The hidden diversity of Archaea and Bacteria in the human microbiome.

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

In

Microbiology

At Massey University, Manawatū,

New Zealand.

Nicholas Tyler Dreisbach

2020

Abstract

Current methods of metagenomic analysis require deep sequencing to identify microorganisms that are present at low abundance in complex microbiomes, including the human gut microbiome. The few known archaeal taxa present in the human gut are low in abundance in comparison to bacteria. This raises the question about whether the full diversity of human gut-associated archaea is known.

To increase the resolution of metagenomic analysis, a new DNA normalization technique utilizing duplex specific nuclease (DSN) was used to enrich for DNA from “rare” archaeal and bacterial taxa isolated from two human metagenomic faecal samples. This DSN based normalization method failed to enrich for archaeal DNA, as it was digested by the DSN, however, it succeeded in enriching for low abundance bacterial DNA. This indicated that further optimization of the normalization method is required to enrich for low abundance archaeal DNA in human metagenomic samples.

Whole metagenome shotgun sequencing was also used to identify a microbial community composition of participants gut microbiota including archaea. WGS identified a higher than anticipated diversity of archaeal taxa in gut microbiomes from both participants. Regardless of higher diversity, the low abundance of archaea in the human gut still render them as a part of rare biosphere.

We envisage that with further optimization of DSN-based normalization, enrichment of “rare” taxa will improve detection resolution and therefore enhance our current understanding of the diversity of both archaeal and bacterial species in human gut-microbiome.

Acknowledgements

This research would not have been possible without the expertise and support of my supervisor Dr. Dragana Gagic. I would also like to thank my co-supervisors Dr. Christina Moon and Dr. Zoe Jordens for providing vital input and feedback during the course of this research.

I would also like to thank Dr. Asad Razzaq for his support and expertise with experimental troubleshooting and genome assembly, for this I am very grateful.

I am immeasurably thankful for the support of Plant & Food Research for allowing me to use their laboratory resources, in particular Dr. Shanthi Parkar for her invaluable assistance, support, and feedback with this research.

Finally, I would like to thank the late Hodis Dreisbach for his irreplaceable support throughout this degree. You will be profoundly missed.

List of Abbreviations

dsDNA	Double-Stranded DNA
DSN	Duplex-Specific Nuclease
GMLA	Genome Metagenomic Linker-Amplified
HMP	Human Microbiome Project
ITS	Internal Transcribed Spacer
k-mers	A Nucleotide Sequence of k length
LCA	Least Common Ancestor
LL	Lone-Linker (oligonucleotide tags)
LLL	Lone-Linker-Ligated
LL-PCR	Lone-Linker-PCR
mcrA	Methyl-coenzyme M reductase subunit alpha
MWCO	Molecular Weight Cut Off
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
Q2	Qiime 2 Software Suite
RE	Restriction Enzyme
SNP	Single Nucleotide Polymorphism
ssDNA	Single-Stranded DNA
WGS	Whole Genome Shotgun

Table of Contents

List of Figures

List of Tables

Chapter One: Literature Review

1.1 Introduction

The human body is a host to many different types of microorganisms which live in a variety of different ecological niches including the skin, nose, mouth, gut, and urogenital system. These microorganisms include bacteria, protozoa, fungi, viruses, and archaea (Kho & Lal, 2018). Each of these environments contains a unique community of microorganisms called microbiota. The role of human microbiota in homeostasis has become an important and active field of research in the last 20 years due to the increasing number of detrimental health effects demonstrated in association with dysbiosis (Bhute et al., 2017; Gaci et al., 2014; Ipci et al., 2017; Kho & Lal, 2018; S. Li et al., 2018). The importance of the microbiota and its role in human health cannot be overstated, even if it is not yet fully understood. For example, these microorganisms in effect act as a functional expansion of their hosts genome (Heintz-Buschart & Wilmes, 2018; Kho & Lal, 2018; Mohammed & Guda, 2015). This relationship between microorganisms and the host is characterized by synthesis of vitamins, regulation of the immune system, fermentation of indigestible food constituents into readily host absorbable metabolites, and regulation of host metabolism (Clooney et al., 2019; Heintz-Buschart & Wilmes, 2018; Kho & Lal, 2018; Mohammed & Guda, 2015).

Identification and determination of the relative abundance of microbial species in the human microbiome was originally accomplished *via* culturing. However, this method is biased toward culturable microorganisms, mostly from Bacteria, and does not identify uncultivable microorganisms. However, recent advances in culturing methods, termed "culturomics", has allowed for the cultivation of previously uncultivable microorganisms (J. C. Lagier et al., 2016). These new methods could revitalize culture-based identification and community profiling in the future (J. C. Lagier et al., 2016). Culture-independent methods including phylogenetic marker sequencing and metagenomic shotgun sequencing have proven to be great tools to discover diversity beyond culturable microorganisms (Roh et al., 2010; Rondon et al., 2000; Yarza et al., 2014). In particular, using 16S rRNA as a phylogenetic marker gene, amplicon sequencing has emerged as a standard practice for detection and taxonomic classification of unculterable microorganisms. Despite the knowledge that has been gained from these methods, there remains a problem in detecting and identifying many low abundance species present in the microbiome (Horz, 2015; Jia et al., 2018; Sogin et al., 2006). An example of this is the selective biases introduced by the choice of 16S rRNA oligonucleotides for PCR amplification, as the whole 16S rRNA gene (1.5-kb) cannot be sequenced on second generation sequencing platforms (Bhute et al., 2017; Johnson et al., 2019). This is important because the 16S rRNA gene contains multiple variable regions used for taxonomic assignment, and no single region is able to differentiate between all prokaryotic taxa (Bhute et al., 2017). Another example of this would be archaeal lineages within the human microbiome (Gaci et al., 2014; Horz, 2015; Mihajlovski et al., 2008). Archaea, the third domain of life, share many similarities with both bacteria and eukaryotes (protists, plants, animals). However, they also differ in significant ways, including cell wall composition, which will be discussed in more detail later in this chapter. Archaea are most well-known for species that exist in extreme environments, such as salt lakes and hydrothermal vents. Despite their presence in humans having been demonstrated for some time, for example methanogens, the full extent of their diversity in the human microbiome and effect on health is largely unknown (Moissl-Eichinger et al., 2018; Nkamga et al., 2017a).

Recent advances in DNA normalization techniques enable enriching microbiome samples for sequences representing rare taxa (Gagic et al., 2015; Shagina et al., 2010). The application of DSN (Duplex Specific Nuclease) normalization methods on samples from the human gut may allow for novel low abundance archaeal lineages to be discovered. If such microorganisms are found, it will enhance our current understanding of the human microbiomes' diversity. This could open new avenues for future research into the impact of archaeal species on human health.

1.2 The Rare Human Microbiome

In a “typical” microbiome, the majority of microorganisms present are represented by a relatively small number of taxa (Bhute et al., 2017; Pedros-Alio, 2012). A study reported by Kraal et al. (2014) found that 80% of the total gut microbiota was represented by only 14 species. These highly abundant species are generally well characterized as they are easier for sampling and therefore identification and genetic analysis. Despite these microorganisms being the most abundant, this does not necessarily mean that less abundant species have no impact on the environment they inhabit. This is an area of ongoing research, as we are not yet able to determine the role these rare microorganisms play in, for example, human health (Bhute et al., 2017; Laura Wegener et al., 2011). Another important consideration is that the relative species abundance in each human microbiome changes significantly with variations in diet, environment, immune system, and other factors, however, the overall diversity remains the same (Bhute et al., 2017; David et al., 2014; J. J. Faith et al., 2013; Zarrinpar et al., 2014). These changes can allow for previously low abundance species to displace dominant species by increasing in numbers, which can be seen in the well-known phenomenon of “blooming” (Bhute et al., 2017; Vincent et al., 2016). Blooming can be defined as a large increase in abundance of a microorganism in an environment. This increase can be either transient or sustained.

There has been an ever-present problem with sampling for rare taxa in microbiome samples (Gagic et al., 2015; Sogin et al., 2006). By being in such low abundance, rare species are mainly overlooked when culture-independent methods are used to detect them as the resolution of these techniques is not high enough to reliably allow detection. An example of this would be attempting to identify a single unique strand of DNA in a sample containing millions of other DNA strands from a few highly abundant microorganisms. The probability of getting a detectible signal for the unique strand, even using techniques like PCR, is low.

1.3 Archaea in the Human Microbiome

Another aspect of the microbiome is archaeal species. Archaea have typically been disregarded in medical microbiology because they have never been shown to be pathogenic (Vianna et al., 2006). However, a study by Vianna et al. (2006) demonstrated an association between methanogenic archaea and endodontic infection, but there was no evidence that it was the causative agent. Currently the primary archaea identified in the human gut are methanogens belonging to the order *Methanobacteriales,* including dominant species such as *Methanobrevibacter smithii* and *Methanosphaera stadtmanae* (Bhute et al., 2017; Gaci et al., 2014). Of the two, *M. smithii* was demonstrated to be the most common species present in 95.7% of gut samples out of test population of 700 people, followed by *M. stadtmanae* which was found in 29.4% (Dridi et al., 2009). The relative prevalence of *M. smithii* in the human population appears to be variable, as subsequent studies have reported its carriage being between 64% and 89% in different populations (Million et al., 2013; Million et al., 2012). Also, it was shown that *M. smithii* was not in low abundance as it was estimated to make up roughly 11.5% of the total gut microbiome when detected (Dridi et al., 2009). Given the relative abundance of the few known archaeal species in the human gut, it is possible that this domain could be more diverse than currently known. If such diversity exists, it is possible that these archaeal taxa are low enough abundance to avoid detection by standard culture-independent methods.

1.4 Difficulties Studying Archaea in Microbiomes

With respect to culture-dependant identification methods, in comparison to most bacteria, archaea are more difficult grow. The nutritional and environmental requirements of archaea can be complex (e.g. specific carbon sources and nutrients) or extreme (e.g. high temperature, high hydrostatic pressure, specific atmospheric composition), this makes culture-based detection and identification a challenging laboratory task (J.-C. Lagier et al., 2015; Sun et al., 2020). Therefore, culture-independent methods are currently preferred for the detection and identification of archaeal species in microbiomes (Sun et al., 2020). However, there are considerations which need to be made to successfully adapt these highly sensitive techniques (phylogenetic marker sequencing and metagenomic shotgun sequencing) to archaeal taxa. As many archaea possess a more durable cell wall compared to bacterial species, lysing archaeal cells for DNA extraction is more difficult and labour intensive than bacterial cells. There are several traits that can give archaea stronger cellular envelopes than bacteria, these include an S-layer (also found in many bacterial species), pseudomurein sheath, and a lipid monolayer membrane. For example, the pseudomurein sheath cannot be cleaved by lysozyme, which is a commonly used reagent in DNA extraction kits and protocols (MirMohammad-Sadeghi et al., 2013; Visweswaran et al., 2010). The organization of the cellular envelope can vary significantly between archaeal species (Figure 1). The S-layer is a lattice made out of repeating protein subunits with strong covalent cysteine linkages, which can make the cell envelope resistant to heat and most typical cell lysis methods that rely on membrane disruption (Albers & Meyer, 2011).

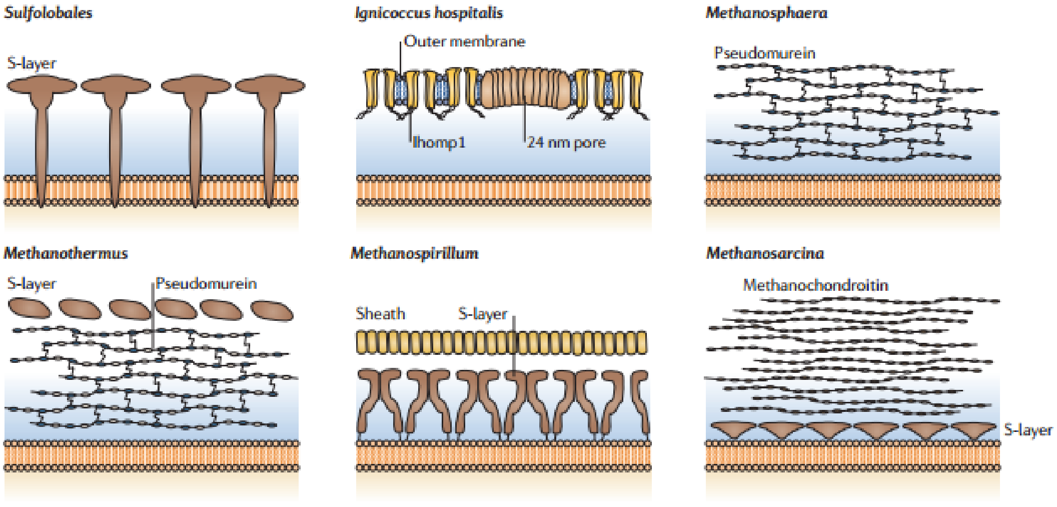


Figure 1: Diagram of Various Archaeal Cell Envelope Arrangements (Adopted with permission from Albers and Meyer (2011)).

As shown above, there are a number of different S-layer proteins, each of which forms a different pseudo-crystalline lattice structure. Another feature of many archaea is a different cell membrane structure. Unlike bacterial and eukaryotic membranes, in which a D-glycerol is ester linked to two fatty acid side chains, archaeal membrane side chains are ether linked to an L-glycerol (Figure 2). Another structural difference is a solid hydrophobic core of a lipid monolayer. Unlike bacterial and eukaryotic membranes which are made up of pairs of diacylglycerols, some archaeal species have membranes where the fatty acid chains are one continuous unit between the glycerol heads; forming a lipid monolayer. This lipid monolayer strengthens the envelope by making it less fluid which improves resistance to harsh environments such as high temperature. Failing to use adequate DNA extraction techniques for metagenomic samples containing archaea can result in low archaeal DNA yields causing inaccurate representation of the microbiome’s diversity (Henderson et al., 2013). To improve lysis of durable archaeal cells, most methods of DNA extraction used on archaea are comparatively harsher than ones used for bacterial cell wall degradation. A common method used on more durable bacteria and also archaea utilizes glass beads to destroy the cellular envelope, although these methods pose a significant risk of shearing any recovered DNA. The outcome could be that recovered DNA is low quality and therefore not suitable for deep sequencing.

Recently there have been several studies carried out to determine an optimal method of metagenomic extraction that produces DNA of high enough quality to allow PCR amplification (Bag et al., 2016; Leuko et al., 2008; Roopnarain et al., 2017; Yu & Morrison, 2004).

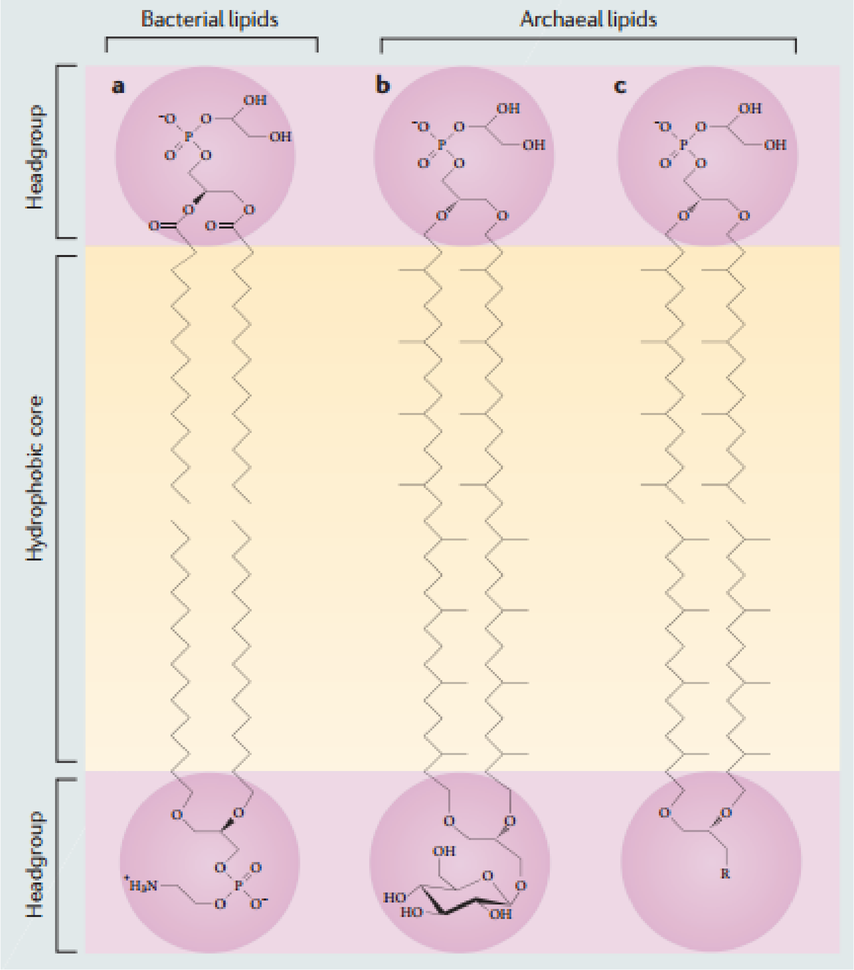


Figure 2: Archaeal and Bacterial Cell Membrane Structure (Adopted with permission from Albers and Meyer (2011)).

One such study carried out by Henderson et al. (2013) demonstrated that statistically significant changes in apparent metagenomic diversity could be directly linked to the extraction method used. The variations between methods, which include various chemical and physical processes to disrupt the microbial cell wall, directly change the amount of DNA recovered from each taxa. Because of the effect of the extraction method, microbiome comparisons should only be done among samples and studies which used the same DNA extraction protocol (Henderson et al., 2013). Of all the methods compared, all have a similar finding in common which is that standard “soft” lysis techniques are inadequate for metagenomic DNA extraction. These “soft” techniques received their name due to their lack of a bead beating step. This bead beating step, while mechanically harsh and which results in varying levels of DNA shearing depending on the method used, is required to get adequate lysis of all microorganisms in the sample (Henderson et al., 2013; Purohit & Singh, 2009; Rondon et al., 2000).

1.5 DNA Normalization to Uncover “Rare Biosphere”

As previously mentioned, in microbiomes the majority of extant microorganisms belong to a relatively small number of taxa. This challenges experiments to determine the diversity of the sampled community as the least abundant taxa are likely to be hidden amongst the highly abundant taxa. To overcome this obstacle, DNA, similarly to subtraction of abundant cDNA transcripts, is normalized. Traditionally, DNA normalization was carried out using hydroxylapatite (HAP) chromatography (Vandernoot et al., 2012). The underlying principle of this method is that low-copy number DNA sequences renature slower (based on spatial separation) than high copy number sequences. After renaturation, double-stranded DNA (dsDNA) is mostly made of abundant sequence duplexes (and rare sequences still as single-strand DNA (ssDNA)) allowing for the physical removal using ion exchange chromatography (Gagic et al., 2015; Shagina et al., 2010; Vandernoot et al., 2012). A newer method, using the enzyme duplex-specific nuclease (DSN) to normalize DNA (Figure 3), similarly relies on DNA renaturation kinetics. However, instead of using a physical removal of dsDNA by HAP, DSN is used to digest dsDNA while leaving ssDNA intact (Gagic et al., 2015).

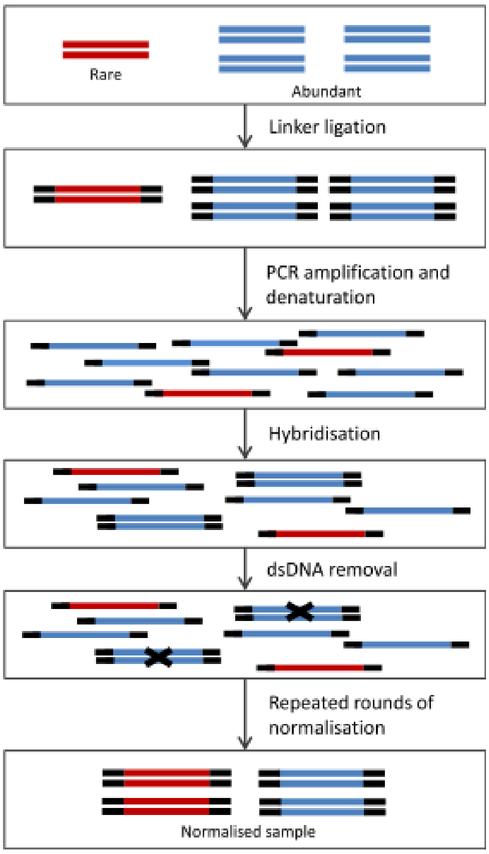


Figure 3: Graphic Diagram of DSN Normalization Method. Initially, all sample DNA is ligated with LL-PCR amplification oligonucleotides. Then the ligated DNA is amplified using LL-PCR. The DNA is then denatured into ssDNA and allowed to slowly re-hybridize, enabling abundant DNA to form dsDNA more quickly than “rare” DNA. At this point the sample is treated with DSN, causing the now hybridized highly abundant DNA to be digested, while leaving the rare DNA intact. (Adopted with permission from Gagic et al. (2015))

The highly abundant DNA sequences are therefore selectively eliminated by DSN, leaving “rare” sequences (ssDNA) intact to increase in proportion over several rounds of normalization (Shagina et al., 2010). Research by Gagic *et al.* (2015) demonstrated that DSN based normalization improved the representation of low abundance microbial species in a synthetic metagenomic sample more than traditional HAP chromatography-based normalization. The relative enrichment of rare species in a synthetic metagenome by DSN normalization in comparison to HAP chromatography is shown below (Figure 4). After a desirable normalization round is achieved (close to equimolar ration, Figure 4 DSN R5), these “rare” sequences are detected and identified *via* high throughput sequencing methods (Gagic et al., 2015).

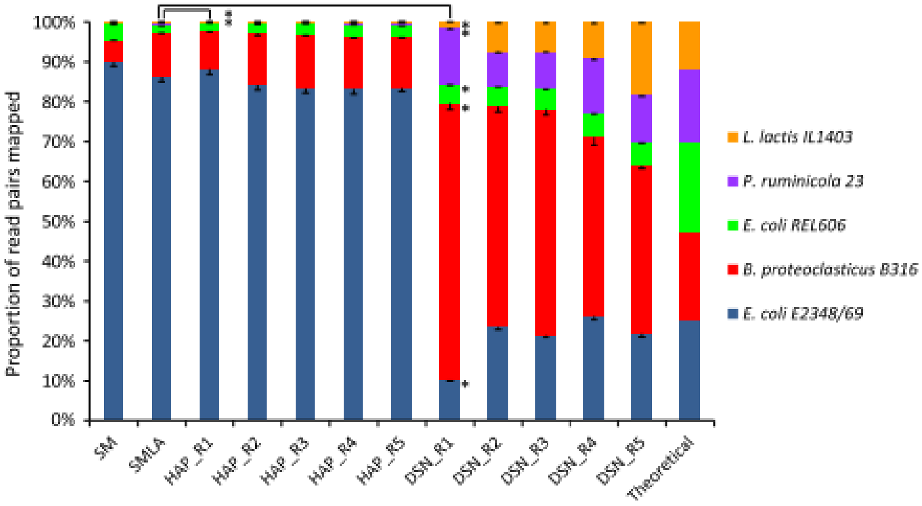


Figure 4: Proportion of each species in the synthetic metagenome (SM), synthetic metagenome linker-amplified (SMLA), HAP normalized (number of rounds indicated by R#), DSN normalized (number of round indicated by R#). This figure compares the effect of multiple rounds of HAP normalization and DSN normalization to the proportion of sequence reads of each species in the synthetic

metagenome. Unlike HAP normalization, DSN normalization shows a higher level of enrichment for low abundances sequences over successive rounds. (Adopted with permission from Gagic et al. (2015)).

