Mining for the Rumen Rare Biosphere

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

Masters

in

Microbiology

Massey University, Manawatu,

New Zealand

Stephanie Baird

2020

Abstract

The microbial diversity present in the gut microbiome of ruminant animals is of great interest due to its effect on the New Zealand economy. The rumen, a forestomach of ruminants, is a large fermentation chamber. The microbiome within the rumen influences production of milk and meat, and additionally impacts on climate change through the emission of enteric methane. Although, the core microbiome has been studied intensely, the rare biosphere, which is comprised of the rare microorganisms present in less than 0.1% of the abundance, is still largely unknown. Recent developments in methods for subtraction, or normalisation, of the dominant microorganisms from analysis of complex microbiomes, including treatment with duplex-specific nuclease (DSN), have enabled the increase of the number of sequences from low abundance microorganisms. Decreasing presence of dominant species and simultaneously increasing low abundant allows the exploration of the rare biosphere and discovery of taxa which otherwise would not have been identified. By applying DSN-based normalisation to a metagenomic DNA isolated from the rumen microbiome, we have demonstrated that the low abundance microorganisms, can be amplified to a detectable level while decreasing the abundance of sequences from dominant species. The outcome of DNA normalisation, primarily taxonomic assignment and phylogeny was assessed by using the gene encoding the β subunit of bacterial RNA polymerase, *rpoB,* as well as the "gold standard" 16S rRNA as phylogenetic markers. We have demonstrated that *rpoB* could be effectively used for determining the rumen microbial community profile and could become by broader adoption from researchers, a valuable resource for microbial ecology studies. We suggest that DSN-based normalisation could be utilised for in-depth exploration of the rare biosphere as a whole, resulting in the discovery of new species, new genes and increasing understanding of the role that these rare microorganisms play in the rumen microbiome. The inclusion of *rpoB*, alone or in combination with 16S rRNA marker, in microbial ecology studies could lead to more accurate classification of the taxa.

Acknowledgements

Firstly, I wish to thank my supervisor Dr Dragana Gagic for the opportunity to do this degree under her supervision and for all her help and support throughout the many ups and downs of this project. I would also like to thank my co-supervisor Associate Professor Jasna Rakonjac, for all her guidance, time and valuable insights. I would also like to thank all the staff at the School of Fundamental Science for all their help in completing this degree.

Secondly, I would like to thank my fellow lab mates: Catrina, Majela, Rayèn, Vuong, Marina, Georgia, Stefanie, Cathy and Nick. Thank you for all the support, banter and only posting one of my meltdowns on Instagram. I'll always fondly remember all the good times we had together and making this a memorable experience.

I would also like to extend thanks to my friends Ella, and Ainsley and my partner Jesse for always being there to listen to my struggles and always having time to remind me that there's a world outside of University. I would not have been able to complete this without you guys to keep me sane.

Finally, I would like to thank my family for all their unwavering support and time spent listening to me whinge without complaint. I would never have been able to complete University without you, let alone a master's degree. Words cannot express my gratitude.

Abbreviations

%	Percent
ºC	Degrees Centigrade
A	Absorbance
AR	After Round, refers to a round of normalisation
ASV	Amplicon Sequence Variants
Bp	Base pairs
cDNA	Coding DNA
CH_3COOH	Acetic acid
CH_4	Methane
CO_2	Carbon Dioxide
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded DNA
DSN	Duplex specific Nuclease
EggNOG	Evolutionary genealogy of genes non-supervised orthologs
h	Hour
H_2	Hydrogen
H_2S	Hydrogen Sulphide
HAP	Hydroxyapatite phosphate
HMW	High Molecular Weight
HTS	High Throughput Sequencing
ID	Identification
Kb	Kilobases
LL	Lone linkers
metDNA	Metagenomic DNA
Min	Minutes
NaOAc	Sodium Acetate
NGS	Next-generation Sequencing
NR	Nucleotide redundant
OTUs	Operation taxonomic Units
PCoA	Principle Coordinate Analysis
PCR	Polymerase Chain Reaction
QIIME	Quantitative Insights Into Molecular Ecology

RNA	Ribonucleic acid
rRNA	Ribosomal RNA
RT	Room Temperature
Sec	Seconds
SO_4^{2-}	Sulphate
ssDNA	Single-Stranded DNA
V	Volts
VFA	Volatile Fatty Acids
Vol	Volumes
w/v	Weight/Volume
α	Alpha
β	Beta

Contents

Acknowledgements iii
Abbreviations iv
Contents vi
List of Figures viii
List of Tables ix
1. Introduction 1
1.2 Rumen Microbiota 3
1.2.1 The rumen microbial metabolism and end-products 4
1.3 The Rare Biosphere 5
1.3.1 Experimental approaches for studying the rare biosphere 7
1.4 Subtractive Nucleic Acid Technologies 8
1.4.1. DSN-based Normalization 10
1.5 Role of HTS sequencing in deciphering "rare biosphere." 11
1.6 Genetic Markers 12
1.6.1 Limitations of 16S rRNA as a phylogenetic marker 13
1.6.2 *RpoB* 14
1.7 Hypothesis and Aims 16
2. Materials and Methods 17
2.1. Materials 17
2.1.1 Oligonucleotides 17
2.1.2. Solutions and Buffers 18
2.1.3 Chemicals, reagents and enzymes 18
2.1.4 Bioinformatics resources and software 18
2.2 Methods 21
2.2.1 General molecular biology techniques 21
2.2.2 Sample Collection 23
2.2.3 Sample Preparation 23
2.2.3.1 Digestion of metagenomic DNA and Lone Linker Ligation. 23
2.2.3.2 Lone Linker Amplification 23
2.2.4. 16S rRNA and *rpoB* phylogenetic markers amplification 24
2.2.5 DSN based normalisation 25
2.2.6 Metagenome Sequencing 25
2.2.7 Bioinformatics Methods 26
2.2.7.1 Taxonomic Classification using 16S rRNA region 26
2.2.7.2 Taxonomic Classification of *rpoB sequences* 26
2.2.8. Statistical Analysis 27

3. Results 28
3.1 DSN-based normalisation of the rumen microbial metagenomic DNA 28
3.2 Sequencing results and initial analysis 32
3.2.1 The rumen "rare biosphere" based on DSN normalisation 16S rRNA amplicons 32
3.2.2 The rumen "rare biosphere" based on DSN-normalisation of *rpoB* amplicons 35
3.3 Analysis of DSN-based Normalisation 37
3.4 Normalisation effect on taxa distribution 39
3.5 Change in the dominant species 42
3.6 Increase in Rare Taxa 45
3.6.1 Rare taxa in PvuII digested DNA 45
3.6.2 PsiI digested DNA 47
3.7 Comparison of 16S rRNA and *rpoB* genetic markers 49
4. Discussion 51
4.1 DNA Normalisation of the Rumen Metagenome 51
4.1.1 Dominant Genera Identified in the Rumen Microbiome 52
4.1.2 The Rare Biosphere 53
4.2 The use of different restriction enzymes in fragmentation of metagenomic DNA 54
4.3 How do 16S rRNA and *rpoB* compare as genetic markers? 55
4.4 Use of Hungate1000 Database 57
4.5 Higher error rate due to methodology could have overestimated effects of DSN-based normalisation 58
4.6 Limitations 59
5. Conclusions 60
6. Future Directions 61
6.1 The whole rumen microbiome 61
6.2 Ruminant variety 61
6.3: Investigation of ecological potential in new species 62
6.4 Improve Sequencing Technology 62
6.4 Restriction Enzymes 62
Appendix 1: Taxonomic Distribution Profiles 64
Appendix 2: Changes in Taxa Classified at higher taxonomic levels 66
References 67

List of Figures.

Figure 1. The ruminant gastrointestinal tract. 2
Figure 2. Schematic overview of the DNA normalisation process. 9
Figure 3: Schematic diagram of the normalisation procedure and NGS sequencing of the rumen microbial metagenome. 22
Figure 4: Preparation of the metagenomic DNA for normalisation. 29
Figure 5: Normalised metagenomic DNA after 5 rounds of normalisation. 30
Figure 6: 16S rRNA and *rpoB* amplicons generated from non-normalised and normalised DNA. 31
Figure 7: Taxonomic profile of the distribution of 16S rRNA taxonomy at genus level. 34
Figure 8: Taxonomic profile of the distribution of *rpoB* taxonomy at genus level. 36
Figure 9: Bray Curtis dissimilarity index PCoA 38
Figure 10: The proportion change of the top ten most dominant genera for 16S rRNA. 43
Figure 11: The proportion change of the top ten most dominant genera for *rpoB*. 44
Figure 12: Rare taxa identified in PvuII digested normalised DNA 46
Figure 13: Rare taxa identified in PsiI digested normalised DNA. 48
Figure 14: Decrease in the dominance of *Eubacterium* 50
Figure S1: Taxonomic profile for the distribution of 16S rRNA taxanomy at the genus level 64
Figure S2: Taxonomic profile for the distribution of *rpoB* taxanomy at the genus level 65

List of Tables

Table 1: Oligonucleotides used in this study. 17
Table 2: Bioinformatics resources and software.....19
Table 3: Number of sequences after demultiplexing 32
Table 4: Shannon's Diversity Index for each round of normalisation 39
Table 5: Dominant and emerging genera before and after five rounds of normalisation. 41

1. Introduction

Ruminants acquire their nutrients from the fermentation of plant-based feeds. Fermentation occurs in a specialized chamber in their forestomach, called the rumen,. New Zealand's agricultural sector is heavily reliant on ruminants, including cattle, sheep and deer, and the rumen has been described as the engine of the New Zealand economy (Ciric, 2014, Clark et al., 2007). As of June 2019, there were 10.3 million cattle (6.4 million dairy and 3.9 million beef), 26.7 million sheep and 800 thousand deer in New Zealand, contributing to a total agricultural export value of $31 billion (Statistics New Zealand, 2019, Beef and Lamb, 2019).

The ruminant gastrointestinal tract (GIT) is made of a series of large chambers (**Figure 1**). Vegetation is swallowed, passing through the oesophagus to the first chamber, the rumen. Muscular contractions mix the ingested plant material with saliva and the rumen microbes. Once the plant matter has been partially broken down by microbial digestion and this churning, it moves to the reticulum where it remains until it is regurgitated as cud, re-masticated, and swallowed again. Once the digesta has been broken down to a liquid, it flows out of the rumen into the omasum, (Grünberg and Constable, 2009). The rumen microbes hydrolyse cellusloses, hemicelluloses, pectins, frucrutosans and other polysaccharides into simple sugars where they are fermented forming the final products of acids, which are absorbed through the rumen epithluem, and the waste products methane and carbon dioxide (CO_2) (Stewart and Hobson, 1997). Lastly, digesta passes from the omasum to the abomasum, also referred to as the "true stomach": where the digesta is further broken down by digestive enzymes, and nutrients are absorbed. Lipids in the feeds are hydrolysed into long chain fatty acids which pass on to be absorbed in the small intestine.

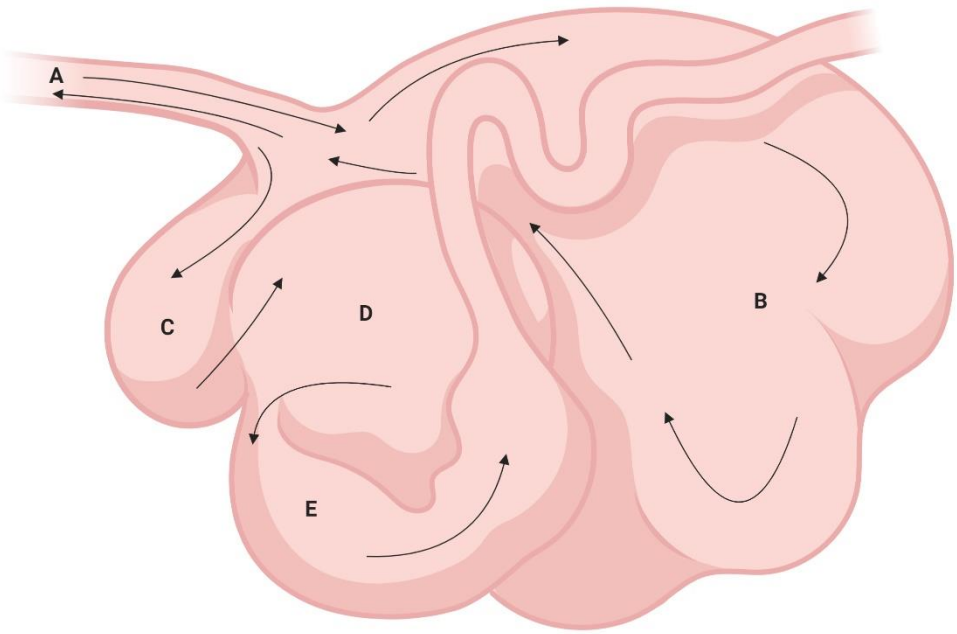

Figure 1. The ruminant gastrointestinal tract.

The feed is swallowed and moves down the oesophagus (A). Digesta enters the rumen (B), then moves to reticulum (C). Once broken to a liquid, digesta moves into the omasum (D) and finally passes through to the abomasum (E). (Adopted from BioRender.com).

In New Zealand, ruminants are predominantly fed on perennial pastures, characterised by a combination of grasses and legumes (Rattray et al., 2007). These pastures are rich in fibre and therefore difficult to digest, and must be broken down by a combination of mastication and microbial digestion (Stafford, 2017). An intrinsic part of the breakdown of fibre is the digestion of cellulose and lignocellulose in the cell walls and structural cells of these plants. As vertebrates lack cellulases, the enzymes required to digest cellulose, are provided by the rumen microbial community (Moon et al., 2014). The digestibility of the lignocellulose component of fibre is a key limiting factor for increasing milk and meat production from ruminants (Ciric, 2014, Azizi-Shotorkhoft et al., 2016). The rumen microbiota (the collection of all microorganisms in the rumen), is therefore of great importance for the wellbeing, and productivity of an animal.

