EXPLORING THE BREADTH AND DEPTH OF DIVERSITY WITHIN THE CANINE GUT MICROBIOME

by

DANIEL HAND

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Biosciences
The University of Birmingham
September 2010

ABSTRACT

The mammalian gut microbiota is an essential factor in intestinal function and thus overall health. In the post genomic era, culture independent studies into the gut microbiota, particularly that of humans have allowed great leaps forward in knowledge of this ecosystem. Furthermore, recent advances in sequencing technologies have allowed acceleration and broadening of work in this research field. Despite this, the canine gut microbiome has remained relatively uncharacterised. This work investigates the faecal microbiota of a diverse multi-breed and multi-location group of 79 dogs, by amplifying and sequencing the 16S rRNA gene from these dogs using both Sanger sequencing of clone libraries and high throughput pyrosequencing. A robust census of the canine faecal microbiota was undertaken. The most abundant genera were the *Bacteroides*, *Prevotella*, *Cetobacterium*, *Fusobacterium*, *Sutterella* and *Megamonas*. A limited core microbiome was defined in 90% of the study population; this represented less than 0.5% of richness but more than 37.4% of abundance. Influences of host sex, diet and age were investigated but were found not significant. Some evidence was found for breed associated richness differences, most marked in Labrador retrievers and miniature Schnauzers. Furthermore, the microbiota of the Labradors appeared to cluster separately from the other breeds.

DEDICATION

This PhD thesis is dedicated to Harry and Ruth Abbott, proud and supportive grandparents, who always wished the best for their grandchildren.

ACKNOWLEDGMENTS

I would like to acknowledge my supervisors Charles W Penn and Corrin Wallis for giving me the chance to study for this PhD. Their advice, support and perseverance were invaluable.

I would also like to acknowledge Mathew Gibbs, Alan Martin and the WALTHAM® Centre for Pet Nutrition (WCPN) lab team for all the help they provided.

A number of colleagues at Birmingham University were especially helpful; Gemma Blackwell, Deborah Eaves, Mala Patel, Chrystala Constantinidou, Jon Hobman, Antony Jones, Lorraine Wallace, Jonathan James, Mohammad Ahasan, Olga Timoshenko and Al-Ali Abdulaziz.

I would like to acknowledge the inspirational lectures delivered by my undergraduate lecturers Charles W Penn, Julia Lodge, Clive Sweet and in particular Jon Hobman that ignited my interest in microbiology.

I would like to acknowledge my father, Geoff Hand and my friend Andrew Hall for the support they provided during the last year of my PhD study.

This project would not have been possible without funding from BBSRC andWCPN.

NON- AUTHOR CONTRIBUTIONS

Chapters 5 and 6 presented as academic papers were written by Daniel Hand, Charles W Penn and Corrin Wallis. The Introductions and discussions were mostly written by Charles W Penn, the material and methods by Daniel Hand, the results were written by Daniel Hand and Charles W Penn. Some of the statistical analyses described in Chapter 6 were performed by George Gettinby and Alison Colyer.

In the practical work, stools from non WCPN animals were collected by Wendy Brown, University of New England, Australia and Corrin Wallis, WCPN. The DNA extractions from all dogs except the miniature Schnauzer were performed by lab workers at WCPN, as was clone library construction for the Papillon, cocker spaniel and Labrador retriever breeds.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF DEFINITIONS AND/OR ABBREVIATIONS

16S rRNA: RNA sequence of the small subunit of the bacterial ribosome

Read: An individual nucleotide sequence

Sanger sequencing: A sequencing method developed by Frederick Sanger using Dideoxynucleotide chain termination

Hit: A single database search result. In the context of Bioinformatics the 'hit' is usually a nucleotide/protein sequence

V1-V9: Denote the nine variable regions within the 16S ribosomal RNA molecule

OTU: Operational taxonomic unit. Units are usually defined using nucleotide identity

WCPN: WALTHAM® Centre for Pet Nutrition

CHAPTER ONE: GENERAL INTRODUCTION

The mammalian gut microbiota is an essential factor in intestinal function and thus overall health. This complex community of bacteria is comprised of thousands of species in numbers up to 10^{13} bacteria per wet gram faeces (Eckburg, Bik et al. 2005). The relevance of the gut microbiota to the health of hosts has been increasing realised over the past 30 years. While much of the gut microbiota is non-harmful (commensal) the synthesis of essential vitamins, enhancement of immunity and fermentation of indigestible carbohydrates has been associated with gut bacteria. Germ-free animal studies have shown that an absence of microbiota is associated with poor immune system development and increased nutrient requirement (Gordon 1971) . Conversely gnotobiotic studies have provided evidence of the mutualistic relationships certain species have with the host. Various species from the *Bifidobacterium* and *Lactobacillus* genera have been identified as mutualists in humans and have been used in probiotics for a number of years.

Some investigators have refered to the intestinal microbiota as an organ (Backhed, Ley et al. 2005; O'Hara and Shanahan 2006). The metabolic activity of the overall microbiota is comparable to that of a mammalian organ, the 'virtual organ' interacting with the rest of the body via the intestinal mucosa. However while the importance of gut microbiota has gained increased recognition, our knowledge of it remains limited.

1.1 Characterising the gut microbiota

The gut microbiota has only recently been characterised at genus level in humans and to a lesser extent in canines and other mammals (Rastall 2004). The studies of the past 50 years

that provided this basic insight into intestinal microbiology were limited by the inherent disadvantages of culture methods. The culture process generally involves preparing an inoculum from a sample of intestinal content which is used to inoculate culture medium. Resulting colonies would be analysed by biochemical, morphological and PCR tests. Different culture mediums are required to grow specific groups of organisms and thus all culture mediums are inherently biased, selecting for the survival of organisms with a certain complement of genes. This selection means that inferences about the overall bacterial community composition from culture based studies can be unsound. Advances in knowledge were coupled to the identification of novel culture methodologies which allowed a marginal increase in ability to characterise the microbiota. Despite these efforts to culture novel gut bacteria the majority remain unculturable or uncultured due to fastidious growth requirements, interdependence on other microbiota components and stresses imposed by the culture process (Green and Keller 2006) .

Culture-independent methodologies for the identification of microorganisms have been developing over the past several decades. Influential work by Carl Woese in molecular phylogeny and the invention of the polymerase chain reaction (PCR) in the late 1980's provided the base for such studies (Woese 1987). Specifically Woese analysed the sequence of the ribosomal small sub unit RNA which is conserved across all kingdoms of life due to its essential role in protein synthesis. Structurally important regions within the molecule are highly conserved due to the lethality of mutants and the lack of silent mutations that can occur in protein coding genes. However, some regions within the molecule are relatively less important structurally and slowly accrue mutations at a phylogenetically useful rate. The sequence thus provides one of the most convenient if not ideal phylogenetic delineators and

together with DNA-DNA reassociation values forms the basis of modern bacterial species classification (Rossello-Mora and Amann 2001; Konstantinidis and Tiedje 2005). Small subunit ribosomal sequence can be referred to as 16S rRNAthe '16S' being derived from centrifugal nomenclature. 'rDNA' is a shorthand term for the DNA sequence that encodes the rRNA. Alternatively the term SSU (small subunit) rRNA may be used. A representation of the secondary structure and 'conserved' and 'variable' regions within the 16S rRNA molecule is shown in Figure 1 and Figure 2.

Figure 1. A model of the secondary structure of Escherichia coli *16S rRNA adapted from (Greetham 2003). Variable regions are highlighted yellow, universally conserved regions are shown in red.*

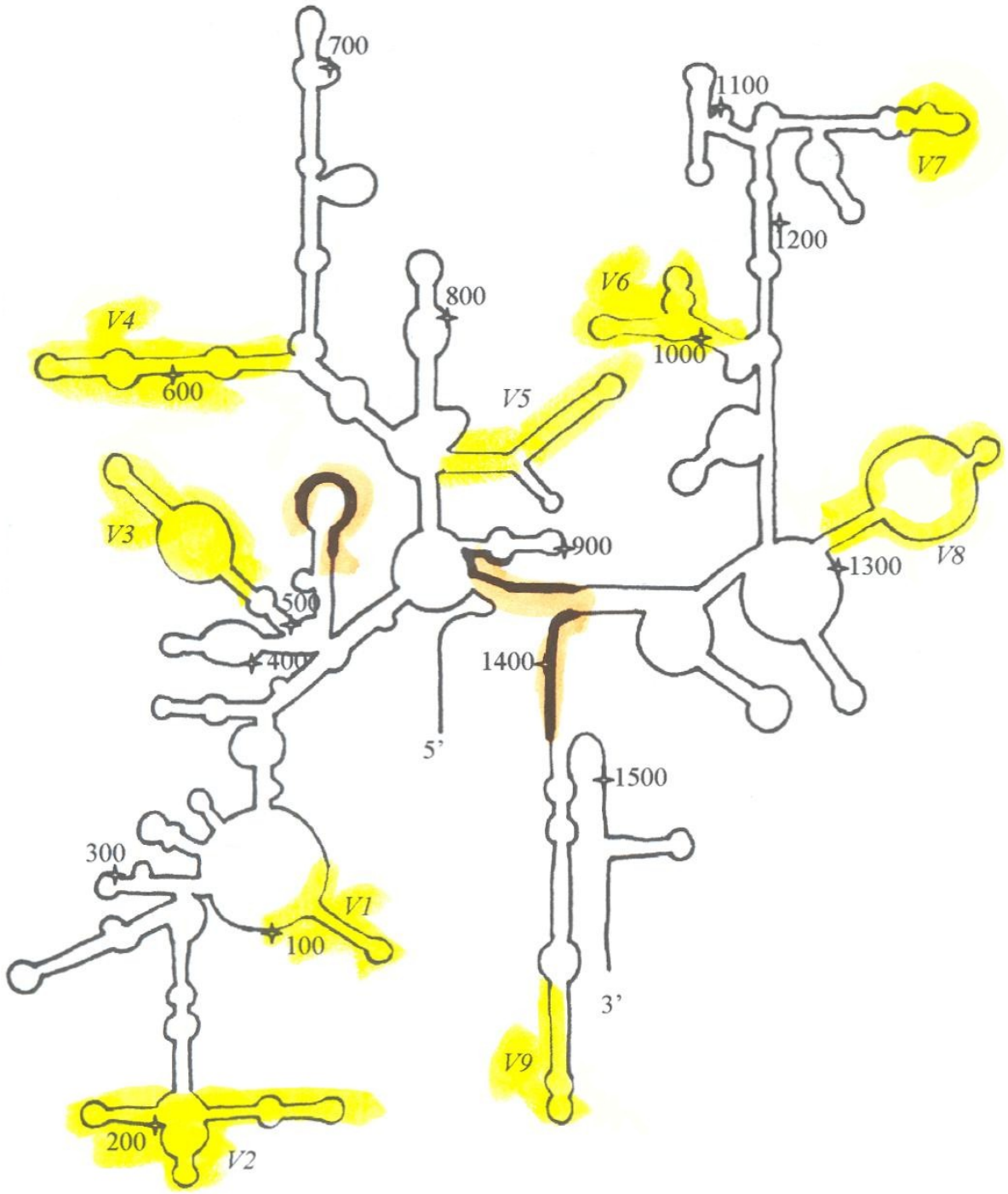
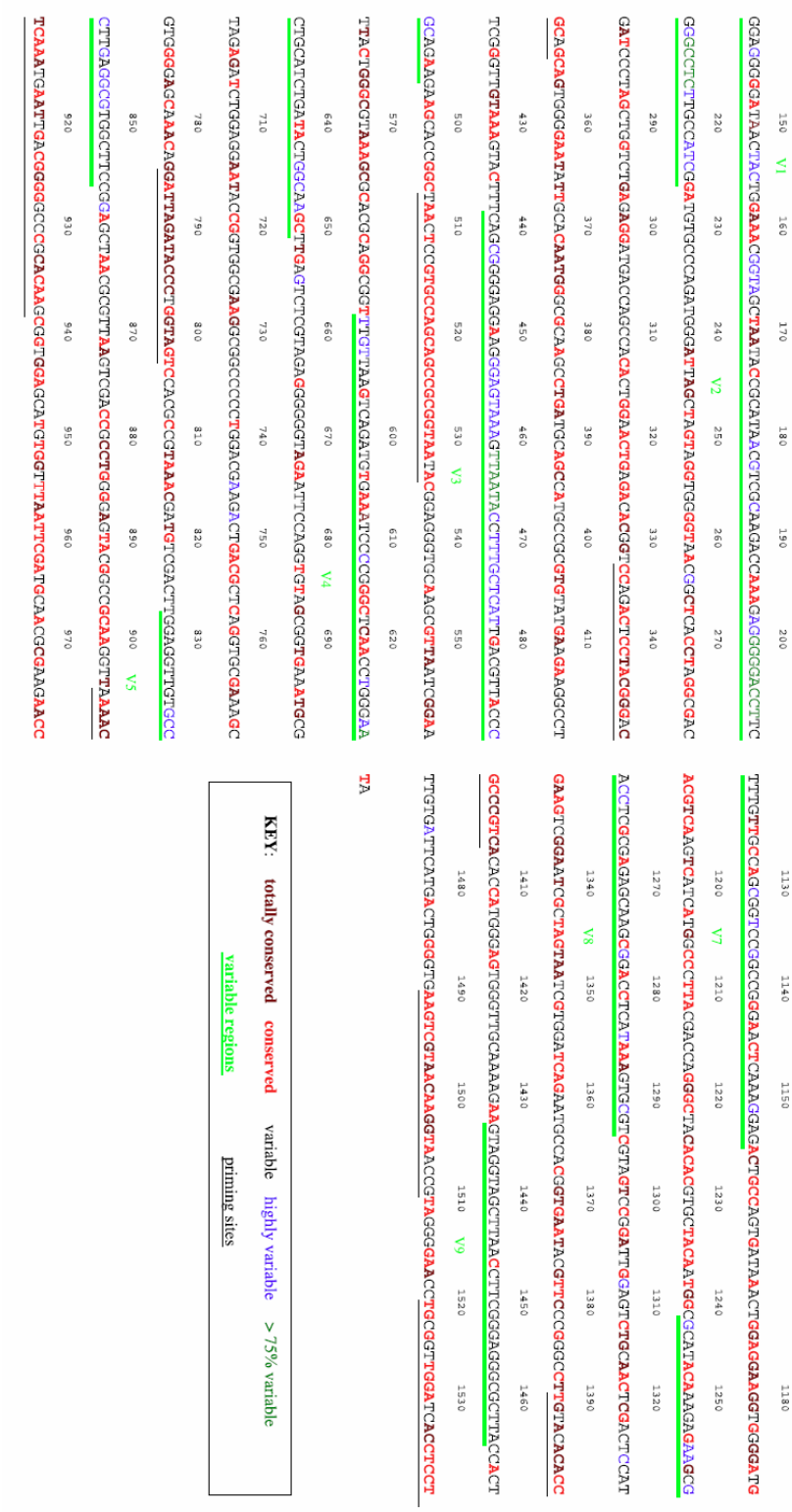


Figure 2. Escherichia coli *16S rRNA gene sequence annotated with a bacterial nucleotide variability map adapted with permission from Baker et al (Van de Peer, Chapelle et al. 1996; Baker, Smith et al. 2003). Variable regions and common priming sites are underlined, while nucleotide variability is denoted by letter colour.*



Analysis of bacterial phylogeny using 16S rRNA gene sequence requires its amplification by PCR using oligonucleotide primers specific to regions conserved throughout specific groups of bacteria that flank variable region(s). Analysis of the resulting amplicons provides microbial ecologists with an estimate of bacterial diversity in the original PCR template (eg. faecal material/bacterial colony). A number of different methodologies have been developed to assaying these 16S amplicons depending on the required information.

An unprecedented emergence of 16S rRNA gene sequence data has allowed microbiologists to understand the substantially biased view of microbial ecosystems that traditional culture-based bacteriology has supplied us. The ribosomal database project (RDP) is a open access annotated 16S rRNA gene sequence database and provides an invaluable tool for microbial ecologists (Cole, Chai et al. 2006; Cole, Wang et al. 2009). The database is rapidly growing, as of August 2010 the dataset contained 1,418,497 SSU rDNA sequences more than ten times that of August 2004 (101,632 sequences)(Cole, Chai et al. 2005). A large proportion of the database sequences are derived from uncultivable strains, which highlights the utility of 16S rRNA gene in culture-independent studies. The size of this database and others will increase as more environmental SSU rRNA gene sequencing projects are completed and will allow investigators an ever increasing toolset for the study of microbial ecology.

Other molecular delineators are available based on protein coding sequences (Santos and Ochman 2004; Ciccarelli, Doerks et al. 2006). Comparative genomic analysis of whole genome sequenced organisms has revealed numerous universally conserved genes. These markers can prove useful in characterising evolutionary/phylogenetic relationships between closely related bacterial strains, which is a limitation of 16S rRNA gene approaches. For

example: One study has used a chaperonin-60 (a protein found throught prokaryotes and eukaryotes) library to characterise pig ileum microbiota (Hill, Hemmingsen et al. 2004). The investigators found that because chaperonin-60 sequence is protein coding, chimera sequences were more readily detectable and a higher phylogenetic resolution was provided than with 16S sequence due to silent mutations. Conversely, this also means that it is harder to design oligonucleotide primers specific to universally conserved regions than in 16S. Studies using protein coding markers are currently limited and do not yet provide microbiologists with a practical tool for the study of microbial ecology. This is in part due to an absence of large phylogenetic databases of marker sequence. Metagenomics has demonstrated the greatest potential for the use of markers other than the 16S rRNA gene. One metagenomics study searched for various markers in environmental metagenomic study datasets (von Mering, Hugenholtz et al. 2007). They searched the shotgun sequence for previously identified markers (Ciccarelli, Doerks et al. 2006). The marker sequence was then aligned to a set of reference markers and taxonomic designations were issued based on mapping to a phylogeny of completely sequenced organisms. The results of the study revealed interesting evolutionary trends; they also showed that the use of protein markers could provide better quantitative measurements of bacteria within the environment.

The use of 16S rRNA gene based methodologies to study gut microbiota has become well established. Initial studies focused on ‘fingerprinting’ the microbiota, as these techniques required less time and resources than the sequencing of SSU rDNA clone libraries. Microbiota fingerprinting techniques such as denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) are well established (Gerard and Kornelia 1998; Zoetendal, Akkermans et al. 1998; Simpson, Martineau et al. 2002). Findings

from microbiota fingerprinting studies have been revealing compared to previous culture work. Several studies in varying mammalian hosts including canines and humans have indicated that microbiota composition is host specific and stable over time (Zoetendal, Akkermans et al. 1998; Simpson, Martineau et al. 2002).

Additionally a DGGE study showed that the microbiota profiles between human mononzygotic twins were significantly similar compared to the profiles of unrelated controls (Zoetendal 2001). The results indicate that host genotype may be a significant factor in the composition of the microbiota.

These profiling technologies while allowing the monitoring of general shifts in microbiota between individuals and conditions do not readily allow the identification of all microbiota components in the fingerprint. The high level of complexity present in intestinal ecosystems inhibits the high throughput identification of bacterial species within a profile. To identify bacteria within the profile, individual DNA bands within the agarose gel are removed, purified and sequenced. The investigator will usually focus on the analysis of DNA bands which differ between two microbiota fingerprints. Confounding band patterns may result from co-migration of DNA fragments and intraspecies rrn operon (the operon containing the 16S rRNA gene) heterogeneity (Gerard and Kornelia 1998). Additionally the techniques do not detect species contributing less than 1% of the microbiota, limiting the value of the approach for the detection of subtle but potentially important changes. This limitation is of significant importance considering the majority of the microbial diversity may be present in the minority of total intestinal microbiota.

The creation and sequencing of 16S rRNA gene clone libraries has proved a more informative method of studying the microbiota. One of the most important studies of this type in humans was the work of Eckburg (Eckburg, Bik et al. 2005). Eckburg's study was of unprecedented scale at the time and established the analysis of 16S rRNA gene sequence as the preferred method for the characterization of microbiota. The study reinforced the findings of DGGE profiling experiments described earlier, that the microbiota is relatively stable over time. However the sequencing of 16S rRNA gene inserts from library clones is resource intensive, requiring automation and time. The quantification of microbial diversity provided by this approach is prone to bias resulting from the DNA extraction, PCR amplification and cloning. Due to this bias a large number of 16S rRNA gene clones need to be sequenced to ensure coverage of minority microbiota components. The resources requirements limited the usefulness of the technique for frequent and practical experimentation compared to DGGE/TGGE profile technologies.

In the post genomic era the availability of microarray technology has provided microbial ecologists with yet an additional tool. While 16S rRNA gene library sequencing was preferred for the characterisation of novel environments, high-density probe microarrays offer increased rapidity and reproducibility. 16S rRNA gene array technology is based on a modification of the comparative genomic hybridization (CGH) concept. A mixed community DNA pool is extracted from a sample and followed by 16S rRNA gene PCR amplification. After the PCR a mixed population of SSU rDNA amplicons is labelled and then hybridised to oligonucleotide probes on the microarray, washed and scanned to detect signals. The probes are designed to be highly specific to individual species 16S rRNA gene sequence. Thus fluorescence at a

particular probe locus indicates the presence of a particular bacterial species in the original sample.

Correct design of the oligonucleotide probe complement of the microarray is of paramount importance. 16S rRNA gene library sequencing projects provide a good source of data for probe design and extensive sequence databases have been built up from their output. Designing a set of probes to assay bacterial species diversity from 16S DNA sequence data is challenging (Peplies, Glockner et al. 2003). Ensuring the hybridisation specificity of probes is vital to ensure accurate results. Validation of probes designed to detect unculturable bacteria is difficult. Additionally the quantification of species present in the microbiota is also difficult due to a number of complex factors including PCR amplification bias and interspecies rrn operon copy number heterogeneity (v. Wintzingerode, Gobel et al. 1997).

1.2 High throughput sequencing technologies

Recent novel advances in sequencing technology have provided the possibility of rapid, massively parallel sequencing directly from complex DNA template. The technology has replaced many microarray based applications. Several principle approaches have currently been developed; however, the general principals of both technologies are similar. Firstly, complex DNA mixtures are separated physically on a solid phase surface and clonally amplified by PCR. Clonal PCR products are referred to as 'molecular colonies' and are in essence equivalent to colonies of bacterial insert libraries used in plasmid-based sequencing projects. Secondly novel 'sequencing by synthesis' chemistries are performed *in situ* on the molecular colonies, whereby a sensor detects massively parallel base calls from the molecular colonies. Complex computation, storage and bioinformatics capabilities are required to

analyse the resulting information output which is in excess of all but the largest traditional sequencing projects. These technologies offer hitherto unachievable levels of rapid sequence output due to the absence of conventional library clone processing.

One of the most mature technologies available was developed by 454 Life Sciences Corporation (Margulies, Egholm et al. 2005), now acquired by Roche. The technology uses a bead based system of DNA separation and PCR amplification. Specifically, DNA is fragmented, ligated to two different linkers and the single-stranded fragments with linkers, attached to beads in titrations that favour one fragment per bead. Emulsion-based PCR is performed in bead-oil droplets resulting in 10 million clonal copies of the original single template molecule per bead. Molecular colony beads are deposited onto a fibre-optic well plate, one bead per well. A pyrosequencing reaction isperformed across the plate using repeated cycles of sequential dATP, dTTP, dCTP and dGTP extension; nucleotide incorporation of one or more complementary bases, depending on the template sequence, is recorded via enzymatic luminometric inorganic pyrophosphate (PPi) detection (Ronaghi, Karamohamed et al. 1996). One disadvantage of 454 sequencing technology is that homopolymeric runs in sequences provide a substantial challenge due to the nature of pyrosequencing. The technical limit of individual sequence reads is improving and currently stands at 500 bp. Although this is much less than the 700 bp provided by a traditional dideoxynucleotide chain termination sequencing run, the massively parallel nature of the reads combined with powerful computing allows reasonable assembly.
Using this approach a 96% coverage de novo assembly of the *Mycoplasma genitalium* genome was achieved (Margulies, Egholm et al. 2005). Impressively, the assembly was

constructed from a dataset of 25 million base reads of shotgun sequence collected in one four hour run. Since this paper hundreds of studies have used 454 pyrosequencing technology.

The second technology currently available is provided by the Illumina Corporation called the Genome Analyser (GA) originally developed by the UK company Solexa. Briefly, the technology creates molecular colonies in a flow cell by applying DNA fragments with attached linkers to a lawn of the complementary linkers on a glass slide surface. Solid phase 'bridge-based' PCR amplification is performed to 'grow' the colonies in the flow cell. Sequencing by synthesis is performed base by base using fluorescently labelled nucleotides with reversible termination properties. This sequencing chemistry means GA technology, unlike 454 technology, can handle long homopolymeric runs without the lost of accuracy. The sequence output in terms of read number is also higher than 454; however sequence reads are currently only 75bp long, thus a greater fold coverage is required for assembly of sequence. The short sequence length compared to 454 pyrosequencing limits its use in sequencing amplicons such as 16S rRNA gene. Consequently, 454 pyrosequencing is used for most amplicon sequencing applications.

Both of the technologies mentioned have already revolutionised the microbial ecology. One of the first studies to use these technologies directly sequenced 16S rDNA gene amplicon pools extracted from a marine ecosystem and achieved sequence coverage far higher than possible in bacterial 16S library clone projects (Sogin, Morrison et al. 2006). Another investigation used a metagenomics approach to extract 16S sequence and study a iron mine ecosystem (Edwards, Rodriguez-Brito et al. 2006). Turnbaugh used a similar approach to characterised human gut microbiota and links with host obesity(Turnbaugh, Ley et al. 2006). Since these

studies the use of high throughput sequencing technologies has become commonplace (Dethlefsen, Huse et al. 2008; Claesson, O'Sullivan et al. 2009; Turnbaugh, Hamady et al. 2009). Analysis of the datasets produced by these projects is challenging and therefore bioinformatics developments are accompanying these studies (Huson, Auch et al. 2007; Cole, Wang et al. 2009; Quince, Lanzén et al. 2009; Schloss, Westcott et al. 2009).

A number of different tools are now available to microbiologists for the study of microbiota. Studies using these tools are fuelling a rapid increase in our knowledge.

1.3 Microbiology of the gut

1.3.1 Commensals, mutualists and parasites

Within the ecosystem of the gut a number of possible relationships and interactions exist between host and bacteria. In traditional ecology terms commensals provide no or neutral benefit upon the host, mutualists provide a positive benefit and parasites a negative effect. Symbiosis defines close and long-term interactions with the host in a commensal, mutualist or parasitic state. The terminology used to define these relationships in the gut is not universally agreed on. It is common for investigators to use the terms symbiotic and commensal to define positive relationships (mutualism). This view is fostered by most symbiotic relationships in the gut being of a commensal or mutualist nature. Conversely, many parasitic relationships are pathogenic and transitory. It is thought that the balance between and development of these states within the mammalian gut microbiome is linked to the co-evolution of mammals with their microbiota (Steinert, Hentschel et al. 2000; Hooper and Gordon 2001; Ley, Hamady et al. 2008).

The study of pathogenic bacteria has and still dominates gut microbiology research. Prominent human gut pathogens such as certain *Escherichia coli* O157:H7 negatively impact their host during infection but it is debatable that this temporary interaction can be defined as a parasitic relationship. Chronic infection by pathogens such as *Clostridium difficile* may be more appropriately considered parasitic.

It seems that the majority of gut bacteria have a commensal relationship with the host. It is clear from germ-free animal studies that an absence of the microbiota is associated with an under developed immune systems and increased nutrition requirements (Gordon 1971). While the host has to be able to quickly react to infection it must also tolerate the commensal and mutualist community that check the growth of pathogens. The immune system employs different mechanisms for control of these groups such as T-cell independent induction of secretory immunoglobulin A (sIgA) in response to commensal antigens (Macpherson, Gatto et al. 2000).

The hygiene hypothesis proposed by Strachan suggests that a lack of exposure to microbiota antigens during early life leads to an imbalance in T-cell development and an increase in allergy in the host (Strachan 1989). Although, extensive evidence of such an still effect remains to be observed (Okada, Kuhn et al. 2010). There has been considerable effort research investigating the relationship of the microbiota with Irritable bowel disease (IBD).

In relation to total number of bacteria within the gut, only a minority are known to be overtly mutualistic. Certain members of the microflora such as some *Bifidobacterium* and

Lactobacillus strains have been traditionally identified as beneficial to humans (Gibson and Roberfroid 1995; Madsen, Doyle et al. 1999; Gionchetti, Rizzello et al. 2000). *Bacteroides thetaiotaomicron* is a well studied mutualist that benefits the host by degrading indigestible dietary polysaccharides (Backhed, Ley et al. 2005). Non-harmful bacteria that attach and colonise the epithelium, prevent pathogenic bacteria from doing so, in return the host provides a nutrient rich habitat and a refuge from the peristaltic movements of the lumen. This form of protection is a mutalistic relationship between host and bacteria.

These relationships are not static, they can change depending on the environment within the gut (Hooper and Gordon 2001). Commensal bacteria can become opportunist pathogens under the correct conditions such as antibiotic treatment and in immunocompromised hosts. The type of substrates fermented by bacterial may influence their role in producing short chain fatty acids (SCFA) or deleterious phenolic compounds (Russell, Gratz et al. 2011).

1.3.2 The canine gut microbiota

Comparatively little work has been done to characterize the canine gut microbiota compared to that of humans. A number of culture based studies have been conducted profiling the canine gut microbiota (Balish, Cleven et al. 1977; Greetham, Giffard et al. 2002; Greetham 2003). While, These studies provide little value in the broad profiling of this ecosystem due to the flaws of cultured based approaches discussed in section 1.1, the bacteria discovered were at least present in a larger uncharacterized microbiome. The significance of these cultured bacteria within the microbiome is unclear.

Balish cultured and identified 45 different genera from the nose, throat and faces of isolated and non-isolated Beagle dogs (Balish, Cleven et al. 1977). From faeces Balish isolated facultative anaerobes from the following genera: *Streptococcus*, *Staphylococcus*, *Corynebacterium*, *Lactobacillus*, *Klebseilla*, *Escherichia*, *Enterobacter*, *Citrobacter* and *Proteus*. Balish also cultured anaerobic isolates from the following genera: *Clostridium*, *Lactobacillus*, *Bifidobacterium*, *Eubacterium*, *Bacteroides*, *Propionibacterium*, *Fusobacterium*, *Butyrivibrio*, *Peptococcus*, *Ruminococcus* and *Veillonella.* Within the *lactobacillus* isolates 15 species were identified biochemically including *L. acidophilus*, *L. casi* and *L. brevis*. Balish observed that isolated dogs had an increased observed bacterial diversity. It was suggested that this may of been due to disruption of the dominate microbiota leading to an increase in previously repressed groups. Furthermore, this might indicate that an important defence mechanism, maintenance of a 'normal' microbiota, might be disrupted leading to the growth of potentially harmful bacteria.

Greetham used molecular methods (16S rRNA gene sequencing) to identify 157 faecal isolates from Labrador dogs using a variety of selective media. Many of the genera frequently isolated were those that Balish observed. Particularly frequent isolates were as follows; *E. coli*, *Streptococcus bovis*, *Lactobacillus ruminus* and *Staphylococcus* sp. Greetham acknowledged the flaws of cultured based enumeration of the microbiota through the use of selective media.

Studies using culture independent approaches profiling the small intestine of canines have been published (Suchodolski, Camacho et al. 2008; Xenoulis, Palculict et al. 2008; Suchodolski, Dowd et al. 2009; Suchodolski, Xenoulis et al. 2009; Suchodolski, Xenoulis et

al. 2010). One study characterized the microbial profile of different sub-anatomical locations: duodenum, jejunum, ileum, and colon (Suchodolski, Camacho et al. 2008). Four phyla; *Firmicutes*, *Proteobacteria*, *Fusobacteria*, and *Bacteroidetes* were predominantly observed. The study showed that members of the *Fusobacteriales* and *Bacteroidales* were abundant in the ileum and colon, while the *Clostridiales* were abundant at all sampled sites. The *Fusobacteria* were observed in high abundance relative to other mammalian gut microbiota communities.

Xenoulis investigated the association of the microbiota with inflammatory bowel disease (IBD) by sampling the duodenum of 10 IBD and 9 control dogs (Xenoulis, Palculict et al. 2008). The microbial communities of IBD dogs were observed to be altered from that of control dogs and enriched with *Enterobacteriaceae*. However, a follow up study comparing these results with a study of mucosal-associated duodenal microbiota and IBD found no such depletion and a number of other differences in IBD associated taxa (Suchodolski, Xenoulis et al. 2010). Suchodolski suggested that was due to the different sampling methods (cytology brushes vs. endoscopic biopsies) or differences in the sample population.

Two high-throughput sequencing studies of the canine gut microflora have been done, albeit with relatively small study populations and read numbers (Suchodolski, Dowd et al. 2009; Middelbos, Vester Boler et al. 2010). Following work into the effects of antibiotics on the human gut microbiota Suchodolski investigated the effect of tylosin on the small intestine microflora of 5 dogs (Dethlefsen, Huse et al. 2008; Suchodolski, Dowd et al. 2009). The results indicated that tylosin treatment had produced significant changes in the microbiota but the study was limited by the short-term 14 day post treatment sampling. Middelbos explored

the effect of dietary fibre on the faecal microbiota of 6 dogs. Changes in dietary fibre were detectable in the study dogs microbiota and *Bacteroides* and *Fusobacteria* were observed as the co-dominate phyla.

1.4 PhD project description

Knowledge of canine intestinal microbiota was limited at the start of the project. Several small studies investigating canine microbiota using culture-independent fingerprinting methods had been published (Simpson, Martineau et al. 2002; Vanhoutte, Huys et al. 2005). However the canine gut microbiota had yet to be systematically characterised at the species level. Much of our knowledge dated back to the 1970's and was of questionable value (Balish, Cleven et al. 1977).

The aim of the project was to conduct a thorough culture-independent investigation of the canine gut microbiota. Ultimately it was envisioned that a high throughput tool would be used for the rapid screening of canine gut microbiota. To do this, a large and diverse population of dogs was sampled from multiple worldwide locations and dog breeds (Table 1). 16S rRNA gene sequence was then amplified and analysed from the extracted faecal DNA of this population.

One envisioned experimental approach was to develop a high density oligonucleotide microarray. In order to design such a tool a substantial census of the canine microbiome would have to be conducted. To this end, design of oligonucleotide probes would be facilitated by access to the results of 16S rRNA gene library sequencing conducted in the first

stages of the project. The library would also provide one of the first large scale 16S rRNA gene profiling studies of this host environment. It was also envisioned that this library would prove invaluable in the validation of probes designed against unculturable bacteria. Alternative phylogenetic markers might also have been included on the array to complement 16S oligonucleotide probes. The large study population would provide a uniquely diverse base from which to design a microarray tool and indeed was deemed vital to the success of the project due to the lack of canine gut microbiome sequences in the public databases.

It became apparent during the early part of the project that microarray design from the sequence libraries would provide significant challenges. Generating enough high quality sequences from which to design a microarray proved to be laborious, expensive and inefficient. Additional issues regarding probe design and signal intensity were also apparent.

The advent of high throughput sequencing technologies and their increasing exploitation during the project provided an alternative to Sanger sequencing of clone libraries. Using the Roche 454 FLX platform half a million reads of 250 nucleotides length could be produced. Pools of faecal DNA from different breeds could be sequenced on the same run using gaskets to separate up to 16 samples. In this way much greater coverage or "deep sequencing" of the microbial diversity could be achieved, albeit sacrificing some phylogenetic information.

