Evaluating molecular methods for human microbiome analysis

by

Katherine Kennedy

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Biology

Waterloo, Ontario, Canada, 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis, except where noted. Original Illumina library construction and DGGE analysis of human gastrointestinal tract samples (Fig. 1) was performed by Dr. Jennifer Stearns at the University of Waterloo. Michael B. Goldberg and Howard C. Tenenbaum collected oral samples, and Kenneth Croitoru collected gastric and intestinal samples at Mount Sinai Hospital (Toronto, Canada). Michael W. Hall provided assistance with AXIOME.

This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In human microbiome analysis, sequencing of bacterial 16S rRNA genes has revealed a role for the gut microbiota in maintaining health and contributing to various pathologies. Novel community analysis techniques must be evaluated in terms of bias, sensitivity, and reproducibility and compared to existing techniques to be effectively implemented. Next-generation sequencing technologies offer many advantages over traditional fingerprinting methods, but this extensive evaluation required for the most efficacious use of data has not been performed previously. Illumina libraries were generated from the V3 region of the 16S rRNA gene of samples taken from 12 unique sites within the gastrointestinal tract for each of 4 individuals. Fingerprint data were generated from these samples and prominent bands were sequenced. Sequenced bands were matched with OTUs within their respective libraries. The results demonstrate that denaturing gradient gel electrophoresis (DGGE) represents relatively abundant bacterial taxa (>0.1%). The β-diversity of all samples was compared using Principal Coordinates Analysis (PCoA) of UniFrac distances and Multi-Response Permutation Procedure (MRPP) was applied to measure sample cluster strength and significance; indicator species analysis of fingerprint bands and Illumina OTUs were also compared. The results demonstrate overall similarities between community profiling methods but also indicate that sequence data were not subject to the same limitations observed with the DGGE method (i.e., only abundant taxa bands are resolved, unable to distinguish disparate samples). In addition, the effect of stochastic fluctuations in PCR efficiency (“PCR drift”) has not been rigorously tested and may differ for DGGE and next-generation sequencing. I compared pooled and individual reactions for samples of high and low template concentration for both Illumina and DGGE using the combined V3-V4 region of the 16S rRNA gene, and demonstrated that template concentration has a greater impact on reproducibility than pooling. This research shows congruity between two disparate molecular methods, identifies sources of bias, and establishes new guidelines for minimizing bias in microbial community analyses.

Acknowledgements

I would like to thank Dr. Josh Neufeld for the opportunity to conduct this research, as well as for his guidance and supervision. I would also like to thank Dr. Barbara J. Butler, Dr. Gabriel Moreno-Hagelsieb, and Dr. Valerie Taylor, for their advice and perspective as committee members. Additionally, I would like to thank Dr. Jennifer Stearns for conducting the biogeography DGGE data that were analyzed in this thesis as referenced in Figure 1.

I would also like to thank the members of the Neufeld lab for their help, and for making the Neufeld office and lab enjoyable places to be. I would especially like to thank Katja Engel for teaching me how to do practically everything during my first month in the lab, Michael D. J. Lynch for his knowledge of bioinformatics, and Michael W. Hall for his invaluable help with AXIOME and all things computers.

Table of Contents

Chapter 4. Conclusions and future considerations

List of Figures

List of Tables

List of Abbreviations and Symbols

%	Percent
°C	degrees Celsius
µM	micro molar
5-HT	5-hydroxytryptamine
A	Effect size
ANOVA	Analysis of Variance
AXIOME	Automation, eXtension, and Integration Of Microbial Ecology
BDNF	Brain Derived Neurotrophic Factor
BLAST	Basic Local Alignment Search Tool
bp	base pairs
CM^{2}BL	Canadian Meta Microbiome Library
CNS	Central nervous system
C-section	Caesarian section
DGGE	Denaturing Gradient Gel Electrophoresis
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside triphosphate
g	gram(s)
GABA	Gamma-aminobutyric acid
GALT	Gut Associated Lymphoid Tissue
HPA	Hypothalamic-Pituitary-Adrenal
IBD	Inflammatory Bowel Disease
Ig	Immunoglobulin
IL-10	Interleukin 10
IV	Indicator Value
kCal	Kilo calorie
MRPP	Multi-Response Permutation Procedure
NAFLD	Non-alcoholic fatty liver disease
ng	nanogram(s)
NMS	Non-metric Multidimensional Scaling
OTU	Operational Taxonomic Unit
PCoA	Principle Coordinate Analysis
PCR	polymerase chain reaction
QIIME	Quantitative Insights Into Microbial Ecology
rRNA	ribosomal ribonucleic acid
SCFA	Short Chain Fatty Acid
SOLiD	Supported Oligonucleotide Ligation and Detection
T	Test statistic
TLR	Toll-Like Receptor
T_{reg}	Regulatory T cell
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
V	Volts
µL	microliter(s)

Chapter 1. Introduction

1.1 The human microbiome

1.1.1 Introduction

The impact that microorganisms have on human health has been a subject of study since the germ theory of disease replaced earlier explanations of illness in the late 19th century (Pasteur 1881). Koch’s postulates, which describe the criteria necessary to link a given microorganism to the disease it causes, shaped future study of medical microbiology to the search for, and defense against, pathogenic microorganisms (Koch 1890). This focus on microorganisms as pathogens brought about a war on Bacteria (“germs”) that is ongoing. Only recently have researchers begun to investigate the abundance of microbial species that inhabit non-pathogenic niches within the environment provided by the human body.

There are over 100 trillion microorganisms inhabiting the human gut, collectively known as the human gut “microbiome” or “microbiota” (Frank and Pace 2008). Though these terms are sometimes used interchangeably, “microbiome” refers to the collective genomes of the microorganisms whereas “microbiota” refers to the organisms themselves. Though frequently referred to as commensals, these microorganisms are more accurately classified as mutualists because of the beneficial relationship that they have with their hosts. Gut microbes metabolize energy sources that are otherwise inaccessible to their hosts, such as cellulose and resistant starches (Sonnenburg et al. 2005; Turnbaugh et al. 2006). Primary fermenters in the large intestine also produce short chain fatty acids that contribute about 10% of daily caloric intake in a typical western diet (Gill et al. 2006). Together, the genomes of gut microorganisms encode genes for the synthesis of essential amino acids and vitamins, and the detoxification of potentially harmful xenobiotics (Gill et al. 2006). Beyond their metabolic contributions, the gut microbiota also influence development of the immune system: germ-free mice have reduced levels of helper T cells and cytotoxic T cells, as well as reduced levels of Immunoglobulin A and other immunological proteins (Round et al. 2010). Because of the magnitude of the impact microorganisms have on humans, many researchers have begun referring to a human “superorganism”, which includes the microbiota (Gill et al. 2006; Li et al. 2008; Sleator 2010).

1.1.2 Normal composition

The gut ecosystem is the densest known microbial habitat (Whitman et al. 1998). It is estimated that one gram of stool contains between 10^{11} and 10^{12} bacteria, making it 60% microbial by mass (O'Hara and Shanahan 2006). Throughout the gastrointestinal tract, both diversity and density of the microbiota increase along two axes: from proximal (mouth) to distal (colon) and from the tissue to the lumen (Sekirov et al. 2010).

Although fluctuations in microbial composition were previously thought to be infrequent and due to factors such as antibiotic use or diarrhoea (Simon and Gorbach 1984), more recent studies have shown this is not the case. The microbial composition of the gut varies over time, even from day to day, but differences between individuals and across body sites persist (Caporaso et al. 2011). The gut microbiota is mainly composed of Bacteria, but also includes Archaea and Eukaryotes, as well as viruses (Ley et al. 2006). Culture-based methods have shown that anaerobic bacteria dominate over aerobic bacteria by a factor of 100 to 1000 (Simon and Gorbach 1984).

The community composition of the gut is characterized by shallow diversity. Most members of the gut belong to the phyla *Firmicutes* and *Bacteroidetes*, with lower abundances of *Proteobacteria*, *Verrucomicrobia*, *Actinobacteria*, *Fusobacteria*, and *Cyanobacteria* (Hugenholtz et al. 1998; Human Microbiome Project Consortium 2012; Stearns et al. 2011). The low diversity of the gut microbiota at a phylum level is juxtaposed with high levels of species and subspecies diversity: there are approximately 1000 species of bacteria in the gut (Human Microbiome Project Consortium 2012). This pattern of diversity may reflect adaptive radiation: a few early colonists resulting in a variety of descendant organisms (Ley et al. 2006).

A variety of factors have been theorized to influence the shallow diversity of the gut microbiota. Compared to other microbial environments, such as soils and oceans, the mammalian gut has existed for a very short time: the microorganisms that inhabit the gut have not diverged sufficiently to form many separate phyla (Ley et al. 2006). The nature of the gut environment also results in a combination of competing selective pressures that differs from abiotic habitats: the host requires functional redundancy for stability whereas the microorganisms favour functional specialization to reduce interspecies competition for nutrients (Ley et al. 2006). The homogenizing force of mixing in the gut due to the muscular contractions of mechanical digestion also reduces niche breadth allowing for greater diversity at a shallow phylogenetic level (Ley et al. 2006).

1.1.3 Development of the microbiota

Though humans were previously thought to be sterile prior to colonization during birth, increasing evidence suggests that this is not the case. Microorganisms have been found in umbilical cord blood (Jiménez et al. 2005), amniotic fluid (DiGiulio et al. 2008), and meconium (Jiménez et al. 2008) of healthy neonates. The inoculum received during birthing also has a large impact on the developing microbiota: that of vaginally delivered infants resembled their mothers' vaginal microbiota (dominated by *Lactobacillus*, *Prevotella*, or *Sneathia* spp.) whereas that of C-section infants resembled that of their mothers' skin microbiota (dominated by *Staphylococcus*, *Corynebacterium*, and *Propionibacterium* spp.; Dominguez-Bello et al. 2010). Transfer of microbiota from mother to child continues after birth with lactic acid bacteria being transferred through breast milk (Martin et al. 2003; Martin et al. 2012). The microbiota of exclusively breastfed infants is characterized by a lower diversity than that of partially formula-fed infants due to lower diversity within the *Firmicutes*, and also by a higher relative abundance of *Bifidobacteria* (Azad et al. 2013).

As the gut microbiota of infants develops into its climax community, the influences of these early factors is obscured, but not eliminated. Early colonizers can shape their environment to be more favourable for their growth by regulating gene expression in epithelial cells and thereby influence the composition and structure of the developing microbiota (Hooper and Gordon 2001). The enduring influence of the initial colonizers is exemplified by the fact that the microbiota can reflect kinship relationships, with inherited genotypes acting on the initial inoculum that is received from the mother as well as subsequent colonizers (Ley et al. 2006).

1.1.4 Metabolic functions

The gut microbiota act as an anaerobic bioreactor in fermenting non-digestible carbohydrates such as cellulose, pectins, and resistant starches as well as unabsorbed sugars and alcohols, and endogenous mucus (Cummings et al. 1996; Roberfroid et al. 1995). This fermentation provides as much as 8% of the total energy harvested by the host (Institute of Medicine of the National Academies 2006) and conventional mice have been shown to have 40% more body fat than germ-free mice on the same diet (Bäckhed et al. 2004). The microbiota also differ between obese and lean individuals in ways that are hypothesized to affect the host's ability to extract energy from food (Hooper and Gordon 2001). These changes are apparent in the relative abundances of the two dominant bacterial phyla, *Bacteroidetes* and *Firmicutes*, resulting in an increased ability to extract energy in obese individuals (Turnbaugh et al. 2006). Jumpertz and coworkers showed that an increase of nutrient load resulted in a 20% increase in *Firmicutes* and a corresponding decrease in *Bacteroidetes*: this shift resulted in an increased capacity for energy harvest of about 150 kcal (Jumpertz et al. 2011).

1.1.5 Impact on the immune system

The purpose of the immune system is to recognize and respond to pathogenic microorganisms, but intestinal microorganisms are not subject to indiscriminate eradication. The symbiosis of humans and their microbiota requires our immune system to be tolerant towards hundreds of species of microorganisms while simultaneously remaining vigilant against potential pathogens and preventing these microbes from penetrating the intestinal epithelium (Sommer and Bäckhed

2013). The microbiota themselves provide some protection from pathogenic organisms by means of niche occupation. The normal microbiota provide stability and prevent colonization by pathogens (Guarner and Malagelada 2003) by competing for epithelial attachment sites (Bernet et al. 1994) and nutrients (Hooper et al. 1999).

The gut microbiota have a demonstrated effect on the immune system, including surface barriers: in the absence of microorganisms, intestinal epithelial cells have decreased cell-turnover rates and altered microvilli (Abrams et al. 1963). The epithelium of germ-free mice also has fewer goblet cells and a thinner mucus layer, which reverts to a normal phenotype upon exposure to the bacterial molecules lipopolysaccharide and peptidoglycan (Macpherson and Uhr 2004). In humans, the gut-associated lymphoid tissue (GALT) contains more immune cells than other body locations (Brandtzaeg et al. 1989) and this is likely due to the gut being so highly populated by microorganisms. The GALT of germ-free pigs has lower levels of immune cells (Butler et al. 2000) and this trend is remediated after colonization of the gut (Umesaki et al. 1993).

As well as influencing the development of the immune system within individuals, gut microbiota likely also influence the evolution of our immune system. Complex, co-evolved microbial communities are a common feature among vertebrates, while invertebrates seem to only harbour small numbers of bacterial species. It has been suggested that these microbial communities are responsible for the evolution of adaptive immunity, the branch of the immune system that learns from previous exposure, which is exclusive to vertebrates (McFall-Ngai 2007).

Tolerance towards our microbiota seems to be achieved by restricting their penetration outside of the lumen of the gut (Macpherson and Uhr 2004). The immunoglobulin IgA has an important role in this regulation (Suzuki et al. 2004): it coats the bacteria of the gut (Round and Mazmanian 2009) and acts as a chemical barrier to prevent them from crossing the gut epithelium (Sommer and Bäckhed 2013). Some microbial cells do penetrate the epithelium where they are taken up by dendritic cells which then migrate to mesenteric lymph nodes and induce naive B cells to produce IgA (Macpherson and Uhr 2004). This process results in an estimated 80% of antibody production in humans (Brandtzaeg 2009). The immune system remains systemically ignorant of the microbiota because the immune response is local, not reaching further than the mesenteric lymph nodes (Macpherson et al. 2005).

1.1.6 The gut microbiota and disease

An increasing number of diseases and disorders have been correlated with an altered gut microbiota. Crohn's disease and ulcerative colitis, the two most common forms of Inflammatory Bowel Disease (IBD), are both associated with increased numbers of bacteria that adhere to the intestinal mucosa such as *Enterobacteriaceae,* including *Escherichia coli,* and other *Proteobacteria* (Nagalingam and Lynch 2012). IBD is characterized by atypical levels of inflammation in the gastrointestinal tract and associated with higher antibody titres against the microbiota than found in healthy controls (Round and Mazmanian 2009). In most cases this systemic immune response is directed against organisms that have the capacity to be pathogenic (pathobionts) such as *Helicobacter, Clostridium,* and *Enterococcus* (Round and Mazmanian

2009). The gut microbiota has also been implicated in metabolic disorders including obesity, type 2 diabetes, and non-alcoholic fatty liver disease (NAFLD; Ley et al. 2005).

1.1.7 The gut-brain axis

The impacts of the human microbiome on the host greatly exceed contributions to nutrient catabolism and the ability to outcompete pathogens. Gut "dysbiosis" has been implicated in an ever-increasing number of outwardly unrelated medical conditions ranging from chronic fatigue syndrome (Lakhan and Kirchgessner 2010) to autism (Parracho et al. 2005). There is a high comorbidity of depressive disorders and functional gastrointestinal ailments such as IBD: up to 90% of subjects with IBD have a comorbid psychiatric condition (Whitehead et al. 2002). This correlation has motivated a variety of murine research studies. In 2004, Sudo and coworkers demonstrated that germ-free mice have a hyperactive hypothalamic-pituitary-adrenal (HPA) axis compared to conventional mice, meaning they have an exaggerated stress-response (Sudo et al. 2004). In addition, changes to the HPA axis due to early life stress result in an altered gut microbiome (O'Mahony et al. 2009), demonstrating the bi-directionality of this relationship. Germ-free mice also have decreased cortical levels of Brain Derived Neurotrophic Factor (BDNF), a molecule responsible for neuronal cell growth and plasticity, as well as the monoamine neurotransmitters norepinephrine and 5-hydroxytryptamine (5-HT, commonly known as serotonin) (Forsythe et al. 2010).

