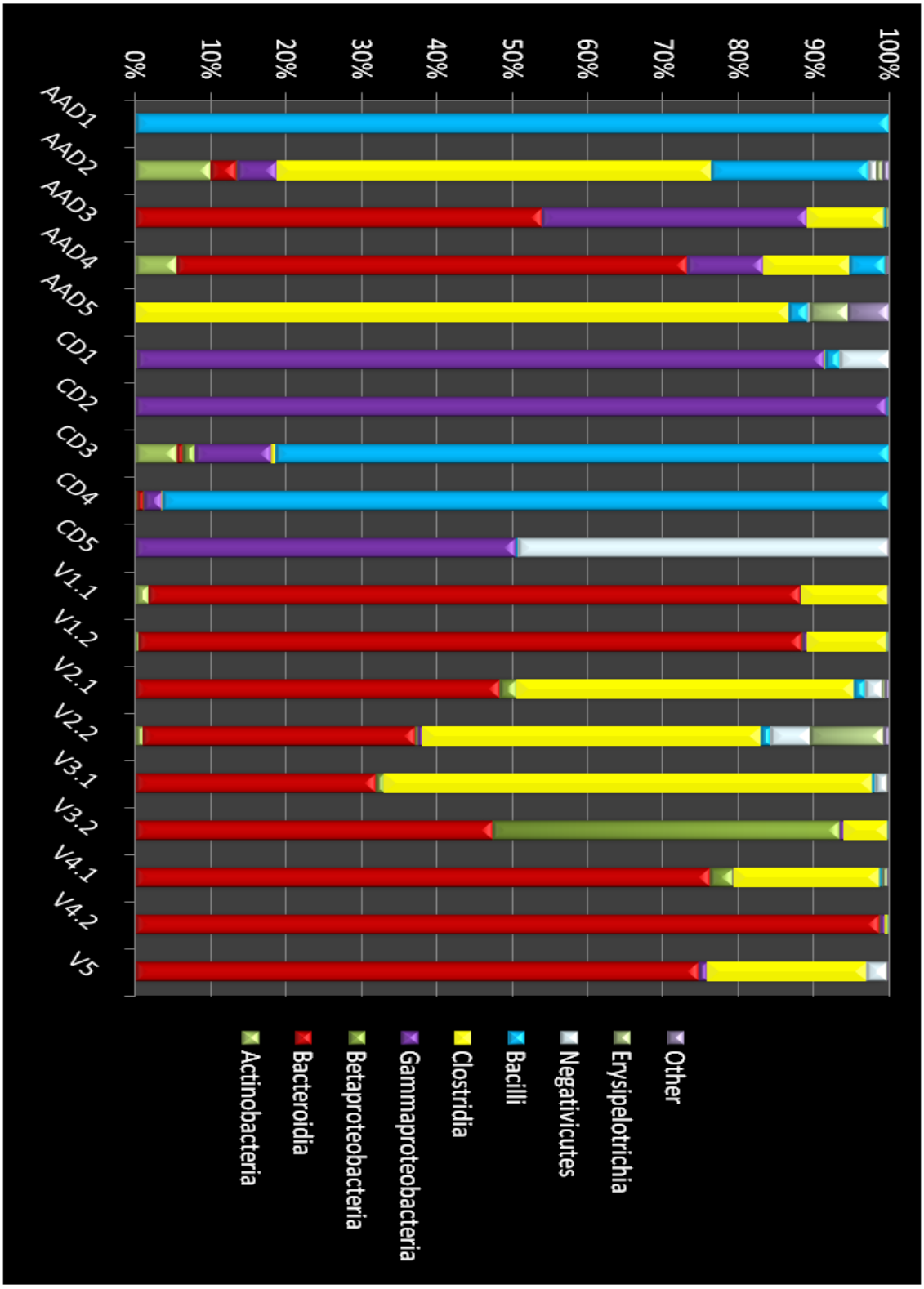


Figure 5.4 shows proportional abundance (percent) at class level per sample asessed using the RDP classifier (Wang *et al*., 2007). Integrated legend shows bacterial classes including Negativicutes, a new taxonomy for Gram-negative Firmicutes. The class Negativicutes includes such families as the Veillonellaceae and Acidaminococcaceae. Other classes are Alphaproteobacteria, Sphingobacteria, Flavobacteria, Synergistia, and Verrucomicrobiae.

Bacteroidia and *Actinobacteria* were the only classes encountered for their respective phyla so statistical tests were unnecessary for these taxonomic ranks, but T-tests were performed for *Clostridium*, *Erysipelotrichi*, *Bacillus*, *Negativicutes*, *Gammaproteobacteria*, and *Betaproteobacteria*, all of which displayed at least 5% abundance in a minimum of 2 samples. Statistically significant differences were evident ($\alpha = 0.05$) for *Clostridium* between the CDAD and AAD groups ($p = 0.023$), and the CDAD and control subjects ($p = 0.000084$), and for *Gammaproteobacteria* between the CDAD and control samples ($p = 0.0143$). Application of the Bonferroni correction adjusted the values for *Gammaproteobacteria*, and *Clostridia* between CDAD and AAD groups, to above the threshold of $\alpha = 0.05$; for the test of *Clostridia* between CDAD and control samples the p-value remained below the threshold ($p=0.003$). In addition, *Erysipelotrichi* were entirely absent from the CDAD group, but relatively low abundance in the control and AAD sets (and the lack of statistical power due to low sample numbers) mean this can be considered no more than an interesting observation.

Figure 5.5 and 5.11 (pg 227) display genus level abundance heatmaps for the samples as derived from classification via Mothur and Qiime respectively. While clustering algorithms for the 2 platforms differ (uclust or average neighbour), the Greengenes core alignment and an OTU cutoff value of 0.03 (97% similarity) were applied in both instances, thus minimising substantive divergence in eventual assignment of a read to an OTU. Inclusion of a taxon in the heatmap was dependent upon a total incidence of more than 50, or proportional abundance in at least 1 sample of greater than 1%.

The Mothur-derived heatmap clearly displays the loss of diversity in the CDAD samples in comparison to both the control and AAD groups, occurrence in less than half of the taxa being detected for this cohort. This is particularly evident with regard to the *Ruminococcaceae* and *Lachnospiraceae* families of the class *Clostridium* (phylum *Firmicutes*), which are all but absent in the CDAD set, while several genera (e.g.

Alistipes) of the *Bacteroidetes* phylum are also undetected in this group. In contrast, the CDAD group displays a relative prevalence of various genera, such as *Klebsiella* of the *Enterobacteriaceae* family (phylum *Proteobacteria*) and *Leuconostoc* and *Lactobacillus* of the *Bacillaceae* family (phylum *Firmicutes*).

Significance tests for the Mothur-derived OTU table were performed using Metastats (White *et al*., 2009) with p and q value thresholds of 0.05. Use of the q value represents a more stringent approach incorporating an adjustment for false discovery rate (FDR; Benjamini and Hochberg, 1995), without being as conservative as the Bonferroni correction (Bender and Lange, 1999; Narum, 2006). Verification of statistical significance observed at the class level for *Gammaproteobacteria* and *Clostridia* provided support for the validity of this correction. OTUs not displayed in the heatmap were also included in the significance testing.

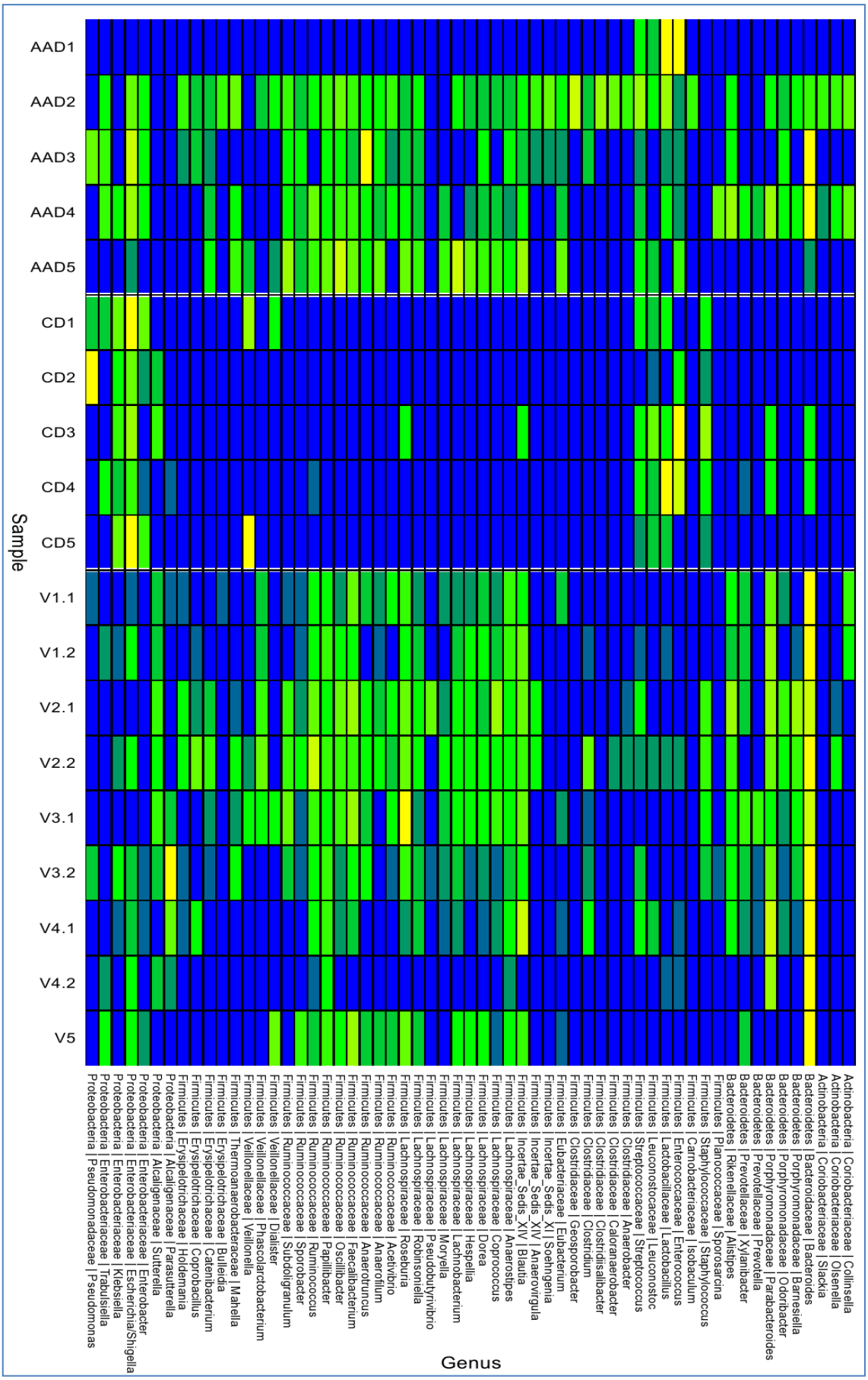
Metastats highlighted 3 OTUs as being differentially abundant (at $\alpha = 0.05$) between the groups: OTU 47 (*Sporobacter*; $q < 0.0001$) for comparison of the CDAD and AAD sets, and OTUs 1 (*Bacteroides*; $q = 0.01$) and 16 (*Papillibacter*; $q = 0.01$) for the test between the CDAD and control groups. The low number of OTUs identified as showing statistically significant difference may be a function of the small sample number and corresponding diminished power of the tests. Other OTUs with a p value of less than 0.05 but a q value greater than the threshold for rejection of the null hypothesis were: 12, 20, 27, 81, and 82, representing *Faecalibacterium*, *Anaerostipes*, *Klebsiella*, *Acetanaerobacterium*, and *Mahella*.

In summary, the CDAD group displays reduced numbers of *Firmicutes* taxa although *Bacillus* were more prevalent in both the CDAD and the AAD sets when compared to the normal group. The diminished representation of the *Bacteroidetes* phylum in the CDAD samples was also apparent, accompanied by a bloom in taxa of the phylum *Proteobacteria*. The AAD samples appeared to represent something of a midpoint

between the control and CDAD groups. Apart from samples AAD1 and V4.2 there is a consistent and discernible fingerprint for a group; this is in agreement with conservation of the microbiotic composition at higher phylogenetic levels while inter-individual variation is expressed at the level of genus and species (Jalanka-Tuovinen *et al*., 2011).

Figure 5.5 displays a genus-level heatmap, produced in 'R' from the Mothur OTU table with logit-transformed abundance. Blue denotes low abundance, while yellow indicates high abundance and green is intermediate. For inclusion, an OTU (shown along the x-axis) needed to represent at least 5% of the total abundance count for at least one sample. Predominant phyla are *Firmicutes*, *Bacteroidetes* and *Gammaproteobacteria*, while the most commonly encountered families are *Clostridiaceae*, *Ruminococcaceae*, *Lachnospiraceae*, and *Enterobacteriaceae*.

Figure 5.5 Logit-transformed genus abundance heatmap derived from Mothur



5.6 Alpha Diversity

Alpha diversity indices for a community can often be partitioned into the components of richness (number of different taxa) and evenness (distribution of individuals amongst the taxa). A 'diverse' value for the index may represent a disproportionate contribution by richness with relatively large numbers of individuals clustering into few of the taxa, or a smaller number of species with no evident dominance (Purvis and Hector, 2000). For this reason, comparison of communities based on values of diversity indices should be approached with caution.

However, the 'fingerprint' provided by the heatmap shown in figure 5.5 indicated that diversity values for the samples would be of interest, potential bias introduced by choice of index minimised by the simple expedient of using multiple indices. The RDP itself provides for calculation of the Chao1 index of diversity (*inter alia*), such that rarefaction curves for a selection of samples could be constructed as in figure 5.6.

Figure 5.6 Selected rarefaction curves at 0.03 cutoff for Chao1 diversity values

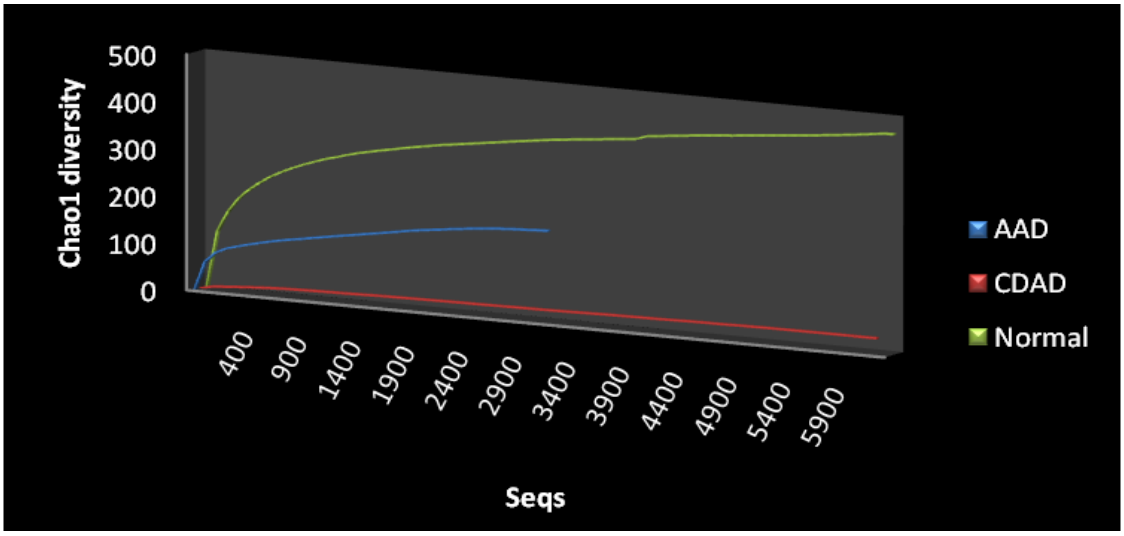


Figure 5.6 shows Chao1 (Chao,1984) rarefaction curves for AAD2, CD2 and V1.2 as examples of their group. The Chao1 index value is shown against the number of sequences used for the rarefaction calculation, 100 iterations being performed at each subsampling level of 100 sequence increments from 0-6000, to arrive at the average value for the index.

While only a selection are shown, all curves displayed a similar pattern in that diversity of the CDAD samples was minimal; expectation that new OTUs would be encountered with further sampling was low. In contrast, Chao1 values for control samples were higher and rarefaction curves had not reached asymptote, even at a sample depth of 5000 reads. Values for the AAD samples were found to be in the mid-range between CDAD and the control group.

A selection of further diversity values were calculated in both Mothur and Qiime and are shown in table 5.1. The table shows the Berger Parker dominance index (evenness), the Fisher Alpha index (diversity), the Shannon index (entropy, or likelihood of the next individual encountered/sampled being from a different OTU or species), and the Simpson index (diversity). All diversity indices display the same trend: values are low for the CDAD group, intermediate for AAD and highest for the control set; the Berger Parker dominance index values are highest for the CDAD samples suggesting that they are 'uneven' with a relatively high incidence of individuals in few clusters. The last 3 rows of the table display the P-values for significance tests on the means of group values. All differences between CDAD and AAD (or control) are considered significant at $\alpha = 0.05$, while none of the means of the AAD and control groups is considered statistically variant.

It is thus clear that diversity of the intestinal microbiota in the CDAD cohort is severely attenuated, even in comparison to those subjects within the AAD group, loss of richness being particularly evident amongst the *Ruminococcaceae*, the *Lachnospiraceae* and the *Bacteroideaceae*. However, it is uncertain whether this is a contributory factor in the pathogenesis of CDAD, or one of the sequelae of hyperinflammatory processes initiated by release of toxins A and B by *Clostridium difficile* subsequent to colonisation.

Table 5.1 Diversity indices, and P-values for pairwise comparisons of the means of indices for AAD, CDAD, and control groups

	Berger Parker	Fisher Alpha	Shannon	Simpson
AAD1	0.496	2.211915	1.905826	0.637352
AAD2	0.168	29.92206	5.079738	0.943256
AAD3	0.432	10.23186	3.098584	0.770456
AAD4	0.18	37.58835	5.228593	0.942544
AAD5	0.202	22.60752	4.726247	0.926056
CD1	0.844	2.44022	1.022878	0.2818
CD2	0.934	0.593724	0.385333	0.124024
CD3	NA	NA	NA	NA
CD4	0.778	3.40203	1.263102	0.375224
CD5	0.486	2.44022	1.566879	0.58316
V1.1	0.248	18.64221	3.91401	0.8644
V1.2	0.654	13.08183	2.624832	0.5668
V2.1	0.132	33.10504	5.286171	0.954712
V2.2	0.248	35.87444	4.637424	0.889808
V3.1	0.24	32.02724	4.715543	0.899552
V3.2	0.452	9.892358	2.592768	0.697704
V4.1	0.286	9.556602	3.328385	0.839016
V4.2	0.688	5.538433	1.955701	0.512736
V5	0.248	14.59592	3.863464	0.863664
CDAD v AAD	0.005176	0.040665	0.00635	0.002444
CDAD v Control	0.005582	0.01545	0.00113	0.000994
AAD v Control	0.57397	0.848361	0.621734	0.515049

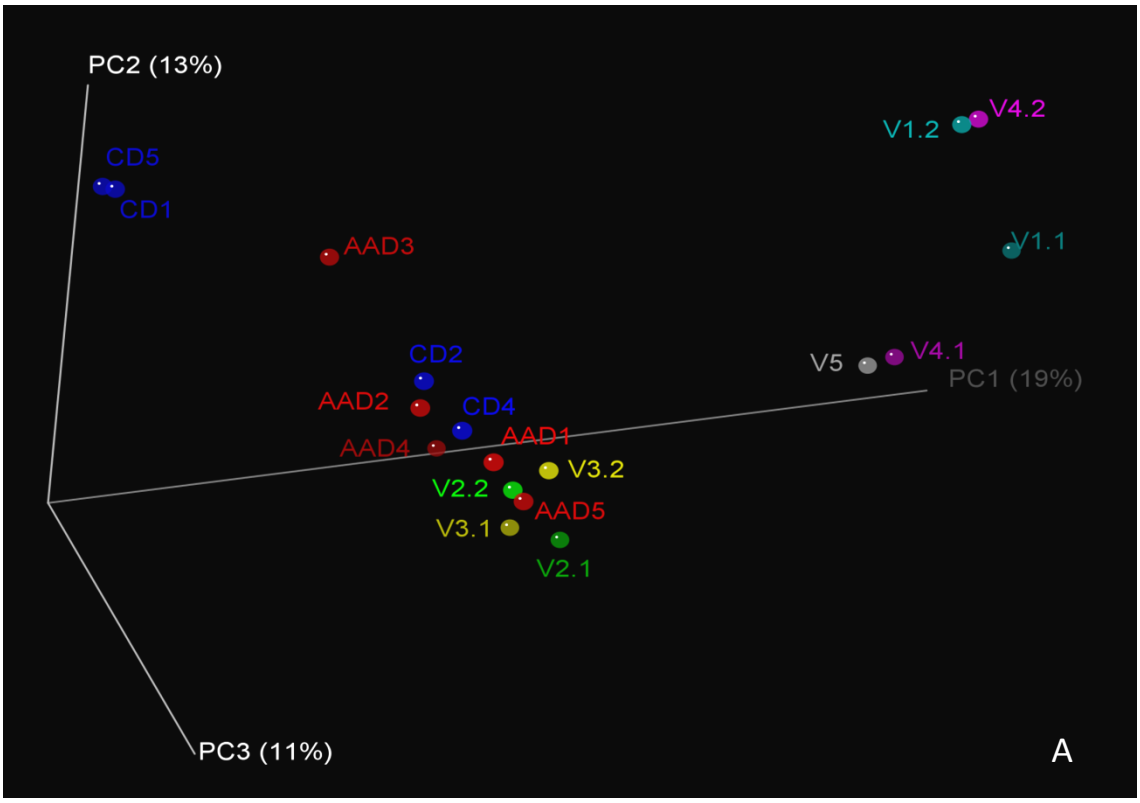
5.7 Beta Diversity

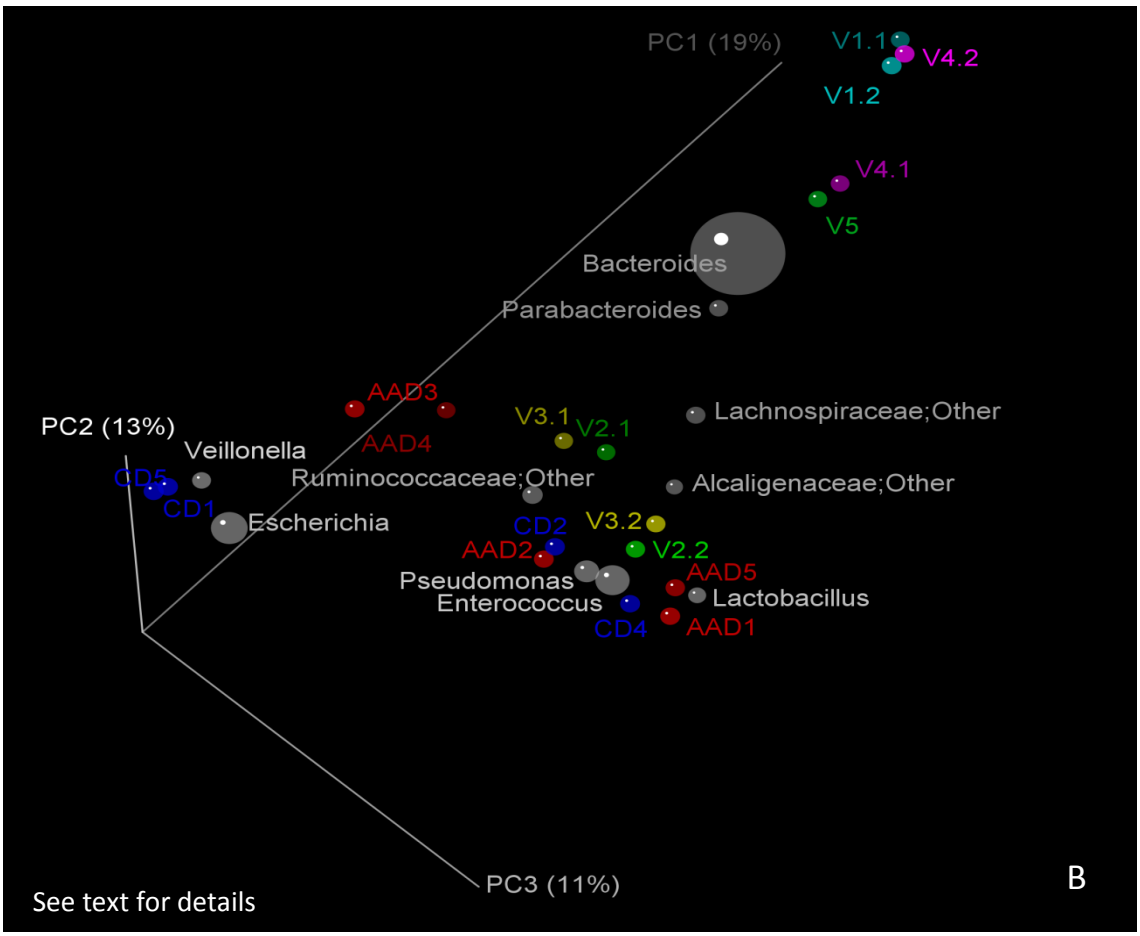
The measure of similarity of communities in terms of shared taxonomic groups (membership), and the relative abundances within those groups (structure), is termed the beta diversity (Schloss and Handelsman, 2006c). Similarity coefficients assessing the former are often termed binary since the equations for evaluation require only presence or absence data (i.e. 1 or 0).

Prior to evaluation of similarity coefficients for the CDAD, AAD and control samples, the OTU table created subsequent to clustering, alignment and classification was subjected to rarefaction at all levels from 200 to 6000 reads in 200 sequence increments. This resulted in exclusion of CD3 above the level of 400 reads and CD1 could only be retained by accepting 1600 sequences as the final rarefaction standard. The final rarefied OTU table was then utilised to evaluate a selection of both binary (Euclidean, Sorensen, Jaccard, and unweighted Unifrac) and membership (Bray-Curtis, Morisita-Horn and weighted Unifrac) similarity coefficients. Preliminary analysis of PCoA and biplots for the various measures suggested minimal bias on the basis of the coefficient selected, so subsequent figures are derived from the Morisita-Horn values. The Morisita-Horn index was chosen for reasons outlined previously, particularly that it is considered robust with respect to differing sample sizes (Wolda, 1981; Magurran, 1988; Faith *et al*., 1987; Peura *et al*., 2012).

Figure 5.7 displays the PCoA biplot for CDAD samples (blue) except CD 3, AAD samples (red), and normal samples (aqua, green,yellow, lilac and grey for V1,V2, V3, V4 and V5 respectively) based on Morisita-Horn coefficients. In transforming the coefficients into principal co-ordinates the contribution of taxa to the eigenvalues is evaluated and can be overlaid onto the primary plot, as shown in 5.7B. For the purposes of clarity only the primary 10 are displayed, but any number could theoretically be incorporated. The three axes shown describe 43% of variation between samples.

Figure 5.7 PCoA biplot derived from Morisita Horn coefficients showing samples and primary contributory taxa





The separation of samples is dominated by PC1, the primary contributory taxa being *Bacteroides* and *Parabacteroides*. Absence, or relatively low incidence, of these 2 genera causes clustering of the majority of samples distant from V1, V4 and V5. CD1 and CD5 cluster closely together at the other extreme of the PC1 axis due to the lack of *Bacteroidetes*. The prevalence of *Escherichia* in CD1 and CD5 explains their position relative to the PC2 axis, although other *Proteobacteria* and the genus *Veillonella* of the *Firmicutes* phylum also contribute to this coordinate. The clustering of the remainder of the samples is along an axis created by PC3. The relative abundance of genera such as *Lactobacillus* and *Enterococcus* (both of the *Firmicutes* phylum), along with other Gram-positive cocci such as bacteria of the genus *Stapylococcus*, determines the relative positions of the samples in this plane.

While clustering of the samples is much as expected the plot suggests a closer relationship between CD2/CD4 and the V2/V3 samples than would be envisaged from the heatmaps (Figures 5.5 and 5.12), particularly in view of the comparative prevalence of *Lachnospiraceae*, *Ruminococacceae* and other *Clostridiales*. It appears that this is primarily due to the logit transformation applied to the heatmaps, which accentuates differences at low frequencies/abundances. While this is visually desirable the logit transformation is most suitable for application to binomial population distributions; microbiotic communities are not easily described by any one distribution but the lognormal approximation is considered viable if sampling is sufficiently extensive (Hughes, 1986; Dunbar *et al*., 2002). In addition, the heatmaps are derived from proportional abundance based on all sequences in a sample; the PCoA biplots, in contrast, are based on rarefied (sub-sampled) OTU tables. Thus the OTU abundance per sample may not precisely match the value calculated using the rudimentary proportional approach.

Figure 5.8 PCoA biplots for Euclidean and binary Euclidean coefficients

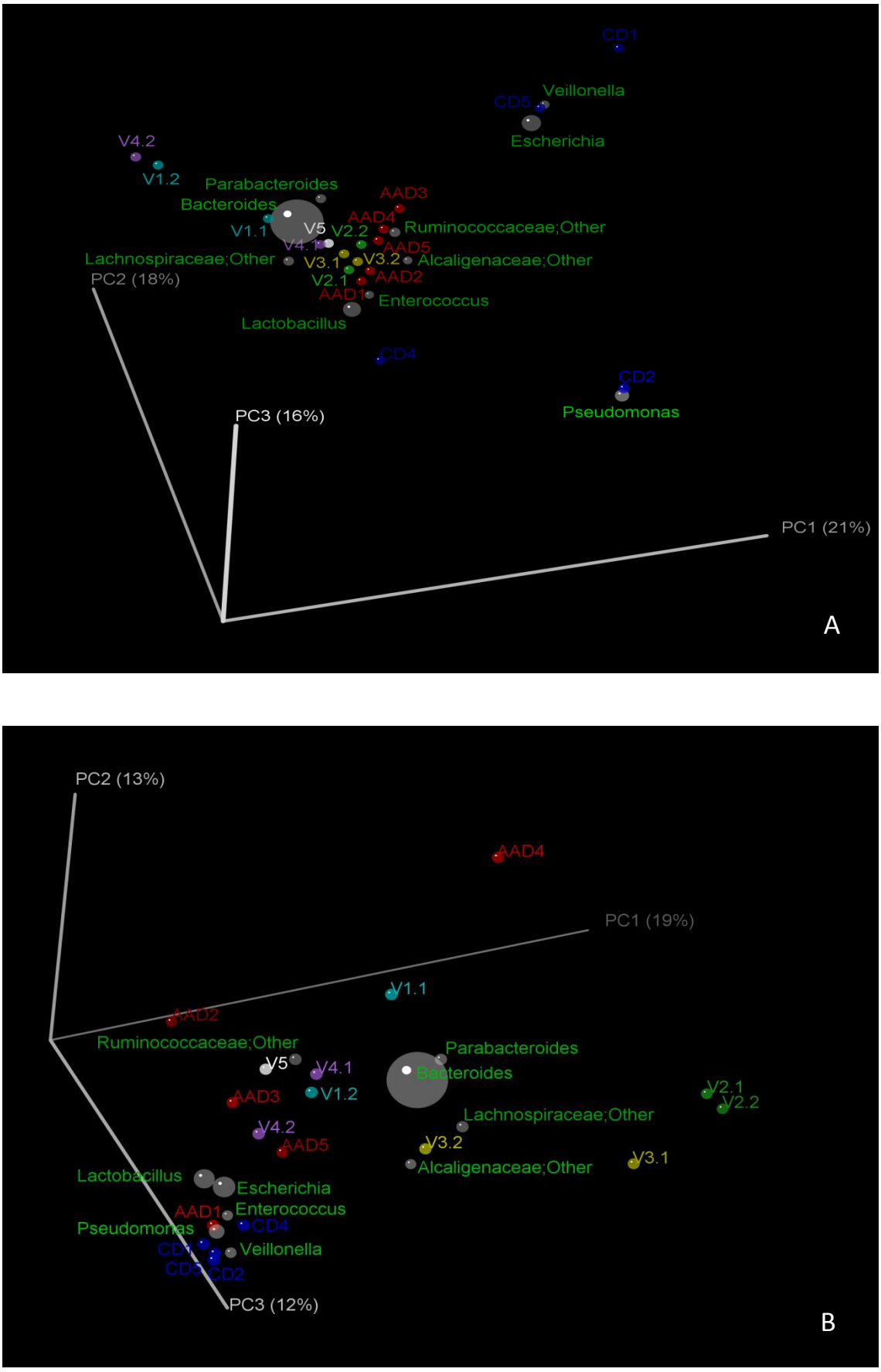


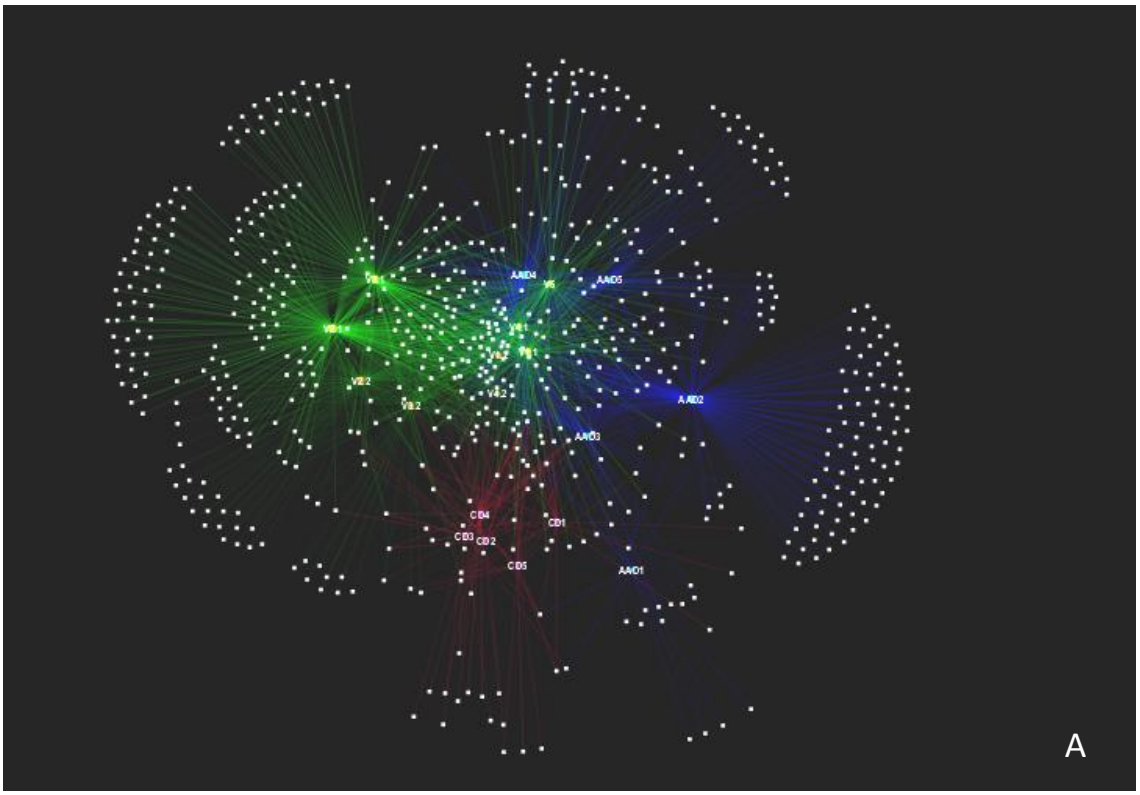
Figure 5.8 shows Euclidean (A) and binary (B) Euclidean distance PCoA biplots for CDAD (blue), AAD (red) and control (aqua, green, yellow, lilac and white) samples. Taxa are overlaid in grey with green labels. Size of spheres represents abundance within taxa.

Figure 5.8 displays the PCoA biplots created through application of different beta-diversity metrics to the dataset: the standard Euclidean measure of distance and its binary counterpart. These have been shown to be less effective in illuminating clustering patterns in multivariate data if depth of sampling is low and associative factors are inconspicuous (Kuczynski *et al*., 2010**).** However, the prominence of differences observed in the OTU heatmaps suggested that clustering should be easily discerned, and there appeared to be some intrinsic value in application of a metric with both binary and membership-related derivations.

The Euclidean biplots show custering of the four CDAD samples away from the other samples, apart from AAD1 whose heatmap fingerprint more closely resembles that for the CD group. Dominant contibutory genera with regard to the partitioning are *Pseudomonas*, *Veillonella*, *Escherichia*, and, to a lesser extent, *Lactobacillus* and *Enterococcus*. It is noteworthy that the 10 taxa identified as explaining the majority of the variation are consistent despite the distance metric utilised for derivation of coordinates for the respective PCoA plots. This indicates that the transformations of the metric values to 3D space, as opposed to the metrics themselves, are responsible for the variability of clustering patterns.

Figure 5.9 displays the Cytoscape visualisation of the relationship between the respective samples. While not strictly displaying beta diversity, the networks are produced through implementations of algorithms which provide 'weighting' of connections based on abundance. Each 'edge' connects a sample node and an OTU node in which that sample is represented, the 'force' inherent in the edge being proportional to the abundance. The force-directed (A) and spring-embedded (B) algorithms displayed in the figure cause superficial differences in the layout but clustering of samples within groups is clearly visible. The AAD samples generally position closer to the control than the CDAD samples, AAD1 again being an expected exception to this observation.