Crucially, using a nucleotide ‘barcode’ or multiplex identifier technique, multiple (>50) 16S rRNA gene samples could be multiplex sequenced on one 454 sequencing run without the prohibitive loss of reads due to gasketting. This provided the possibility of individually

sequencing samples from all dogs in the study population to a level of ~20,000 reads in three 454 sequencing runs.

This technological shift, which would yield complete sequence data without the need for difficulties interpreting what would almost certainly be equivocal and lower resolution probe hybridisation signals, effectively rendered microarrays obsolete in the exploration of novel microbiomes. Despite this argument, microarrays still have merits such as low unit cost and other advantages associated with mature technologies. Nevertheless it was thus decided to adopt a barcoded 16S rRNA gene 454 sequencing approach as a current and future tool to characterise the canine gut microbiota.

1.4.1 Aims and hypotheses

The ultimate objective of the project, to conduct a robust analysis of the canine gut microbiome, was achieved using 454 sequencing. In the process the efficacy of the 454 sequencing technology as a tool to characterise this microbial community was tested and compared with 16S rRNA gene library sequencing. The heterogeneous nature of the sample population was then used to assess the following four hypotheses. One, the effect of host genotype on the microbiota, particularly the breed of dog but also parental relationships was explored. Two, The multiple locations that the study dogs are sourced from was used to access the effect of host location and housing on the gut microbiota. Three, the effect of host diet on the microbiota was also assessed. Four, in the concept of the 'core microbiome', that a shared microbial group can be found in most or all members of a hosts population was investigated (Turnbaugh, Hamady et al. 2009).

1.4.2 Outline of work presented

Four sections of work are presented in this thesis. The first is that of the initial Sanger sequencing work; a detailed analysis of a miniature Schnauzer clone library. The second is a later interbreed comparison of cocker spaniel, Labrador retriever and papillon libraries and a full length 16S rRNA gene phylogenetic analysis of clones within these libraries that were found to be important in the 454 work. The third results chapter, presented as an academic paper describes an initial analysis of individual miniature Schnauzer samples by 454 sequencing. Finally, the fourth results chapter also presented as an academic paper describes an in depth analysis of the total sample dog population (79 dogs representing six breeds).

CHAPTER TWO: GENERAL MATERIALS AND METHODS

The following section describes in detail various methods used during the project.

2.1 Canine sample population

Details of the dog sample population used in this project are shown in Table 1.

Table 1. Details of the study animals used thought this project.

Dog ID	Breed	Age (y)	Sex	Location	Diet
MSs1	Miniature Schnauzer	11.7	MN	WCPN, UK	PEDIGREE STANDARD WET
MSs2	Miniature Schnauzer	11.4	FN	WCPN, UK	LOW FAT DRY+WET (VET DIET)
MSs3	Miniature Schnauzer	6.2	MN	WCPN, UK	PEDIGREE ADULT DRY SB
MSs4	Miniature Schnauzer	6.2	FN	WCPN, UK	PEDIGREE ADULT DRY SB
MSs5	Miniature Schnauzer	6.2	FN	WCPN, UK	CUSTOM WET/DRY DIET
MSs6	Miniature Schnauzer	6.2	FN	WCPN, UK	PEDIGREE ADULT DRY SB
MSs7	Miniature Schnauzer	4.8	MN	WCPN, UK	PEDIGREE ADULT DRY SB
MSs8	Miniature Schnauzer	4.8	FE	WCPN, UK	PEDIGREE ADULT DRY SB
MSs9	Miniature Schnauzer	2.6	FN	WCPN, UK	PEDIGREE ADULT DRY SB
MSs10	Miniature Schnauzer	2.6	MN	WCPN, UK	PEDIGREE ADULT DRY SB
MSs11	Miniature Schnauzer	2.1	FE	WCPN, UK	PEDIGREE ADULT DRY SB
Ls1	Labrador	11.4	FN	WCPN, UK	PEDIGREE ADULT DRY
Ls2	Labrador	12.4	FN	WCPN, UK	RC MOBILITY SUPPORT
Ls3	Labrador	9	MN	WCPN, UK	PEDIGREE ADULT DRY
Ls4	Labrador	9	FN	WCPN, UK	PEDIGREE ADULT DRY
Ls5	Labrador	8.8	FN	WCPN, UK	RC SENSITIVITY CONTROL
Ls6	Labrador	8.7	MN	WCPN, UK	PEDIGREE ADULT DRY SB
Ls7	Labrador	7.8	FN	WCPN, UK	PEDIGREE ADULT DRY
Ls8	Labrador	6.8	MN	WCPN, UK	PEDIGREE ADULT DRY
Ls9	Labrador	6.8	FN	WCPN, UK	PEDIGREE ADULT DRY
Ls10	Labrador	4.4	MN	WCPN, UK	PEDIGREE ADULT DRY
Ls11	Labrador	4.4	FN	WCPN, UK	PEDIGREE ADULT DRY
Ls12	Labrador	3.4	FN	WCPN, UK	PEDIGREE ADULT DRY
Ls13	Labrador	3.4	FE	WCPN, UK	PEDIGREE ADULT DRY
Ls14	Labrador	1.3	FE	WCPN, UK	PEDIGREE JUNIOR DRY
COSs1	Cocker Spaniel	6.9	MN	WCPN, UK	PEDIGREE ADULT DRY
COSs2	Cocker Spaniel	6.9	FN	WCPN, UK	PEDIGREE ADULT DRY SB
COSs3	Cocker Spaniel	6.7	MN	WCPN, UK	CHAPPIE DRY
COSs4	Cocker Spaniel	4.8	FN	WCPN, UK	PEDIGREE ADULT DRY SB
COSs5	Cocker Spaniel	4.8	FN	WCPN, UK	CHAPPIE DRY

COSs6	Cocker Spaniel	4.1	MN	WCPN, UK	CHAPPIE DRY
PAs1	Papillon	7.5	FN	WCPN, UK	PEDIGREE ADULT DRY SB
PAs2	Papillon	7.5	FN	WCPN, UK	PEDIGREE ADULT DRY SB/ PUPPY
PAs3	Papillon	6.9	MN	WCPN, UK	PEDIGREE ADULT DRY SB
PAs4	Papillon	6.9	MN	WCPN, UK	PEDIGREE ADULT DRY SB
PAs5	Papillon	2.7	MN	WCPN, UK	PEDIGREE ADULT DRY SB
PAs6	Papillon	2.7	MN	WCPN, UK	PEDIGREE ADULT DRY SB
GSDs1	German Shepherd	8	FE	RC, France	Varied
GSDs2	German Shepherd	3	FE	RC, France	Varied
GSDs3	German Shepherd	4	FE	RC, France	Varied
GSDs4	German Shepherd	7	FN	RC, France	Varied
GSDs5	German Shepherd	6	FE	RC, France	Varied
GSDs6	German Shepherd	7	FN	RC, France	Varied
GSDs7	German Shepherd	3	FE	RC, France	Varied
GSDs8	German Shepherd	3	FE	RC, France	Varied
GSDs9	German Shepherd	?	?	RC, France	Varied
GSDs10	German Shepherd	7	FE	RC, France	Varied
GSDs11	German Shepherd	6	ME	PSF,UK	Eukanuba maintenance
GSDs12	German Shepherd	11	ME	PSF,UK	Eukanuba maintenance
GSDs13	German Shepherd	5	ME	PSF,UK	Eukanuba ID
GSDs14	German Shepherd	10	ME	PSF,UK	Eukanuba ID
GSDs15	German Shepherd	8	ME	PSF,UK	Eukanuba maintenance
GSDs16	German Shepherd	9	ME	PSF,UK	Eukanuba ID
GSDs17	German Shepherd	9	FE	Site 1, Australia	Kangaroo tails
GSDs18	German Shepherd	3	MN	Site 2, Australia	Bonnie (dry) + meat/bones
GSDs19	German Shepherd	1	FE	Site 2, Australia	Bonnie (dry) + meat/bones
GSDs20	German Shepherd	1.8	FE	Site 3, Australia	Purina Pro Plan - Performance (dry)
GSDs21	German Shepherd	6	ME	Site 3, Australia	Purina Pro Plan - Performance (dry)
GSDs22	German Shepherd	3	FE	Site 3, Australia	Purina Pro Plan - Performance (dry)
GSDs23	German Shepherd	2	ME	Site 4, Australia	Advance Adult (chicken) dry food
GSDs24	German Shepherd	1.8	ME	Site 5, Australia	Raw fresh homemade diet
GSDs26	German Shepherd	1	ME	Site 6, Australia	OPTIMUM™ Adult Large Breeds (dry) - Chicken, Rice & Vegetables
GSDs27	German Shepherd	2.5	FE	Site 7, Australia	Royal Canin 430
GSDs28	German Shepherd	1.8	FE	Site 8, Australia	Royal Canin Maxi Adult (dry)
GSDs29	German Shepherd	6	ME	Site 9, Australia	Purina Pro Plan (chicken & rice) (80% of diet)
HUSs1	Husky	6	FN	Site 10, Australia	1/2 cup kibble (Eureka Export) 1/2 cup meat (raw beef mince)
HUSs2	Husky	6	ME	Site 10, Australia	1/2 cup kibble (Eureka Export) 1/2 cup meat (raw beef mince)
HUSs3	Husky	8	FN	Site 11, Australia	BARF
HUSs4	Husky	8	MN	Site 11, Australia	BARF
HUSs5	Husky	7	ME	Site 12, Australia	Advance Energy (dry) plus sardines added for one week
HUSs6	Husky	5	FN	Site 13, Australia	Advance Energy (dry) plus sardines added for one week
HUSs7	Husky	1	ME	Site 14, Australia	Eukanuba performance (dry)

HUSs8	Husky	2.5	ME	Site 14, Australia	Eukanuba performance (dry)
HUSs9	Husky	5	FE	Site 15, Australia	1 part Advance Energy (dry) to 1 part chicken mince
HUSs10	Husky	3.5	MN	Site 15, Australia	1 part Advance Energy (dry) to 1 part chicken mince
HUSs11	Husky	1	ME	Site 15, Australia	1 part Advance Energy (dry) to 1 part chicken mince
HUSs12	Husky	5	FN	Site 15, Australia	1 part Advance Energy (dry) to 1 part chicken mince
HUSs13	Husky	1.4	ME	Site 16, Australia	Eagle Pack - Power Formula (dry) with small amount of beef mince
HUSs14	Husky	1.4	FE	Site 16, Australia	Eagle Pack - Power Formula (dry) with small amount of beef mince

MN, Male neutered. ME, Male intact. FN, Female neutered. FE, Female intact. SB, Small bite (smaller kibble size). WCPN, WALTHAM® Centre for Pet Nutrition. RC, Royal Canin research facility. PSF, Public sector facility. GSDs25 was not used in this study

2.1.1 Sample collection

Faeces were collected as soon as possible after defecation by scooping a cross-section from the middle portion of the stool into a stool collection tube. Stool collection tubes containing faeces were then frozen on dry ice no more than 15 min after initial defecation. Frozen sample tubes were stored at -80ºC before DNA extraction

2.1.2 Faecal DNA extraction

Faecal DNA was extracted from faecal samples using a QIAamp DNA Stool Mini kit (Qiagen). 190-220 mg of frozen faeces was removed from sample collection tubes using core borers and spatulas in a fume cupboard. These measurements of faeces were then processed following the 'Protocol for Isolation of DNA from Stool for Pathogen Detection' detailed in the manufactures instructions. This protocol provides the option of two different stool lysis temperatures; to ensure full lysis a temperature of 95ºC was chosen for this step.

Extracted DNA was eluted from the spin columns in 200 µl of Qiagen AE buffer (10 mM Tris-Cl and 0.5 mM pH 9.0 EDTA). Extractions were then quantified and checked for purity

on a ND1000 spectrometer (Nanodrop Technologies Inc) and then made up with HPLC grade water into 100 ng aliquots for later PCR.

2.2 PCR primers

Details of the primers used in this project are shown in Table 2.

Table 2. Primers used in this study.

Primer	Nucleotide sequence 5' - 3'	Orientation	*E. coli* 16S rRNA gene position	Source
D88	GAG AGT TTG ATY MGG CTC AG	Forward	7	Paster, Boches et al. 2001
AC84	AGA GTT TGA TYM TGG CTC AG	Forward	8	Frank, Reich et al. 2008
U519F	CAG CMG CCG CGG TAA TWC	Forward	519	Baker, Smith et al. 2003
Bact-968F	GAA CGC GAA GAA CCT TAC	Forward	968	Baker, Smith et al. 2003
E94	GAA GGA GGT GWT CCA DCC	Forward	1525	Paster, Boches et al. 2001
E533	TIA CCG III CTI CTG GCA C	Reverse	533	Watanabe, Kodama et al. 2001
Univ-1492R	GGT TAC CTT GTT ACG ACT T	Reverse	1492	Baker, Smith et al. 2003
Bact-1054R	ACG AGC TGA CGA CAG CCA TG	Reverse	1054	Baker, Smith et al. 2003
Bact-683R	GCA TTT CAC CGC TAC AC	Reverse	683	Baker, Smith et al. 2003

2.3 Agarose gel electrophoresis

Agarose gel electrophoresis enables the visualisation of nucleic acid fragments within a sample. This is achieved by placing a sample in an agarose gel matrix surrounded by an electrophoresis buffer. An electric field is then applied to the buffer and negatively charged nucleic acids migrate though the agarose matrix, smaller fragments migrate faster than larger fragments. Once separated nucleic fragments bound to DNA collators present within the gel matrix may be visualised under UV light.

Different gel concentrations, volumes, DNA collators and equipment were used thought the project, thus a general protocol is provided below. A mixture of agarose (Bioline) and 1X Tris-

borate-EDTA (TBE) buffer was heated in an 850 W microwave oven until homogeneous (generally a 1-2% agarose gel was made). After cooling to 50°C 1-6 µl of Ethidum Bromide (Sigma) 10 mg ml^{-1} was added before the mixture was poured and allowed to form a gel in an appropriate apparatus. Wells within the gel were loaded with molecular weight markers and samples along with a volume of loading dye (Bioline or Qiagen) required to produce 1 X concentration. Captions describing agarose gel figures within this text will exclude loading dye in the volumes stated, so that it is clear how much actual sample was loaded. The gel was placed in electrophoreses apparatus for between one and two hours at 100 V, 500 mAMPs and visualised on a Bio-Rad Gel Doc 2000 imager.

2.4 Bioinformatics.

2.4.1 Ribosomal Database Project (RDPii)

The ribosomal database project (RDP) is an open access annotated 16S rRNA gene sequence database (Cole, Chai et al. 2007). As well as providing access to 16S rRNA gene sequences it also provides a number of tools for analysing 16S rRNA gene clone library and high throughput sequencing data.

2.4.2 RDP Classifier

RDP classifier is tool which uses a naïve Bayesian algorithm to assign taxonomy (Wang, Garrity et al. 2007). The tool assigns taxonomy down to the genus level with confidence levels for each rank.

2.4.3 ClustalX

ClustalX is a multiple sequence alignment application which consists of the ClustalW program with a graphical user interface (Thompson, Gibson et al. 1997; Larkin, Blackshields et al. 2007). ClustalX Version 2.03 was used to create multiple sequence alignments, phylogenetic trees and DNA distance matrixes.Multiple sequence alignments were performed using default settings. Briefly, Clustal X creates sequence alignments by first producing a DNA distance matrix of FASTA format input sequences by performing pairwise alignments of sequences. Secondly a neighbour joining guide tree is produced from this distance matrix. Finally a progressive alignment is performed using the branch order of the guide tree (Larkin, Blackshields et al. 2007).

2.4.4 Rarefaction analysis

Rarefaction curves are generally calculated using the following equation (Heck Jr, van Belle et al. 1975):

$$E(S_n) = \sum_{i=1}^{s}\left[1-(N-\frac{Ni}{n}/(\frac{N}{n})\right]$$

Where n is expected number of species, drawn from a population size N which has S species. Where N_i represents the number of individuals in the i^{th} species in the sample.Rarefaction curves were the produced from this information.

2.4.5 Chao1 minimum richness estimate

The Chao1 minimum richness estimate is based on the following equation:

$$S_{chao1} = S_{obs} + n_1^2 / (2n_2)$$

Where S_{obs} is the observed number of ribotypes (OTU's); and n_1 and n_2 are the number of ribotypes observed once or twice respectively (Chao 1984).

2.4.6 Shannon-Wiener diversity index

The Shannon-Wiener index, a nonparametric diversity index (Shannon and Weaver 1963). The index is calculated using the following equation:

$$H' = -\sum[P_i(\ln P_i)]$$

Where P_i is the proportion of individuals (clone sequences) found in ith ribotype (OTU) of the community (clone library).

CHAPTER THREE: CHARACTERISATION OF MINIATURE SCHNAUZER FAECAL MICROBIOTA VIA 16S CLONE LIBRARY SEQUENCING.

3.1 Introduction

At the start of this project the human microbiota was beginning to be characterised by large scale clone library sequencing projects (Eckburg, Bik et al. 2005). Eckberg's paper was indicative of the revolution occurring in microbial ecology and provided a benchmark for studies into microbiota of mammals at this time.

Briefly, the approach used by Eckberg was to create and sequence 16S rRNA gene clone libraries from six different sub-anatomical sites and faeces of the human colon in 3 different subjects (Figure 3).

The method used in this chapter is similar to that shown above but has some key differences. The primers used in the 16S PCR were designed to amplify the whole 16S rRNA gene molecule to capture its full phylogenetic properties whereas Eckburgs approach only amplified a portion of the molecule. A single 16S rRNA gene library would be created from canine faeces collected from 12 miniature Schnauzer dogs and then sequenced (Figure 4). The library would be a representation of the microbiota of 12 dogs achieved by the cloning of a PCR product pool which was produced by pooling of individual faecal DNA PCRs. This is a significant divergence from Eckburg's study which created one library per PCR reaction. It

was hoped that this approach would reduce the impact of any potentially abnormal microbiota from a single sample.

Figure 3. Flow chart showing the experimental method of (Eckburg, Bik et al. 2005).

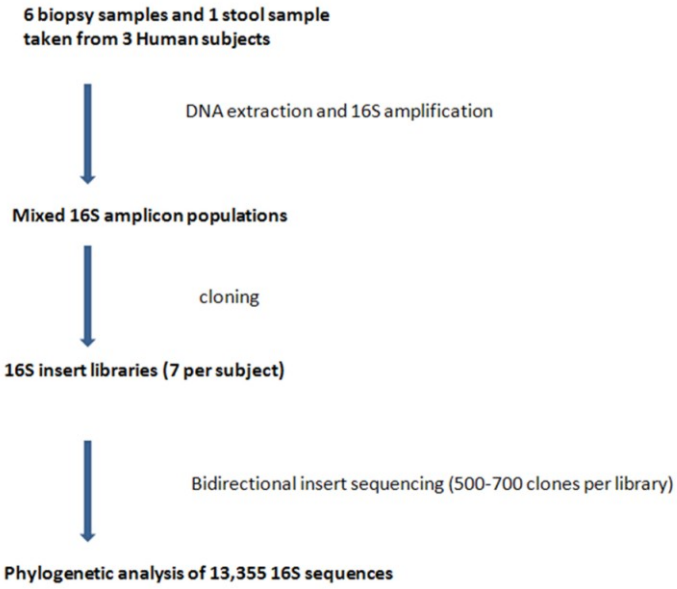
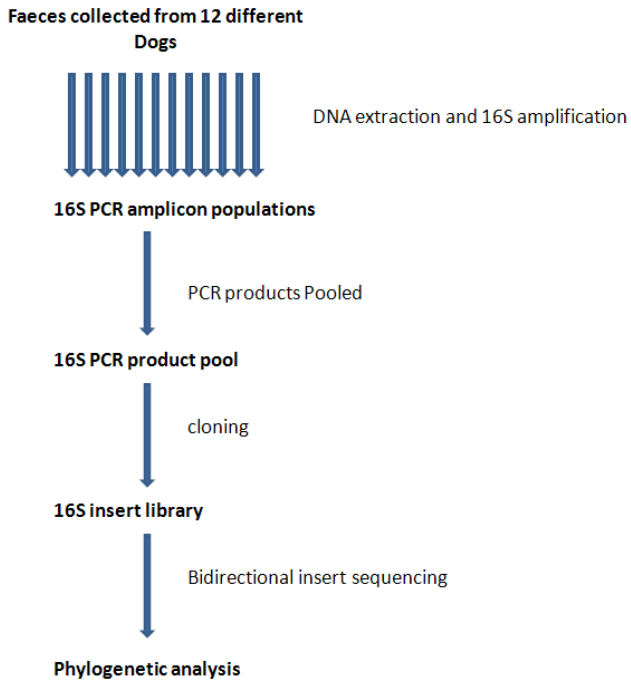


Figure 4. Flow chart showing the planned experimental approach of this study.



The aim was to create and sequence a 16S rRNA gene library as an initial proof of concept. After which further breed libraries would be sequenced using this work as the template.

3.2 Materials and methods

3.2.1 Faeces Collection

One faecal sample per dog was collected from the miniature Schnauzer breed dogs listed in Table 1. The sampling method is described in section 2.1.1.

3.2.2 Faecal DNA extraction

Extractions were performed as described in section 2.1.2.

3.2.3 16S rRNA gene PCR

16S PCR reactions were performed using faecal DNA as template with the D88 and E94 primer pair (Table 2, page 25). The primers are adaptations of those used in a previous study into oral cavity microbiota (Paster, Boches et al. 2001).

PCR reactions consisted of the following regents:

Reagent	Volume/amount
10X Pfx amplification buffer buffer (Invitrogen, final concentration: 2X)	10 µl
D88 Forward primer (10 pmol/µl)	2 µl
E94 Reverse primer (10 pmol/µl)	2 µl
1.5 µl dNTP mixture (Invitrogen, 10 mM)	1.5 µl
$MgSO_4$ (Invitrogen, 1 mM)	1 µl
Pfx Polymerase (Invitrogen 2.5 U/µl)	0.5 µl
H_20 (HPLC grade)	28 µl
faecal DNA (100 ng) or HPLC grade water for negative controls	5 µl
Total volume	**50 µl**

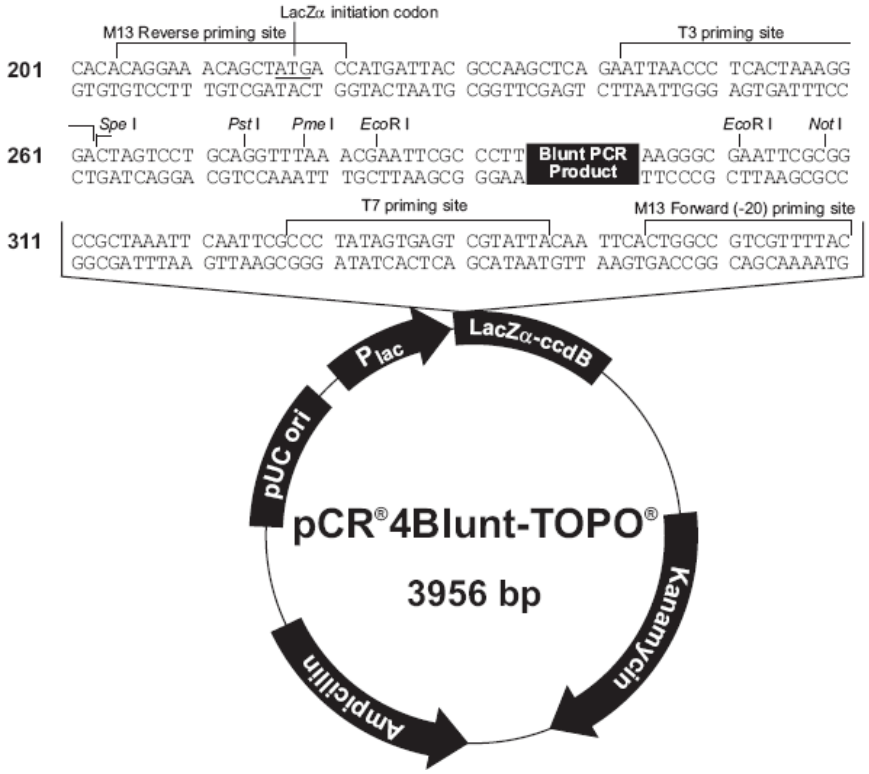
Reactions were incubated in a Tetrad DNA Engine thermal cycler within thin walled 0.2 ml micro centrifuge tubes. Two PCR thermal cycler programs were used in the study. The first was a gradient PCR program used to identify correct PCR conditions: 94ºC for 2 min followed by 25 cycles of 94ºC for 45 s, 50-60ºC gradient (see Figure 6) for 1 min, 68ºC for 2 min, 30 s and a single extension of 72ºC for 10 min. A second optimised PCR program consisted of the following cycles: 94ºC for 2 min followed by 25 cycles of 94ºC for 45 s, 60ºC for 1 min, 68ºC for 2 min, 30 s and a single extension of 72ºC for 10 min.

PCR reaction products were electrophoresied at 100 V on 1% agarose gels either cast or stained overnight with SYBRgreen and then visualised on a UV imager.

3.2.4 Cloning

16S PCR reactions from all but one faecal sample template were pooled by adding 35 µl of each PCR product to a single micro centrifuge tube and mixing thoroughly. Four microlitres of this pool was cloned into invitrogen's pCR®4Blunt-TOPO® vector (Figure 5) using invitrogen's 'Zero Blunt TOPO PCR Cloning Kit for Sequencing'. Cloning was conducted following the general manufactures instructions; however; specific choices were made to maximise the clone numbers of the library.

Figure 5. A map of the pCR®4Blunt-TOPO® vector.



The initial vector construction was done via a topoisomerase reaction in the following mixes and included a positive control and a vector only control:

MS16S- pCR®4Blunt-TOPO Construct		Ve^+ control PCR- pCR®4Blunt-TOPO Construct		pCR®4Blunt-TOPO (Vector only control)	
16SPCR pool	4 µl	Ve^+ control PCR	1 µl		
		H_20 (Invitrogen)	3 µl	H_20 (Invitrogen)	4 µl
Salt solution(Invitrogen)	1 µl	Salt solution(Invitrogen)	1 µl	Salt solution(Invitrogen)	1 µl
pCR®4Blunt-TOPO	1 µl	pCR®4Blunt-TOPO	1 µl	pCR®4Blunt-TOPO	1 µl
Total volume	6 µl	Total volume	6 µl	Total volume	6 µl

The reactions were incubated at room temperature for 30 minutes before being chemically transformed into DH5α-T1 competent *E. coli* as per the manufacture's protocol. This included conducting a transformation control reaction supplied in the kit. Transformation-SOC mixtures were spread on pre-warmed selective agar plates.

For the MS16S- pCR®4Blunt-TOPO Construct transformation 140 µl MS-16S rRNA gene-pCR®4Blunt-TOPO Construct transformation spread onto 2 large LB agar 50 µg/µl kanamycin agar plates each. 1000 µl of SOC medium (Invitrogen) was used to lubricate the agar surface of each plate prior to spreading of the transformation. 15 µl MS16S- pCR®4Blunt-TOPO Construct transformation spread onto a standard LB agar 50 µg/µl kanamycin agar plate. 20 µl of SOC medium (Invitrogen) was used to pre lubricate the agar plate.

For the positive control PCR-pCR®4Blunt-TOPO Construct transformation 140 µl positive control PCR-pCR®4Blunt-TOPO Construct transformation spread onto a large LB agar 50 µg/µl kanamycin plate. 1000 µl of SOC medium (Invitrogen) was used to pre lubricate the agar plate.

For the pCR®4Blunt-TOPO Vector only control transformation, 50 µl Vector only control transformation was spread on a standard LB agar 50 µg/µl kanamycin plate.

For the Transformation Control, 10 µl Transformtion control wsa spread on a standard LB agar 50 µg/µl ampicillin plate. 20 µl of SOC medium (Invitrogen) was used to pre lubricate the agar plate.

Inoculated agar plates were incubated overnight at 37ºC, resulting colonies were then counted. All colonies on the MS16S- pCR®4Blunt-TOPO transformation plates were then sub cultured into individual wells containing LB Hogness modified freezing medium (3.6 mM K2HPO4, 1.3 mM KH2PO4, 2.0 mM Trisodium citrate, 1.0 mM MgSO4.7H2O and 4.4% (v/v) glycerol) supplemented with 50 µg/µl kanamycin. These plates were then incubated overnight at 37ºC and then stored at -80ºC to serve as the clone library stock.

3.2.5 Plasmid DNA extraction

Plasmid DNA was prepared from the clone library using two protocols. In the first protocol plasmid DNA was prepared to analyse clone insert size. Clones from the library were inoculated onto LB agar supplemented with 50 µg/µl kanamycin using a 48 prong inoculator and plates subsequently incubated overnight at 37ºC. Colonies from this agar plate were

individually inoculated into universals containing 10ml LB broth supplemented with 50 µg/µl kanamycin and incubated at 37ºC with 200 rpm agitation for 12-16 hours. Cells were then harvested by centrifugation of 5 ml of culture at 5400 G in a Mistral centrifuge. Plasmid DNA was then extracted from the cell pellet using a QIAprep Minprep kit (Qiagen) following manufactures instructions and eluted in 50 µl EB buffer (10 mM Tris·Cl, pH 8.5).

The second protocol was used to extract plasmid DNA robotically from the library prior to sequencing and is essentially a modification of theabove protocol. Clones from the library were inoculated into Qiagen 96 well culture blocks containing LB broth supplemented with 50 µg/µl kanamycin using a 48 prong inoculator and incubated at 37ºC with shaking at 300 rpm for 24 hours. Plasmid DNA was then extracted from the cultures using a QIAprep 96 turbo Minprep kit (Qiagen) following manufactures instructions and eluted in 100 µl EB buffer (10 mM Tris·Cl, pH 8.5).

3.2.6 Endonuclease Restriction Digests

In order to analyse clone insert size plasmid DNA was digested with *EcoRI* enzyme to excise the insert (Figure 5). Plasmid DNA was digested in the following reaction mix:

Reagent	Volume/amount
Plasmid DNA (~1µg)	10 µl
10X react3 buffer (Invitrogen)	5 µl
H_2O (milipore filtered)	34.8 µl
EcoRI enzyme (10 units/µl) or 0.2 µl H_2O for a no enzyme control	0.2 µl
Total volume	**50 µl**

Reactions were incubated at 37ºC for 25 minutes and then heat inactivated at 65ºC for 20 minutes. 10 µl of restriction digests were run at 100 V on a 1.2% agarose gel cast with ethidum bromide and then visualised with a Bio-Rad Gel Doc 2000 imager.

3.2.7 Sequencing

Cloned plasmid DNA was bidirectionally sequenced using pCR®4Blunt-TOPO insert sequencing primers (Figure 5) M13 forward (5´-GTAAAACGACGGCCAG-3') and M13 reverse (5´-CAGGAAACAGCTATGAC-3´). Sequencing reactions were set with 5 µl Plasmid DNA (200-300 ng) and 5 µl primer (3.2 pmol) either M13 forward or M13 reverse for a total volume of 10 µl.

DNA sequence was then determined using a Bigdye terminator version 3.0 cycle sequencing kit (ABI, Applied Biosystems) and an ABI 3700 capillary sequencer.

3.2.8 Phylogenetic analysis/ bioinformatics

3.2.8.1 Quality processing of Sanger reads

Sanger sequence reads were quality trimmed using CodonCode Aligner (CodonCode Corporation). Bases were called using Phred (Ewing, Hillier et al. 1998) and quality trimed using Phred quality values. Vector and primer sequences were also trimmed and sequences with less than 500 quality bases were removed from further analysis.

To detect chimeras the sequences were first aligned in Mothur (Schloss, Westcott et al. 2009) using the SILVA SEED template alignment of 14,956 bacterial sequences derived from the SILVA database (Pruesse, Quast et al. 2007). Then the Pintail program integrated in Mothur was used to detect chimeras using a reference alignment of bacterial sequences with greater than 1% nucleotide dissimilarity (Ashelford, Chuzhanova et al. 2005). Chimeras were removed from further analysis.

3.2.8.2 OTU determination and statistics

Operational taxonomic units (OTUs) were determined using the RDP infernal aligner and Complete Linkage Clustering tools (Cole, Wang et al. 2009). Rarefaction curves and Chao1 richness estimates (Chao 1984; Hill 2006) were calculated for each library using Mothur version 1.12.3 (Schloss, Westcott et al. 2009).

3.2.8.3 Taxonomic assignments

Reads were assigned to taxa using the online version of the RDP classifier equivalent to the standalone downloaded version used in chapter 7 (Wang, Garrity et al. 2007).

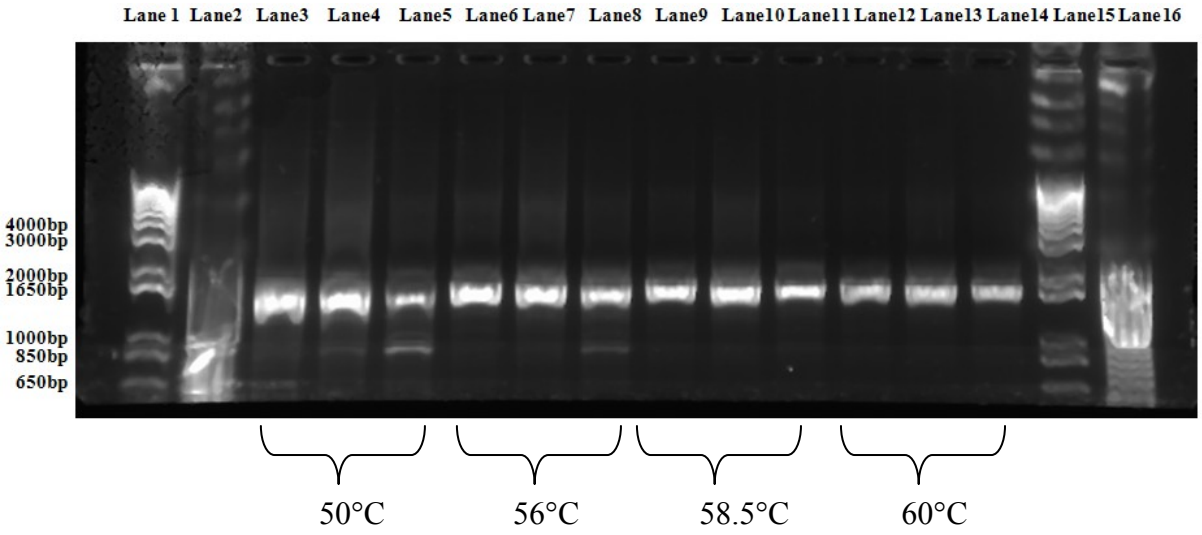
3.3 Results and discussion

Faecal samples were collected from 12 miniature Schnauzers during a two week period. The numerical designations of the dogs were: MS25, MS32, MS40, MS41, MS42, MS43, MS46, MS51, MS52, MS57, MS60 and MS72. In order to minimise a reduction in obligate anaerobe numbers upon contact with air, any stools exposed for greater than 15 minutes before freezing were rejected. Faecal DNA of high purity was successfully extracted from the stool samples with typical yields ranging between 30 ng and 60 ng.