The effects of the gut microbiota are further supported by findings that some of these microorganisms are capable of producing neuroactive molecules such as serotonin, melatonin, acetylcholine, and gamma-aminobutyric acid (GABA), the primary inhibitory neurotransmitter of the central nervous system (CNS; Iyer et al. 2004). These neuroactive compounds may act directly on the human nervous system. Gut microbes also affect the central nervous system indirectly through Toll-like Receptors (TLRs) and immune cells, resulting in altered levels of circulating cytokines, through short chain fatty acids (SCFAs). The SCFA butyrate, a product of anaerobic bacterial metabolism, has been shown to have antidepressant effects when injected systemically (Schroeder et al. 2007). Butyrate is a histone deacetylase inhibitor that has the capacity to modulate epigenetic modifications via chromatin remodelling, which has been linked to mood disorders (Tsankova et al. 2006). This may be a result of free fatty acids interfering with the binding of albumin to tryptophan. Tryptophan is an amino acid precursor to serotonin (5-HT), the depletion of which is associated with anxiety and depression (Graeff et al. 1996). Tryptophan which is bound by albumin is unavailable for uptake by the CNS, so if fatty acids interfere with this binding they may increase the amount of tryptophan available for uptake by the CNS (Gentil et al. 1977; Maes 2011).

The inflammatory immune response observed in cases of depression is also associated with gut dysbiosis. Commensal gut microbes have been linked to the up-regulation of IL-10, an anti-inflammatory cytokine, and development of healthy T_{reg} populations, which are responsible for suppressing activation of the immune system and preventing autoimmune disorders (Forsythe et al. 2010; Macpherson and Uhr 2004; Ostman et al. 2006).

1.2 Microbiome community analysis

1.2.1 Introduction

More than 99% of microorganisms in many natural environments are not readily cultivable (Streit and Schmitz 2004). *E. coli* is thought of as a common gastrointestinal bacterium because it is easily cultured, however this *Gammaproteobacteria* member usually constitutes less than 1% of gut species (Hamady and Knight 2009). Culturing is an invaluable tool, but it falls far short in revealing the microbial diversity of the gut environment. For this reason molecular microbiological methods are necessary for the study of these environments.

When characterizing microbial communities, marker genes are targeted for amplification by PCR. The majority of microbial ecology studies have targeted the 16S rRNA gene because it is found in all prokaryotic microorganisms, is not prone to horizontal transfer, and it has both conserved regions and nine variable regions from which targets can be selected (Inglis et al. 2012; Neufeld and Mohn 2006). Drawbacks of the 16S rRNA gene include that different microbial taxa possess variable gene copy numbers, which hinders assessment of relative abundances (Neufeld and Mohn 2006).

1.2.2 Polymerase chain reaction

Polymerase chain reaction (PCR) is useful for effective analysis of the human microbiome and other microbial communities, but it suffers from inherent biases. These biases can be classified into two general categories: selection, which is the result of inherent differences in amplification efficiencies, and drift, which is the result of stochastic fluctuations and therefore non-reproducible (Wagner et al., 1994). One aspect of selection is the tendency towards a 1:1 ratio of all products due to more abundant templates being less available for amplification because of reannealing (Suzuki and Giovannoni 1996). Although bias caused by differences in primer binding energies is not easily surmounted, reducing the number of rounds of amplification was proposed in order to limit the tendency toward this homogeneous product ratio (Inglis et al. 2012; Polz and Cavanaugh 1998; Wagner et al. 1994). Bias caused by drift can be minimized by pooling replicate reactions because the way in which drift biases the products varies in each replicate as it is non-reproducible (Inglis et al. 2012; Polz and Cavanaugh 1998).

1.2.3 Denaturing gradient gel electrophoresis

Denaturing gradient gel electrophoresis (DGGE) is a fingerprinting technique which exploits two properties of DNA: partially denatured DNA cannot migrate through a polyacrylamide matrix, whereas double-stranded helices can (Lerman et al. 1984), and the melting temperature of a fragment of DNA depends on its sequence (Green et al. 2010; Muyzer and Smalla 1998) . As a result, DNA fragments of the same length can be electrophoretically separated on a gel by sequence, resulting in a visual representation of a sample's community structure. DGGE uses a modified PCR to amplify fragments of the same length consisting of organismal DNA with the addition of a GC-rich region. This GC clamp prevents complete denaturation, which would result in single stranded DNA that would otherwise move rapidly through the gel (Muyzer et al. 1993). DGGE allows multiple samples to be analyzed rapidly and economically (Green et al. 2010), which is imperative in microbial ecology as microbial ecosystems should be studied over time to observe interactions between microorganisms and their environment (Muyzer and Smalla 1998). Compared to sequencing, DGGE is orders of magnitude less expensive and time-consuming (Inglis et al. 2012; Muyzer and Smalla 1998), although this is becoming decreasingly true with advances in sequencing technologies. In addition, DGGE allows for bands that change in intensity with changing conditions to be picked and sequenced (Inglis et al. 2012; Muyzer et al. 1993) and can be very effective at identifying abundant indicator organisms because of this characteristic (Muyzer and Smalla 1998).

Despite the strengths of DGGE, this method has limited resolving power and cannot detect organisms that represent much less than 1% of the sample community, even when using SYBR Green I stain to decrease background staining (Muyzer and Smalla 1998). It also works best with low diversity samples because high diversity samples generate many bands, causing individual bands to not be easily discernible (Green et al. 2010; Inglis et al. 2012). DNA fragments stop in the gel after the portion with the lowest melting temperature denatures, so fragments that vary from each other outside of this area may still result in a single band (Fischer and Lerman 1980; Lerman et al. 1984). Unrelated sequences can also result in bands in the same location on a gel (Muyzer et al. 1993) and one species can produce multiple bands due to multiple 16S rRNA operons (Nübel et al. 1996). DGGE is very flexible in terms of what target region is used (Green et al. 2010), but it is optimally applied to separating fragments of less than 500 base pairs (Myers et al. 1985).

Because of its inability to resolve and detect low abundance organisms, DGGE is not able to completely visualize microbial communities or evaluate the diversity of high diversity samples.

However, it excels in initial screenings of many samples (Green et al. 2010) because it provides a rapid qualitative and semi-quantitative visual representation of community structure (Muyzer et al. 1993). Additionally, it is not capable of providing accurate quantitative diversity analysis due to the indirect relationship of species to bands within a fingerprint, resulting in biases in species richness and relative species abundance (Neilson et al. 2013). The strengths and weaknesses of this fingerprinting technique are well-complemented by sequencing-based characterization (Inglis et al. 2012).

1.2.4 Sequencing

DNA sequencing-based studies of the human microbiome can either be amplicon-based, with marker genes being sequenced to identify community members, or metagenomic, with randomly sheared fragments of DNA being sequenced. Metagenomic sequencing has the advantage of providing more functional information about a community but its random nature results in a sacrifice to taxonomic resolution, so amplicon sequencing is typically included in studies that aim to compare the identities of microorganisms within multiple samples (Kuczynski et al. 2012).

The dideoxy method of sequencing, now known as Sanger sequencing, was introduced in 1977 and remained the method by which almost all sequencing was conducted for approximately 30 years (Shendure and Ji 2008). Sanger sequencing uses chain-terminating nucleotide analogs to generate nucleotide-specific terminated fragments that are electrophoretically separated by size allowing the sequence to be read from terminal nucleotide of each fragment from shortest to longest (Sanger et al. 1977). After decades of gradual improvements, Sanger sequencing is now capable of read lengths up to approximately 1000 bp with 99.999% accuracy (Shendure and Ji 2008). Still, “next-generation” sequencing (NGS) has become the preferred method due to being “massively parallel”, thereby allowing for vastly more sequence reads to be obtained per experiment. This increase comes with sacrifices to read length and accuracy (Hutchison 2007), though these areas are constantly improving as they are solely a function of signal-to-noise ratio and not due in part to gel-related factors as is the case with Sanger sequencing (Mardis 2013). Because of this high degree of parallelism, avoidance of *E. coli* transformation and colony picking, and decreased reagent volumes required due to the immobilization of the molecules on a surface (Mardis 2013), next generation sequencing is significantly less costly and labour intensive than Sanger-based sequencing (Shendure and Ji 2008).

The predominant NGS platforms all utilize sequencing-by-synthesis (Inglis et al. 2012), which involves a step-wise reaction: nucleotide addition, nucleotide detection, and washing to remove fluorescent labels (Mardis 2013). Common to all sequencing-by-synthesis technologies (Table 1) is that PCR amplicons from a single molecule become spatially clustered, allowing for a signal bright enough to be detected upon nucleotide addition (Shendure and Ji 2008).

The first NGS method to become commercially available was 454 pyrosequencing, wherein each of the four nucleotides is added sequentially, and nucleotide additions release a pyrophosphate, which is detected by luciferase activity. The main source of error associated with pyrosequencing is indels resulting from misread homopolymer runs (Hutchison 2007; Shendure and Ji 2008). Illumina uses reversible chain-terminating nucleotides that are unblocked when the nucleotide-specific fluorescent tag is removed by washing. This suffers from increased base substitution errors due to the modified bases (Hutchison 2007). SOLiD (Supported Oligonucleotide Ligation and Detection) differs from 454 and Illumina in that it uses sequencing by ligation instead of by synthesis and as a result every base is matched to a probe twice, resulting in greater accuracy but shorter read lengths and longer run times (Hutchison 2007).

Table 1. Comparison of commonly used next-generation sequencing technologies as of 12/2013 from manufacturers' websites.

	Run time	Read length (average)	# of single reads (amplicon)	cost per run
Illumina MiSeq	~65 hours	2 x 300 bp	25 million	~$1000
454 GS Jr.	~10 hours	~400 bp	70000	~$1000
SOLiD 5500 Wildfire	10 days	2 x 50 bp	1.2 billion	~$2500

http://res.illumina.com/documents/products/datasheets/datasheet_miseq.pdf, http://www.gsjunior.com/instrument-workflow.php,http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_088661.pdf

The output of next-generation targeted amplicon sequencing is a complex dataset that requires extensive downstream analysis to produce interpretable results (Inglis et al. 2012). Bioinformatics tools are available to aid in this analysis. These tools include the software package QIIME (Quantitative Insights Into Microbial Ecology), a tool that takes raw sequence data and performs Operational Taxonomic Unit (OTU) picking, taxonomic assignment, and downstream statistical analysis (Caporaso et al. 2010). More recently, AXIOME (Automation, eXtension, and Integration Of Microbial Ecology) was produced to streamline QIIME (as well as mothur, another microbial analysis package; Lynch et al. 2013).

Next generation sequencing is susceptible to a variety of biases. Multiplex sequencing allows for parallel sequencing of multiple samples using unique, sample-specific “barcodes” that are added to the fragment to be sequenced. These barcodes can be added by ligation, but modified “barcoded primers” are more commonly employed as they introduce less bias (Alon et al. 2011). Berry and colleagues used 11 different barcoded primers in triplicate on the same sample from the mouse gut lumen and demonstrated that these primers still result in less reproducible data sets than primers that are not barcoded, and the bias they introduce cannot be predicted from secondary structure of the primer (Berry et al. 2011).

Sequencing results are biased by the method employed for DNA extraction (Martin-Laurent et al. 2001) and sample storage conditions (Cardona et al. 2012). Frozen samples maintain the highest alpha diversity and differ least in beta diversity (Rubin et al. 2013). Low concentration samples are particularly susceptible to bias due to the increased impact of stochastic processes during PCR (Chandler et al. 1997).

1.3 Hypotheses and objectives

1.3.1 Sensitivity and clustering behaviour of fingerprinting and sequencing

DGGE is known to be capable of distinguishing species that represents as little as ~1% of the sample being analyzed (Green et al. 2010; Muyzer et al. 1993). Samples included in clustering analysis must also be from similar sources because of the comigration of bands of different sequences, and the possibility of a single species forming multiple bands due to multiple variable 16S rRNA operons. Although DGGE has been previously used in the analysis of the human microbiota (Kinross et al. 2008; Scanlan et al. 2006), its sensitivity and clustering behaviour have not been directly compared to those of NGS. One of the objectives of this thesis is to compare the sensitivity and clustering behaviour of DGGE to that of Illumina sequencing in the analysis of the human microbiome. The sensitivity of DGGE is hypothesized to be limited to abundant organisms and this will be tested by comparing the relative abundance of picked DGGE bands to their respective OTU within samples. DGGE bands are also hypothesized to not cluster well by sample type across all samples because of the lack of a direct relationship between bands and their corresponding species within a sample.

1.3.2 Reproducibility of fingerprinting and sequencing

Both DGGE and Illumina sequencing are subject to the biases inherent in PCR amplification, one of which is a result of stochastic fluctuations and therefore non-reproducible. The pooling of replicate PCR amplifications has been suggested in order to minimize the effect of this bias and thereby increase the reproducibility of the technique being used subsequently to PCR (Inglis et al. 2012; Polz and Cavanaugh 1998). The impact that pooling has on increasing the reproducibility of fingerprinting and next generation sequencing methods has not been demonstrated. The reproducibility of DGGE is hypothesized to not be impacted by pooling because the bias is expected to not have a large enough effect to be apparent in abundant species. The reproducibility of Illumina sequencing is similarly hypothesized to be not substantially effected by pooling, though the bias will be more apparent than when using DGGE. Template concentration is hypothesized to be positively correlated with reproducibility.

Chapter 2. Methods

2.1 Assessment of sensitivity and clustering behaviour of fingerprinting and sequencing

2.1.1 Sampling

As described previously (Stearns et al. 2011), all samples used were collected from four healthy adults: two men and two women. Stool samples were frozen at -20°C immediately after collection. Within 24 hours of stool sample collection, each subject was prepared for colonoscopy by Klean Prep. Oral biofilm was collected from supragingival plaque, subgingival plaque, and the tongue prior to gastroscopy and colonoscopy. Biopsies of the transverse colon, sigmoid colon, rectum, gastric antrum, gastric body, and duodenum were collected during gastroscopy and colonoscopy. All samples were stored at -80°C prior to DNA extraction. Samples collected were used in a previous study on the bacterial biogeography of the human gastrointestinal tract (Stearns et al. 2011).

2.1.2 DNA extraction

DNA was extracted from 500 µL of sample storage buffer (or 0.25 g of stool) of each sample using the PowerSoil DNA Isolation Kit (MoBio) according to the manufacturer’s instructions with minor modifications: the addition of a 40 second bead-beating step and heating to 70°C for 10 minutes prior to the contaminant binding step. Purified DNA was run on a 1% agarose gel for densitometric quantification, and spectrophotometrically quantified using a Nanodrop 1000 (Thermo Scientific, USA).

2.1.3 PCR

A nested PCR protocol was used to amplify the V3 region of the 16S rRNA gene. The first round used primers 27F and 1492R (Stackebrandt and Goodfellow 1991) and 25 cycles were used: 5 minutes at 98°C, 25 times 1 minute at 98°C followed by 1 minute at 55°C then 2 minutes at 72°C, and finally 7 minutes at 72°C. The second round used modified 341F and 518R primers (Muyzer et al. 1993) containing a six-base barcode, the Illumina adapter sequence, and regions for binding of the sequencing primers (Bartram et al. 2011). The number of cycles for the second round was 20: 5 minutes at 98°C, 20 times 30 seconds at 98°C followed by 1 minute at 55°C then 1 minute at 72°C, and finally 7 minutes at 72°C. Each reaction mixture was prepared in a 25-µL volume consisting of the following components: 5 µL Phusion buffer, 0.05 µL dNTPs, 0.05 µL of each primer, 0.25 µL Phusion *Taq* polymerase, 1.5 µL BSA (10 mg/mL), 1 µL template, and 18.6 µL PCR water. A separate PCR was performed for use in DGGE except that the forward primer was 341F-GC.