1.2 Rumen Microbiota

Within the rumen microbiota, bacteria are the dominant microorganisms with 10^{10} - 10^{11} viable cells per gram of rumen contents (Alzahal et al., 2017). Other microorganisms include methanogenic archaea (about 10^{10} cells/g), bacteriophages (10^{7} - 10^{9} phage/g), ciliate protozoa (10^{4} - 10^{6} cells/g), and anaerobic fungi (10^{2} - 10^{4} cells/g) (Mackie et al., 2013, Seedorf et al., 2015). Similarly to other GIT microbiota, the dominant taxa of the rumen bacteria differs depending on animal breed, geographical location, diet, and between individuals (Zhemakova et al., 2016). However, a core microbiota is recognised within all GITs and includes the bacterial phyla Bacteroidetes, Firmicutes, Proteobacteria and Actinobacteria (Jewell et al., 2015, Xue et al., 2018, Mizrahi, 2013). The dominant genera within these phyla are often *Fibrobacter*, *Prevotella*, *Ruminococcus*, *Coprococcus*, *Butyrivibrio*, genera from unclassified Clostridiales class, and unclassified Lachnospiraceae family (Henderson et al., 2015a, Xue et al., 2018, Jewell et al., 2015).

Different microbial profiles in the rumen microbiota can influence the fermentation process. For instance, in cattle, the microbial profile differs significantly between individuals with a high-efficiency digestion profile compared to individuals with a less efficient one (Guan et al., 2008). The dominance of some phyla has also been linked to the efficiency and quality of the animal products being produced. Jami et al. (2014) demonstrated a strong correlation between the ratio of Firmicutes to Bacteroidetes and milk fat yield in cattle. Individual cattle which had a decreased ratio of Bacteroidetes to Firmucutes present in their rumen microbiome produced a higher milk fat yeild.

The wide variety of different species in the microbiota also indicates different microorganisms may be involved in the fermentation and digestion of feed. Dominant species such as *Ruminococcus albus*, *Ruminococcus flavefaciens*, and *Fibrobacter succinogenes* are known as "primary colonisers;" species involved in the initial digestion of the highly insoluble forms of cellulose in the plant cell walls (Mizrahi, 2013). Apart from these dominant and abundant species, other metabolically significant species in the rumen microbiota may be present in low abundance. For example, *Cellulosilyticum ruminicola*, which is rarely identified in phylogenetic studies, has also been shown to be a potent lignocellulose digestor in yaks (Cai et al., 2010, Guder and Krishna, 2019, Cai and Dong, 2010, Palevich, 2011).

1.2.1 The rumen microbial metabolism and end-products

The digestion of plant materials by microbial metabolism is an essential process in the rumen. Plant celluloses, hemicelluloses, pectins, fructans, starches and other polysachharides are hydrolysed into monomeric or dimeric sugars. These sugars and simple vegetation sugars and then fermented into acetic, propionic and butyric acids, methane and CO_2. Proteins on the otherhand are hydrolysed into amino acids and peptides which are then further broken down into ammonia and Volatile Fatty Acids (VFAs) (Stewart and Hobson, 1997, Gharechahi and Salekdeh, 2018). VFAs are a significant source of energy for the animal and contribute significantly to milk and meat production and therefore, the productivity of the animal (Henderson et al., 2015b, Jami et al., 2014). Hydrogen (H_2) is also producted as endproduct of fermentation, it can be further metabolised by three different pathways after its production during fermentation. Firstly, sulphate-reducing bacteria use H_2 to reduce sulphate (SO_4^{2-}) to hydrogen sulphide (H_2S). Secondly, hydrogenotrophic methane-producing bacteria and methanogenic archaea use H_2 to reduce CO_2 to methane (CH_4). And finally, reductive acetogenenic bacteria use H_2 to reduce CO_2 to acetic acid (CH_3COOH) (Verstraete, 1996). The bacteria within the rumen microbiome compete with each other, and one of these three pathways generally becomes dominant.

Due to the production of methane from methanogens, the rumen microbiota has a role in greenhouse gas emission. The agricultural sector accounted for 48% of the gross greenhouse gas emission of New Zealand in 2018, with 74% of agricultural emissions from enteric fermentation. (Environment, 2020). A recent study of dairy cattle has found that methane emission from individual cows has a genetic and microbial component, which are primarily independent of each other, and contribute to 21% and 13% of methane production respectively (Difford et al., 2018). Microbiota in sheep, which have low acetate to propionate ratio or contains lactate- and succinate-producing species, have been recorded to have significantly lower methane emissions (Kittelmann et al., 2014, Shi et al., 2014). Ruminants that produce methane as an end-product of fermentation, not only contribute to global greenhouse emissions but also lose 5-19% of the energy content from their feed (Johnson and Ward, 1996).

With the current global warming crisis and the contribution to it from enteric fermentation, understanding the rumen microbiota is of utmost importance. However, the complex interactions and energy flow between all the different microorganisms within the rumen microbiota is not yet fully understood (Mizrahi, 2013). In addition, there is a part of the rumen microbiota, termed the "rare biosphere", that is mainly undiscovered. While the more dominant taxa within the rumen microbiome have known effects on productivity and metabolic pathways, increasing the knowledge of the rare biosphere may be critical in understanding this complex microbiome.

1.3 The Rare Biosphere

The rare biosphere is defined as the collective of the rare microbial taxa, which are found in a given sample at a specific time point (Lynch and Neufeld, 2015). The threshold at which the rare biosphere begins is subject to debate but is generally accepted to be below 0.1-1% of total community relative abundance (Lynch and Neufeld, 2015, Pedrós-Alió, 2012, Ann Reid, 2011). Early evidence of the rare biosphere was found in rank abundance curves of complex microbial communities, where rare taxa are represented by the long tail of the curve (Epstein, 2009, Sogin et al., 2006).

When all of the species of the rare biosphere are combined, they often make up a large proportion of the diversity of taxa present, accounting for the high level of α diversity found in microbial communities (Lynch and Neufeld, 2015). This indicates that there may be large ecological potential in these less abundant taxa. This ecological potential can be seen in the disproportional effects which particular species within the rare biosphere, have on their communities, i.e. the keystone species (Power et al., 1996). These keystone species are often found at very low abundance in microbial communities. For example; *Desulfosporosinus* species represents only 0.006% of the total microbial community in peat soil but play a considerable part in reducing sulphate to CO_2 instead of methane (Pester et al., 2010). Communities with a higher number of rare species were shown to have higher active richness as measured by respiration rates (Dimitriu et al., 2010). Rare species also occupy a key niche and slow down the establishment of new species. This has been shown by experiments where the removal of rare species has resulted in an increased number of new species becoming established (van Elsas et al., 2012, Vivant et

al., 2013). Therefore, it appears the more diverse a community is, even if most of the diversity is not present in high numbers, the more stable it is (McCann, 2000, Shade, 2018).

By its large diversity and abundance, the rare biosphere can be considered as a microbial seed bank which could ensure plasticity of a given microbial community under environmental change. The microbial seed bank represents a vast functional gene pool for the community to access (Jousset et al., 2017). If the selection pressures within an environment change, the less abundant species that are more suited to the new environmental pressure may become dominant or share their advantageous genes *via* horizontal gene transfer. This concept corresponds with the notion that "everything is everywhere, but the environment selects" which is one of the main principles in the field of microbial ecology (Becking, 1934, De Wit and Bouvier, 2006). However, not all species in the rare biosphere are able to be selected for, as there are some species which remain persistently rare. Some of these taxa are assumed to be dormant or in starvation conditions. Taxa within this part of the rare biosphere could also be occupying a small ecological niche, which only provides enough nutrients for a small number of individuals. Although dominant species in microbiomes appear to have the best position in the community due to their ability to grow to high numbers, being rare in a large community also has advantages associated with it. The majority of predation in microbial communities is due to unicellular eukaryotes and viruses. Since viruses, find their prey using encounter probability (Pedrós-Alió, 2007), the probability that a low abundance taxa will be found and predated on is extremely low. This advantage means that life in the rare biosphere gives protection against predation which could otherwise decimate a species if it grew up to higher abundance.

The rare biosphere is an important phenomenon to study due to its roles in the environment, and the health and productivity of its host, including cattle and humans. It represents an enormous pool of unexplored genetic and physiological diversity, which could contribute to advances in biotechnology (e.g.biofuels) and pharmaceuticals. It may also serve an essential role in biological interactions and the succession of different taxa in microbial communities. Study of the rare biosphere in different microbiotas may inform other areas of science and give a pool of genetic variability and novel genes to study. The rare biosphere can be studied by culture-independent techniques, but the focus on this area of research has only increased in recent years due to improvements in high

throughput sequences (HTS) technologies (Shade et al., 2012, Sogin et al., 2006). Currently, culture-independent methods are, the leading way for studying the rare biosphere.

1.3.1 Experimental approaches for studying the rare biosphere

The study of the rare biosphere using culture-independent methods began after small subunit ribosomal RNA was established as a phylogenetic gene marker for the study of microbial communities diversity (Woese and Fox, 1977, Ward et al., 1990). This technique has taken off since the late nineties due to the advances in high-throughput sequencing (HTS) technologies and after the initial discovery of vast sequence diversity at a low relative abundance in marine water and sediment samples (Sogin et al., 2006). Before this work, there was an absence of suitable methods for sampling high diversity microbial habitats, and studies relied mostly on culturing and molecular methods. Gel fingerprinting and clone libraries can give some insight into the rare biosphere, but metagenomics gives a more in-depth look into taxa with a lower relative abundance (Lynch and Neufeld, 2015). Due to the nature of HTS, some of the diversity in early studies was created by PCR errors and artefacts (Kunin et al., 2010). These problems resulted in the creation of aggressive filtering and the development of a dependable computational protocol to minimise the bias towards higher species diversity than is actually present (Goodrich et al., 2014). Clustering sequence reads into meaningful Operational Taxonomic Units (OTUs) is a critical step in this type of analysis (binning). By clustering reads that differ by less than 3% into an OTU, reads that differ by one or two bases due to sequencing error will be clustered into their correct OTU, thus alleviating the problem of sequencing errors overestimating the species diversity (Ann Reid, 2011). This solution has been shown to reduce the OTU richness by as much as 30-60% but does not reduce the amount of OTUs in the long-tailed rank abundance curves that define the rare biosphere (Huse et al., 2010).

Although sequencing technology has enabled us to get more of an insight into the species within the rare biosphere, the estimates of the amount of diversity on earth show that there is still a large proportion of species which has not been identified yet. Subtractive nucleic acid technologies, which reduce the abundance of dominant sequences and amplify less

abundant sequences may be the solution into increasing the ability to sequence deeper into the rare biosphere and identify more unknown species representatives.

1.4 Subtractive Nucleic Acid Technologies

Subtractive nucleic acid methods are commonly applied to cDNA samples before sequencing to facilitate the detection of rare transcripts (Chung et al., 2015). Traditionally, the removal of the dsDNA was achieved by hydroxyapatite (HAP) chromatography, and it is based on DNA renaturation kinetics. Here, the dsDNA and ssDNA molecules are physically separated from each other by the differential charge interaction between the Ca^{2+} ions on the surface of HAP and the negatively charged phosphorus backbone of the nucleic acids (Bernardi, 1965).

More recently, the use of a thermostable duplex-specific nuclease (DSN) isolated from the hepatopancreas of the Kamchatka crab has been shown to be a more efficient method of achieving DNA normalisation (Bogdanova et al., 2008). This method initially involves ligation of specific oligonucleotides, lone linkers (Ko et al., 1990), to the ends of the DNA fragments, amplifying with PCR, and then during the PCR hybridisation step, the more common DNA sequences hybridise, and the double-stranded DNA (dsDNA) is removed, leaving the single-stranded DNA (ssDNA) (**Figure 2**). This method is still based on DNA renaturation kinetics but differs as dsDNA and ssDNA are separated by enzymatic digestion instead of physical separation.

.

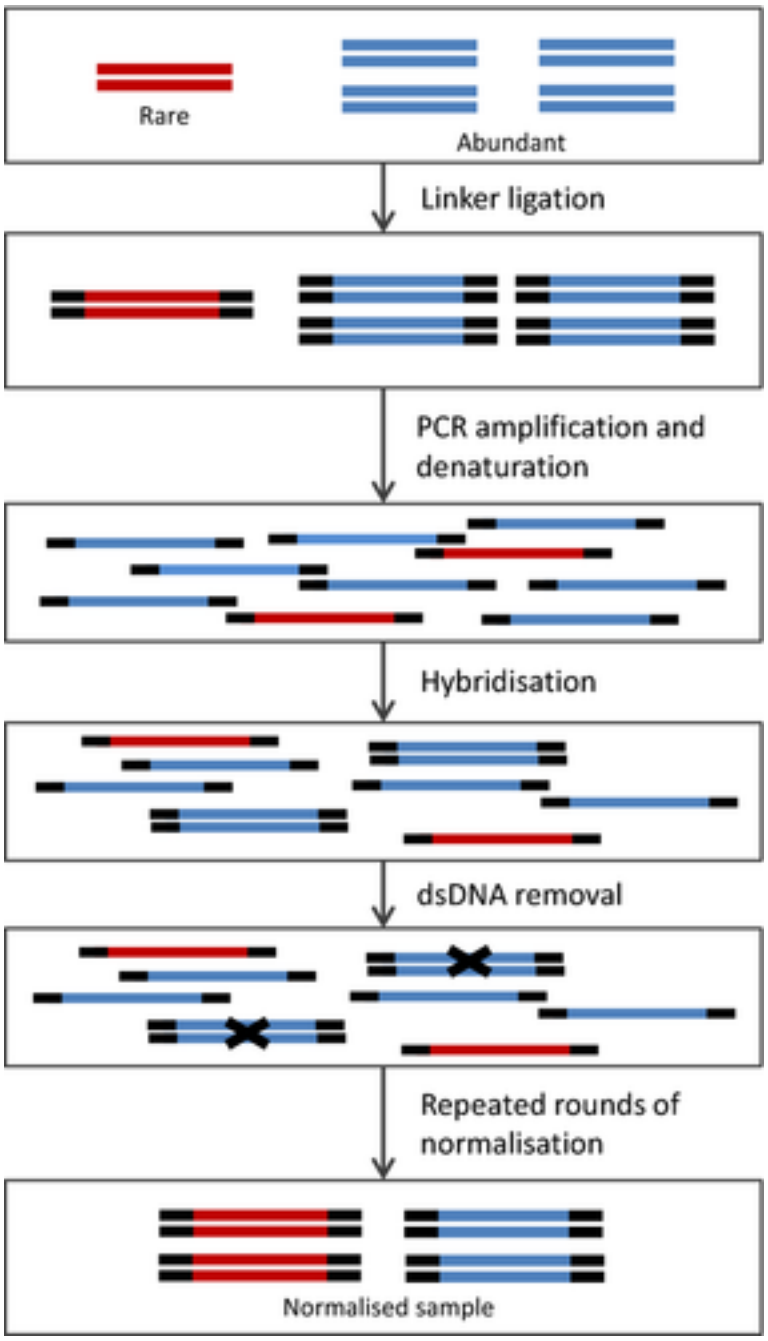

Figure 2. Schematic overview of the DNA normalisation process.