A preliminary PCR was conducted to assess the correct annealing temperature to use with the 16S rRNA gene primers. This step was crucial as PCR with 16S rRNA gene degenerate primers has a potential to bias the amplicon pool phylogenetically depending on specific annealing temperatures and primer pair melting temperature combinations. The PCR products were visualised by agarose gel electrophoresis (

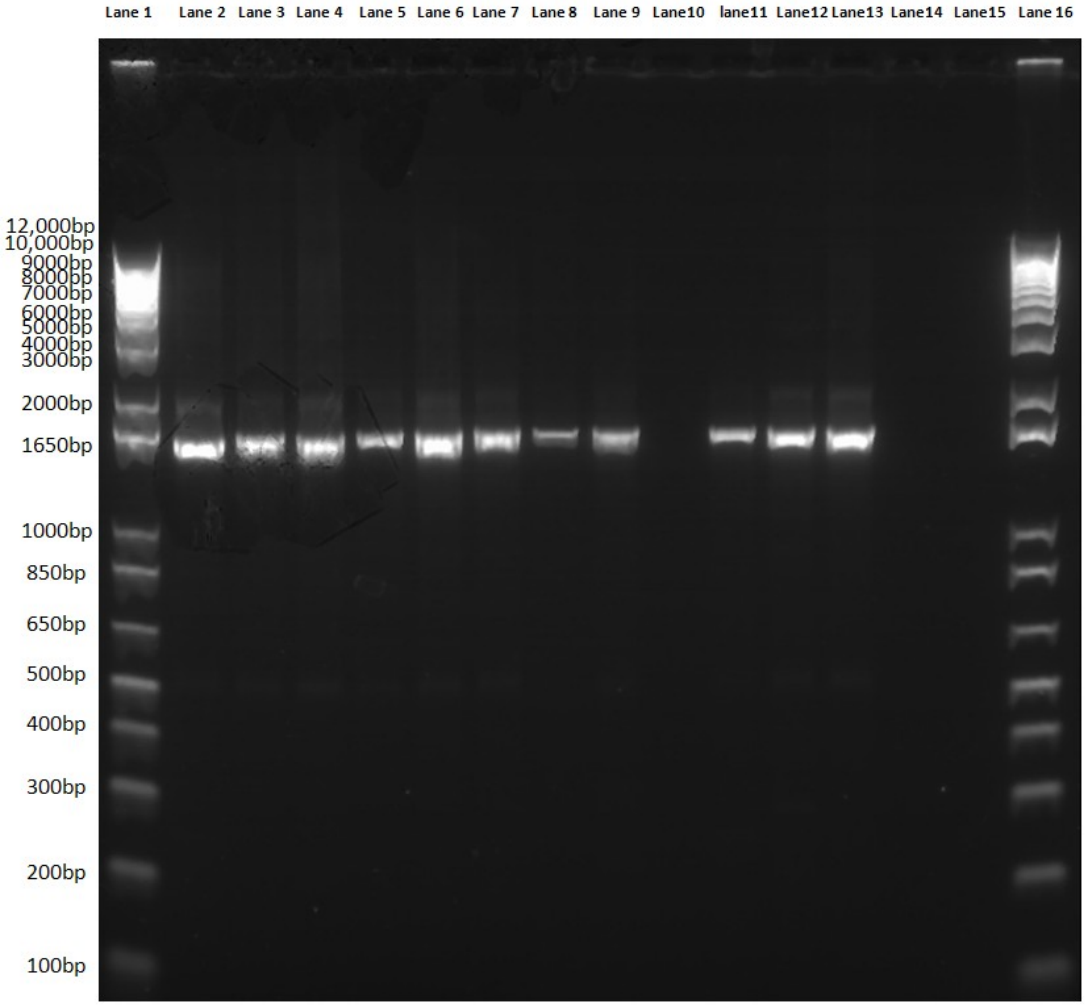
Figure 6). The three different templates used in the reactions produced 16S rRNA gene-sized products approximately 1.5 kb across all annealing temperatures. This 16S rRNA gene sized product was highly heterogeneous most likely containing amplicons from hundreds of different species. Thus the band appears slightly smeared due to slight differences in species amplicon size and the high DNA yield of the PCR. In addition to 16S rRNA gene-sized products, lower molecular weight products are visible in lanes containing PCR reactions conducted with an annealing temperature lower than 58.5°C. It was thus decided that an annealing temperature of 60°C be used for future PCR reactions to amplify up 16S rRNA gene from all the canine faecal DNA extracts.

Figure 6. Optimizing 16S rRNA gene PCRs prior to cloning. 16S rRNA gene gradient PCR products visualized on a 1% agarose gel cast with SYBRgreen. Three different templates were amplified using four annealing temperatures; the temperatures used are indicated below the figure. Gel lanes were loaded as follows: Lane1; 1µl 1k plus ladder (Invitrogen). Lanes 3, 6, 9 and 12; 10 µl of PCR's with MS32 template DNA. Lanes 4, 7, 10 and 13; 10 µl of PCR's with MS40 template DNA. Lanes 5, 8, 11 and 14; 10 µl of PCR's with MS60 template DNA. Lane15; 1 µl 1k plus ladder (Invitrogen) Lane16; distorted ladder.



PCR products were amplified from the DNA extracts using the optimised PCR protocol and visualised by agarose gel electrophoresis (Figure 7). The major visible products were of 16S rRNA gene size ~1.5 kilo bases in addition to a minor higher molecular weight ~ 2 kilo bases products. This 2 kilo bases product was probably not identified from the previous gradient reaction due to the low resolution of the gel used for visualization. A repeat of the PCR with a higher annealing temperature was briefly considered; however, this idea was rejected due to the previously mentioned concerns about introducing phylogenetic bias. These PCR amplicons were considered to represent a good balance between cloning and phylogenetic considerations. No product was produced in the negative controls; these controls are critical in PCR's using broad-range primers such as 16S rRNA gene PCR and indicate that the 2 kilo bases products did not originate from template contamination. One reaction failed to produce a product: the PCR with MS50 template. Further attempts to amplify from this template also failed, thus the library was constructed using 11 faecal samples as staring material instead of 12.

Figure 7. Successful amplification of 16S rRNA gene PCR products. 16S rRNA gene PCR products (10 µl) from 12 faecal DNA extractions visualized on a 1% agarose gel cast with SYBRgreen. Gel lanes were loaded as follows (PCR's are designated by template): Lane1; 1 µl 1k plus ladder (Invitrogen). Lane2; MS25PCR. Lane3; MS32. Lane4; MS40. Lane5; MS41. Lane6; MS42. Lane7; MS43. Lane8; MS46. Lane9; MS51PCR. Lane10; MS52. Lane11; MS57. Lane12, MS60. Lane13; MS72. Lane14; 10 µl no template control PCR. Lane15; no template control PCR. Lane16; 1 µl 1k plus ladder (Invitrogen).



A 16S amplicon pool was created by mixing equimolar amounts of the 11 successful PCR products from the second PCR reaction together. This pool served as the direct basis of the inserts for the library, thus it was hoped that any a-typical microbiotas would be buffered against. A proportion of the pool was cloned into the pCR®4Blunt-TOPO® vector resulting in 229 clones. Table 3 shows the total number of colonies produced from each transformation. The number of vector-only clones in the library was estimated to be 12.8, 5.59% of the library. However, this figure is subject to a large error due to it being derived from only two colonies on the negative ligation control transformation plate.

It had been hoped that the transformation might have yielded over 1000 colonies, to provide a large sequence dataset. The positive ligation control transformation colony yielded a greater number of colonies, however the size of the insert was half that of the 16S insert which most probably explains the increase in ligation efficiency.

Table 3. Results of transformations.

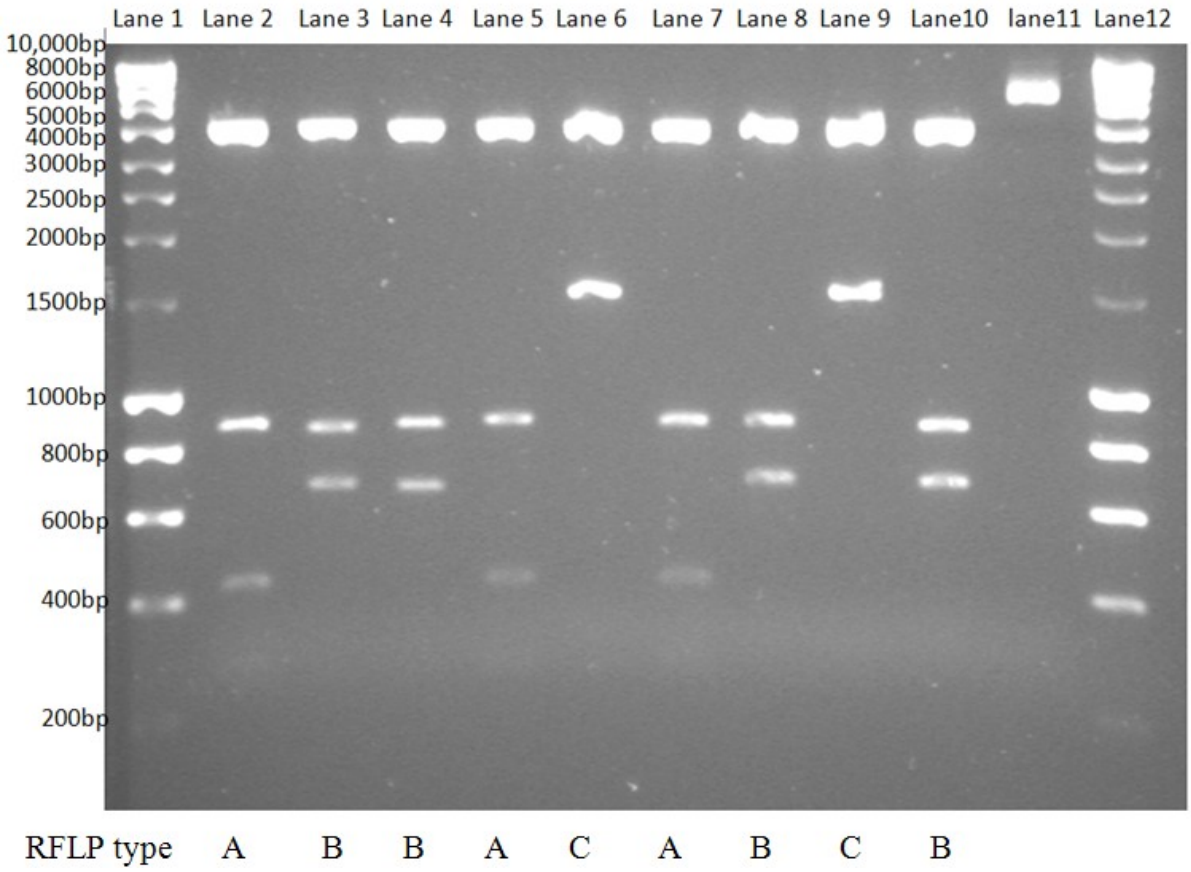
Transformation	**Colonies**	**cfu/µl^{-1} transformation Solution**
MS 16S pool Transformation	229	0.78
Positive ligation control transformation	1000+	7
Negative ligation control transformation	2	0.04
Transformation control	66	6.6

A small proportion of the library was analysed to confirm its viability prior to sequencing. Plasmid DNA was extracted from nine clones and digested with a restriction endonuclease to

excise the insert, the digests were then and visualised by agarose gel electrophoresis (Figure 8). The vector can clearly be seen as a 4 kilo bases sized band in all digest lanes. In addition to this vector band, lower molecular weight band were present across the digest lanes. Excluding the vector band, the bands within all lanes add up to ~1.5 kb and indicated that all clones checked had acquired a 16S rRNA gene-sized insert.

Heterogeneity within the insert sequences produced RFLP's (Restriction Fragment Length Polymorphisms) of 3 types (see bottom of Figure 8). Type A, B and C inserts have two, one and zero EcoRI restriction sites respectively. Although the portion of clones tested would render any solid conclusions statistically unreliable; the digest indicated that the library could contain a large number different of 16S rRNA gene inserts and suggested that the library contained at least three different phylotypes. The phylogenetic resolution of this typing is quite low and would represent diversity ranging many species.

Figure 8. RFLP types showing diversity within the MS 16S rRNA gene clone library. *EcoRI digests of clone plasmid DNA electrophorised on a 1.2% agarose gel. RFLP types are indicated at the bottom of the figure. Gel lanes were loaded as follows: Lane1; 5µl Hyperladder I (Bioline). Lane2; 10 µl clone one digest. Lane3; 10 µl clone two digest. Lane4; 10 µl clone three digest. Lane5; 10 µl clone four digest. Lane6; 10 µl clone five digest. Lane7; 10 µl clone seven digest. Lane8; 10 µl clone eight digest. Lane9; 10 µl clone nine digest. Lane10; 10 µl clone ten digest. Lane11; 10 µl no enzyme control digest. Lane12; Hyperladder I (Bioline).*



The results of the restriction digests indicated that it would be profitable to sequence the library. Plasmid DNA was robotically prepared from 192 clones, the inserts were then bidirectionally sequenced resulting in 384 sequence reads. An inspection of the sequence FASTA files (extracted from ABI 3700 generated chromas data) revealed that out of the 192 M13 forward reactions, 47 failed or produced sequences of poor quality, 41 of the M13 reverse reaction sequences were also unsatisfactory. These figures indicate that the M13 forward and reverse reactions yielded 75.5% and 78.6% potential good sequence respectively.

Unfortunately sequence read length was insufficient to enable accurate assembly of bidirectional sequence reads due to insufficient overlap between sequences. For an accurate assembly approximately 800 bp of high quality sequence was required; however, most sequencing reactions yielded 500-700 bp of good quality sequence.

Due to these difficulties it was decided that the 5' 16S rRNA gene sequence reads would be analysed instead of the whole molecule. Analysis of 5' 16S rRNA gene sequence instead of that at the 3' terminus was viewed to be favoured by the greater abundance of online 16S rRNA gene database entry's covering the 5' region such as the RDPii Website (A histogram of RDP database entry coverage of the 16S rRNA gene is provided in the index). Taxonomic analysis of 3' 16S rRNA gene sequence would be hindered due to the underrepresentation of the 3' terminus in the online datasets. Analysis using the V1, V2 and V3 variable regions of the 16S rRNA gene are also thought to be better for assigning sequences to the bacterial species level (Suau, Bonnet et al. 1999).

Good quality sequence reads were aligned using a multiple sequence alignment algorithm in CLUSTALX (Thompson, Gibson et al. 1997). The blunt ended cloning resulting in two possible 16S rRNA gene insert orientations within the vector, thus sequences within the

alignment clustered into two main groups representing both insert orientations. A small number of vector only sequences were also found. Sequence reads containing 5' 16S rRNA gene sequence were isolated from those containing 3'16S rRNA gene sequence and vector only sequence. Five prime 16S sequence reads were realigned using CLUSTALX. Vector sequence was trimmed from the 5'prime end of reads and poor quality sequence from the 3' end. A number of sequences were also removed due to poor sequence quality identified through alignment anomalies.

This analysis produced a set 141 sequences of 555-610 bp length covering the V1, V2 and V3 hypervariable regions.

The dataset was then analysed to detect chimeric PCR artefacts. Chimera sequences are produced during 16S rRNA gene PCR and represent an important issue for 16S rRNA gene libraries (v. Wintzingerode, Gobel et al. 1997). The dataset was analysed for chimeras using Pintail (Ashelford, Chuzhanova et al. 2005). Chimera checking tools do not positively identify chimeras, they only give a general indication. No chimeras were positively flagged in the dataset (although it is likely some are present).

Operational Taxonomic Unit's (OTU's) within the dataset were determined. At the 97% nucleotide sequence identity level 39 OTUs were observed. OTU classification at 97% sequence identity has been considered an adequate species level indicator. In the post genomic era some authors have started to use a 99% sequence identity OTU species classification, citing the high phenotypic diversity in some bacterial groups sharing high 16S identity matches (Eckburg, Bik et al. 2005). However, others suggest OTU classifications above 98% sequence identify may not be as useful for studying microbial communities as 16S

rRNA gene resolution is limited above genus level (Konstantinidis and Tiedje 2005; Konstantinidis and Tiedje 2007).

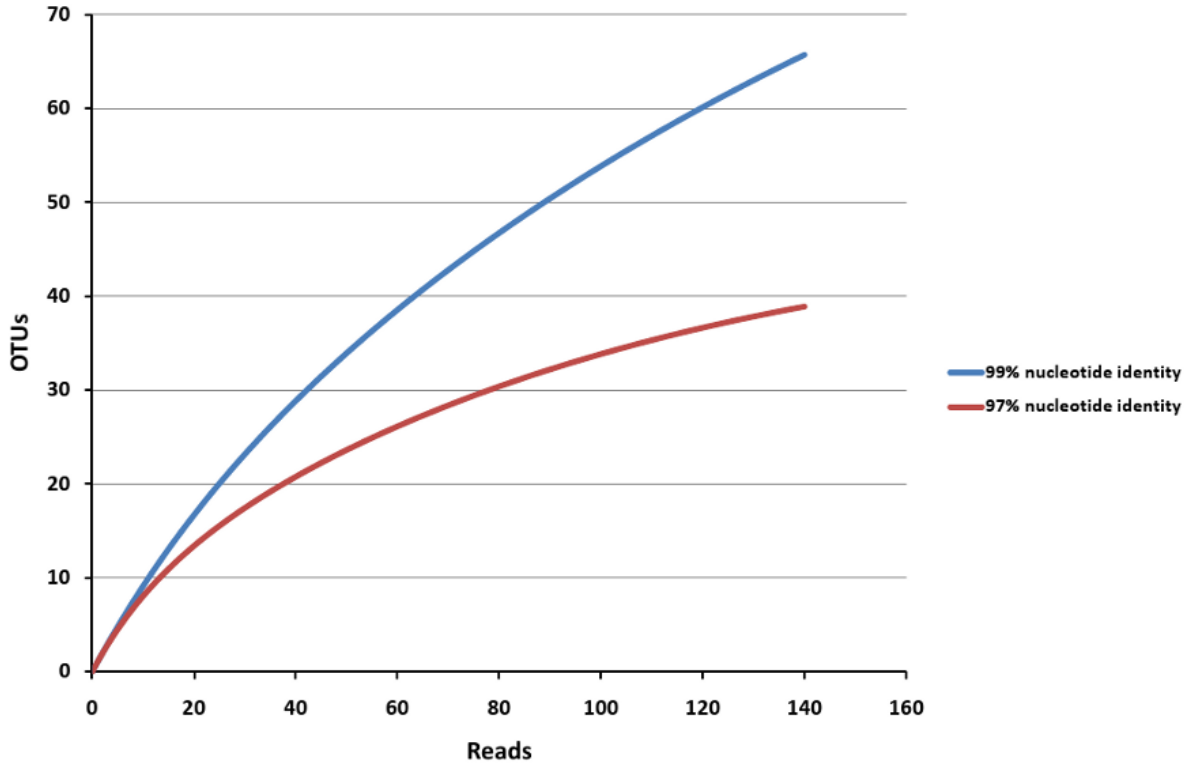
The Chao1 richness estimator (Chao 1984) predicted that in the cloned 16S amplicon pool there was a minimum of 46 OTU's at the 97% sequence identity level. Using this estimator the 141 sequence dataset represents 16S rRNA gene amplicon pool coverage of 84.8%. However, this dataset is relatively small compared to other 16S rRNA gene library projects (Eckburg, Bik et al. 2005; Ley, Bäckhed et al. 2005; Palmer, Bik et al. 2007). It has been noted that Chao1 will underestimate true richness when applied to small datasets, thus the actual coverage may be lower than this (Hughes, Hellmann et al. 2001). Previous work has shown that human intestinal microbiota communities are comprised of hundreds of bacterial species (Eckburg, Bik et al. 2005). While other work has shown that carnivore intestinal communities have lower bacterial diversity than those of omnivores, it is likely that greater sampling would reveal more diversity (Ley, Hamady et al. 2008). In the high throughput sequencing analyses of these miniature Schnauzer samples in chapters 4 and 5 it is clear the true value is much higher.

Rarefaction curves for both sequence identity levels are shown in Figure 9 (Hughes and Hellmann 2005). These curves represent the expected accumulation of novel OTU's with increasing standardized sampling effort. The asymptotic curves suggest that the 16S rRNA gene amplicon pool is relatively well sampled and that further sampling will result in diminishing returns in novel OTU identification. As with the Chao1 estimate the small sample size limits the value of this analysis. Although asymptotic curves are still an important result, rarefaction analysis at low sampling levels can be misleading as it is merely an indicator of

observed richness at a particular sampling effort. For an accurate comparison with other studies, rarefaction curves would have to be produced from these larger datasets at the same sampling level of this study.

The Shannon-Wiener diversity index (Shannon and Weaver 1963) value was determined to be 3.19 at the 97% nucleotide identity level.

Figure 9. Rarefaction curves for the MS clone library. The curves represent the rate at which novel OTU's are classified with increasing sampling effort. Two OTU classifications levels are displayed. The red and blue curves represent OTU's identified at a 99% and a 97% identify cut-off respectively.



Sequences were then analysed using the RDP classifier version 2.2 with a 30% genus level bootstrap filter. A 30% bootstrap filter was used for the reasons outlined in section 5.4.1 on page 91. This taxonomically classified the reads from the phylum down to the genus level (Table 4).

Four phyla were represented in the library the *Firmicutes*, *Fusobacteria*, *Bacteroidetes* and *Proteobacteria*. Assignments at the Phylum level for clone sequences shown in . The majority of sequences belonged to the *Firmicutes* phylum which also contained the most diversity. The most abundant *Firmicutes* family's were the *Lachnospiraceae* and *Veillonellaceae* represented most prominently by the *Coprococcus* and *Phascolarctobacterium* genera respectively. Two sequences belonging to one OTU were classified to the *Lactobacillus* genus, this OTU shared 97.55% sequence identity with the *Lactobacillus acidophlius*. The *Firmicutes* and *Bacteroidetes* are frequently found to be the major constituents of the mammalian intestinal microbiota (Hold, Pryde et al. 2002; Wang, Heazlewood et al. 2003; Eckburg, Bik et al. 2005; Ley, Hamady et al. 2008). However, unlike in other mammalian gut microbiota studies where the *Fusobacteria* are present as a small minority in these results they represent the second largest group. All but two of the *Fusobacteria* classified to the *Fusobacterium* genus. The *Bacteroidetes* were similarly largely represented by one genera. The lack of diversity in the *Bacteroidetes* and *Fusobacteria* as compared to the *Firmicutes* can be seen in the differences between classified sequences and OTUs in Figure 10 and Figure 11 respectively.

Table 4. Taxonomic assignment of MS 16S rRNA gene library reads using the RDP classifier.

Phyla (bold), genera	Reads
Bacteroidetes	**27**
Bacteroides	21
Paraprevotella	3
Prevotella	3
Firmicutes	**80**
Acetanaerobacterium	1
Anaerobacter	1
Anaerofilum	2
Anaerotruncus	1
Anaerovibrio	2
Blautia	7
Bulleidia	2
Clostridium	5
Coprococcus	11
Faecalibacterium	9
Lactobacillus	2
Megamonas	8
Phascolarctobacterium	24
Robinsoniella	1
Roseburia	1
Syntrophococcus	2
Thermotalea	1
Fusobacteria	**29**
Cetobacterium	2
Fusobacterium	27
Proteobacteria	**4**
Sutterella	4
Grand Total	**140**

Reads with genus bootstrap values below 30% were excluded. Genera numbers are proportionally coloured from red to yellow to green based on the lowest and highest genera abundance observed.

Figure 10. Phylum level sequence distribution in the MS clone library.

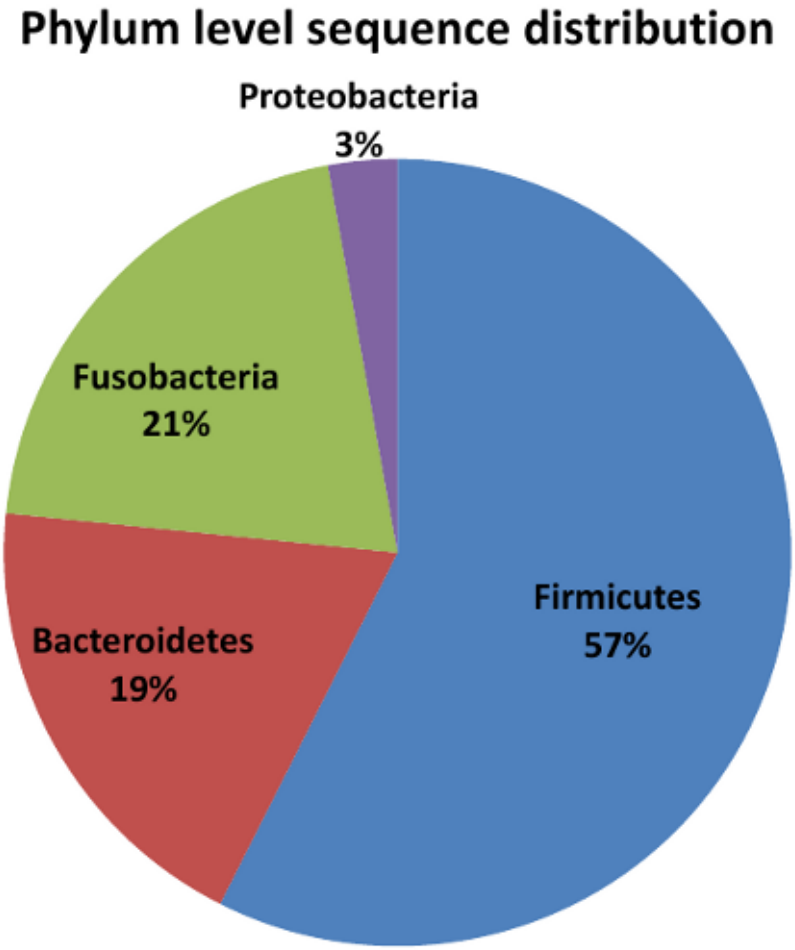
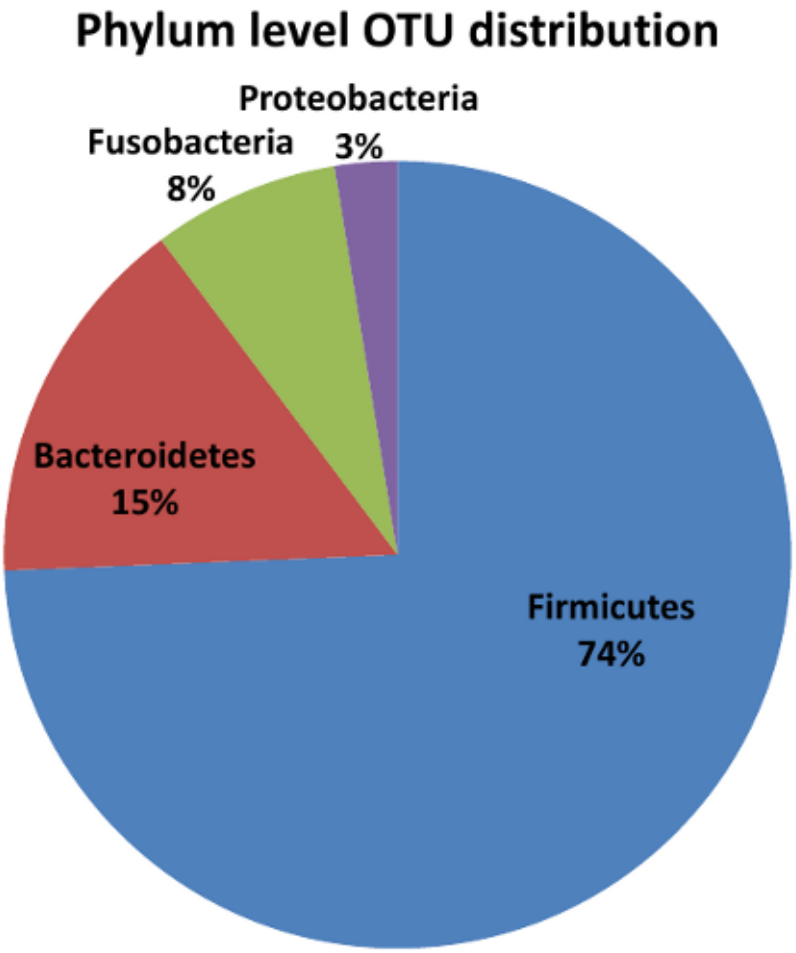


Figure 11. Phylum level OTU classification in the MS clone library showing increased diversity in the Firmicutes *compared to other phyla.*



The *Proteobacteria* were represented by justfour clone sequences assigned to one OTU classified as the genus *Sutterella* in the *Burkholderiales* Order. The *Sutterella* A low abundance of *Proteobacteria* are also see in other faecal 16S rRNA gene microbiota studies (Eckburg, Bik et al. 2005). This is also consistent with the general trend for the *Proteobacteria* to decrease in abundance from the proximal to distal gut, The distal gut being most represented by stool samples (Eckburg, Bik et al. 2005).

This dataset is small compared to that of other experiments of this type (Eckburg, Bik et al. 2005; Ley, Hamady et al. 2008). While limited, the results provided a promising pilot study for the larger and more comprehensive work described in the following chapters.

CHAPTER FOUR: BREED SPECIFIC CHARACTERISATION OF CANINE FAECAL MICROBIOTA BY SANGER SEQUENCING.

4.1 Introduction

It has been observed that host genotype has a significant if not dominating effect on intestinal microbiota composition (Zoetendal 2001; Ley, Hamady et al. 2008). Domesticated canines (*Canis lupus familiaris*) have a large amount of genetic and phenotypic diversity due to extensive artificial selection over periods ranging from hundreds to thousands of years. It follows that any attempt to catalogue canine faecal microbiota fully should include inter-breed sampling to negate potential breed bias. Additionally, discovery of substantial inter-breed microbiota differences may have important implications for breed health and nutrition. The small 16S rRNA gene library of 141 clones originating from miniature Schnauzer faecal microbiota described in chapter 3 was a pilot for larger libraries. Additional canine breed-specific libraries were created from pooled 16S rRNA gene derived from the Labrador retrievers, cocker spaniel and Papillon breeds in the study population shown in Table 1. These libraries were sequenced to further assess the diversity of the canine microbiome and to ascertain whether there was a breed difference.

In order to verify the taxonomic assignments of some of the most widespread and abundant bacteria found throughout this project a targeted full length 16S rRNA gene sequencing programme was employed. The ‘gold standard’ of identification: high quality full length 16S rRNA gene sequence alignment and phylogenetic tree reconstruction, was then used to test the accuracy of RDP classification of the single Sanger and pyrosequencing reads.

4.2 Materials and methods

4.2.1 Faeces Collection

Faeces were collected as described in section 2.1.1 from all the cocker spaniel, Papillon and Labrador retriever breed dogs in Table 1.

4.2.2 Faecal DNA extraction

Faecal DNA was extracted as described in section 2.1.2.

4.2.3 16S rRNA gene PCR

PCR reactions were performed using faecal DNA as template with the following primer pair:

D88 (7-27 forward, 5'-GAG AGT TTG ATY MGG CTC AG-3')
E94 (1525-1541 reverse, 5'-GAA GGA GGT GWT CCA DCC-3')

The primers are adaptations of those used in a previous study into oral cavity microbiota (Paster, Boches et al. 2001).

PCR reactions consisted of the following reagents:

Reagent	Volume/amount
5X HF amplification buffer (Phusion High-Fidelity PCR Master Mix, NEB)	5 µl
D88 Forward primer (10 pmol/µl)	2.5 µl
E94 Reverse primer (10 pmol/µl)	2.5 µl
dNTP mixture (Bioline, 25 mM)	0.4 µl
H_20 (HPLC grade)	29.1 µl
faecal DNA (100 ng) or HPLC grade water for negative controls	5 µl
Total volume	**50 µl**

Reactions were incubated in a BioRad Tetrad DNA Engine thermal cycler within thin walled 0.2 ml micro centrifuge tubes. The following PCR conditions were used: 98ºC for 30 s followed by 25 cycles of 98ºC for 10 s, 60ºC for 30 s, 72ºC for 1 min and a final extension of 72ºC for 10 min.

PCR reaction products were electrophoresed at 100V on 1% agarose gels either incorporating or stained overnight with SYBRgreen and then visualised on a UV imager.

4.2.4 Cloning

Cloning was done as described in section 3.2.4 using breed specific pooled 16S rRNA gene amplicons.

4.2.5 Plasmid DNA extraction

Plasmid extraction was done as described in section 3.2.5.

4.2.6 Sequencing

Clone plasmid DNA was sequenced using the D88 (7-27 forward, 5'-GAG AGT TTG ATY MGG CTC AG-3') primer for the single read analysis. To generate full length 16S rRNA gene sequences three forward and three reverse primers were used (Baker, Smith et al. 2003; Frank, Reich et al. 2008).

The forward primers were AC84 5'-AGAGTTTGATYMTGGCTCAG-3', U519F 5'–CAGCMGCCGCGGTAATWC-3' and Bact-968F 5'-GAACGCGAAGAACCTTAC-3'. The reverse primers used were Univ-1492R 5'-GGTTACCTTGTTACGACTT-3', Bact-1054R 5'-ACGAGCTGACGACAGCCATG-3' and Bact-683R 5'-GCATTTCACCGCTACAC-3' (Table 2, page 25).

Sequencing reactions were set up with 5 µl of Plasmid DNA (200 ng) and 5 µl sequencing primer (3.2-12.8 pmol), for a total volume of 10µl.

DNA sequence was then determined using a Bigdye terminator version 3.0 cycle sequencing kit (ABI, Applied Biosystems) and an ABI 3700 capillary sequencer for single reads and a ABI 3730 for full length 16S rRNA gene sequencing work.

4.2.7 Data analysis

4.2.7.1 Single Sanger read 16S rRNA gene analysis

4.2.7.1.1 Quality processing of Sanger reads

Sanger sequence reads were quality trimmed and chimeras detected as described in section 3.2.8.1 (page 38).

Sequences were then randomly sampled to 212 reads per breed using a Mersenne Twister algorithm implemented in the random module in python (Makoto and Takuji 1998). This allowed standardization of sampling effort and reduced bias in the subsequent comparative data analysis between breeds.

4.2.7.1.2 OTU determination and statistics

Operational taxonomic units (OTUs) were determined as described in section 3.2.8.2 (page 39).

4.2.7.1.3 Taxonomic assignments

Reads were assigned to taxa as described in section 3.2.8.3 (page 39).

4.2.7.2 Full length 16S rRNA gene analysis

4.2.7.2.1 Sequence assembly

Full length sequences were assembled using DNAbaser (HeracleSoftware) into a contig. Each sequence was checked manually for ambiguous bases. Vector primers D88 and E94 (see section 3.2.3) were removed and the alignment/orientation relative to *E. coli* reference sequence was assessed using ClustalW (Thompson, Higgins et al. 1994).

4.2.7.2.2 Chimera analysis

Full length sequences were checked for chimeras using the online version of Pintail (Ashelford, Chuzhanova et al. 2005). Query sequences were BLAST searched against the NR database at the NCBI. The sequence with the best BLAST score to the query sequence was chosen as its nearest neighbour. The query and nearest neighbour were analysed for observed vs expected nucleotide differences across the length of the 16S rRNA gene molecule. Sections of query sequence with significant deviations from expected difference were potential chimeras and were investigated further to confirm this. If anomalous sections of sequence were substantially different phylogenetically, then a chimera was confirmed. Finally the nearest neighbour was also analysed with the above method to confirm that it was not a chimera. To do this it was checked against a BLAST hit produced by a different research group with at least a 2% nucleotide identity difference.

4.2.7.2.3 Phylogenetic analysis

Full length sequences were aligned using the SINA aligner at the SILVA website (Pruesse, Quast et al. 2007). This produced a high quality alignment based on the SIVA curated SEED alignment of reference 16S rRNA gene sequences. The alignment was further improved by manual editing in ARB's alignment editor using 16S rRNA secondary structure models (Ludwig, Strunk et al. 2004).