2.1.4 Illumina library construction

Triplicate PCR amplifications were pooled for each sample and gel purification was performed to remove primers and primer dimers by separating them on a 2% agarose gel and using a QIAquick Gel Extraction Kit (Qiagen, Mississauga, Ontario, Canada). The products for each sample were mixed in equal nanogram amounts, and quantified using a NanoDrop ND2000 spectrophotometer (Thermo Scientific, Wilmington, DE) before being sequenced in two lanes of a Genome Analyzer IIx at the Plant Biotechnology Institute (National Research Council Canada;

Saskatchewan, Canada) using paired-end multiplex sequencing as previously described (Bartram et al. 2011).

2.1.5 Denaturing gradient gel electrophoresis

DGGE used a 30% to 70% denaturing gradient in 10% acrylamide gels that were run for 15 hours at 85V with 10 ng of PCR product loaded in each lane. Gels were stained with SYBR green and imaged using a PharosFX Plus Molecular Imager (Bio Rad, USA). A total of 84 bands were picked (Fig. 1) and sequenced using single-pass Sanger sequencing (Beckman Coulter Genomics, USA).

2.1.6 Indicator species analysis

DGGE images were analysed in Gelcompar II (Applied Maths, USA) where bands were bands and band classes were automatically assigned and manually curated. Band class data was imported into PC-ORD (McCune et al. 2002) where Dufrene-Legendre indicator species analysis was conducted (Dufrene and Legendre 1997). The same analysis was conducted on sequencing data in QIIME through AXIOME (Caporaso et al. 2010; Lynch et al. 2013).

2.2 Assessment of reproducibility of fingerprints and sequencing

2.2.1 Sample selection

Two of the stool samples used in the previous analysis were selected to assess reproducibility of sequencing and fingerprinting. Two soil samples were also selected from the Canadian

MetaMicrobiome Library (CM^{2}BL; Neufeld et al. 2011): one from a temperate deciduous forest (6TD) and one from an agricultural soybean field (10AS). These soil samples were selected because of their varying land usage origins and soil chemistry profiles, most notably pH (6.4 for 6TD and 7.6 for 10AS). The soil samples were included to evaluate the effect of varying levels of diversity within a sample on reproducibility because soils are much more diverse than stool.

2.2.2 DNA extraction

Five extractions for each sample were conducted and then pooled. Each extraction was performed on 0.25 g of sample using the PowerSoil DNA Isolation Kit (MoBio) according to the manufacturer's instructions with minor modifications: the addition of a 40 second bead-beating step at 5 m/s using a FastPrep-24 (MP Bio, USA) and heating to 70°C for 10 minutes prior to the contaminant binding step. An ethanol precipitation concentrated and purified the DNA. Purified DNA was run on a 1% agarose gel for densitometric quantification, and spectrophotometrically quantified using a Nanodrop 1000 (Thermo Scientific, USA) and using Qubit fluorometric quantification (Life Technologies, USA).

2.2.3 Sample preparation

Each sample was diluted to create a high and low concentration template. Samples S3, 6TD, and 10AS were diluted to 10 ng/µL and 0.1 ng/uL and sample S1 was diluted to 5 ng/µL and 0.1 ng/µL. For each of the 8 samples (S1 high, S1 low, S3 high, S3 low, 6TD high, 6TD low, 10AS high, 10AS low) 20 reactions were performed (Table 2). Of these, 5 were not-pooled and the remaining were pooled in triplicate, resulting in 5 pooled products and 5 not pooled products for each sample. SX replicates (Table 2) are from sample S3 but are separated in order to assess cross-lane variability during sequencing.

Table 2. Summary of sample replicates and conditions. Numbers correspond to Illumina barcode tag used for sequencing, but were also used for DGGE with the exclusion of SX replicates.

	S1	6TD	10AS	S3	SX
Pooled, high template concentration	1, 2, 3, 4, 5	11, 12, 13, 14, 15	21, 22, 23, 24, 25	31, 32, 33, 34, 35	81, 82, 83
Pooled, low template concentration	6, 7, 8, 9, 10	16, 17, 18, 19, 20	26, 27, 28, 29, 30	36, 37, 38, 39, 40	84, 85, 86
Not pooled, high template concentration	41, 42, 43, 44, 45	51, 52, 53, 54, 55	61, 62, 63, 64, 65	71, 72, 73, 74, 75	87, 88, 89
Not pooled, low template concentration	46, 47, 48, 49, 50	56, 57, 58, 59, 60	66, 67, 68, 69, 70	76, 77, 78, 79, 80	90, 91, 92

[a.] No-template controls 93, 94, 95, 96

2.2.4 PCR

Reactions were performed on the same thermocycler, a CFX96 Touch Real Time PCR Detection System (Bio Rad, USA), in randomized 96-well plates to limit bias due to possible thermal profile variations across wells. The entire procedure was performed twice on two separate days to assess possible variability introduced by sample manipulation. SX samples were excluded on one of the days in order to run identical products from these samples on both sequencer lanes.

The combined V3-V4 regions were amplified using modified 341F and 816R primers containing a six-base barcode, the Illumina adapter sequence, and regions for binding of the sequencing primers (Bartram et al. 2011). The number of cycles was 30: 30 seconds at 95°C, 30 times 15 seconds at 95°C followed by 30 seconds at 50°C then 30 seconds at 68°C, and finally 5 minutes at 68°C. Each reaction was done in a 25-µL volume consisting of the following components: 2.5 µL Thermal Polymerase buffer (10x; NEB), 0.05 µL dNTPs (100 µM), 0.05 µL 341F primer (100 µM), 0.5 µL 816R primer (10 µM), 0.125 µL *Taq* polymerase, 1.5 µL BSA (10 mg/mL), 1 µL template, and 19.3 µL PCR water. A separate PCR was performed for use in DGGE with slight modification: the forward primer 341F-GC, the reverse primer was 518R, both were 100 µM, and both were added in 0.05 µL volume to each reaction.

2.2.5 Illumina library construction

Products for each sample were gel purified to remove primers and primer dimers by separating them on a 1% agarose gel and using a QIAquick Gel Extraction Kit (Qiagen, Mississauga, Ontario, Canada). The products for each sample were mixed in equal nanogram amounts, and quantified using a NanoDrop ND2000 spectrophotometer (Thermo Scientific, Wilmington, DE) before being sequenced in two lanes of an Illumina MiSeq at Argonne National Labs (Lemont, IL) using paired-end sequencing as previously described (Bartram et al. 2011).

2.2.6 DGGE

Products were separated into those with high template concentration and those with low template concentration. Within these groups, products from each sample were distributed across multiple DGGE gels to reduce bias due to gel effects. For low template concentration gels approximately 15 ng of PCR product was loaded into each well. For high template concentration gels approximately 30-60 ng of PCR product was loaded into each well. No gels had both high and low template concentration samples and therefore this difference in amount of PCR product loaded was compensated for by the exposure time during image capture. DGGE used a 30% to 70% denaturing gradient in 10% acrylamide gels that were run for 15 hours at 85V. Gels were stained with SYBR green and imaged using a PharosFX Plus Molecular Imager (Bio Rad, USA).

2.2.7 Data analysis

DGGE fingerprints were aligned in Gelcompar II (Applied Maths, USA) with bands and band classes assigned automatically and manually curated. A UGPMA dendrogram of the fingerprints was created in Gelcompar II using Pearson correlations. DGGE data were exported and analysed in PC-ORD (McCune et al. 2002) wherein indicator species analysis was performed (Dufrene and Legendre 1997), in addition to PCoA (Gower 2005) using a Bray-Curtis distance metric, NMS (Kruskal 1964) using a Pearson correlation, and MRPP.

Sequencing data were clustered using CD-HIT (version 4.5.4; Li and Godzik 2006) and trees were built using FastTree2 (version 2.1.3; Price 2010) through AXIOME. Indicator species analysis, NMS, MRPP, and PERMDISP were carried out through AXIOME with NMS being performed for all samples together, and both NSM and PERMDISP being performed for each sample group individually (S1, S3, 6TD, 10AS, SX). All samples were rarefied to the fewest number of sequences in a sample within each analysis.

Chapter 3. Results and discussion

3.1 Assessment of sensitivity of DGGE fingerprinting and Illumina sequencing

3.1.1 Relative abundance of bands within community

DGGE can provide an immediate visual representation of complex microbial communities, but it suffers from a lack of sensitivity. Bands must be visible to be included in analyses so low abundance organisms are not detected if the intensity of their band's fluorescence is less than that of the background. This study used human microbiome samples to compare the sensitivity of DGGE to that of Illumina sequencing by matching bands that were picked and sequenced to OTUs within their respective Illumina library and calculating the proportion of that sample that the picked band represented.

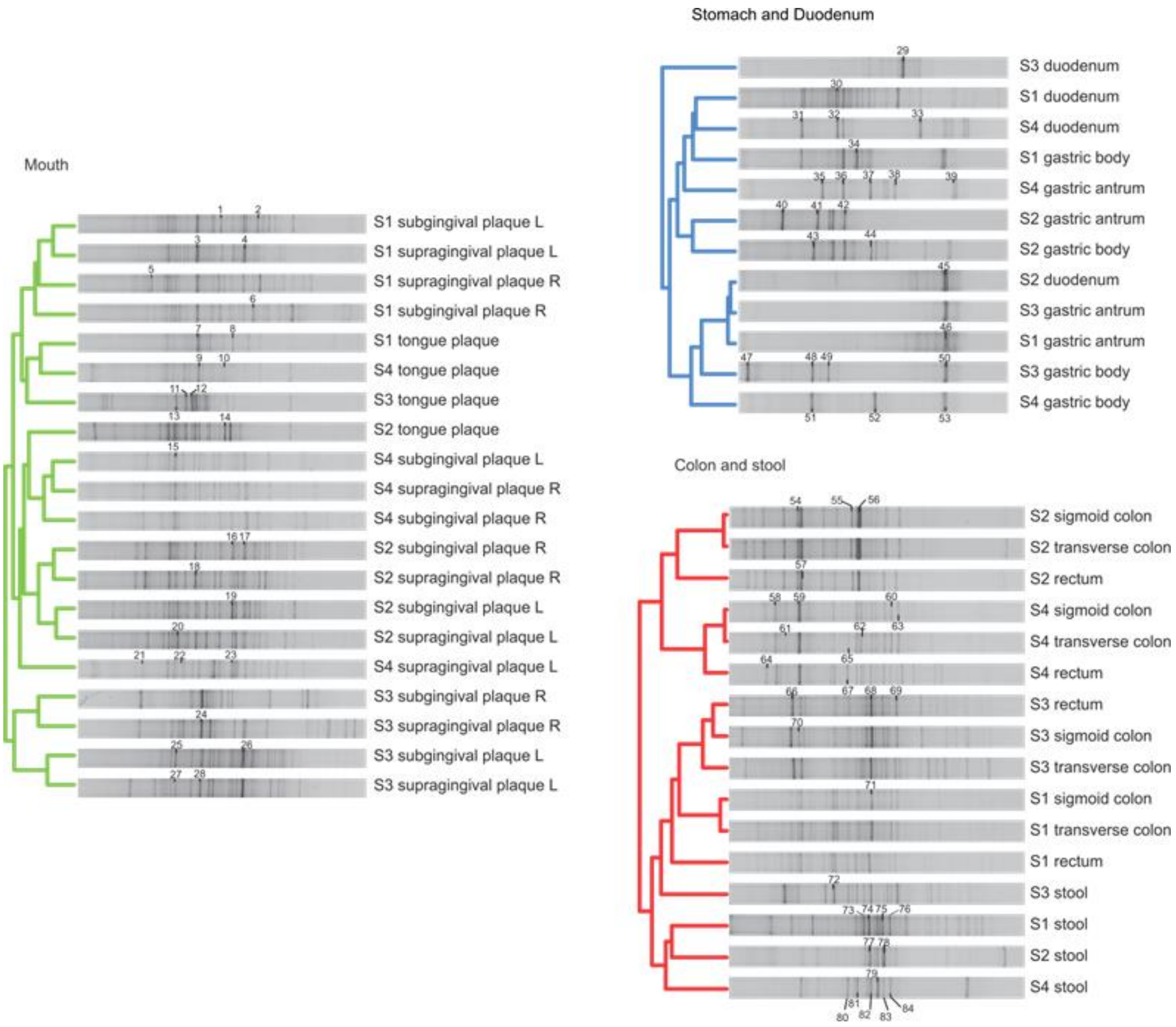

Figure 1. UPGMA dendrograms of DGGE fingerprints of human microbiome samples with picked bands numbered. Samples were taken from subjects 1-4 (S1-S4; Stearns et al. 2011).

DGGE fingerprints of samples from 4 subjects showed the varying levels of microbial diversity found across 12 sites in the gastrointestinal tract. Stomach and duodenal communities had very few bands (Fig. 1), which corroborates previous accounts of low diversity in these environments (Bik et al. 2006; Zoetendal et al. 2008). All bands picked were relatively prominent within their sample.

Bands picked (bands 1-84 in Fig. 1) from each sample were sequenced and matched at 100% sequence identity with an OTU in their respective Illumina library. The proportion of sequences in its sample that the matched OTU represents was plotted on a log scale (Fig. 2). Picked sequences that did not match with any OTU in the sample are represented by vertical grey bars. As well as not matching an OTU at 100%, these unmatched bands did not have any close matches within their respective samples. Out of bands that were matched with an OTU, 48 of 56 represented an OTU with 1% of more of the sequences in their sample. Almost all of the bands that did not match with an OTU in their Illumina library were gastric samples. As the same extracted DNA was used for both the DGGE and the Illumina PCRs, this cannot be due to problems with DNA extraction specific to gastric samples. It could be a result of differences in amplification efficiencies of the primers used for PCR because distinct primer modifications (GC-clamp or barcode) are used for DGGE and Illumina. This bias may be most apparent in the gastric samples due to their low diversity, or due to inhibitors present in extracted gastric DNA that preferentially affect either *Taq* (used for DGGE) or Phusion (used for Illumina libraries for high fidelity) polymerase.

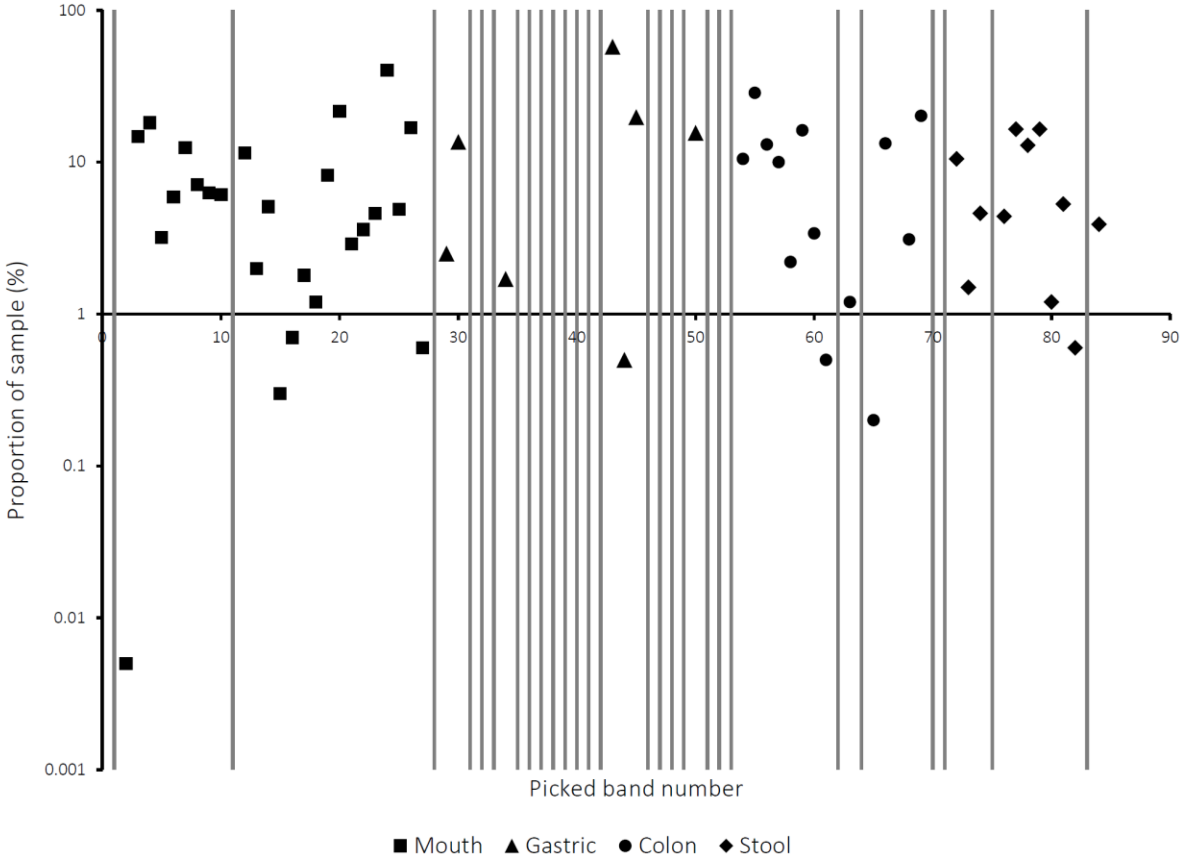

Figure 2. Proportion of sample represented by picked bands (1-84). Picked band sequences that did not match to an OTU in their respective Illumina library are represented by vertical grey bars.