FIGURE 5.9 Cytoscape network relationship of samples and OTUs



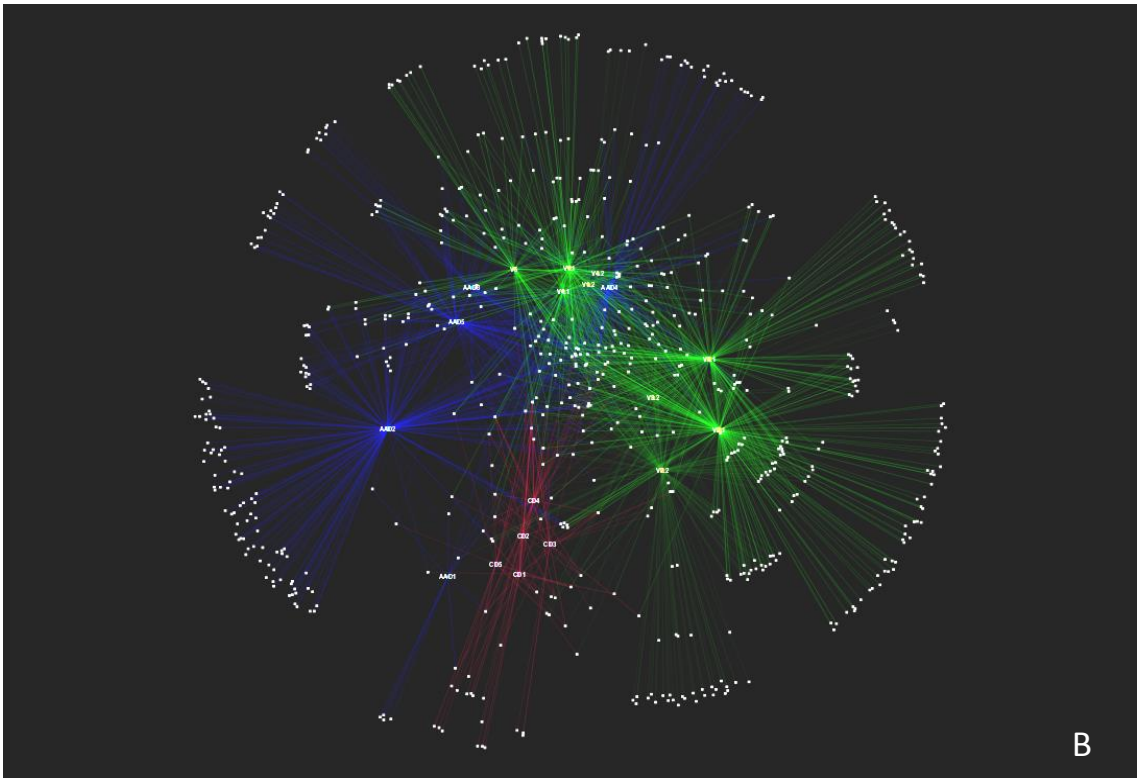


Figure 5.9 shows networks produced by Cytoscape (Shannon et al., 2003) through implementation of a force-directed edge-weighted algorithm (A; Biolayout; Enright and Ouzounis, 2001) or a spring-embedded edge-weighted algorithm (B; Kamada and Kawai, 1988). Magenta nodes and red edges represent CDAD connections, aqua nodes and blue edges are for AAD samples and yellow nodes with green edges denote control samples. OTUs are white and predominantly occupy the periphery of the figure.

Figure 5.10 displays the distance boxplots created for the various groups based on weighted UniFrac values. The UniFrac beta-diversity coefficient can be considered as representing a third group of similarity measures based on phylogenetic distances, and requiring a tree in addition to the OTU table for calculation. Weighted UniFrac incorporates abundances into the calculation as compared to the unweighted or binary version. Distances between the normal volunteers and the CDAD group are found to be maximal while those between the AAD group and CDAD are intermediate. Values for comparisons between pre- and post-antibiotic samples support the decision to merge these two sets into one 'normal' control group.

Figure 5.10 Distance boxplots calculated using weighted UniFrac values

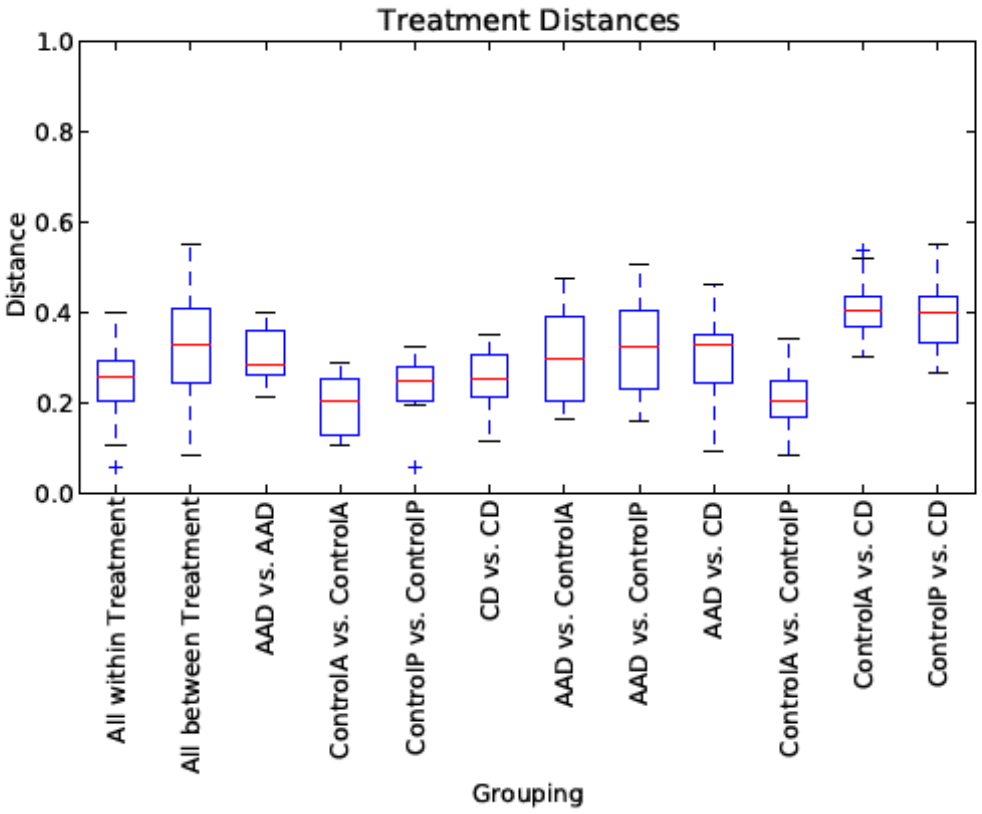


Figure shows distance boxplots for the 'treatment' goups based on weighted UniFrac values. CD is the CDAD group, AAD represents the AAD group, while ControlA represents the normal vounteers prior to antibiotic treatment and Control P represents the volunteer group subsequent to antibiotic administration. 'Within treatment' are the distances calculated between all samples in the same group.

5.8 Statistics

Statistical comparisons to identify OTUs which partition with particular groups were conducted previously on OTU tables created from Mothur or RDP classifications. The Metastats application was utilised for these significance tests and results were presented in section 5.5.

The QIIME platform also incorporates its own statistical tests (ANOVA and G-Test of independence) for identification of OTUs associating with a given set of samples.

Table 5.2 ANOVA test for differential abundance within OTUs across groups

OTU	Genus	CDAD	AAD	Control	P-value	FDR-corrected
21	Bacteroides	Absent	Detected	Detected	0.010783968	0.506846488
1372	Faecalibacterium	Absent	Detected	Detected	0.036166481	0.8499123
428	Bacteroides	Absent	Detected	Detected	0.039704801	0.622041884

Table 5.2 displays P-values and FDR-corrected (q=0.05) values for ANOVA and table 5.3 displays the corresponding values for application of the G-test. Both the OTU number (as designated by the QIIME uclust algorithm) and the associated classification are shown. The presence or absence of an OTU in a set of samples is also noted: 'Present' or 'Detected' indicate identification in all samples of a group, 'Absent' indicates that the OTU was not encountered in a group, while 'Low' correlates with incidence in less than the full complement of group samples. All OTUs which provided a P-value of less than 0.05 are shown, although none of these values remained below the 95% significance threshold once FDR correction had been applied, a corollary of the small number of samples eventually included in the investigation. The prevalence of *Bacteroides* spp. amongst OTUs identified by both the G-test and ANOVA is clearly of

interest, albeit not a significant finding subsequent to FDR correction. The observation that abundance of at least one species of the genus *Faecalibacterium* shows a degree of association with category could also be worthy of more expansive investigation in future.

Table 5.3 G-Test of independence for correlation of presence or absence of OTUs

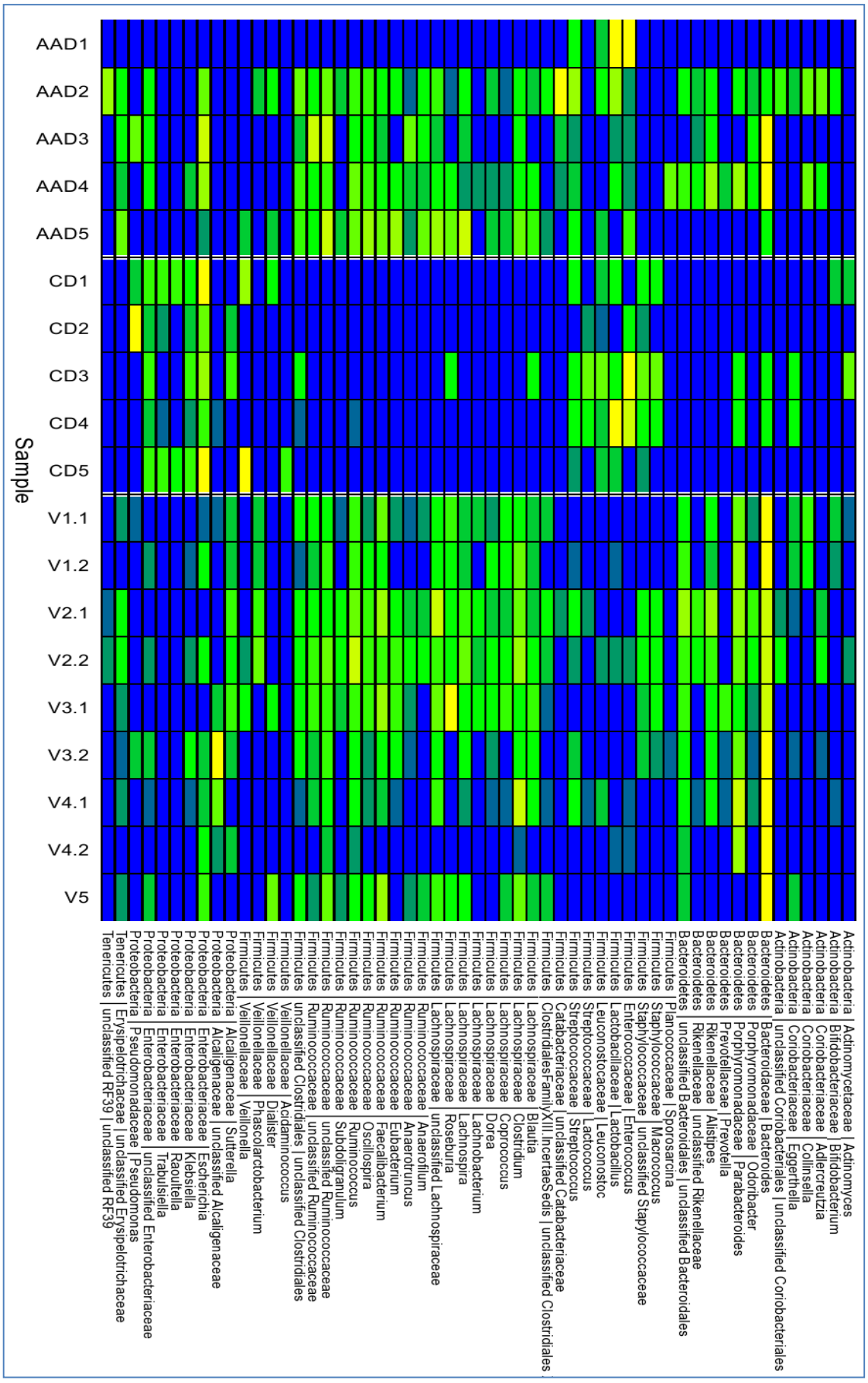
OTU	P-value	FDR_corrected	CDAD	AAD	Control	Lineage
1300	0.00619	1.771	Absent	Low	Present	Bacteroides
330	0.01541	2.202	Absent	Present	Present	Bacteroides
442	0.0294	2.811	Absent	Low	Present	Ruminococcaceae
944	0.0401	2.867	Absent	Low	Present	Bacteroides
757	0.0449	2.571	Absent	Present	Present	Oscillospira
448	0.0478	1.954	Absent	Low	Present	Bacteroides
942	0.0478	2.2801	Absent	Low	Present	Bacteroides

To test for differences between the structure of samples as a whole, the Libshuff statistic (Mothur; Schloss *et al*., 2004) and Martin's P-test for significance (QIIME; Martin, 2002) can be applied. These provide no information on how communities might be dissimilar, simply whether they are different at a given level of significance. Multiple pairwise testing in QIIME with corresponding FDR-correction indicated that the majority of communities were 'different', although the null hypothesis was not rejected ($P > 0.05$) for comparisons between CDAD samples except CD1 and CD4, indicating a greater within-group similarity of these samples. The Libshuff statistic (Singleton et al., 2001) for estimation of statistical difference of microbiotic communities was applied

only for between group comparisons, only CDAD against control providing a significant p-value of 0.0217.

Figure 5.11 displays a genus-level heatmap, produced in 'R' from the Qiime OTU table with logit-transformed abundance. Blue denotes low abundance, while yellow indicates high abundance and green is intermediate. For inclusion, an OTU (shown along the x-axis) needed to represent at least 5% of the total abundance count for at least one sample. Predominant phyla are *Firmicutes*, *Bacteroidetes* and *Gammaproteobacteria*, while the most commonly encountered families are *Clostridiaceae*, *Ruminococcaceae*, *Lachnospiraceae,* and *Enterobacteriaceae*.

Figure 5.11 Logit-transformed genus abundance heatmap derived from Qiime



5.9 Summary

On the whole, the investigation into intestinal microbiotic differences between control subjects and the AAD or CDAD cohorts was disappointing, primarily due to inadequate sampling. The number of subjects in each group did not meet initial expectations, while the depth of sequencing for each representative was also disappointing. As a result, potential patterns in OTU association could be clouded by inter-individual variation, while the power of statistical tests is diminished.

Despite this failing, certain aspects of a differential and distinctive microbiota in patients with CDAD have become evident. Diversity of the microbiota is significantly reduced in comparison with both AAD patients and 'normal' controls, the spectrum of genera in the *Bacteroidetes* and *Firmicutes* phyla displaying the primary impact of the amelioration. Total abundance in the *Bacteroidetes* phylum in CDAD subjects is significantly decreased in comparison to both AAD and control groups. Total incidence in the phylum *Firmicutes* is partially maintained as the loss of *Lachnospiraceae* and *Ruminococcaceae* is offset by a rise in numbers of *Bacillus*, but the CDAD subjects display a significant bloom in the phylum *Protoeobacteria* in comparison to the control (but not AAD) group. It is possible that this is due to a fall in total numbers of bacteria in the colon such that Proteobacteria merely represent a greater proportion of an obliterated microbiota; indeed, certain antibiotics and sequelae of CDAD may lead to this population dynamic (SulliVan *et al*., 2001). However, a real overgrowth of facultative species, concomitant with reduced prevalence of the other phyla while overall numbers remain comparatively stable, has been experimentally shown in at least one study (Hopkins and Macfarlane, 2002).

The difference between the CDAD samples as a group and AAD or control samples is also evident from PCoA analysis. Biplots of the data, derived from application of relativised Manhattan metrics, display that genera such as *Escherichia*, *Pseudomonas*,

and *Lactobacillus* drive the CDAD samples to cluster together, distant from other samples.

However, implementation of significance tests such as ANOVA do not confirm these findings statistically. Abundance of the *Bacteroides* and *Faecalibacterium* genera show a degree of correspondence with the categories, prevalence being diminished in the CDAD group, but this apparent relationship cannot be asserted with any confidence due to the reduced power of the tests. Had the power been sufficient, though, conclusions would still have to be tempered by the caveat that any correlation is likely to be associative rather than causative.

CHAPTER 6

DISCUSSION

6.1 General discussion

In terms of results from the experiments conducted, the research has shed little fresh light on the pathogenesis of CDAD with regard to the intestinal microbiota and colonisation resistance. At higher levels of phylogenetic classification (e.g. phylum), results are in broad concordance with previous studies of the healthy intestine, indicating dominance by *Bacteroidetes* and *Firmicutes* in terms of both abundance and diversity (Eckburg *et al*., 2005; Andersson *et al*., 2008; Qin *et al*., 2010); at the family level, *Bacteroidaceae*, *Lachnospiraceae*, and *Ruminococcaceae* are some of the most prevalent (Nava and Stappenbeck, 2011; Mariat *et al*., 2009), contributing, *inter alia*, to complex nutrient metabolism (Turnbaugh *et al*., 2006; Louis *et al*., 2007; Sonnenburg *et al*., 2010; Lopez-Siles *et al*., 2011), and immune response regulation (Rakoff-Nahoum et al, 2004; O'Hara and Shanahan, 2006; Yoon and Sun, 2011). With greater taxonomic resolution, however, inter-individual variation becomes more apparent (Dethlefsen *et al*.,2008; Qin *et al*., 2010), so, although genera such as *Faecalibacterium* and *Bacteroides* are often encountered, precise species composition and relative abundances may vary considerably (Claesson *et al*., 2011; Lozupone *et al*., 2012). Aside from species identification, which was not achieved in this study, similar patterns were observed in the current experimental data.

Investigations into the microbiota of individuals with AAD or CDAD are not as numerous, but the disruptive impact of antibiotics on the intestinal microbiota is well-established (SulliVan *et al*., 2001; Young and Schmidt, 2004; Jernberg *et al*., 2007; Dethlefsen *et al*., 2008). Antibiotics will clearly reduce numbers of their target organisms so, for instance, Gram-positive anaerobes are seen to be significantly diminished by cephalosporins (Rafii *et al*., 2008). However, concomitant blooms in genera such as *Escherichia* of the phylum *Proteobacteria* (Bartosch *et al*., 2004; Manichanh *et al*., 2010; Young and Schmidt, 2004), and the genus *Enterococcus* of the

phylum *Firmicutes* (Jakobsson *et al*., 2010) have also been observed with both patterns being evident in the current dataset.

The CDAD samples from the final phase of the study (section 5) were seen to have distinct fingerprints from both the AAD group and the control group. This was evident from the OTU tables and heatmaps which displayed the general reduction in diversity and numbers of *Bacteroidetes* and *Firmicutes*, with a relative bloom of *Proteobacteria* and particular *Firmicutes* such as *Lactobacillaceae* and *Enterococcaceae*. Rarefaction curves also evinced the lower diversity of the CDAD microbiota, while network figures and PCoA biplots highlighted the clustering of the CDAD samples.

Variability in numbers of the genus *Veillonella* may be of particular interest (section 4.3.9; section 5.7, figures 5.7 and 5.8), although differential abundances between the CDAD and AAD groups may be caused by contrasting antibiotic administration, *Veillonella* spp. being susceptible to penicillin and cephalosporins, but resistant to aminoglycosides (Bhatti and Frank, 2000); this underlines the need for collation of comprehensive metadata files to accompany samples being investigated.

In broad terms the results presented here correlate with those of previous studies into CDAD. Overall diversity was found to be reduced (Hopkins *et al*., 2001; Hopkins and Macfarlane, 2002) and reduction may be associated with severity of the condition (Chang *et al*., 2008), although one study found that hospitalisation itself causes shifts in the microbiota, and that the CDAD microbiota was no less diverse than that of other patients (Manges *et al*., 2010). The numbers of OTUs corresponding to *Enterobacteriaceae* and other facultative bacteria are raised (Hopkins *et al*., 2001; Bartosch *et al*., 2004), while prevalence of *Bacteroidetes*, *Firmicutes* and *Actinobacteria* (especially the *Bifidobacteriaceae*) is diminished, although abundance of the genus *Enterococcus* is often found to be increased (Hopkins *et al*., 2001; Hopkins and Macfarlane, 2002; Bartosch *et al*., 2004). In addition, incidence of

Faecalibacterium prausnitzii is reduced (Bartosch *et al*., 2004; Rea *et al*., 2011) and it is interesting that this commensal has been shown to have anti-inflammatory properties (Sokol *et al*., 2008). There are, however, certain discordances in the data e.g. one of the studies found that numbers of *Enterococcaceae* were attenuated (Rea *et al*., 2012). Such contrasting results can be explained by the timepoint in the course of infection at which samples were obtained or differing methodologies. In addition, discordance in therapeutic regimes for the subjects may also have contributed, highlighting the difficulties in comparing datasets between microbiotic studies.

However, even if patterns are evident from numerous studies with increased numbers of subjects and depth of sequencing per sample, while correlations are found to be both robust and significant, conclusions about the contribution of the microbiota to the pathogenesis of CDAD would remain tenuous without a defined mechanism.

The multifactorial nature of CDAD has been illuminated by multiple studies, with inter-individual microbiotic variation, polymorphisms in the IL-8 gene (Jiang *et al*., 2007; Garey *et al*., 2010), serum levels of anti-toxin IgG (Kyne et al., 2001) and even up-regulation of *Clostridium difficile* colonisation factor expression by certain antibiotics (Denève *et al*., 2008), contributing to infection outcome and severity of the symptoms. Differences in microbiotic composition may also be consequent to toxin-induced inflammation rather than causative (Walker *et al*., 2011), while the nature of asymptomatic carriage remains unexplained; this cohort probably represents the true control group for CDAD in that exposure to the pathogen has definitely occurred.

However the success of faecal bacteriotherapy (MacConnachie *et al*., 2009; Khoruts *et al*., 2010) in treating intractable recurrent CDAD cannot be ignored, and attests to the value of a broad range of therapeutic approaches. Probiotics which may contribute to restoration of the original commensal population, produce antimicrobial agents and lower the pH, increase expression of mucins, stabilise gut permeability and enhance

phagocytic activity have also been seen to be effective (Gerritsen *et al*., 2011). Both are rooted in direct manipulation of the microbiotic composition, so improved knowledge of the intestinal metagenome in normal, CDAD and asymptomatic groups may still prove to be of immense value, especially in the face of increasing resistance to antibiotics.

While the focus of the current investigation has been the bacterial component of the microbiota it is likely that the microbiome (and its contribution to the intestinal metabolome) will be more illuminating with regard to CDAD and other conditions associated with dysbiosis i.e the unique bacterial signature of an individual probably represents a commonality of potential metabolic pathways (Backhed *et al*., 2005; Ley *et al*., 2005). In this respect, the use of other technologies such as the Illumina systems (Caporaso *et al*., 2012), which are better suited to metagenomics, and perhaps even microfluidics devices allowing separation of individual bacterial cells for analysis (Leung *et al*., 2012) would complement the current approach, while comprehensive databases for allocation of potential metabolic roles to gene fragments are also available (KEGG; Ogata *et al*., 1999).

The contribution of the virome and the archaeal population of the intestine may also be significant and could be investigated in tandem. The archael population appears to be dominated by *Methanobrevibacter smithii* whose genome suggests a capacity to interact with a wide variety of bacterial species (Hansen *et al*., 2011), while the virome may comprise over 1000 phylotypes (Breitbart *et al*., 2003), and has been shown to consist of elements that are both highly specific to the human intestine (Ogilvie *et al*., 2012) and vary with diet and host genetics (Minot *et al*., 2011). It must be remembered that the microbiome consists of genes from more than just bacterial species, other organisms having considerable potential to influence the gene pool of the intestinal milieu and thus contribute to the pathogenesis of CDAD.

6.2 Design and analysis

6.2.1 Experimental design

In addition to some of the errors briefly mentioned in section 6.1 there were a number of avoidable flaws in experimental design which contributed to limitations on the inferences that could be drawn from the data.

Foremost amongst these was the selection of too few samples for analysis, both for pyrosequencing and arrays. In the latter case, clone libraries were not extensive enough to provide for empirical assessment of hybridisation dynamics. Although time and funds were also limiting factors in this respect, a range of percentage sequence identities (between clones and probes) would have been necessary to define the characteristics of duplex formation for each probe, such that interaction with 'unknown' sequences would lead to confident predictions about the sequence of the clones (Binder and Preibisch, 2005). In addition, the number of hybridisation experiments providing images of adequate quality for analysis was insufficient to appreciate this requirement soon enough for it to be met in a comprehensive manner.

For 454 sequencing the flaw was more fundamental, beginning with the selection of only 4 samples (2 CDAD and 2 AAD) for analysis in the first run (section 4.3.1) which prevented use of statistical procedures for analysis. Although this was partially rectified in subsequent experiments, even the final run for comparison of CDAD, AAD and controls suffered from a lack of samples and a correspondingly reduced power of statistical tests. In truth, power (1-β) can be increased by accepting a higher value of α (likelihood of Type I error, or incorrect rejection of a null hypothesis) where β is the likelihood of a type II error or incorrect acceptance of a null hypothesis (Zar, 1996). A value of 0.8 for 1-β is considered a statistically powerful test and this can be approached more easily if α increases; confidence in correctly rejecting the null hypothesis decreases but it is less likely that meaningful relationships will be ignored. If the

number of samples is limited, while knowledge of a system is relatively sparse and conclusions would be used to tentatively inform the direction of further work, then this approach is probably justified.

The problems encountered in obtaining sufficient samples mentioned previously (section 6.1) were also allowed to influence experimental design, such that investigations were less driven by the hypothesis to be addressed than specimen availabilty. Metadata on the samples collected was also deficient, particularly for the CDAD samples. More information about provenance, underlying conditions, clinical status, and associated biochemical tests such as serum anti-toxin A IgG levels (Kyne *et al*., 2000; Babcock *et al*., 2006), would have permitted a more thorough analysis. In addition to greater numbers of samples and technical replicates, effective accumulation of metadata will enhance the potential of future studies (Knight *et al*., 2012)

6.2.2 Analytical issues

Potential for analysis of the datasets improved over the course of the research. At the outset, the RDP (Maidak *et al*., 1997) and other databases/workbenches such as Greengenes (DeSantis *et al*., 2006a) and ARB-Silva (Pruesse *et al*., 2007), combined with applications such as BLAST (Altman *et al*., 1990), were comprehensive resources available for analysis of SSU rRNA data, but none were suitable for the number of sequences provided by 454 sequencing. The ARB software environment (Ludwig *et al*., 2004) was also incapable of handling the short sequence reads provided by the 454 platform and is, in any case, no longer supported. The addition of dedicated SSU rRNA platforms such as Mothur (Schloss *et al*., 2009), Qiime (Caporaso *et al*., 2010), and METAGENAssist (Arndt *et al*., 2012) to the 16S bioinformatics armoury has facilitated the analysis of complex 454 datasets. Coupled with the growth of cloud computing to provide the processing capability (Stein, 2010), the potential to address and answer

questions about the microbiotic composition of any niche seems to increase almost daily.

However, for each of the platforms substantial periods of time and viable datasets are required if competence is to be attained, while upgrades to versions can cause disruptions to analyses. For instance, in the midst of Mothur analysis for the third 454 run (section 4.5) it was revealed that certain versions prior to V1.19.0 included significant flaws in chimera identification algorithms. While not strictly disruptive, it was also disappointing that added functionality in QIIME, including incorporation of hypothesis-testing procedures such as ANOSIM and ADONIS, came subsequent to final analysis of datasets. Portability of data between platforms along with naming conventions and/or OTU definitions could also be problematic, particularly when using a command-line application such as Mothur.

In essence, then, while bioinformatics pipelines have improved significantly, their functionality has perhaps not yet been fully exploited, and convergence through improved coding ability or utilisation of workbenches such as Galaxy (Giardine *et al*., 2005) or Taverna (Hull *et al*., 2006) would be of considerable utility. In addition, incorporation of commercial software such as the Community Analysis Package (CAP; Henderson and Seaby, 2007, Pisces Conservation Ltd, Lymington, UK.) into the pipeline may be of benefit to allow for application of a greater range of ordination techniques and multivariate statistical analyses (Ramette, 2007).

6.3 Potential sources of bias

Aside from the ubiquitous 'operator error' there are numerous well-established sources of bias in PCR-based studies of microbial populations (Forney *et al*., 2004). Consistency of methodological approach limits these biases when evaluating Δ*MS* (the change in microbiotic structure) due to different categories or treatments, but their impact influences the assessment of 'absolute' composition.

6.3.1 DNA extraction

Extraction of DNA from a sample is one of the potential sources of technical bias in community analysis. This is first evident in terms of the sampling site or material collected, where differences between mucosal and luminal populations of the same intestinal compartment have been observed (Zoetendal *et al*., 2002; Eckburg *et al*., 2005; Zoetendal *et al*., 2008; Walker *et al*., 2011) although microbial community profiles along the course of an individual's intestinal tract correlate much more closely than those from the equivalent site in differing individuals (Maukonen *et al*., 2008; Bogert *et al*., 2011). While considered temporally stable (Claesson *et al*., 2010) certain shifts in components of the microbiota over relatively short time periods have been seen to occur (Vanhoutte *et al*., 2006), particularly in reponse to such factors as dietary modulation (Walker *et al*., 2010) which may not be monitored as part of a study. Storage of samples prior to extraction of community DNA may also impact on the determined profile although the contribution may have been overstated (Lauber *et al*., 2010; Wu *et al*., 2010). While all samples for the present study were faecal (caecal for chickens), the *ad hoc* nature of specimen acquisition meant that storage conditions prior to collection could not be monitored or standardised. It is thus possible that certain discrepancies within groups, particularly as noted in the first 454 run (section 4.3), were

attributable to variations in colection and storage conditions rather than characteristics of the samples themselves.

While it is important that pre-processing is of a standardised nature, it is also imperative that a high level of bacterial lysis is achieved and that genomic DNA is of a suitable quality for downstream analyses. In the past lysis was performed by cycles of freezing and thawing, with addition of enzymes such as lysozyme followed by bead-beating and subsequent phenol/chloroform extraction of DNA (Stahl *et al*., 1988). However, in the last few years kits have become available which enable processing of raw faecal material using incorporated buffers and columns. The lysis buffers consist of high-strength chaotropic guanidium salts and detergents, while wash solutions contain lower strength chaotropic compounds and Tris/alcohol/acid buffers for DNA elution (McOrist *et al*., 2002). Investigations into the efficacy of such kits have been conducted and the Qiagen was found to be the most effective and reproducible (Mei Li *et al*., 2003), possibly due to use of a proprietary resin to adsorb inhibitors of PCR such as bile salts (McOrist *et al*., 2002). However, alternative extraction techniques can lead to variable estimations of microbial diversity and the associated protocol includes a differential temperature step dependent on whether Gram-positives or Gram-negatives are considered most likely to be represented in the population: the thicker cell-wall of Gram-positives hinders lysis so a higher temperature is employed. The current investigation adopted this higher temperature as a standard element of the protocol to maximise retrieval of Gram-positive DNA, but there is a concomitant danger that Gram-negative DNA suffers excessive degradation and fragmentation (Wintzingerode *et al*., 1997). It is noteworthy that recent investigations have advised a return to inclusion of a beat-beating step prior to processing of material with a kit (Salonen *et al*., 2010; Yuan *et al*., 2012; Sergeant *et al*., 2012), and future studies will implement this recommendation.

6.3.2 Primers

Once DNA of sufficient quality has been obtained amplification through PCR can be performed. The output of previous investigations provided a catalogue of primers with the potential for annealing to conserved regions of the eubacterial 16S rDNA gene which have been designated 'universal' (Lane, 1991; Weisburg *et al*., 1991; Marchesi *et al*., 1998; Baker *et al*., 2003; Chakravorty *et al*., 2007). However, the universality of such primers is somewhat questionable, absolutely conserved regions of the rDNA gene normally extending only to consecutive tetranucleotides, while assessment of complementarity across the bacterial domain can only be calculated against identified species (Baker *et al*., 2003; Forney *et al*., 2004).

Empirical findings are that even small alterations in primer sequences can result in detection of significantly different microbial members (Schmalenberger *et al*., 2001), although an approximate 70% identity of the primers with regard to the template is considered adequate for successful annealment and amplicon production if PCR stringency is sub-optimal (Baker *et al*., 2003). In addition, implementation of successful PCRs is governed by other primer-associated factors such as formation of primer dimers, optimum annealing temperatures to prevent spurious amplification, and stringency of 3′ complementarity to ensure extension (Baker *et al*., 2003). Design of primers for microbial community analysis is thus constrained by the conflicting demands of representative coverage and PCR-specific considerations.

Many of the issues can be addressed through incorporation of degenerate nucleotides into the primers, although this introduces variability into the T_m of the duplex (Polz and Cavanaugh, 1998); to obviate this, deoxyinosine (I) may be utilised, since it is capable of forming hydrogen bonds with all four of the nucleotides found in DNA (Ben-Dov *et al*., 2006). However, preliminary tests of primers containing inosine were unsatisfactory, PCRs displaying numerous artifacts on AGE-visualisation.

Titration of magnesium concentration (Blanchard *et al*., 1993) and adjustment of annealment temperature (Ishii And Fukui, 2001) should allow primers containing degenerate nucleotides to form duplexes with the majority of community members. However, no primer can be expected to be 100% inclusive (Forney *et al*., 2004), and considerable variation in library composition can be introduced solely as a result of primer choice (Wintzingerode *et al*., 1997). It is for this reason that separate PCRs with differing primer pairs are recommended, the resultant products being quantified and mixed in equimolar ratios to conserve relative population frequencies (Hansen *et al*., 1998; Polz and Cavanaugh, 1998; Qiu *et al*., 2001). While this approach was adopted for certain clone libraries, the cost of multiple barcoded primers for 454, and the potential complexities introduced into analysis pipelines suggested it would be inadvisable for pyrosequencing, and it does not seem to have found widespread use for investigations where 454 is utilised.

In addition to 'universality' of primers, selection of the hypervariable regions of the 16S rDNA gene for an investigation may be of considerable importance. Differing variable regions of the sequence afford varying degrees of discriminatory power when it comes to analysis (Schmalenberger *et al*., 2001; Chakravorty *et al*., 2007; Wang *et al*., 2007; Wu *et al*., 2010), with current investigations tending to favour analysis of the V3 and V6 domains (Liu *et al*., 2008; Huse *et al*., 2008; Wang and Qian, 2009). Experiments detailed in sections 4.3- 4.4 found that apparent differences in microbiotic composition were introduced through choice of differing primer pairs, though this is probably attributable to variations in amplification (rather than classification) bias since the discrepancies are primarily quantitative rather than qualitative. However, since investigations were conducted on samples of unknown composition it is not possible to state the exact nature of the bias introduced by a given primer pair, just that inferences about 'absolute' composition of a microbiota should be made with caution.

6.3.3 PCR

In addition to the potential for introduction of bias through primer choice, PCR amplification is susceptible to further potential errors which can be grouped into 'drift' and 'selection' (Polz and Cavanaugh, 1998). Some of these are inherent to all PCRs, while others are specific to the template being amplified (Wintzingerode *et al*., 1997).

PCR drift is essentially stochastic but may be exacerbated by technical errors and can thus be minimised by performance of replicate reactions (Polz and Cavanaugh, 1998). High starting template concentrations have also been suggested as a means to reduce this source of bias (Polz and Cavanaugh, 1998; Wintzingerode *et al*., 1997), especially if relative proportions of community members are to be inferred (Chandler *et al*., 1997).

PCR selection is related to the interaction between template, primers and polymerase. The chosen polymerase should be of high-fidelity with a correspondingly high processivity (Wintzingerode *et al*., 1997; Speksnijder *et al*., 2001), although differences in secondary structure of templates will affect the ability of the enzyme to perform readthrough (Polz and Cavanaugh, 1998; Qiu *et al*., 2001). Such an influence is not confined to the 16S rDNA gene itself, genomic size and flanking regions having the potential to affect enzyme access to the template, especially if G+C content is high and rates of duplex dissociation would be reduced (Hansen *et al*., 1998; Wintzingerode *et al*., 1997; Suzuki and Giovannoni, 1996). A further genomic factor is the 16S rDNA copy number, known to vary between 1 and 14, and complicating correlation between clone number resultant from PCR and original community representation (Farrelly *et al*., 1995).

Additives to PCRs to reduce the bias caused by differing templates have been suggested and include acetamide, BSA or glycerol (Hansen *et al*., 1998; Reysenbach *et al*., 1992; Marchesi *et al*., 1998); these may also reduce detrimental effects on PCRs caused by contaminating materials carried over from the DNA extraction process

(Wintzingerode *et al*., 1997). Indeed, addition of BSA to PCR mixtures is recommended by Qiagen subsequent to DNA extraction using the QIAamp® DNA Stool Kit (QIAamp DNA Stool Handbook 04/2010) and was included as a matter of course. However, additives are not expected to affect the differential sensitivity of diverse species to amplification, estimates of the minimal cell numbers required for detection ranging from 2 for *Bacteroides* spp. to 1000 for the gram-positive *Eubacterium limosum* (Wang *et al*., 1996).

The most significant bias, though, is associated with cycle number of the amplification recation. While it is common to perform 30 cycles to achieve a satisfactory yield in PCRs, high cycle numbers with mixed templates increasingly lead to attenuation of starting template ratios (Suzuki and Giovannoni, 1996). This is thought to be a result of preferential reannealment of the duplex PCR products after each denaturation step, longer regions of homology associating with each other with greater efficiency than shorter regions such as those represented by 20-mer primers (Suzuki and Giovannoni, 1996). This can be termed reannealment inhibition of PCR and is dependent on product concentration; thus if a template of initial high concentration reaches this effective saturation point at a cycle number beyond which amplification continues for many rounds, concentrations at the end-point of the reaction will not reflect the initial ratio (Suzuki and Giovannoni, 1996; Bonnet *et al*., 2002). To minimise the potential for such a bias the PCR total cycle number is generally maintained below 20 in analyses of microbial community structure (Wintzingerode *et al*., 1997), and investigations have shown greater correlation between PCR and plate counts at these levels (Polz and Cavanaugh, 1998). A further advantage of employing low cycle numbers is that it prevents excessive accumulation of spurious products, artifacts, and the chimeras which result from curtailed extension of a primer with subsequent reannealment of the truncated product to a non-homologous sequence (Kanagawa,

2003; Qiu *et al*., 2001; Speksnijder *et al*., 2001). The corollary of an approach which utilises low-cycle number, though, is that minority members of the community may remain unidentified (Moeseneder *et al*., 1999), and yields may be low (Sipos *et al*., 2007).

All of the above factors were considered when selecting oligonucleotides and establishing PCR protocols for primers utilised in the current study, and where standard 16S community primers were utilised, low cycle numbers (< 25) with 'high' starting template concentrations (Chandler *et al*., 1997) were employed. However, estimation of the concentration of bacterial genomic DNA as a template in PCRs was rudimentary in that host genomic DNA constituted an unknown proportion. Standardised dilutions of templates were employed to approximate to 25 ng per PCR reaction but it is likely that the range of concentrations of 16S rDNA varied across a sample set.

While methods for preparation of libraries or amplicons for sequencing may thus introduce bias in a number of ways, and it was clear from sections 4.3 and 4.4 that primer pair selection is significant in this respect, as a whole experiments were shown to be reproducible. In particular, experiments detailed in section 4.5 revealed that replicate PCRs and extractions led to very little variation in estimation of a sample's microbiotc community structure. While this might be deemed an elementary finding it is important with regard to future studies: while consistency of primers, PCR constituents or cycling conditions, and extraction techniques can be maintained, stochastic factors influencing the outcome of a PCR or set of extractions would be far more problematic to identify and eliminate.