The process starts with a fractionated DNA sample. Lone linkers (black) are ligated to abundant (blue) and rare (red) sequences. The DNA is PCR amplified using primers which bind to the lone linker sequences, the amplicons denatured and then allowed to hybridise under controlled conditions, which allows the most abundant transcripts to form duplexes. The dsDNAs are removed, and the ssDNAs are amplified *via* the linkers. The process repeats for several more rounds to enrich for the rare sequences and yield a normalized sample. Figure obtained with permission from Gagic et al. (2015).

1.4.1. DSN-based Normalization

Duplex-specific nuclease is an enzyme which has been purified from the hepatopancreas of the Kamchatka crab (*Paralithodes camtschaticus)* (Shagin et al., 2002). It is thermostable, shows a strong preference for the digestion of dsDNA and DNA in DNA-RNA duplexes while being inactive against ssDNA and dsRNA (Bogdanova et al., 2008). The enzyme also has the ability to discriminate between perfectly and imperfectly matched DNA duplexes (Shagin et al., 2002). Due to the ability of this enzyme to differ between nucleotides with one nucleotide variation, initially, it was used to create an assay for single nucleotide polymorphism (SNP) typing, which resulted in a more efficient method for the ability to use SNP in diagnostics (Shagin et al., 2002, Shagina et al., 2011). From this, the use of DSN expanded further into genomic studies. DSN has been discovered as a more effective method to normalise cDNA libraries than the previously used HAP chromatography (Bogdanova et al., 2008, Zhulidov et al., 2004, Shagina et al., 2010). The use of this enzyme has been shown to be very effective at reducing the amount of evolutionarily young repetitive sequences with high sequence identity in a range of organisms, such as humans (Shagina et al., 2010). It has also had great success differentiating between closely related pine species and their hybrids, leading to a range of applications in forest seed stock identification which are more cost-efficient than previous methodologies (Cullingham et al., 2013). DSN-based normalisation has also been used successfully to remove repetitive sequences from genomic DNA to facilitate genome sequencing (Yuan et al., 2003).

Recently, DSN was compared against HAP chromatography for its ability to normalise a "mock" metagenome (Gagic et al., 2015). Five different species were distributed at 1000:100:10:10:1 ratio to represent the distribution of taxa in a metagenomic sample. This study showed that HAP was not as efficient as DSN in DNA normalisation as the number of reads from most abundant (1000 and 100 molar ratio) and low abundant genomes had not reached equimolar ratio as expected. Using the HAP method, the proportion of reads for each genome did not significantly differ when compared to the sample before normalisation nor among all five cycles of normalisation. In contrast, DSN normalisation made a marked shift in the representation of the genomes after each round of normalisation.

DSN treatment increased the number of low abundance genomes and decreased the number of high diversity organisms. After five rounds of DSN normalisation, the rare members of the mock metagenome which could not be detected using HAP at the same depth of sequencing were significantly enriched. The rarest member, *Lactococcus lactis* IL1403, increased in abundance of reads from 0.07% to 1.48% after a single round of normalisation. After five rounds this had increased to 18.32% of the mapped sequence reads. The most abundant strain, *Escherichia coli* O127:H6 E2348/69, decreased from 90.05% to 21.76%.

DSN-based normalisation was also more effective in increasing the genome coverage of the microbes used after normalisation. Before normalisation representation of the rare genomes was unable to be reassembled. After 5 rounds of DSN-based normalisation, genes were able to be recovered, increasing to 7.6% for *Pretovella ruminocola* 23 and 13.55% for *L. lactis* IL1403.

The study of this synthetic metagenome has highlighted how much more effective DSN normalisation is in increasing the abundance of reads and enrichment of rare individuals in microbial communities. The significant increase in the abundance of reads and ability to reconstruct the genome of *L. lactis* indicates that the use of DSN for normalisation in future metagenomics studies may be able to reveal sequences of microbial taxa that have not been previously detected, and therefore the use of this method on a natural microbiome may be able to give insight into the rare biosphere of that community.

1.5 Role of HTS sequencing in deciphering “rare biosphere.”

HTS technologies are an important part of microbial ecology studies and have enabled the exploration of microbial communities at an unprecedented scale (Logares et al., 2014). There are three major sequencing approaches that are used in microbial community structure determination. Amplicon sequencing or meta-barcoding, where a phylogenetically informative gene or gene fragment is amplified, their sequence determined and is commonly used to determine the diversity and relative abundance of microorganisms in the sample. Metagenomic sequencing (WGS) reveals the sequences of all genomes present in a sample and can be used for determining the diversity and overall metabolic potential of the microbial community. Finally, metatranscriptome

sequencing (RNA-Seq), where the RNA in a sample is sequenced to determine the expression profile of the microbial community under certain conditions.

Meta-barcoding sequencing at higher sequencing depth (reads per run) is used to determine rare taxa and detect subtle differences in their abundance between different environments or treatments. Each HTS sequencing method has its advantages and disadvantages for use in meta-barcoding sequencing. Until recently Roche 454 pyrosequencing and Illumina were the two main HTS methods for microbiome profiling. Roche 454 pyrosequencing provided a long read length (700 bp) and about 700 000 reads per amplicon run, while Illumina sequencing has a shorter read length (up to 2×300 bp) but a higher number of reads per run (Logares et al., 2014, Margulies et al., 2005, Bennett, 2004). Roche 454 had a higher error rate in regions where one nucleotide is repeated many times because of the way it detects the incorporation of nucleotides (Margulies et al., 2005). In comparison, Illumina sequencing is more accurate but has a smaller read length and a relatively long run time (Luo et al., 2012). As well as suitable HTS sequencing method, amplicon sequencing also requires an appropriate genetic marker to be amplified.

1.6 Genetic Markers

Marker genes which are used in phylogenetic metagenomic studies need to have specific characteristics. The gene used must be ubiquitous, highly conserved, hold enough information for useful analysis, and not be strongly affected by horizontal gene transfer. 16S ribosomal RNA (rRNA) is one of the most commonly used genetic markers as it is conserved across all prokaryotic species (Boughner and Singh, 2016). 16S rRNA is a component of the 30S small subunit of a prokaryotic ribosome that binds to the Shine-Dalgarno sequence found in messenger RNA. It is the gold standard for studies in microbial ecology and has become of significant use in determining phylogenetic relationships, assessing environmental diversity, and for the detection and quantification of communities (Case et al., 2007). However, it does not give an accurate representation of microbial communities due to several limitations.

1.6.1 Limitations of 16S rRNA as a phylogenetic marker

The accuracy of studies using 16S rRNA as a genetic marker is limited by several factors. The most important of these are intragenomic heterogeneity and not having high enough resolution to differentiate between organisms at the species level or lower.

Intragenomic heterogeneity is when a single genome has multiple copies of the 16S rRNA gene, which differ in sequence. The amount of copies of 16S rRNA in a genome has been recorded to vary from a single copy to 15 or more copies, and not all of these copies were identical to each other (Acinas et al., 2004). The study by Case et al. (2007) showed that of 111 genomes, 62% had more than one copy of 16S rRNA with some degree of heterogeneity. For example, *Aeromonas veronii* has up to six copies of the 16S rRNA gene that can differ by 1.5% of all nucleotides among themselves (Janda and Abbott, 2007). Intragenomic heterogeneity becomes a significant factor when the taxonomy levels of family, genus, and species are unable to be determined (Větrovský and Baldrian, 2013). Multiple and variable 16S rRNA copies affect the ability to study the relative abundance by skewing the abundance estimates of individual taxa (Větrovský and Baldrian, 2013). At 97% similarity level, which is the default clustering cut-off to group reads into OTUs (Operation Taxonomic Units), Větrovský and Baldrian (2013) found 21.3% of their OTUs contained sequences of multiple species and 9.3% contained sequences of multiple genera.

The 16S rRNA provides reliable taxonomic classification at higher taxonomic levels, but it has a low resolution at a species level and weak discriminatory power for some genera (Janda and Abbott, 2007). For example, Fox et al. (1992) found two strains of *Bacillus, B. globisporus* and *B. phychrophillus* share 99.5% similarity between their 16S rRNA genes but only show 23-50% affiliation when compared with DNA hybridization. Janda and Abbott (2007) found that three *Edwardsiella* species exhibit 99% similarity between 16S rRNA genes; however, they can be distinguished biochemically and by DNA homology (28-50% relatedness). The genera *Escherichia* and *Shigella* have also been shown to be unable to be differentiated from each other using partial regions of the 16S rRNA gene (Case et al., 2007).

Currently, intragenomic heterogeneity of 16S rRNA as a phylogenetic marker makes it a challenge to define species as a taxonomic level. However, due to the vast number of

partial and full-length 16S rRNA sequences in well-curated 16S rRNA databases, this limitation of 16S rRNA as a phylogenetic marker becomes less pronounced. (Fox et al., 1992, Quast et al., 2013, McDonald et al., 2012b). There are a number of large databases for 16S rRNA including Greengenes, RDP and SILVA as well as resources for specific environments, such as Hungate1000, a collection of rumen genomes (Seshadri et al., 2018, Yilmaz et al., 2014, McDonald et al., 2012b).

1.6.2 RpoB

One way to overcome the limitation of using 16S rRNA in microbial community profiling (1.6.1), is to use other housekeeping genes as alternatives or for use alongside 16S rRNA. One of the most successfully used of these is the *rpoB* gene, which encodes for the β-subunit of DNA dependent RNA polymerase (Mollet et al., 1997, Ogier et al., 2019). It is a required enzyme in the transcription process and the final target for regulatory pathways in controlling gene expression in all living organisms (Borukhov and Nudler, 2003). Like 16S rRNA is it ubiquitous, highly conserved, and large enough to hold enough information to be useful in analysis. It is a monocopy gene, except for in *Norcardia farcinia,* so intragenomic heterogeneity is not of concern (Adékambi et al., 2009, Ishikawa et al., 2004). Phylogenetic trees created from the same populations with 16S rRNA and *rpoB* are consistently in close agreement (Holmes et al., 2004, Case et al., 2007, La Scola et al., 2006, Adékambi and Drancourt, 2004). The *rpoB* gene-sequence similarity between bacterial species also correlates significantly with their DNA-DNA hybridization value, a gold standard method for species determination, indicating that *rpoB* might be a better genetic marker than 16S rRNA gene for defining species (Adékambi et al., 2009).

Protein-encoding genes, such as *rpoB*, have a high genetic resolution. The meta-barcoding analyses using *rpoB* as a phylogenetic marker has shown that microbial relationships from a domain level down to molecular variation at the population level can be determined (Adékambi et al., 2009, Case et al., 2007). When compared with the resolution which can be obtained using 16S rRNA, *rpoB* reveals more species (Vos et al., 2012, Dahllöf et al., 2000, Ogier et al., 2019). For example, as a result of the higher resolution of *rpoB,* species *E. coli* and *E. fergusonii* have been differentiated which was

not the case when the 16S rRNA gene of these two species was used (Adékambi et al., 2009). However, this high genetic resolution also means it cannot be used as a universal marker and can only be used to target a subset of microbial communities (Vos et al., 2012). The marker *rpoB* performs better than 16S rRNA because protein alignments allow for the identification of relationships at the higher taxonomic levels, while nucleotide-level alignments allow for fine-scale identifications at the species level or lower (Case et al., 2007, Adékambi et al., 2009).

There are a few disadvantages with using protein-encoding genes, such as *rpoB,* compared with 16S rRNA. Protein-encoding genes have the saturation of all third codon positions over a long evolutionary time scale, which makes it more difficult to design primers, but primers have been designed and used successfully (Case et al., 2007, Vos et al., 2012, Ogier et al., 2019). The marker *rpoB,* unlike 16S rRNA, does not have a large designated database to obtain sequences from. However, since the completion of the experimental work of this study, a reference database of 45,000 sequences has been constructed (Ogier et al., 2019). For this work sequences had to be obtained from GenBank (www.ncbi.nlm.gov/), the Cluster of Orthologous Groups (COG) database EggNOG (evolutionary genealogy of genes: Non-supervised orthologous groups) or from a genome reference set such as the Hungate collection (Huerta-Cepas et al., 2016, Seshadri et al., 2018).

The use of *rpoB* in combination with 16S rRNA will, therefore, give a more of a complete picture of microbial communities, reducing the number of errors caused by intragenomic heterogeneity and being able to determine the identity of the community down to a species and subspecies level reliably. The use of a protein encoded gene, such as *rpoB,* alongside 16S rRNA will result in a better understanding of microbial diversity and community composition in a range of environmental habitats.

1.7 Hypothesis and Aims

The rumen microbiome is of key interest due to its importance in the agricultural industry and impact on greenhouse gas emissions. The dominant taxa of the rumen microbiome have been studied in detail, but the rare biosphere remains understudied. The recent use of the DSN enzyme to normalise a synthetic metagenome has shown that this method may be the next step for mining the rumen rare biosphere. With the use of DSN-mediated DNA normalisation in combination with meta-barcoding sequence analysis of a traditional genetic marker, 16S rRNA, and the protein-encoding gene, *rpoB,* this study aims to explore the rare biosphere of the rumen microbiome.