3.1.2 Indicator species analysis

Indicator species analysis is commonly used with human microbiome studies to identify microorganisms that are associated with pathologies (Mager et al. 2005; Russell et al. 2012). These pathologies are rarely the result of a single microorganism, so to fully describe them the input data for indicator species analysis must come from techniques with higher sensitivity. Indicator Values (IV) are assigned according to the fidelity (proportion of the species that are within a given group) and specificity (proportion of samples within the group that contain the species; (Dufrene and Legendre 1997). An indicator value of 1 would be assigned to a perfect indicator species that is always present in the group in question and never present in other groups. Groups are assigned *a priori*. In this case multiple sets of groups were examined in order to account for varying niche breadth (De Cáceres et al. 2010): general site (mouth, stomach, bowel, stool), specific site (each of the 13 gastrointestinal sites), and subject (S1, S2, S3, S4).

Indicator species analysis was run on DGGE profiles in PC-ORD and Illumina data in AXIOME. The number of indicator species with an indicator value greater than or equal to 0.5 with a *p* value equal to or less than 0.01 was tallied for both methods (Table 3). The number of indicator bands from DGGE for subjects is also higher than the number for specific site. This is not an expected result; microbial community composition should vary more by site than by subject. This is also contradicted by the Illumina data.

Table 3. Number of indicator species (IV ≧ 0.5 and p ≦ 0.01) found using DGGE (wherein bands stand in for species) and Illumina (wherein OTUs stand in for species)

	General Site	Specific Site	Subject
DGGE	5	2	3
Illumina	2225	1707	53

General site: mouth, gastric, colon, or stool; Specific site: subgingival plaque left, subgingival plaque right, supragingival plaque left, supragingival plaque right, tongue, gastric antrum, gastric body, duodenum, transverse colon, sigmoid colon, rectum, and stool; Subject: subject 1, subject 2, subject 3, subject 4.

Using BLAST, the sequences of the indicator bands from DGGE were compared against a database of indicator species generated in AXIOME from Illumina data. All indicator bands from DGGE matched with an Illumina indicator OTU except for one of the two “specific site” indicator bands, which had not been picked and sequenced.

3.2 Clustering behaviour of data from DGGE fingerprinting and Illumina sequencing

Principal coordinates analysis (PCoA) is a form of multidimensional scaling that seeks to position objects in fewer dimensions while maintaining their distances as accurately as possible (Gower 2005). Non-metric multidimensional scaling (NMS) instead prioritizes representation in two dimensions with the goal of plotting similar objects together and dissimilar objects apart; the variation explained by the two axes is not necessarily maximized (Kruskal 1964). Samples were plotting using PCoA with a Bray-Curtis dissimilarity metric (Bray and Curtis 1957) for both Illumina data (Fig. 3A) and DGGE data (Fig. 3C). Illumina samples clustered by general sampling location, with colon and stool samples separating from gastric and mouth samples more than from each other. If UniFrac, which is based on phylogenetic distances (Lozupone and Knight 2005), is used as a distance metric (Fig. 3B), these patterns are still apparent. In contrast, DGGE samples do not clearly separate by general sampling location when using either PCoA with a Bray-Curtis dissimilarity metric (Fig. 3C), or NMS with Pearson correlation (Fig. 3D).

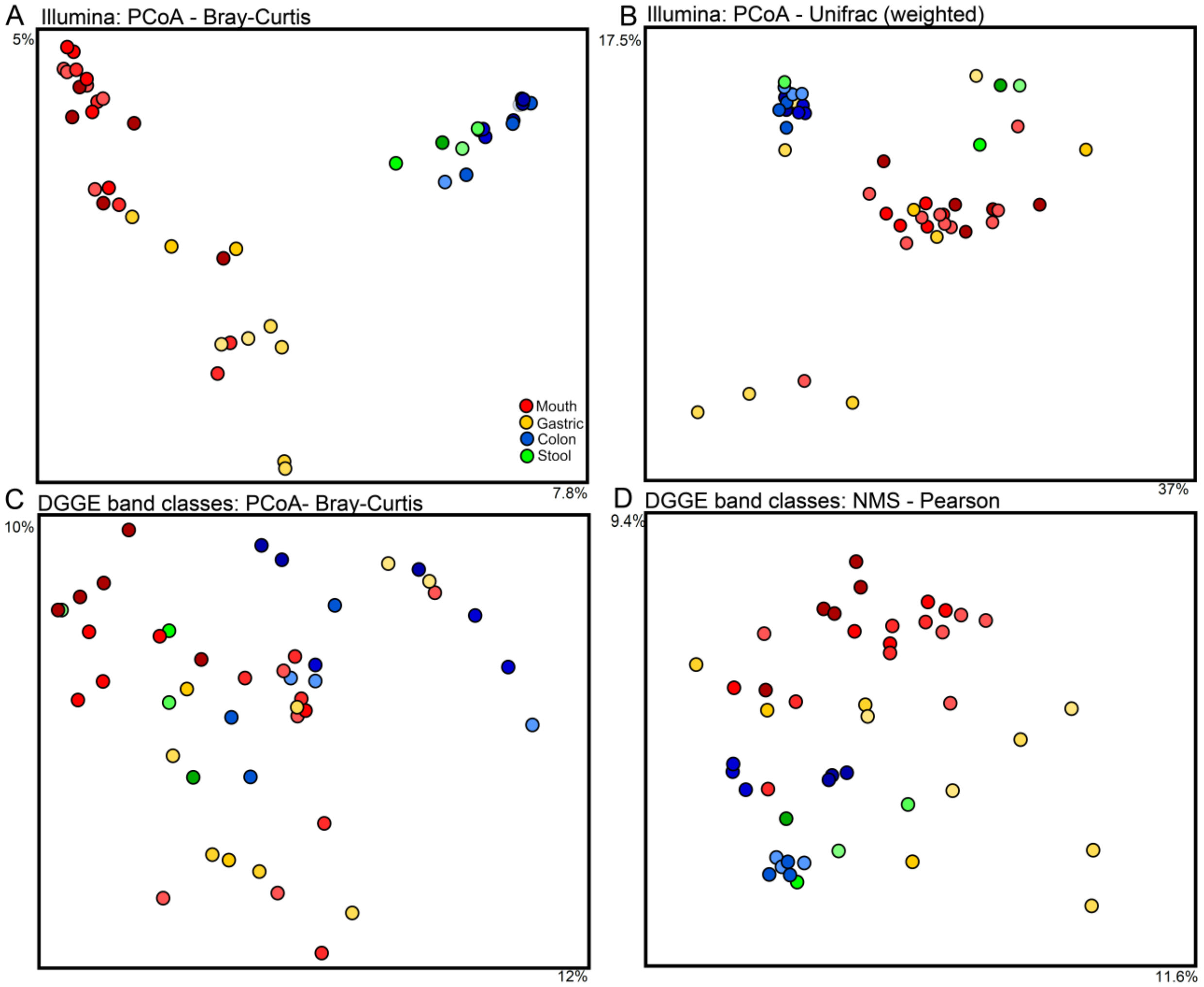

Figure 3. Ordination of DGGE and Illumina data using both PCoA and NMS. Shades of each colour indicates from which subject samples originate (S1-S4 corresponds to dark to light).

One of the weaknesses of DGGE is that one band can correspond to multiple species and one species doesn't necessarily result in one band (Green et al. 2010). Fragments with different sequences can co-migrate to the same position on a gel, and because some bacteria possess multiple 16S rRNA genes two or more bands can result from the same species. As a result of this DGGE fingerprints cannot be compared across different communities: comparisons must be "apples to apples" with samples being measured either over time, or with changing treatments, rather than between unrelated samples.

The poor performance of DGGE for clustering samples based on distinct body sites is evident in both PCoA and NMS plots when compared to ordinations obtained using sequencing data (Fig. 3). This result is confirmed by Multi-Response Permutation Procedure (MRPP), which tests the degree of within group homogeneity (effect size, *A*) and the amount separation between groups (test statistic, *T*, where more negative indicates greater separation; Mielke Jr et al. 1976). MRPP was performed on both Illumina data and DGGE data using *a priori* groupings based on the general location from which a sample was obtained (mouth, gastric, colon, and stool). The effect size of *a priori* groups for Illumina data (*A*=0.14) is much greater than that for DGGE data (*A*=0.02) meaning that groups using Illumina data have much greater within-group homogeneity that those using DGGE data. The separation between groups using Illumina data (*T*=-23.45) is similarly much greater than groups using DGGE data (*T*=-5.73). Groups using data obtained from Illumina have both greater homogeneity within each group and greater separation between groups.

3.3 Assessment of reproducibility of DGGE fingerprint and Illumina sequencing data

Polymerase chain reaction (PCR) is susceptible to two types of bias: selection and drift. Selection results from variable amplification efficiencies that are inherent in a given sample and drift is a result of stochastic fluctuation and therefore non-reproducible. Polz and Cavanaugh performed PCR on mixtures of known concentrations of genomic DNA from different species to determine the relative impact of bias due to selection and drift. Though bias due to drift is less than that due to selection when template concentrations are high, PCR amplifications are usually pooled in triplicate to minimize the effect of drift (Polz & Cavanaugh, 1998).

3.3.1 Impact of pooling triplicate reactions on DGGE

3.3.1.1 Tree construction

The DGGE patterns of all samples (S1 stool, S3 stool, 6TD soil, 10AS soil) were aligned in Gelcompar II (Applied Maths, USA). Both bands and band classes were automatically assigned and manually curated. A UPGMA dendrogram was generated using Pearson correlations (Fig. 4). Both stool samples grouped strongly by sample whereas the two soil samples separated from the stool samples, but not from each other. The high diversity of the soil samples caused Gelcompar to have difficulty discerning discrete fingerprint bands and resulted in this lack of separation. The pooling of triplicate amplifications did not result in any improvement in tree formation over non-pooled amplifications: pooled samples do not separate from non-pooled (Fig. 4). Soil replicates that do group by sample appear to generally have had high template concentration.

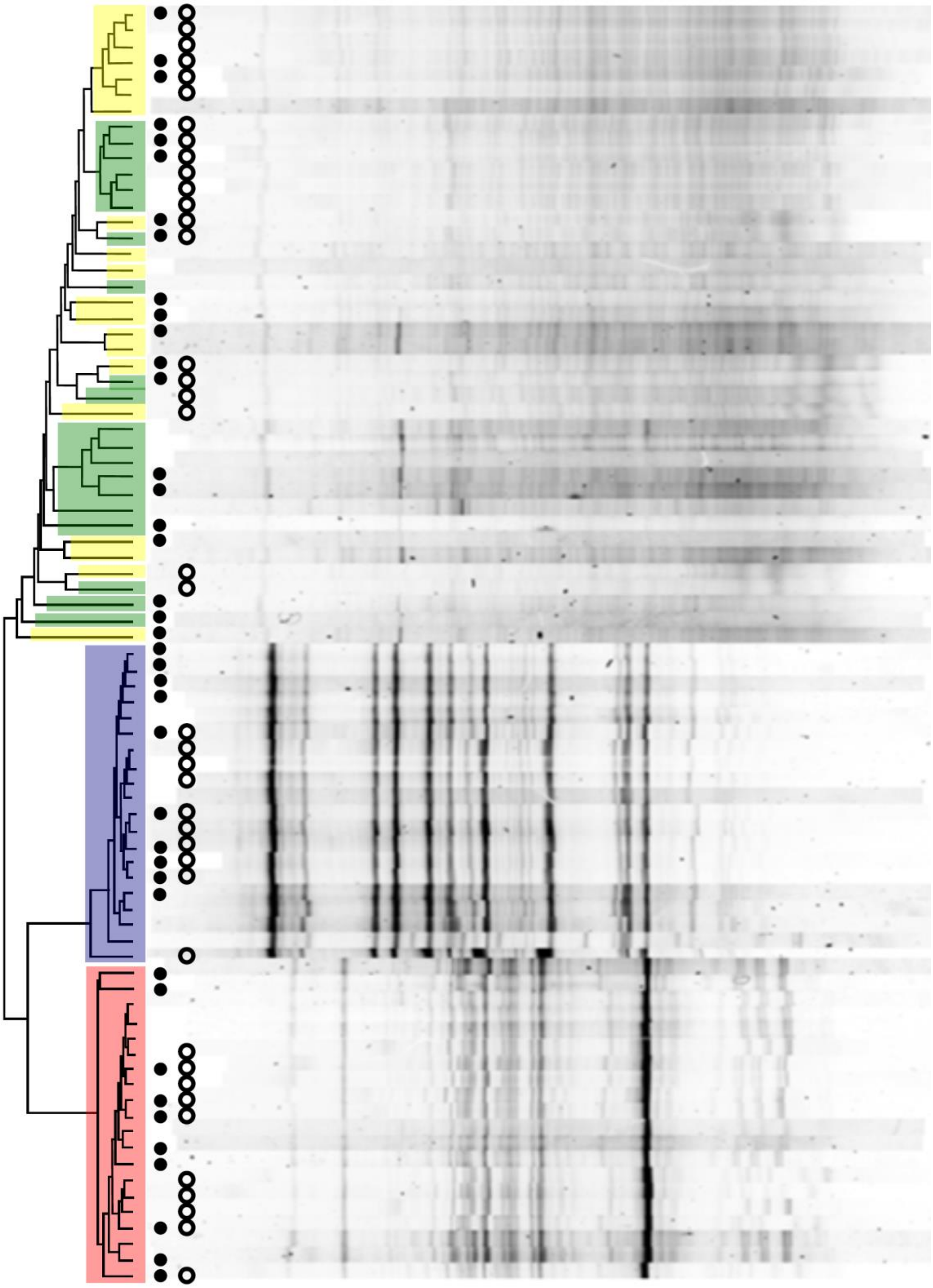

Figure 4. UPGMA dendrogram of replicate DGGE fingerprints shows that stool samples separate clearly, whereas soil samples do not. Replicates that grouped by sample are coloured: S1 in red, S3 in blue, 10AS in yellow, 6TD in green. Pooled samples are indicated by a black circle and high template concentration samples are indicated by a white circle.

3.3.1.2 Multi-response permutation procedure (MRPP)

To further test the effect of pooling PCR amplifications or analyzing individual reactions with DGGE fingerprints, I used MRPP analysis. MRPP tests the hypothesis that there is no difference between two *a priori* groups, assigned by the user. In MRPP, the T statistic is a measure of separation between groups: a more negative value means a stronger separation (Mielke Jr et al. 1976). The effect size is provided by the change-corrected within group agreement (*A*) which is a measure of within-group homogeneity. If all samples within each group are identical (i.e., *A*=1), and if the within group homogeneity is equal to what would be expected by chance (i.e., *A*=0). The groups assigned in this case were the four samples: S1, S3, 6TD, and 10AS. Pooled and not-pooled samples were compared independently. The separation and scatter of pooled samples (*A*=0.40 and *T*=-21.5) was the same as for not-pooled samples (*A*=0.40 and *T*=-21.4). The average within-group distances for each group in the analyses for pooled and not-pooled samples demonstrate no consistent effect of pooling (Fig. 5), indicating that pooling replicate reactions doesn’t increase the accuracy of DGGE, because pooled samples do not cluster more tightly than samples that were not pooled.