6.4 Technical issues with arrays

Development of the OFARG approach, as an alternative to the laborious sequencing of clones, or utilisation of low-resolution techniques such as TGGE, DGGE, or T-RFLP, was one of the main objectives of the research. It was intended to be a parallel and complementary approach to 454 sequencing, verification of its efficacy being through sequencing of selected clones with confirmed fingerprints, and qPCR of certain samples targetting identified bacterial groups. However, even prior to curtailment of optimisation due to time constraints, there were doubts as to its viable application as a 'high-throughput' technique. In particular, a bottle-neck arises between creation of clone libraries and production of amplicons for 'spotting', while the cost of the approach in terms of PCR constituents and array construction was greater than had been envisaged. In addition, the number of hybridisations which provided images of a suitable quality for analysis could not be reduced below about 40%, further supplementing the cost per 'clone'.

In the subsequent sections possible reasons for the poor quality of arrays as a whole, and the unpredictable nature of certain hybridisation events will be discussed. However, a consistent hindrance to effective investigation was the array printing process. For this, either an Arrayjet or a Stanford spot printer (see Appendix 3) were utilised and technical difficulties with the apparatus meant that both were rarely available concurrently; for instance, the Arrayjet was twice returned to the manufacturer. The two printers are fundamentally different in that the Stanford is a contact printer while the Arrayjet head 'sprays' the PCR products onto the slide surface to create a feature. For OFARG to be an effective high-throughput technique required eventual use of the Arrayjet, which can produce arrays with a greater feature density, but it was unavailable for phase I of the research (section 3.3) so early optimisation was conducted using Stanford-printed arrays, the Arrayjet then being utilised for phase II (section 3.7). It

may be that this had no bearing on the outcome but it is interesting that hybridisations were found to be less reproducible with arrays printed by the Arrayjet.

6.4.1 Slide chemistry and spotting

The array platform utilised was a glass substrate coated with poly-L-lysine, the positive charges of the lysine residues interacting electrostatically with the negative phosphate groups of the DNA (Roth and Yarmush, 1999), and cross-linked prior to hybridisation via heating (Dufva, 2005). This relatively non-specific interaction with adsorption of the nucleic acids onto the surface has two important consequences: firstly, the feature may not be in the optimal conformation for hybridisation with probes, and secondly, the chemistry of the slide itself may allow non-specific adsorption of the interrogatory probes (Roth and Yarmush, 1999; Beaucage, 2001). Aldehyde- and aminosilane-coated slides were investigated as alternatives but appeared to reduce the amount of nucleic acids which could be immobilised despite the covalent basis of the interaction (Beaucage, 2001), while also leading to higher background signal. The issue of non-specific adsorption did not appear to be significant for OFARG as detectable intensity of fluoresecence (with poly-L-lysine arrays) was limited to feature boundaries on good quality images, while rejected images tended to suffer from black holes, physical effects such as coverslip scratches, or probe aggregation (hot-spots) possibly caused by evaporation and/or poor diffusion of probes (Bowtell and Sambrook, 2003). In addition, long hybridisation times are expected to minimise adsorption, as probes are predicted to bind non-specifically in the short-term but then approach equilibrium with their complementary sequences as the reaction proceeds (Roth and Yarmush, 1999).

In spotting the microbiotic libraries, DNA was dissolved in betaine-SSC. Betaine (*N*,*N*,*N*-trimethylglycine) is a viscous liquid which prevents evaporation of water and allows for homogeneous distribution of nucleic acids across the feature (which is thus more tightly defined), while it is also a powerful denaturant and generally improves

immobilisation of PCR products on poly-L-lysine slides (Diehl *et al*., 2001; Dufva, 2005). However, its viscosity may have contributed to problems encountered in slide production with the ArrayJet, each machine and pin head combination apparently having an optimal spotting solution.

6.4.2 Steric hindrance and spacers

Steric hindrance is a non-specific interaction of the spotted DNA and substrate surface which can adversely affect accessibility of probes; bases closest to the surface of the support are less accessible than those towards the non-tethered end (Southern *et al*., 1999; Peplies *et al*., 2003). The size of this effect is thought to correlate with the size of the probe and the distance between the support and the immobilised nucleic acid (Southern *et al*., 1999). Although the former may be constrained by issues of specificity, the latter can be increased through incorporation of spacer regions, and it has been found that there is a beneficial effect on signal intensity with increasing distance of the spotted product from the substrate surface (Peplies *et al*., 2003). The spacers are generally poly 'A' or poly 'T' tails of about 10 nucleotides in length (Palmer *et al*., 2006), although the OFARG approach incorporates significant spacer regions (between 20 and 250 bp) merely through PCR from the plasmid, conformation of the product on the spot probably being of greater importance.

6.4.3 Kinetics of duplex formation and probe concentration

The thermodynamics of duplex formation in solution are relatively well-established, but these can not be applied with confidence to immobilised nucleic acids and probe capture (Pozhitkov *et al*., 2006). However, hybridisation kinetics are generally thought to be second-order in that rate is concentration dependent, in the region of 10^4–10^6 M^{-1} s^{-1} for association (Roth and Yarmush, 1999). For short (20-25mer) probes dissociation is also concentration dependent while optimal wash temperatures are low (circa 22°C),

as are hybridisation temperatures (40-60°C; Bodrossy and Sessitich, 2004; Koltai *et al*., 2008). The process of hybridisation can be viewed as a sequence of phases with the first involving multitudinous duplex formation due to high concentrations and high association rates, followed by a period in which equilibrium is approached due to slow dissociation of perfect matches and faster dissociation of mismatches (Dai *et al*., 2002; Koltai *et al*., 2008).

Previous investigations have found that a minimum of 500 ng is required to establish significant SNRs (Loy *et al*., 2002; Palmer *et al*., 2006), while between 1 and 2 μg of labelled probe in a hybridisation volume of approximately 40 μl appears to be optimal (Cho and Tiedje, 2002; De Santis *et al*., 2007). In attempting to achieve a signal more than 5 μg of probe was utilised in phase I, but this was reduced once it became clear that hybridisation temperature and buffer composition were primarily responsible for lack of detectable duplex formation. It is possible that probe concentration should have been further attenuated to improve specificity (Sorokin *et al*., 2006), but it is unlikely that this would have led to improved hybridisation of probes to features with complementary regions of sequence, but not designated as hybridised (Naef *et al*., 2003).

6.4.4 Diffusion

In the majority of array experiments a diverse population of immobilised 'probes' is interrogated with an equally complex mixture of labelled cDNA testers. Diffusion of components is thus an important consideration and there have been suggestions that Cy3 and Cy5 labelled oligonucleotides move at different rates, diffusion of the Cy3 dye thought to be more rapid (Borden *et al*., 2005). Experiments to investigate this found that variations in signal intensity were founded in detection sensitivity of the scanners, agitation of arrays during hybridisation causing a 20% increase in signal intensity for both dyes (Borden *et al*., 2005). The inherent variability in signal intensity between the dyes, Cy3 being detected at concentrations half that of Cy5 (Borden *et al*., 2005), must

be accounted for in future experimental design via quantile normalisation. While the OFARG approach should not suffer from diffusion-dependent effects, small hybridisation volumes (less than 50µl) may be subject to unspecified viscosity factors (Bodrossy and Sessitich, 2004), such that agitated hybridisation has the potential to be more reliable than static (Koltai *et al*., 2008), and may have been of value in eliminating the phenomenon of 'black holes'. Variable viscosity or diffusion rates of probes may also account for aggregation of oligonucleotides (visible as hotspots) on the array surfaces, although evaporation of buffer solution (despite humidification of chambers) is also thought to have contributed to this effect.

6.4.5 Secondary structure

Steric hindrance mediated by undesirable interactions between substrate and nucleic acids limits probe accessibility to complementary regions, but the formation of intramolecular secondary structure can also diminish the potential for hybridisation **(**Pozhitkov *et al*., 2006). In fact, the secondary structure (of probes or features) is thought to be the main limiting factor for duplex yield (Peplies *et al*., 2003). This is of particular relevance to OFARG since formation of secondary structure by the 16S rRNA sequence is inherently related to its *in vivo* function (Gutell *et al*., 1994).

Short probes are more effective in accessing their complementary spotted sequences as smaller molecules can penetrate the densely-packed features more readily (Southern *et al*., 1999), while intramolecular pairing of shorter fragments is statistically less likely and, should secondary structures develop, they are markedly less stable than those of longer fragments (Peplies *et al*., 2003). However, the longer, spotted PCR products constituting the features (~1.5 Kb) would be capable of forming significant regions of secondary structure. Although such interaction may only be of importance at lower hybridisation temperatures, and buffer composition may be adapted to attenuate such reduction in accessibility e.g. through incorporation of tetramethylammonium (TMA)

salts (Southern et al 1999; Peplies *et al*., 2003), it is possible that this mechanism explains a proportion of the unpredictable hybridisation.

6.4.6 Specificity, buffers and hybridisation temperature

In microarray systems specificity is the relationship between output signal from a feature and interaction with its complementary probe sequence, and is intimately linked to both hybridisation temperature and buffer composition, which can be viewed as the analogs of annealing temperature and salt concentration in PCR.

Differentiation between a single base mismatch and a perfect match represents the ultimate in specificity. With regard to probes this is attained by focus on comparable T_m and buffer requirement (Sorokin *et al*., 2006), while features on the array should contain defined amounts of DNA (Koltai *et al*., 2008). However, absolute differentiation of perfect matches from mismatched probes may not be possible within a given system, estimates of contiguity or identity sufficient for stable duplex formation being in the rangeof 75-80% (Kane *et al*., 2000; Tomiuk and Hofmann, 2001; Dai *et al*., 2002).

One of the aims of OFARG was that a single buffer and hybridisation temperature should be applicable for the majority of probes, in addition to the design constraint that probes should target semi-conservative regions of the SSU rRNA. The effective T_m of oligonucleotides may be homogenised through inclusion of chemicals such as tetramethylammonium chloride or betaine in buffers (Peplies *et al*., 2003); certain reports state that specificity is maintained at standardised hybridisation temperatures under these conditions (Peplies *et al*., 2003), while others are less unequivocal in their findings (Loy *et al*., 2002). While absolute specificity of duplex formation was not a requirement in OFARG, this can be improved, subsequent to the hybridisation period, through high stringency washes of the arrays (Pozhitkov *et al*., 2006), a low salt concentration enhancing dissociation of non-specifically bound probes while the more stable perfect matches remain hybridised. With regard to the concentration of such

solutions it is of interest that the distance of the hybridisation event from the slide surface may affect the stringency required in washes: much lower salt concentrations are required for stringency the further away from the surface that the oligo has bound (Poulsen *et al*., 2008).

Results from phase II of OFARG strongly suggested that a combination of localised secondary structure of spotted product, and reduced specificity due to homogenisation of hybridisation conditions for probes with differing thermodynamic characteristics, were responsible for unpredictable hybridisation events (negative and positive). It is also possible that 3DNA flurophores were interacting with probes or spotted products in a non-specific manner to skew recorded feature intensities (Randolph and Waggoner, 1997; Naef and Magnasco, 2003).

Each array could only be interrogated with 2 probes per hybridisation, so there was scope for some adjustment of buffer composition, hybridisation temperatures, and washes, to provide a more optimal environment for specific interactions. For instance, some of the probes were derived from studies using FISH, where buffers with varying concentrations of formamide were required (Giovannoni *et al*., 1988; Harmsen *et al*., 2002; Zhu and Joerger, 2003), while Denhardt's solution, (Denhardt, 1966), Sarkosyl buffer (Valinsky *et al*., 2002a), or other wash solutions such as SET (NaCl, EDTA, Tris) could also have been introduced (Giovannoni *et al*., 1988).

As mentioned, though, such a thorough examination of the thermodynamic properties of the probe set for the OFARG system required significant time and resources, both of which were in short supply once this became apparent.

6.4.7 Probe design

While the final probe set was perhaps one of the successes of the research (in theoretical terms, if not in practical application), the scripts for probe design were not optimally coded. Both the 'Match' and 'Contiguity' scripts (Appendices 7 and 8) were

deficient with regard to a randomisation function (for the order of the tabulated binary fingerprints) linked to an output text file which might have allowed more efficient probe selection. While they enable iteration through a table of the expected interactions between sequences and probes, output was dictated by the order of the fingerprints. So, if 2 sequences produced identical fingerprints with a test set, and occupied the first 2 positions of the table, the iteration would end after a single comparsion. Improved code would also have incorporated a scoring system such that multiple probe sets could be tested and compared without excessive manual curation of files in Excel.

Overall, the approach to probe design was probably somewhat naïve, especially the expectation that such a comprehensive set could be collated without recourse to established software and algorithms (Borneman *et al*., 2001; Militon *et al*., 2007). In truth, though, there is a complex relationship between probe length, probe concentration, buffer composition, hybridisation temperature, T_m of duplexes, and oligo kinetics, which determines the specificity for a probe in an array system (Relógio *et al*., 2002). As a result each probe must be empirically assessed within the array system under a range of conditions so *in silico* approaches, while potentially providing guidelines and parameters for probe design, may be of limited value for prediction of probe behaviour on novel platforms (Pozhitkov *et al*., 2006).

6.5 Technical issues with 454

Two decades of experiments have provided a comparative wealth of publications and technical data on the dynamics of duplex formation on arrays, permitting speculation on causes of idiosyncratic results and thus suggesting a range of options to attempt resolution.

The relative novelty of 454 means that issues with the platform are less well-documented. Indeed, use of the sequencing platform has been embraced by microbial ecologists, and successful investigations of the effects of diet (De Filippo *et al*., 2010; Claesson *et al*., 2012), and antibiotics (Dethlefsen *et al*., 2008; Zwielehner *et al*., 2011), on the intestinal microbiota of humans have been undertaken. Aside from some concerns regarding over-estimation of diversity (Pilloni *et al*., 2012; Schloss *et al*., 2012), and the potential for barcoded primers to bias community PCR (Berry *et al*., 2011), the pitfalls of community analysis with 454 are considered equivalent to those experienced in any PCR-based investigation of microbiomes (Forney *et al*., 2004).

However, while the evidence is mostly anecdoctal, it appears that there are numerous unresolved issues with the reliability of the platform in providing adequate sequence, particularly with regard to the GS FLX Titanium reagents and software. For instance, there are more than 550 threads relating to 454 on the seqanswers.com website, albeit dating back to 2008, on a range of topics, from the acceptable enrichment percentages of beads from emPCR, to the processing pipelines (amplicon or shotgun) used to convert raw signals into nucleotide sequence. Common themes are the variability in quality of reagent batches and acceptance that a percentage of runs will fail to provide expected numbers of good quality sequence reads.

Experience of 454 sequencing throughout the course of research was that the original GS FLX system (~200 bp reads) provided reliable output, while the number of expected reads per PTP from the Titanium (~400 bp reads) system was subject to successive

reductions and the eventual output was unlikely to conform to even the downgraded estimates.

Of particular concern is the difference in output achieved from the same raw data simply through use of the genome (as opposed to the amplicon) sequence processing pipeline - counterintuitively the genome pipeline appears better suited to processing the amplicons produced for studies investigating the SSU rRNA of bacteria (http://pathogenomics.bham.ac.uk/blog/2011/05/curious-results-from-454-amplicon-processing/) and settings for the 'valley filter' seem integral to this anomaly.

Such a problem was encountered with data for the 454 run described in chapter 5: initially read numbers from .sff files were less than 10% of the final downloaded total, the first files being processed with the amplicon pipeline, while for the second batch the genome pipeline was utilised. However, the final number of reads suitable for analysis (subsequent to filtering in QIIME) approximated to the initial total, while a parallel run for COPD samples, prepared concurrently using the same primers, PCR constituents, protocols and quality controls, displayed an equivalent initial discrepancy but retained greater then 80% of reads for clustering and assignment to OTUs.

It is possible that the differences in final output originate from residual contaminants in the faecal DNA samples which were not removed by the extraction protocol, and for which there were no comparable PCR inhibitors in the sputum extracts. The fact that all 3 regions experienced a similar reduction in read numbers through quality filtering in QIIME would support this, since it seems unlikely that the 3 faecal pools would be sequenced using a batch of poor quality emPCR/454 reagents while the single COPD pool remained unaffected. However, one of the faecal samples from the same group had been sequenced a few months earlier, contributing to data presented in section 4.5. The sample had been extracted contemporaneously with samples ultimately utilised for chapter 5 investigations, and differed from samples in that set only by virtue of the

batch of PCR reagents employed for final amplicon preparation, while primer sequences and stocks were identical. It is no simple matter then, to explain the numbers and quality of sequence reads which were obtained for the final 454 sequencing run.

A prevalence of short amplicons in the pools has been suggested as one possible reason for predominance of poor quality reads in 454 sequencing output. It is suspected that the shorter, spurious amplicons are preferentially amplified during the emPCR phase, leading to disproportionate representation of beads with non-target material on the PTP. However, this does not explain the quality-filtering of reads by QIIME across the full range of read lengths (figure 5.3), nor the prevalence of reads filtered due to mismatches in primer sequences, particularly for region B.

It is clear, however, that amplicon preparation for 454 requires specific precautions to ensure viability for sequencing. The Roche fusion sequences at the 5′ end of primers which are necessary for bead-binding, emPCR and sequencing caused reduced yields in PCR, contrary to previous evidence suggesting improved performance of primers with 5′ tails (Regier and Shi, 2005; Afonina *et al*., 2007). As a result, higher cycle numbers were employed to obtain sufficient product for subsequent steps. It is possible that this causes spurious artifacts to accumulate which are not adequately removed by the AMPure procedure. The Agilent BioAnalyser would fail to detect these amplicons if the range of lengths is broad, such that the total molarity is significant but the total mass at any given length is small. In combination with a small fraction of primer dimers which are not eliminated by the AMPure step, there would then be the potential for considerable disruption during the emPCR phase as mentioned above.

Recommendations to minimise carry-over of extraneous amplicons are additional AMPure steps before and after pooling, and gel extraction of the desired amplicon band subsequent to initial 'community' PCR. In all future investigations this approach would be implemented.

6.6 Analysing complex microbial populations

The current research was disappointing in that OFARG remained inchoate despite the theoretical potential of the probe set, while 454 sequencing failed to provide the expected volume of quality data, possibly due to stochastic processes inherent in the procedure combined with unsuitability of first-pass processing algorithms. Combined with difficulties in obtaining sample sets and associated experimental design flaws, significant steps towards identification of a CDAD-specific microbiota, or markers for predisposition, could not be made.

However, the research has not been without merit with regard to continuing investigations into CDAD, or the dynamics of the intestinal microbiota in health conditions ranging from obesity (Ley *et al*., 2005; Ley *et al*., 2006; Turnbaugh *et al*., 2006) to autism (Parracho *et al*., 2005) and allergy (Noverr and Huffnagle, 2005). It has been established that there are differences between the intestinal microbiotic composition of CDAD and AAD patients and that these can be detected and quantified, even if sufficient power of statistical testing could not be achieved to confirm their precise nature.

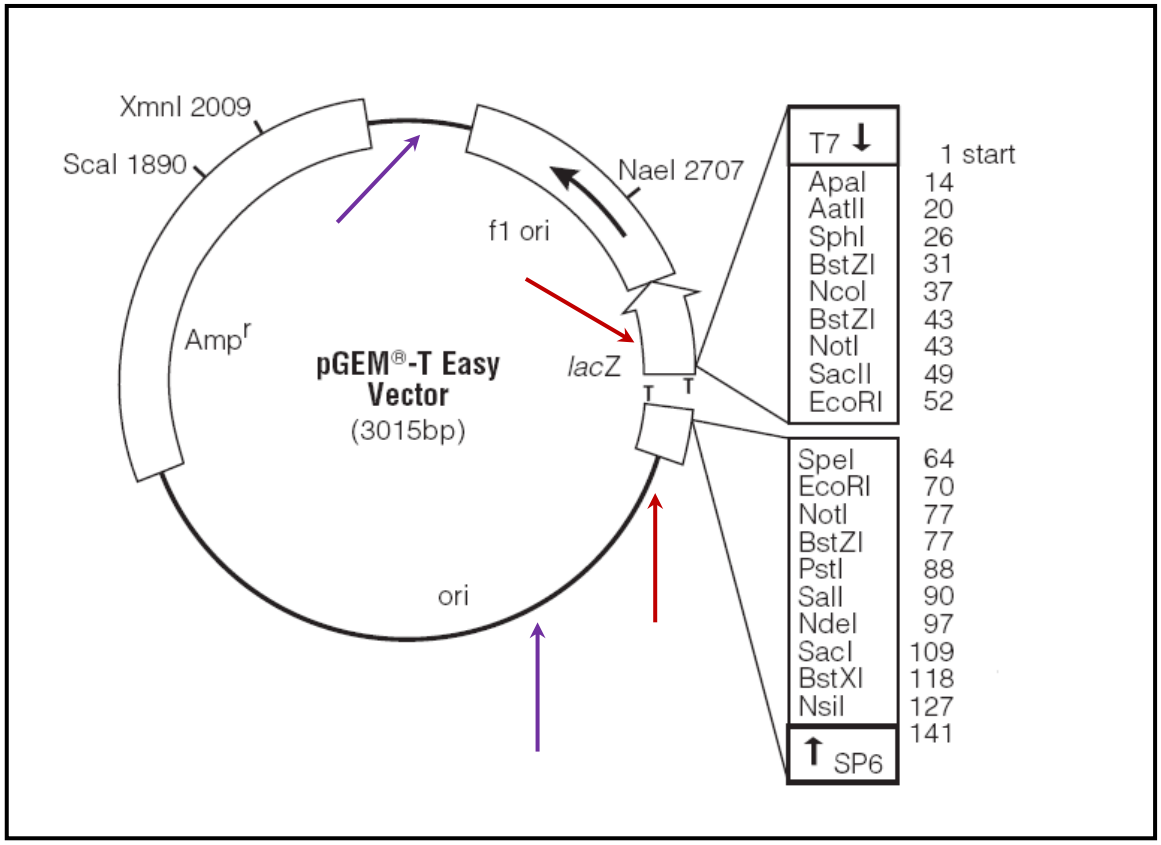
Future studies would address this through recruitment of increased numbers of subjects, while acquisition of 'baseline' samples would be of immense value so that individual dynamics could be determined, thus permitting improved identification of patterns within and between groups. However, animal models of CDAD are now well-established, and have been used to investigate relapse (mouse; Sun *et al*., 2011), and therapeutic potential of antibodies against *Clostridium difficile* toxins (hamster; Babcock *et al*., 2006); for longitudinal studies rodents probably represent a more realistic resource. If trials are conducted with human subjects, collation of comprehensive metadata is essential, while consistent adherence to strict protocols is also vital to minimise introduction of experimental bias.

It is perhaps presumptuous to be considering the next wave of sequencing platforms when issues with the current technologies have not been resolved, but it has essentially arrived and may obviate some of the problems encountered with second generation sequencers. For instance, the PacBio RS (Pacific BioSciences, US) provides reads in excess of 1 kb and does not require a pre-amplification procedure, thus eliminating a potential source of bias (Liu *et al*., 2012), although current protocols for 16S sequencing still incorporate such a step. Undoubtedly, the new technologies will find application in microbiomics, and the current bioinformatics software may require additional features to process the data. Improved coding ability, and a deeper understanding of multivariate statistical methods and ordination, may also permit development of a more comprehensive and integrated pipeline for analysis.

It is also expected that dissemination of data from the HMP (Peterson *et al*., 2009), and from application of techniques such as TAS (Targetted Amplicon Sequencing; Bybee *et al*., 2011) or MLST (Multi-Locus Sequence Typing; Maiden *et al*., 1998; Aanensen and Spratt, 2005), would inform and direct future investigations. While the intention would be to retain focus on analysis of SSU rRNA, awareness of the potential for parallel accumulation of data within the spheres of metabolomics, metagenomics, and viromics should allow for ever more thorough mapping of the biological landscape of the human intestinal ecosystem.

APPENDICES

APPENDIX 1: THE PGEM®-T EASY VECTOR



Above can be seen the vector utilised for the creation of microbial libraries. Reproduced with permission from Promega.

Important stretches of sequence are the multiple cloning region allowing for digestion with a variety of restriction enzymes, *NotI* being of particular utility for complete removal of inserts; the T-tails of the linearised vector are also located within this region of the *lacZ* gene. The pUC/M13 Reverse Sequencing Primer binding site lies external to SP6 while the pUC/M13 Forward Sequencing Primer binding site lies external to T7. (red arrows). Approximate binding sites of pGEM primers are denoted by purple arrows.

APPENDIX 2 : LIST OF CHEMICALS AND CONSUMABLES

Ampicillin sodium salt, 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal), and isopropyl-β-D-thiogalactopyranoside (IPTG) were obtained from Melford Labs (UK).

Betaine inner salt monohydrate, poly-L-lysine, ethylene glycol, N,N-dimethyl formamide (DMF), Orange-G dye, and 3-(*N*-morpholino) propanesulfonic acid (MOPS) were purchased from Sigma-Aldrich (UK).

Sodium dodecyl sulphate, propan-2-ol (isopropanol), Tris base, NaCl, KCl, D-glucose, sodium hydroxide (NaOH), glacial acetic acid, ethidium bromide, diaminoethanetetra-acetic acid disodium salt (EDTA), disodium hydrogen phosphate, potassium dihydrogen phosphate, sodium acetate and sodium citrate were bought from Fisher Scientific (UK).

Glycerol, sodium hypochlorite, and hydrochloric acid (HCl) were obtained from Acros Organics (Belgium). Glycogen was obtained from Roche Diagnostics (USA). Big Dye® version 3.1 and sequencing buffer were provided by Applied Biosystems (USA), while Performa gel Filtration Cartridges for clean-up of sequencing reactions were purchased from Edge Bio (USA). Ethanol was obtained from departmental stocks.

Standard nutient agar, bacto-yeast, and bacto-tryptone were provided by Oxoid (UK). SeaKem® LE Agarose was purchased from Lonza (USA). Hyperladder I marker for estimation of the size of DNA fragments under electrophoresis was supplied by BioLine (UK).

Bio-X-Act™ DNA Polymerase, associated buffers and $MgCl_2$ (50 mM) were supplied by Bioline (UK), while KAPA Taq DNA Polymerase, associated buffers and $MgCl_2$ (25 mM) were supplied by Kapa Biosystems (USA). Bovine serum albumin (BSA), restriction enzymes, and Phusion Hi-Fidelity DNA Polymerase were all supplied by New England Biolabs (UK). dNTPs for PCR were purchased from Promega (USA), as were pGEM®-T Easy, associated buffer and T4 DNA ligase.

For extraction of DNA from agarose gels the following kits were utilised: Genelute™ Gel Extraction Kit from Sigma-Aldrich (UK); Zymoclean™ Gel DNA Recovery Kit from Zymo Research (USA); and the QIAquick® Gel extraction Kit from Qiagen (UK). For standard purification of PCR products the following kits were utilised: E.Z.N.A™

Cycle Pure Kit from Omega-BioTek (USA) and the Zymo DNA Clean and Concentrator™ Kit from Zymo Research (USA). For isolation of plasmid vectors from host bacteria the E.Z.N.A™ Plasmid Mini Kit 1 from Omega-BioTek (USA) was employed, while for extraction of total community bacterial DNA from faecal samples the QIAamp® DNA Stool Mini Kit was used. In addition, array hybridisations were conducted with the Array50® Kit from Genisphere (USA), purification of 454 products was through use of the Agencourt® AMPure® XP kit from Beckman Coulter (USA) and assessment of DNA concentration through fluorescence employed the Quant-iT™ PicoGreen ® Kit from Invitrogen (UK).

Standard Petri-dishes were purchased from Sterilin (UK), while 245 mm x 245 mm low-profile polystyrene bioassay dishes for use with QPix were obtained from Corning (USA). Disposable 5, 10, and 25 ml Costar® pipettes for use with the IBS Pipetboy (Integra BioSciences, USA), and 15ml and 50 ml Falcons™ were also supplied by Corning (USA). 7 ml bijou tubes and 30 ml universal containers were purchased from VWR (UK). 5, 10 and 20 ml disposable syringes were obtained from Becton Dickinson (USA) while attachable Acrodisc ® filters of 0.2 and 0.45 µm pore size were manufactured by Pall (USA). Standard polypropylene tubes of volume 0.5, 1.0, 1.5 and 2.0 ml were purchased from Eppendorf (UK), while 2.0 ml cryotubes were obtained from Thermo Scientific (UK).

Flat-cap 0.3 ml PCR tubes and Thermo-Fast® 96-well non-skirted PCR plates (0.3 ml per well) were purchased from Thermo Scientific (UK), while adhesive PCR film for sealing plates was supplied by Abgene (UK). 96-well polypropylene round-bottom plates and 96-well flat-bottom polystyrene plates (both 0.4 ml per well), for use with AMPure and Pico-green kits respectively, were purchased from Nunc (Denmark). For colony growth and robotic transfer the following plates were utilised: low-profile polystyrene flat-bottom 96-well plates (0.2 ml per well), standard-profile polystyrene flat-bottom 96-well plates (0.4 ml per well), and low-profile polystyrene flat-bottom 384-well plates (50 µl per well) all supplied by Genetix (UK).

Standard glass slides (25 mm x 75 mm x 1 mm) and coverslips (44 mm x 22 mm) were puchased from Fisher Scientific (UK), while aminosilane-coated and aldehyde-coated glass slides of the same dimensions were obtained from Schott (UK). Specialist LifterSlip™ coverslips (40 mm x 22 mm) were supplied by Thermo Scientific (UK) while lens cleaning tissue was provided by Whatman International (UK).

In addition, disposable pipette tips for the full-range of Gilson pipettes and multi-channel pipettes were supplied by Sarstedt (Germany) or Starlab (UK), while filter tips for the same apparatus were provided by Sarstedt (Germany). Eurotubo collection swabs were obtained from DeltaLab (Spain), 2 mm disposable electroporation cuvettes were purchased from Cell Projects (UK), and 0.22µm pore fliters, Stericup® and SteriTop® vacuum-filter accessories were manufactured by Millipore (USA).

APPENDIX 3: LABORATORY EQUIPMENT

Polymerase chain reactions and incubations of volumes of less than 0.5 ml were performed using the Mastercycler Pro VapoProtect (Eppendorf, UK), the GS 00001 (G-Storm, UK) or the MJ Research PTC-225 Peltier Tetrad (Bio-Rad laboratories Inc., USA). The latter was also used in conjunction with a twin-tower block (2 x 16 slides) for incubation of slides at temperatures greater than 80°C.

Incubations of volumes between 0.5 and 2 ml were performed using a Techne DriBlock DB2A (Bibby Scientific Ltd., UK). Larger volumes of up to 50 ml were incubated in a Grant W14 waterbath (Grant Instruments, UK), while volumes above 100 ml were heated in a Techne Hybridisation HB-1D Oven (Bibby Scientific Ltd., UK). In certain instances, rapid heating of solutions was achieved using a microwave oven (Panasonic, UK). Sterilisation of solutions by autoclaving was

Agitation and mixing of small-volume solutions was through use of a Vortex Genie 2 (Scientific Industries, USA), while agitation of larger volumes was achieved using a Stuart SSL4 See-Saw Rocker (Bibby Scientific Ltd., UK), or a HB-SHK1 shaker (Hybaid, UK).

Bench centrifugation of volumes up to 2 ml was through use of the Eppendorf 5415D (Eppendorf, UK). Centrifugation of volumes up to 50 ml was achieved with the Eppendorf 5804, while 96-well plates were subject to centrifugation using the Eppendorf 5810R (both Eppendorf, UK).

Liquid transfer of volumes up tp 1 ml was through employment of P2, P10, P20, P100, P200, or P1000 Pipetman Classic™ pipettes (Gilson Inc., USA). Volumes of up to 100 ml were transferred using an IBS Pipetboy Acu (Integra Biosciences, USA) with associated 5, 10 or 25 ml disposable Costar® pipettes (Corning Inc., USA). Volumes above 100 ml were transferred using Nalgene® measuring cylinders (ThermoScientific, USA), Pyrex®/Boro® glass cylinders (Fisher Scientific, UK) or Schott durans (Schott, UK). For transfer of volumes into 96-well plates the Finnipipette™ T14585, T06470 or V44490 multi-channel pipettes (0.5-10 µl, 5-50 µl, or 5-300 µl respectively) were employed (Thermo LabSystems, UK). For large-scale transfer of volumes from multiple 96-well plates to 384-well plates the Beckman Biomek 2000 Laboratory Automation Workstation was utilised (Beckman Coulter Inc., USA).

Preparation of agarose gels was in perspex trays utilising perspex combs for the wells, both custom-built by a University-based workshop. Electrophoresis was conducted in perspex tanks produced by the same workshop, potential difference being established using a PowerPac300 (Bio-Rad, USA). In addition, the Electro-Fast® 96-well gel system from Abgene (UK) wa sutilised where electrophoresis of entire plates was required. UV visualisation for records and/or extraction of DNA bands was using the Syngene Gene Genius Bio Imaging System (Synoptics, UK).

For measurement of fluorescence a FLUOstar Omega, multi-mode microplate reader (BMG Labtech) was utilised, while DNA purity and quantity were routinely assessed using a NanoDrop 1000 Spectrophotometer (Thermo Scientific, UK), or, in the early stages of the project, a Biophotometer (Eppendorf, UK). For stringent assesment of sample purity and quantity the Agilent 2100 Bioanalyzer from Agilent Technologies (UK) was employed.

Monitoring of pH was performed using the pH210 meter from Hanna Instruments (UK). Weights of chemicals and samples were assessed using a Swiss Quality 125A Balance (Precisa Instruments, Switzerland). The Ultrospec 10 Cell Density Meter (Amersham Biosciences, UK) was utilised to assess the OD_{600} values of bacterial cultures. Class II safety hoods and laminar flow unit were custom-built and supplied by Walker Biological Safety Cabinets (UK), while UV4PCR hood was manufactured by Scie-Plas (UK).

Cold storage of samples at 4°C was in Liebherr (UK) refridgeration units, while -20°C and -80°C storage required the use of freezers supplied by Sanyo Industries (UK).

Miscellaneous glassware including soda-lime staining jars for slide washes were purchased from Fisher Scientific (UK). Other equipment included a FisherBrand FB 7015S vacuum pump (Fisher Scientific, UK) for vacuum sterilisation, a Prestige Medical Autoclave (Prestige Medical, UK) for steam sterilisation of fluids and materials and hybridisation chambers (75 mm x 95 mm x 52 mm) for sealed array hybridisation and incubation (Genetix, UK).

Electroporation was performed using a composite unit comprising Gene Pulser™, Pulse Controller and Capacitance Extender all supplied by Bio-Rad Laboratories (USA). Large-scale blue-white colony selection was performed using a QPix robot (Genetix, UK), while array slides were printed using an Ultra Marathon Arrayjet (Arrayjet, UK)

or a Stanford spot printer, the latter having been custom-built by Dr Tim Gant (University of Leicester, UK) using components supplied by Western Technology Marketing (USA). Analysis of fluorescence of hybridised arrays employed an Axon 4200A 4-channel scanner from Molecular Devices Inc. (UK).

APPENDIX 4: MANUFACTURERS AND SUPPLIERS

Abgene, Abgene House, Blenheim Road, Epsom, Surrey, KT19 9AP, UK.

Acros Organics, Janssen Pharmaceuticalaan 3a, Geel-2440, Antwerpen, Belgium.

Agilent Technologies UK Ltd., 5 Lochside Avenue, Edinburgh Park, Edinburgh, EH12 9DJ, UK.

Amersham Biosciences, part of GE Healthcare Life Sciences, Amersham Place, Little Chalfont, Buckinghamshire, HP7 9NA, UK.

Applied Biosystems, part of Life Technologies Inc., 5791 Van Allen Way, PO BOX 6482, Carlsbad, California 92008, USA.

Arrayjet Ltd., Midlothian Innovation Centre, Pentlandfield, Roslin, EH25 9RE, Scotland, UK.

Beckmann Coulter Inc., 250 South Kraemer boulevard, Brea, CA 92821-6232, USA.

Becton Dickinson, 1 Becton Drive, Franklin Lakes, NJ 07417, USA.

BioLine Ltd., 16 The Edge Business Centre, Humber Road, London, NW2 6EW, UK.

Bio-Rad Laboratories Inc., 590 Lincoln Street, Waltham, MA 02451, USA.

BMG Labtech GmBH, 10 Hanns-Meyer-Martin-Strasse, D-77656, Offenburg, Germany.

Cell Projects Ltd., 2 Roebuck Business Park, Ashford Road, Harrietsham, Kent ME17 1AB, UK.

Corning Inc. Life Sciences, Tower 2, 4th Floor, 900 Chelmsford Street, Lowell, MA 01851, USA.

DeltaLab, Plaza de la Verneda, 1 Pol. Industrial La Llana, PO Box 195, 08191 Rubi, Spain.

Edge Bio, 201 Perry Parkway, Suite 5, Gaithersburg, MD 20877, USA.

Elga LabWater, Marlow international, Parkway, Marlow, SL7 1YL, UK.

Eppendorf Ltd., Endurance House, Chivers Way, Histon, Cambridge, Cambridgeshire, CB4 9ZR, UK.

Fisher Scientific UK Ltd., Bishop Meadow Road, Loughborough, Leicestershire, LE11 5RG, UK.

Genetix Ltd., Queensway, New Milton, Hampshire, BH25 5NN, UK.

Genisphere Inc., 2801 Sterling Drive, Hatfield, 19440 PA, USA.

Gilson Inc., 3000 Parmenter Street, PO Box 620027, Middleton, WI 53562-0027, USA.

Grant Instruments (Cambridge) Ltd., Shepreth, Cambridgeshire, SG8 6GB, UK.

G-Storm, Unit 3 Byfleet Technical Centre, Canada Road, Byfleet, Surrey, KT14 7JX, UK.

Hanna Instruments Ltd., Eden Way, Pages Industrial Park, Leighton Buzzard, Bedfordshire, LU7 4AD, UK.

Hybaid Ltd., Unit 5, The Ringway Centre, Edison Road, Basingstoke, Hampshire, RG21 6YH, UK.

Integra Biosciences AG, Tardisstrasse 201, CH-7205, Zizers, Switzerland.

Invitrogen UK Ltd., 3 Fountain Drive, Inchinann Business Park, Paisley, PA4 9RF, UK.