Methods such as DSN normalization in combination with DNA extraction methods that are optimized for archaea, have the potential to allow for previously undetected archaeal taxa in the “rare biosphere” of the human gut microbiome to be detected. Future studies using such an approach could also enhance our knowledge of the microbial diversity in any metagenomic sample.

1.6 Metagenomic & Bioinformatic Approaches for Microbial Community Profiling

Metagenomics is a science field that focuses on the study of genetic material extracted from complex environmental samples. There are two distinct approaches to metagenomics, whole metagenome shotgun sequencing, where the whole genetic content of a sample is fragmented and sequenced, or targeted metagenomics, in which PCR amplicons generated from phylogenetic marker genes (such as 16S rRNA or ITS) are sequenced (Siegwald et al., 2017). In contrast, bioinformatics is a closely related field that produces and utilizes software for filtering, processing, and interpreting the sequence data generated from metagenomic samples. These processes rely heavily on mathematical and statistical analyses of sequencing reads and their associated quality data to produce meaningful results from highly complex metagenomic samples. The development and subsequent adoption of Next Generation Sequencing technology (NGS) has prompted a wealth of metagenomic studies in the last fifteen years (Roh et al., 2010; Siegwald et al., 2017). This high-throughput sequencing technology stimulated the advancement of many different bioinformatic approaches to profiling human, animal, and environmental microbiomes (D'Argenio, 2018; Sczyrba et al., 2017; Siegwald et al., 2017; Ye et al., 2019). However, the workflow through which these sequencing reads are generated, processed, and evaluated can impact the results obtained and/or their ability to be compared to other studies (Bhute et al., 2017; D'Argenio, 2018; Siegwald et al., 2017). Due to these potential issues it is important to weigh the pros and cons of each program used when choosing an appropriate bioinformatic pipeline for the research objective (Allali et al., 2017; Siegwald et al., 2017).

Targeted 16S rRNA sequencing (or metabarcoding) is a standard practice in taxonomic assignment of bacterial and archaeal metagenomic samples. The 16S rRNA gene is used for this purpose because it is highly conserved throughout prokaryotic life (Coenye & Vandamme, 2003). However, despite its utility this method has some shortcomings. For example, horizontal gene transfer has been demonstrated within the 16S rRNA gene at the intragenus and intraspecies levels, which can result in taxonomic misclassification (Kitahara & Miyazaki, 2013; Tian et al., 2015). NGS technologies are only able to sequence short reads (300 - 700 bp), requiring the selection of specific hypervariable regions of the 16S rRNA gene for amplification and subsequent sequencing, as full length 16S rDNA is approximately 1.5-kb in size (Johnson et al., 2019; Yarza et al., 2014). The selection of hypervariable regions to amplify must be considered carefully, as it has been demonstrated that there is no “universal” hypervariable region that is able to accurately assign taxonomy to the same level for all prokaryotic lineages (Figure 5) (Johnson et al., 2019; Yarza et al., 2014). As it can be seen in Figure 5, the greater the length of the 16S rRNA gene that amplicons are generated from, the more accurate and diverse taxonomic assignments become (Yarza et al., 2014). Similar conclusions were drawn in a latter investigation comparing full-length 16S rRNA sequencing to hypervariable region sequencing (Johnson et al., 2019). This whole-gene sequencing approach has only become practical in the last decade due to third generation sequencing technology, which allows for much longer sequence reads (>10 kb) than previous generations (Rhoads & Au, 2015).

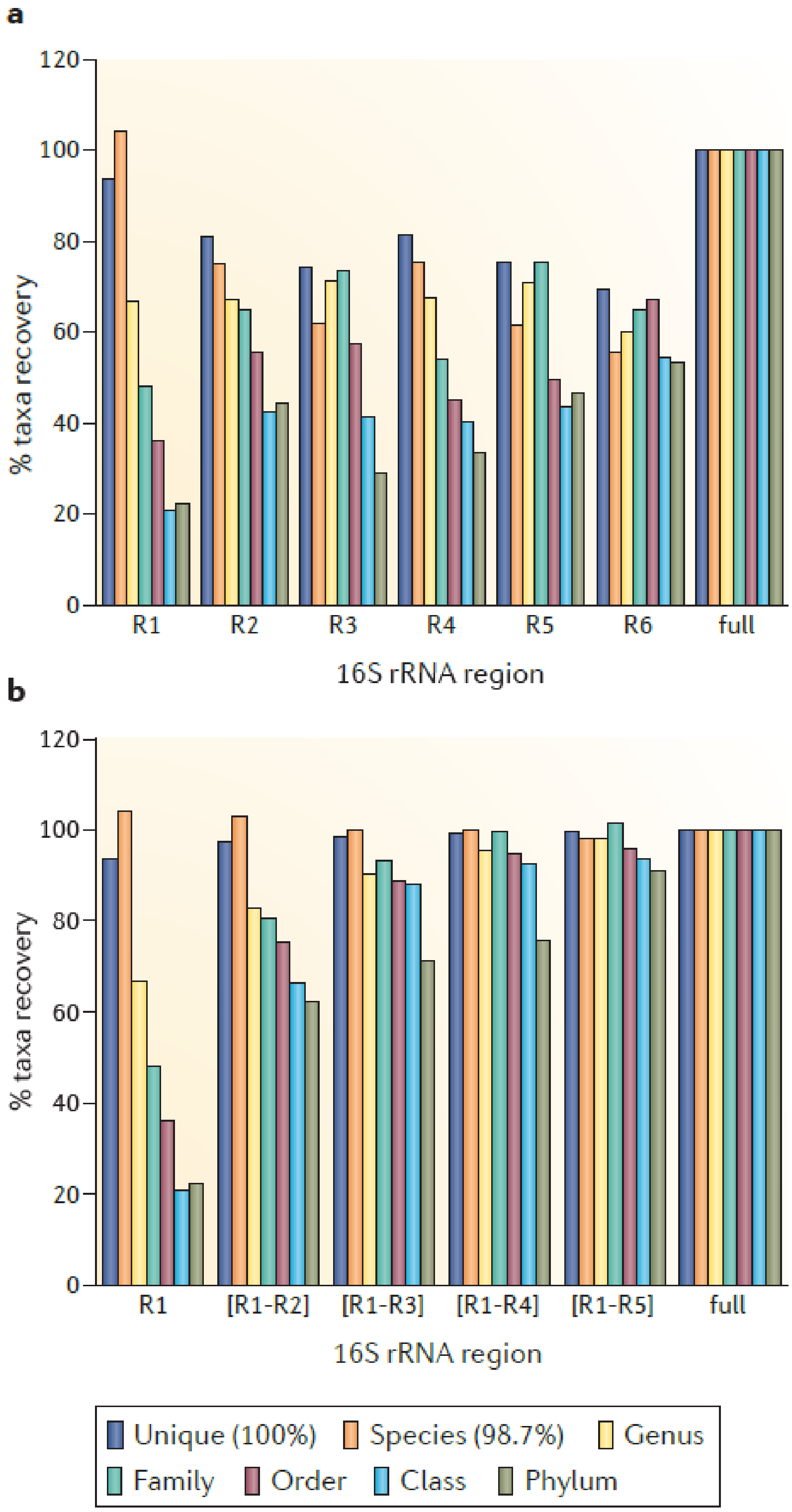


Figure 5: a. Six amplicon fragments (R1 - R6) of approximately 250 bases were generated conforming to the 16S rRNA hypervariable (V) regions, with the complete 16S sequence included for contrast. **b.** Constructed off the previous "R" fragments, four larger amplicons were generated, all starting at the 5′ end of the V1 region with increasing size:

R1 containing the 250 bases of the 5′ end, R1–R2 containing the 5′ 500 bases, R1–R3 containing the 5′ 750 bases, R1–R4 containing the 5′ 1050 bases, R1–R5 containing the 5′ 1300 bases, and 'full' containing the full *E. coli* 16S rRNA gene (1,542 nucleotides). Taxa recovery rate demonstrates a large underestimation of taxa diversity when incomplete sequences are utilized. As extended fragments were generated diversity estimation improved, however near full-length 16S rRNA sequences are necessary for precise diversity estimations and precise classification of high taxa. Figures were generated with data taken from the Living Tree Project release 108 (Adopted with permission from Yarza et al. (2014)).

Despite advancements in sequencing technology enabling a phylogenetic marker whole-gene approach, there are still obstacles with its adoption including cost and higher read error rate of third generation sequencing (~10% error) (Johnson et al., 2019; Rhoads & Au, 2015). Recent improvements in bioinformatic denoising algorithms have allowed (to a degree) for the removal of random sequencing errors, while maintaining intragenomic SNPs (single-nucleotide polymorphisms) (Johnson et al., 2019).

Unlike targeted approaches, metagenomic shotgun sequencing relies on extracting DNA from all cells in a sample, randomly fragmenting these whole genomes, then sequencing the large number of fragments using NGS (Urry, 2018). This method eliminates the amplification biases associated with targeted sequencing methods and can provide an improved taxa detection resolution. However, assembling the resulting sequence fragments so that they can be bioinformatically analysed is computationally challenging (Sczyrba et al., 2017; Ye et al., 2019). Other drawbacks of metagenomic shotgun sequencing are increased cost compared to metabarcoding, higher genomic DNA quality requirements, and increased possibility of false positive and false negative identifications (Chouvarine et al., 2016; Hwang et al., 2019). To address these computational and bioinformatic challenges, numerous software algorithms have been developed (Chouvarine et al., 2016; Sczyrba et al., 2017; Ye et al., 2019). Additionally, the abundance of differing bioinformatic tools and pipelines available for processing metagenomic data results in significant variations between the outputs of the studies (Roy et al., 2018; Sczyrba et al., 2017; Ye et al., 2019). Similarly to 16S rRNA sequencing, the various advantages and limitations of shotgun metagenomics must be weighed carefully when determining if this method is appropriate for the research objective.

1.7 Summary

While the presence of archaeal species in the human gut microbiome has been known for a long time, the diversity of archaea is open for further analysis. This is in part due to the difficulties in DNA extraction from archaeal lineages. While a number of extraction methods have been tested in various studies throughout the years, there still remains a problem in developing a standard method for metagenomic analyses. Because even slight variations in DNA extraction methods can have statistically significant effects on taxa representation, most metagenomic studies on the human gut microbiome (or any other microbiome for that matter) are incompatible for direct comparison. Any attempt to compare even similar metagenomic studies could likely result in misinterpretation of the findings. This will continue to pose a significant problem in aggregating metagenomic data in this field. However, one point has become clear: whichever method is selected as a gold standard for metagenomic sample preparation, it must precisely balance DNA quality with optimal cell lysis. Because a shift too far toward either aspect will have a significant negative effect on the quality of results obtained.

Another hurdle in identifying the diversity of archaea in the human gut microbiome is the “masking effect” highly abundant species have on rare species. Due to the huge difference in relative concentration of DNA for each species, many rare species likely fall below the resolving power of PCR based identification. Recent advances in sample preparation, namely DNA normalization, have shown significant promise in improving the representation of rare taxa in metagenomic samples, which could enable previously undetected gut-associated archaea to be identified. One such technique, DSN normalization, has demonstrated a level of rare species enrichment in synthetic metagenomes that is far superior to traditional HAP chromatography-based methods. The use of such a technique could increase the effective resolving power of PCR based metagenome analysis to a level that allows even the most rare and transient species of a metagenome to be identified.

Further compounding the previously mentioned problems, the use of different sequencing and bioinformatic pipelines can cause large differences in the results obtained. This necessitates the importance of choosing which method of sequencing is best for a particular study by weighing their individual pros and cons carefully. Then considering how the resulting sequencing reads can be bioinformatically processed in such a way as to generate results that can be reasonably compared to other metagenomic studies, while also reducing the amount of false taxonomic identifications.

Finally, a combination of the techniques and methods described in this review could allow for a deep high sensitivity analysis of the human gut microbiome, which could uncover a currently unknown amount of archaeal diversity. The impact of which could have implications for topics of research on the association between gut-associated archaea and human health. Similarly, these methods could be applied to a number of other topics related to metagenomic studies to improve the resolution of analysis.

1.8 Aims

A combination of the techniques and methods described in this review could allow for a deep high sensitivity analysis of the human gut microbiome, which could uncover a currently unknown amount of archaeal diversity. In addition, there has been no previous attempt to utilize DSN based normalization to enrich a human gut microbiome samples for low abundance bacteria. Therefore, the main aims of this study are:

1. To enrich for “rare” archaeal microorganisms from human faecal samples by utilizing DSN-based DNA normalization.
2. To determine whether the diversity and abundance of human gut-associated archaeal species is greater than what is currently known and overlooked due to their low abundance and the limitations of standard culture-independent methods.
3. To enrich for “rare” bacterial microorganisms from human faecal samples by utilizing DSN-based DNA normalization.
4. To demonstrate the potential utility of DSN-based DNA normalization for high resolution surveillance of the human gut microbiome.

The impact of this study could have implications for topics of research on the association between gut-associated archaea and human health. Similarly, experimental and bioinformatic pipeline established in this work could be applied to a number of other topics related to metagenomic studies to improve the resolution of analysis.

Chapter Two: Materials and Methods

2.1 Sample collection

All faecal samples were self-collected by three volunteers ("N", "SP", and "R") (ethics approval number: 13/CEN/144) using a human faecal sample collection kit provided by Plant & Food Research© (Appendix 1). Volunteer and sample designations were derived from the participants initials. This kit included an instruction pamphlet to assure that proper sample collection and handling was maintained. As per the instructions provided, volunteers collected faeces directly into the sample jar, which was immediately sealed and inserted into an air-tight plastic bag containing an anaerobic atmosphere generation sachet. These samples were then placed into an insulated container with an ice pack and delivered to the laboratory for DNA extraction within thirty minutes to minimize degradation.

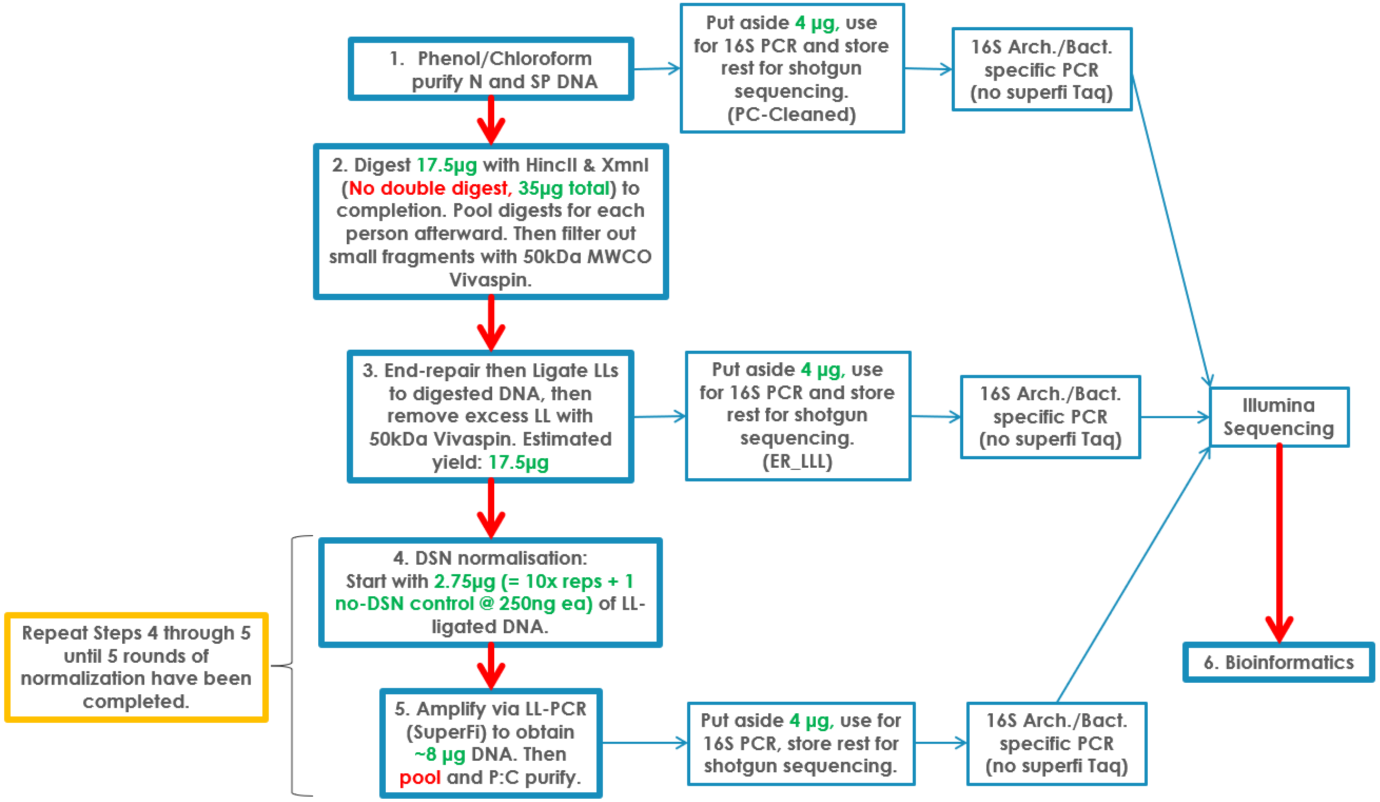
2.2 Preparation of Metagenomic DNA For Normalization

2.2.1 Metagenomic DNA extraction and purification

Each sample was separated into 24 replicates of 250mg faecal aliquots and labelled with an identification code indicating their source; "N", "SP", or "R", respectively. The DNA was extracted from these aliquots using a Qiagen DNeasy PowerLyzer® Power Soil® kit (Qiagen; Germany) with a modified protocol. Samples were bead-beaten for 30 seconds at either 6.5m/s or 4.0 m/s with a FastPrep-24. This DNA extraction kit was selected as it utilizes both chemical and physical means to disrupt cellular membranes. Additionally, because this kit is optimized for soil samples (containing many contaminants), it was decided that it would be suited to a similarly contaminant filled faecal sample. The resulting lysate was then pre-filtered with Zymo-Spin-IV filter columns (Zymo Research, USA) to remove large/undigested particles prior to further purification.

To determine the amount of mechanical shearing after extraction, resulting DNA was run on a 0.8% agarose gel using electrophoresis at 80V for 90 minutes and visualized by UV after staining with ethidium bromide. Further purification of samples was accomplished using a standard phenol:chloroform:isoamyl-alcohol protocol (25:24:1) (Evans, 1990) to remove any remaining contaminants.

Figure 6: Flowchart of Laboratory Work.


1. Phenol/Chloroform purify N and SP DNA
Put aside 4 µg, use for 16S PCR and store rest for shotgun sequencing. (PC-Cleaned)
16S Arch./Bact. specific PCR (no superfi Taq)
2. Digest 17.5µg with HincII & XmnI (No double digest, 35µg total) to completion. Pool digests for each person afterward. Then filter out small fragments with 50kDa MWCO Vivaspin.
3. End-repair then Ligate LLs to digested DNA, then remove excess LL with 50kDa Vivaspin. Estimated yield: 17.5µg
Put aside 4 µg, use for 16S PCR and store rest for shotgun sequencing. (ER_LLL)
16S Arch./Bact. specific PCR (no superfi Taq)
Illumina Sequencing
4. DSN normalisation: Start with 2.75µg (= 10x reps + 1 no-DSN control @ 250ng ea) of LL-ligated DNA.
Repeat Steps 4 through 5 until 5 rounds of normalization have been completed.
6. Bioinformatics
5. Amplify via LL-PCR (SuperFi) to obtain ~8 µg DNA. Then pool and P:C purify.
Put aside 4 µg, use for 16S PCR, store rest for shotgun sequencing.
16S Arch./Bact. specific PCR (no superfi Taq)

2.2.2 Restriction digests

To prepare the purified metagenomic DNA for ligation, restriction digests were carried out using HincII and XmnI (New England Biolabs© (NEB), Ipswich, Massachusetts) restriction endonucleases. Prior to selecting blunt cutting restriction enzymes, *in silico* restriction digests were run in Geneious R10.1 on *Methanobrevibacter smithii* reference genomes obtained from NCBI Genomes (www.ncbi.nlm.nih.gov/genome/) to generate genome fragmentation statistics.

Each reaction used 10U of either XmnI or HincII restriction endonuclease, 1µg of DNA, and 5µl of 10X Cutsmart® or 3.1 buffer, with a final volume of 50µl. Both digests were carried out in six replicates overnight at 37ºC. After digestion, all replicates for both restriction enzymes were pooled according to their respective sample identification ("N" or "SP").

2.2.3 Blunting of DNA ends

After fragmentation *via* restriction digests, sample DNA was processed using a NEB Quick Blunting™ Kit (New England Biolabs) following the manufacturers protocol. This step was included to ensure that no single strand overhangs were present due to shearing from mechanical lysis during the metagenomic DNA extraction protocol. Additionally, the blunting of any existing single strand overhangs increased the efficiency of subsequent blunt ligation of lone linker tags. Following blunting, samples were cleaned using a QIAquick® PCR Purification Kit (Qiagen).

2.2.4 Ligation of Lone-Linker Tags

To prepare the end-repaired DNA samples for the blunt ligation reaction, each sample was washed twice with 10 volumes of sterile water in a Vivaspin 2 – 50 kDa MWCO micro-concentrator (Sartorius Stedim Biotech GmbH, Goettingen, Germany). This also served the purpose of removing small oligonucleotides produced by mechanical shearing during the DNA extraction protocol.

Lone-linker tags were produced by annealing LL-RIA and LL-RIB (Table 1) oligonucleotides overnight at room temperature in 10mM Tris-HCl (pH 8) (Ko et al., 1990). Excess linker was removed by washing twice with 10 volumes of sterile water with a Vivaspin 2 – 50 kDa MWCO micro-concentrator.

The annealed lone-linker tags were then ligated to sample DNA using a T4 DNA Ligase kit (Thermo Scientific™, Massachusetts, United States). Ligation reactions were carried out using a 300:1 (LL:DNA) molar ratio, 5U of T4 ligase, 15% v/v PEG-4000, at 22ºC for 18 hours. The ligations were then washed twice with 10 volumes of sterile water in a Vivaspin 2 – 50 kDa MWCO micro-concentrator to remove excess/un-ligated LL tags.