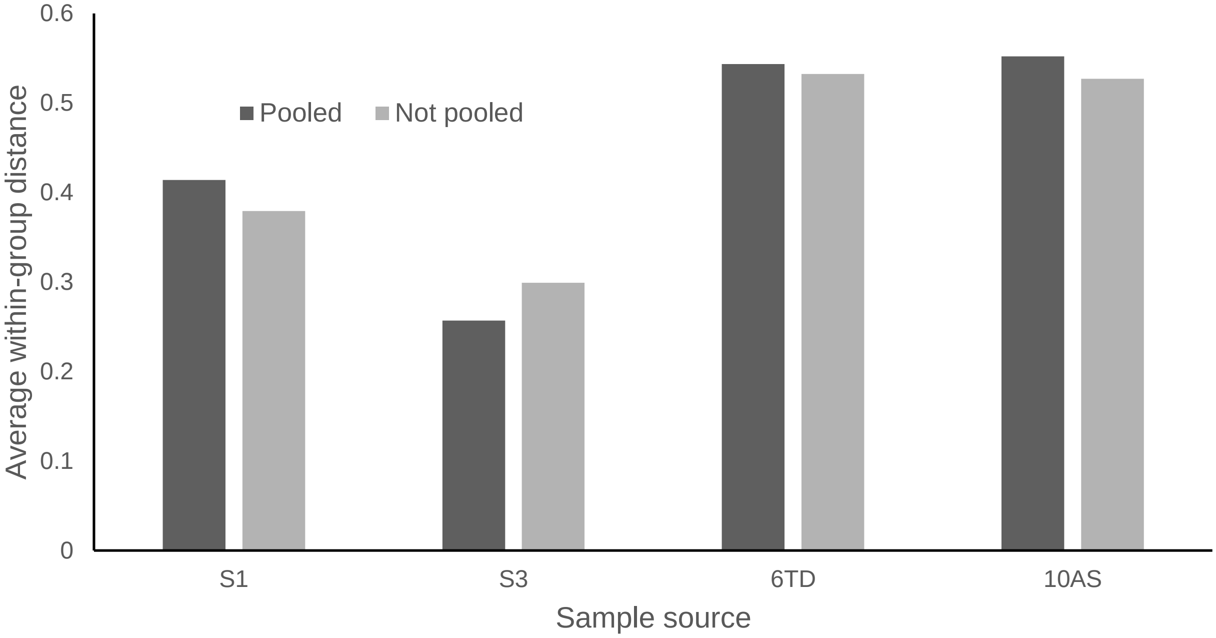

Figure 5. Average within-group distances from MRPP for pooled and not-pooled replicates when grouped by sample type shows no consistent effect of pooling.

3.3.2 Impact of Pooling on Sequencing

3.3.2.1 Impact on Ordination using NMS and MRPP

All samples that contributed fewer than 4000 sequences were removed from further analysis (Table 4; Supplementary Table 1). As the samples removed are fairly evenly distributed between high and low template concentration, sample source (S1, S3, 6TD, 10AS), and pooled and non-pooled samples, their removal should not introduce bias to further analyses. The average number of sequences contributed by a sample was 32000, ranging from 4100 to 143000.

Table 4. Summary of samples removed from analyses due to low number of sequences contributed.

Lane	Barcode	Sample	Pooled	Template Concentration	Barcode	Sequences Contributed
1	V4_7R	S1	Yes	low	CAGATC	222
1	V4_40R	S3	Yes	low	CATCTA	257
1	V4_45R	S1	No	high	GAGAAT	3699
1	V4_46R	S1	No	low	CTCAAT	2279
1	V4_48R	S1	No	low	TTGCAT	1835
1	V4_52R	6TD	No	high	AGTGTT	277
1	V4_64R	10AS	No	high	TAATCG	507
1	V4_69R	10AS	No	low	AGATTC	3006
1	V4_71R	S3	No	high	TGCGAA	147
2	V4_3R	S1	Yes	high	TTAGGC	0
2	V4_40R	S3	Yes	low	CATCTA	458

NMS plots were generated using Bray Curtis dissimilarity. When all sample replicates were analyzed together, they clustered very clearly by sample (Fig. 6), regardless of whether samples are pooled or not pooled. Within sample type, stool samples are closer to each other than to soil samples, and soil samples are closer to each other than to stool samples. Pooled and not pooled samples were compared using MRPP: for pooled samples A=0.64 and T=-45.1; for not pooled samples A=0.58 and T=-41.1. This indicates that pooling samples may result in greater within-group homogeneity and greater separation between groups, though this difference is not large.

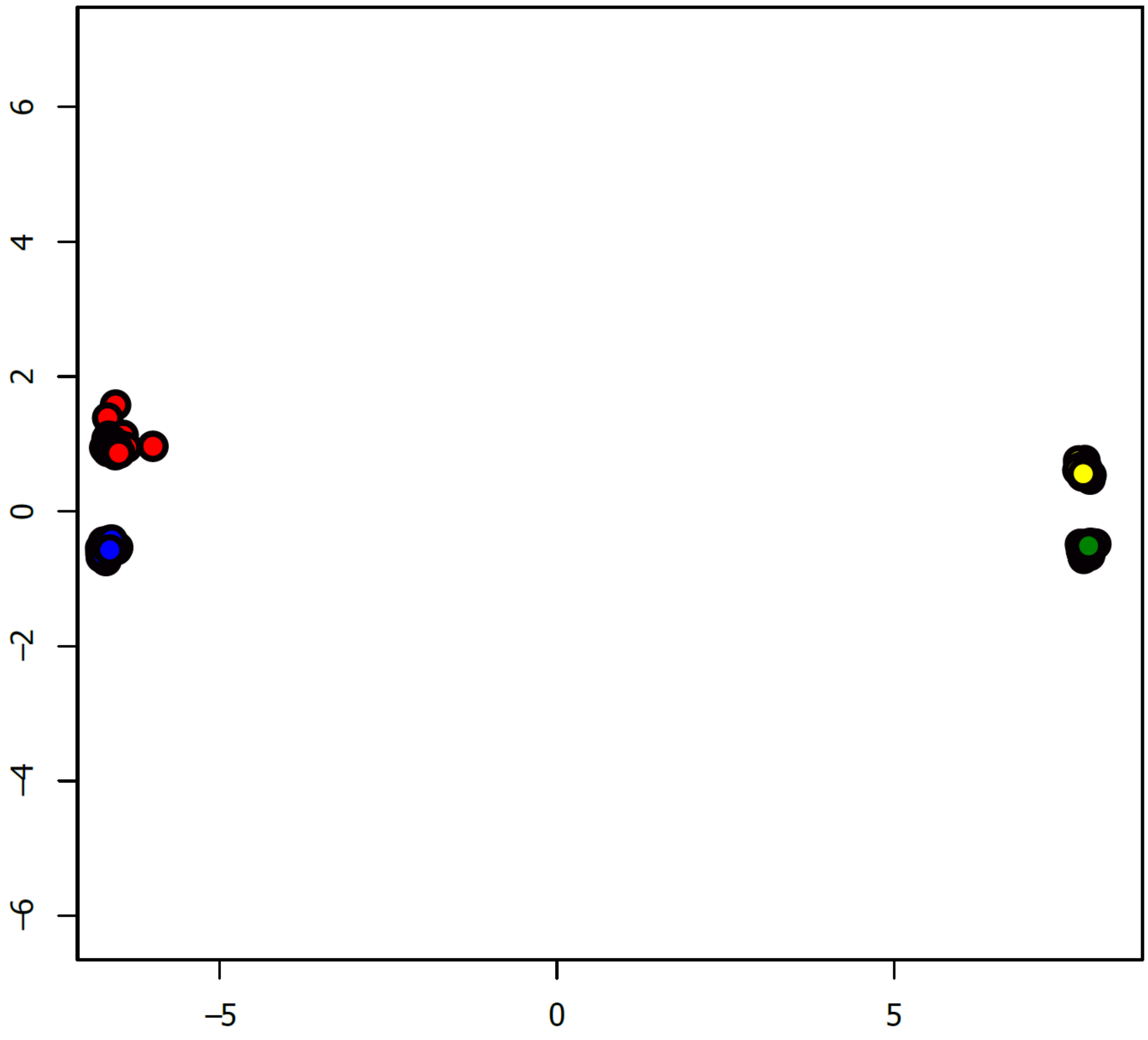

Figure 6. NMS (Bray Curtis dissimilarity) of all sample replicates from Illumina data shows replicates clustering very strongly by sample: S1 (red), S3 (blue), 6TD (green), and 10AS (yellow).

3.3.2.2 NMS and PERMDISP within each sample

Each sample was analyzed independently using AXIOME in order to examine differences that were not visible when all samples were analyzed together. Each amplification originates from the same DNA extract, therefore each point should overlap ideally. However, because PCR and sequencing are susceptible to bias there is variation between replicate amplifications of the same samples. Techniques that minimize bias will result in replicate amplifications clustering more tightly whereas techniques that increase bias with increase the variance between replicate amplifications. As the groups being analyzed within samples are not truly different samples, MRPP is not an appropriate technique to use: it provides a measure of the average homogeneity within groups across all groups rather than comparing the homogeneity of each group.

NMS was also performed on each sample for each grouping (Fig. 7): template concentration (high and low), pooling (pooled and not pooled), and lane (lane 1 and lane 2). The most apparent difference in variance was between high and low template concentrations of the two soil samples (6TD and 10AS), with the high template concentration samples clustering more tightly than low template concentration. This pattern was also present in stool samples (S1, S3, and SX), though less clearly so. One possible explanation for this difference between soil and stool samples is that soils are a much more diverse environment; a low template concentration may bias amplification more than in a less-diverse sample such as stool. Differences in variance were less apparent between pooled and non-pooled samples. Lane 1 and Lane 2 are included to demonstrate variability introduced by sample manipulation as PCR for each lane was performed on a different day for all samples except SX. Only one PCR was performed for SX and this was sequenced on both lanes to assess variability associated with separate lanes.

Although NMS provides a visual representation of differences in variance between groups, it is does not provide quantitative results. Quantitative data were obtained by using the QIIME analysis PERMDISP which tests whether the variances are significantly different between groups (Anderson 2004). Within each group the distance from each sample to the centroid of that group is calculated and these distances are plotted in box plots (Fig. 8). The largest difference between within-group distances to the centroid was between the high and low template concentration groups. Across all samples high template concentration groups and pooled groups are shown to have lower average distances to the group's centroid than low template concentration groups and not pooled groups. Samples from Lane 1 clustered more closely to the centroid, but this difference was inconsistent across samples, and the spread of distances to the centroid mostly overlaps for both lanes for all samples.

As well as having a lower average distance to the centroid of the group, high template concentration replicates are demonstrated lower variability in distance to centroid than either low template concentration replicates, or pooled replicates. This means that high template concentration replicates are more consistently close to their group's centroid, and therefore high template concentration replicates are more consistently reproducible than either low template concentration replicates, or pooled replicates.

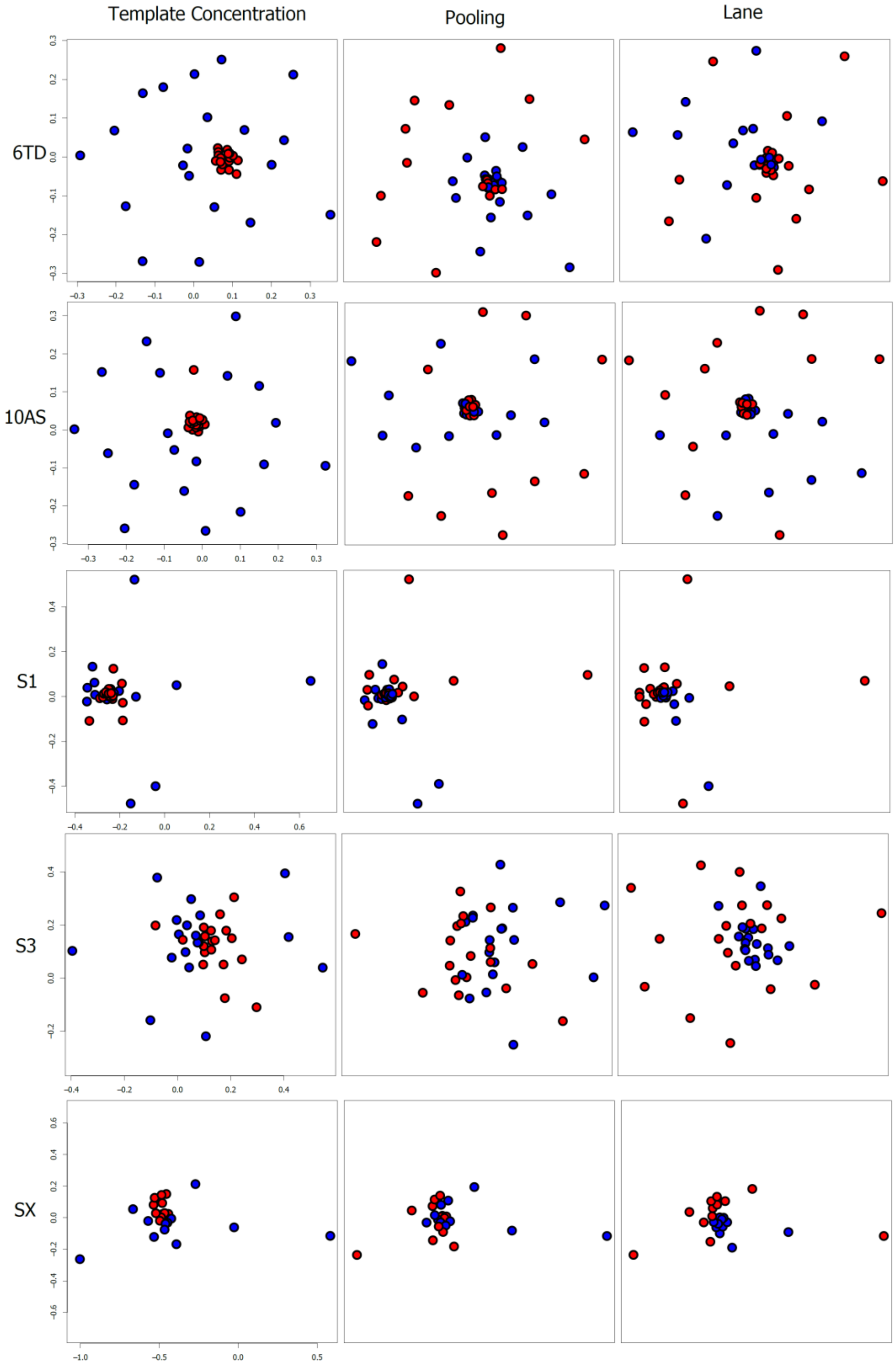

Figure 7. NMS of replicates of each sample using Bray-Curtis dissimilarity. High template concentration, not-pooled, and lane 2 replicates are coloured red, low template concentration, pooled, and lane 1 replicates are coloured blue. Scales repeated across columns from left.