We hypothesise that by using DSN normalisation to increase the proportion of low abundance reads in the metagenome before sequencing, that we will be able to explore a higher proportion of the rare biosphere. By utilising the sequencing and analysis of the two phylogenetic markers, *rpoB* and 16S rRNA, the rare biosphere of the rumen microbiome will be able to be identified to a greater level of detail.

2. Materials and Methods

2.1. Materials

2.1.1 Oligonucleotides

The oligonucleotides used in this thesis for PCR and sequencing are listed in **Table 1**.All were synthesized by Integrated DNA Technologies (Iowa, USA).

Table 1: Oligonucleotides used in this study.

Name	Sequences	Notes	Reference
LL-RIA	5'-pGAGATATTAGAATTCTACTC-3'	Lone Linker A	(Ko et al., 1990)
LL-RIB	5'TATAATCTTAAGATGAGp-3'	Lone Linker B,	(Ko et al., 1990)
16F (Ba9F)	5'-GAGTTTGATCMTGGCTCAG-3'	Forward primer for the 16S rRNA marker	This Study
16R	5'-CTATGCGCCTTGCCAGCCCGCTCAGCCGCGGCKGCTGGCAC-3'	Reverse primer for the 16S rRNA marker,	This Study
rpoB1698f	5'5'- AACATCGGTTTGATCAAC-3'	Forward primer for rpoB	(Dahllöf et al., 2000)
rpoB2041r	5'-CGTTGCATGTTGGTACCCAT -3'	Reserve primer for rpoB	(Dahllöf et al., 2000)

2.1.2. Solutions and Buffers

Laboratory solutions and common buffers were prepared as described in Sambrook and Russell (2001) and sterilized by autoclaving at 121°C for 20 minutes. Solutions and buffers were stored at room temperature (RT) unless otherwise stated.

2.1.3 Chemicals, reagents and enzymes

Common chemicals were purchased from Sigma-Aldrich (Missouri, USA), Merck Ltd (New Jersey, USA), and BDH (Pennsylvania, USA). Restriction endonucleases were obtained from Roche Applied Sciences (Basel, Switzerland) and New England Biolabs Inc. (Massachusetts, USA). DNA polymerases were purchased from Invitrogen (California, USA). The DNA Purification Kits were purchased from Roche Applied Sciences and Qiagen (Hilden, Germany). The DSN enzyme was sourced from Evrogen (Moscow, Russia).

2.1.4 Bioinformatics resources and software

The bioinformatics resources and software used in this thesis are listed below in ***Table 2***.

Table 2: Bioinformatics resources and software.

Resource	Application	Source	Reference
Vector NTI® Advance 11.5.3	DNA sequence display and analysis	Life Technologies, USA	(Lu and Moriyama, 2004)
Geneious	DNA sequence display and demultiplexing	Geneious Biologics	(www.geneious.com)
QIIME2 2019.1	Denoising, taxonomic classification and diversity analysis of sequencing data	QIIME 2 Development Team (qiime2.org)	(Bolyen et al., 2018)
Basic Local Alignment Search Tool (BLAST)	Finding Regions of local similarity between sequences	http://blast.ncbi.nlm.nig.gov/BLAST.cgi	(Altschul et al., 1990)
Double Index Alignment of Next-Generation Sequencing Data (DIAMOND)	Aligning sequences against a translated protein database	https://github.com/bbuchfink/diamond	(Buchfink et al., 2015)
EMBOSS: The European Molecular Biology Open Software Suite	*Insilco* digestion of rumen microbiome genomes to determine which restriction enzymes to use	http://emboss.sourceforge.net/	(Rice et al., 2000)
Evolutionary genealogy of genes: Non-supervised Orthologous Groups (EggNOG) 4.5.1	Database of orthologous groups and functional annotations used for rpoB taxonomic classification	http://eggnogdb.embl.de	(Huerta-Cepas et al., 2016)

EggNOG-mapper 4.5.1	Functional annotations of sequences	http://eggnogdb.embl.de	(Huerta-Cepas et al., 2016)
R	Calculation of Diversity Indices	R Core Team http://CRAN.R-project.org	(R Core Team, 2019)
R package Vegan	Calculation of Diversity Indices	https://CRAN.R-project.org/package=vegan	(Oksanen et al., 2019)
SILVA 132	Database of 16S rRNA sequences for all three domains of life. Used for 16S rRNA taxonomic classification	https://www.arb-silva.de/	(Yilmaz et al., 2014, Quast et al., 2013)
Greengenes 13_8	Database of 16S rRNA sequences. Used for 16S rRNA taxonomic classification	http://greengenes.secondgenome.com	(DeSantis et al., 2006, McDonald et al., 2012b)
Hungate1000	A reference set of rumen microbial genomic sequences	*Joint Genome Institute/* IMG portal: https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=HungateCollection.	(Seshadri et al., 2018)
Perl 5.26.1	Used for running scripts		https://www.perl.org/

2.2 Methods

2.2.1 General molecular biology techniques

General molecular biology techniques were implemented as instructed by Sambrook and Russell (2001). The general overview of the methodology of this study can be viewed in **Figure 3**. Gel electrophoresis was run on 1.1% weight/volume (w/v) agarose gels in 1× TAE buffer, with a 1kb ladder (Axygen, New York, USA) and about 200 ng of each sample, at 70 V for 1 hour (h). DNA was stained with ethidium bromide (1µg/mL) for 20 minutes (min) and visualized using a Gel Doc UV illuminator (Bio-Rad, California, USA).

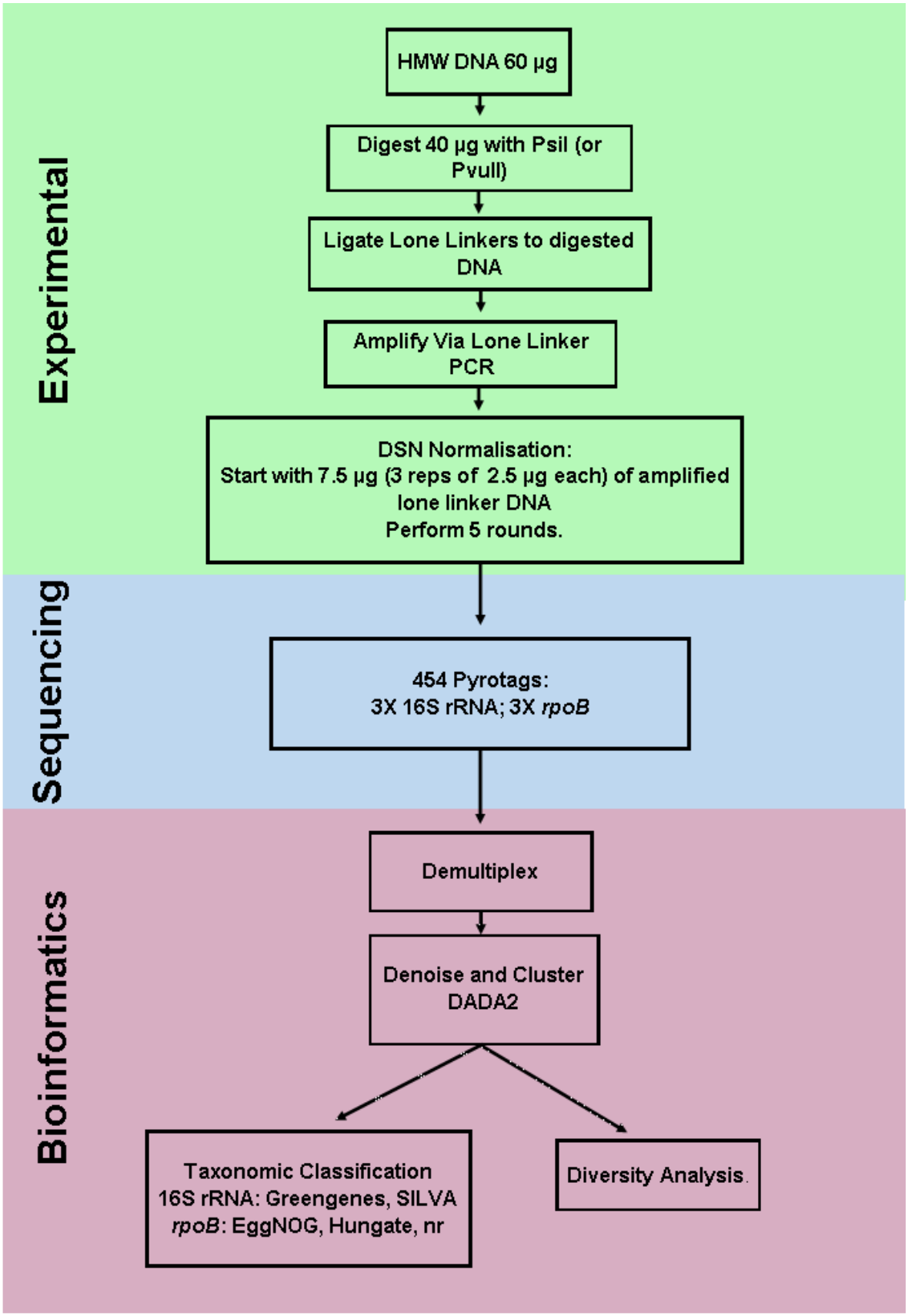

Figure 3: Schematic diagram of the normalisation procedure and NGS sequencing of the rumen microbial metagenome.

2.2.2 Sample Collection

Samples of the whole rumen contents (digesta) from pasture-fed fistulated dairy cows were collected at a DairyNZ research farm (Lye Farm, Waikato, New Zealand) under Animal Ethics permission number AE 11483 granted by the Ruakura Animal Ethics Committee, in May 2009. The rumen plant-adherent microbial fraction and corresponding microbial metagenomic DNA of Animal C (Ciric, 2014) were isolated as previously described (Ciric et al., 2014). The collected digesta was sequezed through a double layer cheesecloth, Plant debris were removed via low speed centrifugation. Plant adherent fraction was obtained by chemically detaching the microbes from the digesta by incubation in a dissociation buffer and centrifugation to remove small plant debris and collect the microbial pellet (Ciric, 2014).

2.2.3 Sample Preparation

2.2.3.1 Digestion of metagenomic DNA and Lone Linker Ligation.

Approximately 60 µg of metagenomic DNA was cleaved with either PsiI or PvuII restriction endonucleases (NEB) overnight at 37°C in a total volume of 50 µL. Digested DNA was purified using a High Pure DNA Purification Kit (Roche). The "lone linker" (LL-RIA, B (**Table 1**) (Ko et al., 1990b)) was ligated to digested DNA (40 µg), in excess of a 100:1 molar ratio using the Rapid DNA Dephosphorylation and Ligation kit (Roche). The lone linker was generated by annealing LL-RIA (**Table 1**), and LL-RIB (**Table 1**) oligonucleotides with the 40 µL digested DNA overnight at RT. Excess linker was removed by washing with 10 volumes of sterile water twice in a microconcentrater (Vivaspin 100, GE Healthcare Biosciences, Sweden), and DNA was resuspended in a final volume of 80 µl of sterile nuclease-free water.

2.2.3.2 Lone Linker Amplification

Using the LL-RIA, as a primer (**Table 1**), the metagenomic DNA was amplified using Polymerase Chain Reaction (PCR). The thermocycling protocol started with an initial

denaturation step for 2 min at 94°C, followed by 25 cycles of a denaturation step at 94°C for 30 seconds (sec), an annealing step at 54°C for 30 sec and an extension step at 68°C for 4 min. A final extension step of 7 min ensured all products were fully amplified. PCR fragments were purified by phenol-chloroform-isoamyl alcohol (25:24:1) method (Sambrook and Russell, 2001), precipitated with 3M NaOAc (pH 5.2) and ice-cold ethanol, and the resulting pellet resuspended in 500 µl 1× TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0).

2.2.4. 16S rRNA and *rpoB* phylogenetic markers amplification

Two phylogenetic markers, 16S rDNA and *rpoB* (Woese and Fox, 1977, Mollet et al., 1997) were chosen to evaluate DSN normalisation of the rumen microbial metagenomic DNA. 16S rDNA amplicons are generated using 16S F and 16S R (**Table 1**) primers and DNA polymerase (5× HOT FIREPol Blend Master Mix (Solis BioDye, Tartu, Estonia)). Thermocycling conditions used were 35 cycles of a denaturation step at 95°C for 30 sec, an annealing step at 53°C for 30 sec and an extension step at 72°C for 1 min. The final extension step was extended to 5 min to ensure a complete synthesis of all the products.

The *rpoB* PCR amplicons were generated as previously described: Sequences from *rpoB* from four bacteria where compared and two regions which were conserved were used to construct primers, of those primers the ones that gave PCR products for the 10 type strains where constructed as used for all bacteria (Dahllöf et al., 2000).

For *rpoB* locus amplification oligonucleotides rpoB1698f, rpoB2041r (**Table 1**) and High-Fidelity Platinum Taq polymerase (Invitrogen) were used. Thermocycling conditions used were: 9 cycles of a denaturation at 94°C for 30 sec, an annealing step at 40°C for 30 sec and an extension step at 68°C for 90 sec followed by 19 cycles with denaturation at 94°C for 30 sec, an annealing step at 50°C for 90 sec and an extension step at 68°C for 90 sec.

2.2.5 DSN based normalisation

2.5 µg of PsiI digested, and lone linker amplified DNA was normalised using modified DSN normalisation method (Gagic et al., 2015). The same procedure and amounts of DNA were used for the PvuII digested, and lone linker amplified DNA. Briefly, DNA was first denatured at 98°C for 3 min, followed by hybridisation at 68°C for 5 (h). DSN normalisation was performed in triplicate for each sample. Prewarmed (68°C) DSN Master buffer and 0.125 Units (U) of DSN (Evrogen) was added to the hybridised DNA and incubated 20 min at 65°C. The reaction was stopped by the addition of EDTA (pH 8.0) to a final concentration of 2.25 mM. Samples were incubated for 10 min and subsequently normalised DNA (5 µl) was used for PCR amplification with LL-RIA primer and Platinum PCR Supermix HiFidelity (Invitrogen). Thermocycling conditions used were 25 cycles of a denaturation step at 94°C for 30 sec, annealing step at 55°C for 30 sec and an extension step at 68°C for 4 min. In total 16 PCR reactions per sample (50 µl) were generated, pooled and amplicon DNA was purified by phenol-chloroform and subsequently precipitated with 3M NaOAc (pH 5.2) and 2.5 volumes (vol) of ice-cold 100% ethanol. The resulting pellet was then resuspended in 100 µl 1× TE Buffer. Before amplification DNA was used as a starting material for the next round of DNA normalisation, PCR product fragment distribution was visualised by agarose gel electrophoresis. In total, five rounds of DSN-based DNA normalisation were performed.