2.3 Lone-Linker Amplification (LL-PCR)

Before beginning the normalization protocol, and after each round of normalization, each sample was amplified *via* LL-PCR using 2x Platinum™ Taq SuperFi™ Polymerase Master Mix (Invitrogen, Carlsbad, CA, USA). PCR reactions used the LL-RIA oligonucleotide as the primer at a final concentration of 1µM, with 100ng (for pre-normalization LL-PCR) or 0.1µl (for post-normalization LL-PCR) of template DNA. Additionally, a negative PCR control that had no template DNA was used. Thermocycling was carried out on a Bioer TC-XP-G Thermal Cycler (Hangzhou, China), with an initial denaturation stage of 98ºC for 30 seconds, followed by 30 cycles of denaturation (98ºC for 7 seconds), annealing (55ºC for 30 seconds), and extension (72ºC for 4 minutes); with a final extension step in the final cycle extended to 5 minutes. After normalization samples were cleaned using a QIAquick® PCR Purification Kit (Qiagen).

2.4 Trial Duplex-Specific Nuclease Digestion of ss/ds DNA

To determine the amount of DSN enzyme required for complete digestion of double-stranded DNA while limiting digestion of single-strand DNA, a trial digest was undertaken using different dilutions of DSN. To simulate DNA under hybridization conditions, 100ng of ssDNA isolated from phage M13 and 500ng of dsDNA isolated from the fosmid pCC2FOS were used for each DSN digestion. The ss and ds DNA was mixed and digested with either 1/8U or 1/16U of DSN enzyme for 20 minutes at 65ºC in 4× hybridization buffer [200mM HEPES (pH 7.5), 2M NaCl, 0.8mM EDTA] made to a final concentration of 1×.

2.5 Duplex-Specific Nuclease (DSN) Normalization

DSN normalization was carried out on 11 replicates of each sample using a modified method described by Shagina et al. (2010) and Gagic et al. (2015). Prior to normalization, 2.75µg of metagenomic DNA from each sample was mixed with 11µl of 4× hybridization buffer [200mM HEPES (pH 7.5), 2M NaCl, 0.8mM EDTA] in nuclease-free water up to a final reaction volume of 44µl. This mixture was then aliquoted equally into 11 – 200µl PCR tubes and overlaid with 2µl of sterile mineral oil. The metagenomic DNA was then denatured at 98ºC for 3 minutes, then slowly renatured at 68ºC for 5 hours in a Bioer TC-XP-G Thermal Cycler (Hangzhou, China).

After the renaturation step, 2µl of pre-warmed 68ºC 5× DSN Master buffer (Evrogen, Moscow, Russia) was added to each reaction tube by pipetting under the mineral oil, without removing the tubes from the thermocycler; to ensure that the hybridization temperature was not disrupted. Subsequently 1/8U of DSN (Evrogen, Moscow, Russia) was added to 10 of the 11 tubes, then incubated at 65ºC for 20 minutes. Reactions were then stopped by inactivating the DSN with the addition of 10mM EDTA to a final concentration of 3mM.

The residual (normalized) DNA was then amplified using the previously mentioned LL-PCR method. After PCR amplification, the metagenomic DNA was purified using a QIAquick® PCR Purification Kit (Qiagen). After each round of normalization an aliquot of normalized DNA was purified and set aside for 16S rRNA PCR amplification.

2.6 16S rRNA PCR

2.6.1 Archaeal 16S rRNA PCR

Prior to sequencing, the presence of archaeal DNA was determined both before and after each round of normalization using PCR with archaea specific (V6-V8 region) 16S rRNA primers Ar915aF and Ar1386R (Table 1) (Kittelmann et al., 2013; Zhou et al., 2017). As positive PCR control, DNA isolated from *M. ruminantium* M1 was used in conjunction with a negative PCR control that had no template DNA added. Amplification reactions were carried out using Hot Start *Taq* 2x Master Mix (New England Biolabs) with 50ng of template DNA and 0.2µM final concentration of each primer. Thermocycling was done on a Bioer TC-XP-G Thermal Cycler (Hangzhou, China) with an initial denaturation step of 95ºC for 30 seconds, followed by 30 cycles of denaturation (95ºC for 15 seconds), annealing (62ºC for 10 seconds), and extension (68ºC for 20 seconds). The extension stage of the final cycle was extended to 30 seconds.

Amplicons were then visualized by electrophoresis using a 1.5% agarose gel made with 1x TAE buffer (Tris-HCl 40 mM, Acetic Acid 20 mM, EDTA 1 mM), run at 70V for 90 minutes at room temperature (Evans, 1990).

Table 1: Oligonucleotides Used.

Oligo Name	Sequence	Source	Target/Use
LL-RIA	5’-GAGATATTAGAATTCTACTC-3’	(Ko et al., 1990)	PCR Amplification Tag
LL-RIB	5’-TATAATCTTAAGATGAG-3’	(Ko et al., 1990)	PCR Amplification Tag
Ar915aF	5’-AGGAATTGGCGGGGGAGCAC-3’	(Kittelmann et al., 2013)	Archaea-Specific Primer
Ar1386R	5’-GCGGTGTGTGCAAGGAGC-3’	(Kittelmann et al., 2013)	Archaea-Specific Primer
V4F1	5’-AYTGGGYDTAAAGNG-3’	(Marsh et al., 2013)	Universal-Bacterial Primer
V5R1	5’-CCGTCAATTYYTTTRAGTTT-3’	(Marsh et al., 2013)	Universal-Bacterial Primer

2.6.2 Bacterial 16S rRNA PCR

The presence of bacterial DNA was also determined both before and after each round of normalization using PCR with universal primers for the V4-V5 region of the 16S rRNA gene V4F1 and V5R1 (Table 1) (Claesson et al., 2010; Marsh et al., 2013). The V4-V5 regions were chosen due to their higher taxonomic identification resolution when used on highly complex microbiota samples, compared to other variable region combinations as demonstrated experimentally by Claesson et al. (2010). As a positive PCR control, DNA isolated from *Lactobacillus rhamnosus* HN001 was used in conjunction with a negative PCR control that had no template DNA added. Amplification reactions were carried out using Hot Start *Taq* 2x Master Mix (New England Biolabs) with 50ng of template DNA and 0.2µM final concentration of each primer. Thermocycling was done on a Bioer TC-XP-G Thermal Cycler (Hangzhou, China) with an initial denaturation step of 95ºC for 30 seconds, followed by 30 cycles of denaturation (95ºC for 15 seconds), annealing (52ºC for 1 minute), and extension (68ºC for 30 seconds). The extension stage of the final cycle was extended to 45 seconds.

Amplicons were then visualized by electrophoresis using a 1.5% agarose gel made with 1x TAE buffer (Tris-HCl 40 mM, Acetic Acid 20 mM, EDTA 1 mM), run at 70V for 90 minutes at room temperature (Evans, 1990).

2.7 Metagenome Sequencing

Aliquots of samples taken after phenol-chloroform-isoamyl alcohol purification and additionally after lone-linker ligation were sequenced using whole metagenome shotgun sequencing prepared using a Nextera XT DNA Library Preparation Kit (Illumina, San Diego, California) on an Illumina MiSeq™ 2× 300bp platform (Massey Genome Service, Palmerston North, New Zealand). The prepared Nextera XT library was loaded at 55% of a standard paired end (PE) run. To control for potential uneven representation of bases at each cycle a PhiX control V3 (Illumina, San Diego, California) library was loaded into the run at 10% volume as a reference.

16S rRNA bacterial amplicons generated after phenol chloroform isoamyl alcohol purified sample DNA, post LL ligated DNA, pre-normalization LL-PCR amplified DNA, and DNA from each completed round of normalization were prepared for sequencing using a MiSeq™ Reagent Kit v3 (Illumina). These libraries were then loaded for sequencing at 35% of a standard PE run on an Illumina MiSeq™ 2× 300bp platform (Massey Genome Service). As previously mentioned, control for potential uneven representation of bases at each cycle a PhiX control V3 (Illumina, San Diego, California) library was loaded into the run at 10% volume.

2.7.1 Metagenome Sequences Processing & Quality Control

Samples were demultiplexed by the Massey Genome Service prior to quality control processes. Controlling for uneven representation of bases at each cycle of sequencing was accomplished by mapping reads against the PhiX genome using Bowtie2 (Langmead & Salzberg, 2012). Any reads that matched the PhiX reference were removed from the generated SAM file, then the fastq file was rebuilt using the SamToFastq.jar plugin of the Picard software suite ("Picard toolkit," 2019). Sequencing adapters were then removed from reads using the *"fastq-mcf"* script in the ea-utils software suite (Erik Aronesty, 2011; E. Aronesty, 2013). Quality control of the resulting metagenomic sequences was then completed using SolexaQA++ (Cox et al., 2010) to remove low quality reads, FastQ Screen (Wingett & Andrews, 2018) to identify potential sample contamination, and FastQC ("FastQC," 2019) as a replicate identifier of low quality reads.

2.8 Metagenomic Shotgun Sequence Bioinformatics

2.8.1 Metagenome Assembly

Quality controlled metagenomic shotgun sequencing reads were assembled using the MEGAHIT (D. Li et al., 2015) metagenomic assembly software. The number of steps between *k*-mer assembly iterations was increased from the default, as was the maximum *k*-mer size, to improve assembly quality.

Command used:

- megahit -1 'Forward-Reads.fastq' -2 Reverse-Reads.fastq' --k-list 21,29,39,59,79,99,119,141,151,161,171,181,191,201,211,221,231,241,251 --no-mercy -t 4 -o '[Assembly-Output-Folder-Location]’

2.8.2 Taxonomic Assignment of Metagenome Assemblies

Metagenomic assemblies were initially classified taxonomically with the Kraken2 software package using the MiniKraken2_v2 database (Wood et al., 2019).

Command used:

- './kraken2' --db MiniDB 'MegaHit-Assembly.fasta' --classified-out Sample-Classified-Kraken2 --output Sample-Kraken2-Output

2.8.3 Bayesian Re-estimation of Taxon Abundance

The output files generated by Kraken2 were then processed to estimate taxon abundance at the genus level using Bracken (Lu et al., 2017).

Command used:

- ‘est_abundance.py' -i 'Sample-K2-report' -k '/MiniDB/database200mers.kmer_distrib' -l G -t 1 -o 'Sample-Bracken-GenusLevel'

Note: For the “-l” option, results were also generated for species and family taxonomic levels by using the “F” and “S” commands, respectively.

2.9 Metagenomic 16S rRNA Amplicon Bioinformatic Analysis Using Qiime2 Pipeline

2.9.1 Importing Sequence Data

To prepare 16S rRNA sequencing reads for analysis with the Qiime2 software suite (Bolyen et al., 2019) the sequences were imported using the following command.

-
  ```
  qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' --input-path 'Manifest-16S-Bact.tsv' --output-path 16S-Bact_Demux_Paired-End/paired-end-demux.qza' --input-format PairedEndFastqManifestPhred33V2
  ```

Importing using this format required the creation of a “manifest” file. The file was made in excel by creating three columns labelled “sample-id”, “forward-absolute-filepath”, and “reverse-absolute-filepath” (with each respective column containing the indicated information). The file was saved as TSV format.

2.9.2 Joining Forward and Reverse Sequence Reads

The imported 16S rRNA forward and reverse reads were then joined using the VSEARCH plugin (Rognes et al., 2016) in Qiime2 with the following command.

Command:

- qiime vsearch join-pairs --i-demultuplexed-seqs ‘paired-end-demux.qza’ --o-joined-sequences ‘demux-joined.qza’

2.9.3 Denoising of Reads

The joined read pairs were then denoised using the Qiime2 plugin DADA2 (Callahan et al., 2016) and the following command.

Command:

- qiime dada2 denoise-single --i-demultiplexed-seqs ‘demux-joined.qza’ --p-trim-left 0 --p-trunc-len 0 --o-representative-sequences ‘rep-seqs-dada2.qza’ --o-table ‘table-dada2.qza’ --o-denoising-stats ‘stats-dada2.qza’

No truncation was required in either the forward or reverse reads as the quality score was high for all bases, and there was sufficient sequence overlap.

2.9.4 Feature Classification

Taxonomy was assigned to denoised reads using the Silva132-97% reference database (Quast et al., 2012; Yilmaz et al., 2013) in an open reference method. All reads that did not map to the Silva reference database were submitted to BLAST (Camacho et al., 2009) for identification.

Command used:

-
  ```
  qiime feature-classifier classify-consensus-blast --i-query 'rep-seqs-dada2.qza' --i-reference-reads 'Silva-132-97-16S.qza' --i-reference-taxonomy ‘Silva-97-Taxonomy.qza' --output-dir '/Silva97-Feature-Classification'
  ```

2.9.5 Taxonomic Composition Analysis

Taxonomic classifications were then compared to their respective number of reads in each sample. The resulting data was then used to generate a taxonomic composition bar plot for each sample using the following command.

Command:

-
  ```
  qiime taxa barplot --i-table ‘table-dada2.qza’ --i-taxonomy ‘classification.qza’ --m-metadata-file ‘sample-metadata.tsv’ --o-visualization ‘Taxa-Barplot.qzv’
  ```

To aid visualization, low abundance taxa (≤ 1000 reads) that were only present in pre-normalization samples, and taxa that were not identified to at least the phylum level were condensed into a “minor taxa” group. This was only for visualization and will not impact any diversity statistics.

2.9.6 Alpha Diversity Statistics

Generation of α diversity statistics for each respective sample was accomplished by using iterations of the following command.

Command:

- qiime diversity alpha --i-table table-dada2.qza --p-metric: faith_pd --o-alpha-diversity faithPD.qza

This command calculated Faith's phylogenetic diversity metric (D. P. Faith, 1992).

Command

- qiime diversity alpha --i-table table-dada2.qza --p-metric: shannon --o-alpha-diversity Shannon.qza

This command calculated the Shannon index (Shannon, 1963).

Command

- qiime diversity alpha-rarefaction --i-table table-dada2.qza --p-max-depth 40000 --o-visualization 40k-alpha-rarefaction.qzv

This command calculated the alpha-rarefaction curves of the sequence data provided.

Chapter Three: Results

3.1 Isolation of Metagenomic DNA From Faecal Samples

Metagenomic DNA was extracted from participants' faecal samples using a commercial kit (see section 2.2.1) with the addition of a physical treatment (bead-beating) to lyse microbial cell walls prior to the DNA extraction. These samples were initially bead-beaten for 30 seconds at 6.5m/s, which (Figure 7) caused the extensive amount of DNA fragmentation as demonstrated by the majority of DNA fragments less than ~10kb in size. Therefore, the intensity of the bead beating was reduced to 4.0 m/s to limit the amount of mechanical shearing of DNA.

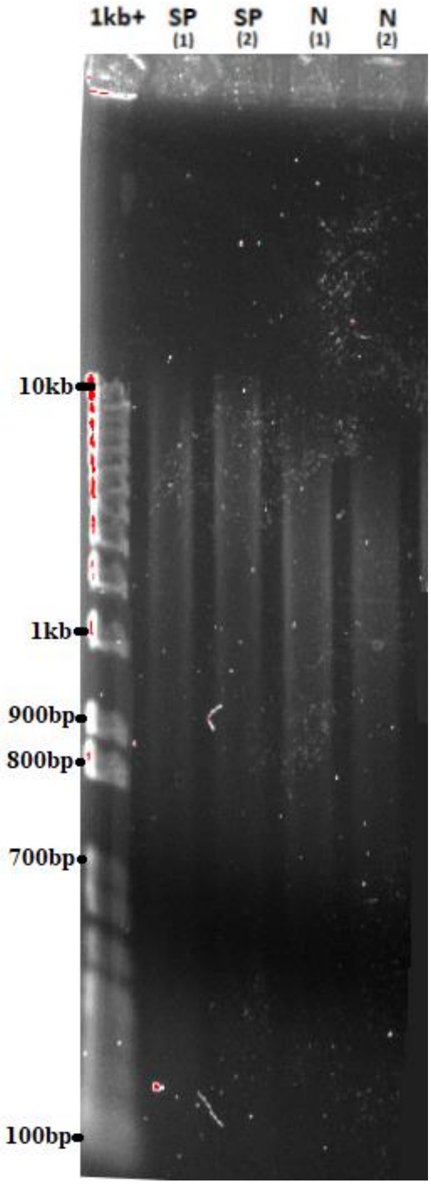


Figure 7: Metagenomic DNA extraction from the human faecal samples. Metagenomic DNA extracted in duplicate from participants “N” and “SP”. The resulting smears represent DNA fragments ranging from ≈10kb to ~100bp for participant “SP”, and ~6kb to 100bp for participant “N”.

After reduction of bead-beating intensity, the DNA extraction procedure indicated the majority of cells were lysed and metagenomic DNA quality (concentration and lack of significant shearing) was sufficient for the next step (Figure 8). The average metagenomic DNA concentration and purity readings are listed in Table 2. An absorbance reading $A_{260}/_{280}$ ratio of ~1.80 indicates samples have no protein contamination. Furthermore, the $A_{260}/_{230}$ readings below 2.0 indicate that there is some residual contamination present in the form of carbohydrates from the faeces, guanidine from the DNA extraction kit. These potential contaminants were then removed with a subsequent phenol:chloroform:isoamyl-alcohol DNA purification as mentioned in section 2.2.1.

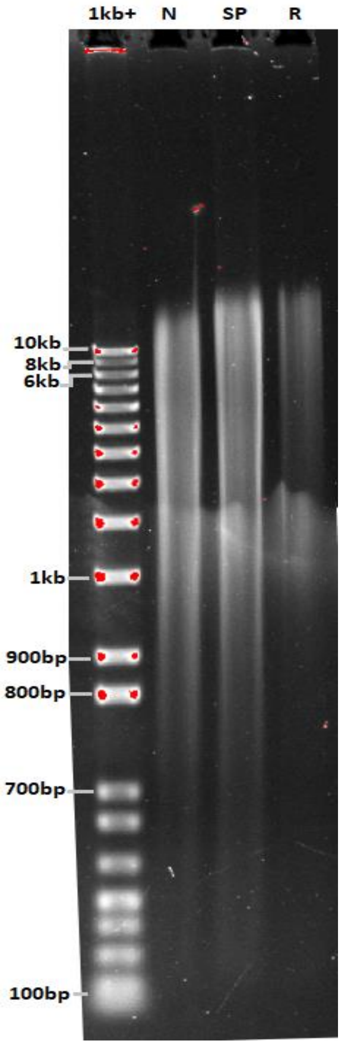


Figure 8: Metagenomic DNA from the human faecal samples using modified method. Metagenomic DNA extracted from participants “N”, “SP”, and “R”. The resulting smears represent DNA fragments ranging from >>10kb to ~100bp.

Table 2: Metagenomic DNA concentrations and quality

Sample	**Avg. Concentration (ng/µl)***	**Avg. $A_{260/280}$**	**Avg. $A_{260/230}$**
N	142	1.78	1.71
SP	221	1.79	1.52
R	94	1.83	1.62

*Readings were calculated based off the average of 24 replicates for each sample after DNA extraction.

3.2 Preliminary PCR Surveillance of Samples for Archaea and Bacteria

As budgetary constraints only allowed for two participant samples to be examined, it was necessary to select from the three available samples ("N", "SP, and "R"). To determine which two participants samples would be prepared for normalization 16S rRNA PCR using both universal bacterial primers (V4F1/V5R1, see Table 1) and archaea specific primers (Ar915aF/Ar1386R, see Table 1) was performed.

Amplicons generated using the archaeal specific oligonucleotides indicated the presence of Archaea in only the "SP" participant sample. The expected amplicon size of 492 bp for the archaeal 16S rRNA encoding gene was observed in the positive archaeal control (genomic DNA from the archaea *Methanobrevibacter ruminantium* M1) and the "SP" sample (Figure 9). Additional PCR fragments are present due to the thermocycler program not being optimized for the Taq polymerase master mix that was used. To eliminate this non-specific amplification, the extension step of the thermocycler program was reduced, and subsequently the temperature of the annealing step was increased.

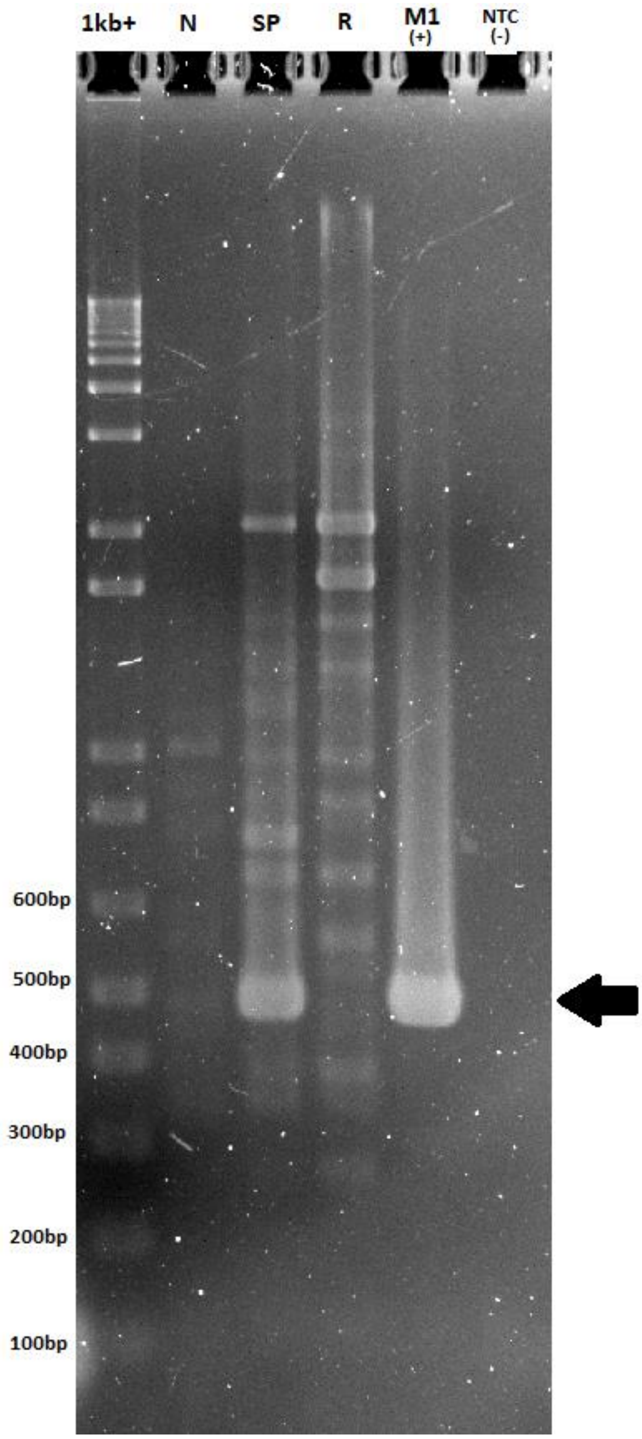


Figure 9: Archaeal 16S rRNA PCR of Samples "N", "SP", and "R". Initial PCR surveillance of all collected samples indicated that only the "SP" sample had detectable archaea using the chosen primer set. The expected amplicon size using archaeal oligonucleotides was 492bp (black arrow), as seen in lanes "SP" and the positive control (M1).