Kapa Biosystems Inc., 600 West Cummings Park, Suite 2250, Woburn, MA 01801, USA.

Lonza Rockland Inc., 191 Thomaston Street, Rockland, ME 04841, USA.

Liebherr-GB Ltd., Normandy Lane, Stratton Business Park, Biggleswade, Biggleswade, SG18 8QB, UK.

Melford Labs Ltd., Bideston Road, Chelsworth, Ipswich, Suffolk, IP7 7LE, UK.

Millipore, 290 Concord Road, Billerica, MA 01821, USA.

Molecular Devices Inc., 1311 Orleans Drive, Sunnyvale, CA 94089-1136, USA.

New England Biolabs (UK) Ltd., 75/77 Knowl Piece, Wilbury Way, Hitchin, Hertfordshire, SG4 0TY, UK.

Nunc A/S, Kamstrupvej 90, PO Box 280, DK-4000, Roskilde, Denmark.

Omega Bio-Tek Inc., 1850-E Beaver Ridge Circle, Norcross, GA 30071, USA.

Oxoid Ltd., Wade Road, Basingstoke, Hampshire, RG24 8PW, UK.

Pall Corporation, 2200 Northern Boulevard, East Hills, NY 11548, USA.

Panasonic UK Ltd., Panasonic House, Willoughby Road, Bracknell, Berkshire, RG12 8FP, UK.

Precisa Instruments AG, Moosmattstrasse, CH-8953, Dietikon, Switzerland.

Prestige Medical Ltd., Unit 1, First Avenue, Maybrook Industrial Estate, Minworth, Sutton Coldfield, West Midlands, B76 1BA, UK.

Promega Corporation, 2800 Woods Hollow Road, Madison, WI 57311, USA.

Purite Ltd., (a subsidiary of Ondeo Industrial Solutions), Bandet Way, Thame, Oxfordshire, OX9 3SJ, UK.

Qiagen Ltd., Qiagen House, Fleming Way, Crawley, West Sussex, RH10 9NQ, UK.

Roche 454 Life Sciences, 15 Commercial Street, Branford, CT 06405, USA.

Roche Diagnostics Ltd., Charles Avenue, Burgess Hill, West Sussex, RH15 9RY, UK.

Sanyo Scientific, Sanyo Sales and Marketing GmbH, 18 Colonial Way, Watford, Hertfordshire, WD24 4PT, UK.

Sarstedt AG and Co., Sarstedtstrasse, Postfach1220, 51582 Numbrecht, Germany.

Schott UK Ltd., Drummond Road, Astonfields Industrial Estate, Stafford, ST16 3EL, UK.

Scientific Industries Inc., 70 Orville Drive, Bohemia, NY 11716, USA.

Scie-Plas Ltd., Unit 3, Gainsborough Trading Estate, Southam, Warwickshire, CV47 1RB, UK.

Sigma-Aldrich Company Ltd., The Old Brickyard, New Road, Gillingham, Dorset, SP8 4XT, UK.

Starlab (UK) Ltd., 4 Tanners Drive, Blakelands, Milton Keynes, MK14 5NA, UK.

Sterilin Ltd., Parkway, Pen-y-Fan Industrial Estate, Newport, NP11 3EF, UK.

Synoptics Ltd., Beacon House, Nuffield Road, Cambridge, CB4 1TF, UK.

Techne Industrial and Stuart Equipment, now parts of Bibby Scientific Ltd., Beacon Road, Stone, Staffordshire, ST15 0SA, UK.

Thermo Scientific, 81 Wyman Street, Waltham, MA 02454, USA.

Thermo Scientific UK, Unit 5, The Ringway Centre, Edison Road, Basingstoke, Hampshire, RG21 6YH, UK.

Thermo LabSystems, 1 St Georges Court, Hanover Business Park, Altrincham, WA14 5TP, UK.

VWR International, Hunter Boulevard, Magna Park, Lutterworth, Leicestershire, LE17 4XN, UK.

Walker Safety Cabinets, Mill Street, Glossop, Derbyshire, SK13 8PT, UK.

Western Technology Marketing, 315 Digital Drive, Morgan Hill, CA 95037, USA.

Whatman International Ltd., Springfield Mill, James Whatman Way, Maidstone, Kent, ME14 2LE, UK.

Zymo Research Corporation, 17062 Murphy Avenue, Irvine, CA 92614, USA.

APPENDIX 5: SOFTWARE AND INTERNET RESOURCES

ProbeBase: http://www.microbial-ecology.net/probebase/

Ribosomal Database Project: http://www.rdp.cme.msu.edu/

Greengenes: http://greengenes.lbl.gov/cgi-bin/nph-index.cgi

ARB-SILVA: http://www.arb-silva.de/

NCBI: http://www.ncbi.nlm.nih.gov/

BLAST: http://www.ncbi.nlm.nih.gov/BLAST

MOTHUR: http://www.mothur.org/

QIIME: http://qiime.org/

R language: http://www.r-project.org/

Oligonucleotide T_m : http://www.basic.northwestern.edu/biotools/oligocalc.html

Spade: http://chao.stat.nthu.edu.tw/softwareCE.html

TreeView: http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

UniFrac: http://bmf.colorado.edu/unifrac/

Metastats: http://metastats.cbcb.umd.edu/

Phylip: http://evolution.genetics.washington.edu/phylip.html

MEGAN: http://ab.inf.uni-tuebingen.de/software/megan/

ITOL: http://itol.embl.de/

HMP: http://www.hmpdacc.org

Cytoscape: http://www.cytoscape.org/

Kinemage: http://kinemage.biochem.duke.edu/

APPENDIX 6: RDP INTESTINAL 214

The following are the species whose 16S sequences were utilised as the reference set for the probescript final iterations. Bold typeface indicates the first member of an OTU.

***Escherichia coli**; Enterobacter cloacae; **Sutterella wadsworthensis**; **Serratia liquefaciens**; Citrobacter freundii; Proteus mirabilis; Providencia stuartii; Providencia rettgeri; **Shigella flexneri**; Klebsiella pneumoniae; Citrobacter youngae; Escherichia fergusonii; **Burkholderia cepacia**; Achromobacter xylosoxidans; Alcaligenes faecalis; **Morganella morganii**; Yersinia enterocolitica; Plesiomonas shigelloides; Hafnia alvei; Yersinia frederiksenii; Enterobacter aerogenes; **Neisseria gonorrhoeae**; **Campylobacter jejuni**; **Aeromonas hydrophila**; **Aeromonas veronii**; **Wolinella succinogenes**; **Vibrio cholerae**; **Succinivibrio dextrinosolvens**; **Acinetobacter calcoaceticus**; **Haemophilus parainfluenzae Pseudomonas aeruginosa**; **Kingella kingae**; **Desulfovibrio vulgaris**; **Desulfovibrio fairfieldensis**; **Desulfovibrio desulfuricans**; Bilophila wadsworthia; **Photorhabdus luminescens**; **Serratia fonticola**; **Bacteroides vulgatus**; **Bacteroides thetaiotaomicron**; Bacteroides fragilis; Cytophaga fermentans; Bacteroides xylanisolvens; **Parabacteroides merdae**; Prevotella enoeca; Parabacteroides johnsonii; Porphyromonas asaccharolytica; **Bacteroides caccae**; **Bacteroides uniformis**; Alistipes putredinis; Alistipes finegoldii; Prevotella tannerae; Prevotella zoogleoformans; Alistipes shahii; Bacteroides dorei; **Bacteroides ovatus**; Porphyromonas levii; **Porphyromonas catoniae**; **Prevotella ruminicola**; **Myroides odoratus**; Mycoplasma pneumoniae; **Parabacteroides distasonis**; **Bacteroides eggerthii**; Bacteroides intestinalis; **Rikenella microfusus**; **Bacteroides stercoris**; **Prevotella pallens**; Prevotella loescheii; Prevotella bryantii; Bacteroides finegoldii; **Bifidobacterium bifidum**; **Bifidobacterium longum**; **Bifidobacterium adolescentis**; Bifidobacterium pseudocatenulatum; **Bifidobacterium angulatum**; Bacteroides capillosus; **Bifidobacterium breve**; **Bifidobacterium catenulatum**; **Actinomyces odontolyticus; Collinsella aerofaciens**; **Micrococcus luteus**; **Propionibacterium acnes**; **Mycobacterium avium**; Corynebacterium durum; **Actinomyces naeslundii**; **Eggerthella lenta**; **Fusobacterium nucleatum**; **Fusobacterium varium**; **Fusobacterium necrophorum**; Clostridium rectum; **Fusobacterium mortiferum**; **Bacillus subtilis**; **Bacillus cereus**; Staphylococcus aureus; Streptococcus salivarius; Streptococcus sanguinis; Streptococcus thermophilus; Streptococcus constellatus;*

Gemella bergeri**; **Gemella sanguinis**; **Enterococcus durans**; **Lactobacillus acidophilus**; Lactobacillus brevis; Lactobacillus delbrueckii; **Lactobacillus fermentum**; Lactobacillus pentosus; Lactobacillus rhamnosus; Lactobacillus gasseri; Lactobacillus plantarum; **Lactobacillus reuteri**; **Enterococcus faecium**; Enterococcus faecalis; Streptococcus mitis; **Streptococcus parasanguinis**; Streptococcus anginosus; Streptococcus mutans Aerococcus viridans; **Streptococcus bovis**; Streptococcus equinus; **Streptococcus gordonii**; **Eubacterium rectale**; **Dorea formicigenerans**; Clostridium leptum; Eubacterium contortum; **Dorea longicatena**; Ruminococcus torques; Coprococcus comes;Clostridium xylanolyticum; Clostridium nexile; Clostridium saccharolyticum; **Eubacterium oxidoreducens**; **Eubacterium eligens**; **Eubacterium hallii**; **Eubacterium xylanophilum**; **Ruminococcus obeum**; **Butyrivibrio fibrisolvens**; Ruminococcus lactaris; **Eubacterium ramulus**; Roseburia intestinalis Roseburia hominis**; Pseudobutyrivibrio xylanivorans**; **Clostridium clostridioforme**; **Clostridium propionicum; Clostridium indolis**; **Lachnospira pectinoschiza**; Roseburia faecis; **Ruminococcus gnavus**; **Roseburia cecicola**; **Coprococcus eutactus Coprococcus cactus**; Blautia producta Blautia hansenii; Blautia luti; **Clostridium symbiosum**; **Clostridium aminovalericum**; **Clostridium scindens**; **Butyrivibrio crossotus**; **Bacteroides pectinophilus**; **Faecalibacterium prausnitzii**; **Ruminococcus callidus**; **Ruminococcus flavefaciens**; **Ruminococcus bromii**; **Eubacterium desmolans**; **Papillibacter cinnamivorans**; **Clostridium orbiscindens**; **Eubacterium siraeum**; **Bacteroides capillosus**; **Anaerotruncus colihominis**; **Mogibacterium timidum**; **Eubacterium ventriosum**; **Eubacterium cylindroides**; **Eubacterium callanderi**; **Eubacterium dolichum**; Eubacterium tortuosum; Clostridium innocuum; **Eubacterium limosum**; Eubacterium biforme; **Clostridium acetobutylicum**; Clostridium bifermentans; **Clostridium haemolyticum**; **Clostridium butyricum**; **Clostridium paraputrificum**; **Clostridium perfringens**; Clostridium disporicum; Clostridium glycolicum; **Peptoniphilus asaccharolyticus**; **Anaerococcus prevotii**; **Parvimonas micra**; **Finegoldia magna**; **Veillonella dispar**; Veillonella atypica; Veillonella parvula; **Acidaminococcus fermentans**; **Veillonella ratti**; **Megasphaera elsdenii**; **Dialister invisus**; **Clostridium sporogenes**; **Sarcina ventriculi**; **Blautia coccoides**; **Peptostreptococcus anaerobius**; **Clostridium difficile**; **Clostridium sordellii**; **Dehalobacter restrictus**; **Peptococcus niger**; **Desulfotomaculum ruminis**; **Clostridium ramosum**; **Holdemania filiformis**; **Fibrobacter succinogenes**; **Mycoplasma pneumoniae**; **Akkermansia muciniphila**; **Verrucomicrobium spinosum

APPENDIX 7: PERL 'MATCH' PROBESCRIPT

The following script was compiled for the current research using the Perl coding language. Input is a text file provided as the output from the downloaded version of Blast assessing the matches between a list of sequences and a list of probes. The script determines whether differentiation of the sequences is possible with the given probe set, and provides a list of those sequences not differentiated if this is not the case.

```
#! usr/bin/perl -w
open ( SEQUENCES, "HITS.txt" ) || die "$!";
@sequences = <SEQUENCES> ;
close ( SEQUENCES );
$word = shift @sequences ;
print $word ;

@fingerprint_max = ();
$pro_row = -1 ;
$seq_col = 0 ;
$col_max = 0 ;

RUNTHROUGH: while (@sequences) {
      $_= shift @sequences;
      print $_ ;
      if (/Query=  probe\d+/) {
            $pro_row ++;
            print "$pro_row\n" ;
      } elsif (/Length=(\d+)/) {
            $probe_length_checker = $1 ;
            if ($probe_length_checker < 100 ) {
            $probe_length = $probe_length_checker ;
            print "$probe_length\n" ;
            }
      } elsif (/> Sequence(\d+)/){
            $seq_col = $1 ;
            print "$seq_col\n" ;
           if ( $seq_col >= $col_max ) {
                  $col_max = $seq_col ;
                  print "$col_max\n" ;
            }
      } elsif (m[Identities = (\d+)/(\d+)]) {
            $numerator = $1 ;
            $percent_match = ($numerator / $probe_length) * 100 ;
            print "$percent_match\n" ;
            if ( $percent_match >
$fingerprint_max[$pro_row][$seq_col] ) {
                  $fingerprint_max[$pro_row][$seq_col] =
$percent_match ;
                  print "$fingerprint_max[$pro_row][$seq_col]\n";
            }
      } else { next RUNTHROUGH ;
      }
```

```
}

open ( MAX, ">>max.txt" ) || die "$!";
for $row (@fingerprint_max) {
      print MAX "@$row\n" ;
      }
close ( MAX );


for $x (0 .. $pro_row) {
      for $y (0 .. $col_max ) {
            if ($fingerprint_max[$x][$y] >= 100) {
                  $fingerprint_max[$x][$y] = 1 ;
                  }
            else { $fingerprint_max[$x][$y] = 0 ;
                  }
      }
}

open ( MAX2, ">>max2.txt" ) || die "$!";
for $row (@fingerprint_max) {
      print MAX2 "@$row\n" ;
      }
close (MAX2) ;

$x = 0 ;
$y = 0 ;
@temp = () ;
$concatenation = 0;

for ($x = 0; $x <=  $pro_row; $x ++ ) {
      @temp = () ;
      for ( $y = 0; $y <= $col_max; $y ++ ) {
            push @temp, $fingerprint_max[$x][$y];
            if ($y == $col_max) {
                  $concatenation = join ( '', @temp );
                  push @compare, $concatenation ;
            }
      }
}

$probe_iteration = -1 ;
while (@compare) {
    $a = shift @compare;
      $probe_iteration ++;
      $probe_comparison = $probe_iteration ;
      foreach (@compare) {
            $probe_comparison ++ ;
            if ( $a eq $_ ) {
                  open ( PROBECOMP, ">>probecomp.txt" ) || die
"$!";
                  print PROBECOMP "Fingerprint of probe is not
unique! Probe $probe_iteration is equivalent to probe
$probe_comparison\n";
                  close (PROBECOMP);
            }
      }
```

```
}

print "Probecomp completed" ;

$x = 0 ;
$y = 0 ;
@temp3 = () ;
$concatenation = 0;

for ($y = 0; $y <=  $col_max; $y ++ ) {
      @temp3 = () ;
      for ( $x = 0; $x <= $pro_row; $x ++ ) {
            push @temp3, $fingerprint_max[$x][$y];
            if ($x == $pro_row) {
                  $concatenation = join ( '', @temp3 );
                  push @compare3, $concatenation ;
            }
      }
}

print "@compare3" ;

$seq_iteration = -1 ;
while (@compare3) {
    $a = shift @compare3;
      $seq_iteration ++;
      $seq_comparison = $seq_iteration ;
      foreach (@compare3) {
            $seq_comparison ++ ;
            if ( $a eq $_ ) {
                  open ( SEQCOMP, ">>seqcomp.txt" ) || die "$!";
                  print SEQCOMP "Fingerprint of sequence is not
unique! Sequence $seq_iteration is equivalent to sequence
$seq_comparison\n";
                  close (SEQCOMP);
            }
      }
}

print "Seqcomp completed" ;


$concatenation = 0 ;
$probe_counter = -1 ;
OUTER: while ($probe_counter <= $pro_row ) {
                  @compare2 = () ;
                  $probe_counter ++ ;
                  for ($y = 0; $y <=  $col_max; $y ++ ) {
                        @temp2 = () ;
                        for ( $x = 0; $x <= $probe_counter ; $x ++
) {
                              push @temp2,
$fingerprint_max[$x][$y];
                              if ($x == $probe_counter) {
                                    $concatenation = join ( '',
@temp2 );
```

```
                                    push @compare2, $concatenation
;
                              }
                        }
                  }
                  $comparison_iteration = 0 ;
                  $sequence_counter = -1 ;
      COMPARE: while (@compare2) {
                        $sequence_counter ++ ;
                        $comparison_iteration = 0 ;
                        $a = shift @compare2;
                        foreach (@compare2) {
                              $comparison_iteration ++ ;
                              if ( $a eq $_ ) {
                                    open ( PROBECOMP2,
">>probecomp2.txt" ) || die "$!";
                                    print PROBECOMP2 "Fingerprint
is not unique for probe count at $probe_counter and iteration
count $comparison_iteration from sequence$sequence_counter\n";
                                    close (PROBECOMP2) ;
                                    next OUTER ;
                              }
                        }
                  }
            open ( PROBECOMP2, ">>probecomp2.txt" ) || die "$!";
            print PROBECOMP2 "Probe $probe_counter completes the
set!\n";
            close (PROBECOMP2);
            last OUTER ;
}

if ($probe_counter == $pro_row) {
      print "Differentiation is not possible with this
probeset!";
      }
```

APPENDIX 8: PERL 'CONTIGUITY' PROBESCRIPT

Script detailed here is equivalent to that from Appendix 6, but assessing the probes and sequences based on a level of contiguity, generally set to 70% for the purposes of this study. Inputs and outpuuts are the same as those for the 'Match' script.

```
#! usr/bin/perl -w

open ( SEQUENCES, "HITS.txt" ) || die "$!";
@sequences = <SEQUENCES> ;
close ( SEQUENCES );

$pro_row = -1 ;
$seq_col = 0 ;
$col_max = 0 ;

RUNTHROUGH: while (@sequences) {
      $_= shift @sequences;
      print $_ ;
      if (/Query=  probe\d+/)                    {
            $pro_row ++;
            print "$pro_row\n" ;
      } elsif (/Length=(\d+)/)                   {
            $probe_length_checker = $1 ;
            if ($probe_length_checker < 100 ) {
            $probe_length = $probe_length_checker ;
            print "$probe_length\n" ;
            }
      } elsif (/> Sequence(\d+)/)                {
          $seq_col = $1 ;
            if ( $seq_col >= $col_max ) {
                  $col_max = $seq_col ;
                  print "$col_max\n" ;
            }
      } elsif (/\|/)                             {
          @contiguity = split (//, $_);
            $max = 0 ;
            $match_counter = 0 ;
            while (@contiguity) {
                  $c = shift @contiguity;
                  if ($c eq '|') {
                        $match_counter ++ ;
                        if ( $max <= $match_counter ) {
                        $max = $match_counter ;
                  }
                        }
                  elsif ($c eq '') {
                        $match_counter = 0 ;
                        }
            }
            $numerator = $max  ;
            $percent_match = ($numerator / $probe_length) * 100 ;
            print "$percent_match\n" ;
```

```
            if ( $percent_match >
$fingerprint_contig[$pro_row][$seq_col] ) {
                  $fingerprint_contig[$pro_row][$seq_col] =
$percent_match ;
                  print "$fingerprint_max[$pro_row][$seq_col]\n";
            }
      }
      else { next RUNTHROUGH ;
      }
}

for $x (0 .. $pro_row) {
      for $y (0 .. $col_max ) {
            if ($fingerprint_contig[$x][$y] <= 1) {
                  $fingerprint_contig[$x][$y] = 0 ;
                  }
            else { $fingerprint_contig[$x][$y] =
$fingerprint_contig[$x][$y] ;
                  }
      }
}

open ( CONT, ">>cont.txt" ) || die "$!";
for $row (@fingerprint_contig) {
      print CONT "@$row\n" ;
      }
close ( CONT );

for $x (0 .. $pro_row) {
      for $y (0 .. $col_max ) {
            if ($fingerprint_contig[$x][$y] >= 60) {
                  $fingerprint_contig[$x][$y] = 1 ;
                  }
            else { $fingerprint_contig[$x][$y] = 0 ;
                  }
      }
}

open ( CONT2, ">>cont2.txt" ) || die "$!";
for $row (@fingerprint_contig) {
      print CONT2 "@$row\n" ;
      }
close (CONT2) ;

$x = 0 ;
$y = 0 ;
@temp = () ;
$concatenation = 0;

for ($x = 0; $x <=  $pro_row; $x ++ ) {
      @temp = () ;
      for ( $y = 0; $y <= $col_max; $y ++ ) {
            push @temp, $fingerprint_contig[$x][$y];
            if ($y == $col_max) {
                  $concatenation = join ( '', @temp );
                  push @compare, $concatenation ;
            }
```

```
      }
}

$probe_iteration = -1 ;
while (@compare) {
    $a = shift @compare;
      $probe_iteration ++;
      $probe_comparison = $probe_iteration ;
      foreach (@compare) {
            $probe_comparison ++ ;
            if ( $a eq $_ ) {
                  open ( PROBECOMP, ">>probecomp.txt" ) || die
"$!";
                  print PROBECOMP "Fingerprint of probe is not
unique! Probe $probe_iteration is equivalent to probe
$probe_comparison\n";
                  close (PROBECOMP);
            }
      }
}

print "Probecomp completed" ;

$x = 0 ;
$y = 0 ;
@temp3 = () ;
$concatenation = 0;

for ($y = 0; $y <=  $col_max; $y ++ ) {
      @temp3 = () ;
      for ( $x = 0; $x <= $pro_row; $x ++ ) {
            push @temp3, $fingerprint_contig[$x][$y];
            if ($x == $pro_row) {
                  $concatenation = join ( '', @temp3 );
                  push @compare3, $concatenation ;
            }
      }
}

print "@compare3" ;

$seq_iteration = -1 ;
while (@compare3) {
    $a = shift @compare3;
      $seq_iteration ++;
      $seq_comparison = $seq_iteration ;
      foreach (@compare3) {
            $seq_comparison ++ ;
            if ( $a eq $_ ) {
                  open ( SEQCOMP, ">>seqcomp.txt" ) || die "$!";
                  print SEQCOMP "Fingerprint of sequence is not
unique! Sequence $seq_iteration is equivalent to sequence
$seq_comparison\n";
                  close (SEQCOMP);
            }
      }
}
```

```
print "Seqcomp completed" ;


$concatenation = 0 ;
$probe_counter = -1 ;
OUTER: while ($probe_counter <= $pro_row ) {
                  @compare2 = () ;
                  $probe_counter ++ ;
                  for ($y = 0; $y <=  $col_max; $y ++ ) {
                        @temp2 = () ;
                        for ( $x = 0; $x <= $probe_counter ; $x ++
) {
                              push @temp2,
$fingerprint_contig[$x][$y];
                              if ($x == $probe_counter) {
                                    $concatenation = join ( '',
@temp2 );
                                    push @compare2, $concatenation
;
                              }
                        }
                  }
                  $comparison_iteration = 0 ;
                  $sequence_counter = -1 ;
      COMPARE: while (@compare2) {
                        $sequence_counter ++ ;
                        $comparison_iteration = 0 ;
                        $a = shift @compare2;
                        foreach (@compare2) {
                              $comparison_iteration ++ ;
                              if ( $a eq $_ ) {
                                    open ( PROBECOMP2,
">>probecomp2.txt" ) || die "$!";
                                    print PROBECOMP2 "Fingerprint
is not unique for probe count at $probe_counter and iteration
count $comparison_iteration from sequence$sequence_counter\n";
                                    close (PROBECOMP2) ;
                                    next OUTER ;
                              }
                        }
                  }
            open ( PROBECOMP2, ">>probecomp2.txt" ) || die "$!";
            print PROBECOMP2 "Probe $probe_counter completes the
set!\n";
            close (PROBECOMP2);
            last OUTER ;
}

if ($probe_counter == $pro_row) {
      print "Differentiation is not possible with this
probeset!";
      }
```

APPENDIX 9: ECOLOGICAL INDICES

The following alpha and beta diversity measures are taken from the Mothur website, Ecological Methods (Southwood and Henderson, 2000), or from Chao *et al*., 2006.

The Shannon-Wiener diversity index:

$$SW = -\sum_{i=1}^{n} P_i \ln P_i$$

where: n is the total OTU number; i is the OTU in question; P_i is the proportion of the sample represented by OTU i. Approaches zero for samples of low diversity.

The Simpson diversity index:

$$SI = \sum_{i=1}^{n} P_i^2$$

where: n is the total OTU number; i is the OTU in question; P_i is the proportion of the sample represented by OTU i. Approaches zero for samples of high diversity so the reciprocal or inverse is often utilised.

The Berger-Parker dominance index:

$$BPI = \frac{n_i}{n}$$

where: n_i is the number of individuals in the most abundant OTU; n is the total number of individuals in the sample. Approaches 1 for samples with high dominance by a single taxon.

The Chao 1 diversity index:

$$Chao1 = \frac{n_1^2}{2n_2} + R$$

where: n_1 is the number of OTUs encountered once; n_2 is the number of OTUs encountered twice; R is the observed richness or total number of OTUs.

Good's coverage estimator :

$$C = 1 - \frac{n_1}{n}$$

where: n_1 = the number of OTUs that have been sampled once and N = the total number of individuals in the sample. Tends towards 1 as coverage improves.

The Sørensen and Jaccard similarity coefficients:

$$SQ = \frac{2i}{a+b} \times 100$$

$$JQ = \frac{i}{a+b-i} \times 100$$

where: i is the number of OTUs found in both samples; a is the number of OTUs in sample 1; b is the number of OTUs in sample 2. These are measures of similarity based purely on membership i.e. presence or absence of an OTU.

The Kulcyzcynski-Cody similarity coefficient:

$$KC = 1 - \frac{1}{2}\left(\frac{S_{AB}}{S_A} + \frac{S_{AB}}{S_B}\right)$$

where: S_{AB} is the number of OTUs in both communities; S_A is the number of OTUs in community A and S_B is the number of OTUs in community B

The Morisita-Horn similarity coefficient:

$$MH = 1 - 2\frac{\sum \frac{S_{Ai}}{n}\frac{S_{Bi}}{m}}{\sum\left(\frac{S_{Ai}}{n}\right)^2 + \sum\left(\frac{S_{Bi}}{m}\right)^2}$$

where; S_{Ai} is the number of individuals from community A in the *i*th OTU; S_{Bi} is the number of individuals from community B in the *i*th OTU; n is the total number of individuals in community A; and m is the total number of individuals in community B

The Bray-Curtis similarity coefficient:

$$BC = 1 - 2\,\frac{\sum \min(S_{Ai}, S_{Bi})}{\sum S_{Ai} + \sum S_{Bi}}$$

where: $\min(S_{Ai},S_{Bi})$ is the lesser of the abundance values for a shared OTU, S_{Ai} is the number of individuals in the *i*th OTU of community A and S_{Bi} is the equivalent for community B

Euclidean measure of similarity:

$$EMS = \sqrt{\sum (S_{Ai} - S_{Bi})^2}$$

where: S_{Ai} is the number of individuals in the *i*th OTU of community A and S_{Bi} is the equivalent for community B

The Smith theta similarity coefficient:

$$S\Theta = 1 - \frac{(\sum_{i=1}^{T} a_i)(\sum_{i=1}^{T} b_i)}{(\sum_{i=1}^{T} a_i) + (\sum_{i=1}^{T} b_i) - (\sum_{i=1}^{T} a_i)(\sum_{i=1}^{T} b_i)}$$

where: T represents the total number of shared OTUs; a_i is the relative abundance of OTU i in community A and b_i is the relative abundance of OTU i in community B

The Yue & Clayton theta similarity coefficient:

$$YC\Theta = 1 - \frac{\sum_{i=1}^{T} a_i\, b_i}{\sum_{i=1}^{T} (a_i - b_i)^2 + \sum_{i=1}^{T} a_i\, b_i}$$

where: T represents the total number of shared OTUs; a_i is the relative abundance of OTU i in community A and b_i is the relative abundance of OTU i in community B

The Jaccard similarity coefficient (abundance-adjusted):

$$JA = \frac{UV}{U + V - UV}$$

$$U = \sum_{i=1}^{T} \frac{A_i}{n} + \frac{m-1}{m}\,\frac{z_1}{2z_2} \sum_{i=1}^{T} \frac{A_i}{n}\, I(B_i = 1)$$

$$V = \sum_{i=1}^{T} \frac{B_i}{m} + \frac{n-1}{n} \frac{y_1}{2y_2} \sum_{i=1}^{T} \frac{B_i}{m} I(A_i = 1)$$

where: T represents the total number of shared OTUs; A_i is the number of individuals in OTU i of community A; B_i is the number of individuals in OTU i of community B; n is the total number of individuals in community A; m is the total number of individuals in community B; z_1 and z_2 are the number of shared OTUs with one or two individuals respectively in community A; y_1 and y_2 are the number of shared OTUs with one or two individuals respectively in community B; and I represents a function whose output is 1 if the following expression is true and zero if the following expression is untrue

The Sørensen similarity coefficient (abundance-adjusted):

$$SA = \frac{2UV}{U+V}$$

$$U = \sum_{i=1}^{T} \frac{A_i}{n} + \frac{m-1}{m} \frac{z_1}{2z_2} \sum_{i=1}^{T} \frac{A_i}{n} I(B_i = 1)$$

$$V = \sum_{i=1}^{T} \frac{B_i}{m} + \frac{n-1}{n} \frac{y_1}{2y_2} \sum_{i=1}^{T} \frac{B_i}{m} I(A_i = 1)$$

where: T represents the total number of shared OTUs; A_i is the number of individuals in OTU i of community A; B_i is the number of individuals in OTU i of community B; n is the total number of individuals in community A; m is the total number of individuals in community B; z_1 and z_2 are the number of shared OTUs with one or two individuals respectively in community A; y_1 and y_2 are the number of shared OTUs with one or two individuals respectively in community B; and I represents a function whose output is 1 if the following expression is true and zero if the following expression is untrue.

Fisher's alpha diversity index:

$$S=\alpha*\ln(1+n/\alpha)$$

where S is number of taxa, *n* is number of individuals and α is the Fisher's alpha.

APPENDIX 10: STATISTICS

Unless otherwise stated, definitions for statistics were obtained from Biostatistical Analysis (Zar, 1996).

Coefficient of variation = $\sigma/\mu \times 100\%$ where σ is the standard error of the mean and μ is the mean value.

Library compare RDP: One underlying assumption for this equation is that x and y are small relative to N1 and N2 (less than 5% of the total), and N1 and N2 are relatively large (above 500). The probability of the observed difference in assignment to taxon T is estimated as:

$$p(y|x) = \left(\frac{N_2}{N_1}\right)^y \frac{(x+y)!}{x!\,y!\left(1+\frac{N_2}{N_1}\right)^{(x+y+1)}}$$

where N_1 and N_2 are the total number of sequences for library 1 and 2 respectively, and x and y are the number of sequences assigned to T from library 1 and 2 respectively (Wang *et al*., 2007; Audic and Claverie, 1997)

Library compare from RDP: For taxa with greater than five sequences assigned, the standard two-population proportions test is used to estimate the probability of the observed differences (Wang *et al*., 2007). The P value is estimated from the Z critical value, where :

$$Z = \frac{\frac{x}{N_1} - \frac{y}{N_2}}{\sqrt{\mu(1-\mu)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$$

where and where N_1 and N_2 are the total number of sequences for library 1 and 2, respectively, x and y are the number of sequences assigned to taxon T from library 1 and 2, respectively, and μ equals $(x + y)/(N_1 + N_2)$

Confidence interval of a mean: CI = μ - (s.e.m × t) to μ + (s.e.m × t)

where $\mathrm{s.e.m} = \frac{\sigma}{\sqrt{\mathrm{n}}}$, s.e.m stands for the standard error of the mean, σ is the standard deviation of the sample, and n is the sample size. The critical value of t depends on the sample size minus one (so-called degrees of freedom), and on the CI you want to calculate (e.g. 95% CI with $P < 0.05$).

Student's T-test for the difference betweeen 2 means:

$$t = \frac{\bar{A}_1 + \bar{A}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{A}_1$ and $\bar{A}_2$ are the means of the two populations; s_1 and s_2 are the standard deviations of the two populations; and n_1 and n_2 are the respective sample sizes.

Analysis of Variance (ANOVA):

k is the number of populations; *n* is the total number of observations; *μ* is the mean of all *n* observations; *nj* is the size of sample from Population *j*; *μj* is the mean of sample from Population *j*; sj^2 variance of sample from Population *j*; T*j* sum of sample data from Population *j; x* is any given value of the *n* observations.