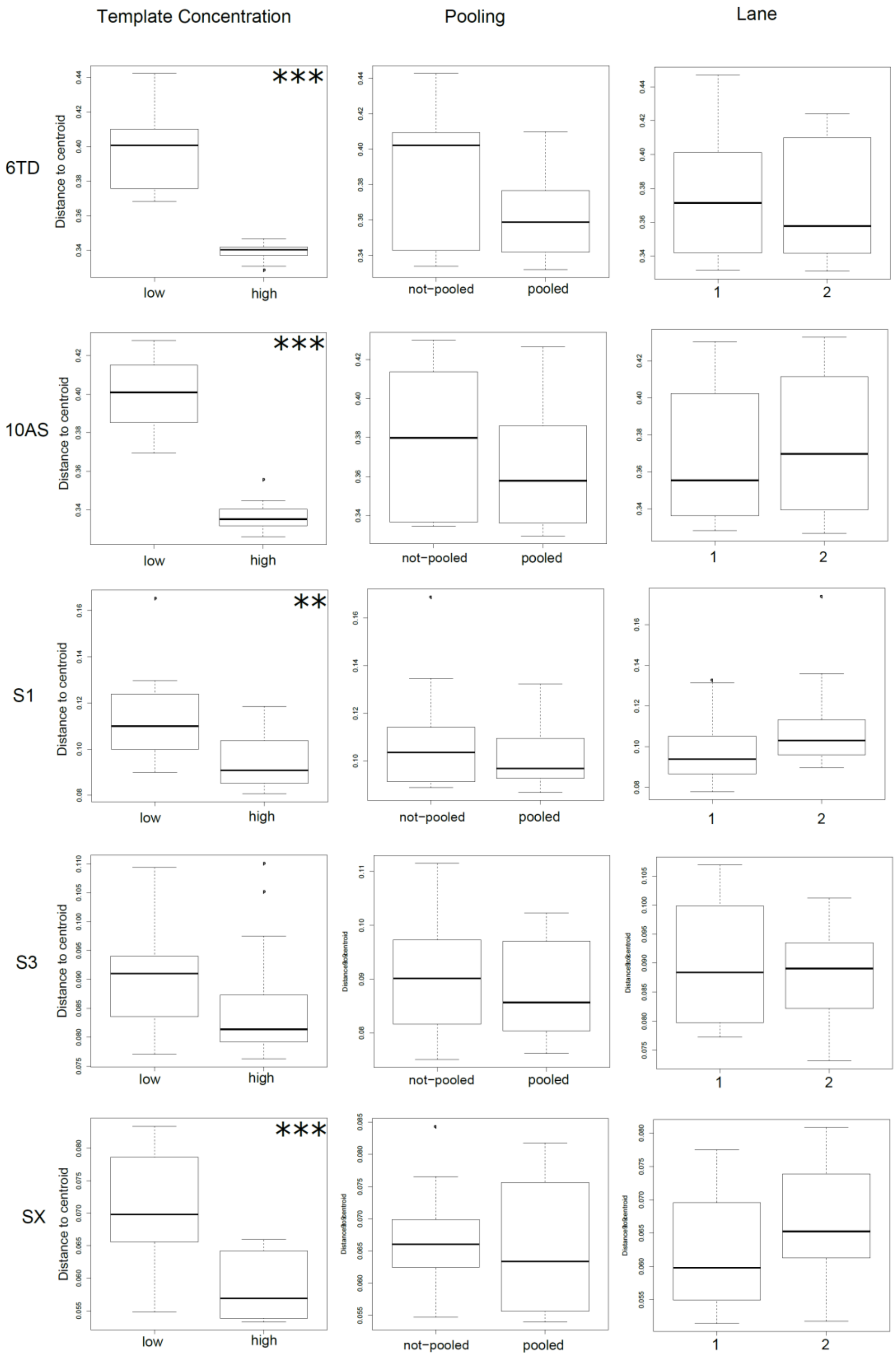

Figure 8. The distance to the centroid of the given group when using NMS ordination is plotted for replicates of each sample. Significance is indicated symbolically: 0-0.001 (***), 0.001-0.01 (**), 0.01-0.05 (*).

Within PERMDISP, Analysis of Variance (ANOVA) was implemented to test the statistical significance of the differences in variance between groups (Table 5). The *F* ratio gives the actual variance over the expected variance: if there is no difference between groups it is equal to 1 and the higher the ratio the larger the effect of the group. The *p* value is the significance of the difference. The results show that the effect of template concentration is both large and significant for both soil samples, significant for samples S1 and SX, and approaches significance for S3. Neither the effect of pooling nor that of lane is significant. The effect of pooling approaches significance for both soil samples though its effect is much less than that of template concentration. The effect of lane approaches significance for sample S1, but this is likely due to more sample replicates from Lane 1 being removed from analysis due to low sequence counts as this effect is not seen in other samples.

Table 5. Analysis of Variance between groups of replicates of each sample. The F-ratio (*F*) gives the actual variance over the expected and *p* gives the significance of the difference. Significance is also indicated symbolically for convenience: 0-0.001 (***), 0.001-0.01 (**), 0.01-0.05 (*).

	Template concentration	Pooling	Lane
6TD	*F* = 139.59 *p* = 6.028e-14 ***	*F* = 1.7699 *p* = 0.1918	*F* = 0.1973 *p* = 0.6596
10AS	*F* = 188.2 *p* = 7.222e-16 ***	*F* = 3.3124 *p* = 0.07686	*F* = 0.0214 *p* = 0.8845
S1	*F* =11.16 *p* = 0.002086 **	*F* = 1.2164 *p* = 0.2781	*F* = 3.1558 *p* = 0.08488
S3	*F* = 3.7378 *p* = 0.06132	*F* = 0.9439 *p* = 0.3379	*F* = 0.2749 *p* = 0.6034
SX	F = 17.832 *p* = 0.0003501 ***	*F* = 0.1687 *p* = 0.6852	*F* = 1.8203 *p* = 0.191

3.3.2.3 To pool or not to pool

Based on the data produced, pooling does reduce the impact of bias due to PCR drift. However, the impact of pooling triplicate amplifications does not have as great an impact as the concentration of the template DNA.

The impact of pooling in reducing bias is not sufficient to impact the clustering of samples, even when they are from similar sources: both stool samples and soil samples separated from each other clearly and clustered by sample type (soil or stool). Pooling triplicate reactions may help separate very similar samples, such as stool samples from the same individual or soil samples from the same location. Pooling also would be of greater value for samples with low template concentrations as they are more susceptible to PCR biases than samples with high template concentration. To test this, I ran MRPP on all samples divided into four groups: all non-pooled high template concentration (*A*=0.70, *T*=-21.11); all pooled high template concentration (*A*=0.75, *T*=-23.12); all non-pooled low template concentration (*A*=0.55, *T*=-20.75); all pooled low template concentration (*A*=0.60, *T*=-21.16). As hypothesized, the difference in within-group homogeneity was greater for low template concentration samples than for high template concentration samples (Δ*A*=0.06 for low template concentration; Δ*A* =0.04 for high template concentration). This is much less than the difference in within-group homogeneity between high template concentration samples and low concentration samples for either pooled or not pooled samples (Δ*A* =0.15 for pooled; Δ*A* =0.16 for non-pooled). Pooling may also be more important for higher-diversity samples, such as soils, or when examining the rare biosphere.

Caporaso and coworkers questioned whether the decreasing cost of sequencing should result in sequencing samples more deeply or sequencing a larger number of samples and concluded that increasing sequencing depth is not likely to provide additional insight comparable to inclusion of more samples (Caporaso et al., 2012). The question of whether to pool triplicate amplifications is similar, though the number of PCR amplifications is not strictly limited and so it is possible to both include more samples and pool triplicate amplifications. Protocols for both the Earth Microbiome Project and the Human Microbiome project require the pooling of triplicate reactions (Gilbert et al. 2010; Peterson et al. 2009). However, the increase in time and effort required to triple the number of amplifications being performed is likely not be warranted in cases where samples are very different.

Chapter 4. Conclusions and Future Considerations

4.1 Conclusions

The study of the human microbiota is currently a very prominent area of research within both microbial ecology and human pathophysiology. The microbial community composition of these environments differ from non-host-associated communities such as those found in soil and water likely due to the bi-directional selective pressures of the host-microbe mutualism (Ley et al. 2006). In order to effectively analyze these microbial communities, available methods must be examined critically in terms of their strengths and weaknesses. My research compared a DGGE fingerprinting method and Illumina paired-end sequencing in terms of sensitivity, clustering behaviour, template concentration, and the value of pooling triplicate reactions in reducing the impact of PCR bias on each method.

DGGE requires that the fluorescence of a band be greater than background fluorescence in order for the band to be included in analyses (Muyzer and Smalla 1998). In order for a band to have enough DNA to produce this level of fluorescence, a species must be abundant within the sample being analyzed. DGGE bands picked from gel fingerprints of human microbiome samples in this thesis were shown to represent more than 0.1% of their respective sample (excluding one outlier, which represented less than 0.1% of its sample; Table 1). This limit of the resolving power of DGGE, due to its relatively low signal-to-noise ratio, restricts DGGE to the analysis of dominant community members (Green et al. 2010). This degree of sensitivity is appropriate for examining changes in abundant microorganisms or varying conditions, but is not sufficient for examining the true extent of biodiversity within a given community. DGGE is also limited by its inability to distinguish individual bands within high-diversity samples, such as soils (Green et al. 2010;

Inglis et al. 2012). Even though DGGE could differentiate replicate amplifications of two stool samples, it could not differentiate between replicate amplifications of soil samples from different environments (Fig. 4).

DGGE separates DNA fragments based on their melting behaviour (i.e. the concentration of denaturant at which they become partially denatured), which is not necessarily unique to a given fragment (Fischer and Lerman 1980). Within a sample, a single band can represent multiple species of bacteria. In addition, a band in the same position for two disparate samples is unlikely to represent the same species (Muyzer et al. 1993). Ordination cannot accurately cluster DGGE fingerprints derived from distinct environments, such as the mouth and colon, because these bands at the same position are interpreted as representing the same species. Because of the lack of direct relationship between species and bands, DGGE is not an appropriate tool for the analysis of samples from disparate environments, but is well suited to analyses of samples from very similar environments, or the same environment over time or varying treatments.

Both DGGE and Illumina sequencing require PCR amplification and are therefore affected by its biases. The pooling of triplicate amplifications was suggested to reduce the impact of the bias due to non-reproducible fluctuations in amplification efficiencies (Polz and Cavanaugh 1998), but the impact of this pooling has not been previously demonstrated. In my thesis research, the impact of pooling on DGGE and sequencing was evaluated by comparing amplifications of samples that were pooled in triplicate to individual reactions for both high (5-10 ng/µL) and low (0.1 ng/µL) template concentration and high (soil) and low (stool) diversity. My results did not demonstrate a measurable effect of pooling replicates on DGGE fingerprint clustering. This is likely because more abundant organisms are less affected by pooling than rarer organisms which are not observable using DGGE due to the limits of this method's sensitivity.

Unlike DGGE, sequencing results can be affected by pooling: groups of sample replicates were more homogeneous and had greater between-group separation when triplicate PCR amplifications were pooled. However, pooling also resulted in less reproducibility within samples than using a higher concentration of template. The impact of pooling seems to be especially important for high-diversity samples, and had a greater impact on samples with low template concentration than those with high template concentration. This corroborates the finding of others that low template concentration results in greater stochastic fluctuations in PCR amplifications (Chandler et al. 1997). For low diversity samples or high template concentrations, the effect of pooling was almost equivalent to the variability that was observed between sequencing lanes. When these conditions are met, or when samples are from sufficiently different environments (e.g. different individuals), the pooling of triplicate amplifications is therefore not required.

Both DGGE fingerprinting and Illumina sequencing are useful tools for examining the human microbiota. DGGE provides a rapid and economical visual representation of the major constituents of a microbial community, and is well suited to tracking large changes in community composition over time or with varying treatments. Illumina sequencing provides a much greater depth of coverage and therefore allows for the analysis of low-abundance organisms within a community. It also allows for comparison and ordination of any samples, unlike DGGE, which requires samples to be from similar communities. The strengths and weaknesses of these methods are somewhat complementary and the methods are best used in tandem to effectively analyse the human microbiota.

4.2 Future Directions

The work presented in this thesis compares the fingerprinting method DGGE and the Illumina sequencing platform in terms of their sensitivity, clustering behaviour of resulting data, and reproducibility for the purposes of studying the human microbiota. As the cost and time requirements associated with amplicon sequencing and data analysis continue to be reduced by technological advances, continued analysis will be needed to weigh the benefits and drawbacks of each of these methods.

In order to evaluate the impact of pooling triplicate amplifications further, several other sources of bias should be examined. In this study, three DNA extractions were pooled for each sample. Further study should evaluate the impact of pooling multiple DNA extractions by comparing replicates from pooled DNA from multiple extractions to replicates from a single extraction. Variability in samples taken from a single community due to spatially varying composition may also result in variability in analyses. The difference between extractions taken from different portions of a donation, or from nearby areas in case of soil or other environmental samples, should also be examined. If significantly more variability results from not pooling multiple extractions, or between two samplings, than results from not pooling triplicate amplifications during PCR, this pooling would not be justified.

Both concentrations of template used in this thesis are fairly low compared to what is expected for high density human host environments, such as the colon. Based on the effect of template concentration observed, it is probable that higher concentrations of DNA (i.e. 50 ng/µL) would further reduce variation between replicates and eliminate the need for pooling multiple PCR amplifications.

4.3 Significance

The scope of microbial ecology encompasses a vast array of research areas from biogeochemical cycles to human health, and molecular methods such as DGGE and NGS continue to be instrumental in the expansion of this field. 16S rRNA gene fingerprinting and sequencing are invaluable tools for the study of microbial ecology. These molecular methods are capable of revealing levels of diversity that are not detectable through culturing. This research tests established practices for performing community analysis from various environments using PCR. Current protocols for both the Human Microbiome Project (Human Microbiome Project Consortium 2012) and the Earth Microbiome Project (Gilbert et al. 2010) include the pooling of triplicate PCR amplifications. However, my research suggests that will likely result in detectable differences. Although eliminating the pooling of triplicate PCR amplifications would decrease the effort associated with adding additional samples to any analysis, pooling should be maintained as an added precautionary “best practice”. That said, my data suggest that maximizing substrate concentrations for PCR (e.g., >10 ng) is a significantly important methodological consideration for PCR-based analyses of marker genes by NGS.

References

Abrams, G.D., Bauer, H., and Sprinz, H. 1963. Influence of the normal flora on mucosal morphology and cellular renewal in the ileum. A comparison of germ-free and conventional mice. Lab. Invest. **12**: 355-364.

Alon, S., Vigneault, F., Eminaga, S., Christodoulou, D.C., Seidman, J.G., Church, G.M., and Eisenberg, E. 2011. Barcoding bias in high-throughput multiplex sequencing of miRNA. Genome Res. **21**: 1506-1511.

Anderson, M. 2004. PERMDISP: a FORTRAN computer program for permutational analysis of multivariate dispersions (for any two-factor ANOVA design) using permutation. Department of Statistics, University of Auckland, New Zealand.

Azad, M.B., Becker, A.B., Guttman, D.S., Sears, M.R., Scott, J.A., Kozyrskyj, A.L., and Canadian Healthy Infant Longitudinal Development Study Investigators. 2013. Gut microbiota diversity and atopic disease: does breast-feeding play a role? J. Allergy Clin. Immunol. **131**: 247-248.

Bäckhed, F., Ding, H., Wang, T., Hooper, L.V., Koh, G.Y., Nagy, A., Semenkovich, C.F., and Gordon, J.I. 2004. The gut microbiota as an environmental factor that regulates fat storage. Proc. Natl. Acad. Sci. U. S. A. **101**: 15718-15723.

Bartram, A.K., Lynch, M.D.J., Stearns, J.C., Moreno-Hagelsieb, G., and Neufeld, J.D. 2011. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Appl. Environ. Microbiol. **77**: 3846-3852.

Bernet, M.F., Brassart, D., Neeser, J.R., and Servin, A.L. 1994. *Lactobacillus acidophilus* LA 1 binds to cultured human intestinal cell lines and inhibits cell attachment and cell invasion by enterovirulent bacteria. Gut **35**: 483-489.

Berry, D., Mahfoudh, K.B., Wagner, M., and Loy, A. 2011. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl. Environ. Microbiol. **77**: 7846-7849.

Bik, E.M., Eckburg, P.B., Gill, S.R., Nelson, K.E., Purdom, E.A., Francois, F., Perez-Perez, G., Blaser, M.J., and Relman, D.A. 2006. Molecular analysis of the bacterial microbiota in the human stomach. Proc. Natl. Acad. Sci. U. S. A. **103**: 732-737.