To confirm that bacterial DNA was also present in all samples, PCR amplification using the bacteria specific 16S rRNA oligonucleotides that amplify a 408 bp variable region between V4 and V5 regions was undertaken (Figure 10). The amplicons generated were of the expected size and matched the positive control *Lactobacillus rhamnosus* HN001, indicating that bacterial DNA was present. Considering that participant SP was positive for the presence of archaea and the N sample was negative, yet had higher DNA concentration than the R sample, these two samples were subjected to DSN normalization.

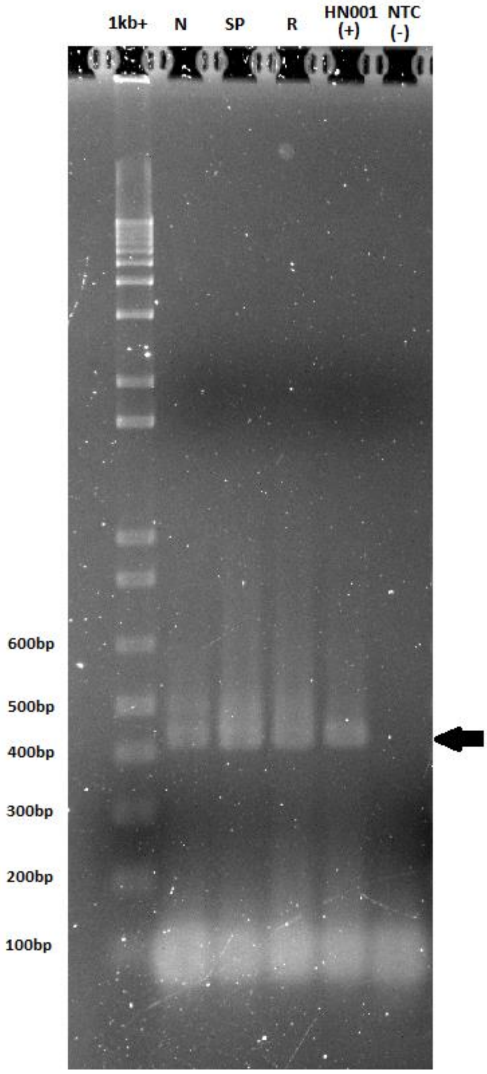
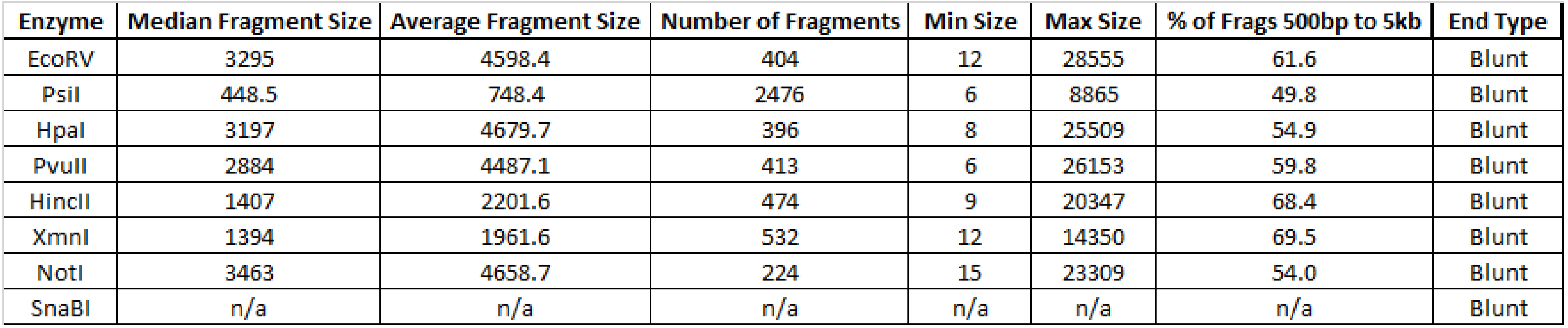


Figure 10: Bacterial 16S rRNA V4-V5 Amplicons Generated from the Human Faecal Samples ("N", "SP", and "R"). Participants samples were amplified using bacterial 16S rRNA primers with *Lactobacillus rhamnosus* HN001 as a positive control. Black arrow - the expected amplicon size of 408bp.

3.3 Preparation of Sample DNA for DSN Normalization

To prepare the previously selected participants samples (N and SP) for normalization, the metagenomic DNA was digested with restriction endonucleases that produce blunt cut ends. The digest was required to cut the metagenomic DNA into 500bp to 5 kb fragments because fragments of this size are optimal for PCR amplification. To verify the size range of fragments generated by the restriction enzymes, an *in silico* restriction digest was undertaken on a reference genome of *Methanobrevibacter smithii* downloaded from the NCBI database (accession number CP000678). The *in-silico* digests were performed with the blunt cutters PsiI and EcoRV. The average fragment size generated by PsiI was 748 bp, and EcoRV was 4598 bp (Table 3). As these two restriction enzymes met requirements, both the "N" and "SP" samples were digested to completion using PsiI and EcoRV.

Table 3. Restriction Enzymes Used for *in silico* and/or Experimental Digests.



Enzyme	Median Fragment Size	Average Fragment Size	Number of Fragments	Min Size	Max Size	% of Frags 500bp to 5kb	End Type
EcoRV	3295	4598.4	404	12	28555	61.6	Blunt
PsiI	448.5	748.4	2476	6	8865	49.8	Blunt
HpaI	3197	4679.7	396	8	25509	54.9	Blunt
PvuII	2884	4487.1	413	6	26153	59.8	Blunt
HincII	1407	2201.6	474	9	20347	68.4	Blunt
XmnI	1394	1961.6	532	12	14350	69.5	Blunt
NotI	3463	4658.7	224	15	23309	54.0	Blunt
SnaBI	n/a	n/a	n/a	n/a	n/a	n/a	Blunt

A standard RE digestion method (NEB) recommended 1 unit of restriction enzyme per µg of DNA, and an overnight digest, failed to fragment the SP sample compared to the N sample when visualized *via* agarose gel electrophoresis (Figure 11).

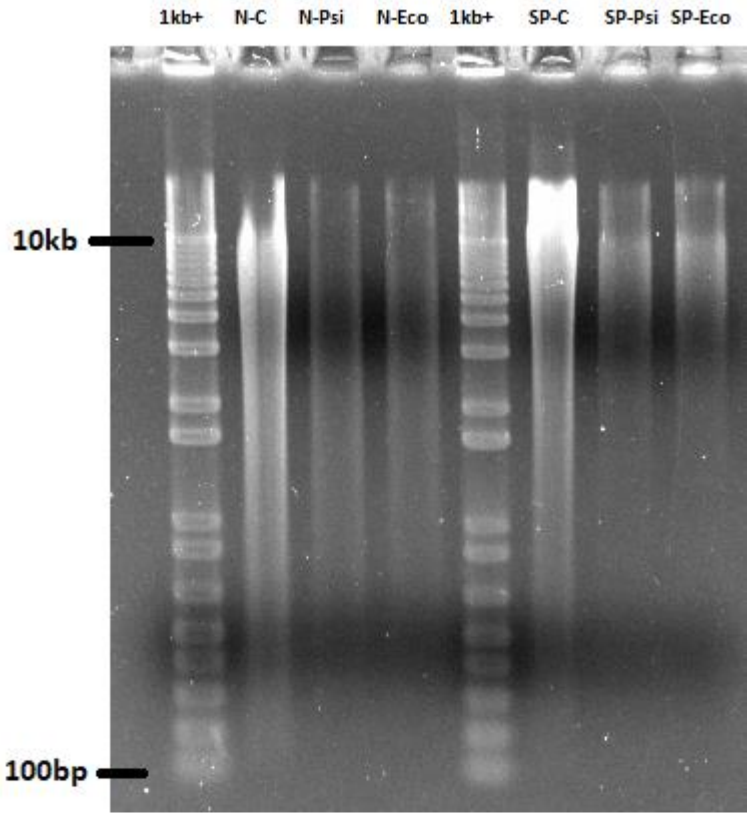


Figure 11: Restriction Digest of "N" and "SP" samples with EcoRV and PsiI restriction endonucleases. Sample conditions are denoted by the labels, with "-Psi" and "-Eco" indicating that these samples were digested by PsiI and EcoRV, respectively. Samples denoted with a "-C" are undigested DNA, used as a negative control.

The concentration of the REs was increased to 10 units to ensure that the digest was to completion. However, a similar result was obtained showing that both enzymes failed to digest the DNA from the SP sample completely (data not shown).

The outcome of a tenfold increase in concentration of RE used for each digest was a limited digestion of the SP sample compared to the N sample. This result ruled out the possibility of an insufficiency in the amount of endonuclease used. In response to the SP samples resistance to digestion, it was suspected that inhibitors may still be present in the sample at a concentration high enough to prevent digestion. Multiple different DNA purification kits and methods were used in attempts to remove any potential inhibitor molecules that remained in the sample; however, this had no significant effect (data not shown).

To overcome this obstacle, the SP sample was digested with a series of different REs both virtually and experimentally (Table 3). Experimental results showed that the SP sample could be digested using HincII and XmnI (Figure 12). As it was unexpected that the SP sample could only be completely digested using two of the eight restriction enzymes tested, REBASE (database listing specific RE sensitivity to DNA modification of cutting sites, such as methylation) (Roberts et al., 2015) was searched to determine if there were any DNA modifications (such as methylation) that these enzymes were sensitive toward and therefore could not digest to completion. Additionally, REBASE was utilized to determine if there were any similarities between the two working restriction enzymes, in terms of insensitivity to specific DNA modifications. However, neither of these searches showed any similarities in sensitivities or insensitivities between any of the working or non-working restriction enzymes. Furthermore, these subsequent *in vitro* RE digests were also repeated after multiple successive DNA purification attempts, using different commercial kits and protocols (data not shown). However, these further attempts at purification had no effect on RE digestion. It remains unknown what was inhibiting RE digestion of the SP sample with most of the tested endonucleases. Additionally, it has not been overlooked that this is unusual for restriction digests on highly purified DNA.

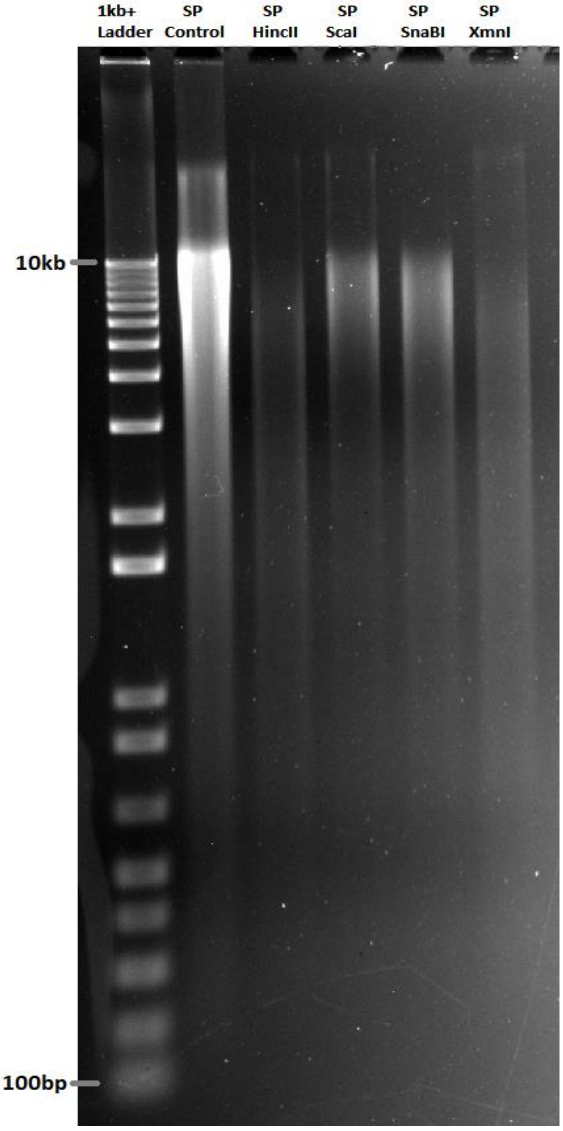


Figure 12: Restriction Endonuclease profiles of SP Sample. Each restriction digest is labelled with the sample and enzyme used, with the control indicating undigested DNA.

To prepare the samples for lone-linker PCR, lone-linker tags were ligated to the restriction digested DNA in a 100:1 (LL:DNA) molar ratio using the method outlined by Gagic et al. (2015). The efficiency of the lone-linker tag ligation was evaluated by PCR using a single lone-linker oligonucleotide (LL-RIA) for amplification (Figure 13 A & B).

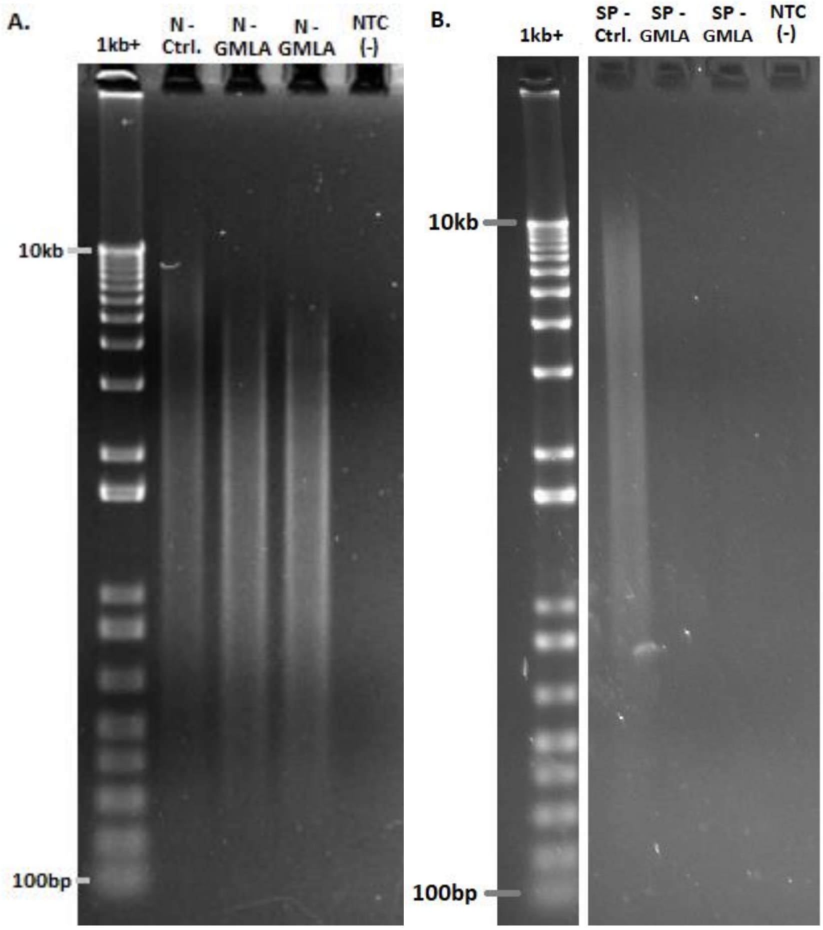


Figure 13 A & B: Evaluation of LL-Tag Ligation Using LL-PCR. Gel lane are labelled by their sample source (participant N or SP) and either “Ctrl.” for control (un-amplified DNA post-ligation), “GMLA” (Genome Metagenomic Linker-Amplified), or “NTC” (no template control). **(A.)** PCR amplicons generated from ligated sample N. **(B.)** PCR amplicons generated from ligated sample SP. No amplification is present in this sample, indicating that the ligation reaction failed.

As the SP sample showed no amplification after LL-PCR, it was hypothesized that DNA ends in that sample could be uneven due to mechanical shearing from the lysis protocol. This would prevent the blunt ligation reaction from working. If this were the case, it could also apply to the N sample to some degree. Although REs produce blunt DNA ends that would be valid for one end of the digested DNA fragment, the other end could still have various overhangs due to DNA shearing produced during isolation from the cells. To tidy up DNA ends, digested SP and N DNA was end-repaired, followed by ligation of lone-linkers in a 100:1 (LL:DNA) molar ratio. The efficiency of the lone-linker tag ligation after end repair was evaluated by LL-PCR (Figure 14).

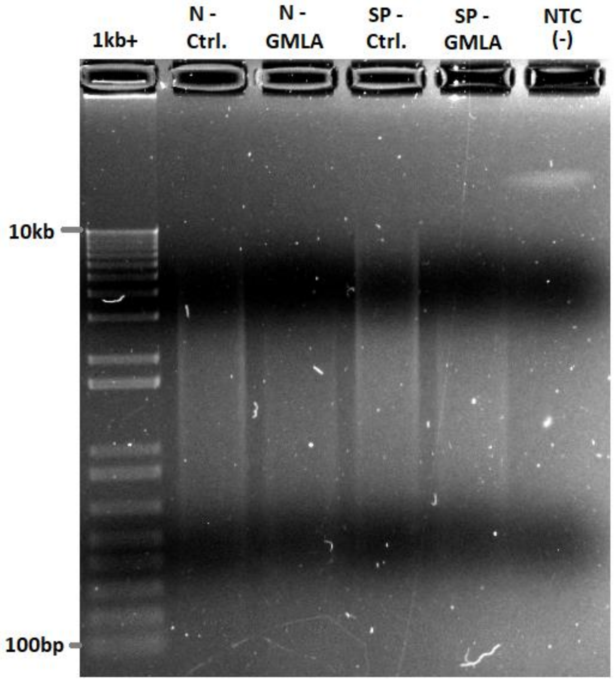


Figure 14: Evaluation of LL-Tag Ligation Using LL-PCR After End Repair. Gel lane are labelled by their sample source (N or SP) and either “Ctrl.” for control (DNA post-ligation), “GMLA” (Genome Metagenomic Linker-Amplified), or “NTC” (no template control).

The LL ligation to DNA was achieved, however, the efficiency of the blunt ligation was marginal based on the amount of amplification quantified. Comparison of the mass of metagenomic template DNA to the mass of amplicons after LL-PCR using fluorometric measurements showed that the level of template DNA amplification was approximately 11-fold. Therefore, optimization of the molar ratio of LL-tag to DNA ends (originally 100:1) was performed using a series of ligations with different LL-tag to DNA ratios. This series compared the relative level of LL-PCR amplification of all samples using 100:1, 300:1, and 500:1 LL to DNA ends molar ratios (Figure 15 and Table 4). Additionally, PEG-4000 was added to the reaction mixes to increase the chance of LL-tags interacting with DNA ends (Teraoka & Tsukada, 1987).

Both N and SP samples show an increase in amplicon DNA concentrations after increasing the LL-tag to DNA ends ratio to 300:1. This ratio of LL-tag to DNA ends was used to prepare DNA for normalization

.

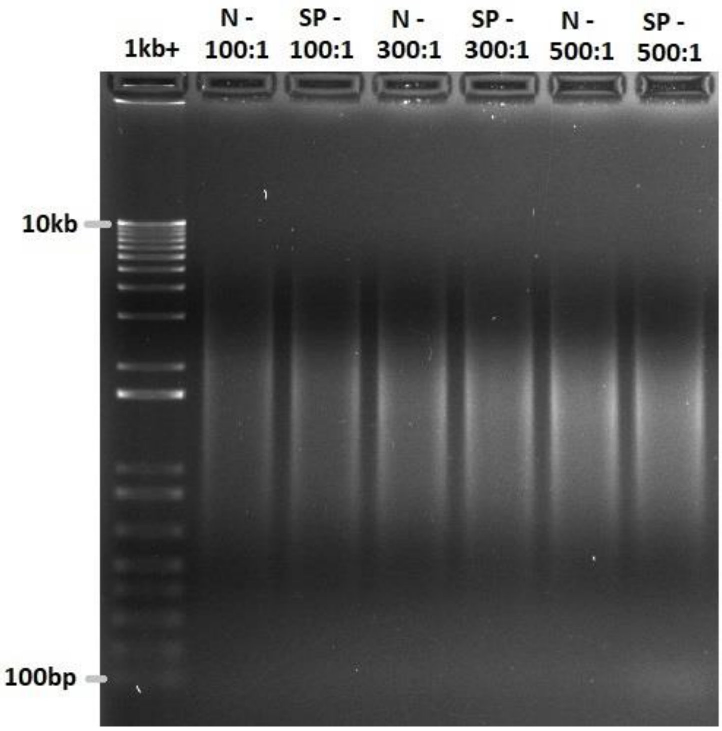


Figure 15: Molar Ratio Series of LL-Tag Ligations. Samples and their corresponding LL:DNA ligation reaction ratio were compared by concentrations of amplicon DNA after LL-PCR.

Table 4: Concentration of LL Amplicons at Different LL-Tag to DNA Molar Ratios.

Sample	Concentration Pre-PCR (ng/µ)	Concentration Post-PCR (ng/µ)	Fold Increase
N 100:1	1.428	15.6	10.92
N 300:1	1.326	15.8	11.92
N 500:1	1.53	15.9	10.39
SP 100:1	1.572	15.8	10.05
SP 300:1	1.519	16.3	10.73
SP 500:1	1.626	14.11	8.68

3.4 Duplex-Specific Nuclease (DSN) Normalization of Metagenomic DNA

Prior to the normalization protocol on the LL-ligated metagenomic DNA samples, the activity of the DSN enzyme on ssDNA and dsDNA was examined. Dilutions of DSN tested were 1/8U and 1/16U to determine which would digest dsDNA to completion while leaving ssDNA intact (Figure 16).

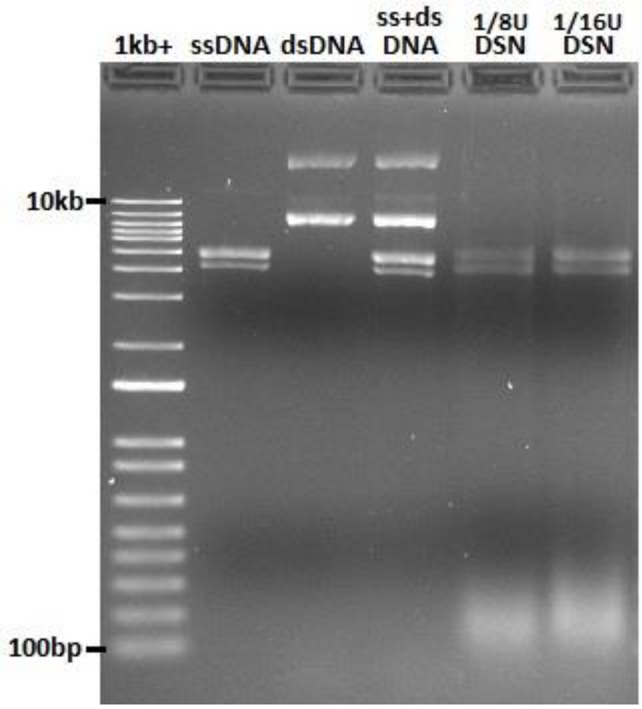


Figure 16: Trial DSN Digestion of ss/ds DNA Using 1/8U and 1/16U. Single strand DNA (ssDNA) isolated from phage M13 and double-strand fosmid pCC2Fos DNA (dsDNA) shown separately and combined. The lanes labelled 1/8U and 1/16U contain both ssDNA and dsDNA, and either 1/8U of DSN or 1/16U of DSN respectively. The 1/8U reaction shows near complete digestion of dsDNA while leaving the ssDNA intact, whereas the 1/16U reaction shows less complete dsDNA digestion.