2.2.6 Metagenome Sequencing

The original uncut metagenomic DNA (uncut metDNA), the lone-linker amplified PsiI and PvuII metagenomic DNA samples (R0) and the PCR amplicons created over 5 rounds of DSN normalisation (R1-R5) were barcoded using a Nextera XT DNA sample preparation kit (Illumina Inc., San Diego, USA). Barcoded sequences were sequenced using Titanium Chemistry on a 454 GS FLX instrument (Roche Applied Science, Germany). Sequencing was carried out by Macrogen Inc. sequencing facility (Seoul, Korea).

2.2.7 Bioinformatics Methods

The raw 454 sequence reads were demultiplexed using Geneious R.8.9.1 and denoised using QIIME2-2019.1 (Quantitative Insights Into Molecular Ecology) (Bolyen et al., 2018) with DADA2 package (Callahan et al., 2016). OTUs were clustered by VSEARCH open reference clustering (Rideout et al., 2014) and chimaeras filtered by VSEARCH UCHIME DENOVO (Rognes et al., 2016).

Amplicon Sequence Variants (ASVs) were aligned with mafft (Katoh et al., 2002) and used to construct a phylogeny with mafft fasttree (via q2 phylogeny).

2.2.7.1 Taxonomic Classification using 16S rRNA region

Taxonomy was assigned to the Amplicon Sequence Variants (ASVs) using the *q2-feature-classifier* (Bokulich et al., 2018). Classification against the SILVA-132 90%, 94% and 99% OTU reference sequences (Quast et al., 2013, Yilmaz et al., 2014) used *classify consensus-blast* (Camacho et al., 2009) and classification against the Greengenes 13_8 99% OTUs used *classify-sklearn* naïve Bayes taxonomy classifier (McDonald et al., 2012a).

NCBI classification of the 16S rRNA region was achieved by using the comparing the Amplicon Sequence variance against the 16S Microbial Database (https://ftp.ncbi.nlm.nih.gov/blast/db/). The was done using the BLAST+® Command Line application using the function *blastn*, with the *max_target_seqs* set to 1. Matches with an Evalue of less than E^{-10} were removed.

2.2.7.2 Taxonomic Classification of *rpoB sequences*

The *rpoB* ASVs and Biom table were exported from QIIME2. The sequences were compared against the EggNOG 4.5.1 database (Huerta-Cepas et al., 2016) by using the EggNOG-mapper (Huerta-Cepas et al., 2017). The mapping mode was set to DIAMOND, the taxonomic scope set to adjust automatically and to search through all

orthologs using non-electronic gene ontology evidence terms to prioritize coverage. The corresponding matches to the sequences were filtered by removing matches which matched to proteins other than rpoB and had an E-value of more than E^{-10}.

The *rpoB* sequences were also taxonomically classified against Hungate1000 (Seshadri et al., 2018), which is a reference set of 410 rumen microbial genome sequences. Using DIAMOND (Buchfink et al., 2015) the database of the amino acids sequences of the Hungate1000 database was created, and the sequences were compared to the database using the *blastx* function. The alignment was set to sensitive, maximum target sequences set to 2 and an output format of BLAST tabular format was used. The corresponding matches were filtered by matching the ASV ID to the taxonomic output with the highest Evalue and removing all matches with an Evalue higher than E^{-10}. This same process was also used for taxonomic classification against the NCBI nr database.

2.2.8. Statistical Analysis

Alpha diversity metrics: observed OTUs and Faith's Phylogenetic Diversity (Faith, 1992), beta diversity metrics, weighted uniFrac (Lozupone et al., 2007), unweighted uniFrac (Lozupone and Knight, 2005), Jaccard distance, and Bray Curtis dissimilarity, and Principal Coordinate Analysis (PCoA) were estimated using q2-diversity, after samples were rarefied to 19685 sequences (16S rRNA) and 60000 (*rpoB*) sequences per sample using QIIME2. Shannon's Diversity Index for each round of normalisation was calculated in R 3.5.3 using the vegan package (Oksanen et al., 2019). The difference between the proportion of reads between rounds was tested using a Z 2-proportion test at 95% significance. The difference between taxa identified in *rpoB* was tested using a dependant sample t test.

3. Results

3.1 DSN-based normalisation of the rumen microbial metagenomic DNA

For normalisation, in particular for restriction digestion step or fragmentation, the DNA needs to be of high quality and high molecular weight (HMW). The HMW DNA (**Figure 4**, Lane 1) from the rumen microbiome was isolated from the pasture-fed cattle in, the Rumen Microbiology laboratory, AgResearch (NZ) in 2015. (Ciric, 2014). It was recognised in a previous study (Gagic et al. 2015) that the restriction digestion by a single enzyme affects both the downstream functional bioactivity screening (since numerous genes will be cleaved) and limits any metagenome sequencing assembly as the restricted fragments do not overlap. Therefore, in this study, one of the aims was to compare the performance of two different restriction enzymes (RE) in microbial community profiling after the normalisation of metagenomic DNA. Both RE enzymes which were used give blunt-end fragments of an average size of 4 kb (**Figure 4**, Lanes 2 and 3). The RE enzymes were chosen based on *in silico* digestion profiles of the most dominant taxa in the Hungate 1000 rumen microbiome collection using the restriction function in EMBOSS suite version 6.6.0.0 (Rice et al., 2000) (*R. flavefaciens, B. proteoclasticus, Prevotella ruminocula* 23; data not shown). This analysis showed that PsiI and PvuII do not cut through 16S rRNA and *rpoB* genes in the analysed genomes and gave the highest proportion of fragments in the size range 1-5kb.

To permit the amplification of all the digested DNA fragments lone linkers (Ko et al., 1990) were ligated to the DNA ends. The digested sample was amplified using a single lone linker (LL-RIA) to obtain the necessary amount of starting sample for normalisation procedure (Round 0).

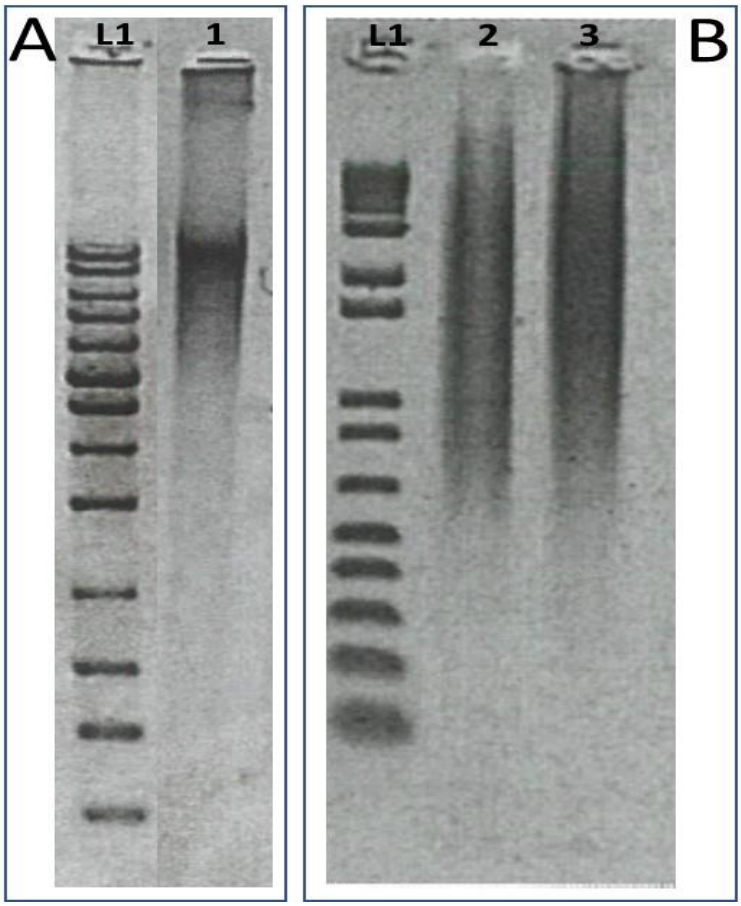

Figure 4: Preparation of the metagenomic DNA for normalisation.

L1: 1 kbp ladder; A) The rumen microbial HMW DNA (1); B) HMW metagenomic DNA with digested with PvuII (2) and PsiI (3)

.

The DNA from Round 0 samples was purified, suspended in hybridization buffer and then denatured in a boiling water bath. The DNA was stringently hybridised by incubation at 68°C (Short and Mathur, 1999). At this point, the dsDNA in the sample had been eliminated by DSN digestion. The resulting ssDNA were amplified by PCR *via* the linker sequence. This process was repeated for five cycles (**Figure 2**) removing a sample after each round, to obtain templates for 16S rRNA and *rpoB* amplification.

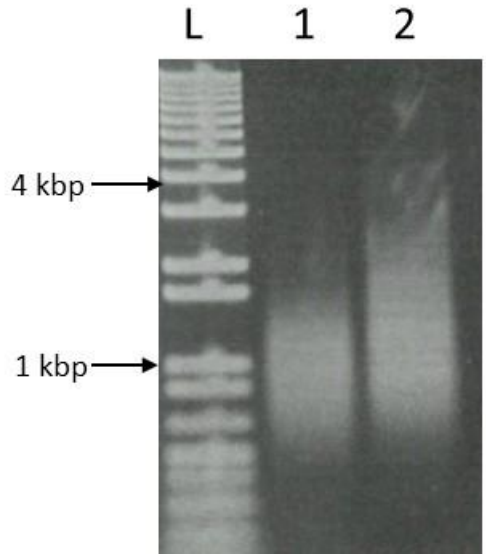

Figure 5: Normalised metagenomic DNA after 5 rounds of normalisation.

L: 1 Kb ladder. 1: PsiI digested, lone linker amplified DNA after 5 rounds of normalisation. 2: PvuII-digested DNA, lone linker amplified after 5 rounds of normalisation.

The oligonucleotides specific for 16S rRNA V1-V3 region and *rpoB* genes (Table 1) were used for amplification of phylogenetic markers after each round of normalisation (**Figure 6**; AR1-5). Round 0 samples (**Figure 6** *rpoB* L7, 16S L7) and undigested metagenomic DNA (**Figure 6**; *rpoB* L1, 16S L1) have fragments lengths of ~ 550 bp and ~380 bp for 16S rRNA V1-V3 region and *rpoB*, respectively as expected. These were generated using barcoded primers, purified, pooled and sent for pyrosequencing.

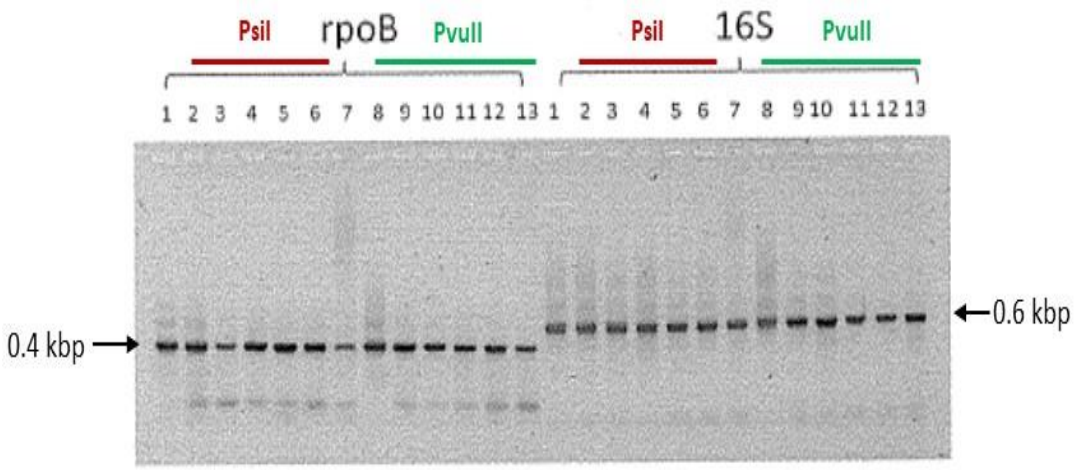

Figure 6: 16S rRNA and *rpoB* amplicons generated from non-normalised and normalised DNA.

Expected sizes of fragments are shown by arrows. The *rpoB* and 16S rRNA labels represented by the 13 samples each denote amplicons generated with specific primers for those phylogenetic markers: L1, uncut metagenomic DNA; L2, PsiI-digested non-normalised (AR0); L3-L7, PsiI-digested DNA normalised from round 1 (AR1, L3) to round 5 (AR5, L7); L8, PvuII-digested DNA non-normalised (AR0); PvuII-digested DNA normalised from round 1 (AR1, L9) to round 5 (AR5, L13).

3.2 Sequencing results and initial analysis

Sequencing generated a total of 1,227,690 16S rRNA and 1,547,923 *rpoB* sequence reads. After demultiplexing, 73,310 sequences were removed from the 16S rRNA reads, and 44,820 were removed from the *rpoB* reads as unable to be mapped to barcoding regions. The number of sequences after demultiplexing for each sample is presented in **Table 3**. After de-noising and chimera removal there were 564,032 sequences made up of 4,071 OTUs for 16S rRNA and 1,278,845 sequences made up of 3,125 OTUs for *rpoB*.

Table 3: Number of sequences after demultiplexing

Sample name	**16S rRNA**	***rpoB***
Uncut metagenomic DNA	98881	111889
PsiI AR0[1]	106011	111481
PsiI AR1[2]	97048	93066
PsiI AR2	66586	153751
PsiI AR3	100455	143041
PsiI AR4	108081	131304
PsiI AR5	95566	79034
PvuII AR0	96455	74806
PvuII AR1	98961	139730
PvuII AR2	111274	127937
PvuII AR3	60740	119734
PvuII AR4	73094	105158
PvuII AR5	41228	112172

[1] AR0, represents the sample which has not been subjected to normalisation; [2] AR(numeral), represents round of normalisation with DSN.