$\mathrm{SST} = \Sigma (x - \mu)^2$ $\mathrm{SSTR} = \Sigma\, nj\, (\mu j - \mu)^2$ $\mathrm{SSE} = \Sigma\, (nj - 1)\, sj^2$

One-way ANOVA identity: SST = SSTR + SSE

$\mathrm{SST} = \Sigma\, x^2 - (\Sigma\, x)^2/n$ $\mathrm{SSTR} = \Sigma\, (\mathrm{T}j^2/nj) - (\Sigma\, x)^2/n$ $\mathrm{SSE} = \mathrm{SST} - \mathrm{SSTR}$

$\mathrm{MSTR} = \mathrm{SSTR}/(k - 1)$ $\mathrm{MSE} = \mathrm{SSE}/(n - k)$

F = MSTR/MSE with df = $(k - 1, n - k)$

Logit transformation: logit $x = \ln[x/(1\text{-}x)]$

G-test of independence:

$$G = 2 \sum_i O_i \cdot ln\left(\frac{O_i}{E_i}\right)$$

where O_i is the observed frequency and E_i is the expected frequency

Metastats

For a number of subjects (n_t) in each treatment (t), and the proportion (f_{ij}) of a given taxon (i) in individual j, the mean value is calculated:

$$\bar{x}_{it} = \frac{1}{n_t} \sum_{1}^{j} f_{ij}$$

along with the variance:

$$s_{it}^2 = \frac{1}{(n_t - 1)} \sum_{1}^{j} \left(f_{ij} - \bar{x}_{it}\right)^2$$

prior to application of a T-test with comparison of values to critical tabulated values. Multiple testing correction is applied using the FDR (false discovery rate) method

APPENDIX 11: MOTHUR COMMANDS FOR RUN 1

Text in red indicate regions where the matrix/group filename needs to be inserted; text in blue indicate where the filename of the original reads needs to be inserted.

```
cd Desktop\MOTHUR ROOT\mothur

mothur

cluster(phylip=filename.dist, cutoff=0.5, precision=100,
method=furthest)

bin.seqs(list=filename.fn.list, fasta=filename.fna
group=filename.groups)

make.shared(list=filename.fn.list, group=filename.groups)

get.oturep(phylip=filename.dist, fasta=filename.fna,
list=filename.fn.list, group=filename.groups,
groups=groups_required_separated_by_dashes)

classify.seqs(fasta=filename.fasta, template=
nogap.bacteria.fasta, taxonomy= silva.bacteria.silva.tax)

classify.otu(taxonomy=filename.silva.tax, list=filename.fn.list,
group=filename.groups)

collect.single(shared=filename.fn.shared, calc=sobs-chao-ace-
jack-bootstrap-simpsoneven-bergerparker-shannon-npshannon-
simpson-invsimpson-coverage-qstat-boneh-logseries-geometric-
bstick)

rarefaction.single(shared=filename.fn.shared, calc=sobs-chao-
ace)

summary.single(shared=filename.fn.shared, calc=sobs-chao-ace-
jack-bootstrap-simpsoneven-bergerparker-shannon-npshannon-
simpson-invsimpson-coverage-qstat-boneh-logseries-geometric-
bstick)

collect.shared(shared=filename.fn.shared, calc=sharedsobs-
sharedchao-sharedace, groups=all)


rarefaction.shared(shared=filename.fn.shared, calc=sharedsobs-
sharedchao-sharedace, groups=all)

summary.shared(shared=filename.fn.shared, calc=sharedsobs-
sharedchao-sharedace-anderberg-hamming-jclass-jest-kulczynski-
kulczynskicody-lennon-memchi2-memchord-memeuclidean-mempearson-
ochiai-sorclass-sorest-whittaker-braycurtis-canberra-gower-
hellinger-jabund-manhattan-morisitahorn-odum-soergel-sorabund-
spearman-speciesprofile-structchi2-structchord-structeuclidean-
structkulczynski-structpearson-thetan-thetayc, groups=all)
```

```
get.sharedseqs(list=filename.fn.list, group=filename.groups)

heatmap.bin(shared=filename.fn.shared, sorted=none,
scale=linear)

tree.shared(shared=filename.fn.shared, groups=
groups_required_separated_by_dashes)

metastats(shared=filename.fn.shared, design=filename.design.txt)

libshuff(phylip=filename.dist, group=filename.groups)

parsimony(tree=filename.fn.distance.tre,
group=filename.design.txt)

unifrac.weighted(tree=filename.fn.distance.tre,
group=filename.design.txt, random=t)

dist.shared(shared=filename.fn.shared)

pcoa(phylip=filename.fn.distance.lt.dist)

nmds(phylip=filename.fn.distance.lt.dist)

amova(phylip=filename.fn.distance.lt.dist,
design=filename.design.txt)

corr.axes(axes=filename.fn.distance.lt.nmds.axes,
shared=filename.fn.shared, method=spearman, numaxes=2)
```

APPENDIX 12: 'R' SCRIPT FOR RUN 2

The following script was adapted from code written in the 'R' coding language by Dr Kelvin Lau and provides heatmaps and PCA plots from input OTU data. Sections in green indicate variables specific to the input data set.

```
source("http://bioconductor.org/biocLite.R")
biocLite()
install.packages("gplots", dependencies = TRUE)

PathPrefix <- "C:/Users/Adam/Desktop/R_ROOT/454Data/Input/"
PathSuffix <- "_trimmed.fasta_download.txt"

SampleNames <- c( "S5rep1", "S5rep2", "S5rep3", "S5rep4",
"S5rep5", "S5rep6", "S5rep7", "S5rep8", "S5rep9", "S4",  "Cj1")

Files <- paste(PathPrefix, SampleNames, PathSuffix, sep="")

for (j in 1:length(Files)) {
      temp <- read.delim(Files[j], skip=19, sep=";", header=F,
na.strings=NA, fill=T)
      temp <- temp[grep("GJVUD7X05", temp[,1]),1:16]
      colnames(temp) <- c("Sequence", "Sample")
      temp[,2] <- SampleNames[j]
      assign(SampleNames [j], temp)
}

CombinedSamples <- rbind(S5rep1, S5rep2, S5rep3, S5rep4, S5rep5,
S5rep6, S5rep7, S5rep8, S5rep9, S4,  Cj1)

CombinedSamples2 <- CombinedSamples
CombinedSamples2[,4] <- as.numeric(sub("%", "",
as.character(levels(CombinedSamples[,4])[as.integer(CombinedSamp
les[,4])])))/100
CombinedSamples2[,6] <- as.numeric(sub("%", "",
as.character(levels(CombinedSamples[,6])[as.integer(CombinedSamp
les[,6])])))/100
CombinedSamples2[,8] <- as.numeric(sub("%", "",
as.character(levels(CombinedSamples[,8])[as.integer(CombinedSamp
les[,8])])))/100
CombinedSamples2[,10] <- as.numeric(sub("%", "",
as.character(levels(CombinedSamples[,10])[as.integer(CombinedSam
ples[,10])])))/100
CombinedSamples2[,12] <- as.numeric(sub("%", "",
as.character(levels(CombinedSamples[,12])[as.integer(CombinedSam
ples[,12])])))/100
CombinedSamples2[,14] <- as.numeric(sub("%", "",
as.character(levels(CombinedSamples[,14])[as.integer(CombinedSam
ples[,14])])))/100
CombinedSamples2[,16] <- as.numeric(sub("%", "",
as.character(levels(CombinedSamples[,16])[as.integer(CombinedSam
ples[,16])])))/100
```

```
colnames(CombinedSamples2) <- c("Sequence", "Sample",
"RootName", "RootConf", "KingdomName", "KingdomConf",
"PhylumName", "PhylumConf", "ClassName", "ClassConf",
"OrderName", "OrderConf", "FamilyName", "FamilyConf",
"GenusName", "GenusConf")



#Function 1

CIStats <- function(CombinedSamples, SampleNames, Confidence =
0.95, perc = T, includeArchaea = T) {
ReturnTable <- NULL
for (i in 1:length(SampleNames)) {
     Data <-
data.frame(CombinedSamples[CombinedSamples[,"Sample"] ==
SampleNames[i],])

     TotalLength <- length(Data[,1])
     PhylumLength <- length(Data[Data[,8] >= Confidence, 1])
     ClassLength <- length(Data[Data[,10] >= Confidence, 1])
     OrderLength <- length(Data[Data[,12] >= Confidence, 1])
     FamilyLength <- length(Data[Data[,14] >= Confidence, 1])
     GenusLength <- length(Data[Data[,16] >= Confidence, 1])

     if (perc == T) {
          ReturnRow <- c(TotalLength, PhylumLength /
TotalLength, ClassLength / TotalLength, OrderLength /
TotalLength, FamilyLength / TotalLength, GenusLength/
TotalLength)
     } else {
          ReturnRow <- c(TotalLength, PhylumLength,
ClassLength, OrderLength, FamilyLength, GenusLength)
     }

     ReturnTable <- rbind(ReturnTable, ReturnRow)

}
     return(ReturnTable)
}


SummaryStats50 <- CIStats(CombinedSamples2, SampleNames,
Confidence = 0.50, perc =T)
rownames(SummaryStats50) <- SampleNames
ColumnNames <- c("Total Sequences", "Phylum", "Class", "Order",
"Family", "Genus")
colnames(SummaryStats50) <- ColumnNames
write.table(SummaryStats50,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/CI50.txt",
sep="\t")

SummaryStats80 <- CIStats(CombinedSamples2, SampleNames,
Confidence = 0.80, perc =T)
rownames(SummaryStats80) <- SampleNames
```

```
ColumnNames <- c("Total Sequences", "Phylum", "Class", "Order",
"Family", "Genus")
colnames(SummaryStats80) <- ColumnNames
write.table(SummaryStats80,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/CI80.txt",
sep="\t")

SummaryStats95 <- CIStats(CombinedSamples2, SampleNames,
Confidence = 0.95, perc =T)
rownames(SummaryStats95) <- SampleNames
ColumnNames <- c("Total Sequences", "Phylum", "Class", "Order",
"Family", "Genus")
colnames(SummaryStats95) <- ColumnNames
write.table(SummaryStats95,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/CI95.txt",
sep="\t")

ArchaeaNames <- CombinedSamples2[gsub("(^ +)|( +$)",
"",CombinedSamples2[,5]) == "Archaea",]
write.table(ArchaeaNames,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/ArchaeaNames.txt",
sep="\t")

BacteriaBelow95Names <- CombinedSamples2[gsub("(^ +)|( +$)",
"",CombinedSamples2[,5]) == "Bacteria" & CombinedSamples2[,6] <
0.95,]
write.table(BacteriaBelow95Names,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/BacteriaBelow95Name
s.txt", sep="\t")

RemoveSeq <- rbind(ArchaeaNames, BacteriaBelow95Names)
RemoveSeqNames <- RemoveSeq[,"Sequence"]
write.table(RemoveSeqNames,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/Working.RemoveSeq.a
ccnos", sep="\t")

CombinedSamples2.filtered <- CombinedSamples2[gsub("(^ +)|(
+$)", "",CombinedSamples2[,5]) != "Archaea",]
CombinedSamples2.filtered <-
CombinedSamples2.filtered[CombinedSamples2.filtered[,6] >=
0.95,]

SummaryStats50.filtered <- CIStats(CombinedSamples2.filtered,
SampleNames, Confidence = 0.50, perc =T)
rownames(SummaryStats50.filtered) <- SampleNames
ColumnNames <- c("Total Sequences", "Phylum", "Class", "Order",
"Family", "Genus")
colnames(SummaryStats50.filtered) <- ColumnNames
write.table(SummaryStats50.filtered,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/CI50.filtered.txt",
sep="\t")

SummaryStats80.filtered <- CIStats(CombinedSamples2.filtered,
SampleNames, Confidence = 0.80, perc =T)
rownames(SummaryStats80.filtered) <- SampleNames
ColumnNames <- c("Total Sequences", "Phylum", "Class", "Order",
"Family", "Genus")
```

```
colnames(SummaryStats80.filtered) <- ColumnNames
write.table(SummaryStats80.filtered,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/CI80.filtered.txt",
sep="\t")

SummaryStats95.filtered <- CIStats(CombinedSamples2.filtered,
SampleNames, Confidence = 0.95, perc =T)
rownames(SummaryStats95.filtered) <- SampleNames
ColumnNames <- c("Total Sequences", "Phylum", "Class", "Order",
"Family", "Genus")
colnames(SummaryStats95.filtered) <- ColumnNames
write.table(SummaryStats95.filtered,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/CI95.filtered.txt",
sep="\t")

#Function 2

AbundanceTable <- function(DataTable, SampleNames, DataCol,
ListNames = unique(DataTable[,DataCol]), perc = T) {
      ReturnTable <- NULL
      CountTable <- NULL

      for (i in 1:length(SampleNames)) {
            Data <- DataTable[ DataTable[,"Sample"]
      ==SampleNames[i], ]
            Data <- Data[!is.na(Data[, DataCol]),]
            TotalLength <- length(Data[,1])
            NewRow <- NULL
            for (j in 1:length(ListNames)) {
                  temp <- Data[Data[,DataCol] == ListNames[j], ]
                  temp <- temp[!is.na(temp[, DataCol]),]
                  NewRow <- cbind(NewRow, length(temp[,1]))
            }
            CountTable <- rbind(CountTable, NewRow)
            if (perc == T) {
                  NewRow <- NewRow / TotalLength
            }
            Total <- sum(NewRow)
            NewRow <- cbind(NewRow, Total, TotalLength)
            ReturnTable <- rbind(ReturnTable, NewRow)
      }
      CountRow <- NULL
      for (j in 1:length(CountTable[1,])) {
            CountRow <- cbind(CountRow, sum(CountTable[,j]))
      }
      CountRow <- cbind(CountRow, sum(ReturnTable[,
length(ReturnTable[1,])-1]), sum(ReturnTable[,
length(ReturnTable[1,])]) )
      ReturnTable <- rbind(ReturnTable, CountRow)
      colnames(ReturnTable) <- c(as.character(ListNames),
"Sum", "Total number of sequences")
      rownames(ReturnTable) <- c(SampleNames, "Total Counts")
      return(ReturnTable)
}
```

```
CombinedSamples2.p <-
CombinedSamples2.filtered[CombinedSamples2.filtered[, 8] >=
0.50, ]
phylumCount <- AbundanceTable(CombinedSamples2.p, SampleNames,
7)
write.table(phylumCount,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/phylumCount.txt",
sep="\t")

CombinedSamples2.c <-
CombinedSamples2.filtered[CombinedSamples2.filtered[, 10] >=
0.50, ]
classCount <- AbundanceTable(CombinedSamples2.c, SampleNames, 9)
write.table(classCount,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/classCount.txt",
sep="\t")

CombinedSamples2.o <-
CombinedSamples2.filtered[CombinedSamples2.filtered[, 12] >=
0.50, ]
orderCount <- AbundanceTable(CombinedSamples2.o, SampleNames,
11)
write.table(orderCount,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/orderCount.txt",
sep="\t")

CombinedSamples2.f <-
CombinedSamples2.filtered[CombinedSamples2.filtered[, 14] >=
0.50, ]
familyCount <- AbundanceTable(CombinedSamples2.f, SampleNames,
13)
write.table(familyCount,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/familyCount.txt",
sep="\t")

CombinedSamples2.g <-
CombinedSamples2.filtered[CombinedSamples2.filtered[, 16] >=
0.50, ]
genusCount <- AbundanceTable(CombinedSamples2.g, SampleNames,
15)
write.table(genusCount,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/genusCount.txt",
sep="\t")

phylumCount.f <- phylumCount[, -c(length(phylumCount[1,]),
length(phylumCount[1,]) -1) ]
phylumCount.f <- phylumCount.f[,
phylumCount.f[length(phylumCount.f[,1]) , ] >0.002]
temp <- phylumCount.f
temp1 <- as.matrix(temp)
temp2 <- temp1[nrow(temp1),]
temp3 <- sort(temp2, decreasing=TRUE, index.return=TRUE)$ix
phylumCount.f <- as.data.frame(temp[,temp3])
phylumCount.f <- phylumCount.f[ -(length(phylumCount.f[,1])),]
phylumCount.f.ln <- log((phylumCount.f + 0.00001)/(1-
phylumCount.f))
```

```
phylumCount.f.lg <- log2((phylumCount.f + 0.00001)/(1-
phylumCount.f))
write.table(phylumCount.f,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/phylumCount.f.txt",
sep="\t")
write.table(phylumCount.f.ln,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/phylumCount.f.ln.tx
t", sep="\t")
write.table(phylumCount.f.lg,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/phylumCount.f.lg.tx
t", sep="\t")

classCount.f <- classCount[, -c(length(classCount[1,]),
length(classCount[1,]) -1) ]
classCount.f <- classCount.f[,
classCount.f[length(classCount.f[,1]) , ] >0.002]
temp <- classCount.f
temp1 <- as.matrix(temp)
temp2 <- temp1[nrow(temp1),]
temp3 <- sort(temp2, decreasing=TRUE, index.return=TRUE)$ix
classCount.f <- as.data.frame(temp[,temp3])
classCount.f <- classCount.f[ -(length(classCount.f[,1])),]
classCount.f.ln <- log((classCount.f + 0.00001)/(1-
classCount.f))
classCount.f.lg <- log2((classCount.f + 0.00001)/(1-
classCount.f))
write.table(classCount.f,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/classCount.f.txt",
sep="\t")
write.table(classCount.f.ln,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/classCount.f.ln.txt
", sep="\t")
write.table(classCount.f.lg,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/classCount.f.lg.txt
", sep="\t")


orderCount.f <- orderCount[, -c(length(orderCount[1,]),
length(orderCount[1,]) -1) ]
orderCount.f <- orderCount.f[,
orderCount.f[length(orderCount.f[,1]) , ] >0.002]
temp <- orderCount.f
temp1 <- as.matrix(temp)
temp2 <- temp1[nrow(temp1),]
temp3 <- sort(temp2, decreasing=TRUE, index.return=TRUE)$ix
orderCount.f <- as.data.frame(temp[,temp3])
orderCount.f <- orderCount.f[ -(length(orderCount.f[,1])),]
orderCount.f.ln <- log((orderCount.f + 0.00001)/(1-
orderCount.f))
orderCount.f.lg <- log2((orderCount.f + 0.00001)/(1-
orderCount.f))
write.table(orderCount.f,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/orderCount.f.txt",
sep="\t")
write.table(orderCount.f.ln,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/orderCount.f.ln.txt
", sep="\t")
```

```
write.table(orderCount.f.lg,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/orderCount.f.lg.txt
", sep="\t")


familyCount.f <- familyCount[, -c(length(familyCount[1,]),
length(familyCount[1,]) -1) ]
familyCount.f <- familyCount.f[,
familyCount.f[length(familyCount.f[,1]) , ] >0.002]
temp <- familyCount.f
temp1 <- as.matrix(temp)
temp2 <- temp1[nrow(temp1),]
temp3 <- sort(temp2, decreasing=TRUE, index.return=TRUE)$ix
familyCount.f <- as.data.frame(temp[,temp3])
familyCount.f <- familyCount.f[ -(length(familyCount.f[,1])),]
familyCount.f.ln <- log((familyCount.f + 0.00001)/(1-
familyCount.f))
familyCount.f.lg <- log2((familyCount.f + 0.00001)/(1-
familyCount.f))
write.table(familyCount.f,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/familyCount.f.txt",
sep="\t")
write.table(familyCount.f.ln,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/familyCount.f.ln.tx
t", sep="\t")
write.table(familyCount.f.lg,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/familyCount.f.lg.tx
t", sep="\t")

genusCount.f <- genusCount[, -c(length(genusCount[1,]),
length(genusCount[1,]) -1) ]
genusCount.f <- genusCount.f[,
genusCount.f[length(genusCount.f[,1]) , ] >0.002]
temp <- genusCount.f
temp1 <- as.matrix(temp)
temp2 <- temp1[nrow(temp1),]
temp3 <- sort(temp2, decreasing=TRUE, index.return=TRUE)$ix
genusCount.f <- as.data.frame(temp[,temp3])
genusCount.f <- genusCount.f[ -(length(genusCount.f[,1])),]
genusCount.f.ln <- log((genusCount.f + 0.00001)/(1-
genusCount.f))
genusCount.f.lg <- log2((genusCount.f + 0.00001)/(1-
genusCount.f))
write.table(genusCount.f,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/genusCount.f.txt",
sep="\t")
write.table(genusCount.f.ln,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/genusCount.f.ln.txt
", sep="\t")
write.table(genusCount.f.lg,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/genusCount.f.lg.txt
", sep="\t")

familyCount.f.lne <- log(familyCount.f + 0.00001)
write.table(familyCount.f.lne,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/familyCount.f.lne.t
xt", sep="\t")
```

```
familyCount.f.lg10 <- log10(familyCount.f + 0.00001)
write.table(familyCount.f.lg10,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/familyCount.f.lg10.
txt", sep="\t")

genusCount.f.lne <- log(genusCount.f + 0.00001)
write.table(genusCount.f.lne,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/genusCount.f.lne.tx
t", sep="\t")
genusCount.f.lg10 <- log10(genusCount.f + 0.00001)
write.table(genusCount.f.lg10,
"C:/Users/Adam/Desktop/R_ROOT/454Data/Output/genusCount.f.lg10.t
xt", sep="\t")

#Heatmaps 1

library(marray)

heatmap(t(familyCount.f[,1:13]), Rowv=NA, Colv=T,
col=maPalette(20, high="yellow", mid="red", low="black"),
scale="none", breaks=seq(from=0, to=1, length=21), na.rm=T,
margins=c(6,6), cexCol=1)

heatmap(t(genusCount.f[,1:15]), Rowv=NA, Colv=T,
col=maPalette(100, high="yellow", mid="red", low="black"),
scale="none", breaks=seq(from=0, to=1, length=101), na.rm=T ,
margins=c(8,8), cexCol=1)

col <- maPalette(100,high="yellow", mid="red", low="black")
breaks=seq(from=0, to=1,length=101)
z <- seq(0, 1, length = length(col))
        image(z = matrix(z, ncol = 1), col = col, breaks =
breaks,
            xaxt = "n", yaxt = "n",
xlab=expression(paste(Proportional, " Abundance")))
box()
        axis(1, at = 0, labels = 0.1)
        axis(1, at = 0.22, labels = 0.25)
        axis(1, at = 0.5, labels = 0.5)
        axis(1, at = 0.78, labels = 0.75)
        axis(1, at = 1, labels = 0.9)

heatmap(t(familyCount.f.ln[,1:13]), Rowv=NA, Colv=T,
col=maPalette(100, high="yellow", mid="green", low="blue"),
scale="none", breaks=seq(from=-12, to=2, length=101), na.rm=T,
margins=c(6,6), cexCol=1)

heatmap(t(genusCount.f.ln[,1:15]), Rowv=NA, Colv=T,
col=maPalette(100, high="yellow", mid="green", low="blue"),
scale="none", breaks=seq(from=-12, to=2, length=101), na.rm=T,
margins=c(6,6), cexCol=1)

col <- maPalette(100, high="yellow", mid="green", low="blue")
breaks=seq(from=-12, to=2, length=101)
z <- seq(-12, 2, length = length(col))
        image(z = matrix(z, ncol = 1), col = col, breaks =
breaks,
```

```
            xaxt = "n", yaxt = "n", xlab=expression(paste(Logit-
transformed, " Abundance")))
box()
        axis(1, at = 0, labels = -12)
        axis(1, at = 0.22, labels = -9)
        axis(1, at = 0.5, labels = -5)
        axis(1, at = 0.78, labels = -1)
        axis(1, at = 1, labels = 2)

heatmap(t(genusCount.f.lg[,1:15]), Rowv=NA, Colv=T,
col=maPalette(128, high="yellow", mid="green", low="blue"),
scale="none", breaks=seq(from=-17, to=3, length=129), na.rm=T,
margins=c(6,6), cexCol=1)

heatmap(t(familyCount.f.lg[,1:13]), Rowv=NA, Colv=T,
col=maPalette(128, high="yellow", mid="green", low="blue"),
scale="none", breaks=seq(from=-17, to=3, length=129), na.rm=T,
margins=c(6,6), cexCol=1)

col <- maPalette(128, high="yellow", mid="green", low="blue")
breaks=seq(from=-17, to=3,length=129)
z <- seq(-17, 0, length = length(col))
        image(z = matrix(z, ncol = 1), col = col, breaks =
breaks,
            xaxt = "n", yaxt = "n", xlab=expression(paste(Log2,
" Abundance")))
box()
        axis(1, at = 0, labels = -17)
        axis(1, at = 0.22, labels = -12)
        axis(1, at = 0.5, labels = -7)
        axis(1, at = 0.78, labels =-2)
        axis(1, at = 1, labels =3)

heatmap(t(familyCount.f.lg10[,1:13]), Rowv=NA, Colv=T,
col=maPalette(20, high="yellow", mid="red", low="black"),
scale="none", breaks=seq(from=-5, to=0, length=21), na.rm=T,
margins=c(6,6), cexCol=1)

heatmap(t(genusCount.f.lg10[,1:15]), Rowv=NA, Colv=T,
col=maPalette(100, high="yellow", mid="red", low="black"),
scale="none", breaks=seq(from=-5, to=0, length=101), na.rm=T ,
margins=c(8,8), cexCol=1)

col <- maPalette(100,high="yellow", mid="red", low="black")
breaks=seq(from=-5, to=0,length=101)
z <- seq(-5, 0, length = length(col))
        image(z = matrix(z, ncol = 1), col = col, breaks =
breaks,
            xaxt = "n", yaxt = "n", xlab=expression(paste(Log10,
" Abundance")))
box()
        axis(1, at = 0, labels = -5)
        axis(1, at = 0.22, labels = -3.75)
        axis(1, at = 0.5, labels = -2.5)
        axis(1, at = 0.78, labels = -1.25)
        axis(1, at = 1, labels = 0)
```

```
heatmap(t(genusCount.f.lne[,1:15]), Rowv=NA, Colv=T,
col=maPalette(128, high="yellow", mid="green", low="blue"),
scale="none", breaks=seq(from=-12, to=0, length=129), na.rm=T,
margins=c(6,6), cexCol=1)

heatmap(t(familyCount.f.lne[,1:13]), Rowv=NA, Colv=T,
col=maPalette(128, high="yellow", mid="green", low="blue"),
scale="none", breaks=seq(from=-12, to=0, length=129), na.rm=T,
margins=c(6,6), cexCol=1)

col <- maPalette(100, high="yellow", mid="green", low="blue")
breaks=seq(from=-12, to=0, length=101)
z <- seq(-12, 0, length = length(col))
        image(z = matrix(z, ncol = 1), col = col, breaks =
breaks,
            xaxt = "n", yaxt = "n",
xlab=expression(paste(Natural-log, " Abundance")))
box()
        axis(1, at = 0, labels = -12)
        axis(1, at = 0.22, labels = -9)
        axis(1, at = 0.5, labels = -6)
        axis(1, at = 0.78, labels = -3)
        axis(1, at = 1, labels = 0)


#PCA

genusCount.f.pca <- prcomp(genusCount.f, retx=T)
genusCount.f.pca$x
genusCount.f.pca$sdev
library(MASS)
eqscplot(genusCount.f.pca$x[,1:2], main="")
text(genusCount.f.pca$x[,1:2], rownames(genusCount.f),cex=0.75,
col="red")

genusCount.f.pcat <- prcomp(t(genusCount.f[,1:2]), retx=T,
scale=T)
genusCount.f.pcat$x
genusCount.f.pcat$sdev
library(MASS)
eqscplot(genusCount.f.pcat$x[,1:2], main="")
text(genusCount.f.pcat$x[,1:2],colnames(genusCount.f[1:11]),cex=
0.5)

genusCount.f.ln.pca <- prcomp(genusCount.f.ln, retx=T)
genusCount.f.ln.pca$x
genusCount.f.ln.pca$sdev
library(MASS)
eqscplot(genusCount.f.ln.pca$x[,1:2], main="")
text(genusCount.f.ln.pca$x[,1:2],
rownames(genusCount.f.ln),cex=0.75, col="red")
```

APPENDIX 13: QIIME COMMANDS FOR RUN 3

Using Virtual Box and Qiime1.4.0

```
wget
http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_file
s/core_set_aligned.fasta.imputed

wget
http://greengenes.lbl.gov/Download/Sequence_Data/lanemask_in_1s_
and_0s

check_id_map.py -m Run3_mapping.txt -o mapping_output/

pick_otus.py -i split_library_output/seqs.fna -m uclust -o otus

pick_rep_set.py -i otus/seqs_otus.txt -f
split_library_output/seqs.fna -o rep_set.fna

align_seqs.py -i rep_set.fna -t core_set_aligned.fasta.imputed -
o pynast_aligned/

assign_taxonomy.py -i rep_set.fna -m rdp -c 0.5

identify_chimeric_seqs.py -m ChimeraSlayer -i
rep_set_aligned.fasta -a core_set_aligned.fasta.imputed -o
chimeric_seqs.txt

filter_fasta.py -f rep_set_aligned.fasta -o
non_chimeric_rep_set_aligned.fasta -s chimeric_seqs.txt -n

filter_alignment.py -i non_chimeric_rep_set_aligned.fasta -m
lanemask_in_1s_and_0s -o filtered_alignment/

make_phylogeny.py -i
non_chimeric_rep_set_aligned_pfiltered.fasta -o rep_phylo.tre

make_otu_table.py -i seqs_otus.txt -o otu_table.txt -e
chimeric_seqs.txt -t rep_set_tax_assignments.txt

per_library_stats.py -i otus/otu_table.txt

make_otu_heatmap_html.py -i otus/otu_table.txt -o
otus/OTU_Heatmap/

#For Cytoscape

make_otu_network.py -m Run3_mapping.txt -i otus/otu_table.txt -o
otus/OTU_Network -b Treatment

summarize_taxa.py -i otu_table2.txt -L 2 -o ./Phylum

summarize_taxa.py -i otu_table2.txt -L 3 -o ./Class

summarize_taxa.py -i otu_table2.txt -L 4 -o ./Order
```

```
summarize_taxa.py -i otu_table2.txt -L 5 -o ./Family

summarize_taxa.py -i otu_table2.txt -L 6 -o ./Genus

#Numbers refer to the RDP classifier taxonomy, Level 1 = Domain
(e.g. Bacteria), 2 = Phylum (e.g. Firmicutes), 3 = Class (e.g.
Clostridia), 4 = Order (e.g. Clostridiales), 5 = Family (e.g.
Clostridiaceae), and 6 = Genus (e.g. Clostridium)

plot_taxa_summary.py -l phylum,class,order,family,genus -i
For_summary/otu_table2_L2.txt,For_summary/otu_table2_L3.txt,For_
summary/otu_table2_L4.txt,For_summary/otu_table2_L5.txt,For_summ
ary/otu_table2_L6.txt -c pie,bar,area -o output_charts/ -n 15 -s

summarize_taxa_through_plots.py -o wf_taxa_sum -i otu_table2.txt
-m Run4_mapping2.txt -c Treatment

summarize_otu_by_cat.py -i Run3_mapping.txt -c otu_table.txt -m
'ColumnHeader'

summarize_taxa.py -L 2 -i ColumnHeader_otu_table.txt -o
ColumnHeader_otu_table_summarized.txt

plot_taxa_summary.py -i ColumnHeader_otu_table_summarized.txt -l
Phylum -o Summary_Graphs/ -c bar,pie,area

multiple_rarefactions.py -i otu_table2.txt -m 500 -x 5000 -s 500
-n 3 -o rarefaction_tables/

alpha_diversity.py -i rarefaction_tables/ -m
ACE,berger_parker_d,brillouin_d,chao1,chao1_confidence,dominance
,doubles,equitability,fisher_alpha,heip_e,kempton_taylor_q,marga
lef,mcintosh_d,mcintosh_e,menhinick,michaelis_menten_fit,observe
d_species,osd,reciprocal_simpson,robbins,shannon,simpson,simpson
_e,singles,strong -o alpha_div/

collate_alpha.py -i alpha_div/ -o collated_alpha/

make_rarefaction_plots.py -i collated_alpha/ -m Run3_mapping.txt
-b 'ColumnHeader' -g png

Metrics =
abund_jaccard,binary_chisq,binary_chord,binary_euclidean,binary_
hamming,binary_jaccard,binary_lennon,binary_ochiai,binary_otu_ga
in,binary_pearson,binary_sorensen_dice,bray_curtis,canberra,chis
q,chord,euclidean,gower,hellinger,kulczynski,manhattan,morisita_
horn,pearson,soergel,spearman_approx,specprof,unifrac,unifrac_g,
unifrac_g_full_tree,unweighted_unifrac,unweighted_unifrac_full_t
ree,weighted_normalized_unifrac,weighted_unifrac

single_rarefaction.py -i otu_table2.txt -o
rarefaction_5000_19.txt -d 5000

make_prefs_file.py -m Run3_mapping.txt -i
OTU_Levels/otu_table2_L6.txt -o prefs_out.txt
```

```
beta_diversity.py -i rarefaction_5000_19.txt -m
abund_jaccard,binary_chisq,binary_chord,binary_euclidean,binary_
hamming,binary_jaccard,binary_lennon,binary_ochiai,binary_otu_ga
in,binary_pearson,binary_sorensen_dice,bray_curtis,canberra,chis
q,chord,euclidean,gower,hellinger,kulczynski,manhattan,morisita_
horn,pearson,soergel,spearman_approx,specprof,unifrac,unifrac_g,
unifrac_g_full_tree,unweighted_unifrac,unweighted_unifrac_full_t
ree,weighted_normalized_unifrac,weighted_unifrac -o beta_div/ -t
repr_set.tre

principal_coordinates.py -i beta_div/ -o beta_div_coords/

make_3d_plots.py -i
beta_div_PCoA_coords/pcoa_weighted_unifrac_rarefaction_5000_19.t
xt -m Run3_mapping.txt -t OTU_Levels/otu_table2_L6.txt --
biplot_output_file biplot.txt --n_taxa_keep=10

make_2d_plots.py -i
beta_div_PCoA_coords/pcoa_bray_curtis_rarefaction_5000_19.txt -o
2d_plots/ -m Run3_mapping.txt -b SampleID

make_distance_histograms.py -d
beta_div/bray_curtis_rarefaction_5000_19.txt -m Run3_mapping.txt
-f Treatment

make_distance_histograms.py -d
beta_div/weighted_unifrac_rarefaction_5000_19.txt -m
Run3_mapping.txt -f Treatment

make_distance_comparison_plots.py -d
beta_div/bray_curtis_rarefaction_5000_19.txt -m Run3_mapping.txt
-f Description -c "PCR_repeat,Extract_repeat" -o
Comparison_Bar_output -t "bar" -g pdf

make_distance_comparison_plots.py -d
beta_div/bray_curtis_rarefaction_5000_19.txt -m Run3_mapping.txt
-f Treatment -c "PCR,Extraction" -o Comparison_Bar_output2 -t
"scatter" -g pdf

make_distance_boxplots.py -d
beta_div/bray_curtis_rarefaction_5000_19.txt -m Run3_mapping.txt
-f "Treatment" -o Distance_Boxplots --y_max=1.2 --
whisker_length=3 --suppress_all_within --suppress_all_between

beta_significance.py -i otu_table2.txt -t rep_phylo.tre -s
weighted_unifrac -o unw_sig.txt -n 100

beta_significance.py -i otu_table2.txt -t rep_phylo.tre -s p-
test -o p_test.txt -n 10

otu_category_significance.py -i otu_table2.txt -m
Run3_mapping.txt -s ANOVA -c Treatment

dissimilarity_mtx_stats.py -i beta_div/ -o dist_stats/

beta_diversity.py -i otu_table2.txt -m euclidean -o
beta_div_for_UPGMA/
```

```
upgma_cluster.py -i beta_div_for_UPGMA/euclidean_otu_table2.txt
-o beta_div_cluster.tre

multiple_rarefactions.py -i otu_table2.txt -m 4000 -x 4500 -s
500 -n 10 -o rarefaction_tables_for_UPGMA/

beta_diversity.py -i rarefaction_tables_for_UPGMA/ -m euclidean
-o beta_div_for_UPGMA2/

upgma_cluster.py -i beta_div_for_UPGMA2/ -o beta_div_clusters

tree_compare.py -m beta_div_for_UPGMA/beta_div_cluster.tre -s
beta_div_for_UPGMA/beta_div_clusters -o jackknife_comparison/

make_bootstrapped_tree.py -m
jackknife_comparison/master_tree.tre -s
jackknife_comparison/jackknife_support.txt -o
jackknife_samples.pdf
```

APPENDIX 14: FINAL WORKFLOW FOR 454

```
wget
http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_file
s/ core_set_aligned.fasta.imputed

wget
http://greengenes.lbl.gov/Download/Sequence_Data/lanemask_in_1s_
and_0s

process_sff.py -i sffs/ -f -o output_dir

split_libraries.py -o Run5C -f Frag.G4ET2Z403.fna -q
Frag.G4ET2Z403.qual -m 5C_Mapping.txt -w 50 -b 10 -l 100 -L 450
-a 5 -c -r -t -n 1

split_libraries.py -o Run6A -f A_HJK0XVJ01.fna -q
A_HJK0XVJ01.qual -m 6A_Mapping.txt -w 50 -b 10 -l 100 -L 450 -a
5 -c -r -t -n 1000001

split_libraries.py -o Run6B -f B_HJK0XVJ02.fna -q
B_HJK0XVJ02.qual -m 6B_Mapping.txt -w 50 -b 10 -l 100 -L 450 -a
5 -c -r -t -n 2000001

#used http://www.n3phele.com/ an access portal for Qiime on
Amazon Cloud services and a bioinformatics platform in its own
right:

denoise_wrapper.py -v -i Frag.G4ET2Z403.txt -f Run5C/5C_seqs.fna
-o Run5C/denoised/ -m 5C_Mapping.txt --titanium –n 8

denoise_wrapper.py -v -i A_HJK0XVJ01.txt -f Run6A/6A_seqs.fna -o
Run6A/denoised/ -m 6A_Mapping.txt --titanium –n 8

denoise_wrapper.py -v -i B_HJK0XVJ02.txt -f Run6B/6B_seqs.fna -o
Run6B/denoised/ -m 6B_Mapping.txt --titanium –n 8

#used Qiime 1.4.0 Virtual Box for the following

inflate_denoiser_output.py -c
centroids_5C.fasta,centroids_6A.fasta,centroids_6B.fasta -s
singletons_5C.fasta,singletons_6A.fasta,singletons_6B.fasta -f
5C_seqs.fna,6A_seqs.fna,6B_seqs.fna -d
denoiser_mapping_5C.txt,denoiser_mapping_6A.txt,denoiser_mapping
_6B.txt -o denoised_seqs.fna

pick_otus.py -s 0.97 -i denoised_seqs.fna -m uclust -n 100 -t

pick_rep_set.py -f denoised_seqs.fna -i denoised_seqs_otus.txt -
m first

align_seqs.py -i denoised_seqs.fna_rep_set.fasta -t
core_set_aligned.fasta.imputed -o pynast_aligned/

assign_taxonomy.py -i denoised_seqs.fna_rep_set.fasta -m rdp -c
0.8 -r gg_97_otus_4feb2011.fasta -t
greengenes_tax_rdp_train_genus.txt
```

```
identify_chimeric_seqs.py -m ChimeraSlayer -i
denoised_seqs.fna_rep_set_aligned.fasta -a
core_set_aligned.fasta.imputed -o chimeric_seqs.txt

filter_fasta.py -f denoised_seqs.fna_rep_set_aligned.fasta -o
non_chimeric_rep_set_aligned.fasta -s chimeric_seqs.txt –n

filter_alignment.py -i non_chimeric_rep_set_aligned.fasta -m
lanemask_in_1s_and_0s -o filtered_alignment/

make_phylogeny.py -i
non_chimeric_rep_set_aligned_pfiltered.fasta -o rep_phylo.tre

make_otu_table.py -i denoised_seqs_otus.txt -o otu_table.biom -e
chimeric_seqs.txt -t
denoised_seqs.fna_rep_set_tax_assignments.txt

#For mothur and RDP

#The original denoised_seqs_otus.txt file contains a list of all
the sequences which were originally clustered. Rather than
repeat chimera-checking in multiple platforms this can be
utilised to create a file containing all the sequences which
progress to be counted for the final OTU table, since the
numbers of those excluded are not in the final .biom file. Open
the otu_table.biom and extract the first column to Word. Use
Find and Replace tool to replace "^p" with "\|" giving a list of
OTU numbers. Then at the command line type the following,
pasting the Word file contents between the inverted commas:

grep –w "" denoised_seqs_otus.txt > temp.txt

filter_fasta.py -f denoised_seqs.fna -o final.fasta -m temp.txt

#In QIIME command line (or any linux system), type the following
command to get a list of sequence names (lines that start with
">"):

grep ">" final.fasta > final.fasta.groups

#Open final.fasta.groups in Excel as a space-delimited file.
Delete every other column except the first column containing the
sequence name. Using the Find and Replace tool in Excel, find
">" and replace with nothing. Use Excel's formula to extract
just the sample name into the second column from the sequence
name in the first column. This can be done using the following
formula in the second column (cell B1): =LEFT(A1,FIND("_",A1)-
1). Fill down to the remaining rows in the second column. Save
as a txt (tab delimited) file and replace the exisiting
final.fasta.groups file, which can also be used to create the
design file required later. Also use Excel to create a .accnos
file for each sample by filtering for each sample name in column
2 and then saving for each sample as "sample.accnos". Repeat
the command on the following line with each individual
sample list to create files for the RDP.

mothur > get.seqs(accnos=sample.accnos, fasta=final.fasta)

mothur > unique.seqs(fasta=final.fasta)
```

```
mothur > classify.seqs(fasta=final.unique.fasta,
template=trainset6_032010.rdp.fasta,
taxonomy=trainset6_032010.rdp.tax)

mothur > align.seqs(candidate=final.unique.fasta,
template=core_set_aligned.imputed.fasta, flip=T)

mothur > filter.seqs(fasta=final.unique.align, vertical=T)

mothur > dist.seqs(fasta=final.unique.filter.fasta)

mothur > cluster(column=final.unique.filter.dist,
name=final.names)

mothur > phylotype(taxonomy=final.unique.rdp.taxonomy,
name=final.names)

mothur > classify.otu(list=final.unique.rdp.tx.list,
name=final.names, taxonomy=final.unique.rdp.taxonomy, label=1)

mothur > classify.otu(list=final.unique.filter.an.list,
name=final.names, taxonomy=final.unique.rdp.taxonomy,
label=0.03)

mothur > make.shared(list=final.unique.filter.an.list,
group=final.fasta.groups)

mothur > make.shared(list=final.unique.rdp.tx.list,
group=final.fasta.groups)

mothur > collect.single(shared=final.unique.filter.an.shared,
calc=chao-invsimpson, freq=100)

mothur >
rarefaction.single(shared=final.unique.filter.an.shared,
calc=chao-invsimpson, freq=100, label=unique-0.01-0.03-0.05)

mothur > summary.single(shared=final.unique.filter.an.shared,
calc=nseqs-coverage-sobs-invsimpson)

mothur >
summary.shared(shared=final.unique.filter.an.shared,calc=braycur
tis-jabund-morisitahorn-sorabund-structeuclidean-
structkulczynski-thetan-thetayc-sharedsobs-
kulczynskicody,label=unique-0.01-0.03-0.05)

mothur > sub.sample(shared=final.unique.filter.an.shared,
size=400)

mothur > sub.sample(shared=final.unique.rdp.tx.shared, size=400)

mothur > metastats(shared=
final.unique.rdp.tx.1.subsample.shared, design=final.design)

#From Mothur take the following files:
final.unique.rdp.tx.1.cons (output of classify.otu) and
final.unique.rdp.tx.shared (output of make.shared) and combine
and edit to create a file with OTUs in rows and samples in
columns, the taxonomy of an OTU split into Phylum, Class, Order,
```