The DSN trial digests, consistent with Gagic et al. (2015), show that 1/8U of DSN is required for complete digestion of dsDNA, while the ssDNA remains mostly intact. However, there was a small decrease in band intensity when using 1/8U compared to 1/16U, indicating that some ssDNA is being digested. For our normalization protocol 1/8U was used, however the test was necessary because despite being called “duplex specific nuclease”, the manufacturers literature indicates that DSN can have minor activity against ssDNA. To reduce the effect of this minor activity in our normalization results it was necessary to decrease the DSN concentration as much as possible, while still maintaining complete digestion of dsDNA (1/8U).

The DSN-based DNA normalization was performed in five cycles on SP and N samples. After completing two rounds of normalization and subsequent LL-PCR, both N and SP samples amplicons were quantified and visualized to verify amplification from normalized samples (Figure 17).

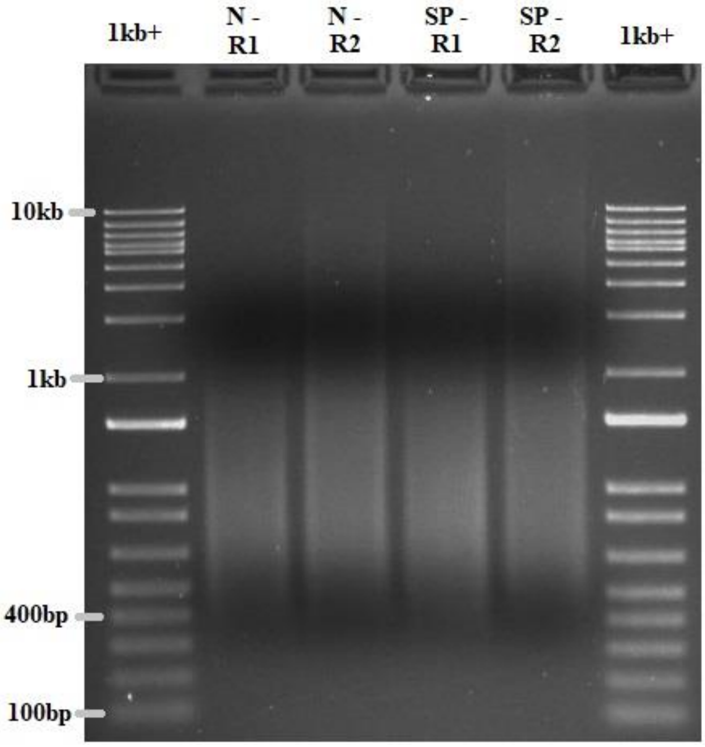


Figure 17: DNA Amplicons After Two Rounds of Normalization. Lanes are designated by their sample name (N or SP) and the round of normalization LL-PCR amplicons were taken from after completion of Round 1 (R1), and Round 2 (R2).

Visualization of samples after two rounds of normalization shows that the samples are amplifying within the expected size range. This information combined with concentration readings after each round of normalization indicate that ssDNA (lower abundance species) remained intact. Had ssDNA been digested the efficiency of amplification would be minimal. An additional three rounds of normalization were then completed, and amplicons were prepared for archaeal and bacterial 16S rRNA marker amplification and sequencing.

3.5 16S rRNA PCR Amplicons Generation

16S rRNA amplicons specific for archaea (Figure 18) and bacteria (Figure 19) were generated from each sample using aliquots taken after each step in DNA normalization and steps prior to it, including the starting metagenomic DNA, the digested DNA, and the LL-amplified DNA (see Section 2.5, Figure 6).

The archaea-specific 16S rRNA PCR shows the presence of archaeal DNA in the SP sample prior to normalization (PC, LLL, and R0). However, after the first round of normalization the band at 492bp is absent. This indicated that the archaeal DNA in the sample has been digested by the DSN enzyme. In contrast, the N samples show no presence of archaeal amplicons even after five rounds of normalization. The result is expected as archaeal-specific 16S rRNA amplicons were absent in the N sample in the preliminary screening for archaea (Figure 8). Due to time constraints, it was not possible to troubleshoot the lack of ssDNA in the SP sample after one round of normalization. The SP sample archaeal amplicons that were generated (PC/LLL/R0) were included for sequencing to determine the archaeal community profile, however, no further insights about the normalization of archaeal DNA could be derived.

Bacterial 16S rRNA amplicons were produced by all samples and the HN001 positive control. This shows that in contrast to archaeal DNA, bacterial DNA remained intact during the DSN normalization protocol. Furthermore, this indicates that DSN is selectively digesting archaeal ssDNA and/or dsDNA. The changes in community profile over each successive round of normalization can be used to determine the relative effectiveness of DSN normalization on human faecal samples despite archaeal DNA being degraded. Like the Archaeal 16S rRNA amplicons, these amplicons were sent for sequencing.

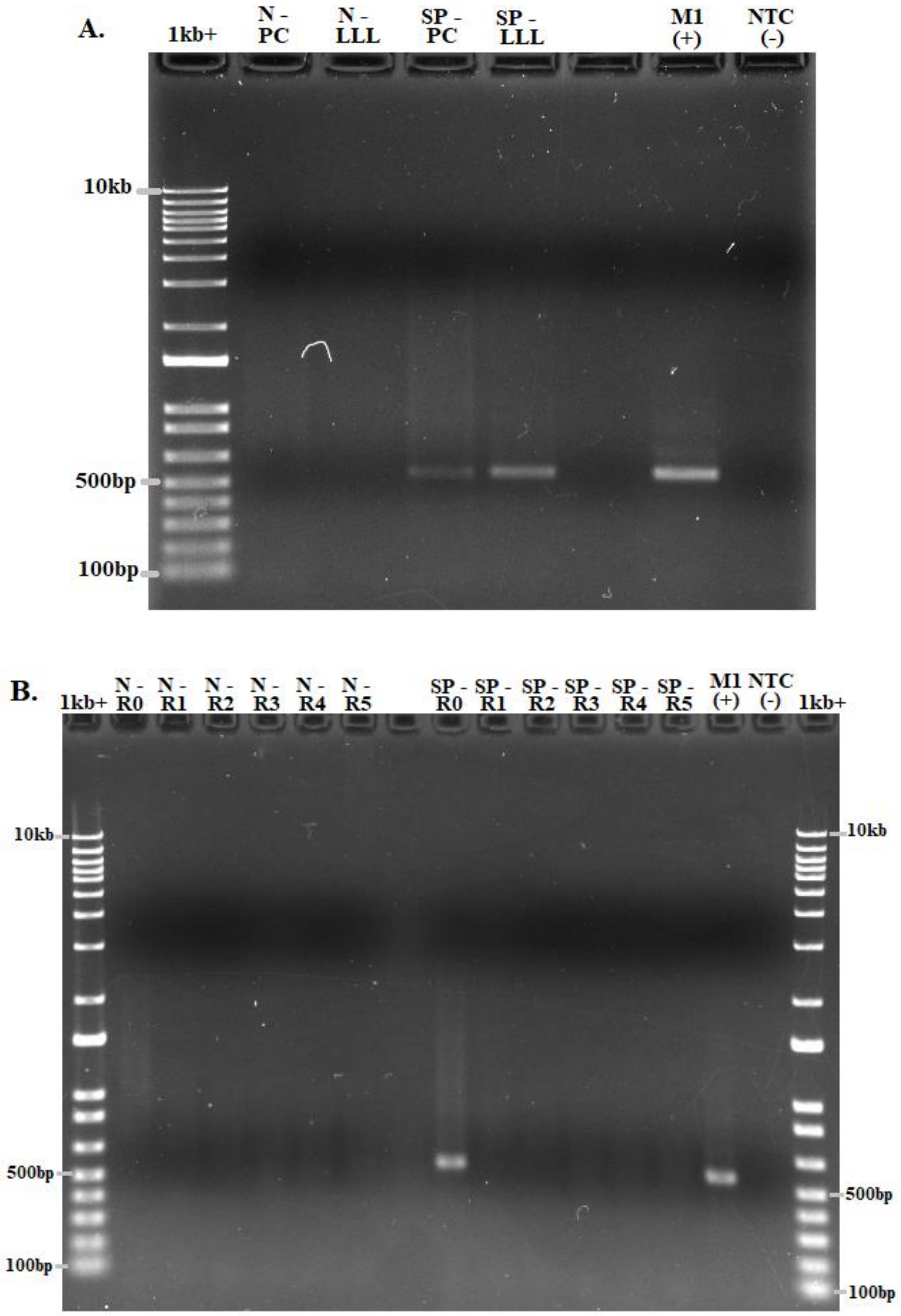


Figure 18 (A). Phenol:Chloroform Purified and LL-Ligated Sample Archaeal Specific 16S rRNA Amplicons. Amplicons generated from metagenomic DNA taken following phenol:chloroform (PC) purification and LL ligation (LL), with *Methanobrevibacter ruminantium* M1 (M1) as a positive control, and a no template (NTC) negative control. Both SP samples show a positive band of the expected size (492bp), matching the positive control, indicating the presence of archaeal DNA. **(B). Pre-Normalization LL-Amplified and**

Normalized Archaeal 16S rRNA PCR. Amplicons generated from metagenomic DNA following pre-normalization LL-PCR amplification (R0), and each successive normalization round (R1 through R5), with *Methanobrevibacter ruminantium* M1 (M1) as a positive control, and a no template (NTC) negative control.

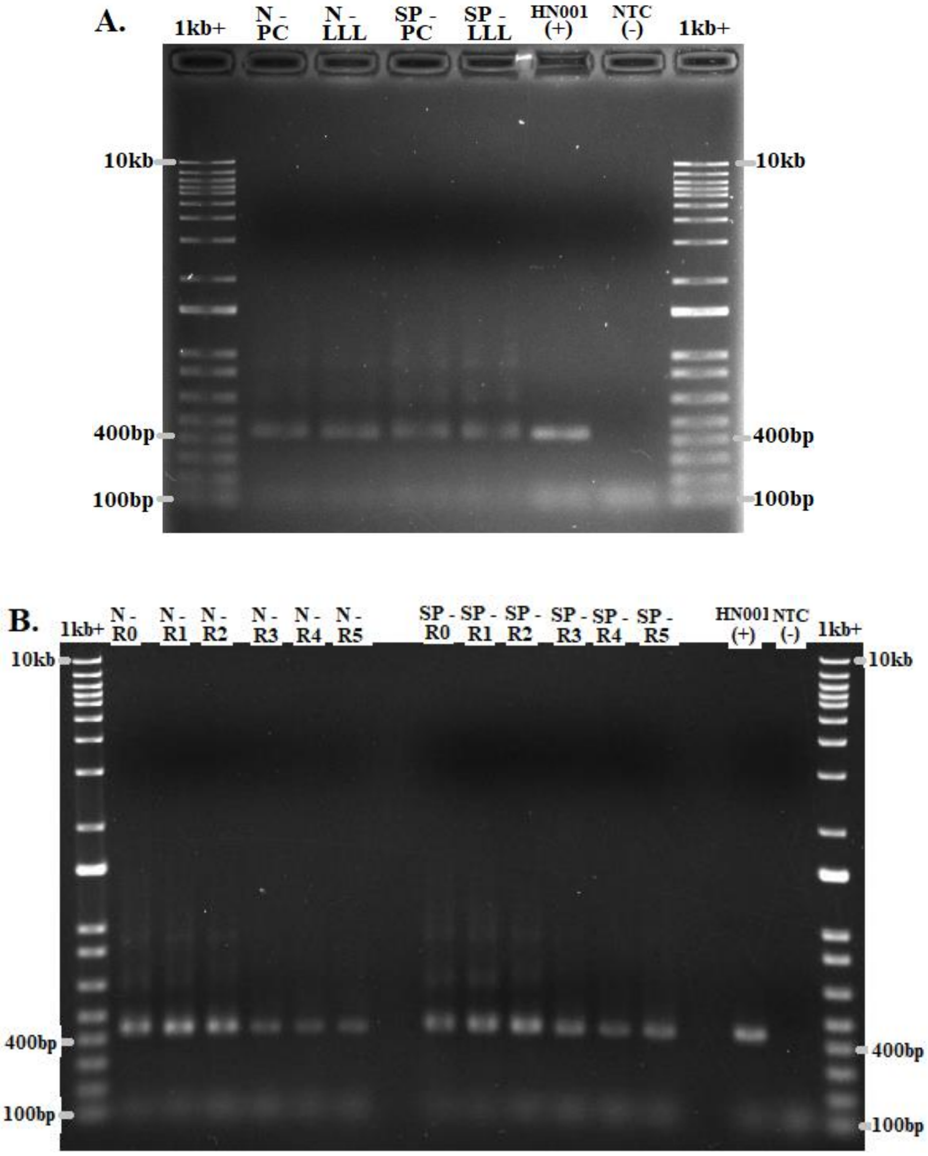


Figure 19 (A). Phenol:Chloroform Purified and LL-Ligated Sample Bacterial 16S rRNA Amplicons. Amplicons generated from metagenomic DNA taken following phenol:chloroform (PC) purification and LL ligation (LL), with *Lactobacillus rhamnosus* HN001 (HN001) as a positive control, and a no template (NTC) negative control. Both N and SP samples show bands of the expected size (408bp), matching the positive control and indicating bacterial DNA presence. **(B). Pre-Normalization LL-Amplified and Normalized Bacterial 16S rRNA PCR.** All samples have a band at 408bp as the positive control.

3.6 Bioinformatic Analysis of 16S rRNA Sequences

The quality of all sequence reads apart from one sample (“N - R2”) was above the cut-off point (quality score of 27) of our quality control software, based on quality score readings. The N-R2 sample consisted primarily of low-quality reads, despite having high purity (A260/280: 1.87, A260/230: 2.03) and average concentration (>20 ng/µl) readings prior to sequencing. Additionally, the number of acceptable quality reads retained from this sample after quality control amounted to less than 1% of the total number of reads generated from other samples.

Analysis of 16S rRNA sequence data was undertaken using the Qiime 2 software suite version 2019.4.0 (Bolyen et al., 2019). Prior to taxonomic classification and community profiling, samples were de-noised and filtered for chimeric reads. Different methods of denoising and chimera filtering were explored to determine their downstream effect on taxonomic classification. The archaeal 16S rRNA amplicon samples (SP-PC, SP-LLL, SP-R0) consisted only of *Methanobrevibacter smithii* reads after denoising, regardless of analysis software or method used (Table 5 and Table 6). The database used for taxonomic assignment also had no effect on the archaeal community profile, as the results were identical (Table 5). Furthermore, the only unique taxon detected in these reads (*M. smithii*) was identified with a 100% sequence similarity and an E-value of 0 (Table 6). In summary, there was no archaeal diversity based on 16S rRNA V4-V5 sequences.

Table 5: Silva and Greengenes Archaeal 16S rRNA Reads Denoising Statistics.

DB Used	Sample	Total Reads	Unique Reads After Derep.	Total Derep. Reads	Total Reads That Hit Ref. DB	Unique Reads That Hit Ref. DB
Silva 132	ArchSP-GMLAR0	13464	582	7476	3781	1
	ArchSP-LLL	13089	582	7215	3767	1
	ArchSP-PC	15063	583	9167	5272	1
Greengenes	ArchSP-GMLAR0	13464	582	7476	3781	1
	ArchSP-LLL	13089	582	7215	3767	1
	ArchSP-PC	15063	583	9167	5272	1

Table 6: BLAST Query of Unique Archaeal 16S rRNA Amplicon Sequenced from the Faecal Metagenome of Participant SP.

Description	Max Score	Total Score	Query Coverage	E-Value*	Percent Identity	Accession
Methanobrevibacter smithii partial 16S rRNA gene, strain C2 CSUR P5816	909	909	100%	0	100.00%	LR590664.1
Methanobrevibacter smithii strain KB11 chromosome, complete genome	909	1819	100%	0	100.00%	CP017803.1
Methanobrevibacter smithii ATCC 35061 strain PS 16S ribosomal RNA, complete sequence	909	909	100%	0	100.00%	NR_074235.1
Uncultured prokaryote clone 08062004-ZSS_YX_Z8_AR_2_49 16S ribosomal RNA gene, partial sequence	909	909	100%	0	100.00%	HQ154702.1
Methanobrevibacter smithii ATCC 35061, complete genome	909	1819	100%	0	100.00%	CP000678.1
Methanobrevibacter smithii ATCC 35061 strain PS 16S ribosomal RNA, complete sequence	909	909	100%	0	100.00%	NR_044786.1
Methanobrevibacter smithii partial 16S rRNA gene, isolate N63	907	907	99%	0	100.00%	LK054636.1
Methanobrevibacter smithii partial 16S rRNA gene, isolate N27	907	907	99%	0	100.00%	LK054635.1
Uncultured archaeon clone Muc-FT8 16S ribosomal RNA gene, partial sequence	907	907	99%	0	100.00%	JX522624.1
Uncultured methanogenic archaeon clone Oran-Met006 16S ribosomal RNA gene, partial sequence	907	907	99%	0	100.00%	JN192467.1

*** E-Value:** Also known as "Expect Value", indicates the number of results expected by chance when querying a large database. Decreases exponentially as score of match increases.

Initial denoising runs of the bacterial 16S rRNA amplicon sequences utilized the Deblur package (Amir et al., 2017) of Q2 with different reference databases (HMP 16S rRNA database, Silva 132 16S rRNA database, or Q2's default Greengenes database). The reference database used had a minor effect on taxonomic identifications, with Silva 132 producing the most diverse taxonomic assignment (942 unique taxa), and Greengenes (DeSantis et al., 2006) producing the least (457 unique taxa). This result is most likely due to using an open reference approach to taxonomic classification, which is discussed later in this section. However, Deblur was discontinued in favour of the DADA2 Q2 package (Callahan et al., 2016). DADA2 was chosen for the bioinformatic pipeline as it produced more reads after denoising compared to Deblur (Table 7). Additionally, the number of taxonomic classifications downstream was twofold higher with DADA2 than Deblur, with only 1% lower mean confidence of assignment (Table 7).

Table 7: Comparison of Taxonomic Assignments and Assignment Confidence Between DADA2 and Deblur Processed Metagenomic Sequence Data.

Statistic	DADA2	Deblur
Total Taxa Assigned	941	457
Mean Confidence of Assignments	0.833	0.843
Median Confidence of Assignments	0.90	0.80
Minimum Confidence of Assignments	0.556	0.6
Maximum Confidence of Assignments	1	1
Standard Deviation of Taxa Assignments	0.1487	0.1366

Prior to denoising and chimera filtering, the poor quality “N - R2” sample consisted of only 632 reads in total. After processing, by either DADA2 or Deblur, this was reduced to 0 reads. Therefore N-R2 was excluded from further analyses, including normalization changes caused between R1 and R2, and comparison of taxonomic diversity with SP-R2.

After the aforementioned pre-processing steps, the resulting reads were assigned taxonomy. Differing methods of taxonomic assignment were used for comparison. The method that, based on the total number of taxa identified in all samples, gave the highest confidence was an open reference approach (Q2 BLAST+ consensus taxonomy) with the Silva 132 reference database.

Microbial community profiles were generated for SP and N and arrayed in the order they were taken from our workflow (Figures 20 and 21). These profiles were generated for every taxonomic level, however, with only the V4-V5 region of the 16S rRNA phylogenetic marker was used for taxonomic assignment, the highest confidence is in assignments at the family level.

The community profile from participant N shows changes in composition with each successive step in the DNA normalization workflow. This is expected as any manipulation of metagenomic DNA (restriction digests, ligations, LL-PCR, etc.) can alter the relative abundance of each taxon’s sequences in a sample. However, at the start of our workflow eight taxa accounted for nearly 88% of the total reads.

These 8 taxa belong to the phyla *Firmicutes*, *Bacteroidetes*, and *Proteobacteria*, all of which are common human gut microbiota (Thursby & Juge, 2017).