3.2.1 The rumen "rare biosphere" based on DSN normalisation 16S rRNA amplicons

The denoised 16S rRNA sequences were compared against the SILVA (Yilmaz et al., 2014) and Greengenes (McDonald et al., 2012b) databases. Greengenes taxonomic

classification identified 2,104 OTUs, all belonging to the Kingdom Bacteria, the rest were discarded as only bacterial databases were used in this study. It identified 20 different phyla, 34 classes, 46 orders, 72 families, 103 genera, and 45 species in these 2,104 OTUs. SILVA taxonomic classification at 94% identified 4,009 OTUs as Bacteria. In these 4,009 OTUs 19 phyla, 30 classes, 50 orders, 75 families, 118 genera, and 4 species were identified. When compared against the NCBI 16S rRNA database, 1932 OTUs were classified. These OTUs consisted of 275 different genera, 129 families, 69 orders, 41 classes, and 19 phyla. All taxonomic classifications belonged to the Kingdom Bacteria.

The taxonomic profile created with Greengenes taxonomic classification shows the effect DSN normalisation has on the proportion of genera (**Figure 7**). The taxonomic profile created by SILVA classification can be seen in **Appendix 1**. As shown in **Figure 7**, *Ruminococcus, Pseudobutyrivibrio* and *Succiniclastium* are the dominant genera in uncut metagenomic DNA. This remains similar in AR0 for PsiI and PvuII with *Rumminococcus* remaining the dominant genera. However, in PsiI digested samples, *Anaerolinaceae* SHD-231 was enriched, and in PvuII digested samples *Shuttleworthia* and *Butyrivibrio* were enriched. After AR1, the dominance of the taxa is beginning to change (as seen in **Figure 7**, darker coloured bands start to be seen in a tightly packed band at the base of the graph). By AR5, *Tissierellaceae* WAL 1855D is dominant in both PsiI and PvuII, and there is a large proportion of the taxa that were not seen in AR0.

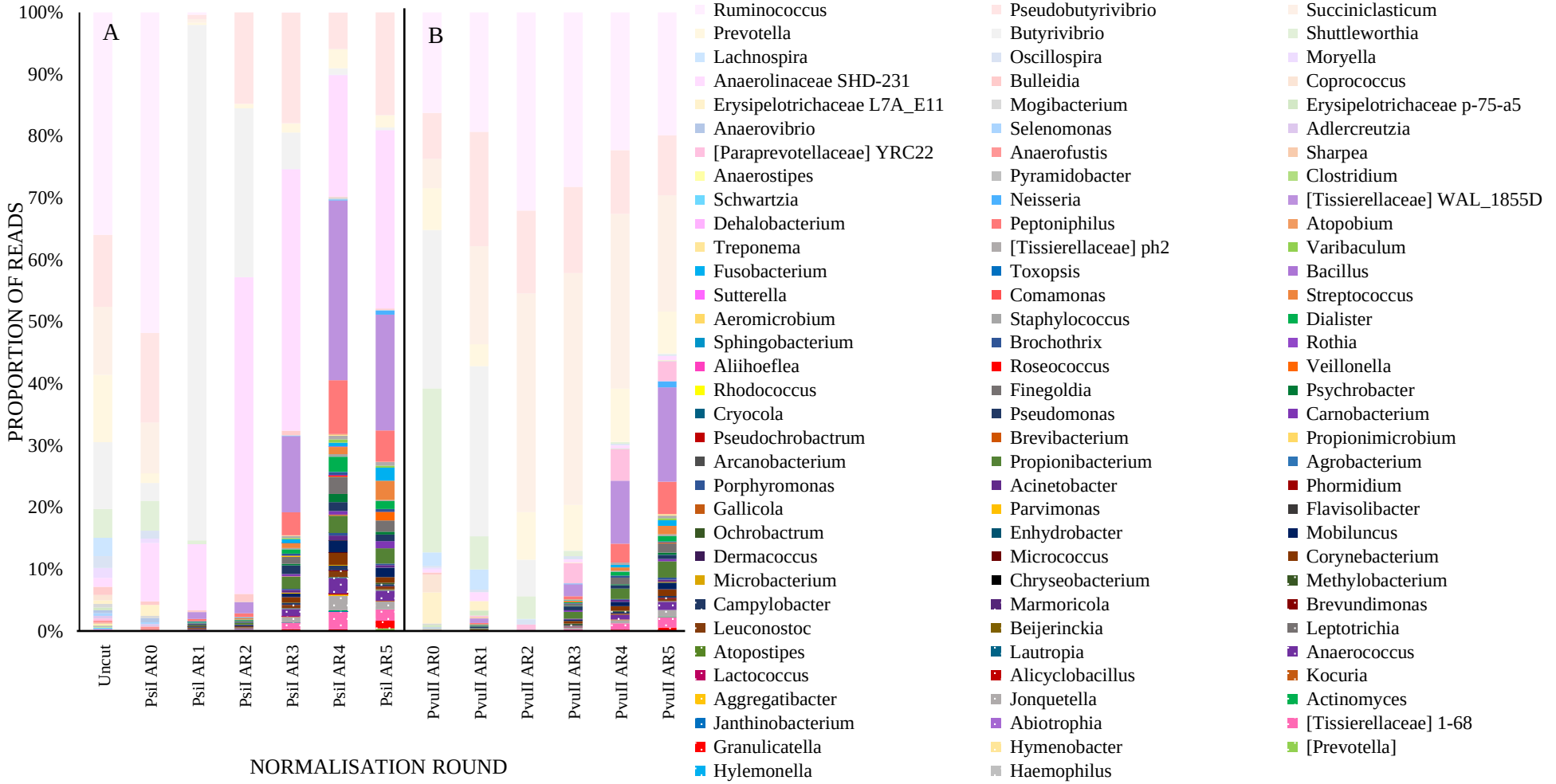

Figure 7: Taxonomic profile of the distribution of 16S rRNA taxonomy at genus level.

Samples of the sum of all matches against the 16S rRNA Greengenes database, grouped at the genus level. Panel A: uncut metagenomic DNA and PsiI digested DNA after normalisation rounds 0-5 (AR). Panel B: PvuII digested DNA AR0-5). The proportion of the genus in each sample is represented by the size of the bar. Colours represent different genera as indicated in the key above.

3.2.2 The rumen "rare biosphere" based on DSN-normalisation of *rpoB* amplicons

Denoised *rpoB* sequences were compared against the EggNOG database (Huerta-Cepas et al., 2016) and Hungate1000 collection (Seshadri et al., 2018). 2979 OTUs were classified against the EggNOG database. These OTUs were classified as belonging to 131 different taxa, made up of 110 species, 71 genera, and 49 different families. Two were unable to be classified at the family level, unclassified Clostridiales and unclassified Rhodospirillales. The Hungate collection classified the sequences into 2,993 OTUs. These were identified as belonging to 136 different taxa, made up of 75 different species and 51 different genera. Six were unable to be classified at a genus level, unclassified Bacteroidales, unclassified Clostridiales, unclassified Erysipelotrichaceae, unclassified Lachnospiraceae, unclassified Porphyromonadacea and unclassified Ruminococcaceae. When compared against the NCBI nr database, 2,980 OTUs were able to be classified as belonging to 102 Genera, 42 families, 19 orders, 11 classes and 6 phyla. All classifications belonged to the Kingdom Bacteria.

The EggNOG database taxonomic profiling (**Figure 8**; see **Appendix 1** for the Hungate taxonomic profile) showed the dominant genera in uncut metagenomic DNA were *Clostridium, Pseudoflavonifractor* and *Coprococcus*. The dominant genera change slightly for AR0; with the PsiI-digested sample selecting *for Lachnoclostrium, Blatuia* and *Anaerobutryicum*, while the PvuII-digested sample is selecting for *Blatuia, Marvinbryantia* and *Butyrivibrio*. After AR1, the dominance of *Clostridium* and *Pseudoflavonifractor*, decreases and less dominant genera such as unclassified *Ruminococcaeae, Haemphillis* and *Granulicatella* are beginning to increase. After AR5, some of the dominant genera such as *Lachnoclostridium, Roseburia* and *Coprococcus* are still very dominant, but there is an increase of the genera that were not seen in uncut metagenomic DNA or in AR0.

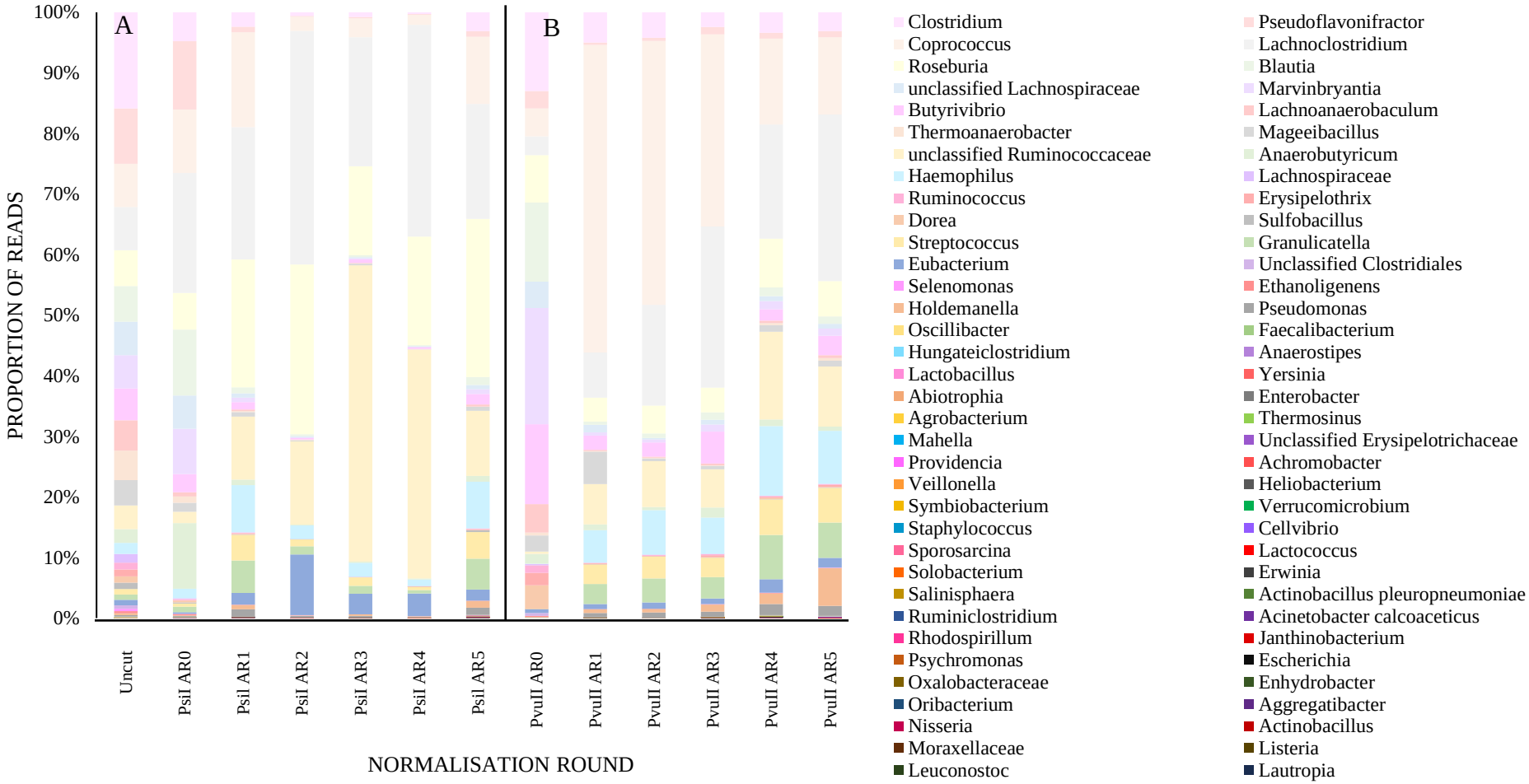

Figure 8: Taxonomic profile of the distribution of *rpoB* taxonomy at genus level.

Samples of the sum of all matches against the EggNOG database, grouped at the genus level. Panel A: uncut metagenomic DNA and PsiI digested DNA after normalisation rounds 0-5 (AR). Panel B: PvuII digested DNA AR0-5. The proportion of the genus in each sample is represented by the size of the bar. The darker the colour the rarer the genus in the uncut metagenomic DNA. Colours represent different genera as indicated in the key above.

3.3 Analysis of DSN-based Normalisation

The similarity of the samples for each round of normalisation was compared using a Principal Coordinate Analysis (PCoA) for both *rpoB,* and 16S rRNA reads (**Figure 9**). As expected, the uncut metagenomic DNA and AR0 for both restriction enzymes are clustered together, near After AR1, the data points move further apart from the uncut and unnormalized samples. For 16S rRNA AR1 reads were not clustered with any of the other data points, lying halfway between the cluster of points before normalisation (AR0) and after (AR2-5). In *rpoB,* the AR1 points can be seen at the top of the clusters on the far side. AR2, AR3, AR4 and AR5 all cluster together. For 16S rRNA, the PsiI cluster at the top of the graph near Axis 2 and PvuII cluster at the top of the graph on the opposite side away from Axis 2. The *rpoB* PsiI normalised points cluster at the bottom of the graph close to Axis 1, while the PvuII points cluster in the middle of the graph, dropping slightly for the final two rounds of normalisation. 16S rRNA and *rpoB* normalisation both show uncut and AR0 DNA clustering together and AR1-AR5 clustering together.