Family and Genus (i.e. 5 columns for an OTU name), the abundance per sample per OTU being represented as a decimal proportion such that each sample column total is equal to 1. Two final columns are 'Total' (total incidences per OTU across all samples) and 'Max' (maximum proportion of any sample represented by a given OTU, i.e. if an OTU is found only once in a single sample or in all samples this figure is the same). A final row contains total counts for each sample. The file should be saved as tab-delimited: Mothur.Input.For.R.Final.txt

#Open 'R' and paste in the following text:

```
source("http://bioconductor.org/biocLite.R")
biocLite()
install.packages("gplots", dependencies = TRUE)

OTUTable <-
read.table("C:/Users/Adam/Desktop/R_ROOT/Mothur.Input.For.R.Fina
l.txt", sep="\t", header=T)

#Remove last row containing total counts and arrange according
to Phylum name, then Class etc

OTUTable <- OTUTable[ 1: (length(OTUTable[,1])-1),]

OTUTable <- OTUTable[order(OTUTable[, "Phylum"], OTUTable[,
"Class"], OTUTable[, "Order"], OTUTable[, "Family"], OTUTable[,
"Genus"]), ]

#Filter out rows that don't meet criteria e.g. total count of
>50 or >1% in at least one sample

OTUTable.Filter <- OTUTable[ OTUTable[, "Total"] >= 50 &
OTUTable[, "Max"] >= 0.01 , ]

#Ignore the first 5 columns containing the names and the last 2
columnsand last row with total counts,

OTUTable.Data <- OTUTable.Filter[, 6:
(length(OTUTable.Filter[1,])-2) ]

#Logit transform and then create names

OTUTable.lg <- log((OTUTable.Data/(1-OTUTable.Data))+0.00001)

OTUTableNames <- paste(OTUTable.Filter[,1], OTUTable.Filter[,4],
OTUTable.Filter[,5], sep=" | ")

library(gplots)
library(marray)

heatmap.2(as.matrix(OTUTable.lg),         Rowv=F,         Colv=F,
col=maPalette(12,   high="yellow",   mid="green",   low="blue"),
labRow=OTUTableNames,    breaks=seq(from=-11,  to=0,  length=13),
scale="none", key=F,  xlab="Sample", ylab="Genus", trace="none",
colsep=c(1:18),         rowsep=c(0:59),         sepcolor="black",
```

```
sepwidth=c(0.001,0.001),    lwid=c(.01,.95),    lhei=c(.01,.95),
margins=  c(4,18),  add.expr=abline(v=c(5.5,  10.5),  col="white",
lwd=3))

#Return to Qiime

per_library_stats.py -i otu_table.biom

make_otu_heatmap_html.py -i otu_table.biom -o Heatmap/

make_otu_network.py -m Mapping.txt -i otu_table.biom -o Network/
-b SampleID

summarize_taxa.py -i otu_table.biom -o ./tax

summarize_taxa_through_plots.py -o taxa_summary -i
otu_table.biom -m Mapping.txt –c Treatment

assign_taxonomy.py -i final.fasta -m rdp -c 0.80 -o
for_R_assignment/ -t greengenes_tax_rdp_train_genus.txt



#Open the for_R_assignment text file from within Excel with TAB
and      SEMICOLON      as      delimiters      and      create
For.R.From.Qiime.Final.txt (a tab-delimited text file) with the
following column headings: Sequence, Sample, Root, Kingdom,
Phylum, Class, Order, Family, and Genus. Copy and paste the
following into 'R':

source("http://bioconductor.org/biocLite.R")

biocLite()

install.packages("gplots", dependencies = TRUE)

TaxonTable <-
read.delim("C:/Users/Adam/Desktop/R_ROOT/For.R.From.Qiime.Final.
txt", sep="\t", header=T)

TaxonTable <- TaxonTable[, 1:9]

SampleNames <- unique(TaxonTable[,"Sample"])

AbundanceTable <- function(DataTable, SampleNames, DataCol,
ListNames = unique(DataTable[,DataCol]), perc = T) {

           ReturnTable <- NULL

           for (i in 1:length(SampleNames)) {

           Data <- DataTable[ DataTable[,"Sample"]
     ==SampleNames[i], ]

           Data <- Data[!is.na(Data[, DataCol]),]

           TotalLength <- length(Data[,1])

           NewRow <- NULL
```

```
            for (j in 1:length(ListNames)) {
                  temp <- Data[Data[,DataCol] == ListNames[j], ]
                  temp <- temp[!is.na(temp[, DataCol]),]
                  NewRow <- cbind(NewRow, length(temp[,1]))
            }
            if (perc == T) {
                  NewRow <- NewRow / TotalLength
            }
            Total <- sum(NewRow)
            NewRow <- cbind(NewRow, Total, TotalLength)
            ReturnTable <- rbind(ReturnTable, NewRow)
            }
            colnames(ReturnTable) <- c(as.character(ListNames),
      "Sum", "Total number of sequences")
            rownames(ReturnTable) <- SampleNames
            return(ReturnTable)
}

genusCount<- AbundanceTable(TaxonTable, SampleNames,
DataCol="Genus")

write.table("C:/Users/Adam/Desktop/R_ROOT/Run5and6_genusCount.tx
t", sep="\t")

#Open table in Excel and edit creating GenusCount.Final

ClassifierTable <-
read.table("C:/Users/Adam/Desktop/R_ROOT/GenusCount.Final.txt",
sep="\t", header=T)

ClassifierTable <- ClassifierTable[order(ClassifierTable[,
"Phylum"], ClassifierTable[, "Class"], ClassifierTable[,
"Order"], ClassifierTable[, "Family"], ClassifierTable[,
"Genus"]), ]

ClassifierTable.Data <- ClassifierTable[,
6:length(ClassifierTable[1,])]
ClassifierTable.lg <- log((ClassifierTable.Data/(1-
ClassifierTable.Data))+0.00001)

GenusTableNames <- paste(ClassifierTable[,1],
ClassifierTable[,4], ClassifierTable[,5], sep=" | ")

heatmap.2(as.matrix(ClassifierTable.lg), Rowv=F, Colv=F,
col=maPalette(12, high="yellow", mid="green", low="blue"),
```

```
labRow=GenusTableNames,  breaks=seq(from=-11, to=0, length=13),
scale="none", key=F,  xlab="Sample", ylab="Genus", trace="none",
colsep=c(1:18),  rowsep=c(0:59),  sepcolor="black",
sepwidth=c(0.001,0.001), lwid=c(.01,.95), lhei=c(.01,.95),
margins= c(4,24), add.expr=abline(v=c(5.5, 10.5), col="white",
lwd=3))

#Continuation of Qiime workflow

multiple_rarefactions.py -i otu_table.biom -m 400 -x 2000 -s 100
-n 5 -o rarefied_otu_tables/

alpha_diversity.py -i rarefied_otu_tables/ -m
ACE,berger_parker_d,chao1,fisher_alpha,observed_species,reciproc
al_simpson,shannon,simpson,simpson_e,PD_whole_tree -o adiv/ -t
rep_phylo.tre

collate_alpha.py -i adiv/ -o collated_alpha/

make_rarefaction_plots.py -i collated_alpha/ -m Mapping.txt –g
pdf –b Treatment

single_rarefaction.py -i otu_table.biom -o
otu_table_even1200.biom -d 1200

make_prefs_file.py -m Mapping.txt –b
"SampleID,Treatment,SampleID&&Treatment" -o prefs_out.txt –i
otu_table_L6.txt

beta_diversity.py -i otu_table_even1200.biom –m
abund_jaccard,binary_euclidean,binary_jaccard,binary_sorensen_di
ce,bray_curtis,euclidean,kulczynski,morisita_horn,spearman_appro
x,unifrac,unifrac_g,unifrac_g_full_tree,unweighted_unifrac,unwei
ghted_unifrac_full_tree,weighted_normalized_unifrac,weighted_uni
frac -o beta_div/ -t rep_phylo.tre

principal_coordinates.py -i beta_div/ -o beta_div_pcoa_results/

make_2d_plots.py -i pcoa_binary_euclidean_otu_table_even1200.txt
-o 2d_plots_be/ -m Mapping_PCA.txt -b 'Treatment,Group'

make_2d_plots.py -i pcoa_euclidean_otu_table_even1200.txt -o
2d_plots_e/ -m Mapping_PCA.txt -b 'Treatment,Group'

make_2d_plots.py -i pcoa_morisita_horn_otu_table_even1200.txt -o
2d_plots_mh/ -m Mapping_PCA.txt -b 'Treatment,Group'

make_2d_plots.py -i pcoa_weighted_unifrac_otu_table_even1200.txt
-o 2d_plots_wu/ -m Mapping_PCA.txt -b 'Treatment,Group'

make_3d_plots.py -i pcoa_binary_euclidean_otu_table_even1200.txt
-m Mapping_PCA.txt -t otu_table_L6.txt -b 'Treatment,Group' –o
3d_plots_be/

make_3d_plots.py -i pcoa_euclidean_otu_table_even1200.txt -m
Mapping_PCA.txt -t otu_table_L6.txt -b 'Treatment,Group' –o
3d_plots_e/
```

```
make_3d_plots.py -i pcoa_morisita_horn_otu_table_even1200.txt -m
Mapping_PCA.txt -t otu_table_L6.txt -b 'Treatment,Group' –o
3d_plots_mh/

make_3d_plots.py -i pcoa_weighted_unifrac_otu_table_even1200.txt
-m Mapping_PCA.txt -t otu_table_L6.txt -b 'Treatment,Group' –o
3d_plots_wu/

beta_diversity.py -i otu_table.biom -m weighted_unifrac -o
beta_div_full/ -t rep_phylo.tre

upgma_cluster.py -i weighted_unifrac_otu_table.biom -o
beta_div_cluster.tre

beta_diversity.py -i rarefied_otu_tables/ -m weighted_unifrac -o
beta_div_for_confidence/ -t rep_phylo.tre

upgma_cluster.py -i beta_div_for_confidence/ -o
beta_div_weighted_clusters/

tree_compare.py -m beta_div_cluster.tre -s
beta_div_weighted_clusters/ -o jackknife_comparison/

consensus_tree.py -i beta_div_weighted_clusters -o
consensus_tree.tre

make_bootstrapped_tree.py -m master_tree.tre -s
jackknife_support.txt -o jackknife_samples.pdf

make_distance_boxplots.py -d weighted_unifrac_otu_table.biom -m
Mapping.txt -f "Treatment" -o out_files

beta_significance.py -i otu_table.biom -t rep_phylo.tre -s
unweighted_unifrac -o unw_sig.txt

beta_significance.py -i otu_table.biom -t rep_phylo.tre -s
weighted_unifrac -o w_sig.txt

beta_significance.py -i otu_table.biom -t rep_phylo.tre -s p-
test -o p_test.txt

transform_coordinate_matrices.py -o out/ -i
pcoa_morisita_horn_otu_table_even1200.txt,pcoa_weighted_unifrac_
otu_table_even1200.txt

compare_3d_plots.py -i 'pc1_transformed.txt,pc2_transformed.txt'
–m Mapping.txt

otu_category_significance.py -i otu_table.biom -m Mapping.txt -s
g_test -c Treatment -f 3 -o single_g_test.txt

otu_category_significance.py -i rarefied_otu_tables -m
Mapping.txt -s ANOVA -c Treatment -f 3 -o multiple_anova.txt
```

REFERENCES AND BIBLIOGRAPHY

REFERENCES

Aagaard, K., Riehle, K., Ma, J., Segata, N., Mistretta, T. A., Coarfa, C. et al. (2012) A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PloS One* **7:** e36466.

Aanensen, D. M., & Spratt, B. G. (2005) The multilocus sequence typing network: mlst. net. *Nucleic Acids Research* **33:** W728-W733.

Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., & Polz, M. F. (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology* **71:** 8966-8969.

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., & Polz, M. F. (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *Journal of Bacteriology* **186:** 2629-2635.

Adlerberth, Marina Cerquetti, Isabelle Poilane, Agnes Wold, Anne Collignon, I. (2000) Mechanisms of colonisation and colonisation resistance of the digestive tract part 1: bacteria/host interactions. *Microbial Ecology in Health and Disease* **12:** 223-239.

Afonina, I., Ankoudinova, I., Mills, A., Lokhov, S., Huynh, P., & Mahoney, W. (2007) Primers with 5' flaps improve real-time PCR. *BioTechniques* **43:** 770-773

Aktories, K., & Wilkins, T. D. (2000) *Clostridium difficile* (Current Topics in Microbiology and Immunology). 2000 edition. New York: Springer Publishing.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215:** 403-410.

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews* **59:** 143-169.

Amann, R. I., Binder, B. J., Olson, R. J., Chisholm, S. W., Devereux, R., & Stahl, D. A. (1990) Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Applied and Environmental Microbiology* **56:** 1919-1925.

Aminov, R. I. (2010) A brief history of the antibiotic era: lessons learned and challenges for the future. *Frontiers in Microbiology* **1/134:** 1-7

Andersson, A. F., Lindberg, M., Jakobsson, H., Bäckhed, F., Nyrén, P., & Engstrand, L. (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One* **3/7:** e2836.

Angiuoli, S. V., White, J. R., Matalka, M., White, O., & Fricke, W. F. (2011) Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. *PloS One* **6/10:** e26624.

Ansorge, W. J. (2009) Next-generation DNA sequencing techniques. *New Biotechnology* **25:** 195-203.

Antonopoulos, D. A., Huse, S. M., Morrison, H. G., Schmidt, T. M., Sogin, M. L., & Young, V. B. (2009) Reproducible community dynamics of the gastrointestinal microbiota following antibiotic perturbation. *Infection and Immunity* **77:** 2367-2375.

Arias, C. A., & Murray, B. E. (2009) Antibiotic-resistant bugs in the 21st century—a clinical super-challenge. *New England Journal of Medicine* **360:** 439-443.

Arndt, D., Xia, J., Liu, Y., Zhou, Y., Guo, A. C., Cruz, J. A. et al. (2012) METAGENassist: a comprehensive web server for comparative metagenomics. *Nucleic Acids Research* **40:** W88-W95.

Ashby, M. N., Rine, J., Mongodin, E. F., Nelson, K. E., & Dimster-Denk, D. (2007) Serial analysis of rRNA genes and the unexpected dominance of rare members of microbial communities. *Applied and Environmental Microbiology* **73:** 4532-4542.

Ashelford, K. E., Weightman, A. J., & Fry, J. C. (2002) PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. *Nucleic Acids Research* **30:** 3481-3489.

Aslam, S., Hamill, R. J., & Musher, D. M. (2005) Treatment of *Clostridium difficile*-associated disease: old therapies and new strategies. *The Lancet Infectious Diseases* **5:** 549-557.

Atherton, J. C., & Blaser, M. J. (2009) Coadaptation of *Helicobacter pylori* and humans: ancient history, modern implications. *The Journal of Clinical Investigation* **119:** 2475-2487.

Audic, S., & Claverie, J. M. (1997) The significance of digital gene expression profiles. *Genome Research* **7:** 986-995.

Babcock, G. J., Broering, T. J., Hernandez, H. J., Mandell, R. B., Donahue, K., Boatright, N. et al. (2006) Human monoclonal antibodies directed against toxins A and B prevent *Clostridium difficile*-induced mortality in hamsters. *Infection and Immunity* **74:** 6339-6347.

Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., & Gordon, J. I. (2005) Host-bacterial mutualism in the human intestine. *Science* **307:** 1915-1920.

Baden, W. F. (1957) Staphylococcal and subsequent *candida albicans* enterocolitis complicating novobiocin therapy. *American Journal of Obstetrics and Gynecology* **74:** 47-52.

Baez, S., & Gordon, H. A. (1971) Tone and reactivity of vascular smooth muscle in germfree rat mesentery. *Journal of Experimental Medicine* **134:** 846-856.

Baker, G., Smith, J., & Cowan, D. A. (2003) Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods* **55:** 541-555.

Balzan, S., De Almeida Quadros, C., De Cleva, R., Zilberstein, B., & Cecconello, I. (2007) Bacterial translocation: overview of mechanisms and clinical impact. *Journal of Gastroenterology and Hepatology* **22:** 464-471.

Barth, H., Aktories, K., Popoff, M. R., & Stiles, B. G. (2004) Binary bacterial toxins: biochemistry, biology, and applications of common *Clostridium* and *Bacillus* proteins. *Microbiology and Molecular Biology Reviews* **68:** 373-402.

Bartlett, J. G. (2002) Antibiotic-associated diarrhea. *New England Journal of Medicine* **346:** 334-339.

Bartlett, J. G., Chang, T. E. W., Gurwith, M., Gorbach, S. L., & Onderdonk, A. B. (1978) Antibiotic-associated pseudomembranous colitis due to toxin-producing clostridia. *England Journal of Medicine* **298:** 531-534.

Bartlett, J., Taylor, N., Chang, T., & Dzink, J. (1980) Clinical and laboratory observations in *Clostridium difficile* colitis. *American Journal of Clinical Nutrition* **33:** 2521-2526.

Bartlett, J., Moon, N., Chang, T., Taylor, N., & Onderdonk, A. (1978) Role of *Clostridium difficile* in antibiotic-associated pseudomembranous colitis. *Gastroenterology* **75:** 778-782

Bartosch, S., Fite, A., Macfarlane, G. T., & McMurdo, M. E. T. (2004) Characterization of bacterial communities in feces from healthy elderly volunteers and hospitalized elderly patients by using real-time PCR and effects of antibiotic treatment on the fecal microbiota. *Applied and Environmental Microbiology* **70:** 3575-3581.

Bauer, H., Horowitz, R. E., Levenson, S. M., & Popper, H. (1963) The response of the lymphatic tissue to the microbial flora. Studies on germfree mice. *The American Journal of Pathology* **42:** 471-483

Beattie, W. G., Meng, L., Turner, S. L., Varma, R. S., Dao, D. D., & Beattie, K. L. (1995) Hybridization of DNA targets to glass-tethered oligonucleotide probes. *Molecular Biotechnology* **4:** 213-225.

Beaucage, S. L. (2001) Strategies in the preparation of DNA oligonucleotide arrays for diagnostic applications. *Current Medicinal Chemistry* **8:** 1213-1244.

Beaugerie, L., & Petit, J. C. (2004) Microbial-gut interactions in health and disease. Antibiotic-associated diarrhoea. *Best Practice & Research: Clinical Gastroenterology* **18:** 337-352.

Bender, R., & Lange, S. (1999) Multiple test procedures other than Bonferroni's deserve wider use. *BMJ: British Medical Journal* **318:** 600-601.

Ben-Dov, E., Shapiro, O. H., Siboni, N., & Kushmaro, A. (2006) Advantage of using inosine at the 3′ termini of 16S rRNA gene universal primers for the study of microbial diversity. *Applied and Environmental Microbiology* **72:** 6902-6906.

Benjamini, Y., & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society.Series B (Methodological)* **57:** 289-300.

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J. et al. (2010) Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proceedings of the National Academy of Sciences* **107:** 18933-18938.

Bent, E., Yin, B., Figueroa, A., Ye, J., Fu, Q., Liu, Z. et al. (2006) Development of a 9600-clone procedure for oligonucleotide fingerprinting of rRNA genes: Utilization to identify soil bacterial rRNA genes that correlate in abundance with the development of avocado root rot. *Journal of Microbiological Methods* **67:** 171-180.

Benveniste, J., Lespinats, G., & Salomon, J. C. (1971) Serum and secretory IgA in axenic and holoxenic mice. *The Journal of Immunology* **107:** 1656-1662.

Berg, J. M., Tymoczko, J., & Stryer, L. (2007) Biochemistry. 6th edition. New York: W H Freeman and Co.

Berg, R. D. (1996) The indigenous gastrointestinal microflora. *Trends in Microbiology* **4:** 430-435.

Berger, W. H., & Parker, F. L. (1970) Diversity of planktonic foraminifera in deep-sea sediments. *Science* **168:** 1345-1347.

Berry, D., Mahfoudh, K. B., Wagner, M., & Loy, A. (2011) Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and Environmental Microbiology* **77:** 7846-7849.

Bhatti, M. A., & Frank, M. O. (2000) *Veillonella parvula* meningitis: case report and review of Veillonella infections. *Clinical Infectious Diseases* **31:** 839-840.

Bidet, P., Lalande, V., Salauze, B., Burghoffer, B., Avesani, V., Delmée, M. et al. (2000) Comparison of PCR-ribotyping, arbitrarily primed PCR, and pulsed-field gel electrophoresis for typing Clostridium difficile. *Journal of Clinical Microbiology* **38:** 2484-2487.

Binder, H., & Preibisch, S. (2005) Specific and nonspecific hybridization of oligonucleotide probes on microarrays. *Biophysical Journal* **89:** 337-352.

Blanchard, M., Taillon-Miller, P., Nowotny, P., & Nowotny, V. (1993) PCR buffer optimization with uniform temperature regimen to facilitate automation. *Genome Research* **2:** 234-240.

Blaser, M. J., & Falkow, S. (2009) What are the consequences of the disappearing human microbiota? *Nature Reviews Microbiology* **7:** 887-894.

Bochud, P. Y., Glauser, M. P., & Calandra, T. (2001) Antibiotics in sepsis. *Intensive Care Medicine* **27:** 33-48.

Bodrossy, L., & Sessitsch, A. (2004) Oligonucleotide microarrays in microbial diagnostics. *Current Opinion in Microbiology* **7:** 245-254.

Bongaerts, G. P. A., & Lyerly, D. M. (1997) Role of bacterial metabolism and physiology in the pathogenesis of *Clostridium difficile* disease. *Microbial Pathogenesis* **22:** 253-256.

Bongaerts, G., & Lyerly, D. M. (1994) Role of toxins A and B in the pathogenesis of *Clostridium difficile* disease. *Microbial Pathogenesis* **17:** 1-12.

Borden, J. R., Paredes, C. J., & Papoutsakis, E. T. (2005) Diffusion, mixing, and associated dye effects in DNA-microarray hybridizations. *Biophysical Journal* **89:** 3277-3284.

Borneman, J., Chrobak, M., Della Vedova, G., Figueroa, A., & Jiang, T. (2001) Probe selection algorithms with applications in the analysis of microbial communities. *Bioinformatics* **17:** S39-S48.

Borriello, S. P. (1990) The influence of the normal flora on *Clostridium difficile* colonisation of the gut. *Annals of Medicine* **22:** 61-67.

Borriello, S. P., & Honour, P. (1981) Simplified procedure for the routine isolation of Clostridium difficile from faeces. *Journal of Clinical Pathology* **34:** 1124-1127.

Borriello, S. (1988) Clostridial toxins and the gastrointestinal tract. *Current Opinion in Infectious Diseases* **1:** 906-911.

Borriello, S., & BARCLAY, F. E. (1985) Protection of hamsters against *Clostridium difficile* ileocaecitis by prior colonisation with non-pathogenic strains. *Journal of Medical Microbiology* **19:** 339-350.

Borriello, S., Welch, A., Barclay, F. E., & Davies, H. A. (1988) Mucosal association by *Clostridium difficile* in the hamster gastrointestinal tract. *Journal of Medical Microbiology* **25:** 191-196.

Borriello, S., Davies, H., Kamiya, S., Reed, P., & Seddon, S. (1990) Virulence factors of *Clostridium difficile*. *Review of Infectious Diseases* **12:** S185-S191.

Bowtell, D., & Sambrook, J. (2003) DNA microarrays: a molecular cloning manual. New York: Cold Spring Harbour Laboratory Press.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T. et al. (2008) The potential and challenges of nanopore sequencing. *Nature Biotechnology* **26:** 1146-1153.

Bray, J. R., & Curtis, J. T. (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs* **27:** 325-349.

Brazier, J. S. (1993) Role of the laboratory in investigations of *Clostridium difficile* diarrhea. *Clinical Infectious Diseases* **16:** S228-S233.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P., & Rohwer, F. (2003) Metagenomic analyses of an uncultured viral community from human feces. *Journal of Bacteriology* **185:** 6220-6223.

Brewer, A., & Williamson, M. (1994) A new relationship for rarefaction. *Biodiversity and Conservation* **3:** 373-379.

Brosius, J., Palmer, M. L., Kennedy, P. J., & Noller, H. F. (1978) Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli*. *Proceedings of the National Academy of Sciences* **75:** 4801-4805.

Brown, T. A. (2006) Genomes 3. New York: Garland Science Publishing

Brüggemann, H. (2005) Genomics of clostridial pathogens: implication of extrachromosomal elements in pathogenicity. *Current Opinion in Microbiology* **8:** 601-605.

Busse, H. J., Denner, E., & Lubitz, W. (1996) Classification and identification of bacteria: current approaches to an old problem. Overview of methods used in bacterial systematics. *Journal of Biotechnology* **47:** 3-38.

Bybee, S. M., Bracken-Grissom, H., Haynes, B. D., Hermansen, R. A., Byers, R. L., Clement, M. J. et al. (2011) Targeted Amplicon Sequencing (TAS): A Scalable Next-Gen Approach to Multilocus, Multitaxa Phylogenetics. *Genome Biology and Evolution* **3:** 1312–1323

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2010a) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26:** 266-267.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J. et al. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences* **108:** 4516-4522.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N. et al. (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal* **6:** 1621-4

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K. et al. (2010b) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7:** 335-336.

Cario, E., Gerken, G., & Podolsky, D. (2007) Toll-like receptor 2 controls mucosal inflammation by regulating epithelial barrier function. *Gastroenterology* **132:** 1359-1374.

Casadevall, A., & Pirofski, L. (1999) Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. *Infection and Immunity* **67:** 3703-3713.

Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods* **69:** 330-339.

Chandler, D., Fredrickson, J., & Brockman, F. (1997) Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. *Molecular Ecology* **6:** 475-482.

Chang, J. Y., Antonopoulos, D. A., Kalra, A., Tonelli, A., Khalife, W. T., Schmidt, T. M., & Young, V. B. (2008) Decreased Diversity of the Fecal Microbiome in Recurrent *Clostridium difficile*—Associated Diarrhea. *Journal of Infectious Diseases* **197:** 435-438.

Chao, A. (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11:** 265-270.

Chao, A., & Shen, T. (2010) SPADE (species prediction and diversity estimation). Program and User's Guide published at http://chao.stat.nthu.edu.tw.

Chao, A., Chazdon, R. L., Colwell, R. K., & Shen, T. J. (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* **62:** 361-371.

Chao, A., C Li, P., Agatha, S., & Foissner, W. (2006) A statistical approach to estimate soil ciliate diversity and distribution based on data from five continents. *Oikos* **114:** 479-493.

Chao, A., Chazdon, R. L., Colwell, R. K., & Shen, T. J. (2004) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters* **8:** 148-159.

Chapin III, F. S., Sala, O. E., Burke, I. C., Grime, J. P., Hooper, D. U., Lauenroth, W. K. et al. (1998) Ecosystem consequences of changing biodiversity. *Bioscience* **48:** 45-52.

Cho, J. C., & Tiedje, J. M. (2002) Quantitative detection of microbial genes by using DNA microarrays. *Applied and Environmental Microbiology* **68:** 1425-1430.

Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., & O'Toole, P. W. (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research* **38:** e200.

Claesson, M. J., O'Sullivan, O., Wang, Q., Nikkilä, J., Marchesi, J. R., Smidt, H. et al. (2009) Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS One* **4:** e6669.

Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S. et al. (2012) Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488:** 178-184.

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. *Proceedings of the National Academy of Sciences* **108:** 4586-4591.

Clemente, C. D. (1987) Anatomy: a regional atlas of the human body. 3rd Edition. Baltimore: Urban and Schwarzenberg.

Cloud, J., & Kelly, C. P. (2007) Update on *Clostridium difficile* associated disease. *Current Opinion in Gastroenterology* **23:** 4-9.

Coconnier, M. H., Liévin, V., Lorrot, M., & Servin, A. L. (2000) Antagonistic activity of Lactobacillus acidophilus LB against intracellular Salmonella enterica serovar Typhimurium infecting human enterocyte-like Caco-2/TC-7 cells. *Applied and Environmental Microbiology* **66:** 1152-1157.

Coconnier, M. H., Bernet, M. F., Kernéis, S., Chauvière, G., Fourniat, J., & Servin, A. L. (1993) Inhibition of adhesion of enteroinvasive pathogens to human intestinal Caco-2 cells by *Lactobacillus acidophilus* strain LB decreases bacterial invasion. *FEMS Microbiology Letters* **110:** 299-305.

Cole, J., Chai, B., Farris, R., Wang, Q., Kulam, S., McGarrell, D. et al. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Research* **33:** D294-D296.

Collins, M., Lawson, P., Willems, A., Cordoba, J., Fernandez-Garayzabal, J., Garcia, P. et al. (1994) The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *International Journal of Systematic Bacteriology* **44:** 812-826.

Cookson, B. (2007) Hypervirulent strains of *Clostridium difficile*. *Postgraduate Medical Journal* **83:** 291-295.

Cotran, R. S., Kumar, V., & Robbins, S. (1999) Pathological basis of disease. 4th Edition. Philadelphia: W. B. Saunders Company.

Crawley, M. J. (2007) The R book. Chichester: John Wiley and Sons Ltd.

Croswell, A., Amir, E., Teggatz, P., Barman, M., & Salzman, N. H. (2009) Prolonged impact of antibiotics on intestinal microbial ecology and susceptibility to enteric Salmonella infection. *Infection and Immunity* **77:** 2741-2753.

Cummings, J., & Macfarlane, G. (1991) The control and consequences of bacterial fermentation in the human colon. *Journal of Applied Microbiology* **70:** 443-459.

Dabard, J., Bridonneau, C., Phillipe, C., Anglade, P., Molle, D., Nardi, M. et al. (2001) Ruminococcin A, a New Lantibiotic Produced by a *Ruminococcus gnavus* Strain Isolated from Human Feces. *Applied and Environmental Microbiology* **67:** 4111-4118.

Dai, H., Meyer, M., Stepaniants, S., Ziman, M., & Stoughton, R. (2002) Use of hybridization kinetics for differentiating specific from non-specific binding to oligonucleotide microarrays. *Nucleic Acids Research* **30:** e86-e86.

Daly, K., & Shirazi-Beechey, S. P. (2006) Design and evaluation of group-specific oligonucleotide probes for quantitative analysis of intestinal ecosystems: their application to assessment of equine colonic microflora. *FEMS Microbiology Ecology* **44:** 243-252.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S. et al. (2010) 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.

De Queiroz, K. (2005) Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences* **102:** 6600-6607.

Deneve, C., Janoir, C., Poilane, I., Fantinato, C., & Collignon, A. (2009) New trends in *Clostridium difficile* virulence and pathogenesis. *International Journal of Antimicrobial Agents* **33:** S24-S28.

Denève, C., Deloménie, C., Barc, M. C., Collignon, A., & Janoir, C. (2008) Antibiotics involved in *Clostridium difficile*-associated disease increase colonization factor gene expression. *Journal of Medical Microbiology* **57:** 732-738.

Denhardt, D. (1966) A membrane-filter technique for the detection of complementary DNA. *Biochemical and Biophysical Research Communications* **23:** 641-646.

Deplancke, B., & Gaskins, H. R. (2001) Microbial modulation of innate defense: goblet cells and the intestinal mucus layer. *American Journal of Clinical Nutrition* **73:** 1131S-1141S.

DeSantis Jr, T., Hugenholtz, P., Keller, K., Brodie, E., Larsen, N., Piceno, Y. et al. (2006a) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Research* **34:** W394-W399.

DeSantis, T. Z., Brodie, E. L., Moberg, J. P., Zubieta, I. X., Piceno, Y. M., & Andersen, G. L. (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microbiology Ecology* **53:** 371-383.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K. et al. (2006b) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72:** 5069-5072.

Dethlefsen, L., Huse, S., Sogin, M. L., & Relman, D. A. (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biology* **6:** e280.

Dethlefsen, L., Eckburg, P. B., Bik, E. M., & Relman, D. A. (2006) Assembly of the human intestinal microbiota. *Trends in Ecology & Evolution* **21:** 517-523.

Dicksved, J., Flöistrup, H., Bergström, A., Rosenquist, M., Pershagen, G., Scheynius, A. et al. (2007) Molecular fingerprinting of the fecal microbiota of children raised according to different lifestyles. *Applied and Environmental Microbiology* **73:** 2284-2289.

Diehl, F., Grahlmann, S., Beier, M., & Hoheisel, J. D. (2001) Manufacturing DNA microarrays of high spot homogeneity and reduced background signal. *Nucleic Acids Research* **29:** e38-e38.

Dillon, S. T., Rubin, E. J., Yakubovich, M., Pothoulakis, C., LaMont, J. T., Feig, L. A., & Gilbert, R. J. (1995) Involvement of Ras-related Rho proteins in the mechanisms of action of *Clostridium difficile* toxin A and toxin B. *Infection and Immunity* **63:** 1421-1426.

Donskey, C. J. (2004) The role of the intestinal tract as a reservoir and source for transmission of nosocomial pathogens. *Clinical Infectious Diseases* **39:** 219-226.

Donskey, C. J., Hujer, A. M., Das, S. M., Pultz, N. J., Bonomo, R. A., & Rice, L. B. (2003) Use of denaturing gradient gel electrophoresis for analysis of the stool microbiota of hospitalized patients. *Journal of Microbiological Methods* **54:** 249-256.

Doolittle, W. F., & Papke, R. T. (2006) Genomics and the bacterial species problem. *Genome Biology* **7:** 116.

Drmanac, S., Stavropoulos, N., Labat, I., Vonau, J., Hauser, B., Soares, M., & Drmanac, R. (1996) Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. *Genomics* **37:** 29-40.

Dufva, M. (2005) Fabrication of high quality microarrays. *Biomolecular Engineering* **22:** 173-184.

Dunbar, J., Barns, S. M., Ticknor, L. O., & Kuske, C. R. (2002) Empirical and theoretical bacterial diversity in four Arizona soils. *Applied and Environmental Microbiology* **68:** 3035-3045.

Duncan, S. H., Louis, P., & Flint, H. J. (2004) Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. *Applied and Environmental Microbiology* **70:** 5810-5817.

Duncan, S. H., Belenguer, A., Holtrop, G., Johnstone, A. M., Flint, H. J., & Lobley, G. E. (2007) Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Applied and Environmental Microbiology* **73:** 1073-1078.

Duncan, S., Louis, P., & Flint, H. (2007) Cultivable bacterial diversity from the human colon. *Letters in Applied Microbiology* **44:** 343-350.

Duncan, S., Lobley, G., Holtrop, G., Ince, J., Johnstone, A., Louis, P., & Flint, H. (2008) Human colonic microbiota associated with diet, obesity and weight loss. *International Journal of Obesity* **32:** 1720-1724.

Durai, R. (2007) Epidemiology, pathogenesis, and management of *Clostridium difficile* infection. *Digestive Diseases and Sciences* **52:** 2958-2962.

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.

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

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194-2200.

Elliott, B., Chang, B., Golledge, C., & Riley, T. (2007) *Clostridium difficile*-associated diarrhoea. *Internal Medicine Journal* **37:** 561-568.

Evertsz, E., Au-Young, J., Ruvolo, M., Lim, A., & Reynolds, M. (2001) Research Report Hybridization Cross-Reactivity within Homolo-gous Gene Families on Glass cDNA Microarrays. *BioTechniques* **31:** 1182-1192.

Faith, D. P., Minchin, P. R., & Belbin, L. (1987) Compositional dissimilarity as a robust measure of ecological distance. *Plant Ecology* **69:** 57-68.

Farrelly, V., Rainey, F. A., & Stackebrandt, E. (1995) Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Applied and Environmental Microbiology* **61:** 2798-2801.

Fawley, W. N., Underwood, S., Freeman, J., Baines, S. D., Saxton, K., Stephenson, K. et al. (2007) Efficacy of hospital cleaning agents and germicides against epidemic *Clostridium difficile* strains. *Infection Control and Hospital Epidemiology* **28:** 920-925.

Ferreira, T. M., Leonel, A. J., Melo, M. A., Santos, R. R. G., Cara, D. C., Cardoso, V. N. et al. (2012) Oral Supplementation of Butyrate Reduces Mucositis and Intestinal Permeability Associated with 5-Fluorouracil Administration. *Lipids* **47:** 1-10.

Finegold, S. M., & Sutter, V. L. (1978) Fecal flora in different populations, with special reference to diet. *American Journal of Clinical Nutrition* **31:** S116-S122.

Fleming, A. (1929) On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of *B. influenzae. British Journal of Experimental. Pathology* **10:** 226-236.

Flint, H. J. (2004) Polysaccharide breakdown by anaerobic microorganisms inhabiting the mammalian gut. *Advances in Applied Microbiology* **56:** 89-120.

Flint, H. J., Duncan, S. H., Scott, K. P., & Louis, P. (2007) Interactions and competition within the microbial community of the human colon: links between diet and health. *Environmental Microbiology* **9:** 1101-1111.

Folkerts, G., Walzl, G., & Openshaw, P. J. M. (2000) Do common childhood infections 'teach'the immune system not to be allergic? *Immunology Today* **21:** 118-120.

Fons, Ana Gomez, Tuomo Karjalainen, M. (2000) Mechanisms of colonisation and colonisation resistance of the digestive tract Part 2: Bacteria/bacteria interactions. *Microbial Ecology in Health and Disease* **12:** 240-246.

Forney, L. J., Zhou, X., & Brown, C. J. (2004) Molecular microbial ecology: land of the one-eyed king. *Current Opinion in Microbiology* **7:** 210-220.

Frank, D. N., Robertson, C. E., Hamm, C. M., Kpadeh, Z., Zhang, T., Chen, H., Zhu, W., Sartor, R. B., Boedeker, E. C., Harpaz, N., Pace, N. R., & Li, E. (2011) Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflammatory Bowel Diseases* **17:** 179-184

Franks, A. H., Harmsen, H. J. M., Raangs, G. C., Jansen, G. J., Schut, F., & Welling, G. W. (1998) Variations of bacterial populations in human feces measured by fluorescent in situ hybridization with group-specific 16S rRNA-targeted oligonucleotide probes. *Applied and Environmental Microbiology* **64:** 3336-3345.

Fredericks, D., & Relman, D. A. (1996) Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clinical Microbiology Reviews* **9:** 18-33.

Freeman, J., Bauer, M. P., Baines, S. D., Corver, J., Fawley, W. N., Goorhuis, B., Kuijper, E. J., & Wilcox, M. H. (2010) The changing epidemiology of *Clostridium difficile* infections. *Clinical Microbiology Reviews* **23:** 529-549

Fu, C., Carter, J., Li, Y., Porter, J., & Kerley, M. (2006) Comparison of agar plate and real-time PCR on enumeration of *Lactobacillus*, *Clostridium perfringens* and total anaerobic bacteria in dog faeces. *Letters in Applied Microbiology* **42:** 490-494.