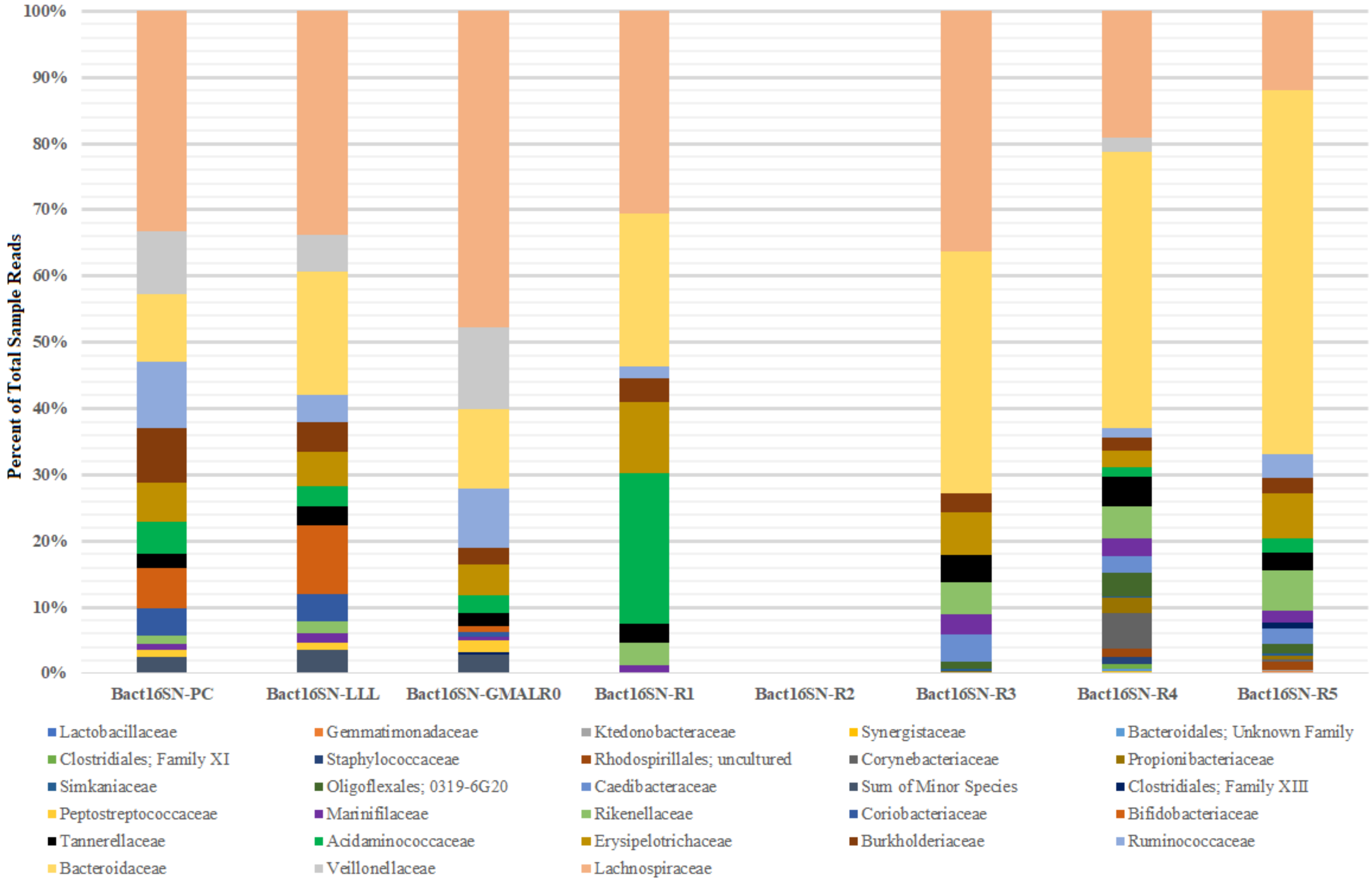

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percent of Total Sample Reads
Bact16SN-PC
Bact16SN-LLL
Bact16SN-GMALR0
Bact16SN-R1
Bact16SN-R2
Bact16SN-R3
Bact16SN-R4
Bact16SN-R5
Lactobacillaceae
Gemmatimonadaceae
Ktedonobacteraceae
Synergistaceae
Bacteroidales; Unknown Family
Clostridiales; Family XI
Staphylococcaceae
Rhodospirillales; uncultured
Corynebacteriaceae
Propionibacteriaceae
Simkaniaceae
Oligoflexales; 0319-6G20
Caedibacteraceae
Sum of Minor Species
Clostridiales; Family XIII
Peptostreptococcaceae
Marinifilaceae
Rikenellaceae
Coriobacteriaceae
Bifidobacteriaceae
Tannerellaceae
Acidaminococcaceae
Erysipelotrichaceae
Burkholderiaceae
Ruminococcaceae
Bacteroidaceae
Veillonellaceae
Lachnospiraceae

Figure 20: 16S rRNA Based Microbial Community (Family Level) Profile of Participant "N" Faecal Microbiome. Samples labels correspond to their type/sample participant (Bact16SN - Bacterial 16S rRNA from participant N) and the part of the workflow they were taken from: post DNA extraction phenol chloroform purification (-PC), lone-linker ligated (-LLL), pre-normalization LL-PCR (-GMLAR0), and post normalization rounds 1 through 5 (-R1, -R2, -R3, -R4, -R5) respectively. Taxonomic classifications are colour coded between he bar-plot and legend.

Further analysis of the community profile shows the initial dominance of *Lachnospiraceae* reduced by subsequent rounds of normalization from nearly 50% abundance, to approximately 12% abundance after five rounds of normalization. A similar decrease in abundance is observed with *Veillonellaceae,* from 9.5% of total reads to undetectable, and with *Ruminococcaceae* from 10.3% of total reads to 3.4% after the last round of normalization. In contrast, *Bacteroidaceae* showed an increasing abundance over subsequent normalization rounds, resulting in an overabundance after five rounds of normalization. This shift in abundance from 10.13% to 54.95% indicates that DSN normalization is not digesting all highly abundant dsDNA equally.

More than a dozen (14) rare taxa were identified after three or more rounds of normalization. The least abundant OTU from *Lactobacillaceae*, was detectible only after five rounds of normalization, and accounting for 0.2058% of the total sample reads. Another low abundance taxon from *Gemmatimonadaceae*, was only detectible after four rounds of normalization, at a relative abundance of 0.076%, and after five rounds of normalization increasing to 0.177%. These extremely low abundance taxa represent a small fraction of the total number of reads, that many cannot be visualized on the bar chart (Figure 19).

Alpha diversity analysis of the normalization workflow shows an overall increase in sample diversity after four rounds of normalization (Table 8). The α-diversity metric chosen for this analysis was Faith's

Phylogenetic Diversity metric (Faith PD) (D. P. Faith, 1992), which represents the minimum total length of all phylogenetic branches necessary to cover a set of taxa on a phylogenetic tree. Faith PD was chosen over others including Shannon Index (Ortiz-Burgos, 2016), because it is not influenced by large unevenness in taxa abundance, which can result in artificially low diversity scores.

Table 8: Alpha Diversity of Samples Derived from Participant N Throughout Normalization Workflow Using Faith's Phylogenetic Diversity (Faith PD) Metric.

Sample Name	Faith PD
Bact16SN-PC	7.01
Bact16SN-LLL	6.85
Bact16SN-GMLAR0	6.5
Bact16SN-R1	3.87
Bact16SN-R3	6.99
Bact16SN-R4	11.84
Bact16SN-R5	17.59

The participant N faecal microbiome shows an initial phylogenetic diversity score of 7.01, which decreased with each step in pre-normalization until round 1 (R1). After round 1 of normalization, the overall phylogenetic diversity of the sample increases greatly, peaking at 17.59 after five rounds of normalization. The phylogenetic diversity of the sample after five rounds of normalization is 2.5-fold greater than before normalization.

The community profile generated from participant SP (Figure 21), as with participant N, shows similar changes in taxa composition with each successive step in the project workflow. Starting metagenomic DNA samples from SP and N participants consisted of eight taxa represent approximately 96% of the total sample reads. These high abundance taxa belong to *Firmicutes* and *Proteobacteria*, however, in contrast to participant N, taxa belonging to *Bacteroidetes* were very low abundance at the beginning of the workflow (0.23% of total reads). Of the two most highly abundant taxa prior to normalization, *Lachnospiraceae* showed a high abundance of 42.5%, and after five rounds of normalization was reduced to 11.3%. Similarly, *Ruminococcaceae* had a high initial abundance of 47.8% of total reads prior to normalization, and after five rounds of normalization was reduced to 30.3% abundance in participant SP. In contrast, six low abundance taxa were only detectable after three or more rounds of normalization, and an additional five taxa were either detectable after R1 to R2 of normalization, or their relative abundance increased largely over five rounds of normalization. One example of these originally undetectable taxa being enriched after normalization is from order Oligoflexales, which after four rounds of normalization had increased in abundance to 2.66% of total sample reads.

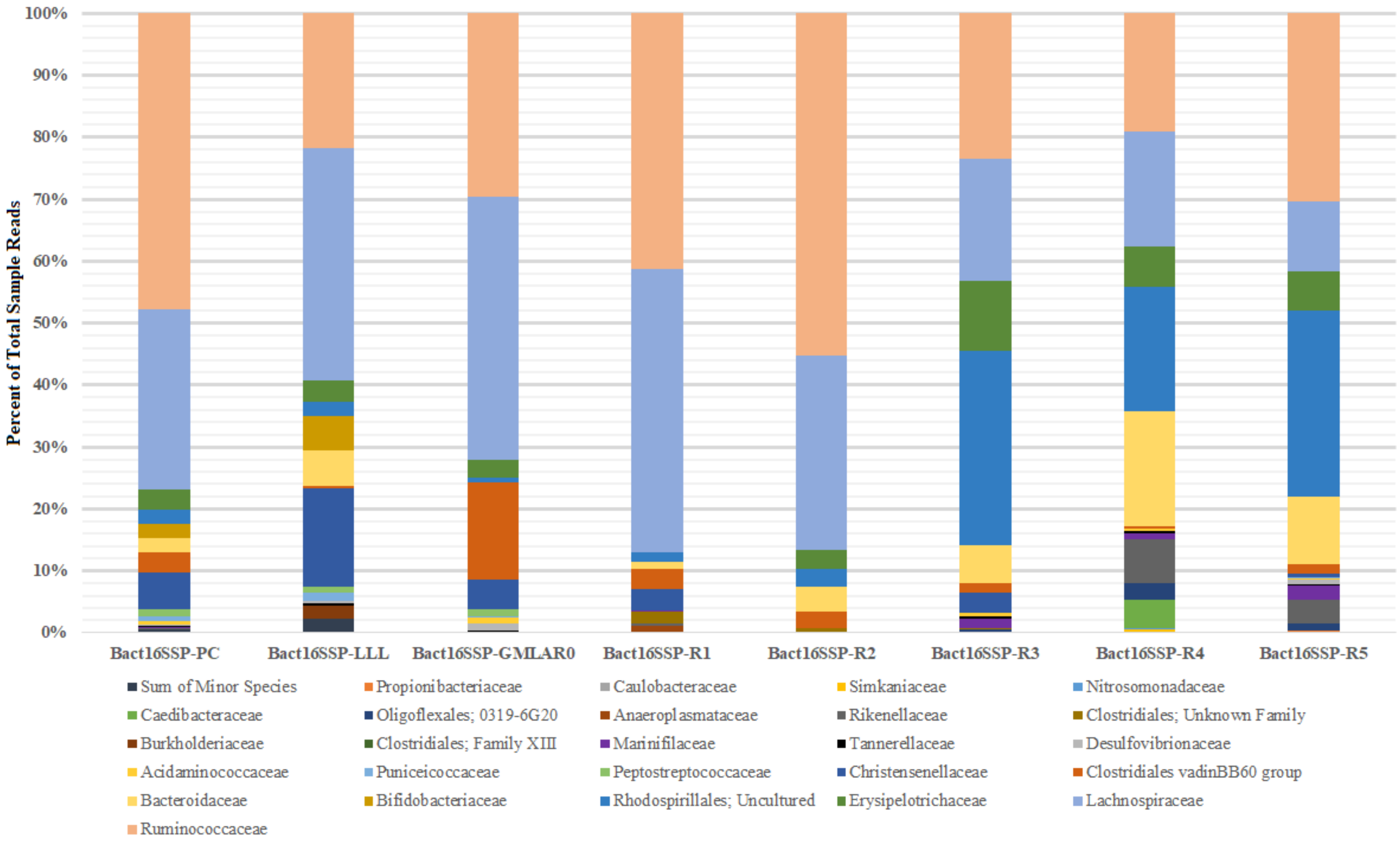

Percent of Total Sample Reads
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Bact16SSP-PC
Bact16SSP-LLL
Bact16SSP-GMLAR0
Bact16SSP-R1
Bact16SSP-R2
Bact16SSP-R3
Bact16SSP-R4
Bact16SSP-R5
Sum of Minor Species
Propionibacteriaceae
Caulobacteraceae
Simkaniaceae
Nitrosomonadaceae
Caedibacteraceae
Oligoflexales; 0319-6G20
Anaeroplasmataceae
Rikenellaceae
Clostridiales; Unknown Family
Burkholderiaceae
Clostridiales; Family XIII
Marinifilaceae
Tannerellaceae
Desulfovibrionaceae
Acidaminococcaceae
Puniceicoccaceae
Peptostreptococcaceae
Christensenellaceae
Clostridiales vadinBB60 group
Bacteroidaceae
Bifidobacteriaceae
Rhodospirillales; Uncultured
Erysipelotrichaceae
Lachnospiraceae
Ruminococcaceae

Figure 21: 16S rRNA Based Microbial Community (Family Level) Profile of Participant “SP” Faecal Microbiome. Samples labels correspond to their type/sample participant (Bact16SSP - Bacterial 16S rRNA from participant SP) and the part of the workflow they were taken from: post DNA extraction phenol chloroform purification (-PC), lone-linker ligated (-LLL), pre-normalization LL-PCR (-GMLAR0), and post normalization rounds 1 through 5 (-R1, -R2, -R3, -R4, -R5) respectively. Taxonomic classifications are colour coded between he bar-plot and legend.

Alpha diversity was determined for each sample from participant SP (Table 9). Despite some fluctuation in diversity (Faith PD score) in samples taken from the pre-normalization steps, sample diversity increased with each subsequent round of normalization. Following five rounds of normalization the diversity of participant SP's sample was nearly double that of the baseline sample (Bact16SSP-PC). Faith PD taken in conjunction with the changes in microbial community profile after normalization indicates that the normalization method successfully enriched for low abundance species.

Table 9: Alpha Diversity of Participant SP Samples Throughout Normalization Using Faith's Phylogenetic Diversity (Faith PD) Metric.

Sample Name	Faith PD
Bact16SSP-PC	7.61
Bact16SSP-LLL	8.16
Bact16SSP-GMLAR0	6.43
Bact16SSP-R1	5.29
Bact16SSP-R2	7.54
Bact16SSP-R3	8.41
Bact16SSP-R4	12.78
Bact16SSP-R5	15.03

Participant SP shows an initial phylogenetic diversity score of 7.61, which fluctuated with each step in the normalization workflow until round 1 (R1). Beginning after round 1 of normalization, the overall phylogenetic diversity of the sample increased, peaking at 15.03 after five rounds of normalization. The phylogenetic diversity of the sample after five rounds of normalization was 1.98-fold greater than before normalization.

3.7 Metagenomic Shotgun Sequencing Assembly

Whole metagenome shotgun sequencing was undertaken on the faecal metagenomic DNA extracted from participants N and SP to provide an alternative method of detecting low abundance archaeal species and community profiling the human gut microbiome. This method was chosen because it would provide a less biased estimation of taxa abundance in the human gut microbiome, due to less manipulation of DNA in preparation for sequencing, compared to metabarcoding methods such as 16S rRNA sequencing. Regardless of method used, each separate bioinformatic workflow introduces some bias to the results obtained.

Whole metagenome shotgun sequencing reads that were previously quality controlled (section 3.8) were assembled using the software MEGAHIT (D. Li et al., 2015). This particular software was chosen after trialling different assembly programs to determine which produced more contigs, longer assemblies, and which assemblies provided more diverse taxonomic classifications downstream. The results of each sample assembly run are listed below (Table 10).

Table 10: MEGAHIT Assembly Statistics.

Sample	Number of Contigs	Min. Size (bp)	Max. Size (bp)	Avg. Size (bp)	N50 (bp)
N - PC	74598	252	257826	1025	1318
N - LLL	73376	252	88297	920	1098
SP - PC	87419	252	208363	1028	1436
SP -LLL	84340	252	31606	878	1085

3.8 Taxonomic Classification & Community Profiling of Assemblies

The process of taxonomically classifying assemblies was completed using the software Kraken2 (Wood et al., 2019). The software compared sample *k-mers* to the miniKraken2 sequence database to determine the lowest common ancestor (LCA) of each query sequence. This resulted in a number of taxa with an estimation of relative abundance for each. As the relative abundances are important to determine the original microbial community profile of the N and SP samples, these abundance estimations were further processed with the Bracken (Lu et al., 2017) software package. Processing the Kraken2 outputs with Bracken (Bayesian Re-estimation of Abundance after Classification with KrakEN) allowed for more refined abundance estimations and generated the profiles (Figures 22 and 23).

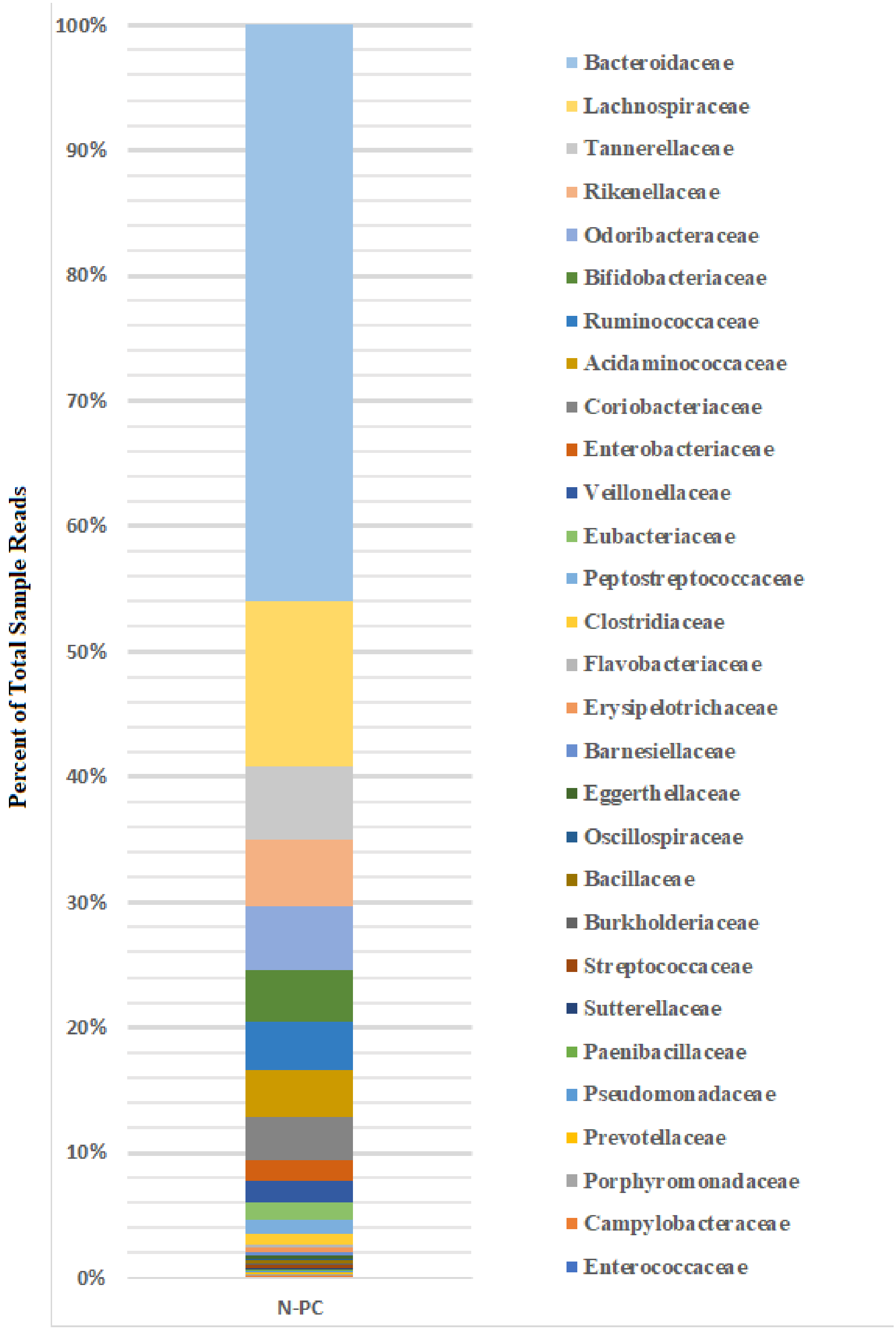


Figure 22: Metagenomic Shotgun Sequence Community Profile of Sample N - PC Using Bracken. Family names represented with <0.1% reads were omitted from the legend.

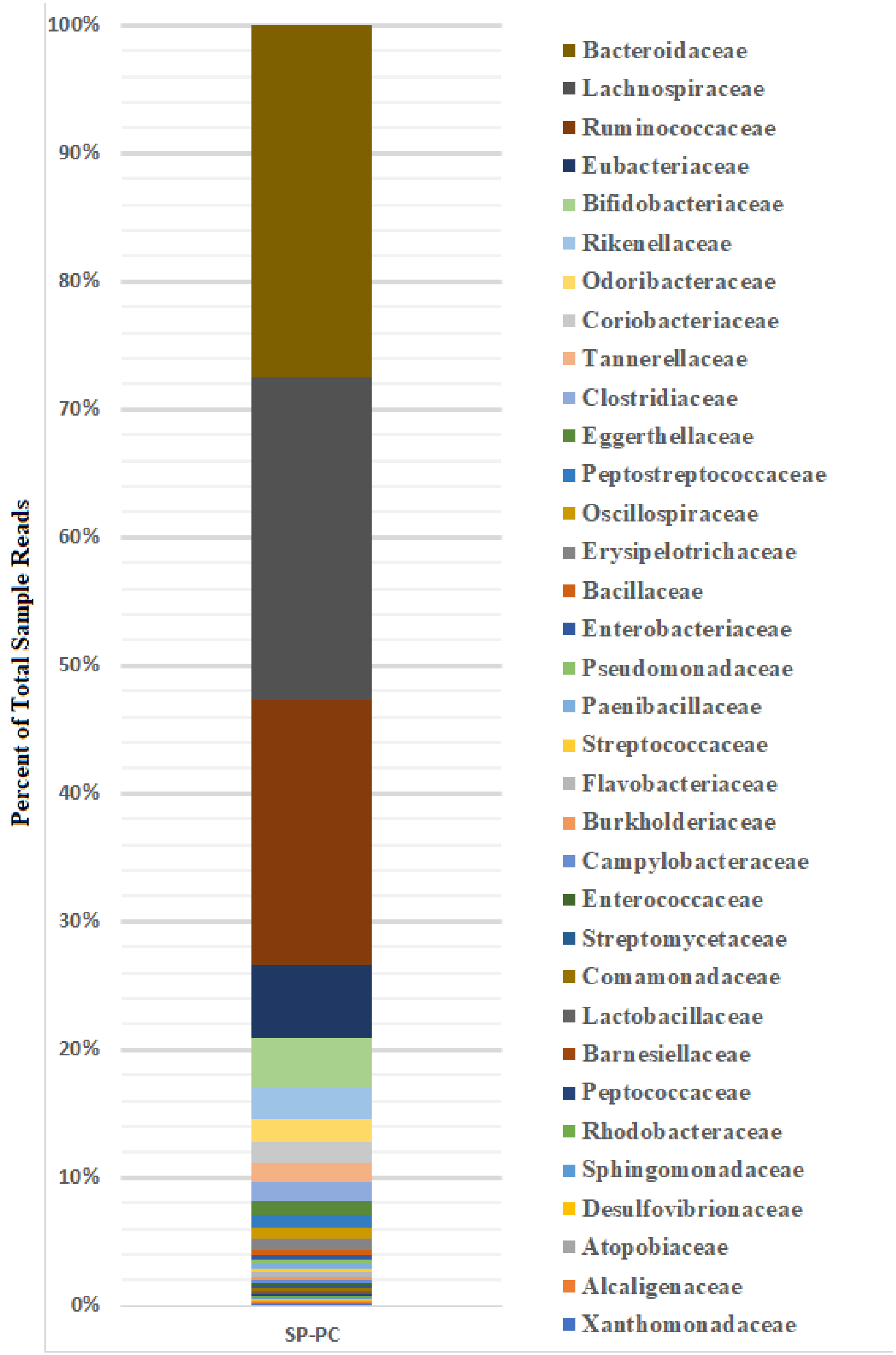
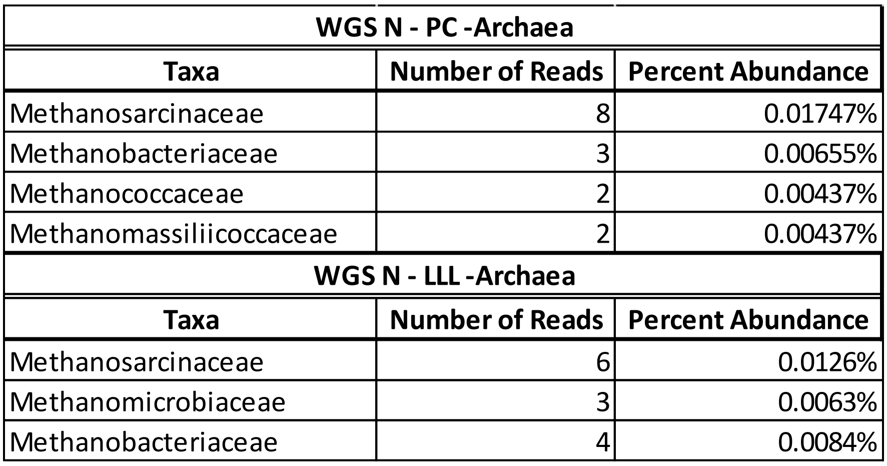


Figure 23: Metagenomic Shotgun Sequence Community Profile of Sample SP - PC Using Bracken. Family names represented with <0.1% reads were omitted from the legend.