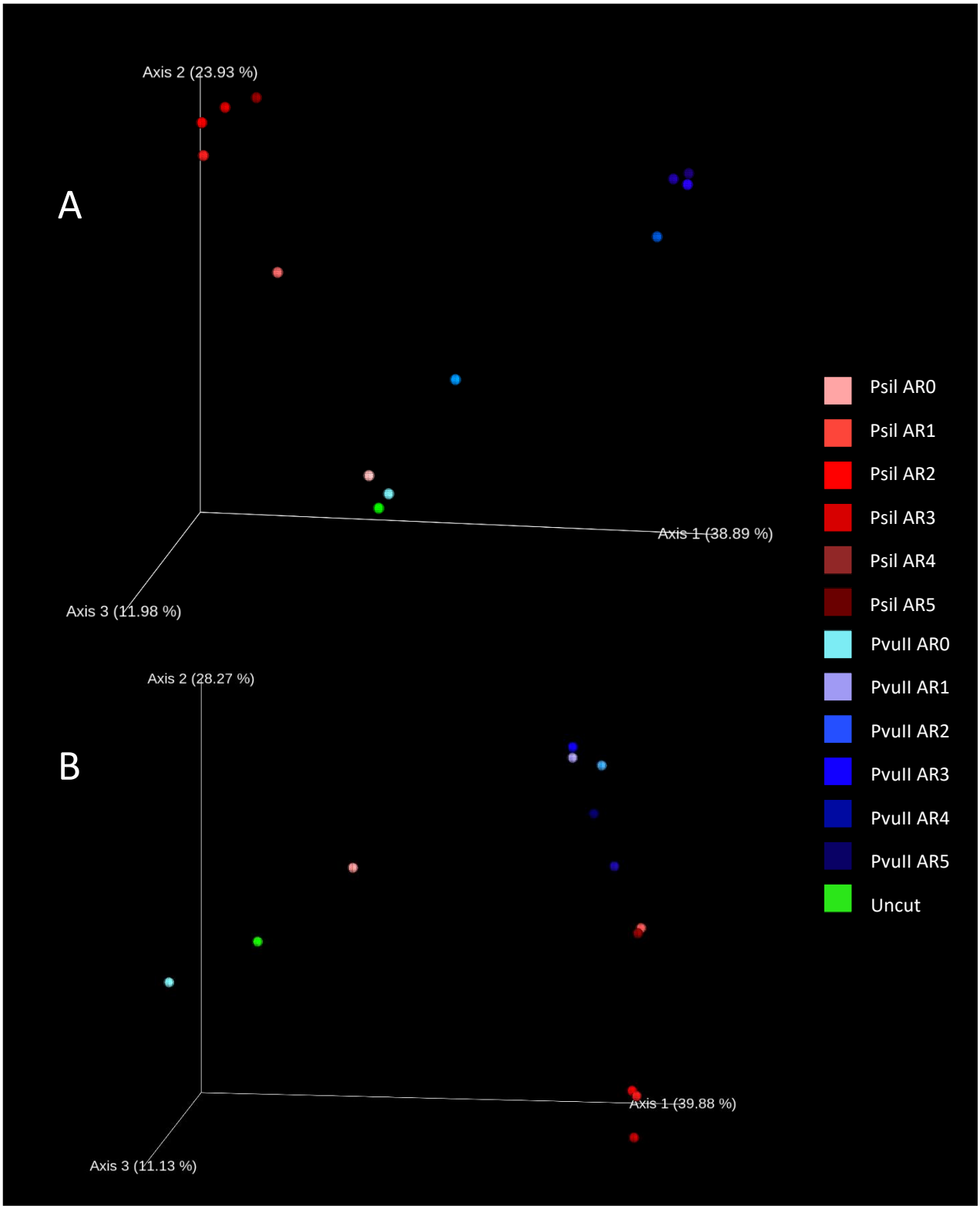

Figure 9: Bray Curtis dissimilarity index PCoA

Principal Coordinate Analysis (PCoA) of the similarities between rounds of normalisation for 16S rRNA (A) and *rpoB* (B). The similarity between each round of normalisation (AR0-AR5), as seen by shade of colour (light to dark with increasing rounds) and the restriction enzymes PsiI, (Red) and PvuII (Blue) and uncut metagenomic DNA (Green) by different colours.

3.4 Normalisation effect on taxa distribution

The diversity of the metagenome for each round of normalisation and the genetic markers were tested using Shannon's Diversity Index, and the results are illustrated in **Table 4**. There is an insignificant drop in diversity for both 16S rRNA and *rpoB* between uncut metagenomic DNA and DNA, which has only been digested by a restriction enzyme (AR0). Normalisation does not decrease the amount of diversity in the metagenome as the Shannon's Index for each round of normalisation remains around 2 for all rounds of normalisation except for in 16S rRNA PvuII AR1 and AR2 and *rpoB* PsiI AR2, AR3 and AR4.

Table 4: Shannon's Diversity Index for each round of normalisation

	16S rRNA	***rpoB***
Uncut metagenomic DNA	2.415959	2.37221
PsiI AR0	2.109723	2.063445
PsiI AR1	2.260834	1.909902
PsiI AR2	2.357199	1.542141
PsiI AR3	2.595157	1.503027
PsiI AR4	2.728173	1.434688
PsiI AR5	2.845236	1.942998
PvuII AR0	2.368749	2.133203
PvuII AR1	1.116522	1.712001
PvuII AR2	1.670931	1.764996
PvuII AR3	2.220588	1.891345
PvuII AR4	2.742622	2.04511
PvuII AR5	2.300844	1.993508

DSN-based normalisation decreased the abundance of the dominant taxa and increased the abundance of rarer taxa (**Table 5**). Genera, which are represented with less than 0.001% of the identified reads, were unable to be detected. The genera *Butyrivibrio* and *Clostridium* decreased in the proportion of reads consistently across both of the restriction enzymes and genetic markers used. The genera *Granulicatella, Streptococcus, Leuconostoc,* and *Veillonella* all consistently increased in the proportion of reads detected after AR5. The other genera (**Table 5**) had a different response to normalisation depending on the marker or restriction enzyme used. *Pretovella, Shuttleworthia, Succinclastiucum, Anaerolinaceae SDH-231, Tissierellaceae WAL_1855D,* and *Peptoniphulis* were not detected in reads where *rpoB* was used as a genetic marker. *Blautia, Marvinbryantia, Holdmanella* and *Hungateiclostridium* were only detected in reads using 16S rRNA as a genetic marker. The response to normalisation also differed depending on the restriction enzyme used to cut the metagenomic DNA before

sequencing. *Ruminococcus, Succinoclasticum* increased the proportion of reads when digested with PsiI but decreased when digested with PvuII.

Table 5: Dominant and emerging genera before and after five rounds of normalisation.

Genus	PsiI				PvuII			
	16S rRNA		*rpoB*		16S rRNA		*rpoB*	
	AR0	AR5	AR0	AR5	AR0	AR5	AR0	AR5
Butyrivibrio	26%	0.1%	3.0%	1.7%	13.2%	3.2%	2.9%	0.4%
Ruminococcus	16%	20%	0.2%	0.09%	1.2%	0.2%	52%	<0.001%
Prevotella	6.8%	6.8%	<0.001%	<0.001%	1.6%	1.9%	<0.001%	<0.001%
Shuttleworthia	26%	0.09%	<0.001%	<0.001%	4.8%	<0.001%	<0.001%	<0.001%
Succinclasticum	4.7%	19%	<0.001%	<0.001%	8.2%	0.08%	<0.001%	<0.001%
Blautia	<0.001%	<0.001%	11%	1.3%	<0.001%	<0.001%	13%	1.3%
Marvinbryantia	<0.001%	<0.001%	7.5%	0.8%	<0.001%	<0.001%	19%	1.2%
Coprococcus	2.9%	<0.001%	10%	11%	<0.001%	<0.001%	4.6%	12%
Clostridium	0.05%	<0.001%	4.8%	3.1%	0.05%	<0.001%	13%	3.1%
Haemophillus	<0.001%	<0.001%	1.6%	7.7%	<0.001%	0.1%	0.03%	8.8%
Granulicatella	<0.001%	0.4%	0.9%	5.1%	<0.001%	5.8%	<0.001%	1.2%
Streptococcus	<0.001%	1.3%	0.6%	4.4%	0.01%	5.6%	<0.001%	3.1%
Leuconostoc	<0.001%	0.3%	<0.001%	0.004%	<0.001%	0.04%	<0.001%	0.3%
Veillonella	<0.001%	0.14%	0.01%	0.06%	<0.001%	1.4%	<0.001%	0.05%
Holdmanella	<0.001%	<0.001%	0.09%	1.1%	<0.001%	<0.001%	0.06%	6.2%
Hungateiclostridium	<0.001%	<0.001%	0.06%	0.04%	<0.001%	<0.001%	0.01%	0.06%
Anaerolinaceae SHD-231	0.5%	0.7%	<0.001%	<0.001%	9.5%	29%	<0.001%	<0.001%
Tissierellaceae WAL_1855D	0.1%	15%	<0.001%	<0.001%	<0.001%	19%	<0.001%	<0.001%
Peptoniphulis	<0.001%	5.2%	<0.001%	<0.001%	<0.001%	5.1%	<0.001%	<0.001%

3.5 Change in the dominant species

To compare the taxonomic composition, at the genus level, between OTUs identified from 16S rRNA and *rpoB* genera, they were both searched against the NCBI nr and 16S rRNA microbial databases. The top ten dominant genera in 16S rRNA AR0 mostly decreased in their abundance after normalisation (**Figure 10**). There is a statistically significant difference between the proportion of total reads in AR0 and after AR5 for all the top ten dominant species (P-value <0.05, Z 2 proportion test). There is also a statistically significant between AR0 and AR3 for all genera apart from *Ruminococcus.* The top seven most dominant genera (with the highest proportion of total reads) in AR0 all decreased in dominance after normalisation. *Butyricicoccus* decreased in its proportion after AR3 from 2.6% to 1.9%, but after AR5, it increased to a higher proportion than in AR0 of 4.1%. *Pseudobutyrivibrio* increased in the proportion of reads between AR0 and AR3 by 0.4% but then decreased in the proportion of reads between AR3 and AR5 by 0.7%. *Prevotella* increased in proportion between AR1 and AR3 from 2.5% to 3.5% then decreased after AR5 to 3%.

The top ten genera based on *rpoB* taxonomy had shown more variation during normalisation than for 16S rRNA based taxonomy (**Figure 10**). There was a significant difference in the proportion of reads between AR0 and AR3 and between AR0 and AR5 for all ten taxa (Z 2 proportion test P<0.05). The proportion of total reads for unclassified Firmicutes, unclassified Lachnospiraceae, *Pseudobutryivibrio, Eubacterium, Butyrivibrio and Roseburia* all decreased in the proportion of reads after normalisation. *Clostridium* and unclassified Clostridiales increased in the proportion of total reads between AR1 and AR3, from 7.6% to 15% and from 1.8% to 20% respectively. *Merdimonas* decreased in the proportion of totals reads after AR3 and AR4 down to 1.4% but increased after AR5 to 9%. The proportion change in the ten most dominant genera shows normalisation appears to continue to either increase or decrease for the first four rounds of normalisation and then after AR5 the trend varies depending on the genus. This trend can be seen in all of these genera except for *Clostridium*, which continues to increase in the proportion of total reads throughout all rounds of normalisation. The AR5 was an arbitrary stop of the normalisation process and it could be expected that after at least four rounds of PCR, the amplification errors would start to contribute to a number of false-positive OTUs and therefore microbial diversity.

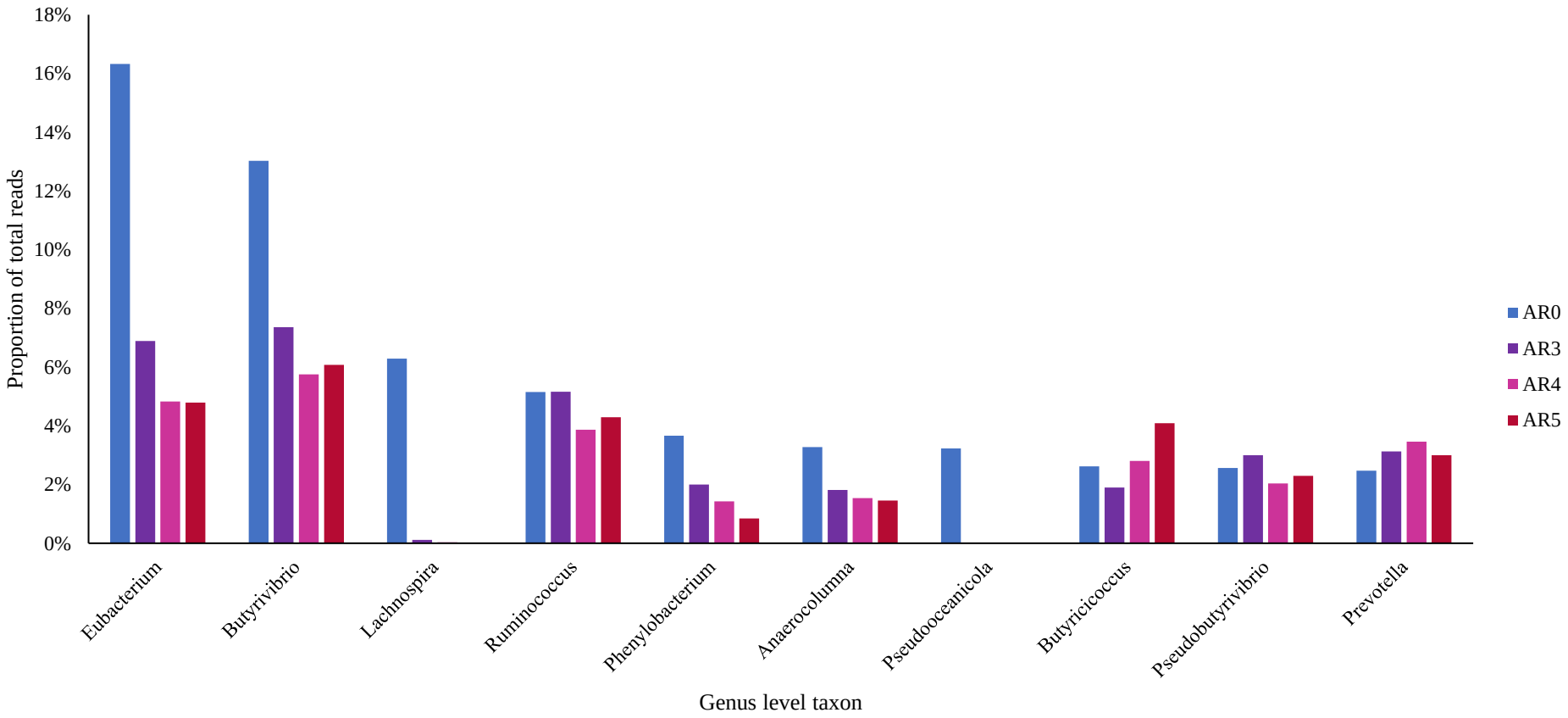

Figure 10: The proportion change of the top ten most dominant genera for 16S rRNA.