The aligned sequences were then added to the RAxML (randomized axelerated maximum likelihood) reconstructed tree of the living tree dataset (Yarza, Richter et al. 2008) using ARB's Parsimony tool. Highly variable bases in the alignment were masked out with a positional variability filter supplied with the Silva database release version 102 (variability values 1-7). This was based on a consensus sequence of 391,116 bacteria taxa.

4.1 Results and discussion

4.1.1 Single Sanger read 16S rRNA gene analysis

A total of 772 high quality reads were Sanger sequenced from 16S rRNA gene libraries derived from cocker spaniel, Labrador retriever and Papillon faecal DNA pools. These sequences were checked for chimeras using pintail software, and 16 (2.1%) were found and removed from analysis yielding 756 sequences (cocker spaniels reads: 244 reads, Labrador: 300 reads, Papillons:212 reads). Sequences were then sampled to 212 reads per breed to normalize sampling effort to the lowest yielding library.

Sequences were then analysed using the RDP classifier version 2.2 with a 30% genus level bootstrap filter. A 30% bootstrap filter was used for the reasons outlined in section 5.4.1 on page 91. This taxonomically classified the reads from the phylum down to the genus level (Table 5). At the phylum level taxa were represented by the *Firmicutes* at 60.51%, *Fusobacteria* 19.26%, *Bacteroidetes* 15.89%, *Tenericutes* 2.41%, *Proteobacteria* 1.61% and *Deferribacteres* 0.32%. The Labrador retriever library had a substantially larger proportion of *Firmicutes* (72.26%) compared to the cocker spaniels (51.43%) and Papillons (56.40%). This was largely due to the genera *Lactobacillus* and *Allobaculum*. *Allobaculum* was first defined with the isolation of *Allobaculum stercoricanis* in canine faeces by the work of Greetham (Greetham, Gibson et al. 2004). The Labrador retriever library also had approximately half the number of the *Fusobacteria* (10.89%) found in the others (cocker spaniels (21.90%), Papillons (24.64%). The cocker spaniel library had almost twice as many *Bacteroidetes* (cocker spaniels: 22.38%, Labrador retriever: 12.87%, Papillons: 12.32%)). This correlated with the presence of both *Bacteroides* and *Prevotella*, whereas the other libraries had only one

or the other in substantial numbers. The Papillon library had almost double the number of *Faecalibacterium* than other libraries.

Comparing the composition of these breed libraries with the miniature Schnauzer library described in chapter 3 is hampered by the difference in sampling effort (141 reads of the MS compared with 212 reads for each of the remaining libraries). However, at the phylum level the taxa composition of the MS library was broadly consistent with the COS, L and PA libraries, being most similar the COS library. As with the COS and PA libraries the *Bacteroidetes* were largely represented by the *Bacteroides*. The Fusobacteria and *Proteobacteria* were similarly dominated by the *Fusobacterium* and *Sutterlla* genera as was observed in the COS, L and PA libraries. Within the Firmicutes the MS library diverged from the other libraries with an increased level of *Phascolarctobacterium* (17.14% of reads) as compared to the COS (7.62% of reads), L (8.42% of reads) and PA (9.48% of reads) libraries. However, the next most frequent genera of the *Firmicute*s were *Coprococcus*, *Faecalibacterium*, and *Blautia* as obeserved in the COS, L and PA libraries.

Table 5.Taxonomic assignment of breed library reads using the RDP classifier.

Phyla (bold), genera	Breed Library COS[1]	L[2]	PA[3]	Total
Bacteroidetes	**47**	**26**	**26**	**99**
Anaerophaga	*0*	*1*	*0*	*1*
Bacteroides	*28*	*3*	*20*	*51*
Paraprevotella	*0*	*0*	*1*	*1*
Prevotella	*15*	*19*	*5*	*39*
Pseudosphingobacterium	*1*	*1*	*0*	*2*
Tannerella	*3*	*2*	*0*	*5*
Deferribacteres	**0**	**2**	**0**	**2**
Mucispirillum	*0*	*2*	*0*	*2*
Firmicutes	**108**	**150**	**119**	**377**
Acetanaerobacterium	*1*	*3*	*0*	*4*
Allisonella	*0*	*0*	*1*	*1*
Allobaculum	*6*	*18*	*2*	*26*
Anaerofilum	*3*	*1*	*0*	*4*
Anaerovorax	*0*	*1*	*1*	*2*
Blautia	*11*	*13*	*11*	*35*
Bulleidia	*3*	*0*	*1*	*4*
Catenibacterium	*0*	*2*	*2*	*4*
Clostridium	*4*	*9*	*2*	*15*
Coprobacillus	*4*	*0*	*1*	*5*
Coprococcus	*19*	*26*	*18*	*63*
Dorea	*1*	*0*	*3*	*4*
Eubacterium	*4*	*0*	*2*	*6*
Faecalibacterium	*20*	*19*	*37*	*76*
Hespellia	*0*	*1*	*0*	*1*
Holdemania	*2*	*0*	*0*	*2*
Lactobacillus	*2*	*17*	*2*	*21*
Megamonas	*7*	*1*	*6*	*14*
Oribacterium	*2*	*0*	*0*	*2*
Oscillibacter	*1*	*4*	*1*	*6*
Papillibacter	*0*	*0*	*1*	*1*
Peptostreptococcus	*0*	*4*	*5*	*9*
Phascolarctobacterium	*16*	*17*	*20*	*53*
Robinsoniella	*0*	*1*	*0*	*1*
Roseburia	*1*	*0*	*0*	*1*
Sporacetigenium	*1*	*7*	*1*	*9*
Streptococcus	*0*	*3*	*1*	*4*
Syntrophococcus	*0*	*1*	*0*	*1*
Tepidibacter	*0*	*1*	*1*	*2*
Thermotalea	*0*	*1*	*0*	*1*

Fusobacteria	**46**	**22**	**52**	**120**
Cetobacterium	*7*	*2*	*7*	*16*
Fusobacterium	*38*	*20*	*44*	*102*
Propionigenium	*1*	*0*	*1*	*2*
Proteobacteria	**6**	**1**	**3**	**10**
Anaerobiospirillum	*2*	*0*	*0*	*2*
Sutterella	*3*	*1*	*3*	*7*
Wautersia	*1*	*0*	*0*	*1*
Tenericutes	**3**	**1**	**11**	**15**
Anaeroplasma	*3*	*1*	*11*	*15*
Grand Total	**210**	**202**	**211**	**623**

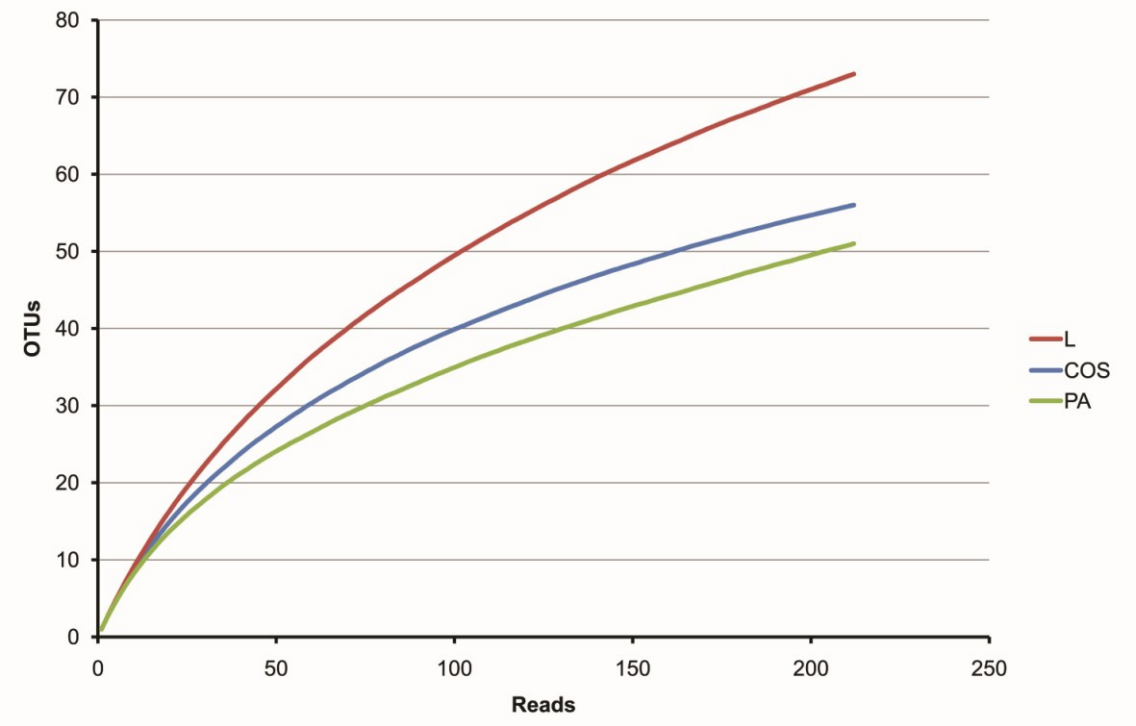
Reads with genus bootstrap values below 30% were excluded. Genera numbers are proportionally coloured from red to yellow to green based on the lowest and highest genera abundance observed.

1. Cocker spaniel. 2. Labrador. 3. Papillion.

OTUs at the 97% nucleotide identity level were determined for the dataset. Across the total dataset 119 OTUs were observed. Within the cocker spaniel library 56 OTUs were observed with 73.8 predicted (Chao1 95%, lower confidence interval (CI): 62.3, upper CI: 105.8). The Labrador retriever library contained 73 observed OTUs and 110.4 predicted (Chao1 95%, lower CI: 88.9, upper CI: 161). Finally, 51 OTUs were observed within the Papillon library 105.2 predicted (Chao1 95%, lower CI: 70.4, upper CI: 202.6).

Rarefaction curves for the three libraries are shown in (Figure 12). The Labrador retriever library had a higher richness than both the cocker spaniel and Papillon libraries. When these data are compared to the miniature Schnauzer library (Figure 9) the cocker spaniel and Papillon richness is similar at the sampling effort of 141 reads.

Figure 12. Rarefaction curves for cocker spaniel, Labrador and Papillon libraries, showing Labradors with increased observed diversity compared to Papillions and cocker spaniels OTUs are defined at the 97% nucleotide identity level.



L, Labrador. COS, Cocker spaniel. PA, Papillion.

It would also be tempting to compare these richness data with the 454 analysis in chapter 6 (Figure 29 and Figure 30). It can be seen in the 454 analysis that the Labrador retriever samples have significantly (Figure 30) higher richness estimates. However, the sampling effort was ~40 times higher and the read length, library creation and sequencing were different.

The pooling of individual sample PCR into breed libraries limits the possibility of investigating the core microbiome hypothesis (Turnbaugh, Hamady et al. 2009). Nevertheless it was observed that only 19 OTUs (16%) were shared across the three libraries. Furthermore, in a pairwise comparisons of the libraries there was not a substantial difference between the number of shared OTUs (COS and L; 28, COS and PA; 25, L and PA; 27).

It was discovered that there was a base deletion in the D88 primer sequence when compared against the *E. coli* priming site sequence (deletion at position 13). This might of lead to a loss in specificity against bacterial 16S rRNA gene template and a reduction in captured diversity. However, it was determined by a pyrosequencing experiment (results not shown) that any loss in diversity was not conspicuous. The pyrosequencing chapters 5 and 6 use a slightly different primer sequence that does not have this deletion.

These results represent a valuable insight into the gut microbiota of 3 dog breeds. The Labrador retrievers were observed to have a substantially richer faecal microbiota with key differences in *Lactobacillus* and *Allobaculum* genera numbers and a depletion of *Fusobacteria*. Cocker spaniel and Labrador retriever libraries had lower richness levels comparable to that seen in the MS library (Chapter 3). The individual sampling and increased

sampling effort of the experiments shown in chapter 6 and 7 provide a greatly expanded exploration of these canine microbiomes.

4.1.2 Full length 16S rRNA gene analysis

The RDP classifier has been used extensively throughout microbial ecology to taxonomically classify high throughput datasets. The classifier uses a 'training' set of reference sequences from the Bergey's taxonomic outline to calibrate the taxonomic identification. The accuracy of the classifier was assessed in the initial publication and in other papers (Wang, Garrity et al. 2007; Claesson, O'Sullivan et al. 2009). However, it is important to independently assess the accuracy of classification of both the Sanger and pyrosequencing sequence data. This is especially important considering the lack of coverage of the canine gut microbiota in the public sequence databases. The clone libraries produced during this project presented the opportunity to test how some of the most abundant pyrosequencing taxa compared to those derived from full length sequences from the same starting source.

Clones from the cocker spaniel, Labrador retriever and Papillons were searched against sequences representative of OTUs identified in the 454 dataset (see chapter 6) using BLAST. The results were then filtered to select clones with >97% nucleotide identity matches to the top 'shared' OTUs of a pre-AmpliconNoise 454 dataset. While a BLAST hit against just the first 250 bases of the full length sequence does not provide evidence of a match for the remainder of the 16S rRNA gene , it is likely these highly abundant 'core microbiome' members are congruent. Seventeen of these clones were then resequenced with a high quality full length 16S rRNA gene sequencing strategy.

Following an improved analysis using AmpliconNoise (see chapter 6) to de-noise the 454 pyroseqeuncing reads, the BLAST hits to these de-noised OTUs were calculated. This resulted in 6 of the full length sequences that originally matched the most widespread OTUs in the previous analysis being relegated to a less widespread OTU but nevertheless shared by ≥50% of dogs. The results of this analysis are shown in Table 6 as an adaptation of Table 10. The RDP classification of the full length sequences and the corresponding 454 sequence BLAST hits was also compared. RDP classifications of the full length Sanger and 454 sequences in Table 6 agreed to the family level for all sequences and to the genus level in 6/11 instances.

Table 6. A summary of BLAST analysis of the most highly shared OTUs (core microbiome members) in the de-noised 454 pyrosequencing dataset Table 10 against the Labrador retriever, cocker spaniel and Papillon clone libraries.

OTU97 no.	% dogs with this OTU	average reads per dog	%total reads in OTU	Cumulative fraction of total reads	Classifier genus associated with OTU	FullLengthSeqID	Matched 454 sequence ID	454 sequence genus classification	% identity	alignment length	mismatches
OTU97_68	100.00%(79)	669.57	7.88%	7.88%	Fusobacterium	P3_D07_clone_RDPFusobacterium	1FUOFXOM01A63T1	Fusobacterium	98.89	180	2
OTU97_112	98.73%(78)	581.18	6.75%	14.63%	Cetobacterium	L1_D09_clone_RDPFusobacterium	1FUOFXOM01A0V34	Cetobacterium	100	188	0
						L1_E07_clone_RDPFusobacterium	1FUOFXOM01A0CK3	Cetobacterium	100	179	0
OTU97_105	97.47%(77)	308.7	3.54%	18.17%	Cetobacterium	P3_E08_clone_RDPFusobacterium	1FUOFXOM01A323U	Cetobacterium	100	179	0
OTU97_108	97.47%(77)	30.45	0.35%	18.52%	Blautia	P5_B09_clone_RDPBlautia	1FUOFXOM01A3VGX	Blautia	100	173	0
OTU97_23	97.47%(77)	69.87	0.80%	19.32%	Sporactigenium	P5_D08_clone_RDPPeptostreptococcus	1FUOFXOM01A3KFQ	Sporactigenium	99.04	208	1
OTU97_4	97.47%(77)	35.42	0.41%	19.72%	Blautia	P3_C02_clone_RDPBlautia	1FUOFXOM01A1LBD	Blautia	100	180	0
OTU97_215	96.20%(76)	288.66	3.27%	22.99%	Cetobacterium	L1_D09_clone_RDPFusobacterium	1FUOFXOM01A026X	Cetobacterium	100	188	0
OTU97_32	96.20%(76)	257.26	2.91%	25.90%	Bacteroides						
OTU97_133	94.94%(75)	51.23	0.57%	26.48%	Blautia						
OTU97_16	94.94%(75)	202.77	2.26%	28.74%	Faecalibacterium	P4_E04_clone_RDPFaecalibacterium	1FUOFXOM01A2OBI	Faecalibacterium	99.06	212	0
						P3_D04_clone_RDPFaecalibacterium	1FUOFXOM01A4QKI	Faecalibacterium	99.06	212	1
OTU97_157	94.94%(75)	21.57	0.24%	28.98%	Blautia/Dorea						
OTU97_19	93.67%(74)	21.69	0.24%	29.22%	LactonifactorCoprococcus						
OTU97_21	93.67%(74)	270.86	2.98%	32.21%	Sutterella						
OTU97_1	93.67%(74)	102.34	1.13%	33.33%	Roseburia						
OTU97_124	93.67%(74)	214.43	2.36%	35.70%	Bacteroides/Parapprevotella	P4_A02_clone_RDPBacteroides	1FUOFXOM01A1NIZ	Bacteroides	98.95	190	2
OTU97_9	92.41%(73)	145.97	1.59%	37.28%	Bacteroides						
OTU97_110	91.14%(72)	10.67	0.11%	37.40%	Roseburia						

The full length sequence ID of the matched clone is shown with a RDP classifier genus suffix. Grey shaded cells indicate OTUs that did not match a full length sequence at ≥97% nucleotide identity.

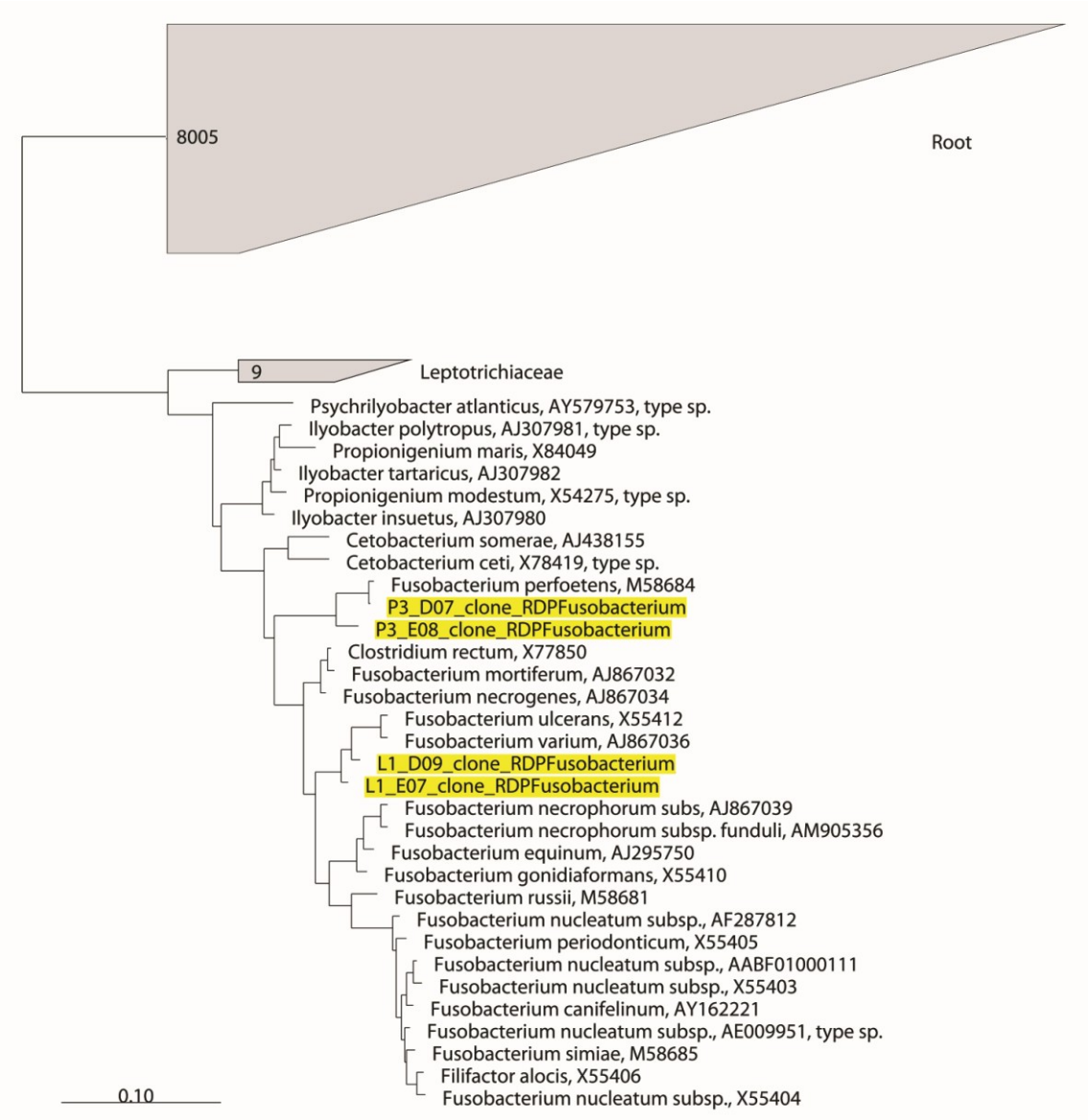
In order to further investigate the classification of these sequences a phylogenetic tree was reconstructed. The full length sequences (including the 6 sequences not shown in *Table 6*) were aligned using the SINA aligner at the SILVA website (Pruesse, Quast et al. 2007). This alignment was then manually curated using ARB's alignment editor and then the sequences were integrated into Living tree using ARB's parsimony tool (Ludwig, Strunk et al. 2004). The living tree is a high quality 16S rRNA gene phylogenetic tree of 8029 type strains and other cultured bacteria (Yarza, Richter et al. 2008). Individual trees shown in Figure 13 through Figure 20 display the clades in which the full length sequences were integrated into the tree.

All full length sequences clustered into clades congruent with their RDP classifier taxonomy at the family level and in most cases down to the genus level. This provided high quality phylogenetic validation of the RDP classifier classifications, albeit a validation limited to the groups with close matches to the small number of full length sequences. Nevertheless, this showed that the 200 nt pyrosequencing reads of the 454 dataset were relatively accurately classified when compared to a full length 16S rRNA gene phylogenetic analysis for some of the most widespread and abundant taxa.

Members of the *Fusobacteria* have been consistently discovered in abundance throughout this project. The substantial presence of these bacteria is a relative novelty compared to other mammalian gut microbiome studies. Although, their presence in the canine gut has been observed previously (Balish, Cleven et al. 1977; Greetham, Giffard et al. 2002; Suchodolski, Camacho et al. 2008). The relationship of the full length sequences with *Fusobacteriaceae* culture strains is shown in Figure 13. Two groups were apparent, one associated with

Fusobacterium perfoetens and a second associated with *Fusobacterium ulcerans* and *Fusobacterium varium*. *Fusobacterium varium* was isolated in Greetham's canine gut microbiota culture work (Greetham, Giffard et al. 2002).

Figure 13. RAXML phylogenetic tree showing the Fusobacteriaceae *clade within the living tree. Integrated full length sequence clones are highlighted yellow and have trailing associated RDP classifier genus identifications. All clones are associated with 'Core microbiome' members detailed in Table 6.*



Clone P3_D07 clustered closely to *Fusobacterium perfoetens*, while the others clustered more distantly to their nearest culture strain. Clones with associated 454 sequence read OTUs classified as *Cetobacterium* clustered with other members of the *Fusobacterium.*

Figure 14 shows how the single *Sutterella* classified full length sequence clustered within the *Alcaligenaceae* clade. This single clone did not cluster closely to *Sutterella stercoricanis* which was also isolated and defined in dogs suggesting that it was novel (Greetham, Collins et al. 2004).

Figure 15 shows how the *Bacteroides* classified full length sequences clustered within the *Bacteroidales* clade.

Figure 14. RAXML phylogenetic tree showing the Betaproteobacteria *clade within the living tree. Integrated full length sequence clones are highlighted yellow and have trailing associated RDP classifier genus identifications.*

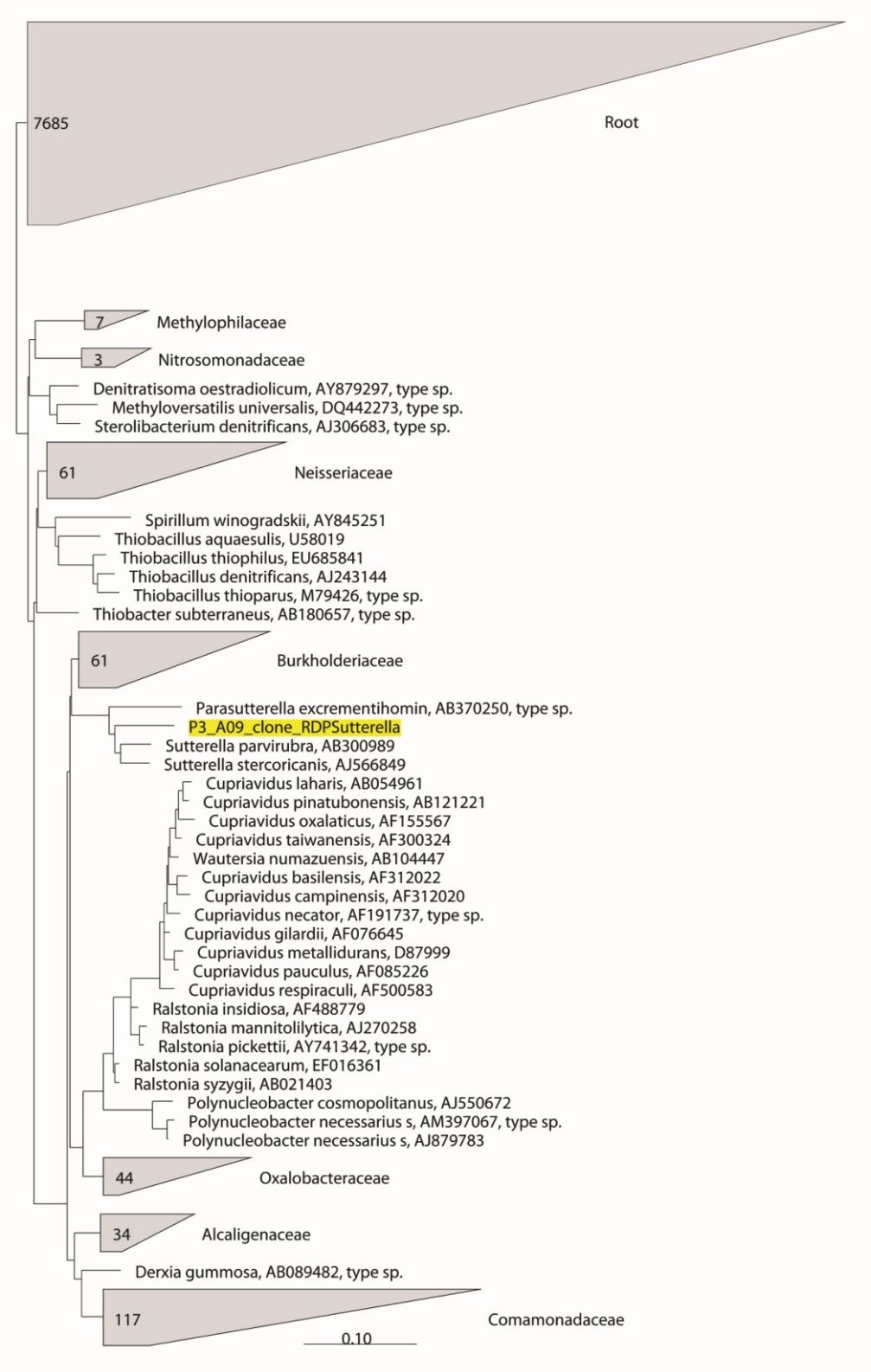


Figure 15. RAXML phylogenetic tree showing the Bacteroidales *clade within the living tree. Integrated full length sequence clones are highlighted yellow and have trailing associated RDP classifier genus identifications. Clones P4_A02 is associated with 'Core microbiome' members detailed in Table 6.*

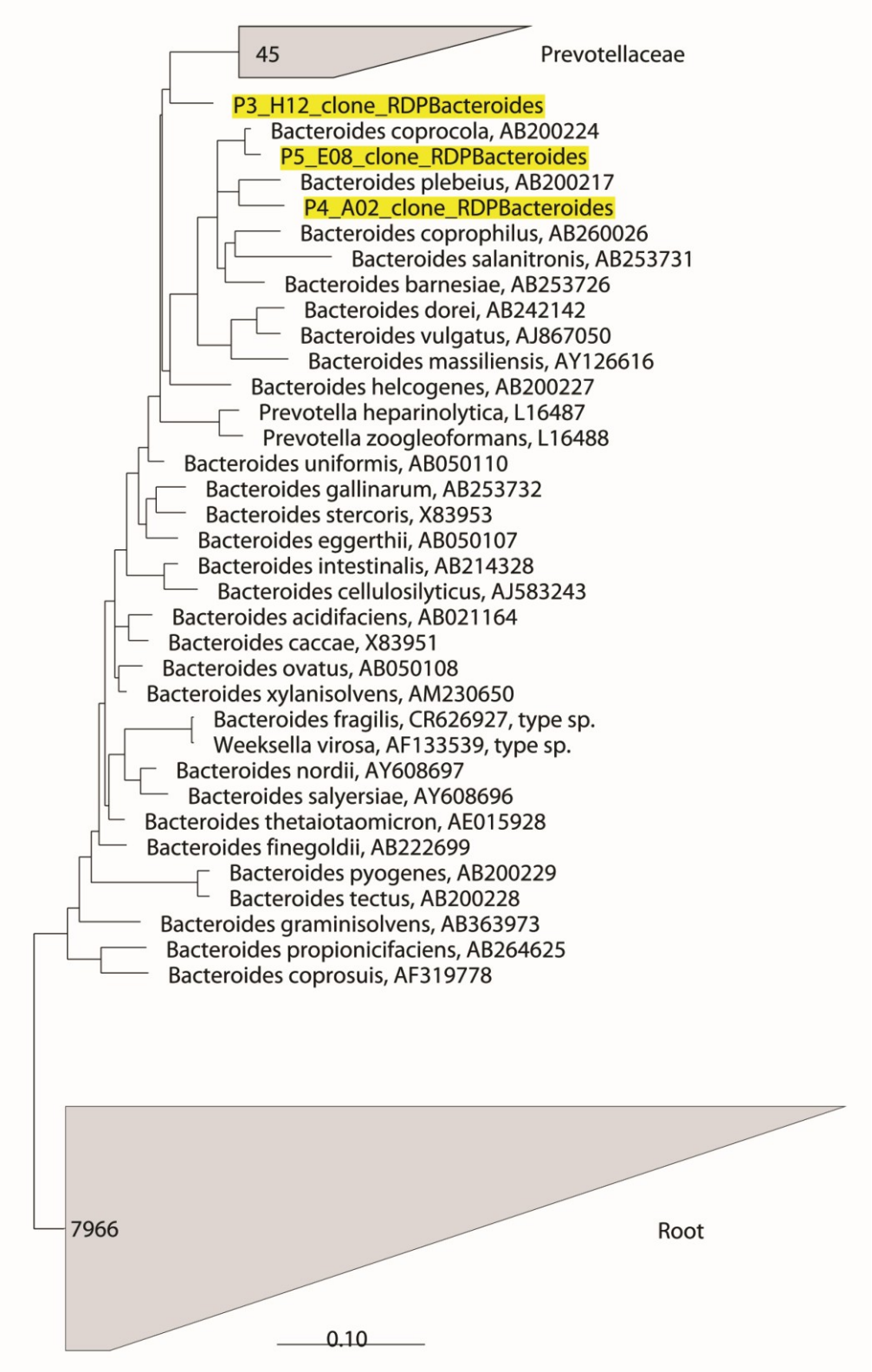


Figure 16 shows the *Lachnospiraceae*, *Ruminococcaceae*, *Peptostreptococcaceae* and *Veillonellaceae* clades within the *Clostridia* which incorporated full length sequences. Figure 17 through Figure 20 show these clades in detail.

Within the *Lachnospiraceae* the *Blautia* classified full length sequences clustered with *Blautia schinkii* and *Blautia producta* (Figure 17).

Within the *Ruminococcaceae* the *Subdoligranulum* and *Faecalibacterium* classified full length sequences clustered with *Subdoligranulum variabile* and *Faecalibacterium prausnitzi* respectively (Figure 18).

Among the *Peptostreptococcaceae* the *Peptostreptococcus* classified full length sequence clustered with *Clostridium hiranonis* (Figure 19).

Finally, within the *Veillonellaceae*, *Blautia* classified full length sequences clustered with *Megamonas funiformis* (Figure 20).