Whole-genome shotgun reads of the faecal metagenomic DNA from participants SP and N was also searched for archaeal taxa prior to normalization. In contrast to the lack of amplification of 16S rRNA in N sample (Figures 17 A & B), WGS showed the presence of reads from low abundance archaea (Table 11).

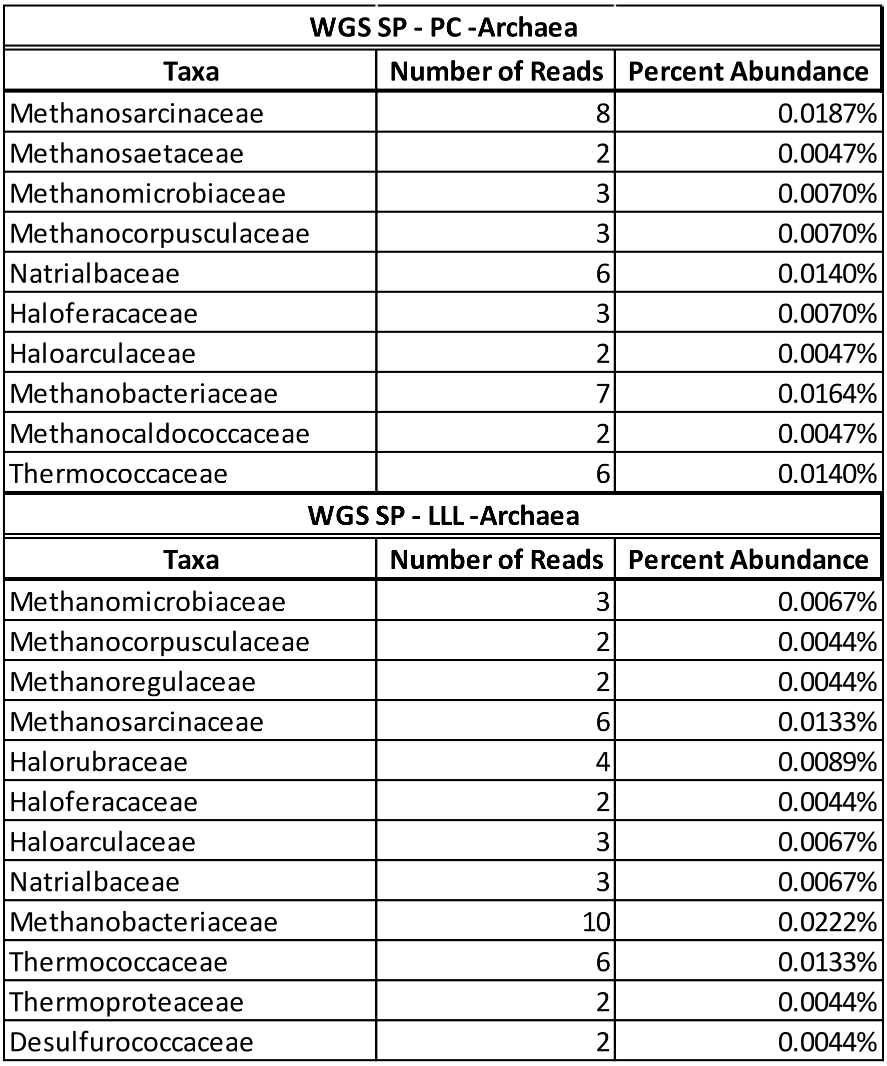
Table 11: Archaeal Reads Identified in the Faecal Metagenome of Participant N in Phenol/Chloroform (PC) cleaned and PC followed by lone-linker ligation (LLL).



WGS N - PC -Archaea		
Taxa	**Number of Reads**	**Percent Abundance**
Methanosarcinaceae	8	0.01747%
Methanobacteriaceae	3	0.00655%
Methanococcaceae	2	0.00437%
Methanomassiliicoccaceae	2	0.00437%
WGS N - LLL -Archaea		
Taxa	**Number of Reads**	**Percent Abundance**
Methanosarcinaceae	6	0.0126%
Methanomicrobiaceae	3	0.0063%
Methanobacteriaceae	4	0.0084%

In addition, participant SP's faecal metagenome showed more archaeal diversity (Table 12) than what have been observed by analysis of 16S rRNA sequences, which taxonomically assigned all reads to *M. smithii* (Table 6). Taxa identifications for both of participant SP's WGS samples indicate the presence of methanogenic, halophilic, and other archaeal lineages. However, all archaeal taxa identified in the sample are in low abundance (below 0.02% of total reads).

Table 12: Archaeal Reads Identified in Faecal Metagenome from Participant SP Phenol/Chloroform (PC) cleaned and PC followed by lone-linker ligation (LLL).



WGS SP - PC -Archaea

Taxa	Number of Reads	Percent Abundance
Methanosarcinaceae	8	0.0187%
Methanosaetaceae	2	0.0047%
Methanomicrobiaceae	3	0.0070%
Methanocorpusculaceae	3	0.0070%
Natrialbaceae	6	0.0140%
Haloferacaceae	3	0.0070%
Haloarculaceae	2	0.0047%
Methanobacteriaceae	7	0.0164%
Methanocaldococcaceae	2	0.0047%
Thermococcaceae	6	0.0140%

WGS SP - LLL -Archaea

Taxa	Number of Reads	Percent Abundance
Methanomicrobiaceae	3	0.0067%
Methanocorpusculaceae	2	0.0044%
Methanoregulaceae	2	0.0044%
Methanosarcinaceae	6	0.0133%
Halorubraceae	4	0.0089%
Haloferacaceae	2	0.0044%
Haloarculaceae	3	0.0067%
Natrialbaceae	3	0.0067%
Methanobacteriaceae	10	0.0222%
Thermococcaceae	6	0.0133%
Thermoproteaceae	2	0.0044%
Desulfurococcaceae	2	0.0044%

Chapter Four: Discussion

4.1 Archaeal DNA Shows Unexpected Sensitivity to DSN Treatment

Amplification after DSN based normalization failed to generate any amplicons using archaea-specific 16S rRNA oligonucleotides for the V6-V8 regions of this phylogenetic marker. This is in spite of the participant SP microbiome showing the presence of archaea using this same oligonucleotide pair prior to normalization. Although the method was optimised in several instances, conditions used, including DNA denaturation and renaturations, were not changed. The lack of archaeal 16S rRNA amplicons after DSN normalisation could be explained by digestion of the gene upon treatment with this enzyme. However, the reasons why archaeal DNA would renature with same kinetics for ssDNA from rare and dominant sequences is difficult to envisage.

Previous studies using DSN for normalisation have been performed on bacterial DNA or cDNA, thus this study was the first to assess utilisation of this enzyme against dsDNA in Archaea. When cDNA is generated from RNA it lacks introns, which can contain many repetitive bases and therefore cross-hybridization can occur. The hybridization conditions we used allow for cDNA with up to 87% sequence identity to not cross-hybridize (Shagina et al., 2010), however, the presence of introns and their repeat sequences could potentially reduce this threshold for cross-hybridization sufficiently to enable DSN digestion during our normalization procedure. Until recently few rRNA introns have been described in archaea, however, the latest bioinformatic analyses of archaeal genomes have uncovered the presence of many group-I self-splicing introns in both the 16S- and 23S rRNA genes (Nawrocki et al., 2018; Tocchini-Valentini et al., 2011). Hybridization of these repetitive elements could offer a possible explanation for the complete digestion of archaeal 16S rRNA sequences in the SP sample after the first round of normalization. Furthermore, the presence and/or absence of these introns in various bacterial genomes could also explain the irregularities in normalization between different taxa in our bacterial 16S rRNA data. For example, some species in participant N (*Bacteroidaceae*) became dominant over subsequent rounds of normalization, while others (*Lachnospiraceae*) were normalized over subsequent rounds.

Optimization of conditions to reduce the formation of heteroduplexes, including decreasing NaCl concentration of the hybridization buffer, could potentially be a way forward. This would lessen the likelihood of unwanted hybridization because salt cations reduce the repulsion of the negatively charged phosphate backbone of DNA strands, allowing complimentary strands to more easily hybridize (Sikorav & Church, 1991). Reducing salt concentration would therefore decrease hybridization efficiency but also increase stringency. A similar result may also be obtainable by increasing the hybridization reaction temperature, which would also increase the stringency of the hybridization reaction (in the same way it does for PCR) (Lorenz, 2012). Both of these potential solutions to this issue would result in a less efficient normalization per round, as they would also affect the hybridization of high abundance DNA in the same way (preventing it from being digested by the DSN). This could necessitate the need for more than five rounds of normalization to compensate.

4.2 The Rare Bacterial Biosphere of the Human Gut Microbiome

Enrichment of low abundance DNA belonging to the rare bacterial biosphere of participants gut metagenomes showed that the DSN normalization workflow was successful. These rare taxa included members of *Ktedonobacteraceae*, *Synergistaceae*, *Simkaniaceae*, *Oligoflexales*, and *Caedibacteraceae*, from participant N; and *Propionibacteriaceae*, *Caulobacteraceae*, *Simkaniaceae*, *Nitrosomonadaceae*, *Caedibacteraceae*, *Oligoflexales*, and *Anaeroplasmataceae* from participant SP. Notably, these taxa were only detectable after one or more rounds of normalization.

Some of these bacterial families have been previously reported in human gut microbiomes, for example, species belonging to *Synergistaceae*, *Propionibacteriaceae*, and *Caulobacteraceae* are reported to be ubiquitous as a minor member of diverse microbiota (Abraham et al., 2014; Stackebrandt, 2014; Vartoukian et al., 2007). Furthermore, species belonging to *Simkaniaceae* have been reported in the human microbiome as early as 1993, and due its phylogenetic similarity to other *Chlamydia*-related taxa, it is suspected to potentially be pathogenic (Vouga et al., 2017). Other families, such as *Nitrosomonadaceae* and *Caedibacteraceae* have been reported in soil samples, the former representing species that are important members of the nitrogen cycle (Prosser et al., 2014), and the latter found both in soil microbiota of forests (Sridevi et al., 2012) and in association with COPD (Chronic Obstructive Pulmonary Disease) exacerbations in humans (Huang et al., 2010). Taxa belonging to *Anaeroplasmataceae* have previously been reported in human, mouse, and ruminant gut microbiomes (Du et al., 2019; Loh & Blaut, 2012; Yang et al., 2017). Furthermore, *Anaeroplasmataceae* has been reported to be detected in higher abundance in the gut microbiota of human patients with colonic Crohns disease (CCD) (Loh & Blaut, 2012). The detection of reads mapping to *Oligoflexales* was also noteworthy, as species belonging to this family have previously only been identified in sand gravels from the Sahara Desert (Nakai et al., 2014), and neither participant had recently travelled to that region. Finally, *Ktedonobacteraceae* are a family of bacteria found in soils, that have not been previously identified in the human gut microbiome (Cavaletti et al., 2006; Yabe et al., 2017).

Bioinformatic analysis of whole-metagenome samples (“PC” samples that have not been digested, ligated, normalized and amplified) resulted in the detected and taxonomic identification of 21 and 22 unique taxa at the family level from participants N and SP, respectively. In comparison, when we pool all detected and taxonomically classified (family level) sequence reads across all normalization workflow sample rounds (PC, LLL, R0 - R5), 37 and 34 unique taxa were uncovered in participants N and SP, respectively. This amounts to a 1.76-fold increase in detection sensitivity for participant N and a 1.54-fold increase in detection sensitivity for participant SP.

Although there is need for further optimization of the normalization pipeline to allow for archaeal DNA to be enriched, this study concurs with the previous report by Gagic et al. (2015) that DSN normalization could be used to increase the detection resolution of bacterial standard metabarcoding sequencing of complex metagenomic samples.

4.3 Archaea Are Elusive and Rare

Based on results from WGS sequence analysis, the diversity of archaea in the human gut microbiome does appear to be low in participant N (five families), but two-fold higher in participant SP (14 different families) (Tables 8 and 9). These archaeal families are primarily represented by methanogens and halophiles, however, they made up less than 0.03276% (participant N) and 0.0912% (participant SP) of total reads, making them part of the human gut microbiome rare biosphere. The failure to normalise archaeal DNA prevented answering the question of whether other rare species are present, aside from the dominant *Methanobrevibacter* species, rendering human gut archaea still elusive. This further underscores the potential of DSN normalization to possibly detect these elusive microorganisms if hybridisation conditions could be optimised or an archaeal phylogenetic marker without introns could be found (such as *mcrA* for methanogens).

It needs to be taken into consideration that results of this study have a statistical limitation as only two participants were studied (due to budgetary constraints). Previously, studies with a larger number of participants showed that methanogens, particularly *M. smithii* and *M. stadtmanae (Bhute et al., 2017; Dridi et al., 2009; Gaci et al., 2014)*, and halophilic archaea are present in human gut microbiomes (Oxley et al., 2010), but their diversity and abundance are dependent on diet, health, and geographical location (Hoffmann et al., 2013; Horz, 2015; Nkamga et al., 2017b). In this study participant N reported a diet high in protein and fat, and low in carbohydrates. This could elucidate the lower abundance and diversity of archaeal reads detected from that participants gut microbiome compared to participant SP, as studies have shown a negative association between high protein, high fat, diets and prevalence of methanogens in the human gut microbiome (Hoffmann et al., 2013). Furthermore, participant SP reported a vegan diet, which could explain the increased archaeal diversity in that participants gut microbiome, as a positive association between diets high in carbohydrates (which is typical of vegan diets (Key et al., 2006; Zimmer et al., 2012)) and methanogen prevalence has also been previously reported (Hoffmann et al., 2013; Nkamga et al., 2017a). The presence of halophilic archaeal DNA reads in participant SP can also potentially be explained by diet, as halophilic archaea have previously reported in the gut microbiomes of people who consumed salt-fermented seafood (Horz, 2015). In addition, viable halophilic archaea have been found in unrefined food-grade sea salt (Henriet et al., 2014). These halophiles could have potentially been introduced to the gut microbiome *via* a similar kind of salty or salt-fermented food.

Chapter 5: Conclusions

Enriching human gut metagenomic samples for rare archaeal DNA using DSN normalization remains an avenue of continuing research, as this first attempt suggests that the DSN enzyme is potent in digestion of archaeal DNA derived from a human metagenome. With further optimization of the hybridization conditions used in normalization, the possibility to uncover previously unknown diversity of archaeal lineages in the human gut microbiota remains.

In contrast to archaeal DNA, the DSN normalization workflow was able to enrich for previously undetectable bacterial DNA belonging to the rare biosphere of the human gut microbiome. These low abundance bacterial reads mapped to a range of taxa found in human and ruminant gut microbiomes, soil samples from forests and deserts, and also microorganisms associated with human disease such as COPD and CCD. This further underscores the potential utility of this workflow for both sensitive diagnostics in a medical setting, to environmental microbiome profiling.

Analysis of whole metagenomic shotgun sequence reads from our participant gut microbiota samples uncovered a greater amount of archaeal diversity than anticipated based off the archaeal 16S rRNA PCR. The archaea that were detected were in low abundance in both participants, in total representing less than 0.1% of total sample reads, and therefore constituted a portion of the rare biosphere of the human gut microbiome. As archaea could only be detected by PCR in participant SP, we can conclude than archaea are both rare, and still elusive.

Normalization of metagenomic DNA samples using duplex-specific nuclease is a promising approach for enriching for low abundance microorganisms. However, as this technique has not previously been utilized on human faecal samples, it requires further optimization before it's true potential can be realized. One of the primary advantages DSN normalization is the relatively low cost while providing improved detection resolution over standard targeted sequencing approaches, compared to costly deep WGS sequencing. Furthermore, it has previously been discussed that DSN normalization would be relatively easy to automate, which could potentially make this method a standard laboratory procedure for enrichment and subsequent detection of low abundance taxa in complex metagenomic samples (Gijavanekar et al., 2012). In addition to metagenomics DSN normalization has also shown potential for use in forensic DNA analysis, as it can be used to enrich for low copy number DNA from evidence swabs and blood/tissue samples (Sambol & Creecy, 2013).

Chapter 6: Future Steps

Despite the promising potential of the DNA normalization technique in metagenomics, obstacles still remain before it can be utilized to enrich for low abundance archaea. To mitigate these obstacles, the hybridization reaction needs to be adjusted either by an increase in temperature or in stringency of the hybridisation buffer. A series of hybridization reactions using decreasing NaCl concentrations, followed by a DSN digestion, could be performed until archaeal 16S rRNA amplicons are generated. Additionally, a series of hybridization reactions at increasing temperatures could be undertaken to further refine hybridization stringency, and therefore reduce cross-hybridisation between introns in archaeal 16S rRNA. These two optimizations should be undertaken separately at first, then in unison, until an optimal balance for amplification of ssDNA from rare archaeal sequences is achieved. This could be accomplished by constructing a synthetic archaeal metagenome consisting of different molar ratios of *M. smithii*, *M. stadtmanae*, *M. ruminantium*, and a few halophilic archaea; then testing the different hybridization conditions that were previously mentioned until post-normalization amplification of archaeal ssDNA is accomplished.

Another approach to identify rare archaeal taxa could be by utilisation of a phylogenetic marker specific for archaea and without introns from either group I or group II such as *mcrA*, RadA, or RadB. The *mcrA* phylogenetic marker is a gene that codes for the α subunit of methyl coenzyme M reductase, the enzyme which catalyses the final step of methanogenesis in methanogenic archaea. In particular, *mcrA* has previously been used in other studies to compliment 16S rRNA phylogenetic analysis of methanogens (Luton et al., 2002; Mihajlovski et al., 2008; Vianna et al., 2006). Other archaea-specific phylogenetic marker genes such as RadA and RadB are more universal for archaea, as they are homologous to the highly conserved RecA family of recombinases in bacterial lineages (Guy et al., 2006; Haldenby et al., 2009).

.

References

Abraham, W.-R., Rohde, M., & Bennasar, A. (2014). The Family Caulobacteraceae. In E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.), *The Prokaryotes: Alphaproteobacteria and Betaproteobacteria* (pp. 179-205). Berlin, Heidelberg: Springer Berlin Heidelberg. 10.1007/978-3-642-30197-1_259

Albers, S.-V., & Meyer, B. H. (2011). The Archaeal Cell Envelope. *Nature Reviews Microbiology, 9*(6), 414-426. 10.1038/nrmicro2576

Allali, I., Arnold, J. W., Roach, J., Cadenas, M. B., Butz, N., Hassan, H. M., Azcarate-Peril, M. A. (2017). A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiology, 17*(1), 194-194. 10.1186/s12866-017-1101-8

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Knight, R. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems, 2*(2), e00191-00116. 10.1128/mSystems.00191-16

Aronesty, E. (2011). ea-utils : "Command-line tools for processing biological sequencing data": https://github.com/ExpressionAnalysis/ea-utils.

Aronesty, E. (2013). Comparison of sequencing utility programs. *Open Bioinformatics Journal, 7*(1), 1-8. 10.2174/1875036201307010001

Bag, S., Saha, B., Mehta, O., Anbumani, D., Kumar, N., Dayal, M., & Pant, A. (2016). An Improved Method for High Quality Metagenomics DNA Extraction from Human and Environmental Samples. *Scientific Reports, 6,* 26775. 10.1038/srep26775

Bhute, S. S., Ghaskadbi, S. S., & Shouche, Y. S. (2017). *Rare biosphere in human gut: A less explored component of human gut microbiota and its association with human health*: Springer Singapore. 10.1007/978-981-10-5708-3_8

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology, 37*(8), 852-857. 10.1038/s41587-019-0209-9

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods, 13*(7), 581-583. 10.1038/nmeth.3869

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*(1), 421. 10.1186/1471-2105-10-421

Cavaletti, L., Monciardini, P., Bamonte, R., Schumann, P., Rohde, M., Sosio, M., & Donadio, S. (2006). New lineage of filamentous, spore-forming, gram-positive bacteria from soil. *Appl Environ Microbiol, 72*(6), 4360-4369. 10.1128/aem.00132-06

Chouvarine, P., Wiehlmann, L., Moran Losada, P., DeLuca, D. S., & Tümmler, B. (2016). Filtration and Normalization of Sequencing Read Data in Whole-Metagenome Shotgun Samples. *PLoS One, 11*(10), 1-16. 10.1371/journal.pone.0165015

Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., & O'Toole, P. W. (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research, 38*(22), e200-e200. 10.1093/nar/gkq873

Clooney, A. G., Sutton, T. D. S., Shkoporov, A. N., Holohan, R. K., Daly, K. M., O’Regan, O., Hill, C. (2019). Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host & Microbe, 26*(6), 764-778.e765. https://doi.org/10.1016/j.chom.2019.10.009

Coenye, T., & Vandamme, P. (2003). Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiology Letters, 228*(1), 45-49. 10.1016/S0378-1097(03)00717-1

Cox, M. P., Peterson, D. A., & Biggs, P. J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics, 11*(1), 485. 10.1186/1471-2105-11-485

D'Argenio, V. (2018). Human Microbiome Acquisition and Bioinformatic Challenges in Metagenomic Studies. *International Journal of Molecular Sciences, 19*(2) 10.3390/ijms19020383

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Turnbaugh, P. J. (2014). Diet Rapidly and Reproducibly Alters the Human Gut Microbiome. *Nature, 505*(7484), 559 - 576.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Andersen, G. L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology, 72*(7), 5069-5072. 10.1128/aem.03006-05

Dridi, B., Henry, M., El Khechine, A., Raoult, D., & Drancourt, M. (2009). High Prevalence of *Methanobrevibacter smithii* and *Methanosphaera stadtmanae* Detected in the Human Gut Using an Improved DNA Detection Protocol. *PLoS One, 4*(9)

Du, H., Erdene, K., Chen, S., Qi, S., Bao, Z., Zhao, Y., Ao, C. (2019). Correlation of the rumen fluid microbiome and the average daily gain with a dietary supplementation of Allium mongolicum Regel extracts in sheep 1. *Journal of Animal Science, 97*(7), 2865.