Galvez, J., Rodríguez-Cabezas, M. E., & Zarzuelo, A. (2005) Effects of dietary fiber on inflammatory bowel disease. *Molecular Nutrition & Food Research* **49:** 601-608.

Ganz, T. (2003) Defensins: antimicrobial peptides of innate immunity. *Nature Reviews Immunology* **3:** 710-720.

Garey, K. W., Jiang, Z. D., Ghantoji, S., Tam, V. H., Arora, V., & DuPont, H. L. (2010) A common polymorphism in the interleukin-8 gene promoter is associated with an increased risk for recurrent *Clostridium difficile* infection. *Clinical Infectious Diseases* **51:** 1406-1410.

Garrett, W. S., Gallini, C. A., Yatsunenko, T., Michaud, M., DuBois, A., Delaney, M. L. et al. (2010) Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. *Cell Host & Microbe* **8:** 292-300..

George, W., Sutter, V., Citron, D., & Finegold, S. (1979) Selective and differential medium for isolation of *Clostridium difficile*. *Journal of Clinical Microbiology* **9:** 214-219.

Geric, B., Carman, R. J., Rupnik, M., Genheimer, C. W., Sambol, S. P., Lyerly, D. M. et al. (2006) Binary toxin–producing, large clostridial toxin–negative *Clostridium difficile* strains are enterotoxic but do not cause disease in hamsters. *Journal of Infectious Diseases* **193:** 1143-1150.

Gerritsen, J., Smidt, H., Rijkers, G. T., & de Vos, W. M. (2011) Intestinal microbiota in human health and disease: the impact of probiotics. *Genes & Nutrition* **6:** 209-240.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P. et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Research* **15:** 1451-1455.

Gich, F., Garcia-Gil, J., & Overmann, J. (2001) Previously unknown and phylogenetically diverse members of the green nonsulfur bacteria are indigenous to freshwater lakes. *Archives of Microbiology* **177:** 1-10.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S. et al. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312:** 1355-1359.

Gillespie, D., & Spiegelman, S. (1965) A quantitative assay for DNA-RNA hybrids with DNA immobilized on a membrane. *Journal of Molecular Biology* **12:** 829-42.

Giovannoni, S. J., DeLong, E. F., Olsen, G. J., & Pace, N. R. (1988) Phylogenetic group-specific oligodeoxynucleotide probes for identification of single microbial cells. *Journal of Bacteriology* **170:** 720-726.

Good, I. J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika* **40:** 237-264.

Gordon, J. I., Ley, R. E., Wilson, R., Mardis, E., Xu, J., Fraser, C. M., & Relman, D. A. (2005) Extending our view of self: the human gut microbiome initiative (HGMI). Bethesda: National Human Genome Research Institute

Gumerlock, P. H., Tang, Y. J., Meyers, F. J., & Silva, J. (1991) Use of the Polymerase Chain Reaction for the Specific and Direct Detection of *Clostridium difficile* in Human Feces. *Review of Infectious Diseases* **13:** 1053-1060.

Gutell, R. R., Larsen, N., & Woese, C. R. (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiology Review* **58:** 10-26.

Guyton, A. C. (1991) Textbook of medical physiology. 8th Edition. Philadephia: W. B. Saunder company.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G. et al. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* **21:** 494-504.

HALL, I. C., & O'TOOLE, E. (1935) Intestinal flora in new-born infants: with a description of a new pathogenic anaerobe, *Bacillus difficilis*. *Archives of Pediatric and Adolescent Medicine* **49:** 390.

Hamady, M., Lozupone, C., & Knight, R. (2009) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *The ISME Journal* **4:** 17-27.

Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* **68:** 669-685.

Hansen, E. E., Lozupone, C. A., Rey, F. E., Wu, M., Guruge, J. L., Narra, A. et al. (2011) Pan-genome of the dominant human gut-associated archaeon, *Methanobrevibacter smithii*, studied in twins. *Proceedings of the National Academy of Sciences* **108:** 4599-4606.

Hansen, M. C., Tolker-Nielsen, T., Givskov, M., & Molin, S. (2006) Biased 16S rDNA PCR amplification caused by interference from DNA flanking the template region. *FEMS Microbiology Ecology* **26:** 141-149.

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

Harmsen, H. J. M., Raangs, G. C., He, T., Degener, J. E., & Welling, G. W. (2002) Extensive set of 16S rRNA-based probes for detection of bacteria in human feces. *Applied and Environmental Microbiology* **68:** 2982-2990.

Harmsen, H. J. M., Wildeboer-Veloo, A. C. M., Grijpstra, J., Knol, J., Degener, J. E., & Welling, G. W. (2000) Development of 16S rRNA-Based Probes for the *Coriobacterium* Group and the *Atopobium* Cluster and Their Application for

Enumeration of *Coriobacteriaceae* in Human Feces from Volunteers of Different Age Groups. *Applied and Environmental Microbiology* **66:** 4523-4527.

Harmsen, H. J. M., Wildeboer-Veloo, A. C. M., Raangs, G. C., Wagendorp, A. A., Klijn, N., Bindels, J. G., & Welling, G. W. (2000) Analysis of intestinal flora development in breast-fed and formula-fed infants by using molecular identification and detection methods. *Journal of Pediatric Gastroenterology and Nutrition* **30:** 61-67.

Harmsen, H., Gibson, G., Elfferich, P., Raangs, G., Wildeboer-Veloo, A., Argaiz, A. et al. (2006) Comparison of viable cell counts and fluorescence in situ hybridization using specific rRNA-based probes for the quantification of human fecal bacteria. *FEMS Microbiology Letters* **183:** 125-129.

Harrison, O. J., & Maloy, K. J. (2011) Innate immune activation in intestinal homeostasis. *Journal of Innate Immunity* **3:** 585-593.

Hayashi, H., Takahashi, R., Nishi, T., Sakamoto, M., & Benno, Y. (2005) Molecular analysis of jejunal, ileal, caecal and recto-sigmoidal human colonic microbiota using 16S rRNA gene libraries and terminal restriction fragment length polymorphism. *Journal of Medical Microbiology* **54:** 1093-1101.

He, M., Sebaihia, M., Lawley, T. D., Stabler, R. A., Dawson, L. F., Martin, M. J. et al. (2010) Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proceedings of the National Academy of Sciences* **107:** 7527-7532.

Hill, D. A., & Artis, D. (2009) Intestinal bacteria and the regulation of immune cell homeostasis. *Annual Review of Immunology* **28:** 623-667.

Hill, M. (1997) Intestinal flora and endogenous vitamin synthesis. *European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP)* **6:** S43.

Hill, T. C. J., Walsh, K. A., Harris, J. A., & Moffett, B. F. (2006) Using ecological diversity measures with bacterial communities. *FEMS Microbiology Ecology* **43:** 1-11.

Högenauer, C., Hammer, H. F., Krejs, G. J., & Reisinger, C. (1998) Mechanisms and management of antibiotic-associated diarrhea. *Clinical Infectious Diseases* **27:** 702-710.

Hold, G. L., Schwiertz, A., Aminov, R. I., Blaut, M., & Flint, H. J. (2003) Oligonucleotide probes that detect quantitatively significant groups of butyrate-producing bacteria in human feces. *Applied and Environmental Microbiology* **69:** 4320-4324.

Holzapfel, W. H., Haberer, P., Snel, J., & Schillinger, U. (1998) Overview of gut flora and probiotics. *International Journal of Food Microbiology* **41:** 85-101.

Hooper, L. V. (2004) Bacterial contributions to mammalian gut development. *Trends in Microbiology* **12:** 129-134.

Hooper, L. V., & Gordon, J. I. (2001) Commensal host-bacterial relationships in the gut. *Science* **292:** 1115-1118.

Hopkins, M., & Macfarlane, G. (2002) Changes in predominant bacterial populations in human faeces with age and with *Clostridium difficile* infection. *Journal of Medical Microbiology* **51:** 448-454.

Hopkins, M., & Macfarlane, G. (2001) Evaluation of 16s rRNA and cellular fatty acid profiles as markers of human intestinal bacterial growth in the chemostat. *Journal of Applied Microbiology* **89:** 668-677.

Hopkins, M., Sharp, R., & Macfarlane, G. (2002) Variation in human intestinal microbiota with age. *Digestive and Liver Disease* **34:** S12-S18.

Hopkins, M., Sharp, R., & Macfarlane, G. (2001) Age and disease related changes in intestinal bacterial populations assessed by cell culture, 16S rRNA abundance, and community cellular fatty acid profiles. *Gut* **48:** 198-205.

Horn, H. S. (1966) Measurement of" overlap" in comparative ecological studies. *The American Naturalist* **100:** 419-424.

Huber, J. A., Morrison, H. G., Huse, S. M., Neal, P. R., Sogin, M. L., & Mark Welch, D. B. (2009) Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environmental Microbiology* **11:** 1292-1302.

Hugenholtz, P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biology* **3:** 1-0003.8.

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., & Bohannan, B. J. M. (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology* **67:** 4399-4406.

Hughes, R. (1986) Theories and models of species abundance. *The American Naturalist* **128:** 879-899.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., & Oinn, T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Research* **34:** W729-W732.

Hurley, B. W., & Nguyen, C. C. (2002) The spectrum of pseudomembranous enterocolitis and antibiotic-associated diarrhea. *Archives of Internal Medicine* **162:** 2177.

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., & Welch, D. M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8:** R143.

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., & Sogin, M. L. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genetics* **4:** e1000255.

Huws, S. A., Edwards, J. E., Kim, E. J., & Scollan, N. D. (2007) Specificity and sensitivity of eubacterial primers utilized for molecular profiling of bacteria within complex microbial ecosystems. *Journal of Microbiological Methods* **70:** 565-569.

Ingle, J. D. J., and Crouch, S. R. (1988) The Beer Lambert Law. In Spectrochemical Analysis. New Jersey: Prentice Hall

Ishii, K., & Fukui, M. (2001) Optimization of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR. *Applied and Environmental Microbiology* **67:** 3753-3755.

Jaccard, P. (1901) Distribution of the alpine flora in the dranse's basin and some neighbouring regions. *Bulletin de la Société Vaudoise des Sciences Naturelles* **37:** 241-272.

Jakobsson, H. E., Jernberg, C., Andersson, A. F., Sjölund-Karlsson, M., Jansson, J. K., & Engstrand, L. (2010) Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLoS One* **5:** e9836.

Jalanka-Tuovinen, J., Salonen, A., Nikkilä, J., Immonen, O., Kekkonen, R., Lahti, L. et al. (2011) Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. *PloS One* **6:** e23035.

Jampachaisri, K., Valinsky, L., Borneman, J., & Press, S. J. (2005) Classification of oligonucleotide fingerprints: application for microbial community and gene expression analyses. *Bioinformatics* **21:** 3122-3130.

Jares, P. (2006) DNA microarray applications in functional genomics. *Ultrastructural Pathology* **30:** 209-219.

Jernberg, C., Löfmark, S., Edlund, C., & Jansson, J. K. (2007) Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *The ISME Journal* **1:** 56-66.

Jiang, Z. D., Garey, K. W., Price, M., Graham, G., Okhuysen, P., Dao-Tran, T. et al. (2007) Association of interleukin-8 polymorphism and immunoglobulin G anti-toxin A in patients with *Clostridium difficile*-associated diarrhea. *Clinical Gastroenterology and Hepatology* **5:** 964-968.

Johnson, I. S. (1983) Human insulin from recombinant DNA technology. *Science* **219:** 632-637.

Johnson, J., & FRANCIS, B. S. (1975) Taxonomy of the clostridia: ribosomal ribonucleic acid homologies among the species. *Journal of General Microbiology* **88:** 229-244.

Johnson, S., & Gerding, D. N. (1998) *Clostridium difficile*-associated diarrhea. *Clinical Infectious Diseases* **26:** 1027-1034.

Jost, L. (2007) Partitioning diversity into independent alpha and beta components. *Ecology* **88:** 2427-2439.

Jurasinski, G., & Koch, M. (2011) Commentary: do we have a consistent terminology for species diversity? We are on the way. *Oecologia* **167:** 893-902.

Kanagawa, T. (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering* **96:** 317-323.

Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D., & Madore, S. J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* **28:** 4552-4557.

Keesing, F., Holt, R. D., & Ostfeld, R. S. (2006) Effects of species diversity on disease risk. *Ecology Letters* **9:** 485-498.

Kelly, D., Conway, S., & Aminov, R. (2005) Commensal gut bacteria: mechanisms of immune modulation. *Trends in Immunology* **26:** 326-333.

Khoruts, A., Dicksved, J., Jansson, J. K., & Sadowsky, M. J. (2010) Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent *Clostridium difficile*-associated diarrhea. *Journal of Clinical Gastroenterology* **44:** 354.

Kikuchi, E., Miyamoto, Y., Narushima, S., & Itoh, K. (2002) Design of species-specific primers to identify 13 species of *Clostridium* harbored in human intestinal tracts. *Microbiology and Immunology* **46:** 353-358.

Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* **217:** 624.

Kitihara, K., Yasutake, Y., & Miyazaki, K. (2012) Mutational robustness of 16S ribosomal RNA, shown by experimental horizontal gene transfer in *Escherichia coli*. *Proceedings of the National Academy of Sciences* **109:** 19220–19225

Klemm, P., Hancock, V., & Schembri, M. A. (2007) Mellowing out: adaptation to commensalism by *Escherichia coli* asymptomatic bacteriuria strain 83972. *Infection and Immunity* **75:** 3688-3695.

Kluyver, A. J., & Niel, C. B. Van (1956) The microbe's contribution to biology. Cambridge, Massachusetts: Harvard University Press.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J. A. et al. (2012) Unlocking the potential of metagenomics through replicated experimental design. *Nature Biotechnology* **30:** 513-520.

Koltai, H., & Weingarten-Baror, C. (2008) Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. *Nucleic Acids Research* **36:** 2395-2405.

Kontani, M., Ono, H., Shibata, H., Okamura, Y., Tanaka, T., Fujiwara, T. et al. (1996) Cysteine protease of *Porphyromonas gingivalis* 381 enhances binding of fimbriae to cultured human fibroblasts and matrix proteins. *Infection and Immunity* **64:** 756-762.

Koonin, E. V., & Novozhilov, A. S. (2009) Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61:** 99-111.

Kotlowski, R., Bernstein, C. N., Sepehri, S., & Krause, D. O. (2007) High prevalence of *Escherichia coli* belonging to the B2 D phylogenetic group in inflammatory bowel disease. *Gut* **56:** 669-675.

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., & Knight, R. (2010) Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nature Methods* **7:** 813-819.

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., & Knight, R. (2011) Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics* **13:** 47-58.

Kuczynski, J., Costello, E. K., Nemergut, D. R., Zaneveld, J., Lauber, C. L., Knights, D. et al. (2010) Direct sequencing of the human microbiome readily reveals community differences. *Genome Biology* **11:** 210.

Küsel, K., Pinkart, H. C., Drake, H. L., & Devereux, R. (1999) Acetogenic and sulfate-reducing bacteria inhabiting the rhizoplane and deep cortex cells of the sea grass Halodule wrightii. *Applied and Environmental Microbiology* **65:** 5117-5123.

Kyne, L., Warny, M., Qamar, A., & Kelly, C. P. (2001) Association between antibody response to toxin A and protection against recurrent Clostridium difficile diarrhoea. *Lancet* **357:** 189.

Kyne, L., Warny, M., Qamar, A., & Kelly, C. P. (2000) Asymptomatic carriage of Clostridium difficile and serum levels of IgG antibody against toxin A. *New England Journal of Medicine* **342:** 390-397.

Labbé, A-C., Poirier, L., MacCannell, D., Louie, T., Savoie, M., Béliveau, C., Laverdière, M., & Pépin, J. (2008) *Clostridium difficile* infections in a Canadian tertiary care hospital before and during a regional epidemic associated with the BI/NAP1/027 strain. *Antimicrobial Agents and Chemotherapeutics* **52:** 3180-3187.

Laheij, R., Van Ijzendoorn, M., Janssen, M., & Jansen, J. (2003) Gastric acid-suppressive therapy and community-acquired respiratory infections. *Alimentary Pharmacology & Therapeutics* **18:** 847-851.

Lander, E. S. (1999) Array of hope. *Nature Genetics* **21:** 3-4.

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., & Pace, N. R. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences* **82:** 6955-6959.

Lane, D. (1991) 16S/23S rRNA sequencing, p. 115–175 In Stackebrandt E., Goodfellow M., editors.(ed.), Nucleic acid techniques in bacterial systematics. New York: Wiley

Langendijk, P. S., Schut, F., Jansen, G. J., Raangs, G. C., Kamphuis, G. R., Wilkinson, M., & Welling, G. W. (1995) Quantitative fluorescence in situ hybridization of *Bifidobacterium* spp. with genus-specific 16S rRNA-targeted probes and its application in fecal samples. *Applied and Environmental Microbiology* **61:** 3069-3075.

Lauber, C. L., Zhou, N., Gordon, J. I., Knight, R., & Fierer, N. (2010) Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiology Letters* **307:** 80-86.

Lay, C., Sutren, M., Rochet, V., Saunier, K., Doré, J., & Rigottier-Gois, L. (2005) Design and validation of 16S rRNA probes to enumerate members of the *Clostridium leptum* subgroup in human faecal microbiota. *Environmental Microbiology* **7:** 933-946.

Lay, C., Rigottier-Gois, L., Holmstrøm, K., Rajilic, M., Vaughan, E. E., De Vos, W. M. et al. (2005) Colonic microbiota signatures across five northern European countries. *Applied and Environmental Microbiology* **71:** 4153-4155.

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Osteras, M. et al. (2009) Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *Journal of Microbiological Methods* **79:** 266.

Lederberg, J. (2000) Infectious history. *Science* **288:** 287-293.

Legendre, P., & Gallagher, E. D. (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia* **129:** 271-280.

Legendre, P., & Legendre, L. (1998) 1998. Numerical ecology. Second English Edition. Amsterdam: Elsevier.

Lennon, N. J., Lintner, R. E., Anderson, S., Alvarez, P., Barry, A., Brockman, W. et al. (2010) A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome Biology* 2010, **11**: R15

Leser, T. D., Amenuvor, J. Z., Jensen, T. K., Lindecrona, R. H., Boye, M., & Møller, K. (2002) Culture-independent analysis of gut bacteria: the pig gastrointestinal tract microbiota revisited. *Applied and Environmental Microbiology* **68:** 673-690.

Leung, K., Zahn, H., Leaver, T., Konwar, K. M., Hanson, N. W., Pagé, A. P. et al. (2012) A programmable droplet-based microfluidic device applied to multiparameter analysis of single microbes and microbial communities. *Proceedings of the National Academy of Sciences* **109:** 7665-7670.

Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124:** 837-848.

Ley, R. E., Turnbaugh, P. J., Klein, S., & Gordon, J. I. (2006) Microbial ecology: human gut microbes associated with obesity. *Nature* **444:** 1022.

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., & Gordon, J. I. (2005) Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences* **102:** 11070-11075.

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S. et al. (2008) Evolution of mammals and their gut microbes. *Science* **320:** 1647-1651.

Li, E., Hamm, C. M., Gulati, A. S., Sartor, R. B., Chen, H., Wu, X. et al. (2012) Inflammatory bowel diseases phenotype, *C. difficile* and NOD2 genotype are associated with shifts in human ileum associated microbial composition. *PloS One* **7:** e26284.

Li, F., Hullar, M. A. J., & Lampe, J. W. (2007) Optimization of terminal restriction fragment polymorphism (TRFLP) analysis of human gut microbiota. *Journal of Microbiological Methods* **68:** 303.

Li, M., Gong, J., Cottrill, M., Yu, H., de Lange, C., Burton, J., & Topp, E. (2003) Evaluation of QIAamp® DNA Stool Mini Kit for ecological studies of gut microbiota. *Journal of Microbiological Methods* **54:** 13-20.

Li, M., Wang, B., Zhang, M., Rantalainen, M., Wang, S., Zhou, H. et al. (2008) Symbiotic gut microbes modulate human metabolic phenotypes. *Proceedings of the National Academy of Sciences* **105:** 2117-2122.

Liévin-Le Moal, V., & Servin, A. L. (2006) The front line of enteric host defense against unwelcome intrusion of harmful microorganisms: mucins, antimicrobial peptides, and microbiota. *Clinical Microbiology Reviews* **19:** 315-337.

Limaye, A. P., Turgeon, D. K., Cookson, B. T., & Fritsche, T. R. (2000) Pseudomembranous colitis caused by a toxin A− B strain of *Clostridium difficile*. *Journal of Clinical Microbiology* **38:** 1696-1697.

Linz, B., Balloux, F., Moodley, Y., Manica, A., Liu, H., Roumagnac, P. et al. (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445:** 915-918.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R. et al. (2012) Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* **2012:**

Liu, W. T., Mirzabekov, A. D., & Stahl, D. A. (2008) Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. *Environmental Microbiology* **3:** 619-629.

Liu, Z., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research* **36:** e120-e120.

Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D., & Knight, R. (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Research* **35:** e120.

Lloyd, J. R., & Lovley, D. R. (2001) Microbial detoxification of metals and radionuclides. *Current Opinion in Biotechnology* **12:** 248-253.

Loman, N. (2011) Curious results from 454 amplicon processing. http://pathogenomics.bham.ac.uk/blog/2011/05/curious-results-from-454-amplicon-processing/

Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., & Murray, C. J. L. (2006) Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *The Lancet* **367:** 1747-1757.

Lopez-Siles, M., Khan, T. M., Duncan, S. H., Aldeguer Mante, X., Harmsen, H. J. M., Garcia-Gil, L. J., & Flint, H. J. (2011) Gut Environmental Factors May Shape the Persistence of *Faecalibacterium Prausnitzii* in the Healthy and Diseased Large Intestine. *Gastroenterology* **140:** S-665.

Louis, P., Scott, K., Duncan, S., & Flint, H. (2007) Understanding the effects of diet on bacterial metabolism in the large intestine. *Journal of Applied Microbiology* **102:** 1197-1208.

Loy, A., Maixner, F., Wagner, M., & Horn, M. (2007) probeBase—an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Research* **35:** D800-D804.

Loy, A., Arnold, R., Tischler, P., Rattei, T., Wagner, M., & Horn, M. (2008) probeCheck–a central resource for evaluating oligonucleotide probe coverage and specificity. *Environmental Microbiology* **10:** 2894-2898.

Loy, A., Lehner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J. et al. (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Applied and Environmental Microbiology* **68:** 5064-5081.

Lozupone, C., & Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71:** 8228-8235.

Lozupone, C., Hamady, M., & Knight, R. (2006) UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7:** 371.

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2010) UniFrac: an effective distance metric for microbial community comparison. *The ISME Journal* **5:** 169-172.

Lozupone, C., Faust, K., Raes, J., Faith, J. J., Frank, D. N., Zaneveld, J. et al. (2012) Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. *Genome Research* **22:** 1974-84.

Lozupone, C. A., & Knight, R. (2007) Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences* **104:** 11436-11440.

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007) Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* **73:** 1576-1585.

Lu, J., Santo Domingo, J., & Shanks, O. C. (2007) Identification of chicken-specific fecal microbial sequences using a metagenomic approach. *Water Research* **41:** 3561-3574.

Lu, J., Idris, U., Harmon, B., Hofacre, C., Maurer, J. J., & Lee, M. D. (2003) Diversity and succession of the intestinal bacterial community of the maturing broiler chicken. *Applied and Environmental Microbiology* **69:** 6816-6824.

Ludwig, W., & Klenk, H. P. (2005) Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics. *Bergey's Manual® of Systematic Bacteriology* 49-66.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Buchner, A. et al. (2004) ARB: a software environment for sequence data. *Nucleic Acids Research* **32:** 1363-1371.

Ma, C., Wu, X., Nawaz, M., Li, J., Yu, P., Moore, J. E., & Xu, J. (2011) Molecular Characterization of Fecal Microbiota in Patients with Viral Diarrhea. *Current Microbiology* **63:** 259-266.

MacConnachie, A., Fox, R., Kennedy, D., & Seaton, R. (2009) Faecal transplant for recurrent *Clostridium difficile*-associated diarrhoea: a UK case series. *QJM: An International Journal of Medicine* **102:** 781-784.

Mackie, R. I., Sghir, A., & Gaskins, H. R. (1999) Developmental microbial ecology of the neonatal gastrointestinal tract. *American Journal of Clinical Nutrition* **69:** 1035s-1045s.

Macpherson, A. J., & Slack, E. (2007) The functional interactions of commensal bacteria with intestinal secretory IgA. *Current Opinion in Gastroenterology* **23:** 673-678.

Macpherson, A. J., & Harris, N. L. (2004) Interactions between commensal intestinal bacteria and the immune system. *Nature Reviews Immunology* **4:** 478-485.

Macpherson, A. J., Hunziker, L., McCoy, K., & Lamarre, A. (2001) IgA responses in the intestinal mucosa against pathogenic and non-pathogenic microorganisms. *Microbes and Infection* **3:** 1021-1035.

Magurran, A. E. (2004) Measuring biological diversity. Oxford: blackwell Publishing Ltd.

Magurran, A. E., & Magurran, A. E. (1988) Ecological diversity and its measurement. Princeton: Princeton University Press

Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., & Woese, C. R. (1997) The RDP (ribosomal database project). *Nucleic Acids Research* **25:** 109-110.

Maidak, B. L., Larsen, N., McCaughey, M. J., Overbeek, R., Olsen, G. J., Fogel, K. et al. (1994) The ribosomal database project. *Nucleic Acids Research* **22:** 3485-3487.

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R. et al. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences* **95:** 3140-3145.

Manges, A. R., Labbe, A., Loo, V. G., Atherton, J. K., Behr, M. A., Masson, L. et al. (2010) Comparative metagenomic study of alterations to the intestinal microbiota and risk of nosocomial *Clostridum difficile*-associated disease. *Journal of Infectious Diseases* **202:** 1877-1884.

Manichanh, C., Reeder, J., Gibert, P., Varela, E., Llopis, M., Antolin, M. et al. (2010) Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. *Genome Research* **20:** 1411-1419.

Marchesi, J. R., Sato, T., Weightman, A. J., Martin, T. A., Fry, J. C., Hiom, S. J., & Wade, W. G. (1998) Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. *Applied and Environmental Microbiology* **64:** 795-799.

Mardis, E. R. (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9:** 387-402.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A. et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Mariat, D., Firmesse, O., Levenez, F., Guimarães, V., Sokol, H., Dore, J. et al. (2009) The *Firmicutes/Bacteroidetes* ratio of the human microbiota changes with age. *BMC Microbiology* **9:** 123.

Martin, A. P. (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and Environmental Microbiology* **68:** 3673-3682.

Maslowski, K. M., Vieira, A. T., Ng, A., Kranich, J., Sierro, F., Yu, D. et al. (2009) Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. *Nature* **461:** 1282-1286.

Maukonen, J., Mättö, J., Suihko, M. L., & Saarela, M. (2008) Intra-individual diversity and similarity of salivary and faecal microbiota. *Journal of Medical Microbiology* **57:** 1560-1568.

Mazmanian, S. K., Round, J. L., & Kasper, D. L. (2008) A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature* **453:** 620-625.

Mazmanian, S. K., Liu, C. H., Tzianabos, A. O., & Kasper, D. L. (2005) An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* **122:** 107-118.

McCann, K. S. (2000) The diversity–stability debate. *Nature* **405:** 228-233.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A. et al. (2011) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal* **6:** 610-618.

McDonnell, G., & Russell, A. D. (1999) Antiseptics and disinfectants: activity, action, and resistance. *Clinical Microbiological Reviews* **12:** 147-179.

McFarland, L. V., Surawicz, C. M., & Stamm, W. E. (1990) Risk factors for *Clostridium difficile* carriage and C. difficile-associated diarrhea in a cohort of hospitalized patients. *Journal of Infectious Diseases* **162:** 678-684.

McKenna, P., Hoffmann, C., Minkah, N., Aye, P. P., Lackner, A., Liu, Z. et al. (2008) The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathogens* **4:** e20.

McLeod, C. (2009) History of DNA Sequencing & Current Applications. www.roche-applied-science.com

McOrist, A. L., Jackson, M., & Bird, A. R. (2002) A comparison of five methods for extraction of bacterial DNA from human faecal samples. *Journal of Microbiological Methods* **50:** 131-139.

Meier, H., Amann, R., Ludwig, W., & Schleifer, K. H. (1999) Specific oligonucleotide probes for in situ detection of a major group of gram-positive bacteria with low DNA G C content. *Systematic and Applied Microbiology* **22:** 186-196.

Menard, S., Candalh, C., Bambou, J., Terpend, K., Cerf-Bensussan, N., & Heyman, M. (2004) Lactic acid bacteria secrete metabolites retaining anti-inflammatory properties after intestinal transport. *Gut* **53:** 821-828.

Metchnikoff, E. (1907) Essais optimistes. Paris. The prolongation of life. Optimistic studies. Translated and edited by P. Chalmers Mitchell.

Metzker, M. L. (2009) Sequencing technologies—the next generation. *Nature Reviews Genetics* **11:** 31-46.

Militon, C., Rimour, S., Missaoui, M., Biderre, C., Barra, V., Hill, D. et al. (2007) PhylArray: phylogenetic probe design algorithm for microarray. *Bioinformatics* **23:** 2550-2557.

Miller, N., Gong, Q., Bryan, R., Ruvolo, M., Turner, L., & LaBrie, S. (2002) Cross-hybridization of closely related genes on high-density macroarrays. *BioTechniques* **32:** 620-625.

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 Research* **21:** 1616-1625.

Minton, N. P., & Clarke, D. J. (1989) *Clostridia.* New York: Springer Publishing.

Moeseneder, M. M., Arrieta, J. M., Muyzer, G., Winter, C., & Herndl, G. J. (1999) Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis. *Applied and Environmental Microbiology* **65:** 3518-3525.

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J., & Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution* **60:** 174-182.

Molecular Devices. (2005) Gene Pix Pro Version 6.0 User's Guide.

Moncrief, J. S., Barroso, L. A., & Wilkins, T. D. (1997) Positive regulation of *Clostridium difficile* toxins. *Infection and Immunity* **65:** 1105-1108.

Moore, W., & Holdeman, L. V. (1974) Human fecal flora: the normal flora of 20 Japanese-Hawaiians. *Applied Microbiology* **27:** 961-979.

Mora, J. R., Zielinski, T. L., Nelson, B. P., & Getts, R. C. (2008) 1, Genisphere, Inc., Hatfield, PA 2, GenTel BioSciences, Inc., Madison, WI, USA BioTechniques, Vol. 44, No. 6, May 2008, pp. 815–818. *BioTechniques* **44:** 815-818.

Morris, J. J., Johnson, Z. I., Szul, M. J., Keller, M., & Zinser, E. R. (2011) Dependence of the Cyanobacterium *Prochlorococcus* on Hydrogen Peroxide Scavenging Microbes for Growth at the Ocean's Surface. *PloS One* **6:** e16805.

Moyer, C. L., Dobbs, F. C., & Karl, D. M. (1994) Estimation of diversity and community structure through restriction fragment length polymorphism distribution analysis of bacterial 16S rRNA genes from a microbial mat at an active, hydrothermal vent system, Loihi Seamount, Hawaii. *Applied and Environmental Microbiology* **60:** 871-879.

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. *Applied and Environmental Microbiology* **72:** 1027-1033.

Mullard, A. (2008) Microbiology: the inside story. *Nature* **453:** 578-580.

Muyzer, G. (1999) DGGE/TGGE a method for identifying genes from natural ecosystems. *Current Opinion in Microbiology* **2:** 317-322.

Muyzer, G., & Smalla, K. (1998) Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. *Antonie Van Leeuwenhoek* **73:** 127-141.

Muyzer, G., De Waal, E. C., & Uitterlinden, A. G. (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology* **59:** 695-700.

Muyzer, G., Teske, A., Wirsen, C. O., & Jannasch, H. W. (1995) Phylogenetic relationships of *Thiomicrospira* species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments. *Archives of Microbiology* **164:** 165-172.

Mylonakis, E., Ryan, E. T., & Calderwood, S. B. (2001) *Clostridium difficile*-associated diarrhea: a review. *Archives of Internal Medicine* **161:** 525.

Naaber, P., Smidt, I., Štšepetova, J., Brilene, T., Annuk, H., & Mikelsaar, M. (2004) Inhibition of *Clostridium difficile* strains by intestinal *Lactobacillus* species. *Journal of Medical Microbiology* **53:** 551-554.

Naef, F., & Magnasco, M. O. (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Physical Review E* **68:** 011906.

Naef, F., Socci, N. D., & Magnasco, M. (2003) A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations. *Bioinformatics* **19:** 178-184.

Nagy-Szakal, D., & Kellermayer, R. (2011) The remarkable capacity for gut microbial and host interactions. *Gut Microbes* **2:** 178.

Narum, S. R. (2006) Beyond Bonferroni: less conservative analyses for conservation genetics. *Conservation Genetics* **7:** 783-787.

Nava, G. M., & Stappenbeck, T. S. (2011) Diversity of the autochthonous colonic microbiota. *Gut Microbes* **2:** 99-104.

Nawrocki, E. P., & Eddy, S. R. (2007) Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Computational Biology* **3:** e56.

Neish, A. S., Gewirtz, A. T., Zeng, H., Young, A. N., Hobert, M. E., Karmali, V. et al. (2000) Prokaryotic regulation of epithelial responses by inhibition of IκB-α ubiquitination. *Science* **289:** 1560-1563.

Neufeld, J. D., Li, J., & Mohn, W. W. (2008) Scratching the surface of the rare biosphere with ribosomal sequence tag primers. *FEMS Microbiology Letters* **283:** 146-153.

Norén, T. (2010) *Clostridium difficile* and the disease it causes. *Methods in Molecular Biology* **646:** 9-35.

Noverr, M. C., & Huffnagle, G. (2005) The 'microflora hypothesis' of allergic diseases. *Clinical & Experimental Allergy* **35:** 1511-1520.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27:** 29-34.

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 *φB124-14*. *PLoS One* **7:** e35053.

O'Hara, A. M., & Shanahan, F. (2006) The gut flora as a forgotten organ. *EMBO Reports* **7:** 688-693.

Okada, H., Kuhn, C., Feillet, H., & Bach, J. F. (2010) The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. *Clinical & Experimental Immunology* **160:** 1-9.

Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R., & Stahl, D. A. (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annual Reviews in Microbiology* **40:** 337-365.

Olsen, G. J., Overbeek, R., Larsen, N., Marsh, T. L., McCaughey, M. J., Maciukenas, M. A. et al. (1992) The ribosomal database project. *Nucleic Acids Research* **20:** 2199.

O'Neill, G., Ogunsola, F., Brazier, J., & Duerden, B. (1996) Modification of a PCR Ribotyping Method for Application as a Routine Typing Scheme for *Clostridium difficile*. *Anaerobe* **2:** 205-209.

Osborn, A. M. (2005) Molecular microbial ecology. New York: Taylor and Francis Group.

Osborn, A. M., Moore, E. R. B., & Timmis, K. N. (2000) An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. *Environmental Microbiology* **2:** 39-50.

O'Toole, P. W., & Claesson, M. J. (2010) Gut microbiota: changes throughout the lifespan from infancy to elderly. *International Dairy Journal* **20:** 281-291.

Ott, S. J., Musfeldt, M., Ullmann, U., Hampe, J., & Schreiber, S. (2004) Quantification of intestinal bacterial populations by real-time PCR with a universal primer set and minor groove binder probes: a global approach to the enteric flora. *Journal of Clinical Microbiology* **42:** 2566-2572.

Pace, N. R. (1997) A molecular view of microbial diversity and the biosphere. *Science* **276:** 734-740.

Pallen, M. J., Loman, N. J., & Penn, C. W. (2010) High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Current Opinion in Microbiology* **13:** 625-631.

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., & Brown, P. O. (2007) Development of the human infant intestinal microbiota. *PLoS Biology* **5:** e177.

Palmer, C., Bik, E. M., Eisen, M. B., Eckburg, P. B., Sana, T. R., Wolber, P. K. et al. (2006) Rapid quantitative profiling of complex microbial populations. *Nucleic Acids Research* **34:** e5-e5.

Park, H., Shim, S., Kim, S., Park, J., Park, S., Kim, H. et al. (2005) Molecular analysis of colonized bacteria in a human newborn infant gut. *Journal of Microbiology* **43:** 345-353.

Parracho, H. M. R. T., Bingham, M. O., Gibson, G. R., & McCartney, A. L. (2005) Differences between the gut microflora of children with autistic spectrum disorders and that of healthy children. *Journal of Medical Microbiology* **54:** 987-991.

Paulsen, I., Banerjei, L., Myers, G., Nelson, K., Seshadri, R., Read, T. et al. (2003) Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science* **299:** 2071-2074.

Peng, L., Li, Z. R., Green, R. S., Holzman, I. R., & Lin, J. (2009) Butyrate enhances the intestinal barrier by facilitating tight junction assembly via activation of AMP-activated protein kinase in Caco-2 cell monolayers. *Journal of Nutrition* **139:** 1619-1625.

Peplies, J., Glöckner, F. O., & Amann, R. (2003) Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. *Applied and Environmental Microbiology* **69:** 1397-1407.

Peterson, A. W., Heaton, R. J., & Georgiadis, R. M. (2001) The effect of surface probe density on DNA hybridization. *Nucleic Acids Research* **29:** 5163-5168.

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A. et al. (2009) The NIH human microbiome project. *Genome Research* **19:** 2317-2323.

Peura, S., Eiler, A., Hiltunen, M., Nykänen, H., Tiirola, M., & Jones, R. I. (2012) Bacterial and Phytoplankton Responses to Nutrient Amendments in a Boreal Lake Differ According to Season and to Taxonomic Resolution. *PloS One* **7:** e38552.

Pilloni, G., Granitsiotis, M. S., Engel, M., & Lueders, T. (2012) Testing the limits of 454 pyrotag sequencing: reproducibility, quantitative assessment and comparison to t-RFLP fingerprinting of aquifer microbes. *PloS One* **7:** e40467.

Polz, M. F., & Cavanaugh, C. M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology* **64:** 3724-3730.

Pond, S. K., Wadhawan, S., Chiaromonte, F., Ananda, G., Chung, W. Y., Taylor, J., & Nekrutenko, A. (2009) Windshield splatter analysis with the Galaxy metagenomic pipeline. *Genome Research* **19:** 2144-2153.