Figure 16. RAXML phylogenetic tree showing the Clostridia *clade within the living tree. Groups containing integrated full length sequence clones are highlighted yellow. Each of these four groups is shown in detail in Figure 17 through Figure 20.*

Figure 17. RAXML phylogenetic tree showing the Lachnospiraceae *clade within the living tree. Integrated full length sequence clones are highlighted yellow and have trailing associated RDP classifier genus identifications. Clone L1_A12 is associated with 'Core microbiome' members detailed in Table 6.*

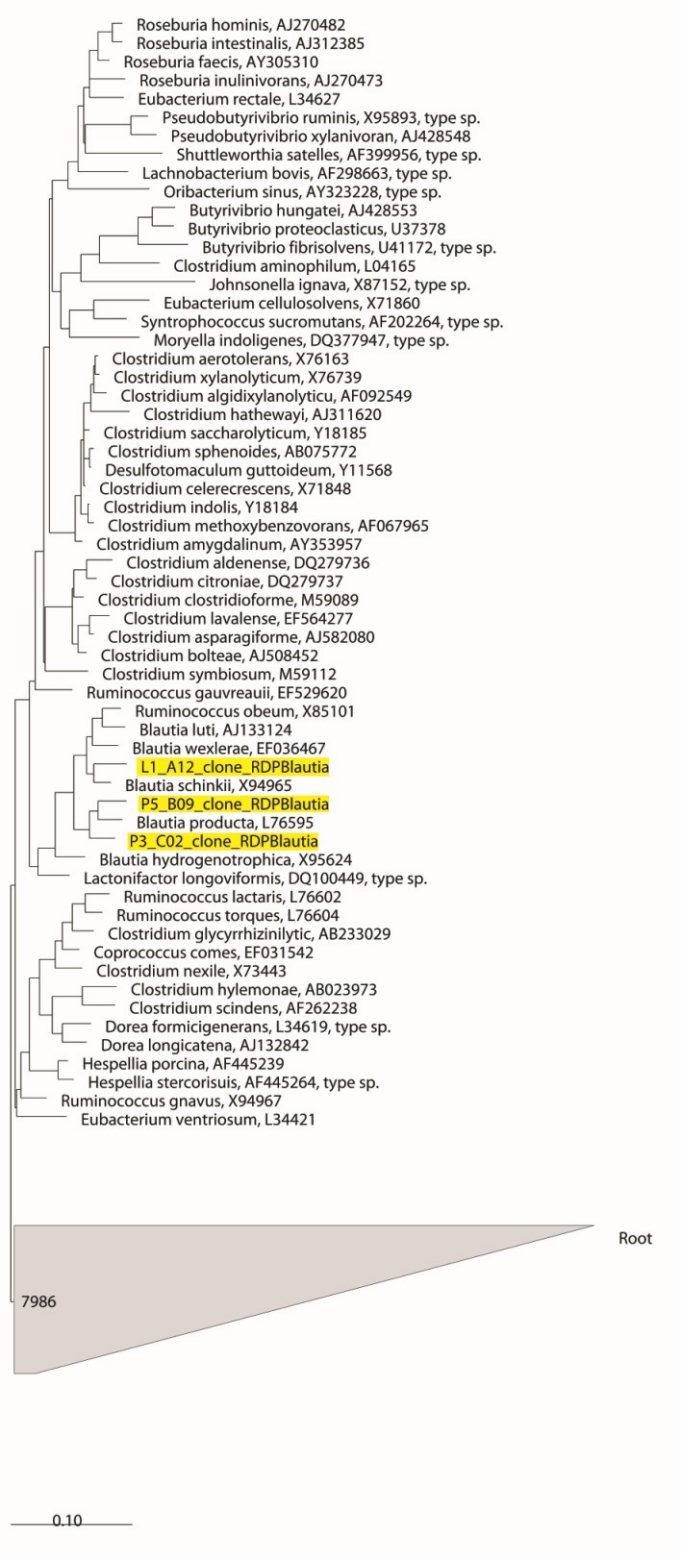


Figure 18. RAXML phylogenetic tree showing the Ruminococcaceae *clade within the living tree. Integrated full length sequence clones are highlighted yellow and have trailing associated RDP classifier genus identifications are associated with 'Core microbiome' members detailed in Table 6.*

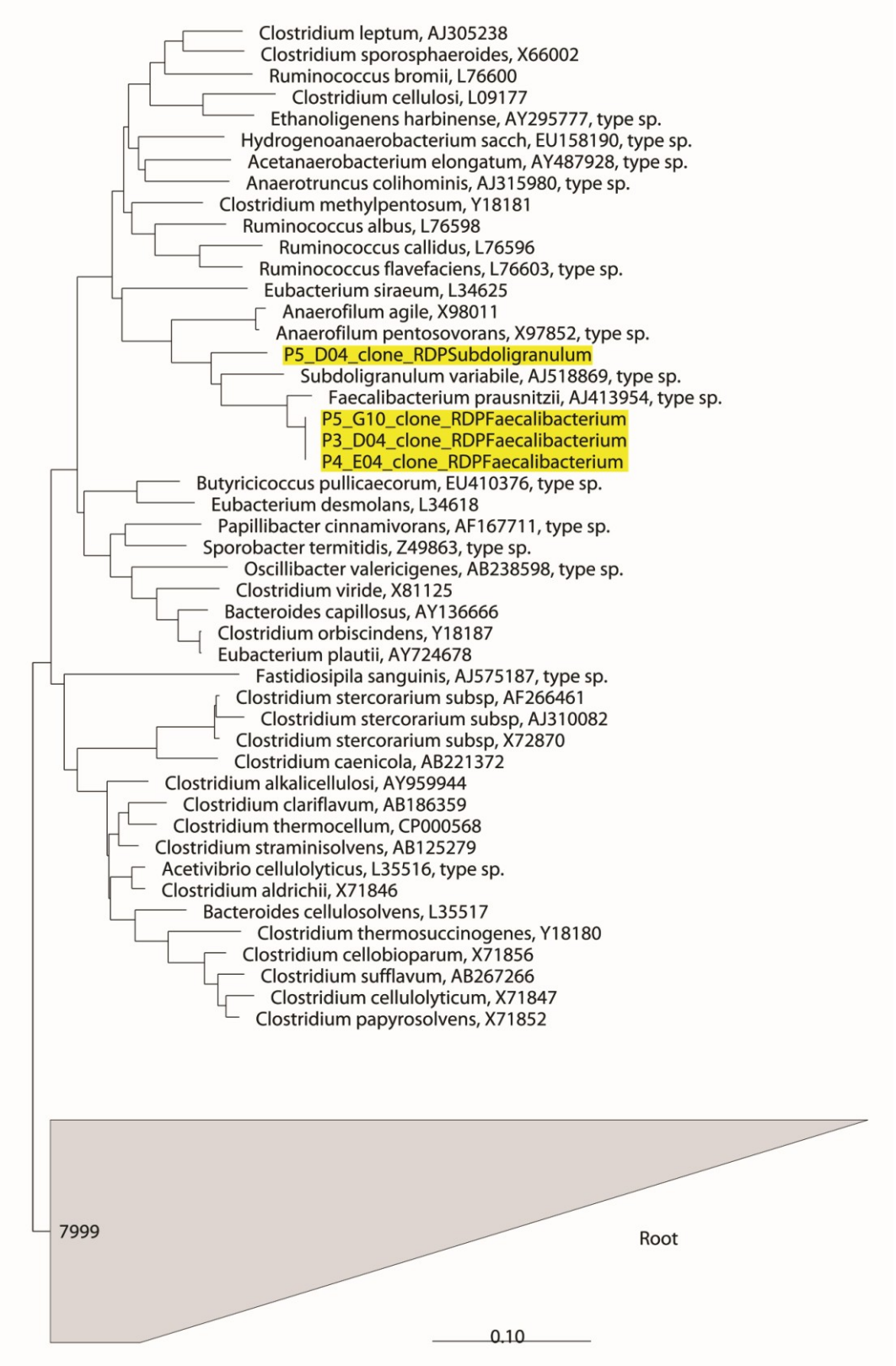


Figure 19. RAXML phylogenetic tree showing the Peptostreptococcaceae *clade within the living tree. Integrated full length sequence clones are highlighted yellow and have trailing associated RDP classifier genus identifications. Clone P5_D08 is associated with 'Core microbiome' members detailed in Table 6.*

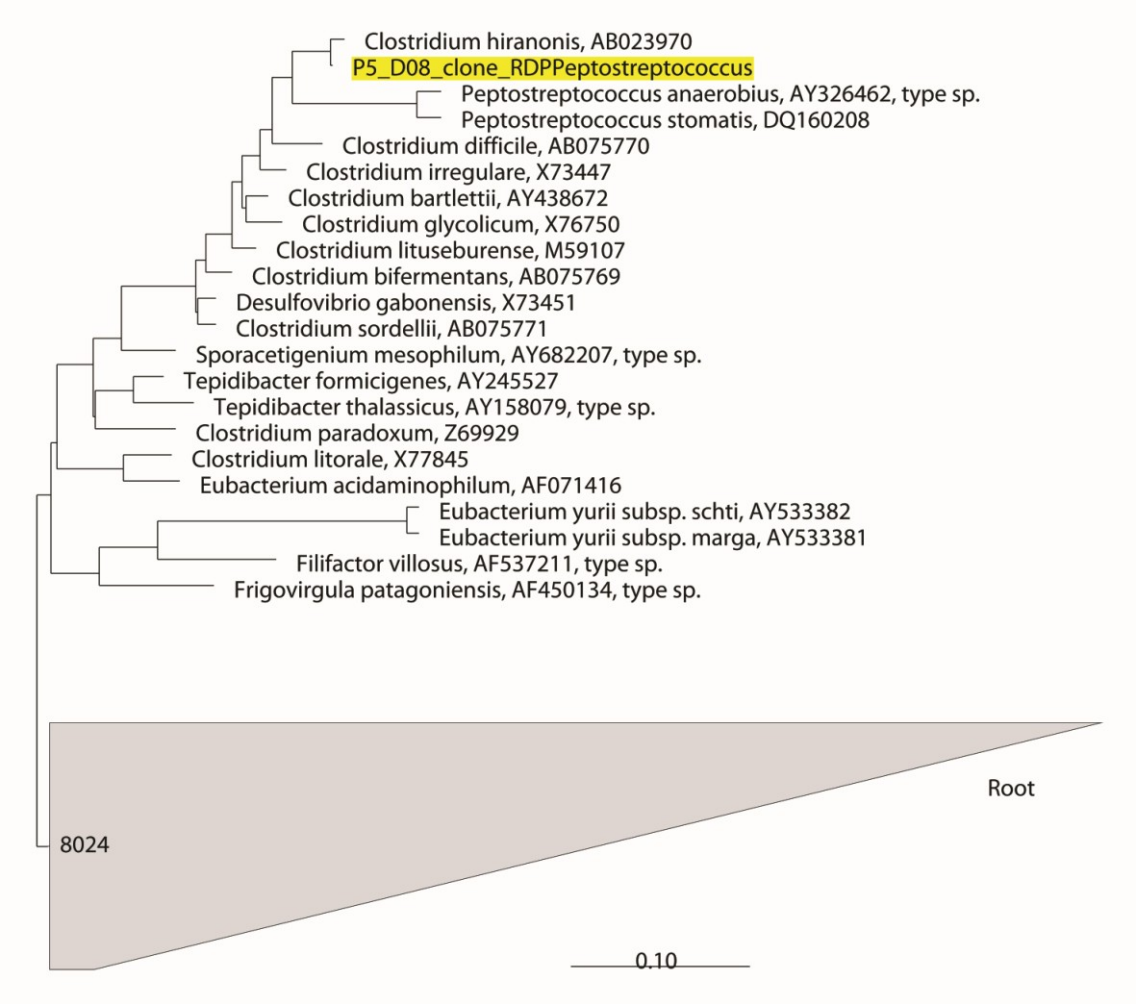


Figure 20. RAXML phylogenetic tree showing the Veillonellaceae *clade within the living tree. Integrated Full length sequence clones are highlighted yellow and have trailing associated RDP classifier genus identifications.*

No clones had an identical match to culture representatives. However, the culture representatives with the least distance to clones were *Clostridium hiranonis*, *Fusobacterium perfoetens* and *Bacteroides coprocola*. It is difficult to determine the biological significance of how the clones cluster with the culture strains within the living tree. While 16S rRNA gene sequence is the best phylogenetic marker available to investigators, in isolation it does not provide enough resolution to definitely and consistently link closely matching sequences to a phenotype.

This analysis of taxa assignment of full length 16S rRNA gene sequence using the RDP classifier and phylogenetic trees as compared to the RDP classification of pyrosequencing reads, allowed an insight into the accuracy of the taxonomic classifications of chapters 5 and 6. While the results are limited to the some of the more abundant and widespread OTUs of these datasets it is reassuring that there were no deviations at the family level and only a few at the genus level. This indicates that the V1 and V2 regions sequenced in the pyrosequencing work provide an adequate amount of information to assign taxonomy reasonably accurately. Although, deviations such as those associated with *Cetobacterium* and *Fusobacterium* demonstrate the limit of this accuracy between full length and partial length 16S rRNA gene analysis.

The creation of clone libraries from the same starting source as of that used in the 454 work allowed this validation to be conducted. It is unlikely that investigators will do this in future due to the limitations of clone library sequencing. In addition the read length of high throughput technologies is increasing. Current Roche 454 FLX titanium pyrosequencing technology can sequence ~500 nucleotides and future improvements of read lengths of ~1000

nucleotides are reported to be in development. With read lengths of 1000 nucleotides and beyond the issue of taxonomic assignment with short 16S rRNA gene read lengths will be eliminated.

CHAPTER FIVE: Pyrosequencing the Canine Faecal Microbiome: Breadth and Depth of Biodiversity

Daniel Hand, Corrin Wallis and Charles W Penn

5.1 Abstract

Mammalian intestinal microbiota remain poorly characterised despite decades of interest and investigation by culture-based and other long-established methodologies. Using high throughput sequencing technology we now report a detailed analysis of the canine faecal microbiome. The study group of animals comprised eleven healthy adult miniature Schnauzer dogs of mixed sex and age, all housed in kennel and pen accommodation on the same premises with similar although not identical feeding and exercise regimes. DNA was extracted from faecal specimens and subjected to PCR amplification of 16S rRNA gene sequence that included variable regions V1 and V2. Barcoded amplicons were sequenced by Roche-454 FLX high-throughput pyrosequencing. Sequences were assigned to taxa using the Ribosomal Database Project Bayesian classifier, which revealed the dominance of *Fusobacterium* and *Bacteroidetes* phyla. Differences between animals in the proportions of different taxa were clear, even among littermates. Diversity of the microbiota was also assessed by assignment of sequence reads into operational taxonomic units (OTUs) at the level of 97% sequence identity. The OTU data were then subjected to rarefaction analysis and determination of Chao1 richness estimates. The data indicated that faecal specimens contained between 500 and 1500 OTUs at the level of 97% identity. Sequences indicative of the presence of some genera that include known pathogens were detected at low levels. The

data described comprise a unique survey of the faecal microbiome in an environmentally and genetically closely related group of animals. The variability seen is not strongly supportive of the concept of a 'core microbiome', at least as determined by analysis of faecal material.

5.2 AUTHOR SUMMARY

Until recently, it was not possible to analyse comprehensively the complex bacterial community or 'microbiota' that flourishes in the digestive tracts of animals. This was because culture and characterisation in the laboratory of the hundreds of species present presented enormous technical challenges and a very substantial workload. In contrast, the determination of 'signature' DNA sequences characteristic of each bacterial species present, has made it possible to perform these analyses thoroughly and rigorously. Amplification by PCR and molecular cloning and sequencing of the DNA sequences encoding 16S ribosomal RNA was the first molecular approach adopted, but remained laborious and expensive. Now it is possible to perform this analysis far faster and more cheaply using high throughput DNA 'pyrosequencing'. We have exploited this new technology in one of the most detailed studies yet of the 'normal' faecal microbiota in healthy adult dogs. We report evidence for the presence and relative abundance of hundreds of different bacterial genera, many of them never before cultured or characterised. We have observed substantial qualitative and quantitative variations in the microbiota even within the same breed of dog. These findings have great potential in enhancing understanding of the physiological and nutritional contribution of the microbiota to animal health.

5.3 INTRODUCTION

The intestinal microbiome can be defined as the totality of microbial species that inhabit the digestive tract. It is increasingly recognised as a major contributor to the digestion and utilisation of foods in the gastrointestinal tract, and a key factor in digestion and nutrition, development, immune function and other aspects of host physiology that contribute to health and wellbeing (Flint 2007; Stecher 2008; Garrett 2010).

This microbial community is very large, diverse and complex, comprising at least hundreds, perhaps thousands of interdependent and/or competing species, and is generally poorly characterised (Eckburg, Bik et al. 2005; Ley, Hamady et al. 2008). The most diverse and abundant component of the digestive tract microbiota in monogastric mammals is the community associated with the contents of the large intestine. This microbial community is usually dominated numerically by strictly anaerobic species, typically including members of the *bacteroidetes*, *bifidobacteria* and *clostridia* (Savage 1977; Gill, Pop et al. 2006). These groups include numerous diverse and often uncultured genera and species, seldom of any great significance as pathogens. *Firmicutes* (low G+C Gram-positives) including anaerobic cocci are often also abundant (Savage 1977). The better-known *proteobacteria*, including many of the medically important pathogenic genera, are generally orders of magnitude less abundant apart from during overt infection.

The complexity of this community lies not only in its richness (the breadth of phylogenetic diversity) but also in its quantitative aspects such as abundance of different phylogenetic

groups, usefully defined as evenness. The population is extremely large at about 10^{11} per gram of intestinal contents, and it is often claimed that in humans the total number of bacterial cells is about ten times that of the host. Furthermore, the abundance distribution of individual species covers many orders of magnitude. The reasons for lack of knowledge and understanding of this microbial community are largely technical. Many species are essentially uncharacterised, due to the difficulty of culturing them routinely in the laboratory (Greetham, Giffard et al. 2002; Green and Keller 2006), although efforts are continuing to culture new species, with considerable success (Duncan, Louis et al. 2007). Often it is claimed that many are unculturable, but failure to culture may simply reflect the diversity and complexity of media and growth conditions required to recover the majority of species in pure culture (hence uncultured is a preferred term). This is particularly relevant for microbial populations from non-human sources, cultured on agars frequently developed for growth and isolation of human microbial taxa. Nevertheless the growth and characterisation of hundreds of often fastidious and highly diverse species has in practice made comprehensive and systematic culture-based studies of mammalian gut microbiota an unattainable goal.

In contrast, more recent efforts have exploited the methodology of PCR amplification and plasmid cloning and dideoxy chain-termination sequencing of 16S rRNA gene amplicons, pioneered by Pace and others (Pace 1997) in studies of environmental microbial communities and the basis of Venter's well known studies of oceanic microbiota collected from the Sargasso sea (Venter, Remington et al. 2004). However, while it gets round problems with culturability, this methodology remains laborious, expensive and limited in scope, for example in relation to quantification of different members of complex communities. Hence the recent development of 'next generation' or more correctly high throughput DNA

sequencing, specifically Roche-454 pyrosequencing, creates a major new opportunity. Direct deep sequencing of the 16S rRNA gene amplicon pool now enables coverage increased by orders of magnitude, with the ability to quantify different microbial groups directly based on the numbers of copies of their signature sequences obtained (Sogin, Morrison et al. 2006). Typically, a million or more sequence reads can be obtained per instrument run, and sequence barcodes can be used to enable reads from numerous different samples to be binned and analysed thus enabling simultaneous processing of large numbers of samples.

A number of recently published studies now describe the exploitation of this technology to analyse the gut microbiota of humans and other animal species (Dethlefsen, Huse et al. 2008; Suchodolski, Dowd et al. 2009; Turnbaugh, Hamady et al. 2009). Often however there is little insight from these publications into the variability within and between individuals, and hence the data generally represent snapshots of a landscape that may be intrinsically very diverse as well as subject to great variation over time. In humans, where interest in the microbiota is intense as its importance in dietary processing and health is increasingly recognised (Turnbaugh, Hamady et al. 2009), the complexity of rigorous studies is compounded by genetic diversity and the difficulty of defining and controlling dietary, behavioural, environmental and other variables. In contrast, studies in defined animal populations offer opportunities to control these variables. Dogs for example include inbred lines represented by different breeds. Animals housed and fed under similar and controlled regimes are not subject to anything like the confounding variables of human study populations.

We now describe the analysis of faecal microbiota in a group of miniature Schnauzer dogs housed on the same site and with known dietary intakes and genetic relatedness, and illustrate the potential for such an approach to generate fundamental new insights into gut microbiology. Knowledge of canine intestinal microbiota is limited. Several studies investigating canine microbiota using culture-independent methods have now been published (Simpson, Martineau et al. 2002; Greetham 2003; Beasley, Manninen et al. 2006; Suchodolski, Camacho et al. 2008; Suchodolski, Dowd et al. 2009; Suchodolski, Xenoulis et al. 2009). However, the canine gut microbiota is yet to be systematically characterised at the species level, and much of our knowledge based on culture methods dates back to the 1970's and is difficult to relate to the newer culture-independent data (Balish, Cleven et al. 1977).

5.4 RESULTS AND DISCUSSION

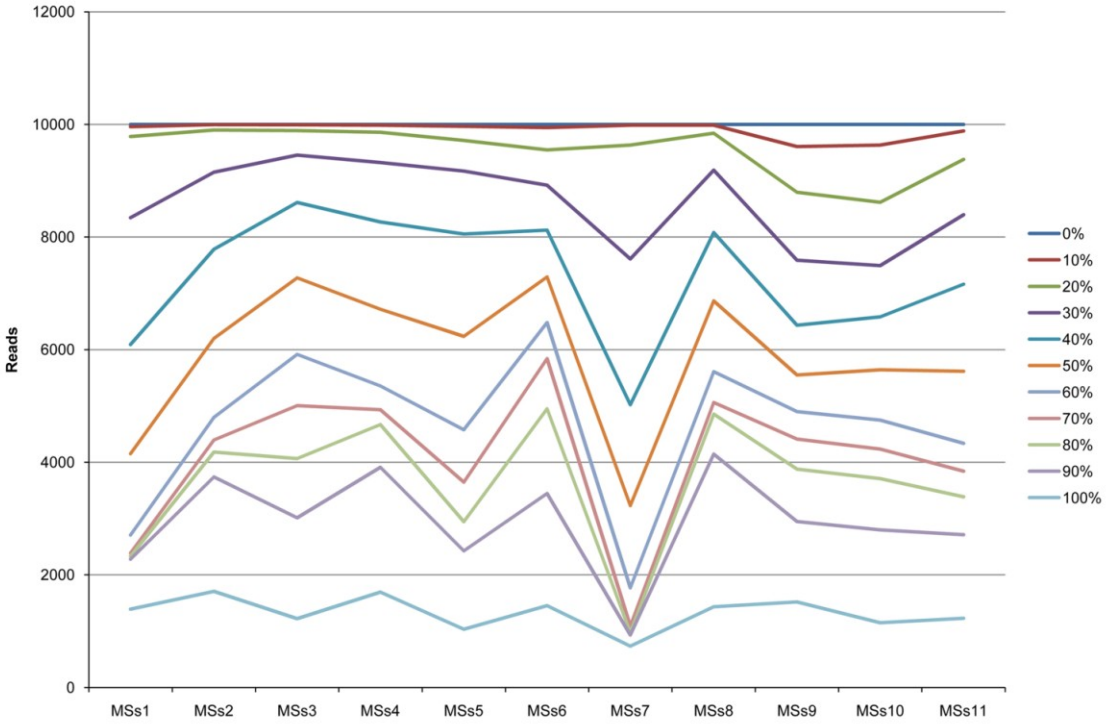
5.4.1 Data acquisition and analysis

Pyrosequencing by Roche-454 GS FLX of faecal rDNA amplicons from 11 miniature Schnauzer dogs, including a replicate of the MSs1 sample, yielded 247501 reads representing 56.7 Mb of sequence. The average read length was 229 bp. After the raw sequence data were filtered for poor quality reads, the average yield per barcode was 17899 ± SD 1434 reads.

To make an initial assessment of the nature of the microbiota in each animal, the reads were analysed using the RDP Bayesian classifier. This method was chosen to take advantage of the straightforward RDP pipeline, fast analysis and widely used, high quality sequence database.

Identification of reads to the genus level with strict RDP classifier bootstrap values resulted in a large and uneven drop in resulting read numbers across the dogs (Figure 21). Dogs MSs1 and MSs7 in particular were subject to disproportionate reductions in proportion of reads that could be classified with increasing filtering stringency based on the bootstrap score. This may in part be due to an under-representation of canine microbiome sequences in the RDP classifier training dataset. Detailed scrutiny shows that members of the Fusobacteria phylum are mainly responsible; specifically the reads assigned to the genera Fusobacterium, Cetobacterium and Ilyobacter were disproportionately associated with classifier scores below 50%. MSs1 and MSs7 had more Fusobacteria reads than other dogs leading to the disproportionate drop observed. A bootstrap score of 50% has been recommended for genus level identification (Claesson, O'Sullivan et al. 2009). However, we decided to reduce this to 30% to maintain a more even representation of reads across the group of dogs, accepting the trade off that a minority of classifications will be incorrect.

Figure 21. Effect of increasing RDP classifier genus bootstrap score stringency on the number of reads passed per dog, justifying a 30% genus score filtered for taxa classifications.



5.4.2 Composition of microbial communities

First we combined RDP classified data from all eleven dogs to generate an overview of the microbiota composition, as represented by the read sequences. In the following discussion we assume that the presence of sequences reflects the presence of the corresponding organisms in the faecal biomass, with the caveat that DNA from dead bacteria or even potentially from ingested foods or other sources may persist and be detectable in the total faecal DNA fraction. There is very little literature available on which to estimate the stability of dietary DNA during passage through the digestive tract, but there is some evidence of persistence in the GI tract of recombinant (GM) DNA from dietary sources (Martin-Orue, O'Donnell et al. 2002; Wiedemann, Lutz et al. 2006).

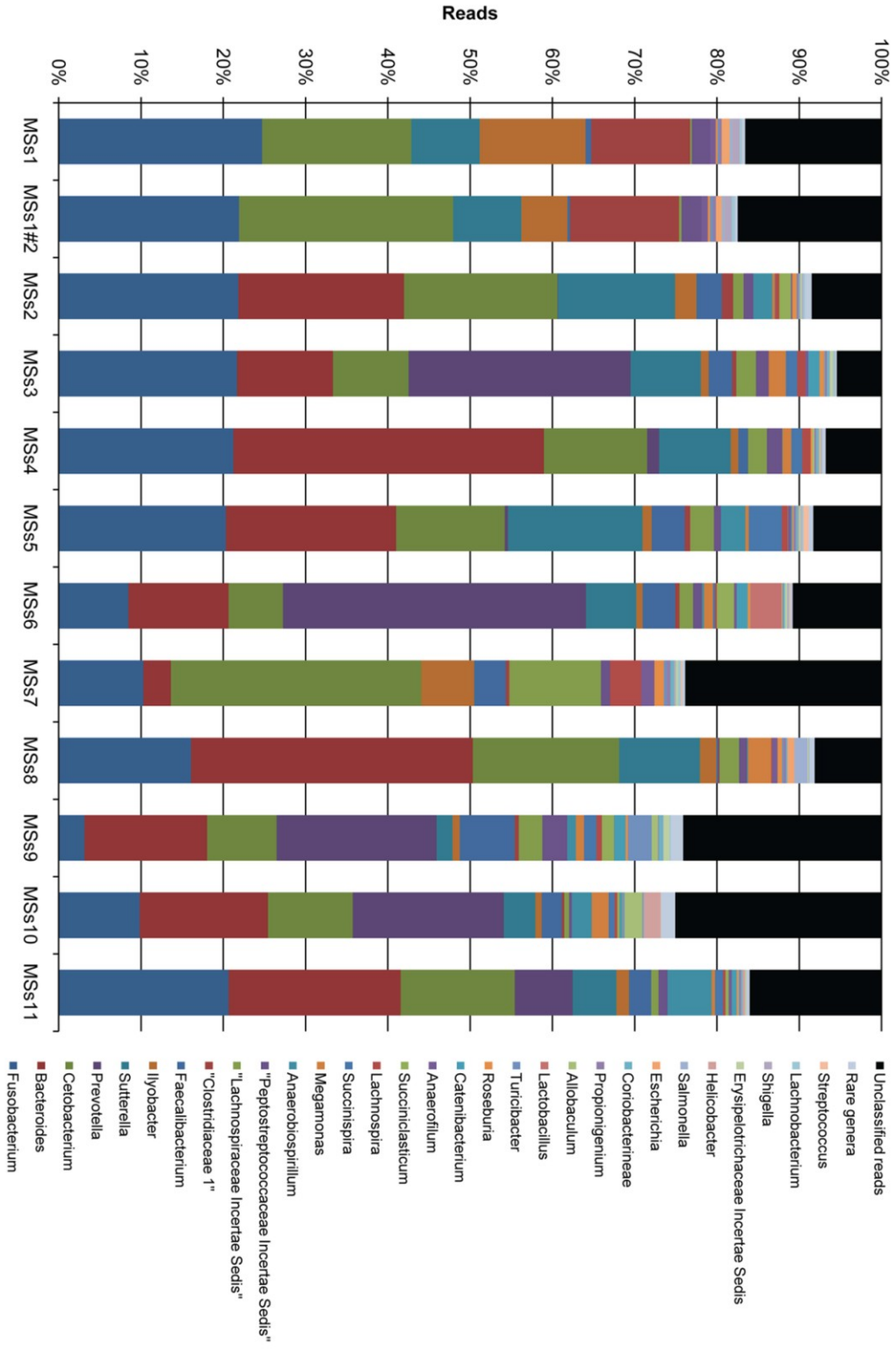
We discovered that at the phylum level, the combined community was dominated by the *Fusobacteria* (42.3% of reads) followed by the *Bacteroidetes* (30.9%), *Firmicutes* (15.2%), *Proteobacteria* (11.3%), *Actinobacteria* (0.3%) and various other phyla at lower abundances. The clear abundance of *Fusobacteria* is relatively unusual for a mammalian gut community (Turnbaugh, Ley et al. 2006; Dowd, Sun et al. 2008; Ley, Hamady et al. 2008; Ritchie, Burke et al. 2010) and confirms findings of others (Suchodolski, Camacho et al. 2008).

Analyses of the faecal microbiota from individual animals revealed considerable divergence in abundances of the major taxonomic groups, analysed at the genus level (Figure 22). In most of the animals the five most abundant genera represented approximately 60-80% of the bacteria present. However even among the most abundant genera the data were highly

variable between animals; only the prominent members of the *Fusobacteria* phylum (*Fusobacterium*, *Cetobacterium* and *Ilyobacterium*) were among the most abundant groups in all animals. None of the other most abundant genera, the *Bacteroides*, *Prevotella* and *Sutterella*, were highly abundant in every one of the dogs tested. This observation does not support the concept of a 'core microbiome' (Turnbaugh, Hamady et al. 2009) in which the major constituents of the community will be more or less universally present at comparable levels in different individual hosts. Among prominent genera, the combined *Prevotella* and *Bacteroide*s (phylum *Bacteroidetes*) abundances tended to be inversely related to phylum *Fusobacteria* abundance. We hypothesise that this distribution may relate to 'competition' for the same niche by these groups of bacteria.

There is considerable variation between animals in the diversity of the less abundant genera. It is apparent from Figure 22 that, for example, while genera comprising fewer than 5% of total numbers make up 25% of the total in MSs9, in MSs1 fewer than 10% of the total is made up from such 'minority' members of the population, which is highly dominated by just 5 genera that comprise more than 90% of the total. This variation in relative abundances of different genera is characteristic of the data as a whole.

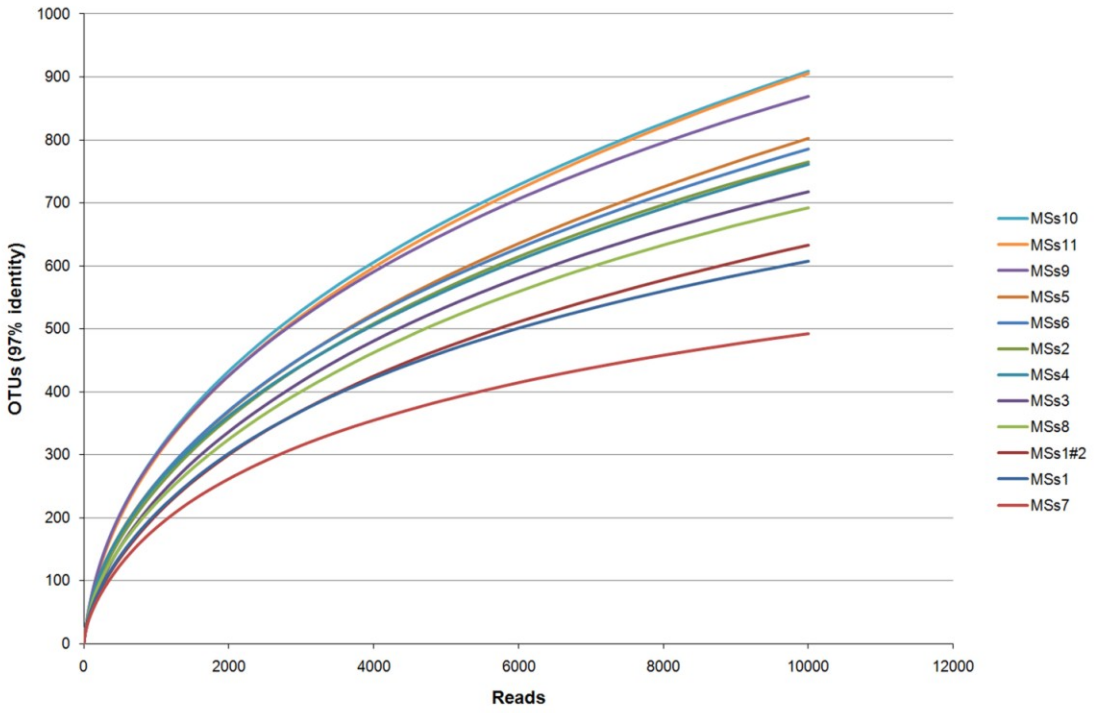
Figure 22. Classification of reads to the genus level. Genera with fewer than 100 reads were pooled and are shown as 'rare genera'. Reads with genus bootstrap scores lower than 30% were pooled as 'unclassified reads'. MSs1#2 denotes replicated read data on the same DNA sample for dog MSs1.



5.4.3 Extent of coverage of microbiota diversity

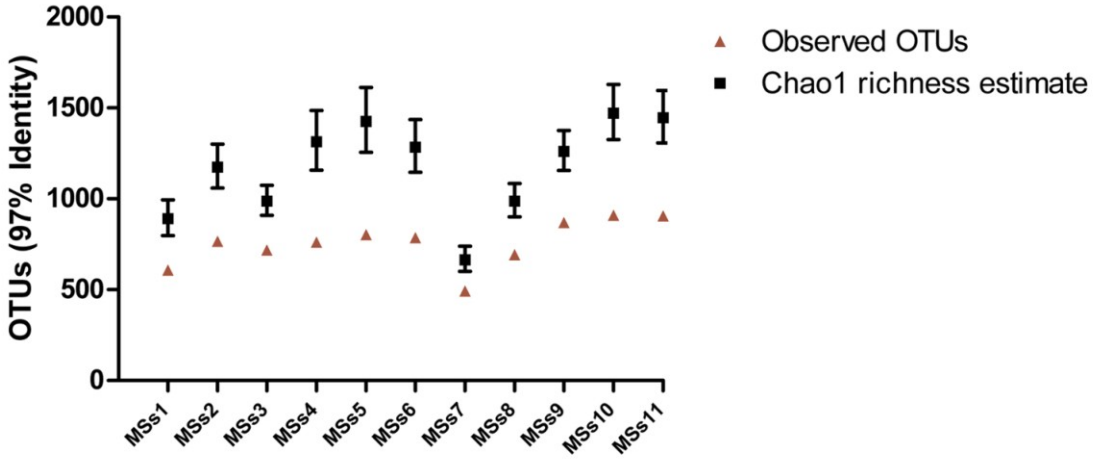
The overall extent of diversity of the digestive tract microbiota is a fascinating question in terms of understanding its complexity and biological role. The identification of sequence reads by matching them to known sequences in the RDP database clearly has limitations in determining overall diversity, since sequences that do not match well to known organisms have been discarded from the analysis above. This is particularly relevant for identification of microbes from lesser studied hosts. We therefore determined the diversity of sequences based on the allocation of reads to 'operational taxonomic units' or OTUs, independently from any known sequence homologies, using RDP infernal aligner and complete linkage clustering tools (Cole, Wang et al. 2009). The OTUs represent notional taxa approximating to species for example, depending on the level of similarity to others, expressed as a percentage similarity in sequence to other groups in the data set. Data showing diversity in an ecological context, where a habitat may not have been exhaustively sampled as here, can conveniently be analysed by mathematically modelling the occurrence of 'new' and repeat sequences in the sample to predict the total diversity in the system. Such analyses can be based on rarefaction curves, shown in Figure 23, where it can be seen that by these analyses there appear to be significant additional numbers of OTUs yet to be determined.

Figure 23. Rarefaction curves for each study animal showing observed diversity within the miniature Schnauzers with increasing sampling effort. Also shown are the technical replicates of the analysis of DNA from MSs1.



It is also possible to predict total numbers of OTUs present in such samples by Chao1 richness estimate as shown in Figure 24. As might be expected from the smaller proportion of the total microbiota taken up by the 5 most abundant genera in this animal, MSs9 shows a larger estimate of total microbiota diversity than, for example, MSs1. Overall it appears that in terms of OTUs, faecal microbiota diversity is likely to range from a few hundred to a thousand or more OTUs per animal. However, these estimates should be viewed with caution, as errors in 454 sequencing at homopolymeric runs may contribute to an overestimate of diversity in the determination of OTU numbers (Quince, Lanzen et al. 2009). Nevertheless, it is clear that there is substantial variation between animals in the diversity of the faecal microbiota.

Figure 24. Observed versus expected diversity of the miniature Schnauzers. Chao1 richness estimates based on the 10,000 reads analysed from each animal. Error bars show upper and lower 95% confidence intervals.



5.4.4 Genetic relatedness between animals and variation in microbiota

Table 7 shows how the subject animals were related. There are several closely related individuals, for example MSs3, MSs4, MSs5, MSs6 and MSs10 are siblings, the first four being littermates. Similarly MSs5 is the mother of MSs7 and MSs8, while MSs8 is the mother of MSs9 and MSs11. MSE is the father of six of the study dogs.