Evans, G. A. (1990). Molecular cloning: A laboratory manual. Second edition. Volumes 1, 2, and 3. Current protocols in molecular biology. Volumes 1 and 2: By J. Sambrook, E. F. Fritsch, and T. Maniatis. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. (1989). 1626 pp. $115.00. Edited by F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl. New York: Greene Publishing Associates and John Wiley & Sons. (1989). 1120 pp. *Cell, 61*(1), 17-18. 10.1016/0092-8674(90)90210-6

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation, 61*(1), 1-10. https://doi.org/10.1016/0006-3207(92)91201-3

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., Rosenbaum, M. (2013). The Long-Term Stability of the Human Gut Microbiota. *Science, 341*(6141), 45 - 52. 10.1126/science.1237439

FastQC. (2019). Barbraham Institute: Babraham Bioinformatics Group. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Gaci, N., Borrel, G., Tottey, W., O’Toole, P. W., & Brugère, J.-F. (2014). Archaea and the Human Gut: New Beginning of an Old Story. *World Journal of Gastroenterology, 20*(43), 16062-16078. 10.3748/wjg.v20.i43.16062

Gagic, D., Maclean, P. H., Li, D., Attwood, G. T., & Moon, C. D. (2015). Improving the Genetic Representation of Rare Taxa Within Complex Microbial Communities Using DNA Normalization Methods. *Molecular Ecology Resources, 15*(3), 464-476. 10.1111/1755-0998.12321

Gijavanekar, C., Strych, U., Fofanov, Y., Fox, G. E., & Willson, R. C. (2012). Rare target enrichment for ultrasensitive PCR detection using cot–rehybridization and duplex-specific nuclease. *Analytical Biochemistry, 421*(1), 81-85. https://doi.org/10.1016/j.ab.2011.11.010

Guy, C. P., Haldenby, S., Brindley, A., Walsh, D. A., Briggs, G. S., Warren, M. J., Bolt, E. L. (2006). Interactions of RadB, a DNA repair protein in archaea, with DNA and ATP. *Journal of Molecular Biology, 358*(1), 46-56. 10.1016/j.jmb.2006.02.010

Haldenby, S., White, M. F., & Allers, T. (2009). RecA family proteins in archaea: RadA and its cousins. *Biochemical Society Transactions, 37*(Pt 1), 102-107. 10.1042/bst0370102

Heintz-Buschart, A., & Wilmes, P. (2018). Human Gut Microbiome: Function Matters. *Trends in Microbiology, 26*(7), 563-574. https://doi.org/10.1016/j.tim.2017.11.002

Henderson, G., Cox, F., Kittelmann, S., Miri, V. H., Zethof, M., Noel, S. J., Janssen, P. H. (2013). Effect of DNA Extraction Methods and Sampling Techniques on the Apparent Structure of Cow and Sheep Rumen Microbial Communities. *PLoS One, 8*(9), e74787. 10.1371/journal.pone.0074787

Henriet, O., Fourmentin, J., Delincé, B., & Mahillon, J. (2014). Exploring the diversity of extremely halophilic archaea in food-grade salts. *International Journal of Food Microbiology, 191*, 36-44. https://doi.org/10.1016/j.ijfoodmicro.2014.08.019

Hoffmann, C., Dollive, S., Grunberg, S., Chen, J., Li, H., Wu, G. D., Bushman, F. D. (2013). Archaea and fungi of the human gut microbiome: correlations with diet and bacterial residents. *PLoS One, 8*(6), e66019-e66019. 10.1371/journal.pone.0066019

Horz, H.-P. (2015). Archaeal Lineages Within the Human Microbiome: Absent, Rare or Elusive? *Life, 5*(2), 1333-1345. 10.3390/life5021333

Huang, Y. J., Kim, E., Cox, M. J., Brodie, E. L., Brown, R., Wiener-Kronish, J. P., & Lynch, S. V. (2010). A persistent and diverse airway microbiota present during chronic obstructive pulmonary disease exacerbations. *Omics : a journal of integrative biology, 14*(1), 9-59. 10.1089/omi.2009.0100

Hwang, K.-B., Lee, I.-H., Li, H., Won, D.-G., Hernandez-Ferrer, C., Negron, J. A., & Kong, S. W. (2019). Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. *Scientific Reports, 9*(1), 3219. 10.1038/s41598-019-39108-2

Ipci, K., AltAntoprak, N., Muluk, N. B., Senturk, M., & Cingi, C. (2017). The Possible Mechanisms of the Human Microbiome in Allergic Diseases. *European Archives of Oto-Rhino-Laryngology*(2), 617 - 626. 10.1007/s00405-016-4058-6

Jia, X., Dini-Andreote, F., & Falcão Salles, J. (2018). Community Assembly Processes of the Microbial Rare Biosphere. *Trends in Microbiology, 26*(9), 738-747. https://doi.org/10.1016/j.tim.2018.02.011

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications, 10*(1), 5029. 10.1038/s41467-019-13036-1

Key, T. J., Appleby, P. N., & Rosell, M. S. (2006). Health effects of vegetarian and vegan diets. *Proceedings of the Nutrition Society, 65*(1), 35-41. 10.1079/PNS2005481

Kho, Z. Y., & Lal, S. K. (2018). The Human Gut Microbiome – A Potential Controller of Wellness and Disease. *Frontiers in Microbiology, 9*(1835) 10.3389/fmicb.2018.01835

Kitahara, K., & Miyazaki, K. (2013). Revisiting bacterial phylogeny. *Mobile Genetic Elements, 3*(1), e24210. 10.4161/mge.24210

Kittelmann, S., Seedorf, H., Walters, W. A., Clemente, J. C., Knight, R., Gordon, J. I., & Janssen, P. H. (2013). Simultaneous Amplicon Sequencing to Explore Co-Occurrence Patterns of Bacterial, Archaeal and Eukaryotic Microorganisms in Rumen Microbial Communities. *PLoS One, 8*(2), e47879. 10.1371/journal.pone.0047879

Ko, M. S., Ko, S. B., Takahashi, N., Nishiguchi, K., & Abe, K. (1990). Unbiased amplification of a highly complex mixture of

DNA fragments by 'lone linker'-tagged PCR. *Nucleic Acids Research, 18*(14), 4293-4294.

Kraal, L., Abubucker, S., Kota, K., Fischbach, M. A., & Mitreva, M. (2014). The Prevalence of Species and Strains in the Human Microbiome: A Resource For Experimental Efforts. *PLoS One, 9*(5), e97279. 10.1371/journal.pone.0097279

Lagier, J.-C., Edouard, S., Pagnier, I., Mediannikov, O., Drancourt, M., & Raoult, D. (2015). Current and Past Strategies for Bacterial Culture in Clinical Microbiology. *Clinical Microbiology Reviews, 28*(1), 208-236. 10.1128/CMR.00110-14

Lagier, J. C., Khelaifia, S., Alou, M. T., Ndongo, S., Dione, N., Hugon, P., Raoult, D. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat Microbiol, 1,* 16203. 10.1038/nmicrobiol.2016.203

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods, 9*(4), 357-359. 10.1038/nmeth.1923

Leuko, S., Goh, F., Ibáñez-Peral, R., Burns, B. P., Walter, M. R., & Neilan, B. A. (2008). Lysis efficiency of standard DNA extraction methods for Halococcus spp. in an organic rich environment. *Extremophiles, 12*(2), 301-308. 10.1007/s00792-007-0124-8

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics, 31*(10), 1674-1676. 10.1093/bioinformatics/btv033

Li, S., Shao, Y., Li, K., HuangFu, C., Wang, W., Liu, Z., Zhao, B. (2018). Vascular Cognitive Impairment and the Gut Microbiota. *Journal of Alzheimer's Disease, 63*, 1209-1222. 10.3233/JAD-171103

Loh, G., & Blaut, M. (2012). Role of commensal gut bacteria in inflammatory bowel diseases. *Gut Microbes, 3*(6), 544-555. 10.4161/gmic.22156

Lorenz, T. C. (2012). Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. *Journal of visualized experiments : JoVE*(63), e3998-e3998. 10.3791/3998

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science, 3*, e104. 10.7717/peerj-cs.104

Luton, P. E., Wayne, J. M., Sharp, R. J., & Riley, P. W. (2002). The *mcrA* gene as an alternative to 16S rRNA in the phylogenetic analysis of methanogen populations in landfill. *Microbiology, 148*(11), 3521-3530. doi:10.1099/00221287-148-11-3521

Marsh, A. J., O'Sullivan, O., Hill, C., Ross, R. P., & Cotter, P. D. (2013). Sequencing-Based Analysis of the Bacterial and Fungal Composition of Kefir Grains and Milks from Multiple

Sources. *PLoS One, 8*(7), e69371. 10.1371/journal.pone.0069371

Mihajlovski, A., Alric, M., & Brugère, J.-F. (2008). A putative new order of methanogenic Archaea inhabiting the human gut, as revealed by molecular analyses of the mcrA gene. *Research in Microbiology, 159*(7), 516-521. 10.1016/j.resmic.2008.06.007

Million, M., Angelakis, E., Maraninchi, M., Henry, M., Giorgi, R., Valero, R., Raoult, D. (2013). Correlation Between Body Mass Index and Gut Concentrations of *Lactobacillus reuteri, Bifidobacterium animalis, Methanobrevibacter smithii* and *Escherichia coli*. *International Journal of Obesity, 37*(11), 1460-1466. 10.1038/ijo.2013.20

Million, M., Henry, M., Armougom, F., Richet, H., Raoult, D., Maraninchi, M., Raccah, D. (2012). Obesity-associated Gut Microbiota is Enriched in *Lactobacillus reuteri* and Depleted in *Bifidobacterium animalis* and *Methanobrevibacter smithii*. *International Journal of Obesity, 36*(6), 817-825. 10.1038/ijo.2011.153

MirMohammad-Sadeghi, H., Abedi, D., Mohmoudpour, H. R., & Akbari, V. (2013). Comparison of five methods for extraction of genomic DNA from a marine Archaea, Pyrococcus furiosus. *Pakistan Journal of Medical Sciences, 29*

Mohammed, A., & Guda, C. (2015). Application of a hierarchical enzyme classification method reveals the role of gut microbiome in human metabolism. *BMC Genomics, 16*(7), S16. 10.1186/1471-2164-16-S7-S16

Moissl-Eichinger, C., Pausan, M., Taffner, J., Berg, G., Bang, C., & Schmitz, R. A. (2018). Archaea Are Interactive Components of Complex Microbiomes. *Trends in Microbiology, 26*(1), 70-85. https://doi.org/10.1016/j.tim.2017.07.004

Nakai, R., Nishijima, M., Tazato, N., Handa, Y., Karray, F., Sayadi, S., Naganuma, T. (2014). Oligoflexus tunisiensis gen. nov., sp. nov., a Gram-negative, aerobic, filamentous bacterium of a novel proteobacterial lineage, and description of Oligoflexaceae fam. nov., Oligoflexales ord. nov. and Oligoflexia classis nov. *International Journal of Systematic and Evolutionary Microbiology, 64*(Pt 10), 3353-3359. 10.1099/ijs.0.060798-0

Nawrocki, E. P., Jones, T. A., & Eddy, S. R. (2018). Group I introns are widespread in archaea. *Nucleic Acids Research, 46*(15), 7970-7976. 10.1093/nar/gky414

Nkamga, V. D., Henrissat, B., & Drancourt, M. (2017a). Archaea: Essential Inhabitants of the Human Digestive Microbiota. *Human Microbiome Journal, 3*, 1-8. 10.1016/j.humic.2016.11.005

Ortiz-Burgos, S. (2016). Shannon-Weaver Diversity Index. In M. J. Kennish (Ed.), *Encyclopedia of Estuaries* (pp. 572-573). Dordrecht: Springer Netherlands. 10.1007/978-94-017-8801-4_233

Oxley, A. P., Lanfranconi, M. P., Wurdemann, D., Ott, S., Schreiber, S., McGenity, T. J., Nogales, B. (2010). Halophilic archaea in the human intestinal mucosa. *Environ Microbiol, 12*(9), 2398-2410. 10.1111/j.1462-2920.2010.02212.x

Parfrey, L. W., Walters, W. A., & Knight, R. (2011). Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Frontiers in Cellular and Infection Microbiology, Vol 2 (2011)* 10.3389/fmicb.2011.00153

Pedros-Alio, C. (2012). The Rare Bacterial Biosphere. *Annual Review of Marine Science, 4*, 449-466. 10.1146/annurev-marine-120710-100948

Picard toolkit. (2019). Broad Institute, GitHub repository: Broad Institute. Retrieved from http://broadinstitute.github.io/picard/

Prosser, J. I., Head, I. M., & Stein, L. Y. (2014). The Family Nitrosomonadaceae. In E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.), *The Prokaryotes: Alphaproteobacteria and Betaproteobacteria* (pp. 901-918). Berlin, Heidelberg: Springer Berlin Heidelberg. 10.1007/978-3-642-30197-1_372

Purohit, M. K., & Singh, S. P. (2009). Assessment of Various Methods for Extraction of Metagenomic DNA from Saline Habitats of Coastal Gujarat (India) to Explore Molecular Diversity. *Letters in Applied Microbiology, 49*(3), 338-344. 10.1111/j.1472-765X.2009.02663.x

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research, 41*(D1), D590-D596. 10.1093/nar/gks1219

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, proteomics & bioinformatics, 13*(5), 278-289. 10.1016/j.gpb.2015.08.002

Roberts, R. J., Vincze, T., Posfai, J., & Macelis, D. (2015). REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research, 43*(Database issue), D298-D299. 10.1093/nar/gku1046

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ, 4*, e2584. 10.7717/peerj.2584

Roh, S. W., Abell, G. C., Kim, K. H., Nam, Y. D., & Bae, J. W. (2010). Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends in Biotechnology, 28*(6), 291-299. 10.1016/j.tibtech.2010.03.001

Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., Goodman, R. M. (2000). Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. *Applied and Environmental Microbiology, 66*(6), 2541-2547.

Roopnarain, A., Mukhuba, M., Adeleke, R., & Moeletsi, M. (2017). Biases during DNA extraction affect bacterial and archaeal community profile of anaerobic digestion samples. *3 Biotech, 7*(6), 1-12. 10.1007/s13205-017-1009-x

Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., Carter, A. B. (2018). Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *The Journal of Molecular Diagnostics, 20*(1), 4-27. 10.1016/j.jmoldx.2017.11.003

Sambol, N., & Creecy, J. (2013). Duplex-specific nuclease (DSN): A method for the rehabilitation of low-copy number DNA profiles. *Forensic Science International: Genetics Supplement Series, 4*(1), e59-e60. https://doi.org/10.1016/j.fsigss.2013.10.030

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Kyrpides, N. C. (2017). Critical Assessment of Metagenome Interpretation - a benchmark of metagenomics software.

Shagina, I., Bogdanova, E., Mamedov, I., Lebedev, Y., Lukyanov, S., & Shagin, D. (2010). Normalization of Genomic DNA Using Duplex-specific Nuclease. *BioTechniques, 48*(6), 455-459. 10.2144/000113422

Shannon, C. E. (1963). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Siegwald, L., Touzet, H., Lemoine, Y., Hot, D., Audebert, C., & Caboche, S. (2017). Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics. *PLoS One, 12*(1), e0169563. 10.1371/journal.pone.0169563

Sikorav, J.-L., & Church, G. M. (1991). Complementary recognition in condensed DNA: Accelerated DNA renaturation. *Journal of Molecular Biology, 222*(4), 1085-1108. https://doi.org/10.1016/0022-2836(91)90595-W

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences of the United States of America, 103*(32), 12115-12120. 10.1073/pnas.0605127103

Sridevi, G., Minocha, R., Turlapati, S. A., Goldfarb, K. C., Brodie, E. L., Tisa, L. S., & Minocha, S. C. (2012). Soil bacterial communities of a calcium-supplemented and a reference watershed at the Hubbard Brook Experimental Forest (HBEF), New Hampshire, USA. *FEMS Microbiology Ecology, 79*(3), 728-740. 10.1111/j.1574-6941.2011.01258.x

Stackebrandt, E. (2014). The Family Propionibacteriaceae: Genera other than Propionibacterium. In E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.), *The Prokaryotes: Actinobacteria* (pp. 725-741). Berlin, Heidelberg: Springer Berlin Heidelberg. 10.1007/978-3-642-30138-4_194

Sun, Y., Liu, Y., Pan, J., Wang, F., & Li, M. (2020). Perspectives on Cultivation Strategies of Archaea. *Microbial Ecology, 79*(3), 770-784. 10.1007/s00248-019-01422-7

Teraoka, H., & Tsukada, K. (1987). Influence of Polyethylene Glycol on the Ligation Reaction with Calf Thyinus DNA Ligases I and II. *Journal of Biochemistry, 101*, 225-231. 10.1093/oxfordjournals.jbchem.a121895

Thursby, E., & Juge, N. (2017). Introduction to the human gut microbiota. *The Biochemical journal, 474*(11), 1823-1836. 10.1042/BCJ20160510

Tian, R.-M., Cai, L., Zhang, W.-P., Cao, H.-L., & Qian, P.-Y. (2015). Rare Events of Intragenus and Intraspecies Horizontal Transfer of the 16S rRNA Gene. *Genome Biology and Evolution, 7*(8), 2310-2320. 10.1093/gbe/evv143

Tocchini-Valentini, G. D., Fruscoloni, P., & Tocchini-Valentini, G. P. (2011). Evolution of introns in the archaeal world. *Proceedings of the National Academy of Sciences, 108*(12), 4782. 10.1073/pnas.1100862108

Urry, L. A. (2018). *Campbell biology* (11th edition ed.): Pearson Australia.

Vandernoot, V. A., Langevin, S. A., Solberg, O. D., Lane, P. D., Curtis, D. J., Bent, Z. W., Lane, T. W. (2012). cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications. *BioTechniques, 53*(6), 373-380. 10.2144/000113937

Vartoukian, S. R., Palmer, R. M., & Wade, W. G. (2007). The division "Synergistes". *Anaerobe, 13*(3-4), 99-106. 10.1016/j.anaerobe.2007.05.004

Vianna, M. E., Conrads, G., Gomes, B. P. F. A., & Horz, H. P. (2006). Identification and Quantification of Archaea Involved in Primary Endodontic Infections. *Journal of Clinical Microbiology, 44*(4), 1274-1282. 10.1128/JCM.44.4.1274-1282.2006

Vincent, C., Miller, M. A., Edens, T. J., Mehrotra, S., Dewar, K., & Manges, A. R. (2016). Bloom and bust: intestinal microbiota dynamics in response to hospital exposures and Clostridium difficile colonization or infection. *Microbiome, 4*(1), 12. 10.1186/s40168-016-0156-3

Visweswaran, G. R. R., Dijkstra, B. W., & Kok, J. (2010). Two major archaeal pseudomurein endoisopeptidases: PeiW and PeiP. *Archaea (Vancouver, B.C.), 2010*, 480492-480492. 10.1155/2010/480492

Vouga, M., Baud, D., & Greub, G. (2017). Simkania negevensis, an insight into the biology and clinical importance of a novel member of the Chlamydiales order. *Critical Reviews in Microbiology, 43*(1), 62-80. 10.3109/1040841X.2016.1165650

Wingett, S., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control: F1000 Research

Limited. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30254741

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6124377/

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology, 20*(1), 257. 10.1186/s13059-019-1891-0

Yabe, S., Sakai, Y., Abe, K., Yokota, A., Take, A., Matsumoto, A., Otsuka, S. (2017). Dictyobacter aurantiacus gen. nov., sp. nov., a member of the family Ktedonobacteraceae, isolated from soil, and emended description of the genus *Thermosporothrix*. *International Journal of Systematic and Evolutionary Microbiology, 67*(8), 2615-2621. 10.1099/ijsem.0.001985

Yang, T., Ahmari, N., Schmidt, J. T., Redler, T., Arocha, R., Pacholec, K., Zubcevic, J. (2017). Shifts in the Gut Microbiota Composition Due to Depleted Bone Marrow Beta Adrenergic Signaling Are Associated with Suppressed Inflammatory Transcriptional Networks in the Mouse Colon. *Frontiers in Physiology, 8*(220) 10.3389/fphys.2017.00220

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology, 12*(9), 635-645. 10.1038/nrmicro3330

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell, 178*(4), 779-794. 10.1016/j.cell.2019.07.010

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Glöckner, F. O. (2013). The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research, 42*(D1), D643-D648. 10.1093/nar/gkt1209

Yu, Z., & Morrison, M. (2004). Improved extraction of PCR-quality community DNA from digesta and fecal samples. *BioTechniques, 36*(5), 808 - 812.

Zarrinpar, A., Chaix, A., Yooseph, S., & Panda, S. (2014). Article: Diet and Feeding Pattern Affect the Diurnal Dynamics of the Gut Microbiome. *Cell Metabolism, 20*, 1006-1017. 10.1016/j.cmet.2014.11.008

Zhou, Z., Fang, L., Meng, Q., Li, S., Chai, S., Liu, S., & Schonewille, J. T. (2017). Assessment of Ruminal Bacterial and Archaeal Community Structure in Yak (Bos grunniens). *Frontiers in Microbiology, 8*, 179-179. 10.3389/fmicb.2017.00179

Zimmer, J., Lange, B., Frick, J. S., Sauer, H., Zimmermann, K., Schwiertz, A., Enck, P. (2012). A vegan or vegetarian diet substantially alters the human colonic faecal microbiota. *European Journal of Clinical Nutrition, 66*(1), 53-60. 10.1038/ejcn.2011.141

Appendix 1: Instructions for Participants: Collection of faecal material

Donation kit contains:

i. Cooler bag
ii. Ice pack
iii. Gloves (3)
iv. Sterile jar
v. Rectangular container (optional)
vi. Spoon (optional)
vii. Paper bag for waste

1. Prior to collection, place the ice pack in freezer.
2. On day of collection, place the ice pack inside cooler bag.
3. Wear the supplied gloves during sample collection.
4. Use the provided pre-labelled faecal specimen container for sample collection.
5. The sample must be collected directly into the sample container. Try to fill ¾ or more of the container *via* the following. two options:
    a. Collect sample directly into plastic bag-lined jar.
    b. Or use the bigger rectangular container to collect all faeces then transfer to the plastic bag-lined jar.
6. Seal the container and place inside the cooler bag.
7. Remove gloves and wash your hands.
8. Deliver cooler-bag to the drop-box outside the laboratory and notify the lab worker immediately.

Note: The faeces were obtained from donors who, self-assessed as healthy and had not taken antibiotics for last 3 months, with approval from Central Health and Disabilities Ethics Committee, New Zealand (13/CEN/144).