Brandtzaeg, P., Haslstensen, T.S., Kett, K., Krajci, P., Kvale, D., Rognum, T.O., Scott, H., and Sollid, L.M. 1989. Immunobiology and immunopathology of human gut mucosa: humoral immunity and intraepithelial lymphocytes. Gastroenterology **97**: 1562-1584.

Brandtzaeg, P. 2009. Mucosal immunity: induction, dissemination, and effector functions. Scand. J. Immunol. **70**: 505-515.

Bray, J.R., and Curtis, J.T. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. **27**: 325-349.

Butler, J.E., Sun, J., Weber, P., Navarro, P., and Francis, D. 2000. Antibody repertoire development in fetal and newborn piglets, III. Colonization of the gastrointestinal tract selectively diversifies the preimmune repertoire in mucosal lymphoid tissues. Immunology **100**: 119-130.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., and Knight, R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U. S. A. **108**: 4516-4522.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods **7**: 335-336.

Cardona, S., Eck, A., Cassellas, M., Gallart, M., Alastrue, C., Dore, J., Azpiroz, F., Roca, J., Guarner, F., and Manichanh, C. 2012. Storage conditions of intestinal microbiota matter in metagenomic analysis. BMC Microbiol. **12**: 158.

Chandler, D.P., Fredrickson, J.K., and Brockman, F.J. 1997. Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. Mol. Ecol. **6**: 475-482.

Cummings, J.H., Beatty, E.R., Kingman, S.M., Bingham, S.A., and Englyst, H.N. 1996. Digestion and physiological properties of resistant starch in the human large bowel. Br. J. Nutr. **75**: 733-747.

De Cáceres, M., Legendre, P., and Moretti, M. 2010. Improving indicator species analysis by combining groups of sites. Oikos **119**: 1674-1684.

DiGiulio, D.B., Romero, R., Amogan, H.P., Kusanovic, J.P., Bik, E.M., Gotsch, F., Kim, C.J., Erez, O., Edwin, S., and Relman, D.A. 2008. Microbial prevalence, diversity and abundance in amniotic fluid during preterm labor: a molecular and culture-based investigation. PLoS One **3**: e3056.

Dominguez-Bello, M.G., Costello, E.K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. 2010. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc. Natl. Acad. Sci. U. S. A. **107**: 11971-11975.

Dufrene, M., and Legendre, P. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol. Monogr. **67**: 345-366.

Fischer, S.G., and Lerman, L.S. 1980. Separation of random fragments of DNA according to properties of their sequences. Proc. Natl. Acad. Sci. U. S. A. **77**: 4420-4424.

Forsythe, P., Sudo, N., Dinan, T., Taylor, V.H., and Bienenstock, J. 2010. Mood and gut feelings. Brain Behav. Immun. **24**: 9-16.

Frank, D., and Pace, N. 2008. Gastrointestinal microbiology enters the metagenomics era. Curr. Opin. Gastroenterol. **24**: 4-10.

Gentil, V., Lader, M.H., Kantamaneni, B.D., and Curzon, G. 1977. Effects of adrenaline injection on human plasma tryptophan and non-esterified fatty acids. Clin. Sci. Mol. Med. **53**: 227-232.

Gilbert, J.A., Meyer, F., Jansson, J., Gordon, J., Pace, N., Tiedje, J., Ley, R., Fierer, N., Field, D., and Kyrpides, N. 2010. The Earth Microbiome Project: Meeting report of the "1st EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6th 2010. Stand. Genomic Sci. **3**: 249.

Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., and Nelson, K.E. 2006. Metagenomic analysis of the human distal gut microbiome. Science **312**: 1355-1359.

Gower, J.C. 2005. Principal coordinates analysis. *Encyclopedia of Biostatistics*.

Graeff, F.G., Guimarães, F.S., De Andrade, T.G., and Deakin, J.F. 1996. Role of 5-HT in stress, anxiety, and depression. Pharmacology Biochemistry and Behavior **54**: 129-141.

Green, S.J., Leigh, M.B., and Neufeld, J.D. 2010. Denaturing gradient gel electrophoresis (DGGE) for microbial community analysis. *In* Handbook of Hydrocarbon and Lipid Microbiology. *Edited by Timmis, K.N.* Springer, Berlin: pp. 4137-4158.

Guarner, F., and Malagelada, J. 2003. Gut flora in health and disease. Lancet **361**: 512-519.

Hamady, M., and Knight, R. 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res. **19**: 1141-1152.

Hooper, L.V., and Gordon, J.I. 2001. Commensal host-bacterial relationships in the gut. Science **292**: 1115-1118.

Hooper, L.V., Xu, J., Falk, P.G., Midtvedt, T., and Gordon, J.I. 1999. A molecular sensor that allows a gut commensal to control its nutrient foundation in a competitive ecosystem. Proc. Natl. Acad. Sci. U. S. A. **96**: 9833-9838.

Hugenholtz, P., Goebel, B.M., and Pace, N.R. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. **180**: 4765-4774.

Human Microbiome Project Consortium. 2012. A framework for human microbiome research. Nature **486**: 215-221.

Hutchison, C.A. 2007. DNA sequencing: bench to bedside and beyond. Nucleic Acids Res. **35**: 6227-6237.

Inglis, G.D., Thomas, M.C., Thomas, D.K., Kalmokoff, M.L., Brooks, S.P.J., and Selinger, L.B. 2012. Molecular methods to measure intestinal bacteria: a review. J. AOAC Int. **95**: 5-23.

Institute of Medicine (US). 2006. Microbial communities of the gut. Forum on Microbial Threats. Ending the war methaphor: the changing agenda for unraveling the host-microbiome relationship: workshop summary. Washington (DC): National Academies Press (US).

Iyer, L.M., Aravind, L., Coon, S.L., Klein, D.C., and Koonin, E.V. 2004. Evolution of cell-cell signaling in animals: did late horizontal gene transfer from bacteria have a role? Trends Genet. **20**: 292-299.

Jiménez, E., Marín, M.L., Martín, R., Odriozola, J.M., Olivares, M., Xaus, J., and Rodríguez, J.M. 2008. Is meconium from healthy newborns actually sterile? Res. Microbiol. **159**: 187-193.

Jiménez, E., Fernández, L., Marín, M.L., Martín, R., Odriozola, J.M., Nueno-Palop, C., and Rodríguez, J.M. 2005. Isolation of commensal bacteria from umbilical cord blood of healthy neonates born by Cesarean section. Curr. Microbiol. **51**: 270-274.

Jumpertz, R., Le, D.S., Turnbaugh, P.J., Trinidad, C., Bogardus, C., Gordon, J.I., and Krakoff, J. 2011. Energy-balance studies reveal associations between gut microbes, caloric load, and nutrient absorption in humans. Am. J. Clin. Nutr. **94**: 58-65.

Kinross, J.M., von Roon, A.C., Holmes, E., Darzi, A., and Nicholson, J.K. 2008. The human gut microbiome: implications for future health care. Curr. Gastroenterol. Rep. **10**: 396-403.

Koch, R. 1890. Uber bakteriologische Forschung. 10th International Congress of Medicine, Berlin.

Kruskal, J.B. 1964. Nonmetric multidimensional scaling: a numerical method. Psychometrika **29**: 115-129.

Kuczynski, J., Lauber, C.L., Walters, W.A., Parfrey, L.W., Clemente, J.C., Gevers, D., and Knight, R. 2012. Experimental and analytical tools for studying the human microbiome. Nat. Rev. Genet. **13**: 47-58.

Lakhan, S.E., and Kirchgessner, A. 2010. Gut inflammation in chronic fatigue syndrome. Nutr. Metab. **7**: 79.

Lerman, L.S., Fischer, S.G., Hurley, I., Silverstein, K., and Lumelsky, N. 1984. Sequence-determined DNA separations. Annu. Rev. Biophys. Bioeng. **13**: 399-423.

Ley, R.E., Peterson, D.A., and Gordon, J.I. 2006. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell **124**: 837-848.

Ley, R.E., Bäckhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., and Gordon, J.I. 2005. Obesity alters gut microbial ecology. Proc. Natl. Acad. Sci. U. S. A. **102**: 11070-11075.

Li, M., Wang, B., Zhang, M., Rantalainen, M., Wang, S., Zhou, H., Zhang, Y., Shen, J., Pang, X., Zhang, M., Wei, H., Chen, Y., Lu, H., Zuo, J., Su, M., Qiu, Y., Jia, W., Xiao, C., Smith, L.M., Yang, S., Holmes, E., Tang, H., Zhao, G., Nicholson, J.K., Li, L., and Zhao, L. 2008. Symbiotic gut microbes modulate human metabolic phenotypes. Proc. Natl. Acad. Sci. U. S. A. **105**: 2117-2122.

Li, W., Godzik, A. 2006. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.

Lozupone, C., and Knight, R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. **71**: 8228-8235.

Lynch, M.D., Masella, A.P., Hall, M.W., Bartram, A.K., and Neufeld, J.D. 2013. AXIOME: automated exploration of microbial diversity. Giga Sci. **2**: 3.

Macpherson, A.J., and Uhr, T. 2004. Induction of protective IgA by intestinal dendritic cells carrying commensal bacteria. Science **303**: 1662-1665.

Macpherson, A.J., Geuking, M.B., and McCoy, K.D. 2005. Immune responses that adapt the intestinal mucosa to commensal intestinal bacteria. Immunology **115**: 153-162.

Maes, M. 2011. Depression is an inflammatory disease, but cell-mediated immune activation is the key component of depression. Prog. Neuropsychopharmacol. Biol. Psychiatry **35**: 664-675.

Mager, D., Haffajee, A., Devlin, P., Norris, C., Posner, M., and Goodson, J. 2005. The salivary microbiota as a diagnostic indicator of oral cancer: a descriptive, non-randomized study of cancer-free and oral squamous cell carcinoma subjects. J. Transl. Med. **3**: 27.

Mardis, E.R. 2013. Next-generation sequencing platforms. Annu. Rev. Anal. Chem. **6**: 287-303.

Martin, R., Langa, S., Reviriego, C., Jiminez, E., Marin, M.L., Xaus, J., Fernandez, L., and Rodriguez, J.M. 2003. Human milk is a source of lactic acid bacteria for the infant gut. J. Pediatr. **143**: 754-758.

Martin, V., Maldonado-Barragan, A., Moles, L., Rodriguez-Banos, M., Campo, R.D., Fernandez, L., Rodriguez, J.M., and Jimenez, E. 2012. Sharing of bacterial strains between breast milk and infant feces. J. Hum. Lact. **28**: 36-44.

Martin-Laurent, F., Philippot, L., Hallet, S., Chaussod, R., Germon, J., Soulas, G., and Catroux, G. 2001. DNA extraction from soils: old bias for new microbial diversity analysis methods. Appl. Environ. Microbiol. **67**: 2354-2359.

McCune, B., Grace, J., and Urban, D. 2002. Analysis of ecological communities. MjM Software Design, USA.

McFall-Ngai, M. 2007. Adaptive immunity: care for the community. Nature **445**: 153.

Mielke Jr, P.W., Berry, K.J., and Johnson, E.S. 1976. Multi-response permutation procedures for a priori classifications. Communications in Statistics-Theory and Methods **5**: 1409-1424.

Muyzer, G., and Smalla, K. 1998. Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. Anton. Leeuw. Int. J. G. **73**: 127-141.

Muyzer, G., de Waal, E.C., and 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. Appl. Environ. Microbiol. **59**: 695-700.

Myers, R.M., Fischer, S.G., Lerman, L.S., and Maniatis, T. 1985. Nearly all single base substitutions in DNA fragments joined to a GC-clamp can be detected by denaturing gradient gel electrophoresis. Nucleic Acids Res. **13**: 3131-3145.

Nagalingam, N.A., and Lynch, S.V. 2012. Role of the microbiota in inflammatory bowel diseases. Inflamm. Bowel Dis. **18**: 968-984.

Neilson, J.W., Jordan, F.L., and Maier, R.M. 2013. Analysis of artifacts suggests DGGE should not be used for quantitative diversity analysis. J. Microbiol. Meth. **92**: 256-263.

Neufeld, J.D., Engel, K., Cheng, J., Moreno-Hagelsieb, G., Rose, D., and Charles, T. 2011. Open resource metagenomics: a model for sharing metagenomic libraries. Stand. Genomic Sci. **5**: 203.

Neufeld, J.D., and Mohn, W.W. 2006. Assessment of microbial phylogenetic diversity based on environmental nucleic acids. *In* Molecular identification, systematics, and population structure of prokaryotes. *Edited by* Anonymous. Springer, Berlin. pp. 219-259.

Nübel, U., Engelen, B., Felske, A., Snaidr, J., Wieshuber, A., Amann, R.I., Ludwig, W., and Backhaus, H. 1996. Sequence heterogeneities of genes encoding 16S rRNAs in *Paenibacillus polymyxa* detected by temperature gradient gel electrophoresis. J. Bacteriol. **178**: 5636-5643.

O'Hara, A.M., and Shanahan, F. 2006. The gut flora as a forgotten organ. EMBO Rep. **7**: 688-693.

O'Mahony, S.M., Marchesi, J.R., Scully, P., Codling, C., Ceolho, A., Quigley, E.M.M., Cryan, J.F., and Dinan, T.G. 2009. Early life stress alters behavior, immunity, and microbiota in rats: implications for irritable bowel syndrome and psychiatric illnesses. Biol. Psychiatry **65**: 263-267.

Ostman, S., Rask, C., Wold, A.E., Hultkrantz, S., and Telemo, E. 2006. Impaired regulatory T cell function in germ-free mice. Eur. J. Immunol. **36**: 2336-2346.

Parracho, H.M.R.T., Bingham, M.O., Gibson, G.R., and McCartney, A.L. 2005. Differences between the gut microflora of children with autistic spectrum disorders and that of healthy children. J. Med. Microbiol. **54**: 987-991.

Pasteur, L. 1881. On the Germ Theory. Science **2**: 420-422.

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., and Deal, C. 2009. The NIH human microbiome project. Genome Res. **19**: 2317-2323.

Polz, M.F., and Cavanaugh, C.M. 1998. Bias in template-to-product ratios in multitemplate PCR. Appl. Environ. Microbiol. **64**: 3724-3730.

Price, M.N., Dehal, P.S., and Arkin, A.P. 2010. FastTree2—approximately maximum-likelihood trees for large alignments. PLoS One. 5: e9490. doi:10.1371/journal.pone.0009490.

Roberfroid, M.B., Bornet, F., Bouley, C., and Cummings, J.H. 1995. Colonic microflora: nutrition and health. Summary and Conclusions of an International Life Sciences Institute (ILSI), Barcelona, Spain, Vol. 53, pp. 127-130.

Round, J.L., and Mazmanian, S.K. 2009. The gut microbiota shapes intestinal immune responses during health and disease. Nat. Rev. Immunol. **9**: 313-323.

Round, J.L., O'Connell, R.M., and Mazmanian, S.K. 2010. Coordination of tolerogenic immune responses by the commensal microbiota. J. Autoimmun. **34**: 220-225.

Rubin, B.E., Gibbons, S.M., Kennedy, S., Hampton-Marcell, J., Owens, S., and Gilbert, J.A. 2013. Investigating the impact of storage conditions on microbial community composition in soil samples. PloS One **8**: e70460.

Russell, S.L., Gold, M.J., Hartmann, M., Willing, B.P., Thorson, L., Wlodarska, M., Gill, N., Blanchet, M., Mohn, W.W., and McNagny, K.M. 2012. Early life antibiotic-driven changes in microbiota enhance susceptibility to allergic asthma. EMBO Rep. **13**: 440-447.

Sanger, F., Nicklen, S., and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. **74**: 5463-5467.

Scanlan, P.D., Shanahan, F., O'Mahony, C., and Marchesi, J.R. 2006. Culture-independent analyses of temporal variation of the dominant fecal microbiota and targeted bacterial subgroups in Crohn's disease. J. Clin. Microbiol. **44**: 3980-3988.

Schroeder, F.A., Lin, C.L., Crusio, W.E., and Akbarian, S. 2007. Antidepressant-like effects of the histone deacetylase inhibitor, sodium butyrate, in the mouse. Biol. Psychiatry **62**: 55-64.