Table 7. Details of miniature Schnauzer dogs used in the study, showing genetic relatedness.

Dog ID	Sex[1]	Mother ID[2]	Father ID[2]	DOB	Age (yrs: mo)
MSs1	MN	MSA	MSC	01/06/1995	11:4
MSs2	FN	MSA	MSD	16/09/1995	11:1
MSs3	MN	MSB	MSE	25/11/2000	5:11
MSs4	FN	MSB	MSE	25/11/2000	5:11
MSs5	FN	MSB	MSE	25/11/2000	5:11
MSs6	FN	MSB	MSE	25/11/2000	5:11
MSs7	MN	MSs5	MSF	18/04/2002	4:6
MSs8	FN	MSs5	MSF	18/04/2002	4:6
MSs9	FN	MSs8	MSE	23/06/2004	2:4
MSs10	MN	MSB	MSE	04/07/2004	2:4
MSs11	FI	MSs8	MSG	12/08/2005	1:2

Note that MSs3, MSs4, MSs5 and MSs6 are littermates and also siblings of MSs10 although from a different litter. MSs7 and MSs8 are also littermates. MSs denotes animals sampled during the study.

[1]Sex is shown as MN/FN indicating male or female neutered (N) or intact (I) at the time of the study.
[2]MSA-E denote parents not sampled in the study.

Inspection of the proportions of prominent genera in the microbiota of individual dogs did not reveal any obvious correlation between genetic relatedness of the animals and the gross distribution of genera present. There does not appear to be any robust association between genetically closely related animals, e.g. littermate siblings, and distribution of the most abundant genera within the microbiota, at least in terms of the most abundant 80% of genera present. Thus, for example littermates MSs3, MSs4, MSs5 and MSs6 did not share similar distributions of the abundant genera *Bacteroides*, *Fusobacterium*, *Prevotella* and *Sutterella*

(Table 7, Figure 22). A control was included in the analysis, whereby DNA from the same sample (MSs1) was analysed in duplicate; whilst not identical, the distribution of genera within these replicate data sets was far more similar than between any two individual animals, indicating that the variations seen are not solely due to technical error in the analysis. It should be noted that in generating these control data, the random resampling of reads was included in the analysis; this in itself would have contributed to the relatively minor differences seen between these replicate determinations on the MSs1 DNA.

5.4.5 Occurrence of potentially pathogenic bacterial genera

Although all the dogs in the study were healthy and lacked any clinical evidence of gastrointestinal disease, small numbers of sequences attributable to pathogenic genera were detected (Table 8). Of three genera of the *Enterobacteriaceae*, sequences representing Salmonella were the most abundant and present in eight of the eleven dogs. However there was a greater than 100-fold range in abundance between animals such that most of the reads were derived from only two individuals. A similar pattern was seen for *Escherichia*, with the same two animals dominating, and in one of these (MSs1) the highest number of reads identified as *Shigella spp.* was also detected. There was a more even distribution of lower levels of *Shigella spp.* in several of the animals. *Helicobacter spp.* were detected in six animals, again with a very uneven distribution – one animal yielding more than 80% of the

248 reads detected. *Campylobacter spp.* were detected very rarely, just two reads from one individual animal.

Table 8. Numbers of reads from potentially pathogenic genera, indicating the presence of potentially pathogenic genera in each dog.

Pathogenic genus	**Dog number (MSs…)**										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Helicobacter	7	4	0	4	0	0	0	0	2	201	30
Salmonella	39	1	2	3	4	0	13	159	0	3	0
Escherichia	96	1	4	8	9	0	5	83	0	0	0
Shigella	78	14	3	10	7	17	2	7	0	0	0
Campylobacter	0	0	0	0	0	0	0	2	0	2	2

The detection of *Escherichia spp.* is not surprising as it would be expected, perhaps more universally, as a member of the normal microbiota; the 16SrDNA sequences do not indicate pathogenic or commensal status. The presence of sequences attributable to *Shigella spp.* is of doubtful meaning when based on these relatively short reads of sequences that are nearly identical to those of *Escherichia*, hence these organisms may not represent pathogenic Shigellae. However it was surprising to find evidence of Salmonella, as this dog population is regularly monitored for all aspects of health and welfare and the organism has not been detected on the study site by conventional bacteriological testing. Furthermore recent literature indicates that the organism is rare among healthy dogs as detected by culture methods (Lenz, Joffe et al. 2009). It is conceivable that dietary intake of *Salmonella* DNA derived from feed ingredients prior to heat treatment during manufacture might account for the signals detected. However if this was so, a more even distribution of these reads between individual animals would be expected.

Regarding the presence of *Helicobacter spp.*, little is known of the occurrence of this genus in dogs, but in view of the widespread occurrence of these organisms in a variety of mammals its

presence in some dogs is not surprising. Again, the uneven distribution between individuals suggests true colonisation of the GI tract rather than dietary intake of DNA. The negligible incidence of Campylobacter 16SrDNA sequences is surprising, as organisms from the species *C. uppsaliensis* and *C. lari* have been isolated previously from dogs at the WALTHAM Centre for Pet Nutrition and studies suggest that *Campylobacter spp.* may be part of the resident faecal microbiota in dogs (Acke, McGill et al. 2009).

5.4.6 Summary and biological significance

The data described comprise one of the first studies to date of the faecal microbiomes a significant number of closely related animals, in terms of both genetic background and family relationships. Perhaps surprisingly, there are considerable differences between host individuals, especially in quantitative terms, between the major groups of bacteria detected, albeit a broad similarity is also perceptible in the dominance of anaerobic organisms in the *Fusobacteria* and *Bacteroidetes* phyla. *Proteobacteria*, including some of the well known potentially pathogenic pathogenic genera, were sparse in number and uneven in distribution.

5.5 MATERIALS AND METHODS

5.5.1 Ethics statement

Animal studies were conducted at the WALTHAM Centre for Pet Nutrition and the project was approved by Waltham's ethics committee.

5.5.2 Animals and faeces collection

The eleven study animals were healthy adult miniature Schnauzer dogs housed at the WALTHAM Centre for Pet Nutrition, and many of the group were closely related genetically. Details are given in Table 1 and also in the general methods section (Table 1). The dogs were housed in high quality kennel accommodation on the same site with varying degrees of cross-contact e.g. from sharing pens, although none was physically isolated, hence direct or indirect contact could occur during exercise, for example. Diet and living conditions were broadly comparable between animals although not identical between all animals. One faecal sample was collected per dog during daily exercise and a cross-section of stool (approx 2 g) frozen on dry ice no more than 15 min following defaecation. The frozen samples were stored at -80ºC for between one and two weeks before DNA extraction.

5.5.3 Faecal DNA extraction

Faecal DNA was extracted from faecal samples using a QIAamp DNA Stool Mini kit (Qiagen). A subsample of 190-220 mg of frozen faeces was processed following the 'Protocol

for Isolation of DNA from Stool for Pathogen Detection' detailed in the manufacturer's instructions, using a lysis temperature of 95ºC.

Extracted DNA was eluted from the spin columns in 200 µl of Qiagen AE buffer (10 mM Tris-Cl and 0.5 mM pH 9.0 EDTA). Extracted DNA was then quantified and checked for purity (based on UV absorption spectrum and 260:280 nm and 260:230 nm absorption ratios) on a ND1000 spectrometer (Nanodrop Technologies Inc).

5.5.4 Amplification by PCR of 16S rRNA gene

A barcoded 16S rRNA gene tag approach was used to amplify a ~550 bp region covering the V1-V3 regions of the 16S rRNA gene sequence. The forward and reverse primers, AC84A_BX and E533B respectively were as follows: AC84A_BX, 5'-GCCTCCCTCGCGCCATCAG[N8]AGAGTTTGATYMTGGCTCAG-3' and E533B, 5'-GCCTTGCCAGCCCGCTCAGTIACCGIIICTICTGGCAC-3'. AC84A_BX comprised, from the 5' end, the 454 sequencing adapter A, sample barcode octamer (denoted by N8) and rDNA-specific sequence 27f-YM (A88, Table 2, page 25)(Frank, Reich et al. 2008). E533B comprised from the 5' end, 454 sequencing adapter B and rDNA-specific sequence E533 (E533, Table 2, page 25)(Watanabe, Kodama et al. 2001). Individual 50 µl PCRs were setup as follows; 25 µl Extensor ready mix (Thermo Scientific), 3 µl AC84A_BX 10 pmol/µl, 3 µl E533B 10 pmol/µl, 1.5 µl Nuclease-Free Water (Promega) and 17.5 µl (35 ng) faecal DNA. Amplification was for 30 cycles with the following conditions: 94°C for 3 min followed by 10 cycles of 94°C for 45 s, 55°C for 30 s, 72°C for 1 min, then 19 cycles of 94°C for 45 s, 55°C for 30 s, 72°C for 1 min, 30 s, and a final cycle of 94°C for 45 s, 55°C for 30 s and 72°C for 7

min. Amplicon abundance and size were checked using agarose gel electrophoresis and were then purified using a QIAquick PCR Purification Kit (Qiagen).

5.5.5 Roche-454 sequencing

Purified PCR amplicons from different dogs were pooled on an equimolar basis based on ND1000 spectrometer readings. This pool was then sequenced from a primer annealing to adapter A on a 454 GS FLX sequencer (Roche) using FLX chemistry and picotitre plates following the manufacturer's protocols.

5.5.6 Data analysis

Using local databases and code written in Python, the raw sequence read data were initially filtered to remove sequences below 150 nt, those containing one or more ambiguous bases and those with a mismatch against the AC84 primer sequence. Sequences were then binned by barcode sequence and each bin was randomly resampled using the Random module in Python, based on the Mersenne Twister algorithm (Makoto and Takuji 1998). This allowed standardization of sequence read number to 10,000 reads per dog to reduce bias in the subsequent comparative data analysis between dogs.

Taxonomic assignment of reads was done using a downloaded copy of the 2.0 version of the Bayesian classifier algorithm from the Ribosomal Database Project (RDP, http://rdp.cme.msu.edu/) (Wang, Garrity et al. 2007). Classification at the phylum and genus levels was done using bootstrap scores of 0% to 100%. A value of 30% was chosen because although a significant number of genus identifications are likely to be incorrect, a

substantially greater number of correctly identified genera are then included in the output data (Dr C. Wallis, unpublished observation). We believe that on balance this increases the validity of the determination of proportions of different genera present.

Operational taxonomic units (OTUs) were determined for the total of 120,000 reads combined from the 10,000 resampled reads per dog, using the RDP infernal aligner and Complete Linkage Clustering tools (Cole, Wang et al. 2009). Rarefaction curves and Chao1 richness estimates (Chao 1984; Hill 2006) were calculated for each dog using Mothur version 1.8.1 (Schloss, Westcott et al. 2009).

5.6 ACKNOWLEDGEMENTS

We acknowledge the support of the Biotechnology and Biological Sciences Research Council for a CASE studentship to DH, supported by WALTHAM Centre for Pet Nutrition. We also acknowledge the contribution of Michigan State University and the Ribosomal Database Project. We thank the staff of the Centre for Genomic Research, University of Liverpool for sequencing work.

5.7 AUTHOR CONTRIBUTIONS

All authors contributed to concept and design of experiments. DH performed the experiments and analysed the data. All authors contributed to writing the paper.

CHAPTER SIX: Inter-breed comparison of canine faecal microbiota by high throughput sequence analysis

Daniel Hand[1], Corrin Wallis[2] and Charles W. Penn[1]*

6.1 Abstract

Faecal microbiota from 79 dogs representing 6 breeds were characterized and quantified by Roche-454 FLX high throughput sequencing of the V1 and V2 regions of 16SrDNA amplicons from faecal DNA, in a uniquely large and comprehensive study of the microbiota of a single species of non-human mammal. Sequence data quality was enhanced by use of the AmpliconNoise algorithm, and the biological significance of 8,500 sequence reads per animal assessed. Sequences were assigned to known taxa using the Bayesian Classifier of the Ribosomal Database Project (RDP), and to operational taxonomic units (OTUs) *de novo* using the RDP Infernal aligner. Assessment of technical error demonstrated a highly significant pairwise correlation between technical replicates. Sequence data indicated that the faecal microbiota were dominated numerically by the five Gram-negative genera *Bacteroides*, *Prevotella*, *Cetobacterium*, *Fusobacterium* and *Sutterella*, with much diversity between individuals. The prominence of *Cetobacterium* and *Fusobacterium,* in the phylum *Fusobacteria*, is unusual in mammals studied to date. There was an indication of breed

specificity in the microbiota, Labrador retrievers forming a statistically separate cluster from all other breed groups. Evidence for the existence of a 'core microbiome' characteristic of the species was very limited: only one of more than 3,500 OTUs identified at a 97% identity cut-off was universally present, and only 20 OTUs were present among 90% or more of the animals. The latter featured several genera of the *Firmicutes* as well as the numerically dominant Gram-negatives. Some evidence was also obtained for environmental and/or dietary influences on microbiota composition.

Key words: Microbiota/canine/microbiome/phylogenetic/pyrosequence

6.2 Introduction

Spearheaded by the Human Microbiome Project (Peterson, Garges et al. 2009) and the MetaHIT consortium (Qin, Li et al. 2010), interest in the GI tract microbiota of mammals has exploded over the past two or three years as High Throughput (HT) sequencing technology (Marchesi 2010; Pallen, Loman et al. 2010) has transformed research. The HT approach has been exploited both to assess the genetic composition of these extremely complex microbial populations by metagenomic analysis (Qin, Li et al. 2010) , and to define and quantify the immense phylogenetic diversity of known and novel microorganisms present (Tringe and Hugenholtz 2008; Marchesi 2010). These organisms are crucial in GI tract function and in the development and maintenance of host physiology and health (Chung and Kasper 2010; Garrett, Gordon et al. 2010; Kelly 2010; Sekirov, Russell et al. 2010; Winter, Keestra et al. 2010), yet until the introduction of HT sequencing approaches, their definition and investigation has been highly problematic as many are difficult to culture, quantify and characterise by conventional bacteriological methods. Plasmid cloning and Sanger sequencing

of phylogenetically informative amplicons (Tringe and Hugenholtz 2008), the fore-runner of HT sequencing to define microbial diversity, lacked the numerical power to impact greatly on the unraveling of these extremely diverse and complex microbial communities.

Most commonly, 16S rRNA gene has been the phylogenetic marker amplified and sequenced as a signature molecule (Tringe and Hugenholtz 2008). Sequences indicate the presence and relative proportion of hundreds or thousands of different known taxa, and of undefined novel organisms identified only as 'operational taxonomic units' (OTUs). The most powerful HT technology for analysis of rDNA sequences is Roche-454 pyrosequencing (Margulies, Egholm et al. 2005), due to its long reads which to a large extent circumvent difficulties arising from generation of chimeric sequences that can occur frequently if assembly of shorter reads is attempted. The analysis of Roche-454 sequence reads has been greatly improved recently by the introduction of algorithms to correct sequencing errors, especially those resulting from erroneous reading of runs of the same nucleotide, and to eliminate chimeric sequences (Quince, Lanzén et al. 2009; Turnbaugh, Quince et al. 2010). To date, exploration of the diversity of GI tract microbiota using HT sequencing approaches have demonstrated major diversity in both numbers and identities of taxa (richness) and the relative abundance of different taxa (evenness) in relatively small cohorts from a diverse range of animal species (Gill, Pop et al. 2006; Dowd, Sun et al. 2008; Hildebrandt, Hoffmann et al. 2009; Bailey, Dowd et al. 2010; Callaway, Dowd et al. 2010; De Filippo, Cavalieri et al. 2010; Friswell, Gika et al. 2010; Middelbos, Vester Boler et al. 2010; Ritchie, Burke et al. 2010). Few studies published to date have extended these surveys to evaluate systematically the effects of variables such as age, diet or environment. Variations observed in different individuals in different circumstances cannot be rationally interpreted without some comprehension of

intrinsic variation and instability in these complex ecosystems. Furthermore, key theoretical issues remain uncertain. These include the long-postulated existence of a 'core microbiome' characteristic of a host species, of microorganisms that are invariably present and perhaps make up a population adapted for optimal association with a host species to best mutual advantage (Turnbaugh and Gordon 2009; Turnbaugh, Hamady et al. 2009; Zaura, Keijser et al. 2009). It is also unclear, for example from studies in monozygotic and dizygotic twins, to what extent host genotype influences the microbiota (Turnbaugh, Quince et al. 2010).

We now report an extensive phylogenetic analysis of the faecal microbiota, using Roche-454 FLX pyrosequencing to sequence 200 bp of the 5' region of the 16S rRNA gene amplicon pool, including V1 and V2 variable regions, in a total study population of 79 dogs representing 6 different breeds. We believe the use of dogs in such a study carries major advantages over many other host species, as environmental, dietary and other conditions are readily controlled and/or monitored. Different breeds represent genetically distinct groups within the species that may enable dissection of the influence of genetic factors or resulting physiological differences in determining variation in the microbiota.

6.3 Methods

6.3.1 Animals and faeces collection

The study population consisted of a diverse group of 79 dogs from a variety of geographical locations and environments (Table 1). The majority of dogs were housed in high quality, managed kennel accommodation at three locations, the WALTHAM® Centre for Pet Nutrition (WCPN), a UK public sector facility (PSF) and Royal Canin research facility, France (RC). The remainder were located in mixed accommodation with a variety of small

business and private owners across Australia. One faecal sample was collected per dog during daily exercise and a cross-section of stool (~2 g) frozen on dry ice no more than 15 min following defaecation. The frozen samples were stored and transported for analysis at -80ºC.

6.3.2 Faecal DNA extraction

DNA was extracted from faecal samples using a QIAamp DNA Stool Mini kit (Qiagen, Crawley, UK). A subsample of 190-220 mg of frozen faeces was processed following the 'Protocol for Isolation of DNA from Stool for Pathogen Detection' detailed in the manufacturer's instructions, using a lysis temperature of 95^{o}C to maximize recovery of DNA from Gram-positive organisms. Extracted DNA was eluted from the spin columns in 200 µl of Qiagen AE buffer (10 mM Tris-Cl and 0.5 mM pH 9.0 EDTA). Extracted DNA was then quantified and checked for purity (based on UV absorption spectrum and 260:280 nm and 260:230 nm absorption ratios) on a Nanodrop ND1000 spectrometer (Thermo Fisher Scientific, Loughborough, UK).

6.3.3 Amplification by PCR of 16S rRNA gene

A barcoded 16S rRNA gene tag approach was used to amplify approximately 550 bp covering the V1-V3 regions of the 16S rRNA gene sequence. The forward and reverse primers, AC84A_BX and E533B respectively were as follows: AC84A_BX, 5'-GCCTCCCTCGCGCCATCAG[N8]AGAGTTTGATYMTGGCTCAG-3' and E533B, 5'-GCCTTGCCAGCCCGCTCAGTIACCGIIICTICTGGCAC-3'. AC84A_BX comprised, from the 5' end, the 454 sequencing adapter A, sample barcode octamer (denoted by N8) and rDNA-specific sequence 27f-YM (A88, Table 2, page 25)(Frank, Reich et al. 2008). E533B comprised from the 5' end, 454 sequencing adapter B and rDNA-specific sequence E533

(E533, Table 2, page 25)(Watanabe, Kodama et al. 2001). Individual 50 µl PCRs were setup as follows; 25 µl Extensor ready mix (Thermo Scientific), 3 µl AC84A_BX 10 pmol/µl, 3 µl E533B 10 pmol/µl, 1.5 µl Nuclease-Free Water (Promega UK Ltd, Southampton, UK) and 17.5 µl (35 ng) faecal DNA. Amplification was for 30 cycles with the following conditions: 94°C for 3 min followed by 10 cycles of 94°C for 45 s, 55°C for 30 s, 72°C for 1 min, then 19 cycles of 94°C for 45 s, 55°C for 30 s, 72°C for 1 min, 30 s and a final cycle of 94°C for 45 s, 55°C for 30 s and 72°C for 7 min. Amplicon abundance and size were checked using agarose gel electrophoresis and were then purified using a QIAquick PCR Purification Kit (Qiagen).

6.3.4 Roche-454 sequencing

Purified PCR amplicons were pooled into three groups on an equimolar basis based on ND1000 spectrometer readings. These pools were then sequenced in three runs from a primer annealing to adapter A on a 454 GS FLX sequencer (Roche Diagnostics, Burgess Hill, UK) using FLX chemistry and picotitre plates following the manufacturer's protocols.

6.3.5 Data analysis

Using local databases and code written in Python, the raw sequence read data were initially filtered to remove sequences below 150 nt, those containing one or more ambiguous bases, those with low base call quality scores and those with a mismatch against the AC84 primer sequence. The primer sequences were removed and sequence length truncated to 200 nt.
The dataset was de-noised using the AmpliconNoise algorithm (Quince, Lanzén et al. 2009; Turnbaugh, Quince et al. 2010). This procedure reduced the noise associated with both pyrosequencing and replication errors during PCR, and created a set of de-noised reference sequences to which the original sequences were assigned. Sequences were then binned by

barcode sequence and each bin was randomly sampled using the Random module in Python, based on the Mersenne Twister algorithm (Makoto and Takuji 1998). This allowed standardization of sequence read number to 8,500 reads per dog to reduce bias in the subsequent comparative data analysis between dogs.

Taxonomic assignment of the reference reads was done using a downloaded copy of the 2.2 version of the Bayesian classifier algorithm with hierarchy model training set 6 from the Ribosomal Database Project (RDP, http://rdp.cme.msu.edu/) (Wang, Garrity et al. 2007). Classification at the phylum and genus levels was done using a range of bootstrap confidence values between 0% and 100%. A bootstrap confidence value of 30% was chosen for use of the classifier because although a minority (approximately 15%) of genus identifications are likely to be incorrect, a substantially greater number of correctly identified genera are then included in the output data. This decision was based on modelling of the outcomes of classifying a set of sequences of known taxa both at full length and trimmed to 250 bp spanning regions V1 and V2, using differing bootstrap confidence values (data not shown). We believe that on balance the use of the confidence value of 30% increases the validity of the determination of proportions of different genera present, to achieve a 'broad brush' view of the data.

Operational taxonomic units (OTUs) were determined using the RDP infernal aligner and Complete Linkage Clustering tools (Cole, Wang et al. 2009). Rarefaction curves and Chao1 richness estimates (Chao 1984; Hill 2006) were calculated for each dog using Mothur version 1.12.3 (Schloss, Westcott et al. 2009). Average linkage hierarchical clustering was done using complete linkage clustering/average linkage clustering using the hclust and heatmap functions and the Bioconductor package.

6.3.6 Statistical analyses

A number of statistical analyses were undertaken to determine the effect of genetics, diet, geographical location, age and sex on the faecal microbiota of dogs. Pearson correlation coefficient was used to determine if inter- and intra-breed measurements across genera are correlated. This analysis was also used to determine the correlation of genera across dogs. The Pearson correlation coefficient measures the strength of the linear relationship between two variables. This correlation can take on values from -1.0 to 1.0, where -1.0 indicates a perfect negative (inverse) correlation, 0.0 no correlation, and 1.0 a perfect positive correlation. As a guide, values >0.7 are deemed strongly positively correlated and values <-0.7 strongly negatively correlated. Analysis of variance (ANOVA) was used to determine if differences existed in the number of observed and estimated OTUs between breeds, using Bonferroni adjustments for pairwise comparisons for an overall 5% test level. Principle component analysis (PCA) was also performed to visualise correlations between genera and associations with breed.

6.4 Results

6.4.1 Read numbers, filtering and processing

Amplicons were generated by PCR of the 16SrDNA from faecal DNA from each member of the study population. Roche-454 sequencing yielded a total of 1,772,263 raw reads. After filtering out those with incorrect MID or primer sequences, short length or one or more ambiguous bases, and applying noise reduction and PCR error correction using AmpliconNoise, 1,083,864 reads remained. Reads were then binned by sequence MID resulting in an average of 12,751 (SD 2572) per MID. To enable comparable analyses

resulting from similar depths of sampling per dog, 8,500 reads were randomly sampled from each bin (see Methods).

6.4.2 Sequence classification

Sequence reads were first analysed using the Ribosomal Database Project (RDP) Bayesian Classifier, version 2.2. Reads were assigned to taxa at genus level down to a bootstrap confidence value of 30%. This cut-off resulted in an average of approximately 10% of reads remaining unclassified across all dogs, ranging from about 2% to 30% with one outlier at 50% (*Figure 25*).

Figure 25A. Distribution by % of total of the most abundant genera (RDP Classifier data) by dog. The key to the right shows genera in ascending order of abundance across all dogs. Columns identified by #n are technical replicates of analyses shown to their left. Rare groups (grey) include those present at <0.1% of the total number of reads for that dog. Black: unclassified reads

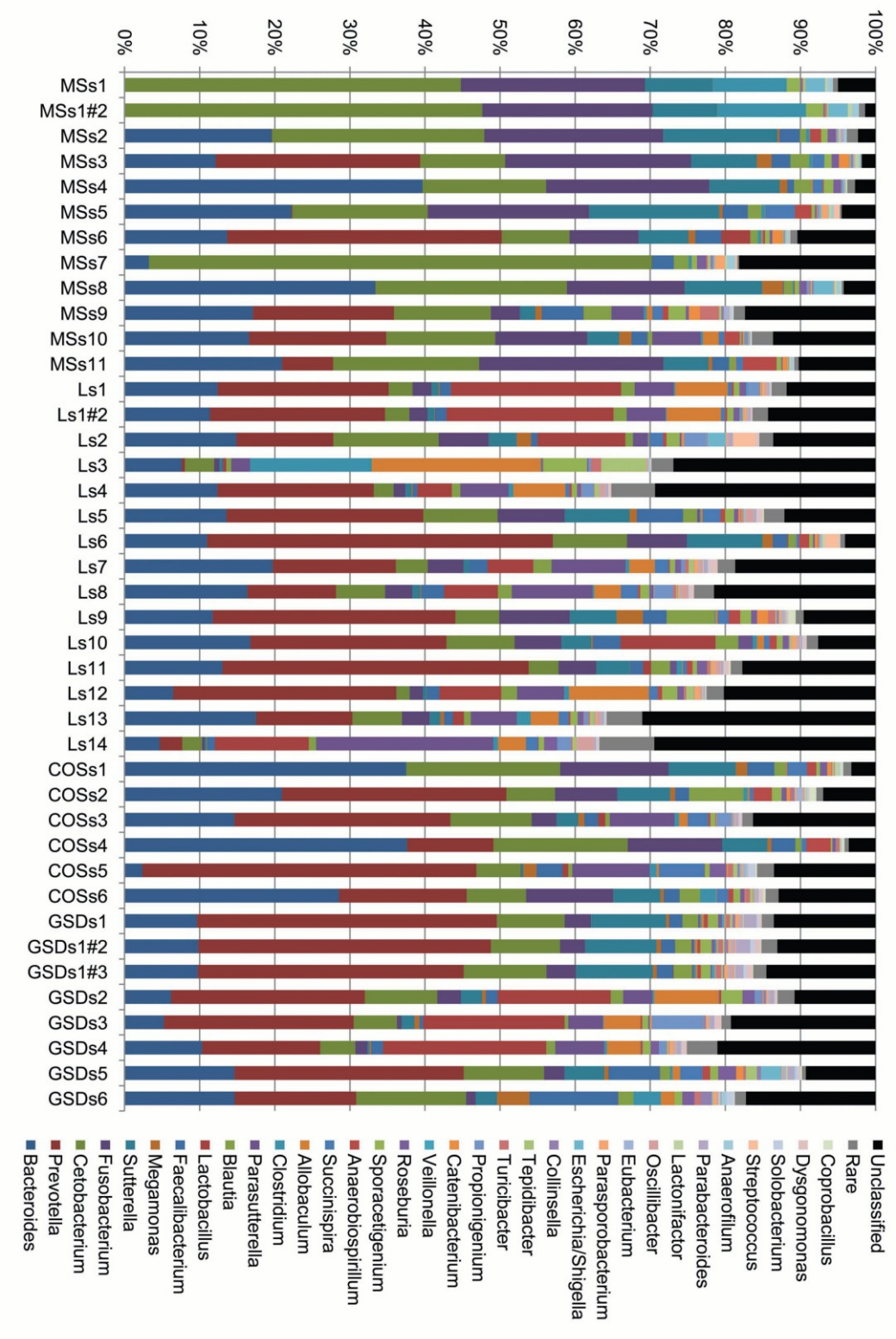
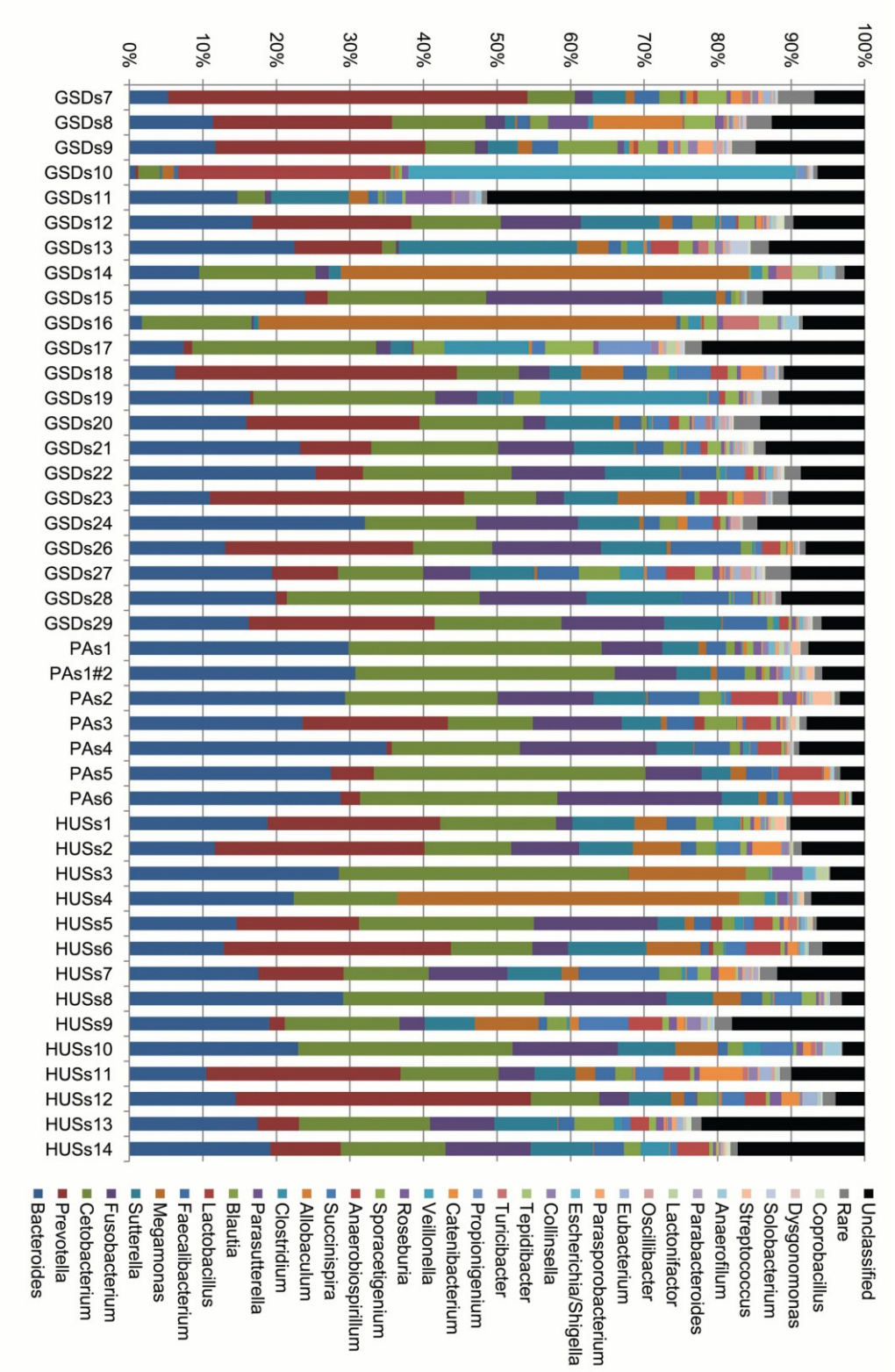


Figure 25B. Distribution by % of total of the most abundant genera (RDP Classifier data) by dog. The key to the right shows genera in ascending order of abundance across all dogs. Columns identified by #n are technical replicates of analyses shown to their left. Rare groups (grey) include those present at <0.1% of the total number of reads for that dog. Black: unclassified reads



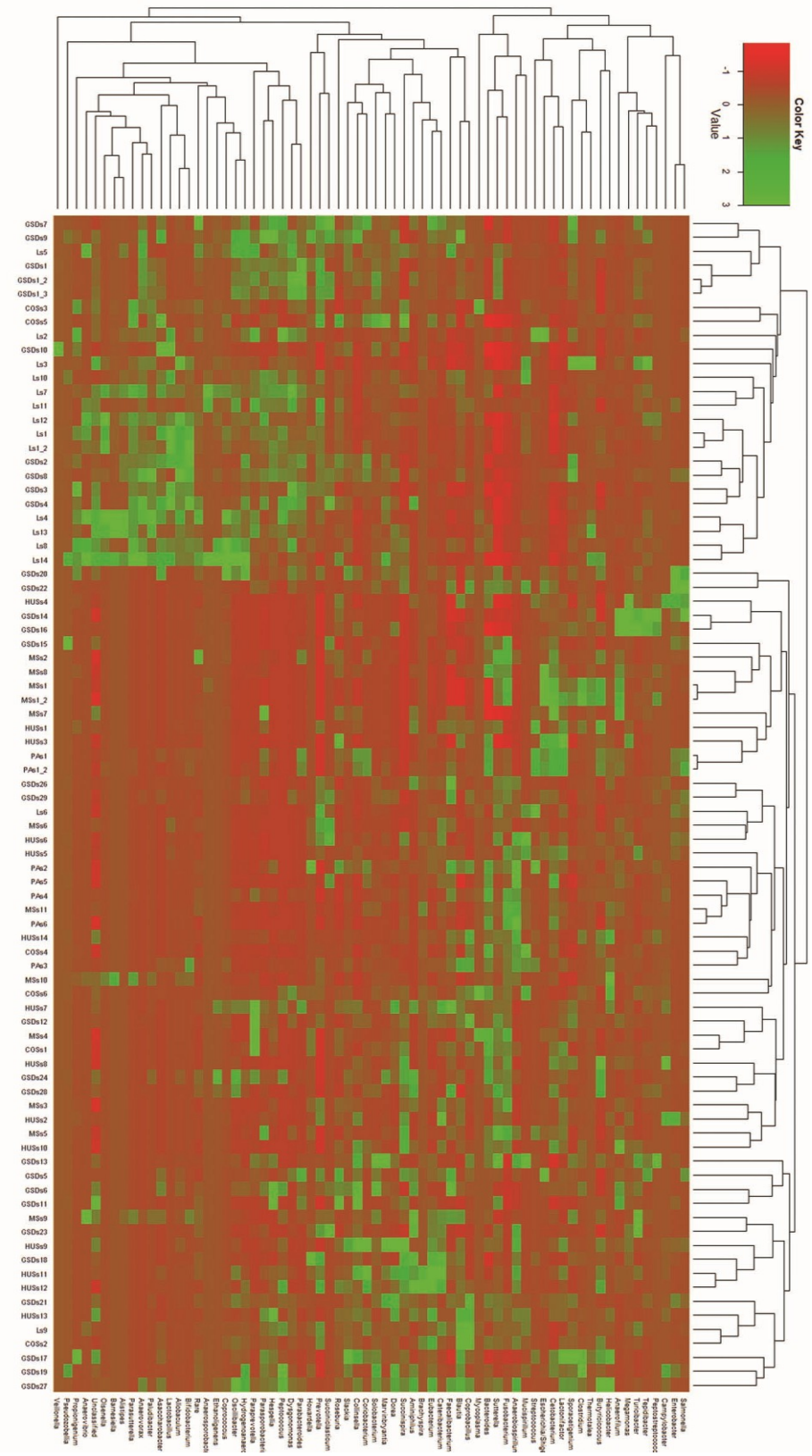
From the data shown in *Figure 25* we indicate the following main conclusions. Populations were dominated numerically in the majority of dogs by members of the Gram-negative genera *Bacteroides*, *Prevotella*, *Cetobacterium*, *Fusobacterium* and *Sutterella*. Within the Firmicutes, genera *Megamonas* and *Faecalibacterium* were among the most prominent although the distribution of *Megamonas* was very uneven; also prominent in the majority of Labrador retrievers and a small number of German shepherd dogs (GSDs) were *Lactobacilli*, but these were generally not abundant in other breeds. There were substantial differences between dogs, even within breeds, in the numbers of the most dominant genera, evenness of distribution of genera, and proportions of unclassified reads. There appeared to be a tendency towards greater evenness in the majority of Labrador retrievers, in other words a less marked numerical dominance of the more abundant genera and hence an apparently greater diversity of genera among the more abundant members of the population. It is noticeable that numbers of *Prevotella* were highly variable, and there is an indication of a possible inverse relationship of their numbers with *Cetobacterium* numbers. Statistical analysis showed a moderate negative correlation between *Prevotella* and *Cetobacterium*, with a Pearson correlation coefficient value of -0.545, $P<0.001$. Additional specific observations include a markedly aberrant distribution of genera and very large number of unclassified reads in dog GSDs11 compared with the norm among all other dogs; and very large numbers of the anaerobic *Firmicute Megamonas* in a small number of animals.