Poulsen, L., Søe, M. J., Snakenborg, D., Møller, L. B., & Dufva, M. (2008) Multi-stringency wash of partially hybridized 60-mer probes reveals that the stringency along the probe decreases with distance from the microarray surface. *Nucleic Acids Research* **36:** e132-e132.

Poxton, I. R. (2005) *Clostridium difficile*. *Journal of Medical Microbiology* **54:** 97-100.

Poxton, I., McCoubrey, J., & Blair, G. (2001) The pathogenicity of *Clostridium difficile. Clinical Microbiology and Infection* **7:** 421-427.

Pozhitkov, A., Noble, P. A., Domazet-Lošo, T., Nolte, A. W., Sonnenberg, R., Staehler, P. et al. (2006) Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. *Nucleic Acids Research* **34:** e66-e66.

Pozhitkov, A. E., Tautz, D., & Noble, P. A. (2007) Oligonucleotide microarrays: widely applied—poorly understood. *Briefings in Functional Genomics & Proteomics* **6:** 141-148.

Pozhitkov, A. E., Stedtfeld, R. D., Hashsham, S. A., & Noble, P. A. (2007) Revision of the nonequilibrium thermal dissociation and stringent washing approaches for identification of mixed nucleic acid targets by microarrays. *Nucleic Acids Research* **35:** e70.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5:** e9490.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* **26:** 1641-1650.

Probert, H., & Gibson, G. (2002) Bacterial biofilms in the human gastrointestinal tract. *Curr Issues in Intestinal Microbiology* **3:** 23-27.

Promega Corporation. (1998) pGEM®-T and pGEM®-T Easy Vector Systems. Technical Manual

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35:** 7188-7196.

Purvis, A., & Hector, A. (2000) Getting the measure of biodiversity. *Nature* **405:** 212-219.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C. et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464:** 59-65.

Qiu, X., Wu, L., Huang, H., McDonel, P. E., Palumbo, A. V., Tiedje, J. M., & Zhou, J. (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Applied and Environmental Microbiology* **67:** 880-887.

Quince, C., Lanzen, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M. et al. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods* **6:** 639-641.

Raes, J., & Bork, P. (2008) Molecular eco-systems biology: towards an understanding of community function. *Nature Reviews Microbiology* **6:** 693-699.

Rafii, F., Sutherland, J. B., & Cerniglia, C. E. (2008) Effects of treatment with antimicrobial agents on the human colonic microflora. *Journal of Therapeutics and Clinical Risk Management* **4:** 1343-1358

Rajilić-Stojanović, M., Smidt, H., & De Vos, W. M. (2007) Diversity of the human gastrointestinal tract microbiota revisited. *Environmental Microbiology* **9:** 2125-2136.

Rajilić-Stojanović, M., Heilig, H. G. H. J., Molenaar, D., Kajander, K., Surakka, A., Smidt, H., & De Vos, W. M. (2009) Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environmental Microbiology* **11:** 1736-1751.

Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S., & Medzhitov, R. (2004) Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell* **118:** 229-241.

Ramette, A. (2007) Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology* **62:** 142-160.

Randolph, J. B., & Waggoner, A. S. (1997) Stability, specificity and fluorescence brightness of multiply-labeled fluorescent DNA probes. *Nucleic Acids Res* **25:** 2923-2929.

Rawls, J. F., Mahowald, M. A., Ley, R. E., & Gordon, J. I. (2006) Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell* **127:** 423-433.

Rea, M. C., O'Sullivan, O., Shanahan, F., O'Toole, P. W., Stanton, C., Ross, R. P., & Hill, C. (2012) Clostridium difficile carriage in elderly subjects and associated changes in the intestinal microbiota. *Journal of Clinical Microbiology* **50:** 867-875.

Reeder, J., & Knight, R. (2010) Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. *Nature Methods* **7:** 668-669

Reeder, J., & Knight, R. (2009) The 'rare biosphere': a reality check. *Nature Methods* **6:** 636-637.

Regier, J. C., & Shi, D. (2005) Increased yield of PCR product from degenerate primers with nondegenerate, nonhomologous 5′ tails. *BioTechniques* **38:** 34, 36, 38.

Relman, D. A., Schmidt, T. M., MacDermott, R. P., & Falkow, S. (1992) Identification of the uncultured bacillus of Whipple's disease. *New England Journal of Medicine* **327:** 293-301.

Relógio, A., Schwager, C., Richter, A., Ansorge, W., & Valcárcel, J. (2002) Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Research* **30:** e51-e51.

Reysenbach, A. L., Wickham, G. S., & Pace, N. R. (1994) Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Applied and Environmental Microbiology* **60:** 2113-2119.

Reysenbach, A. L., Giver, L. J., Wickham, G. S., & Pace, N. R. (1992) Differential amplification of rRNA genes by polymerase chain reaction. *Applied and Environmental Microbiology* **58:** 3417-3418.

Reysenbach, A., & Pace, N. (1995) Reliable amplification of hyperthermophilic archaeal 16S rRNA genes by the polymerase chain reaction. 101–105 in Robb F. T., Place A. R., editors. Archaea: a laboratory manual. New York: Cold Spring Harbor Laboratory Press

Riesenfeld, C. S., Schloss, P. D., & Handelsman, J. (2004) Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics* **38:** 525-552.

Riley, T. (1998) *Clostridium difficile*: a pathogen of the nineties. *European Journal of Clinical Microbiology & Infectious Diseases* **17:** 137-141.

Rinttilä, T., Kassinen, A., Malinen, E., Krogius, L., & Palva, A. (2004) Development of an extensive set of 16S rDNA-targeted primers for quantification of pathogenic and indigenous bacteria in faecal samples by real-time PCR. *Journal of Applied Microbiology* **97:** 1166-1177.

Roche Life Sciences. (2011) 454 Sequencing System Guidelines for Amplicon Experimental Design.

Roche Life Sciences. (2009) Amplicon Fusion Primer Design Guidelines.

Roche Life Sciences. (2009) GS FLX Titanium General Library Preparation Method Manual.

Roche Life Sciences. (2007) GS FLX Amplicon DNA Library Preparation Method Manual.

Roitt, I. (1994) Essential Immunology. 8th Edition. Oxford: Blackwell Scientific Publications.

Rolfe, R. D., Helebian, S., & Finegold, S. M. (1981) Bacterial interference between *Clostridium difficile* and normal fecal flora. *Journal of Infectious Diseases* **143:** 470-475.

Ronaghi, M. (2000) Improved performance of pyrosequencing using single-stranded DNA-binding protein. *Analytical Biochemistry* **286:** 282-288.

Roth, C. M., & Yarmush, M. L. (1999) Nucleic acid biotechnology. *Annual Review of Biomedical Engineering* **1:** 265-297.

Round, J. L., & Mazmanian, S. K. (2009) The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology* **9:** 313-323.

Rudi, K., Skulberg, O. M., Larsen, F., & Jakobsen, K. S. (1997) Strain characterization and classification of oxyphotobacteria in clone cultures on the basis of 16S rRNA sequences from the variable regions V6, V7, and V8. *Applied and Environmental Microbiology* **63:** 2593-2599.

Rupnik, M., Pabst, S., Rupnik, M., von Eichel-Streiber, C., Urlaub, H., & Söling, H. D. (2005) Characterization of the cleavage site and function of resulting cleavage fragments after limited proteolysis of *Clostridium difficile* toxin B (TcdB) by host cells. *Microbiology* **151:** 199-208.

Rupnik, M., Dupuy, B., Fairweather, N. F., Gerding, D. N., Johnson, S., Just, I. et al. (2005) Revised nomenclature of *Clostridium difficile* toxins and associated genes. *Journal of Medical Microbiology* **54:** 113-117.

Rutala, W. A., & Weber, D. J. (2004) Disinfection and sterilization in health care facilities: what clinicians need to know. *Clinical Infectious Diseases* **39:** 702-709.

Sachs, J., & Hollowell, A. (2012) The Origins of Cooperative Bacterial Communities. *MBio* **3:** e00099-12.

Saiki, R., Gelfand, D., Stoffel, S., Scharf, S., Higuchi, R., Horn, G. et al. (1988) Primer-directed enzymatic amplification of DNA. *Science* **239:** 487-491.

Salonen, A., Nikkilä, J., Jalanka-Tuovinen, J., Immonen, O., Rajilić-Stojanović, M., Kekkonen, R. A. et al. (2010) Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *Journal of Microbiological Methods* **81:** 127-134.

Sambrook, J., & Russell, D. W. (2001) Molecular cloning: a laboratory manual. New York: CSHL press.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74:** 5463-5467.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E. *et al*., (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **39** (suppl 1): D38-D51.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467-470.

Schloss, P. D. (2009) A high-throughput DNA sequence aligner for microbial ecology studies. *PLoS One* **4:** e8230.

Schloss, P. D. (2008) Evaluating different approaches that test whether microbial communities have the same structure. *The ISME Journal* **2:** 265-275.

Schloss, P. D., & Handelsman, J. (2006a) Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Applied and Environmental Microbiology* **72:** 6773-6779.

Schloss, P. D., & Handelsman, J. (2006b) Introducing TreeClimber, a test to compare microbial community structures. *Applied and Environmental Microbiology* **72:** 2379-2384.

Schloss, P. D., & Handelsman, J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology* **71:** 1501-1506.

Schloss, P. D., & Handelsman, J. (2004a) Status of the microbial census. *Microbiology and Molecular Biology Reviews* **68:** 686-691.

Schloss, P. D., & Handelsman, J. (2004b) Toward a census of bacteria in soil. *PLoS Computational Biology* **2:** e92.

Schloss, P. D., Larget, B. R., & Handelsman, J. (2004) Integration of microbial ecology and statistics: a test to compare gene libraries. *Applied and Environmental Microbiology* **70:** 5485-5492.

Schloss, P. D., Schubert, A. M., Zackular, J. P., Iverson, K. D., Young, V. B., & Petrosino, J. F. (2012) Stabilization of the murine gut microbiome following weaning. *Gut Microbes* **3:** 0--1.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B. et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75:** 7537-7541.

Schmalenberger, A., Schwieger, F., & Tebbe, C. C. (2001) Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling. *Applied and Environmental Microbiology* **67:** 3557-3563.

Schuster, S. C. (2008) Next-generation sequencing transforms today's biology. *Nature Methods* **5:** 16-17,18.

Schwartz, R. L., & Phoenix, T. (2008) Learning Perl. 4th Edition. Sebastopol: O'Reilly Media, Inc.

Scupham, A. J. (2007) Succession in the intestinal microbiota of preadolescent turkeys. *FEMS Microbiology Ecology* **60:** 136-147.

Scupham, A. J., Patton, T. G., Bent, E., & Bayles, D. O. (2008) Comparison of the cecal microbiota of domestic and wild turkeys. *Microbiology Ecology* **56:** 322-331.

Scupham, A. J., Presley, L. L., Wei, B., Bent, E., Griffith, N., McPherson, M. et al. (2006) Abundant and diverse fungal microbiota in the murine intestine. *Applied and Environmental Microbiology* **72:** 793-801.

Seaby, R., & Henderson, P. (2007) Community analysis package 4.0. Lymington: PISCES Conservation Ltd.

Sebaihia, M., & Thomson, N. (2006) Colonic irritation. *Nature Reviews Microbiology* **4:** 882-883.

Sebaihia, M., Wren, B. W., Mullany, P., Fairweather, N. F., Minton, N., Stabler, R. et al. (2006) The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat Genet* **38:** 779-786.

Segal, S., & Hill, A. V. S. (2003) Genetic susceptibility to infectious disease. *Trends in Microbiology* **11:** 445-448.

Sekirov, I., & Finlay, B. B. (2009) The role of the intestinal microbiota in enteric infection. *Journal of Physiology (London)* **587:** 4159-4167.

Sekirov, I., Tam, N. M., Jogova, M., Robertson, M. L., Li, Y., Lupp, C., & Finlay, B. B. (2008) Antibiotic-induced perturbations of the intestinal microbiota alter host susceptibility to enteric infection. *Infection and Immunity* **76:** 4726-4736.

Sergeant, M. J., Constantinidou, C., Cogan, T., Penn, C. W., & Pallen, M. J. (2012) High-Throughput Sequencing of 16S rRNA Gene Amplicons: Effects of Extraction Procedure, Primer Length and Annealing Temperature. *PloS One* **7:** e38094.

Sghir, A., Gramet, G., Suau, A., Rochet, V., Pochart, P., & Dore, J. (2000) Quantification of bacterial groups within human fecal flora by oligonucleotide probe hybridization. *Applied and Environmental Microbiology* **66:** 2263-2266.

Shanahan, F. (2002) The host–microbe interface within the gut. *Best Practice & Research Clinical Gastroenterology* **16:** 915-931.

Shannon, C. E. & Weaver, W. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal* **27:** 379–423.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D. et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13:** 2498-2504.

Shen, X. J., Rawls, J. F., Randall, T. A., Burcall, L., Mpande, C., Jenkins, N. et al. (2010) Molecular characterization of mucosal adherent bacteria and associations with colorectal adenomas. *Gut Microbes* **1:** 138-147.

Shim, J. K., Johnson, S., Samore, M. H., Bliss, D. Z., & Gerding, D. N. (1998) Primary symptomless colonisation by *Clostridium difficile* and decreased risk of subsequent diarrhoea. *Lancet* **351:** 633.

Simpson, E. H. (1949) Measurement of diversity. *Nature* **163:** 688

Singleton, D. R., Furlong, M. A., Rathbun, S. L., & Whitman, W. B. (2001) Quantitative comparisons of 16S rDNA sequence libraries from environmental samples. *Applied Environmental Microbiology* **67:** 4373-4376.

Sipos, R., Székely, A. J., Palatinszky, M., Révész, S., Márialigeti, K., & Nikolausz, M. (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiology Ecology* **60:** 341-350.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P-L., & Ideker, T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**: 431-432.

Smith, W., Solow, A. R., & Preston, P. E. (1996) An estimator of species overlap using a modified beta-binomial model. *Biometrics* 1472-1477.

Snyder, L., & Champness, W. (2007) Molecular genetics of bacteria. 2nd Edition. Washington: American Society for Microbiology.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R. et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences* **103:** 12115-12120.

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L. G., Gratadoux, J. J. et al. (2008) *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proceedings of the National Academy of Sciences* **105:** 16731-16736.

Sokurenko, E. V., Chesnokova, V., Dykhuizen, D. E., Ofek, I., Wu, X. R., Krogfelt, K. A. et al. (1998) Pathogenic adaptation of *Escherichia coli* by natural variation of the FimH adhesin. *Proceedings of the National Academy of Sciences* **95:** 8922-8926.

Song, H. J., Shim, K. N., Jung, S., Choi, H. J., Lee, M., Ryu, K. H. et al. (2008) Antibiotic-associated diarrhea: candidate organisms other than *Clostridium difficile*. *Korean Journal of Internal Medicine* **23:** 9-15.

Song, Y., Liu, C., McTeague, M., Vu, A., Liu, J. Y., & Finegold, S. M. (2003) Rapid identification of Gram-positive anaerobic coccal species originally classified in the genus *Peptostreptococcus* by multiplex PCR assays using genus-and species-specific primers. *Microbiology* **149:** 1719-1727.

Sonnenburg, E. D., Zheng, H., Joglekar, P., Higginbottom, S. K., Firbank, S. J., Bolam, D. N., & Sonnenburg, J. L. (2010) Specificity of polysaccharide use in intestinal *Bacteroides* species determines diet-induced microbiota alterations. *Cell* **141:** 1241-1252.

Sonnenburg, J. L., Xu, J., Leip, D. D., Chen, C. H., Westover, B. P., Weatherford, J. et al. (2005) Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science* **307:** 1955-1959.

Sørensen, K. B., Canfield, D. E., Teske, A. P., & Oren, A. (2005) Community composition of a hypersaline endoevaporitic microbial mat. *Applied and Environmental Microbiology* **71:** 7352-7365.

Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter* **5:** 1-34.

Sorokin, N., Chechetkin, V., Pan'kov, S., Somova, O., Livshits, M., Donnikov, M. et al. (2006) Kinetics of hybridization on surface oligonucleotide microchips: theory, experiment, and comparison with hybridization on gel-based microchips. *Journal of Biomolecular Structure and Dynamics* **24:** 57-66.

Sousa, T., Paterson, R., Moore, V., Carlsson, A., Abrahamsson, B., & Basit, A. W. (2008) The gastrointestinal microbiota as a site for the biotransformation of drugs. *International Journal of Pharmaceutics* **363:** 1-25.

Southern, E., Mir, K., & Shchepinov, M. (1999) Molecular interactions on microarrays. *Nature Genetics* **21:** 5-9.

Southern, E. (1975) Rapid transfer of DNA from agarose gels to nylon membranes. *Journal of Molecular Biology* **98:** 503-517.

Southwood, T., & Henderson, P. (2000) Ecological methods. 3rd Edition. Oxford: Blackwell Science Ltd.

Speksnijder, A. G. C. L., Kowalchuk, G. A., De Jong, S., Kline, E., Stephen, J. R., & Laanbroek, H. J. (2001) Microvariation artifacts introduced by PCR and cloning of closely related 16S rRNA gene sequences. *Applied and Environmental Microbiology* **67:** 469-472.

Spor, A., Koren, O., & Ley, R. (2011) Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology* **9:** 279-290.

Stackebrandt, E., & Ebers, J. (2006) Taxonomic parameters revisited: tarnished gold standards. *Microbiology Today* **33:** 152-155

Stackebrandt, E., & Goebel, B. (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic Bacteriology* **44:** 846-849.

Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A. D., Peter, K., Maiden, M. C. J. et al. (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology* **52:** 1043-1047.

Stahl, D. A., Flesher, B., Mansfield, H. R., & Montgomery, L. (1988) Use of phylogenetically based hybridization probes for studies of ruminal microbial ecology. *Applied and Environmental Microbiology* **54:** 1079-1084.

Staley, J. T. (2006) The bacterial species dilemma and the genomic–phylogenetic species concept. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361:** 1899-1909.

Stears, R. L., Getts, R. C., & Gullans, S. R. (2000) A novel, sensitive detection system for high-density microarrays using dendrimer technology. *Physiological Genomics* **3:** 93-99.

Stevens, A., & Lowe, J. (1997) Histology. London: Gower Medical Publishing.

Stoddart, B., & Wilcox, M. H. (2002) *Clostridium difficile*. *Current Opinion in Infectious Diseases* **15:** 513-518.

Suau, A., Gibson, G. R., & Collins, M. D. (2002) Differences in rDNA libraries of faecal bacteria derived from 10-and 25-cycle PCRs. *International Journal of Systematic and Evolutionary Microbiology* **52:** 757-763.

Suau, A., Bonnet, R., Sutren, M., Godon, J. J., Gibson, G. R., Collins, M. D., & Doré, J. (1999) Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Applied and Environmental Microbiology* **65:** 4799-4807.

Suau, A., Rochet, V., Sghir, A., Gramet, G., Brewaeys, S., Sutren, M. et al. (2001) *Fusobacterium prausnitzii* and Related Species Represent a Dominant Group Within the Human Fecal Flora. *Systematic and Applied Microbiology* **24:** 139-145.

Sullivan, Å., Edlund, C., & Nord, C. E. (2001) Effect of antimicrobial agents on the ecological balance of human microflora. *The Lancet Infectious Diseases* **1:** 101-114.

Sun, X., Wang, H., Zhang, Y., Chen, K., Davis, B., & Feng, H. (2011) Mouse relapse model of *Clostridium difficile* infection. *Infection and Immunity* **79:** 2856-2864.

Surawicz, C. M. (2003) Probiotics, antibiotic-associated diarrhoea and *Clostridium difficile* diarrhoea in humans. *Best Practice in Research and Clinical Gastroenterology* **17:** 775-783.

Suzuki, M. T., & Giovannoni, S. J. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology* **62:** 625-630.

Talpaert, M. J., Rao, G. G., Cooper, B. S., & Wade, P. (2011) Impact of guidelines and enhanced antibiotic stewardship on reducing broad-spectrum antibiotic usage and its effect on incidence of *Clostridium difficile* infection. *Journal of Antimicrobial Chemotherapy* **66:** 2168-2174.

Teske, A., Hinrichs, K. U., Edgcomb, V., de Vera Gomez, A., Kysela, D., Sylva, S. P. et al. (2002) Microbial diversity of hydrothermal sediments in the Guaymas Basin: evidence for anaerobic methanotrophic communities. *Applied and Environmental Microbiology* **68:** 1994-2007.

Tewhey, R., Warner, J. B., Nakano, M., Libby, B., Medkova, M., David, P. H. et al. (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology* **27:** 1025-1031.

Thomas, C. M., & Versalovic, J. (2010) Probiotics-host communication: modulation of signaling pathways in the intestine. *Gut Microbes* **1:** 1-15

Thompson, C. L., Hofer, M. J., Campbell, I. L., & Holmes, A. J. (2010) Community dynamics in the mouse gut microbiota: a possible role for IRF9-regulated genes in community homeostasis. *PloS One* **5:** e10335.

Tomiuk, S., & Hofmann, K. (2001) Microarray probe selection strategies. *Briefings in Bioinformatics* **2:** 329-340.

Tonna, I., & Welsby, P. (2005) Pathogenesis and treatment of *Clostridium difficile* infection. *Postgraduate Medical Journal* **81:** 367.

Topping, D. L., & Clifton, P. M. (2001) Short-chain fatty acids and human colonic function: roles of resistant starch and nonstarch polysaccharides. *Physiological Reviews* **81:** 1031-1064.

Trüper, H. (1992) Prokaryotes: an overview with respect to biodiversity and environmental importance. *Biodiversity and Conservation* **1:** 227-236.

Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., & Gordon, J. I. (2009) The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine* **1:** 6ra14.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007) The human microbiome project. *Nature* **449:** 804-810.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444:** 1027-1131.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E. et al. (2008) A core gut microbiome in obese and lean twins. *Nature* **457:** 480-484.

Turner, S., Pryer, K. M., Miao, V. P. W., & Palmer, J. D. (1999) Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis 1. *Journal of Eukaryotic Microbiology* **46:** 327-338.

Tyrer, P., Foxwell, A. R., Cripps, A. W., Apicella, M. A., & Kyd, J. M. (2006) Microbial pattern recognition receptors mediate M-cell uptake of a gram-negative bacterium. *Infection and Immunity* **74:** 625-631.

Valinsky, L., Della Vedova, G., Jiang, T., & Borneman, J. (2002a) Oligonucleotide fingerprinting of rRNA genes for analysis of fungal community composition. *Applied and Environmental Microbiology* **68:** 5999-6004.

Valinsky, L., Della Vedova, G., Scupham, A. J., Alvey, S., Figueroa, A., Yin, B. et al. (2002b) Analysis of bacterial community composition by oligonucleotide fingerprinting of rRNA genes. *Applied and Environmental Microbiology* **68:** 3243-3250.

Van den Bogert, B., de Vos, W. M., Zoetendal, E. G., & Kleerebezem, M. (2011) Microarray analysis and barcoded pyrosequencing provide consistent microbial profiles depending on the source of human intestinal samples. *Applied and Environmental Microbiology* **77:** 2071-2080.

Van der Waaij, D., Berghuis-de Vries, J. M., & Lekkerkerk-Van der Wees, J. (1971) Colonization resistance of the digestive tract in conventional and antibiotic-treated mice. *Journal of Hygiene* **69:** 405-411.

Van der Waaij, L. A., Harmsen, H. J. M., Madjipour, M., Kroese, F. G. M., Zwiers, M., Van Dullemen, H. et al. (2006) Bacterial population analysis of human colon and terminal ileum biopsies with 16S rRNA-based fluorescent probes: Commensal bacteria live in suspension and have no direct contact with epithelial cells. *Inflammatory Bowel Diseases* **11:** 865-871.

Van der Wielen, P. W. J. J., Lipman, L. J. A., Van Knapen, F., & Biesterveld, S. (2002) Competitive exclusion of *Salmonella enterica* serovar Enteritidis by *Lactobacillus crispatus* and *Clostridium lactatifermentans* in a sequencing fed-batch culture. *Applied and Environmental Microbiology* **68:** 555-559.

Van Tongeren, S. P., Slaets, J. P. J., Harmsen, H., & Welling, G. W. (2005) Fecal microbiota composition and frailty. *Applied and Environmental Microbiology* **71:** 6438-6442.

Vanhoutte, T., Huys, G., Brandt, E., & Swings, J. (2006) Temporal stability analysis of the microbiota in human feces by denaturing gradient gel electrophoresis using universal and group-specific 16S rRNA gene primers. *FEMS Microbiology Ecology* **48:** 437-446.

Various. (2008-2013) Issues with 454 pyrosequencing. http://seqanswers.com/

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. et al. (2001) The sequence of the human genome. *Science* **291:** 1304-1351

Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I., & Marden, J. H. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Moleular Ecology* **17:** 1636-1647.

Vinolo, M. A. R., Rodrigues, H. G., Hatanaka, E., Sato, F. T., Sampaio, S. C., & Curi, R. (2011) Suppressive effect of short-chain fatty acids on production of proinflammatory mediators by neutrophils. *J Nutr Biochem* **22:** 849-855.

Vollaard, E., & Clasener, H. (1994) Colonization resistance. *Antimicrobial Agents and Chemotherapy* **38:** 409.

Voth, D. E., & Ballard, J. D. (2005) *Clostridium difficile* toxins: mechanism of action and role in disease. *Clinical Microbiological Reviews* **18:** 247-263.

Wagner, M., Smidt, H., Loy, A., & Zhou, J. (2007) Unravelling microbial communities with DNA-microarrays: challenges and future directions. *Microbiology Ecology* **53:** 498-506.

Walker, A. W., Sanderson, J. D., Churcher, C., Parkes, G. C., Hudspith, B. N., Rayment, N. et al. (2011) High-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. *BMC Microbiology* **11:** 7.

Walker, A. W., Ince, J., Duncan, S. H., Webster, L. M., Holtrop, G., Ze, X. et al. (2010) Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *The ISME Journal* **5:** 220-230.

Walker, C. B. (2007) Selected antimicrobial agents: mechanisms of action, side effects and drug interactions. *Periodontology 2000* **10:** 12-28.

Walter, J., Britton, R. A., & Roos, S. (2011) Host-microbial symbiosis in the vertebrate gastrointestinal tract and the *Lactobacillus reuteri* paradigm. *Proceedings of the National Academy of Sciences* **108:** 4645-4652.

Wang, M., Ahrné, S., Jeppsson, B., & Molin, G. (2005) Comparison of bacterial diversity along the human intestinal tract by direct cloning and sequencing of 16S rRNA genes. *FEMS Microbiology Ecology* **54:** 219-231.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73:** 5261-5267.

Wang, R. F., Cao, W. W., & Cerniglia, C. E. (1996) PCR detection and quantitation of predominant anaerobic bacteria in human and animal fecal samples. *Applied and Environmental Microbiology* **62:** 1242-1247.

Wang, X., Heazlewood, S., Krause, D., & Florin, T. (2003) Molecular characterization of the microbial species that colonize human ileal and colonic mucosa by using 16S rDNA sequence analysis. *Journal of Applied Microbiology* **95:** 508-520.

Wang, Y., & Qian, P. Y. (2009) Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS One* **4:** e7401.

Watanabe, K., Hamamura, N., & Kaku, N. (2004) Molecular identification of microbial populations in petroleum-contaminated groundwater. *Methods in Biotechnology* **16:** 235-242.

Wayne, L., Brenner, D., Colwell, R., Grimont, P., Kandler, O., Krichevsky, M. et al. (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic Bacteriology* **37:** 463-464.

Wehkamp, J., Harder, J., Weichenthal, M., Schwab, M., Schäffeler, E., Schlee, M. et al. (2004) NOD2 (CARD15) mutations in Crohn's disease are associated with diminished mucosal α-defensin expression. *Gut* **53:** 1658-1664.

Wei, D., Mona A, S., Olgica, M., & Richard G, B. (2009) Compressive sensing DNA microarrays. *EURASIP Journal on Bioinformatics and Systems Biology* **2009:** 168284

Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991) 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology* **173:** 697-703.

Weller, R., Glöckner, F. O., & Amann, R. (2000) 16S rRNA-Targeted Oligonucleotide Probes for the *in situ* Detection of Members of the Phylum *Cytophaga-Flavobacterium-Bacteroides*. *Systematic and Applied Microbiology* **23:** 107-114.

Wetmur, J. G., & Fresco, J. (1991) DNA probes: applications of the principles of nucleic acid hybridization. *Critical Reviews in Biochemistry and Molecular Biology* **26:** 227-259.

White, J. R., Arze, C., Matalka, M., Team, T., & White, O. (2011) CloVR-16S: Phylogenetic microbial community composition analysis based on 16S ribosomal RNA

amplicon sequencing–standard operating procedure, version1.0. *Nature Precedings* 10.1038

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998) Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences* **95:** 6578-6583.

White, J. R., Nagarajan, N., & Pop, M. (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Computational Biology* **5:** e1000352.

Whittaker, R. H. (1972) Evolution and measurement of species diversity. *Taxon* 213-251.

Wiley, G., Macmil, S., Qu, C., Wang, P., Xing, Y., White, D. et al. (2009) Methods for Generating Shotgun and Mixed Shotgun/Paired-End Libraries for the 454 DNA Sequencer. *Current Protocols in Human Genetics* **61:** 18.1.1-18.1.21

Willey, J., Sherwood, L., & Woolverton, C. (2008) Prescott, Harley, and Klein's Microbiology-7th international Editiond. New York: McGraw-Hill Higher Education

Wilson, J., Schurr, M., LeBlanc, C., Ramamurthy, R., Buchanan, K., & Nickerson, C. (2002) Mechanisms of bacterial pathogenicity. *Postgraduate Medical Journal* **78:** 216-224.

Wilson, K., & Walker, J. (2010) Principles and techniques of biochemistry and molecular biology. Cambridge: Cambridge University Press.

Wilson, K. H., & Blitchington, R. B. (1996) Human colonic biota studied by ribosomal DNA sequence analysis. *Appl Environ Microbiol* **62:** 2273-2278.

Wilson, K. H., & Perini, F. (1988) Role of competition for nutrients in suppression of Clostridium difficile by the colonic microflora. *Infect Immun* **56:** 2610-2614.

Wilson, K. H., & Freter, R. (1986) Interaction of *Clostridium difficile* and *Escherichia coli* with microfloras in continuous-flow cultures and gnotobiotic mice. *Infect Immun* **54:** 354-358.

Wilson, K. H., Blitchington, R., & Greene, R. (1990) Amplification of bacterial 16S ribosomal DNA with polymerase chain reaction. *J Clin Microbiol* **28:** 1942-1946.

Wilson, K. H., Silva, J., & Fekety, F. R. (1981) Suppression of *Clostridium difficile* by normal hamster cecal flora and prevention of antibiotic-associated cecitis. *Infection and Immunity* **34:** 626-628.

Wilson, M. (2009) Bacteriology of humans: an ecological perspective. Oxford: Blackwell Publishing Ltd.

Wilson, M., McNab, R., & Henderson, B. (2002) Bacterial disease mechanisms: an introduction to cellular microbiology. Cambridge: Cambridge University Press.

Win, D. T. (2009) Chemical allergies/chemical sensitivities. *AU Journal of Technology* **12:** 245-250.

Wintzingerode, F. V., Göbel, U. B., & Stackebrandt, E. (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews* **21:** 213-229.

Wise, M. G., & Siragusa, G. R. (2005) Quantitative detection of *Clostridium perfringens* in the broiler fowl gastrointestinal tract by real-time PCR. *Applied and Environmental Microbiology* **71:** 3911-3916.

Woese, C. R. (1987) Bacterial evolution. *Microbiological Reviews* **51:** 221-271.

Woese, C. R., & Fox, G. E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences* **74:** 5088-5090.

Wolda, H. (1981) Similarity indices, sample size and diversity. *Oecologia* **50:** 296-302.

Woodmansey, E. J., McMurdo, M. E. T., Macfarlane, G. T., & Macfarlane, S. (2004) Comparison of compositions and metabolic activities of fecal microbiotas in young adults and in antibiotic-treated and non-antibiotic-treated elderly subjects. *Applied and Environmental Microbiology* **70:** 6113-6122.

Woodmansey, E. (2007) Intestinal bacteria and ageing. *Journal of Applied Microbiology* **102:** 1178-1186.

Wostmann, B. S., Bruckner-Kardoss, E., & Pleasants, J. R. (1982) Oxygen consumption and thyroid hormones in germfree mice fed glucose-amino acid liquid diet. *Journal of Nutrition* **112:** 552-559.

Wren, B. (2006) A family of clostridial and streptococcal ligand-binding proteins with conserved C-terminal repeat sequences. *Molecular Microbiology* **5:** 797-803.

Wu, G. D., Lewis, J. D., Hoffmann, C., Chen, Y. Y., Knight, R., Bittinger, K. et al. (2010) Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiology* **10:** 206.

Wuyts, J., Van de Peer, Y., & De Wachter, R. (2001) Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. *Nucleic Acids Research* **29:** 5017-5028.

Xu, J., & Gordon, J. I. (2003) Honor thy symbionts. *Proceedings of the National Academy of Sciences* **100:** 10452-10459.

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.

Xu, J., Mahowald, M. A., Ley, R. E., Lozupone, C. A., Hamady, M., Martens, E. C. et al. (2007) Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biology* **5:** e156.

Yachi, S., & Loreau, M. (1999) Biodiversity and ecosystem productivity in a fluctuating environment: the insurance hypothesis. *Proceedings of the National Academy of Sciences* **96:** 1463-1468.

Yamamoto, S., & Harayama, S. (1995) PCR amplification and direct sequencing of gyrB genes with universal primers and their application to the detection and taxonomic analysis of *Pseudomonas putida* strains. *Applied and Environmental Microbiology* **61:** 1104-1109.

Yamamoto-Osaki, T., Kamiya, S., Sawamura, S., Kai, M., & Ozawa, A. (1994) Growth inhibition of *Clostridium difficile* by intestinal flora of infant faeces in continuous flow culture. *Journal of Medical Microbiology* **40:** 179-187.

Yan, F., & Polk, D. B. (2002) Probiotic bacterium prevents cytokine-induced apoptosis in intestinal epithelial cells. *The Journal Of Biological Chemistry* **277:** 50959–50965

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.

Yin, B., Valinsky, L., Gao, X., Becker, J. O., & Borneman, J. (2003) Bacterial rRNA genes associated with soil suppressiveness against the plant-parasitic nematode Heterodera schachtii. *Applied and Environmental Microbiology* **69:** 1573-1580.

Yoon, S. S., & Sun, J. (2011) Probiotics, nuclear receptor signaling, and anti-inflammatory pathways. *Gastroenterology Research and Practice* **2011:** 971938.

Young, V. B., & Schmidt, T. M. (2004) Antibiotic-associated diarrhea accompanied by large-scale alterations in the composition of the fecal microbiota. *Journal of Clinical Microbiology* **42:** 1203-1206.

Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z., & Forney, L. J. (2012) Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PloS One* **7:** e33865.

Yue, J. C., & Clayton, M. K. (2005) A similarity measure based on species proportions. *Communications in Statistics-Theory and Methods* **34:** 2123-2131.

Zar, J. (1996) Biostatistical Analysis. 3rd Edition. Upper Saddle River: Prentice Hall, Inc.

Zhang, C., Zhang, M., Wang, S., Han, R., Cao, Y., Hua, W. et al. (2009) Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice. *The ISME Journal* **4:** 232-241.

Zhu, X. Y., Zhong, T., Pandya, Y., & Joerger, R. D. (2002) 16S rRNA-based analysis of microbiota from the cecum of broiler chickens. *Applied and Environmental Microbiology* **68:** 124-137.

Zhu, X., & Joerger, R. (2003) Composition of microbiota in content and mucus from cecae of broiler chickens as measured by fluorescent in situ hybridization with group-specific, 16S rRNA-targeted oligonucleotide probes. *Poultry Science* **82:** 1242-1249.

Zimmer, J., Lange, B., Frick, J. S., Sauer, H., Zimmermann, K., Schwiertz, A. et al. (2011) A vegan or vegetarian diet substantially alters the human colonic faecal microbiota. *European Journal of Clinical Nutrition* **66:** 53-60.

Zoetendal, E. G., Vaughan, E. E., & De Vos, W. M. (2006) A microbial world within us. *Molecular Microbiology* **59:** 1639-1650.

Zoetendal, E. G., Akkermans, A. D. L., & De Vos, W. M. (1998) Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. *Applied and Environmental Microbiology* **64:** 3854-3859.

Zoetendal, E. G., Cheng, B., Koike, S., & Mackie, R. I. (2004) Molecular microbial ecology of the gastrointestinal tract: from phylogeny to function. *Current Issues in Intestinal Microbiology* **5:** 31-48.

Zoetendal, E. G., Collier, C. T., Koike, S., Mackie, R. I., & Gaskins, H. R. (2004) Molecular ecological analysis of the gastrointestinal microbiota: a review. *Journal of Nutrition* **134:** 465-472.

Zoetendal, E. G., Akkermans, A. D. L., Akkermans-Van Vliet, W. M., de Visser, J. A. G. M., & de Vos, W. M. (2001) The host genotype affects the bacterial community in the human gastronintestinal tract. *Microbial Ecology in Health and Disease* **13:** 129-134.

Zoetendal, E. G., Von Wright, A., Vilpponen-Salmela, T., Ben-Amor, K., Akkermans, A. D. L., & De Vos, W. M. (2002) Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces. *Applied and Environmental Microbiology* **68:** 3401-3407.

Zoetendal, E. G., Ben-Amor, K., Harmsen, H. J. M., Schut, F., Akkermans, A. D. L., & De Vos, W. M. (2002) Quantification of uncultured Ruminococcus obeum-like bacteria in human fecal samples by fluorescent in situ hybridization and flow cytometry using 16S rRNA-targeted probes. *Applied and Environmental Microbiology* **68:** 4225-4232.

Zoetendal, E., Rajilić-Stojanović, M., & De Vos, W. (2008) High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut* **57:** 1605-1615.

Zuckerkandl, E., & Pauling, L. (1962) Molecular disease, evolution and genetic heterogeneity. In Horizons in Biochemistry: Albert Szent-Györgyi dedicatory volume. New York: Academic Press. pp. 189–225.

Zwielehner, J., Lassl, C., Hippe, B., Pointner, A., Switzeny, O. J., Remely, M. et al. (2011) Changes in human fecal microbiota due to chemotherapy analyzed by TaqMan-PCR, 454 sequencing and PCR-DGGE fingerprinting. *PLoS One* **6:** e28654.