Sekirov, I., Russell, S.L., Antunes, L.C.M., and Finlay, B.B. 2010. Gut microbiota in health and disease. Physiol. Rev. **90**: 859-904.

Shendure, J., and Ji, H. 2008. Next-generation DNA sequencing. Nat. Biotechnol. **26**: 1135-1145.

Simon, G.L., and Gorbach, S.L. 1984. Intestinal flora in health and disease. Gastroenterology **86**: 174-193.

Sleator, R.D. 2010. The human superorganism: Of microbes and men. Med. Hypotheses **74**: 214-215.

Sommer, F., and Bäckhed, F. 2013. The gut microbiota: masters of host development and physiology. Nat. Rev. Microbiol. **11**: 227-238.

Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C., Westover, B.P., Weatherford, J., Buhler, J.D., and Gordon, J.I. 2005. Glycan foraging in vivo by an intestine-adapted bacterial symbiont. Science **307**: 1955-1959.

Stackebrandt, E., and Goodfellow, M. 1991. Nucleic acid techniques in bacterial systematics (Vol. 4). John Wiley & Son Ltd, USA.

Stearns, J.C., Lynch, M.D.J., Senadheera, D.B., Tenenbaum, H.C., Goldberg, M.B., Cvitkovitch, D.G., Croitoru, K., Moreno-Hagelsieb, G., and Neufeld, J.D. 2011. Bacterial biogeography of the human digestive tract. Sci. Rep. **1**: 170.

Streit, W.R., and Schmitz, R.A. 2004. Metagenomics–the key to the uncultured microbes. Curr. Opin. Microbiol. **7**: 492-498.

Sudo, N., Chida, Y., Aiba, Y., Sonoda, J., Oyama, N., Yu, X., Kubo, C., and Koga, Y. 2004. Postnatal microbial colonization programs the hypothalamic-pituitary-adrenal system for stress response in mice. J. Physiol. (Lond.) **558**: 263-275.

Suzuki, K., Meek, B., Doi, Y., Muramatsu, M., Chiba, T., Honjo, T., and Fagarasan, S. 2004. Aberrant expansion of segmented filamentous bacteria in IgA-deficient gut. Proc. Natl. Acad. Sci. U. S. A. **101**: 1981-1986.

Suzuki, M.T., and Giovannoni, S.J. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl. Environ. Microbiol. **62**: 625-630.

Tsankova, N.M., Berton, O., Renthal, W., Kumar, A., Neve, R.L., and Nestler, E.J. 2006. Sustained hippocampal chromatin regulation in a mouse model of depression and antidepressant action. Nat. Neurosci. **9**: 519-525.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444**: 1027-1031.

Umesaki, Y., Setoyama, H., Matsumoto, S., and Okada, Y. 1993. Expansion of alpha beta T-cell receptor-bearing intestinal intraepithelial lymphocytes after microbial colonization in germ-free mice and its independence from thymus. Immunology **79**: 32-37.

Wagner, A., Blackstone, N., Cartwright, P., Dick, M., Misof, B., Snow, P., Wagner, G.P., Bartels, J., Murtha, M., and Pendleton, J. 1994. Surveys of gene families using polymerase chain reaction: PCR selection and PCR drift. Syst. Biol. **43**: 250-261.

Whitehead, W.E., Palsson, O., and Jones, K.R. 2002. Systematic review of the comorbidity of irritable bowel syndrome with other disorders: What are the causes and implications? Gastroenterology **122**(4): 1140-1156.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. 1998. Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. U. S. A. **95**(12): 6578-6583.

Zoetendal, E., Rajilić-Stojanović, M., and De Vos, W. 2008. High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. Gut **57**(11): 1605-1615.

Appendix A: Supplementary Table 1. All samples sequenced with number of sequences contributed to analysis. Samples in grey were excluded due to a low number of sequences and samples in dark grey were negative controls.

Lane	Barcode	Sample	Pooled	Template Concentration	Barcode	Sequences Contributed
1	V4_1R	S1	Yes	high	ATCACG	29907
1	V4_2R	S1	Yes	high	CGATGT	40939
1	V4_3R	S1	Yes	high	TTAGGC	40012
1	V4_4R	S1	Yes	high	TGACCA	24218
1	V4_5R	S1	Yes	high	ACAGTG	43532
1	V4_6R	S1	Yes	low	GCCAAT	16085
1	V4_7R	S1	Yes	low	CAGATC	222
1	V4_8R	S1	Yes	low	ACTTGA	15911
1	V4_9R	S1	Yes	low	GATCAG	36273
1	V4_10R	S1	Yes	low	TAGCTT	15054
1	V4_11R	6TD	Yes	high	GGCTAC	41985
1	V4_12R	6TD	Yes	high	CTTGTA	77059
1	V4_13R	6TD	Yes	high	AGTACG	70548
1	V4_14R	6TD	Yes	high	TCAGTC	50318
1	V4_15R	6TD	Yes	high	TTGAGC	97766
1	V4_16R	6TD	Yes	low	AAGCGA	15909
1	V4_17R	6TD	Yes	low	TCCTCA	15161
1	V4_18R	6TD	Yes	low	GGTTGT	12052
1	V4_19R	6TD	Yes	low	TGAGGT	23508
1	V4_20R	6TD	Yes	low	TACCGT	24562
1	V4_21R	10AS	Yes	high	CCAACT	71393
1	V4_22R	10AS	Yes	high	AGAGAG	42154
1	V4_23R	10AS	Yes	high	CACTTG	69607
1	V4_24R	10AS	Yes	high	TCAAGG	63817
1	V4_25R	10AS	Yes	high	AGTGGT	54384
1	V4_26R	10AS	Yes	low	GACACT	13211
1	V4_27R	10AS	Yes	low	CCTTCT	27644
1	V4_28R	10AS	Yes	low	GGATAA	20895
1	V4_29R	10AS	Yes	low	CCTTAA	14797
1	V4_30R	10AS	Yes	low	CAAGAA	12505
1	V4_31R	S3	Yes	high	GTTGAA	58081
1	V4_32R	S3	Yes	high	TCACAA	58995
1	V4_33R	S3	Yes	high	AGTCAA	54104
1	V4_34R	S3	Yes	high	CGAATA	70280
1	V4_35R	S3	Yes	high	GCTATA	60227
1	V4_36R	S3	Yes	low	GAGTTA	17204
1	V4_37R	S3	Yes	low	TTGGTA	17907
1	V4_38R	S3	Yes	low	AACGTA	29850
1	V4_39R	S3	Yes	low	GTACTA	16578

1	V4_40R	S3	Yes	low	CATCTA	257
1	V4_41R	S1	No	high	TGTAGA	56188
1	V4_42R	S1	No	high	ATCAGA	48967
1	V4_43R	S1	No	high	ACATGA	44309
1	V4_44R	S1	No	high	TAGACA	42964
1	V4_45R	S1	No	high	GAGAAT	3699
1	V4_46R	S1	No	low	CTCAAT	2279
1	V4_47R	S1	No	low	AGGTAT	9222
1	V4_48R	S1	No	low	TTGCAT	1835
1	V4_49R	S1	No	low	TGGATT	36714
1	V4_50R	S1	No	low	ACCATT	5904
1	V4_51R	6TD	No	high	CTAGTT	39474
1	V4_52R	6TD	No	high	AGTGTT	277
1	V4_53R	6TD	No	high	TCTCTT	46185
1	V4_54R	6TD	No	high	GTAAGT	137519
1	V4_55R	6TD	No	high	CAATGT	34483
1	V4_56R	6TD	No	low	ATTCGT	26978
1	V4_57R	6TD	No	low	ATGACT	14975
1	V4_58R	6TD	No	low	ACTTCT	43590
1	V4_59R	6TD	No	low	CATAAG	4315
1	V4_60R	6TD	No	low	TTCTAG	20219
1	V4_61R	10AS	No	high	AAGATG	73682
1	V4_62R	10AS	No	high	TATGTG	52649
1	V4_63R	10AS	No	high	AATTGG	27085
1	V4_64R	10AS	No	high	TAATCG	507
1	V4_65R	10AS	No	high	ACTAAC	80471
1	V4_66R	10AS	No	low	TGTTAC	4272
1	V4_67R	10AS	No	low	ATACAC	32074
1	V4_68R	10AS	No	low	CTTATC	45677
1	V4_69R	10AS	No	low	AGATTC	3006
1	V4_70R	10AS	No	low	ACGGAA	30973
1	V4_71R	S3	No	high	TGCGAA	147
1	V4_72R	S3	No	high	GACCAA	49796
1	V4_73R	S3	No	high	CTGTCA	27838
1	V4_74R	S3	No	high	GCAGAT	27438
1	V4_75R	S3	No	high	TCGTGT	42015
1	V4_76R	S3	No	low	GAACCT	26648
1	V4_77R	S3	No	low	GTCATG	85676
1	V4_78R	S3	No	low	GATAGC	4595
1	V4_79R	S3	No	low	AAGTCC	31185
1	V4_80R	S3	No	low	ATTGCC	6566
1	V4_81R	SX	yes	high	CCGAGA	53661
1	V4_82R	SX	yes	high	CGCTGA	143162
1	V4_83R	SX	yes	high	GGCACA	52548
1	V4_84R	SX	yes	low	CGTGCA	27642

1	V4_85R	SX	yes	low	GGCCTT	27417
1	V4_86R	SX	yes	low	CCTGGT	20185
1	V4_87R	SX	yes	high	CAGGCT	41885
1	V4_88R	SX	yes	high	GTCGCT	53682
1	V4_89R	SX	yes	high	GCGTAG	35433
1	V4_90R	SX	yes	low	CTGGAG	37627
1	V4_91R	SX	yes	low	CTACGG	35151
1	V4_92R	SX	yes	low	ACACCG	29416
1	V4_93R	negcont	n/a	n/a	GTTCCG	241
1	V4_94R	negcont	n/a	n/a	CAGCAC	642
1	V4_95R	negcont	n/a	n/a	CCGTTC	464
1	V4_96R	negcont	n/a	n/a	GCATCC	371
2	V4_1R	S1	Yes	high	ATCACG	23312
2	V4_2R	S1	Yes	high	CGATGT	30580
2	V4_3R	S1	Yes	high	TTAGGC	0
2	V4_4R	S1	Yes	high	TGACCA	28383
2	V4_5R	S1	Yes	high	ACAGTG	31086
2	V4_6R	S1	Yes	low	GCCAAT	19715
2	V4_7R	S1	Yes	low	CAGATC	24126
2	V4_8R	S1	Yes	low	ACTTGA	19623
2	V4_9R	S1	Yes	low	GATCAG	20025
2	V4_10R	S1	Yes	low	TAGCTT	16685
2	V4_11R	6TD	Yes	high	GGCTAC	31789
2	V4_12R	6TD	Yes	high	CTTGTA	37765
2	V4_13R	6TD	Yes	high	AGTACG	41893
2	V4_14R	6TD	Yes	high	TCAGTC	34343
2	V4_15R	6TD	Yes	high	TTGAGC	46193
2	V4_16R	6TD	Yes	low	AAGCGA	13917
2	V4_17R	6TD	Yes	low	TCCTCA	18927
2	V4_18R	6TD	Yes	low	GGTTGT	8183
2	V4_19R	6TD	Yes	low	TGAGGT	16947
2	V4_20R	6TD	Yes	low	TACCGT	9733
2	V4_21R	10AS	Yes	high	CCAACT	44346
2	V4_22R	10AS	Yes	high	AGAGAG	36900
2	V4_23R	10AS	Yes	high	CACTTG	44436
2	V4_24R	10AS	Yes	high	TCAAGG	43535
2	V4_25R	10AS	Yes	high	AGTGGT	50034
2	V4_26R	10AS	Yes	low	GACACT	14967
2	V4_27R	10AS	Yes	low	CCTTCT	15885
2	V4_28R	10AS	Yes	low	GGATAA	22784
2	V4_29R	10AS	Yes	low	CCTTAA	14842
2	V4_30R	10AS	Yes	low	CAAGAA	10201
2	V4_31R	S3	Yes	high	GTTGAA	42576
2	V4_32R	S3	Yes	high	TCACAA	38499
2	V4_33R	S3	Yes	high	AGTCAA	43807

2	V4_34R	S3	Yes	high	CGAATA	51703
2	V4_35R	S3	Yes	high	GCTATA	57034
2	V4_36R	S3	Yes	low	GAGTTA	16813
2	V4_37R	S3	Yes	low	TTGGTA	20506
2	V4_38R	S3	Yes	low	AACGTA	20498
2	V4_39R	S3	Yes	low	GTACTA	12793
2	V4_40R	S3	Yes	low	CATCTA	458
2	V4_41R	S1	No	high	TGTAGA	33518
2	V4_42R	S1	No	high	ATCAGA	30673
2	V4_43R	S1	No	high	ACATGA	16622
2	V4_44R	S1	No	high	TAGACA	26813
2	V4_45R	S1	No	high	GAGAAT	23804
2	V4_46R	S1	No	low	CTCAAT	4158
2	V4_47R	S1	No	low	AGGTAT	8406
2	V4_48R	S1	No	low	TTGCAT	21612
2	V4_49R	S1	No	low	TGGATT	29469
2	V4_50R	S1	No	low	ACCATT	25817
2	V4_51R	6TD	No	high	CTAGTT	29569
2	V4_52R	6TD	No	high	AGTGTT	22184
2	V4_53R	6TD	No	high	TCTCTT	23581
2	V4_54R	6TD	No	high	GTAAGT	57618
2	V4_55R	6TD	No	high	CAATGT	24669
2	V4_56R	6TD	No	low	ATTCGT	14641
2	V4_57R	6TD	No	low	ATGACT	18141
2	V4_58R	6TD	No	low	ACTTCT	30553
2	V4_59R	6TD	No	low	CATAAG	16383
2	V4_60R	6TD	No	low	TTCTAG	15738
2	V4_61R	10AS	No	high	AAGATG	35874
2	V4_62R	10AS	No	high	TATGTG	32479
2	V4_63R	10AS	No	high	AATTGG	22976
2	V4_64R	10AS	No	high	TAATCG	11814
2	V4_65R	10AS	No	high	ACTAAC	46183
2	V4_66R	10AS	No	low	TGTTAC	15111
2	V4_67R	10AS	No	low	ATACAC	16895
2	V4_68R	10AS	No	low	CTTATC	19676
2	V4_69R	10AS	No	low	AGATTC	13698
2	V4_70R	10AS	No	low	ACGGAA	15970
2	V4_71R	S3	No	high	TGCGAA	23264
2	V4_72R	S3	No	high	GACCAA	33802
2	V4_73R	S3	No	high	CTGTCA	27377
2	V4_74R	S3	No	high	GCAGAT	30528
2	V4_75R	S3	No	high	TCGTGT	20118
2	V4_76R	S3	No	low	GAACCT	28758
2	V4_77R	S3	No	low	GTCATG	55074
2	V4_78R	S3	No	low	GATAGC	14274

2	V4_79R	S3	No	low	AAGTCC	16168
2	V4_80R	S3	No	low	ATTGCC	15946
2	V4_81R	SX	yes	high	CCGAGA	44568
2	V4_82R	SX	yes	high	CGCTGA	41293
2	V4_83R	SX	yes	high	GGCACA	35443
2	V4_84R	SX	yes	low	CGTGCA	16991
2	V4_85R	SX	yes	low	GGCCTT	16868
2	V4_86R	SX	yes	low	CCTGGT	12231
2	V4_87R	SX	yes	high	CAGGCT	32029
2	V4_88R	SX	yes	high	GTCGCT	35515
2	V4_89R	SX	yes	high	GCGTAG	25929
2	V4_90R	SX	yes	low	CTGGAG	24184
2	V4_91R	SX	yes	low	CTACGG	21652
2	V4_92R	SX	yes	low	ACACCG	19598
2	V4_93R	-ve ctrl	n/a	n/a	GTTCCG	248
2	V4_94R	-ve ctrl	n/a	n/a	CAGCAC	866
2	V4_95R	-ve ctrl	n/a	n/a	CCGTTC	429
2	V4_96R	-ve ctrl	n/a	n/a	GCATCC	610