An analysis using the RDP Classifier at phylum level indicated that the major phyla present were *Bacteroidetes* (37.54%), *Fusobacteria* (27.09%), *Firmicutes* (24.16%), *Proteobacteria* (10.55%), *Actinobacteria* (0.55%) and others (0.10%).

The analysis of four samples (MSs1, Ls1, GSDs1, PAs1) was replicated through the entire process of PCR amplification, sequencing and data analysis, including random sampling. The resulting reads were subjected to analysis, to provide an indication of the technical error in the analytical procedure (but excluding sampling error due to any sample heterogeneity during extraction of faecal DNA). In one case (GSDs1) the analysis was performed in triplicate. Statistical analysis showed technical replicates to be highly correlated, with Pearson correlation coefficients of >0.99 at $P<0.001$.

The diversity of microbiota between dogs was further investigated in a hierarchical clustering analysis. *Figure 26* illustrates an average linkage and heatplot analysis of RDP Classifier data performed in the statistical package R, of all dogs *vs* the most abundant bacterial genera identified. The analysis indicates several interesting findings. Twelve of the 14 Labrador retrievers fell into a separate cluster with a smaller number of GSDs and one cocker spaniel, apparently due in part to larger numbers of *Lactobacilli* and *Bifidobacteria* combined with a marked absence of anaerobes such as *Fusobacterium*, *Sutterella* and *Bacteroides*. Among the bacterial genera, there was an interesting association of *Bifidobacteria* and *Lactobacilli*, both lactic acid producers, and clustering of *Proteobacteria* including members of the *Enterobacteriaceae*, which also tended to cluster with a variety of other unrelated organisms including *Megamonas*.

Figure 26. Heatmap of average linkage hierarchical cluster analysis of RDP Classifier data comparing the study animals and bacterial genera.



6.4.3 De novo assignment to OTUs

Reads were assigned to OTUs at the level of 97% sequence identity, using the RDP infernal aligner and complete linkage clustering tools. At this level of identity, 3 786 OTUs were identified among a total of 714,000 reads from 79 dogs and including the 42,500 reads from 5 replicated analyses. The extent of coverage of the microbiota achieved was assessed by rarefaction analysis. The shape of the rarefaction curves (Figure 27 and Figure *28*) indicates that the full bacterial richness has not been revealed at a sequencing depth of 8 500 sequences per individual. Rarefaction analysis by breed average (Figure 28) suggested substantial differences in coverage between breeds, the miniature Schnauzers appearing better covered and the Labrador retrievers less well covered than the other breeds which appeared to cluster together. The data were also modeled using the Chao1 richness index to predict total numbers of OTUs present per animal (Figure 29). Predicted totals ranged from fewer than 200 to more than 800, with a mean of approximately 400 OTUs per animal. There appeared to be some correlation between numbers of OTUs and breed, notably for the miniature Schnauzers which tended to have a low level of richness and the Labrador retriever group with a high level. Analysis of observed OTU numbers by ANOVA (Figure 30) showed that only the Labrador retriever group was significantly different from all the other breed groups after Bonferoni correction for multiple testing. In this analysis the GSDs were separated into groups according to location (PSF, RC and Australian) and some additional differences between groups were statistically significant as follows. The GSD UK group was significantly different from the GSD Australia and GSD France groups; and the miniature Schnauzer group was significantly different from the GSD France group. In a Principal Component Analysis of the data analysed in *Figure 26*, the Labrador retriever group again showed clear differences from other

groups in the distribution of different genera (Figure 31). No clustering was observed in relation to age of animals or sex.

Figure 27. Rarefaction curves of OTUs determined for individual animals showing observed diversity with increasing sampling effort. Colours indicate breeds as shown to the right.

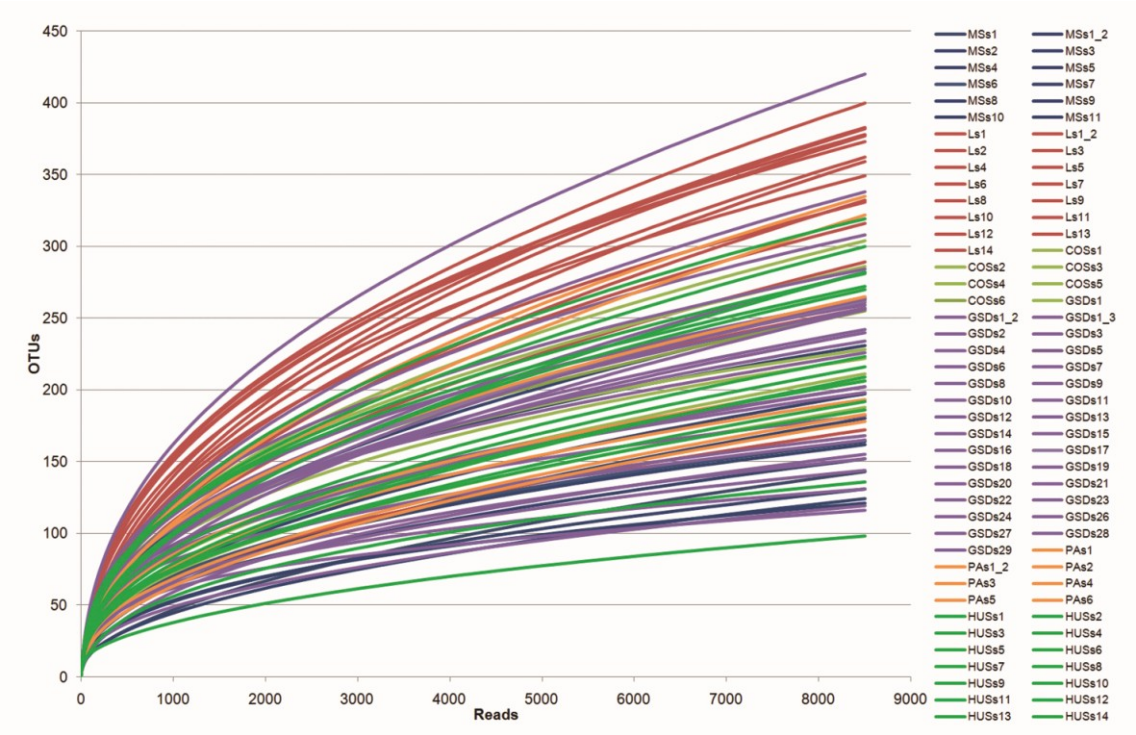


Figure 28. Rarefaction curves showing observed diversity with increasing sampling effort of OTU data by breed average (solid lines) showing standard deviations for animals within each breed group (dashed lines).

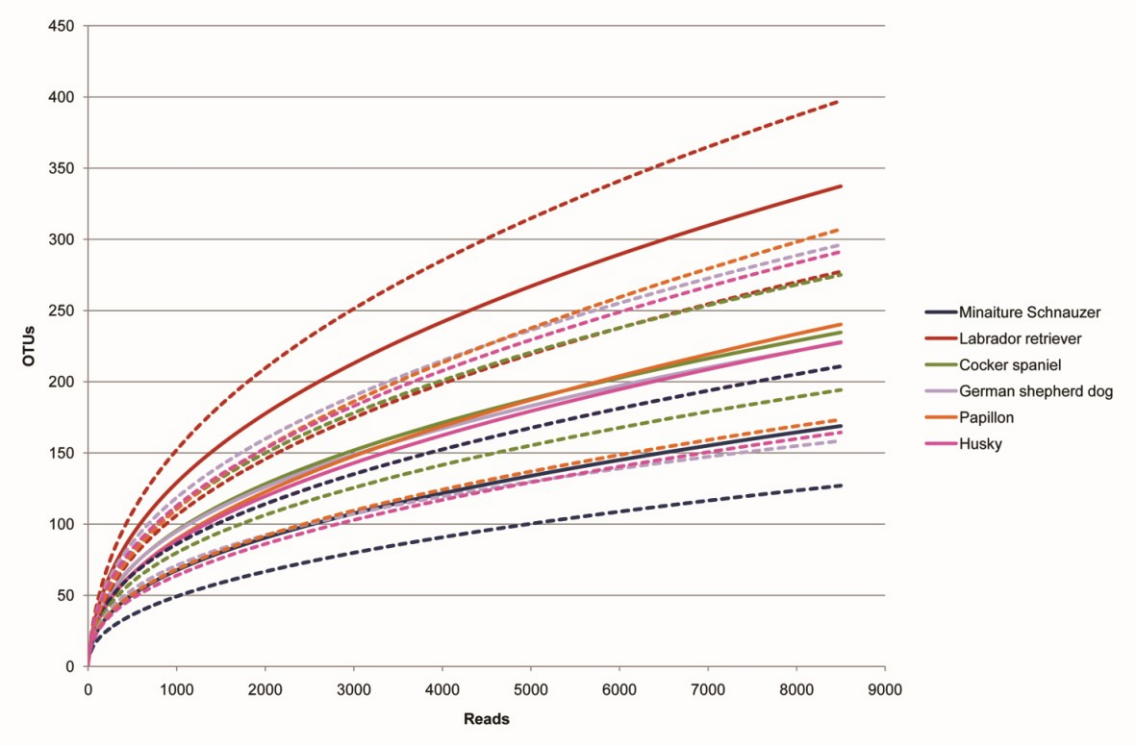


Figure 29. Observed OTU numbers and Chao1 predicted richness estimates by dog, with upper and lower 95% confidence intervals.

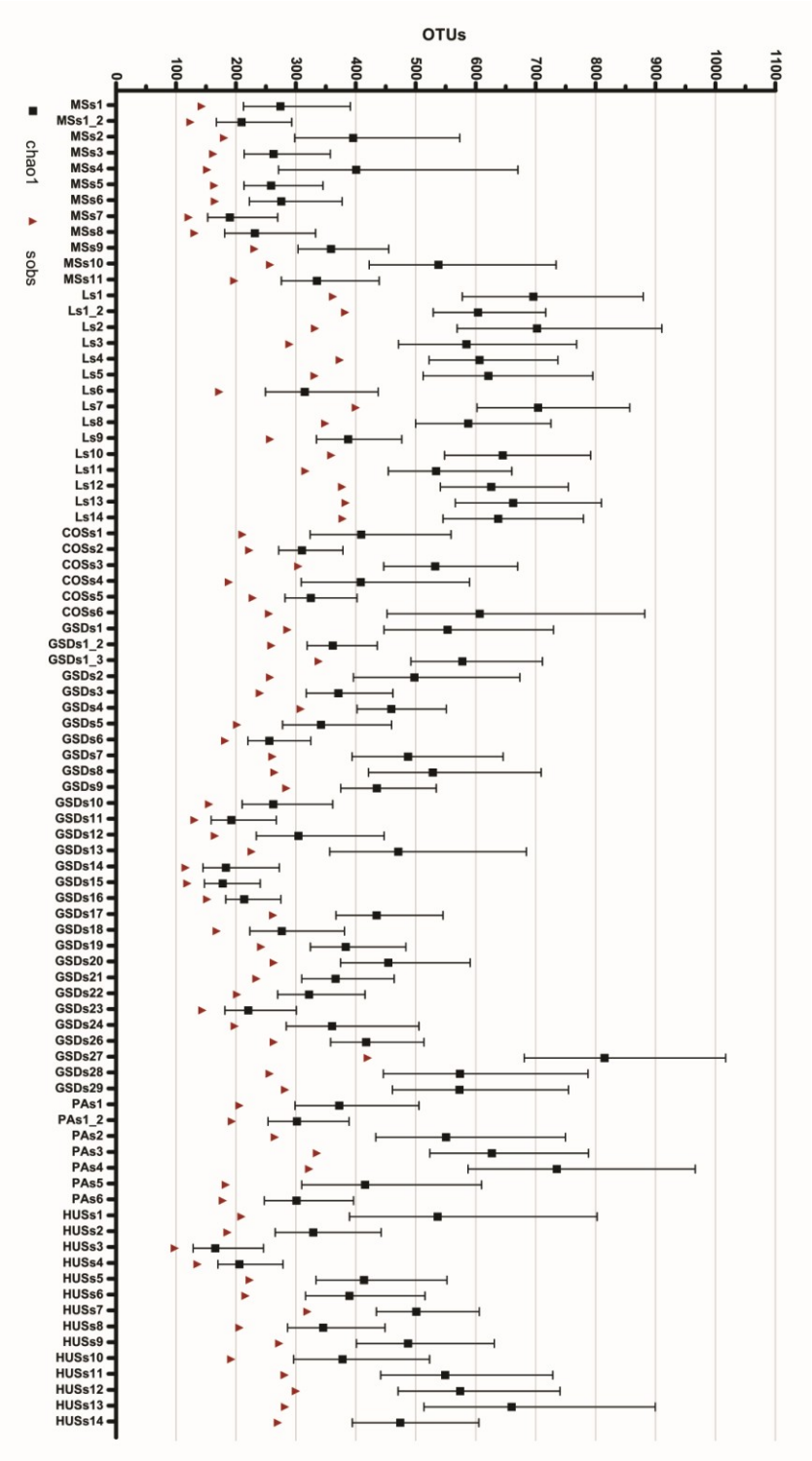


Figure 30. ANOVA analysis of observed numbers of OTUs between breeds and in three geographical groups of GSDs, showing significantly increased Labrador diversity.

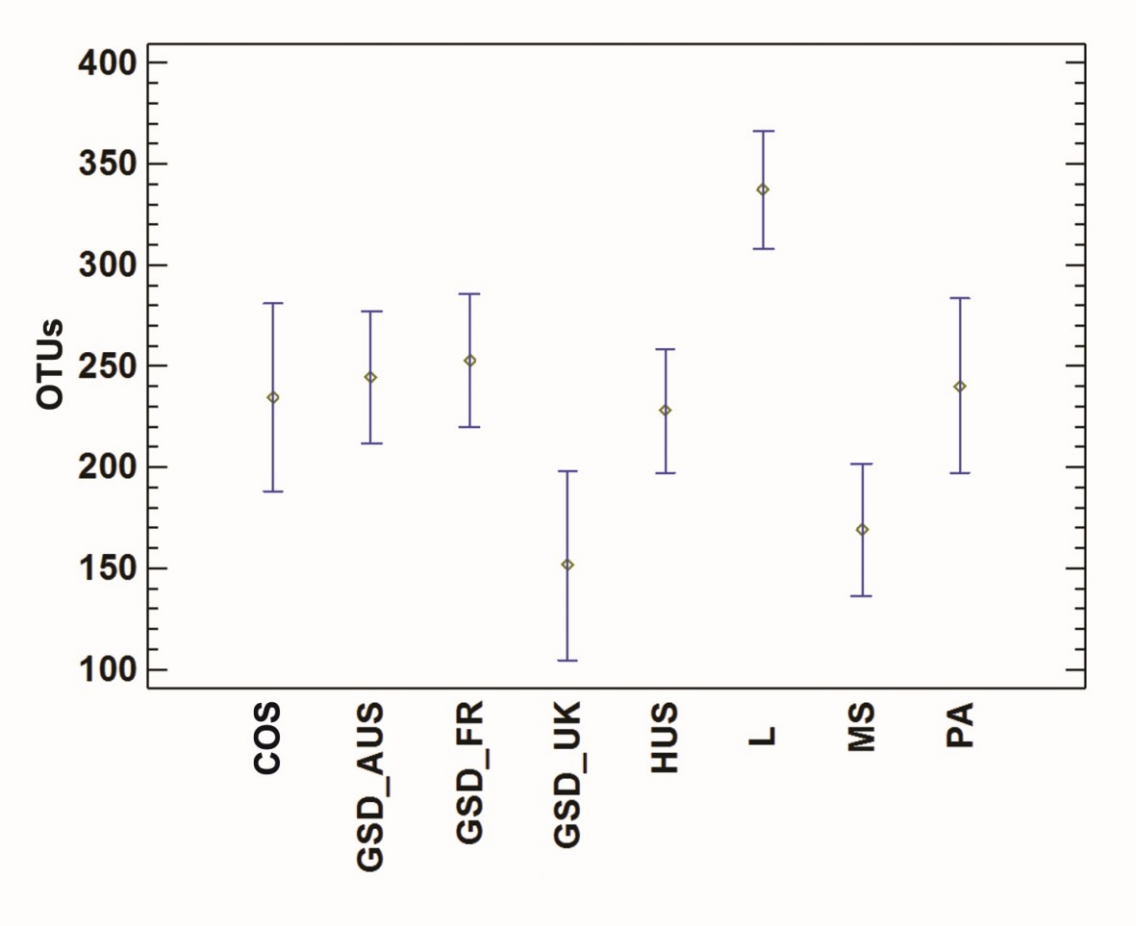
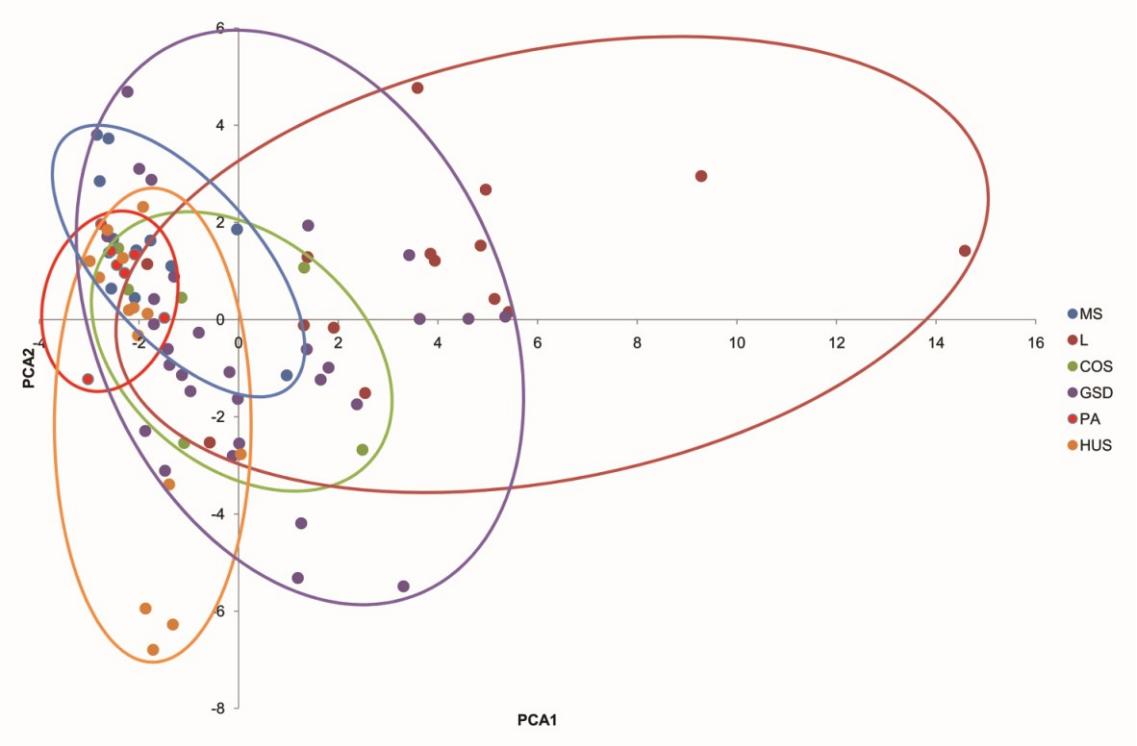


Figure 31. PCA comparison of dog genera profiles and how they cluster by breed showing dissociation of Labradors with the main cluster.



A rank abundance analysis of all OTUs from all dogs was also performed (Figure 32) and it shows that overall, a very small number of OTUs accounted for the bulk of reads: approximately 20 OTUs accounted for about 85% of all reads. However as shown in Table 9, which excludes replicate data to avoid bias in the coverage of the replicated animals, only one OTU was universally present (albeit only one sequence read was detected in two animals) and this accounted for 7.88% of all reads. In Table 10, the OTUs that were present in 90% or more of animals are shown ranked by distribution across animals. The OTUs that were present in 90% or more of animals numbered only 17 and accounted for 37.4% of reads, while those present in 75% or more of animals numbered 44 and accounted for 66% of all reads. However it is also notable that there was only a very weak correspondence between distribution across large numbers of animals and total number of reads: some of the conserved OTUs represent well below 1% of the population. In considering the definition of a core microbiome, these data suggest that abundance is not a pre-requisite for wide distribution. Sequences within the 17 OTUs that were present in at least 90% of animals, which might be considered as approximating a core microbiome, could be assigned to a total of 12 genera in the 5 orders *Fusobacteriales* (*Fusobacterium* (1), *Cetobacterium* (3)); *Clostridiales* (*Blautia* (3), *Sporacetigenium* (1), *Faecalibacterium* (1), *Lactonifactor* (1), *Roseburia* (2), *Megamonas* (1)); Bacteroidales (*Bacteroides* (3), *Prevotella* (1)); *Burkholderiales* (*Sutterella* (2)); and *Flavobacteriales* (*Pseudozobellia* (1)). Numbers in brackets, a total of 20, indicate for each genus how many OTUs included sequences assigned to that genus. These OTUs could be considered to represent numbers of taxa approximating to ‘species’ present within this putative core group.

Figure 32. Rank abundance curve for all OTUs present in the complete dataset showing the extent of the rare biosphere of the canine microbiota.

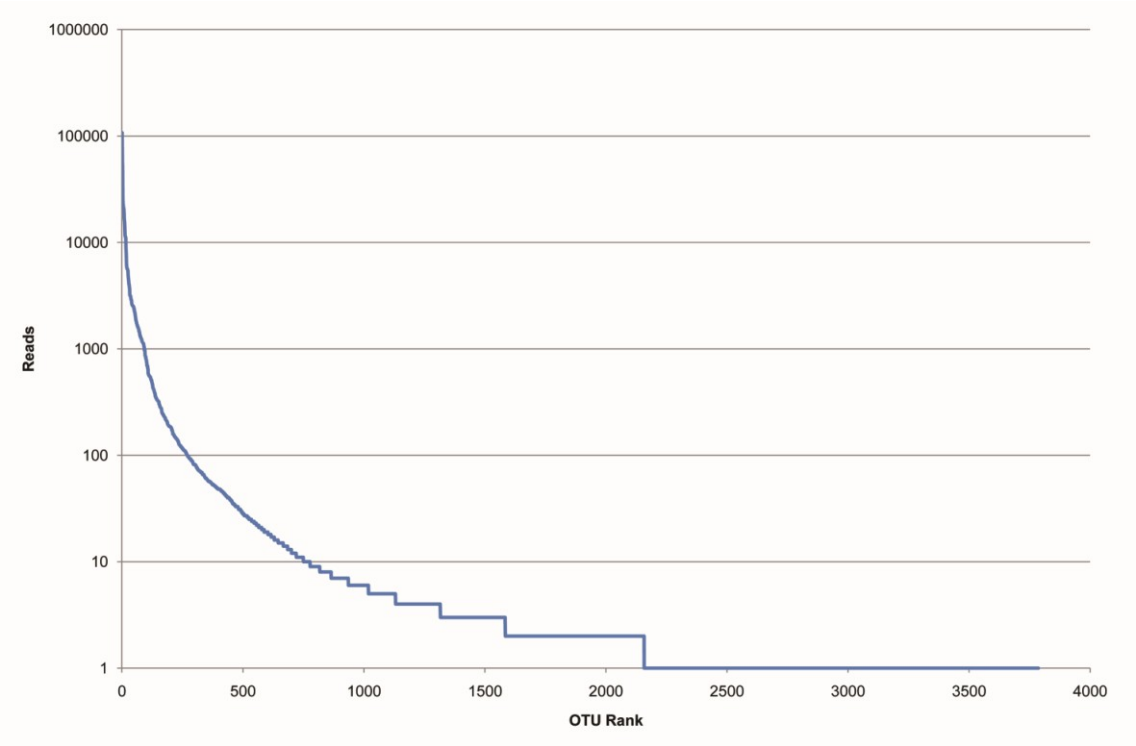


Table 9. OTUs present by proportion of dogs (coverage) carrying each OTU.

Coverage (dogs)	OTUs (3 661)	Reads (671 500)
100% (79)	0.03% (1)	7.88%
90-98.73% (71-78)	0.44% (16)	29.52%
80-90% (63-77)	0.52% (19)	24.10%
70-80% (55-66)	0.55% (20)	9.81%
60-70% (47-54)	0.44% (16)	4.51%
50-60% (40-46)	0.63% (23)	2.36%
40-50% (32-39)	0.93% (34)	2.84%
30-40% (24-31)	1.48% (54)	8.29%
20-30% (16-30)	2.65% (97)	4.45%
10-20% (3-15)	6.50% (238)	3.08%
1.27-10% (2)	36.33% (1330)	2.76%
1.26 (1)	49.52% (1813)	0.4%

Table 10. Distribution and assignment of most widespread OTUs between animals: those present in >90% of animals.

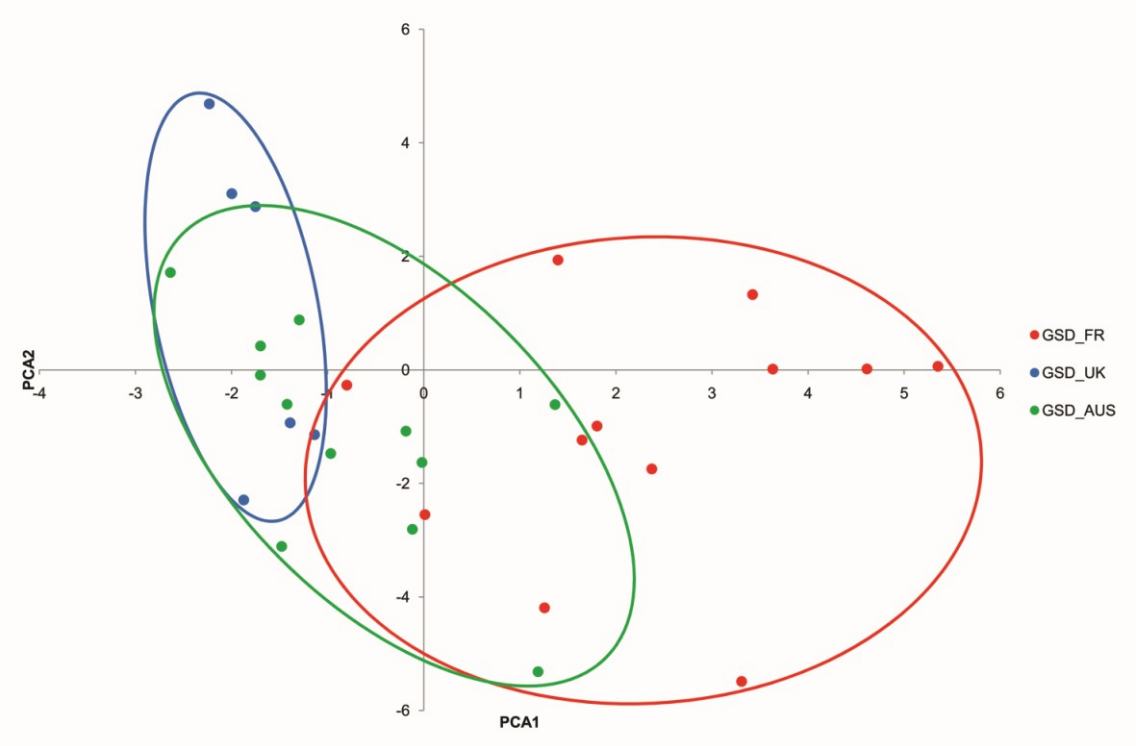
% dogs with this OTU	ave reads per dog	%total reads in OTU	Cumulative fraction of total reads	OTU97 no.	Classifier genus associated with OTU[1]
100.00%(79)	669.57	7.88%	7.88%	OTU97_68	Fusobacterium
98.73%(78)	581.18	6.75%	14.63%	OTU97_112	Cetobacterium
97.47%(77)	308.70	3.54%	18.17%	OTU97_105	Cetobacterium
97.47%(77)	30.45	0.35%	18.52%	OTU97_108	Blautia
97.47%(77)	69.87	0.80%	19.32%	OTU97_23	Sporacetigenium
97.47%(77)	35.42	0.41%	19.72%	OTU97_4	Blautia
96.20%(76)	288.66	3.27%	22.99%	OTU97_215	Cetobacterium
96.20%(76)	257.26	2.91%	25.90%	OTU97_32	Bacteroides
94.94%(75)	51.23	0.57%	26.48%	OTU97_133	Blautia
94.94%(75)	202.77	2.26%	28.74%	OTU97_16	Faecalibacterium
94.94%(75)	21.57	0.24%	28.98%	OTU97_157	Blautia/Dorea
93.67%(74)	21.69	0.24%	29.22%	OTU97_19	Lactonifactor/Coprococcus
93.67%(74)	270.86	2.98%	32.21%	OTU97_21	Sutterella
93.67%(74)	102.34	1.13%	33.33%	OTU97_1	Roseburia
93.67%(74)	214.43	2.36%	35.70%	OTU97_124	Bacteroides/Paraprevotella
92.41%(73)	145.97	1.59%	37.28%	OTU97_9	Bacteroides
91.14%(72)	10.67	0.11%	37.40%	OTU97_110	Roseburia

[1]*The genus assignment is the most frequent named genus of sequences within the OTU identified by the RDP classifier; where another genus is identified for >20% of sequences within the OTU, the second ranking genus name follows the first.*

6.4.4 Environmental influences on dog microbiota

One breed group, the GSDs, was distributed among three different locations and environments, the UK and France groups being housed in uniformly managed kennel environments albeit with different management regimes including diet, while the Australian group were housed and cared for in much more heterogeneous conditions. These groups uniquely can therefore be assessed for any indication of environmental influences on microbiota composition. A PCA analysis of classifier data (Figure 33) indeed shows that the UK GSDs cluster more tightly and can be differentiated from the Australian group, while the France group has some overlap with both clusters. This analysis supports the ANOVA findings outlined above.

Figure 33. PCA analysis of three geographically separated groups of GSDs by genera profile showing weak clustering by location.



6.5 Discussion

These data represent a uniquely large and diverse study of the microbiota of a single species of mammal with extensive and documented variables in genetic lineage (breed), age, sex, diet, environment and management regime. High throughput sequencing has yielded substantial numbers of sequence reads per animal. Use of the AmpliconNoise algorithms for noise reduction and error correction resulted in significant reductions in sequence numbers analysed and in the numbers of reads that could not be classified using the RDP Bayesian Classifier which were reduced by about 50% after AmpliconNoise treatment (data not shown).

Genera identified using the RDP Bayesian Classifier were dominated numerically by the *Bacteroides*, *Prevotella*, *Cetobacterium*, *Fusobacterium* and *Sutterella* genera, all of which are Gram-negative anaerobes. *Cetobacterium* and *Fusobacteria* are related genera within the family *Fusobacteriaceae*; their numerical dominance in dogs is notable, has also been reported elsewhere (Middelbos, Vester Boler et al. 2010) and is not generally echoed in other mammals studied (Eckburg, Bik et al. 2005; Dowd, Sun et al. 2008; Ley, Hamady et al. 2008; Bailey, Dowd et al. 2010; Callaway, Dowd et al. 2010; De Filippo, Cavalieri et al. 2010; Ritchie, Burke et al. 2010). Members of the genus *Fusobacterium* are more widely known as oral species (Karkos, Asrani et al. 2009), and have been implicated as opportunistic pathogens (Huggan and Murdoch 2008). *Sutterella* (Wexler, Reeves et al. 1996) is a little-investigated taxon that with *Parasutterella* branches deeply as family *Sutterellaceae* within the order *Burkholderiaceae* of the beta-*proteobacteria* (Morotomi, Nagai et al. 2010). Such numerical abundance of any member of the *Proteobacterial* phylum is unusual among other mammalian intestinal microbiomes. The *Firmicutes* are noticeably fewer in number than the most abundant taxa, in marked contrast to observations in other host species including humans where they are often the dominant phylum numerically (Ley, Hamady et al. 2008; De Filippo,

Cavalieri et al. 2010). This raises the possibility of bias in our DNA extraction or in primer design for example, that may have skewed numbers in favour of Gram-negatives. In support of our findings however, numerical dominance of Gram-negative taxa in dogs has also been noted by others (Middelbos, Vester Boler et al. 2010). Furthermore, in a minority of animals the *Firmicutes* were much more abundant, suggesting that there was not a systematic failure to recover their DNA adequately. At more moderate levels of abundance, a range of the *Firmicutes* are generally present and make a strong contribution to those widely occurring genera that might be judged to constitute a core microbiome (see below). The more abundant genera of the *Firmicutes* (*Megamonas*, *Faecalibacterium*, *Lactobacillus*, *Blautia*, *Clostridium*) are similar to those observed in other studies.

Definitive knowledge of the richness of the intestinal microbiota is a long-sought goal of many studies and analytical approaches. In reality it is an elusive goal due to the difficulty of culturing many of the organisms present, extreme ranges in abundance of different taxa, and not least variability in both host biology and environmental exposure. In this study we have focused on the *de novo* identified OTUs at 97% sequence similarity in an evaluation of taxon richness. Estimates of richness indicate a wide range between animals but a minimum of approaching 200 and a maximum of 800+ of OTUs, which approximate to species in terms of their sequence divergence. These estimates confirm the widely held view of the extreme complexity of the intestinal microbial community (Ley, Hamady et al. 2008; Marchesi 2010). It is striking that as carnivores, dogs might be expected to exhibit less richness in faecal microbiota than herbivores (Ley, Hamady et al. 2008); hence the diversity seen here is likely to be modest compared with that of some herbivores. However even at a read number of 8,500 per sample, rarefaction curves clearly show that diversity is not fully sampled and it is

clear that organisms present at fewer than 0.01% of the total microbial population will be subject to large stochastic effects in their detection.

A vigorous debate is ongoing in the literature regarding the existence of 'core microbiomes' in digestive tract contents of animals and especially humans (Turnbaugh and Gordon 2009; Turnbaugh, Hamady et al. 2009; Zaura, Keijser et al. 2009; Qin, Li et al. 2010). The rationale for belief in such a microbial community is that many of the intestinal bacteria are postulated to have evolved to occupy a specific host-associated niche and to be vertically acquired soon after birth from family or associated host individuals. A mutualistic relationship with the host is sometimes also postulated and would imply an expectation that such a core population would be widespread among individuals. There is no clear definition of such a community and some have merely sought evidence for taxa that are present in all members of a particular study group, regardless of their abundance. In contrast it could be argued that presence of taxa is biologically meaningful only when numbers of bacteria are sufficient to exert a biological impact on the host. In this study we have sought evidence of a core community with only limited success, since only one OTU was universally present (although not always numerous), and other widely present OTUs were few in number and often not numerically dominant. While this evidence for a core structure is thin, few other studies have provided better. A more fruitful approach to definition of a 'functional' core to the microbiota at the whole population level may be to seek a core metagenome, whereby microbial gene functions that contribute to a mutualistic relationship with the host may be more easily identified than specific microbial taxa that carry these genes. Such an approach has been informative in a recent metagenomics study of the human intestinal metagenome (Qin, Li et al. 2010).

The dominant influences on intestinal microbiome composition remain unclear, partly due to the lack of comprehensive knowledge of the exact makeup of the microbiota in different circumstances. Numerous variables have been postulated to play a part: diet (Hildebrandt, Hoffmann et al. 2009; Barry, Wojcicki et al. 2010; Callaway, Dowd et al. 2010; De Filippo, Cavalieri et al. 2010; Middelbos, Vester Boler et al. 2010), environmental exposure (De Filippo, Cavalieri et al. 2010) and age (Claesson, Cusack et al. 2010) being leading candidates among others (Dethlefsen, Huse et al. 2008; Garner, Antonopoulos et al. 2009; Suchodolski, Dowd et al. 2009; Bailey, Dowd et al. 2010). Host effects, perhaps correlated to genetic determinants in host individuals, have also been proposed (Turnbaugh, Hamady et al. 2009). In this study the animal population included different breeds which are relatively inbred and genetically distinct lines, of a variety of ages, and maintained on a range of diets and in several different environments, with numbers of animals that may in some cases give sufficient power to discern statistically significant effects. We have extensively tested the data to seek such correlations and in many cases trends that could be perceived by eye were not statistically significant. Exceptions however were in several lines of analysis that showed Labrador retrievers to differ from all other breeds in the composition and richness of their microbiota; however, the majority of these dogs were all on the same diet so it is unclear whether diet or breed were driving these differences. In addition, GSDs from the UK differed in their microbiota from the other two GSD groups located in France and Australia. Again, pinpointing the key differences was confounded by differences in both environment and diet.

We have provided clear evidence of technical robustness in the process of data acquisition from faecal DNA samples, by replicating the technical processes of amplification, sequencing and data analysis. The technical variation seen in these procedures, which resemble those

widely used by others in the field, was clearly acceptable in comparison with the variations between samples; as far as we are aware this is the first report of an assessment of technical reproducibility in analyses of this kind.

This study illustrates the power of high throughput sequencing in generating data of a depth and complexity that far exceeds any derived from previously available technologies. The challenges of mining and fully interpreting the data remain significant, but we believe the conclusions presented here contribute significantly to understanding of the canine microbiota.

6.6 Acknowledgements

We acknowledge Nick Loman and Chris Quince for bioinformatics support, Alison Colyer and George Gettinby for statistical support and advice, BBSRC and WALTHAM® Centre for Pet Nutrition for financial support, and Ribosomal Database Project for computing facilities.

CHAPTER SEVEN: GENERAL DISCUSSION

The study of the gut microbiota has been a continuing challenge for microbiologists for decades. With hindsight it is easy to see why efforts were so hampered. The limitations of culture methodologies restricted investigators to assaying a small minority of the thousands of bacterial species present.

The development of large scale 16S rRNA gene clone sequencing projects in the post-genomic era provided our first culture independent look into the mammalian gut microbiota (Eckburg, Bik et al. 2005). Chapters 3 and 4 present data from the amplification, cloning and sequencing of 4 breed faecal DNA pools that produced 777 DNA sequence reads (miniature Schnauzer: 141 reads and cocker spaniels , Labrador and Papillons: 212 reads each). Across the breeds 44 genera were identified. Many of these genera had been previous identified in previous work on the canine gut microbiota (Greetham, Giffard et al. 2002; Suchodolski, Camacho et al. 2008; Middelbos, Vester Boler et al. 2010). Members of the *Firmicutes*, *Fusobacteria* and *Bacteroides* dominated the microbiota. The prevalence of *Fusobacteria* within the dataset, unusual for mammalian gut microbiotas confirmed the findings of others (Suchodolski, Camacho et al. 2008; Middelbos, Vester Boler et al. 2010). Although, the *Firmicutes* were more abundant than observed by Middelbos and Suchodolski.

Because samples from individuals were pooled the hypotheses of a, dietary effects and host genotype were could not be tested on these datasets. Testing for a core microbiome, only 16% of OTUs were found to be shared between the cocker spaniel, Labrador and Papillion pooled libraries. Even with the positive bias of using pooled breed libraries little diversity was shared

indicating that the 'core microbiome' could be smaller in the individual dogs sampled for the pool. Associations between breed and microbiota were assessed. The Labrador group seemed to differ from the rest of the breeds by having higher microbial diversity and enrichment for *Lactobacillus* and *Allobaculum* genera while being depleted of *Fusobacteria*.

It was clear that to begin testing the hypotheses outlined in section 1.4.1 (page 20) a large increase in sampling capability was needed as 16S rRNA gene library sequencing was laborious and expensive. With the advent of throughput sequencing technologies real progress is being made within the field. Key questions such as the existence of a core microbiome at the species or gene level, influence of the host diet and the association of the microbiota with disease are beginning to be assessed (Turnbaugh, Hamady et al. 2009; Zaura, Keijser et al. 2009; Turnbaugh, Quince et al. 2010). However, while we now have the tools to begin to answer these questions, current technologies are still insufficient. Bias is still introduced when using amplified 16S rRNA gene sequence for analysis (v. Wintzingerode, Gobel et al. 1997). One answer is to use shotgun sequencing to gain a metagenomic gene profile plus a phylotype profile with 16S rRNA gene and other phylogenetic markers (Huson, Auch et al. 2007). This approach would seem to be the method of the future. However, currently the coverage of this approach is much lower than that seen in 16S rRNA gene amplicon studies.

These developments in sequencing technology and associated data analysis have had a strong influence on this project. The Sanger sequencing work described in chapters 3 and 4 is largely superseded by the later pyrosequencing work of chapters 5 and 6. There is a ~785 fold increase in sampling effort between the miniature Schnauzers clone library and

pyrosequencing datasets (chapters 3 and 5). The ability to individually sequence every sample from the study population, rather than breed sample pools provided substantial advantages.

Chapter 5 presents the results of a high-throughput 16S rRNA gene pyrosequencing analysis of the faecal microbiota of 11 miniature Schnauzer that were originally analysed as a pool in Chapter 2. This work, excluding that of Chapter 6 represents the largest study of the canine microflora to date. An analysis of identified genera with known pathogens was also done. This study population was regularly monitored for these pathogens via culture techniques and some had not been detected before. This is unsurprising considering the comparatively deep sequencing these samples received. Indeed, culturable genera that were not readily detectable in previous 16S rRNA gene sequencing studies were detected as part of the "rare biosphere" in the dataset (Sogin, Morrison et al. 2006). However, little significance can be attributed to the detection of these potentially pathogenic genera at low levels. Considerable variation was observed between individuals, indicating the flaws in the pooling approach of the clone library sequencing of Chapters 3 and 4. Despite having detailed knowledge of genetic relatedness of the miniature Schnauzer group, no significant clustering by host genotype was observed as has been seen by others (Zoetendal 2001).

The development of high throughput sequencing technologies also provided a number of challenges in terms of sequence error and data analysis. Full length 16S rRNA gene sequencing described in chapter 4 assesses the accuracy of the taxa classification of the pyrosequencing work as compared to taxa assignment of full length 16S rRNA gene clones by Phylogenetic tree reconstruction. The results indicated that pyrosequencing read classifications agreed with the full length sequence analysis at the family level and frequently

at the genus level. It also shows the relationship of some of the most abundant pyrosequencing taxa with culture strains. The sheer amount of sequence data produced by pyrosequencing exceeded the capacity of the previously developed analysis tools used by the community. The determination of OTUs and reconstruction of phylogenetic trees has become particularly computationally intensive.

The progression in data analysis capabilities can be seen between chapters 5 and 6. Initially, the dataset was limited to just miniature Schnauzers (chapter 5) due to the computation limits of OTU determination for the entire 365 megabase sequence dataset. However, in solving the problem of pyrosequencing noise in chapter 6, the de-noised data set became more manageable with regard to total dataset OTU determination. This was due to a substantial reduction in unique reads by using the AmpliconNoise algorithm (Quince, Lanzén et al. 2009). This enabled OTUs to be determined for the whole dataset and allowed, along with a more in depth statistical analysis, a substantial improvement on chapter 5 or any of the previous results chapters.

Chapter 6 details the high-throughput sequencing of 79 dogs across 6 breeds. This is the largest canine and one of the largest mammalian populations to have been profiled to date. This finally allowed the core microbiome hypothesis to be robustly tested. A limited core microbiome was defined in 90% of the study population; this represented less than 0.5% of richness but more than 37.4% of abundance. How biologically significant this core is remains to be seen. Relatively small core microbiomes have been observed in humans at the organismal level and there has been a move to redefine the hypothesis at the gene level (Turnbaugh, Hamady et al. 2009; Turnbaugh, Quince et al. 2010). One proposed reason that a

large and universal core microbiome is not present at the organismal level is that a large amount of functional redundancy exists within the microbiota. It has been shown that individuals acquire distinct and persistent microbial communities from birth (Palmer, Bik et al. 2007; Dethlefsen, Huse et al. 2008). Any number of microbial colonisers might establish within an individual, performing the same role that a one or more bacteria fulfil in another.

Some evidence was found for breed associated richness differences, most marked in Labradors and miniature Schnauzers. Out of all the breeds the Labradors also seemed to have the most distinct flora. Although, this distinction was not particularly strong. The most abundant genera of the faecal microbiota were the *Bacteroides*, *Prevotella*, *Cetobacterium*, *Fusobacterium*, *Sutterella* and *Megamonas*.

The work described in this PhD thesis is the largest and most comprehensive study of the canine gut microbiota to date. The multi-breed, multi-location study population allowed a robust census of the canine microbiota to be conducted. This diverse dataset also allowed the concept of a core microbiome to be rigorously assessed. It is clear that this core is quite small in terms of richness, there nonetheless exists an abundant core spread throughout the study population. While such a diverse study population has its value, assessment of the effects of factors such as diet and location on the microbiota were hampered by lack of specific experimental design to test these hypotheses. Future work will benefit from access to better sequencing technology that will enable a quicker and more numerous work in the area.

REFERENCES

Acke, E., K. McGill, et al. (2009). "Prevalence of thermophilic *Campylobacter* species in household cats and dogs in Ireland." Veterinary Record **164**(2): 44-47.

Ashelford, K. E., N. A. Chuzhanova, et al. (2005). "At Least 1 in 20 16S rRNA Sequence Records Currently Held in Public Repositories Is Estimated To Contain Substantial Anomalies." Applied and Environmental Microbiology **71**(12): 7724-7736.

Backhed, F., R. E. Ley, et al. (2005). "Host-Bacterial Mutualism in the Human Intestine." Science **307**(5717): 1915.

Bailey, M. T., S. E. Dowd, et al. (2010). "Stressor exposure disrupts commensal microbial populations in the intestines and leads to increased colonization by Citrobacter rodentium." Infection and Immunity **78**(4): 1509-1519.

Baker, G. C., J. J. Smith, et al. (2003). "Review and re-analysis of domain-specific 16S primers." Journal of Microbiological Methods **55**(3): 541-555.

Balish, E., D. Cleven, et al. (1977). "Nose, throat, and fecal flora of beagle dogs housed in "locked" or "open" environments." Applied and Environmental Microbiology **34**(2): 207-221.

Barry, K. A., B. J. Wojcicki, et al. (2010). "Dietary cellulose, fructooligosaccharides, and pectin modify fecal protein catabolites and microbial populations in adult cats." Journal of Animal Science **88**(9): 2978-2987.

Beasley, S. S., T. J. K. Manninen, et al. (2006). "Lactic acid bacteria isolated from canine faeces." Journal of Applied Microbiology **101**(1): 131-138.

Callaway, T. R., S. E. Dowd, et al. (2010). "Evaluation of the bacterial diversity in the rumen and feces of cattle fed diets containing levels of dried distiller's grains plus solubles using bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP)." Journal of Animal Science.

Chao, A. (1984). "Nonparametric estimation of the number of classes in a population." Scandinavian journal of statistics **11**(4): 265-270.

Chung, H. and D. L. Kasper (2010). "Microbiota-stimulated immune mechanisms to maintain gut homeostasis." Current Opinion in Immunology **22**(4): 455-460.

Ciccarelli, F. D., T. Doerks, et al. (2006). "Toward Automatic Reconstruction of a Highly Resolved Tree of Life." Science **311**(5765): 1283-1287.

Claesson, M. J., S. Cusack, et al. (2010). "Microbes and Health Sackler Colloquium: Composition, variability, and temporal stability of the intestinal microbiota of the elderly." Proceedings of the National Academy of Sciences

Claesson, M. J., O. O'Sullivan, et al. (2009). "Comparative Analysis of Pyrosequencing and a Phylogenetic Microarray for Exploring Microbial Community Structures in the Human Distal Intestine." PLoS ONE **4**(8): e6669.

Cole, J. R., B. Chai, et al. (2006). "The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data." Nucleic Acids Research: gkl889.

Cole, J. R., B. Chai, et al. (2007). "The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data." Nucl. Acids Res. **35**(suppl_1): D169-172.

Cole, J. R., B. Chai, et al. (2005). "The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis." Nucleic Acids Research **33**(suppl_1): D294-296.

Cole, J. R., Q. Wang, et al. (2009). "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis." Nucleic Acids Research **37**(suppl_1): D141-145.
De Filippo, C., D. Cavalieri, et al. (2010). "Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa." Proceedings of the National Academy of Sciences **107**(33): 14691-14696.
Dethlefsen, L., S. Huse, et al. (2008). "The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing." PLoS Biology **6**(11): e280.
Dowd, S. E., Y. Sun, et al. (2008). "Bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP) for microbiome studies: bacterial diversity in the ileum of newly weaned Salmonella-infected pigs." Foodborne Pathog Dis **5**(4): 459-472.
Duncan, S. H., P. Louis, et al. (2007). "Cultivable bacterial diversity from the human colon." Letters in Applied Microbiology **44**(4): 343-350.
Eckburg, P. B., E. M. Bik, et al. (2005). "Diversity of the Human Intestinal Microbial Flora." Science **308**(5728): 1635-1638.
Edwards, R. A., B. Rodriguez-Brito, et al. (2006). "Using pyrosequencing to shed light on deep mine microbial ecology." BMC Genomics **7**(1): 57.
Ewing, B., L. D. Hillier, et al. (1998). "Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment." Genome research **8**(3): 175.
Flint, H. J., Duncan, S.H., Scott, J.P., Louis, P. (2007). "Interactions and competition within the microbial community of the human colon: links between diet and health." Environmental Microbiology **9**(5): 1101-1111.
Frank, J. A., C. I. Reich, et al. (2008). "Critical Evaluation of Two Primers Commonly Used for Amplification of Bacterial 16S rRNA Genes." Applied and Environmental Microbiology **74**(8): 2461-2470.
Friswell, M. K., H. Gika, et al. (2010). "Site and strain-specific variation in gut microbiota profiles and metabolism in experimental mice." PLoS ONE **5**(1): e8584.
Garner, C. D., D. A. Antonopoulos, et al. (2009). "Perturbation of the small intestine microbial ecology by streptomycin alters pathology in a *Salmonella enterica* serovar typhimurium murine model of infection." Infection and Immunity **77**(7): 2691-2702.
Garrett, W. S., J. I. Gordon, et al. (2010). "Homeostasis and inflammation in the intestine." Cell **140**(6): 859-870.
Gerard, M. and S. Kornelia (1998). "Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology." Antonie van Leeuwenhoek **73**(1): 127-141.
Gibson, G. R. and M. B. Roberfroid (1995). "Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics." The Journal of nutrition **125**(6): 1401-1412.
Gill, S. R., M. Pop, et al. (2006). "Metagenomic analysis of the human distal gut microbiome." Science **312**(5778): 1355-1359.
Gionchetti, P., F. Rizzello, et al. (2000). "Oral bacteriotherapy as maintenance treatment in patients with chronic pouchitis: a double-blind, placebo-controlled trial." Gastroenterology **119**(2): 305-309.
Gordon, H. L. a. L. P. (1971). "The Gnotobiotic Animal as a Tool in the Study of Host Microbial Relationships." Bacteriological Reviews **35**(4): 390-429.
Green, B. D. and M. Keller (2006). "Capturing the uncultivated majority." Current Opinion in Biotechnology
Environmental biotechnology/Energy biotechnology **17**(3): 236-240.

Greetham, H., G. Gibson, et al. (2004). "*Allobaculum stercoricanis* gen. nov., sp. nov., isolated from canine feces." Anaerobe **10**(5): 301-307.
Greetham, H. L. (2003). "Diversity studies of the canine gastrointestinal microbiota." Theses. Reading University. School of Food Biosciences.
Greetham, H. L., M. D. Collins, et al. (2004). "*Sutterella stercoricanis* sp. nov., isolated from canine faeces." International Journal of Systematic and Evolutionary Microbiology **54**(5): 1581-1584.
Greetham, H. L., C. Giffard, et al. (2002). "Bacteriology of the Labrador dog gut: a cultural and genotypic approach." Journal of Applied Microbiology **93**(4): 640-646.
Heck Jr, K. L., G. van Belle, et al. (1975). "Explicit Calculation of the Rarefaction Diversity Measurement and the Determination of Sufficient Sample Size." Ecology **56**(6): 1459-1461.
Hildebrandt, M. A., C. Hoffmann, et al. (2009). "High-fat diet determines the composition of the murine gut microbiome independently of obesity." Gastroenterology **137**(5): 1716-1724 e1711-1712.
Hill, J. E., S. M. Hemmingsen, et al. (2004). "Comparison of Ileum Microflora of Pigs Fed Corn-, Wheat-, or Barley-Based Diets by Chaperonin-60 Sequencing and Quantitative PCR." Applied and Environmental Microbiology **71**(2): 867-875.
Hill, T. C. J., Walsh, K.A., Harris, J.A., Moffett, B.F. (2006). "Using ecological diversity measures with bacterial communities." FEMS Microbiology Ecology **43**(1): 1-11.
Hold, G. L., S. E. Pryde, et al. (2002). "Assessment of microbial diversity in human colonic samples by 16S rDNA sequence analysis." FEMS Microbiology Ecology **39**(1): 33-39.
Hooper, L. V. and J. I. Gordon (2001). "Commensal host-bacterial relationships in the gut." Science **292**(5519): 1115.
Huggan, P. J. and D. R. Murdoch (2008). "Fusobacterial infections: clinical spectrum and incidence of invasive disease." Journal of Infection **57**(4): 283-289.
Hughes, J. B. and J. J. Hellmann (2005). "The application of rarefaction techniques to molecular inventories of microbial diversity." Methods in Enzymology **397**: 292-308.
Hughes, J. B., J. J. Hellmann, et al. (2001). "Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity." Applied and Environmental Microbiology **67**(10): 4399.
Huson, D. H., A. F. Auch, et al. (2007). "MEGAN analysis of metagenomic data." Genome Research **17**(3): 377-386.
Karkos, P. D., S. Asrani, et al. (2009). "Lemierre's syndrome: A systematic review." Laryngoscope **119**(8): 1552-1559.
Kelly, P. (2010). "Nutrition, intestinal defence and the microbiome." Proceedings of the Nutrition Society **69**(2): 261-268.
Konstantinidis, K. T. and J. M. Tiedje (2005). "Genomic insights that advance the species definition for prokaryotes." Proceedings of the National Academy of Sciences **102**(7): 2567-2572.
Konstantinidis, K. T. and J. M. Tiedje (2005). "Towards a Genome-Based Taxonomy for Prokaryotes." Journal of Bacteriology **187**(18): 6258-6264.
Konstantinidis, K. T. and J. M. Tiedje (2007). "Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead." Current Opinion in Microbiology **10**(5): 504-509.
Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947.

Lenz, J., D. Joffe, et al. (2009). "Perceptions, practices, and consequences associated with foodborne pathogens and the feeding of raw meat to dogs." Canadian Veterinary Journal **50**(6): 637-643.
Ley, R. E., F. Bäckhed, et al. (2005). "Obesity alters gut microbial ecology." Proceedings of the National Academy of Sciences **102**(31): 11070-11075.
Ley, R. E., M. Hamady, et al. (2008). "Evolution of Mammals and Their Gut Microbes." Science **320**(5883): 1647-1651.
Ludwig, W., O. Strunk, et al. (2004). "ARB: a software environment for sequence data." Nucleic Acids Research **32**(4): 1363.
Macpherson, A. J., D. Gatto, et al. (2000). "A primitive T cell-independent mechanism of intestinal mucosal IgA responses to commensal bacteria." Science **288**(5474): 2222.
Madsen, K. L., J. S. Doyle, et al. (1999). "Lactobacillus species prevents colitis in interleukin 10 gene-deficient mice." Gastroenterology **116**(5): 1107-1114.
Makoto, M. and N. Takuji (1998). "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator." ACM Transactions on Modeling and Computer Simulation **8**(1): 3-30.
Marchesi, J. R. (2010). "Prokaryotic and eukaryotic diversity of the human gut." Advances in Applied Microbiology **72**: 43-62.
Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.
Martin-Orue, S. M., A. G. O'Donnell, et al. (2002). "Degradation of transgenic DNA from genetically modified soya and maize in human intestinal simulations." British Journal of Nutrition **87**(6): 533-542.
Middelbos, I. S., B. M. Vester Boler, et al. (2010). "Phylogenetic characterization of fecal microbial communities of dogs fed diets with or without supplemental dietary fiber using 454 pyrosequencing." PLoS ONE **5**(3): e9768.
Morotomi, M., F. Nagai, et al. (2010). "Parasutterella secunda sp. nov., isolated from human faeces and proposal of Sutterellaceae fam. nov. in the order Burkholderiales." International Journal of Systematic and Evolutionary Microbiology.
O'Hara, A. M. and F. Shanahan (2006). "The gut flora as a forgotten organ." EMBO Reports **2006**: 688-693.
Okada, H., C. Kuhn, et al. (2010). "The 'hygiene hypothesis' for autoimmune and allergic diseases: an update." Clinical & Experimental Immunology **160**(1): 1-9.
Pace, N. R. (1997). "A molecular view of microbial diversity and the biosphere." Science **276**(5313): 734-740.
Pallen, M. J., N. J. Loman, et al. (2010). "High throughput sequencing and clinical microbiology: progress, opportunities and challenges." Current Opinion in Microbiology **13**: 1-7.
Palmer, C., E. M. Bik, et al. (2007). "Development of the human infant intestinal microbiota." PLoS Biology **5**(7): e177.
Paster, B. J., S. K. Boches, et al. (2001). "Bacterial Diversity in Human Subgingival Plaque." Journal of Bacteriology **183**(12): 3770-3783.
Peplies, J., F. O. Glockner, et al. (2003). "Optimization Strategies for DNA Microarray-Based Detection of Bacteria with 16S rRNA-Targeting Oligonucleotide Probes." Applied and Environmental Microbiology **69**(3): 1397-1407.
Peterson, J., S. Garges, et al. (2009). "The NIH Human Microbiome Project." Genome Research **19**(12): 2317-2323.

Pruesse, E., C. Quast, et al. (2007). "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB." Nucleic Acids Research **35**(21): 7188.
Qin, J., R. Li, et al. (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65.
Quince, C., A. Lanzén, et al. (2009). "Accurate determination of microbial diversity from 454 pyrosequencing data." Nature Methods **6**(9): 639-641.
Rastall, R. A. (2004). "Bacteria in the Gut: Friends and Foes and How to Alter the Balance." The Journal of nutrition **134**(8): 2022S-2026.
Ritchie, L. E., K. F. Burke, et al. (2010). "Characterization of fecal microbiota in cats using universal 16S rRNA gene and group-specific primers for Lactobacillus and Bifidobacterium spp." Vet Microbiol **144**(1-2): 140-146.
Ronaghi, M., S. Karamohamed, et al. (1996). "Real-Time DNA Sequencing Using Detection of Pyrophosphate Release." Analytical Biochemistry **242**(1): 84-89.
Rossello-Mora, R. and R. Amann (2001). "The species concept for prokaryotes." FEMS Microbiology Reviews **25**(1): 39-67.
Russell, W. R., S. W. Gratz, et al. (2011). "High-protein, reduced-carbohydrate weight-loss diets promote metabolite profiles likely to be detrimental to colonic health." The American Journal of Clinical Nutrition **93**(5): 1062.
Santos, S. R. and H. Ochman (2004). "Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins." Environmental Microbiology **6**(7): 754-759.
Savage, D. C. (1977). "Microbial ecology of the gastrointestinal tract." Annual Review of Microbiology **31**: 107-133.
Schloss, P. D., S. L. Westcott, et al. (2009). "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." Applied and Environmental Microbiology **75**(23): 7537-7541.
Sekirov, I., S. L. Russell, et al. (2010). "Gut microbiota in health and disease." Physiological Reviews **90**(3): 859-904.
Shannon, C. E. and W. Weaver (1963). "The mathematical theory of communication." Urbana: University of Illinois Press.
Simpson, J. M., B. Martineau, et al. (2002). "Characterization of Fecal Bacterial Populations in Canines: Effects of Age, Breed and Dietary Fiber." Microbial Ecology **V44**(2): 186-197.
Sogin, M. L., H. G. Morrison, et al. (2006). "Microbial diversity in the deep sea and the underexplored" rare biosphere"." Proceedings of the National Academy of Sciences **103**(32): 12115.
Stecher, B., Dardt, W.D. (2008). "The role of microbiota in infectious disease." Trends in Microbiology **16**(3): 107-114.
Steinert, M., U. Hentschel, et al. (2000). "Symbiosis and pathogenesis: evolution of the microbe-host interaction." Naturwissenschaften **87**(1): 1-11.
Strachan, D. P. (1989). "Hay fever, hygiene, and household size." British Medical Journal **299**(6710): 1259.
Suau, A., R. Bonnet, et al. (1999). "Direct Analysis of Genes Encoding 16S rRNA from Complex Communities Reveals Many Novel Molecular Species within the Human Gut." Applied and Environmental Microbiology **65**(11): 4799-4807.

Suchodolski, J., P. Xenoulis, et al. (2010). "Molecular analysis of the bacterial microbiota in duodenal biopsies from dogs with idiopathic inflammatory bowel disease." Veterinary microbiology **142**(3-4): 394-400.
Suchodolski, J. S., J. Camacho, et al. (2008). "Analysis of bacterial diversity in the canine duodenum, jejunum, ileum, and colon by comparative 16S rRNA gene analysis." FEMS Microbiology Ecology **66**(3): 567-578.
Suchodolski, J. S., S. E. Dowd, et al. (2009). "The effect of the macrolide antibiotic tylosin on microbial diversity in the canine small intestine as demonstrated by massive parallel 16S rRNA gene sequencing." BMC Microbiology **9**: 210.
Thompson, J. D., T. J. Gibson, et al. (1997). "The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality tools." Nucleic Acids Research **24**: 4876-4882.
Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Research **22**(22): 4673-4680.
Tringe, S. G. and P. Hugenholtz (2008). "A renaissance for the pioneering 16S rRNA gene." Current Opinion in Microbiology **11**(5): 442-446.
Turnbaugh, P., C. Quince, et al. (2010). "Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins." Proceedings of the National Academy of Sciences **107**(16): 7503.
Turnbaugh, P. J. and J. I. Gordon (2009). "The core gut microbiome, energy balance and obesity." The Journal of Physiology **587**(Pt 17): 4153-4158.
Turnbaugh, P. J., M. Hamady, et al. (2009). "A core gut microbiome in obese and lean twins." Nature **457**(7228): 480-484.
Turnbaugh, P. J., R. E. Ley, et al. (2006). "An obesity-associated gut microbiome with increased capacity for energy harvest." Nature **444**(7122): 1027-1031.
v. Wintzingerode, F., U. B. Gobel, et al. (1997). "Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis." FEMS Microbiology Reviews **21**(3): 213-229.
Van de Peer, Y., S. Chapelle, et al. (1996). "A quantitative map of nucleotide substitution rates in bacterial rRNA." Nucleic Acids Research **24**(17): 3381.
Vanhoutte, T., G. Huys, et al. (2005). "Molecular monitoring and characterization of the faecal microbiota of healthy dogs during fructan supplementation." FEMS Microbiology Letters **249**(1): 65-71.
Venter, J. C., K. Remington, et al. (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science **304**(5667): 66-74.
von Mering, C., P. Hugenholtz, et al. (2007). "Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments." Science **315**(5815): 1126.
Wang, Q., G. M. Garrity, et al. (2007). "Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy." Applied and Environmental Microbiology **73**(16): 5261-5267.
Wang, X., S. P. Heazlewood, et al. (2003). "Molecular characterization of the microbial species that colonize human ileal and colonic mucosa by using 16S rDNA sequence analysis." Journal of Applied Microbiology **95**(3): 508-520.
Watanabe, K., Y. Kodama, et al. (2001). "Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting." Journal of Microbiological Methods **44**(3): 253-262.

Wexler, H. M., D. Reeves, et al. (1996). "Sutterella wadsworthensis gen. nov., sp. nov., bile-resistant microaerophilic Campylobacter gracilis-like clinical isolates." International Journal of Systematic and Evolutionary Microbiology **46**(1): 252-258.

Wiedemann, S., B. Lutz, et al. (2006). "In situ studies on the time-dependent degradation of recombinant corn DNA and protein in the bovine rumen." Journal of Animal Science **84**(1): 135-144.

Winter, S. E., A. M. Keestra, et al. (2010). "The blessings and curses of intestinal inflammation." Cell Host Microbe **8**(1): 36-43.

Woese, C. R. (1987). "Bacterial evolution." Microbiology and Molecular Biology Reviews **51**(2): 221-271.

Xenoulis, P., B. Palculict, et al. (2008). "Molecular phylogenetic characterization of microbial communities imbalances in the small intestine of dogs with inflammatory bowel disease." FEMS Microbiology Ecology **66**(3): 579-589.

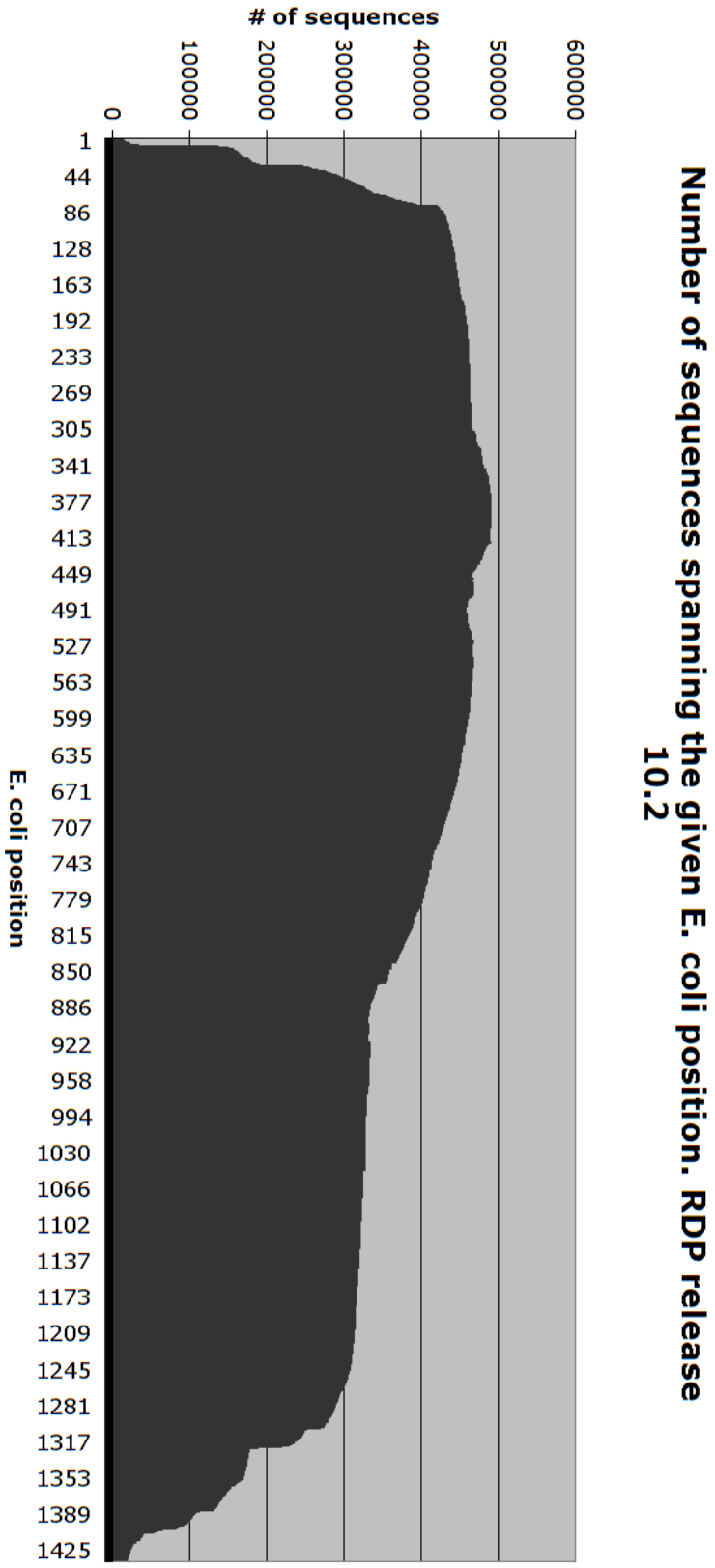
Yarza, P., M. Richter, et al. (2008). "The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains." Systematic and Applied Microbiology **31**(4): 241-250.

Zaura, E., B. J. Keijser, et al. (2009). "Defining the healthy "core microbiome" of oral microbial communities." BMC Microbiology **9**: 259.

Zoetendal, E. G., A. D. L. Akkermans, et al. (1998). "Temperature Gradient Gel Electrophoresis Analysis of 16S rRNA from Human Fecal Samples Reveals Stable and Host-Specific Communities of Active Bacteria." Applied and Environmental Microbiology **64**(10): 3854-3859.

Zoetendal, E. G., Akkermans, A.D.L., Akkermans van-Vliet, W. M., de Visser, J. A.G.M & de Vos, W.M (2001). "The Host Genotype Affects the Bacterial Community in the Human Gastronintestinal Tract." Microbial Ecology in Health and Disease(13): 129-134.

APPENDIX


# of sequences
0
100000
200000
300000
400000
500000
600000
E. coli position
1
44
86
128
163
192
233
269
305
341
377
413
449
491
527
563
599
635
671
707
743
779
815
850
886
922
958
994
1030
1066
1102
1137
1173
1209
1245
1281
1317
1353
1389
1425
Number of sequences spanning the given E. coli position. RDP release 10.2