The ten most dominant genera before normalisation (Blue) compared with their proportion after three (purple), four (pink) and five (red) rounds of normalisation when mapped against the NCBI 16S rRNA microbial database.

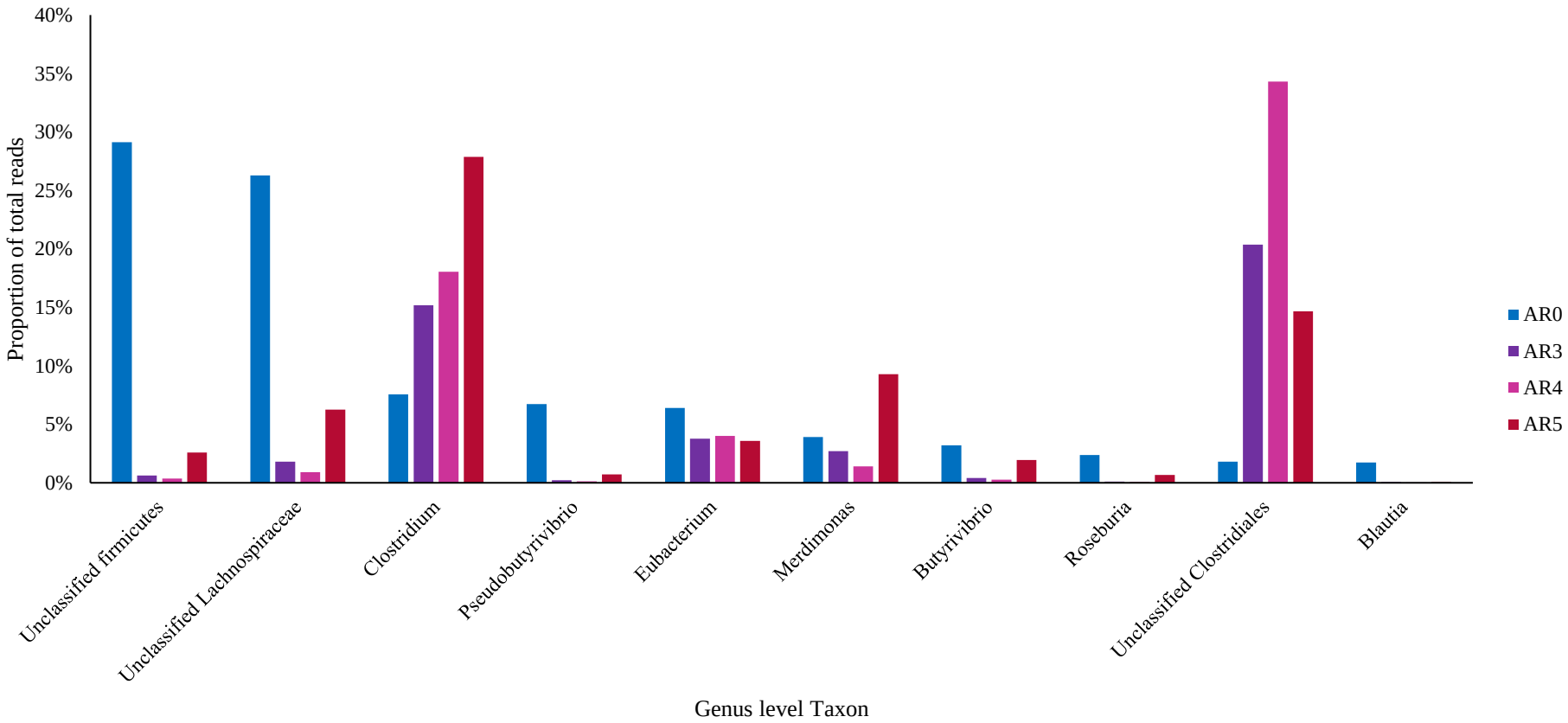

Figure 11: The proportion change of the top ten most dominant genera for *rpoB*.

The ten most dominant genera before normalisation (blue) compared with their proportion after three (purple), four (pink), and five (red) rounds of normalisation for *rpoB* reads when mapped against the NCBI nr database.

3.6 Increase in Rare Taxa

DSN normalisation increases the number of rare species. This can be seen as the general increase in the number of taxa which are unable to be classified down to the genus level, except for unclassified Lachnospiraceae and unknown Ruminococcaceae (**Appendix 2**, **Figure S1**), and the number of taxa which were unable to be detected in the sequencing reads before normalisation but were detectable after normalisation. The OTUs from the *rpoB* sequences were compared against the Hungate1000 database to determine the taxonomy of these rare taxa for both PvuII, and PsiI digested DNA.

3.6.1 Rare taxa in PvuII digested DNA

There were 26 different species that were only detected after normalisation in PvuII digested DNA (**Figure 12**). From these 26 species, 10 were detected after AR1, five after AR2, seven after AR3, 11 after AR4 and 13 after AR5. 14 of these species are detected in more than one round of normalisation. *Enterobacter sp.* KPR-6 was the only one of these species detected in all rounds of normalisation, the rest of the species were detected in three or fewer rounds of normalisation. 16 of the 26 of the detected species were seen in the first three rounds of normalisation.

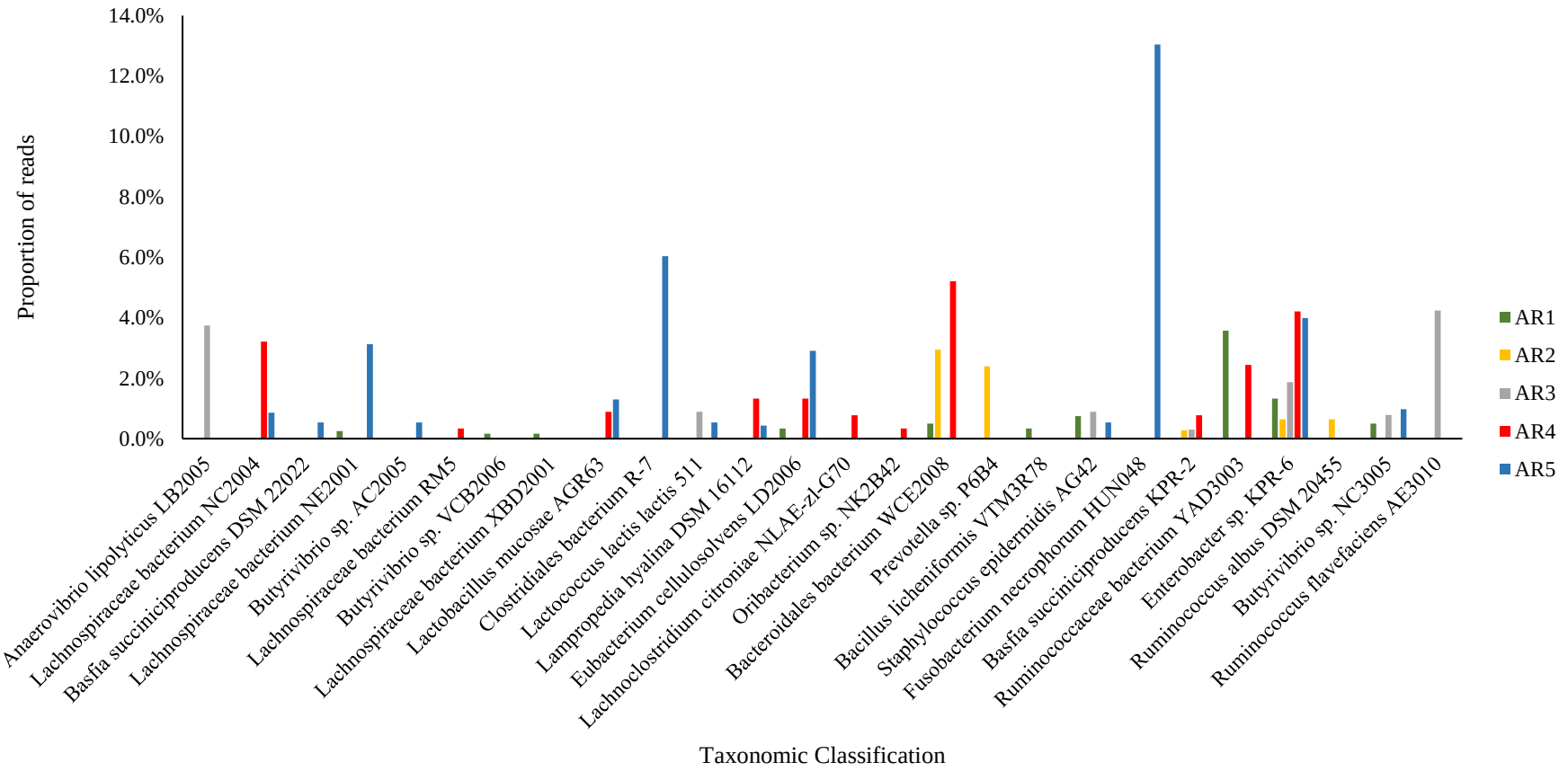

Figure 12: Rare taxa identified in PvuII digested normalised DNA

The proportion of the rumen community of taxa that were only present after normalisation for PvuII digested DNA, mapped against the Hungate1000 database. Different colours represent increasing rounds of normalisation. (AR1: green, AR2: yellow, AR3: grey, AR4: red, AR5: blue).

3.6.2 PsiI digested DNA

There were 25 different taxa which were not seen before normalisation but were detectable after at least one round of normalisation in PsiI digested DNA (**Figure 13**). Of these species, 10 were detected after AR1, seven after AR2, eight after AR3, three after AR4 and nine after AR5. Eight species were detected in multiple rounds of normalisation. *Lachnospiraceae bacterium* AD3010 is the only species which was detected after all five rounds of normalisation. 19 of these rare species were detected in the first three rounds of normalisation. Eleven of these were only able to be detected after normalisation when DNA is digested with PsiI and PvuII.

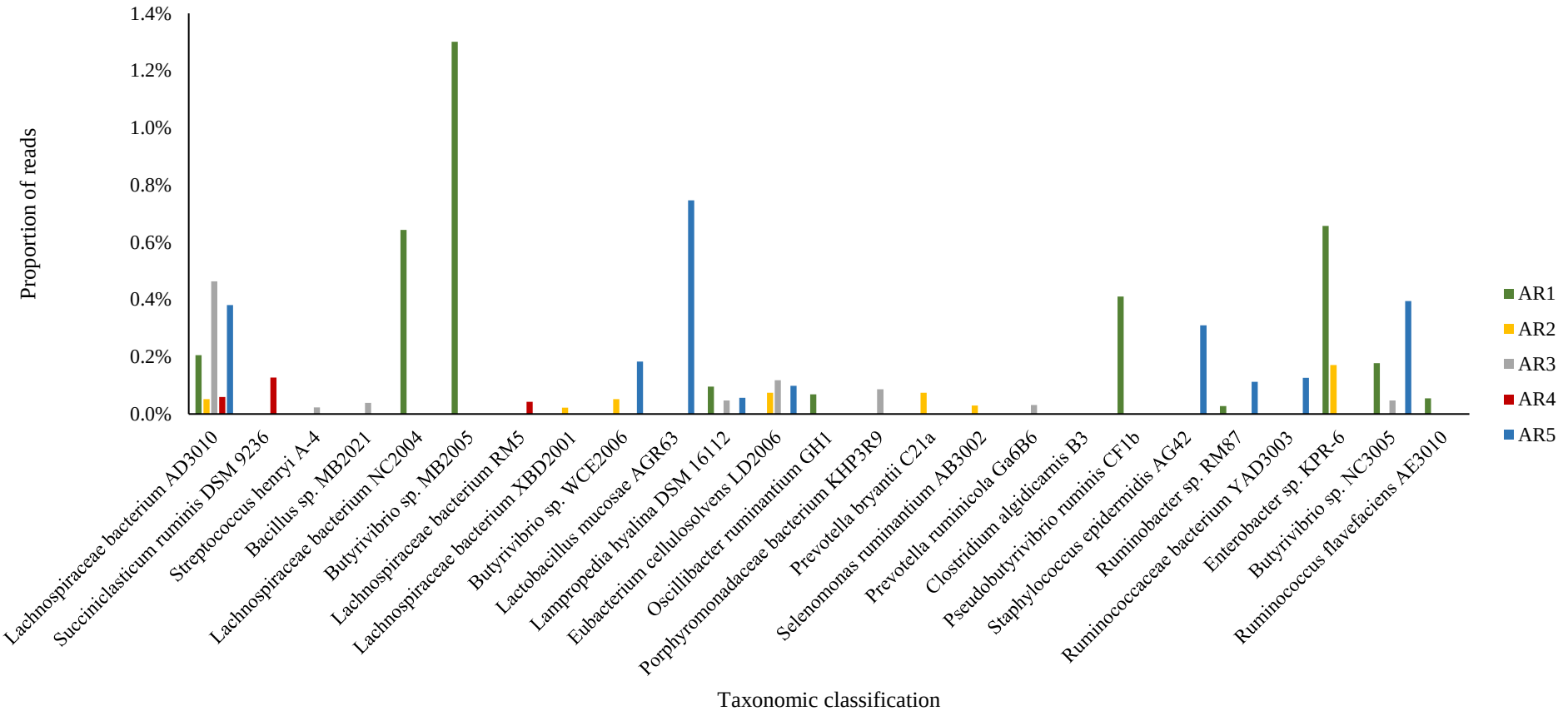

Figure 13: Rare taxa identified in PsiI digested normalised DNA.

The proportion of the rumen community of taxa that were only present after normalisation for PsiI digested DNA, mapped against the Hungate1000 database. Different colours represent increasing rounds of normalisation. (1: green, 2: yellow, 3: grey, 4: red, 5: blue)

3.7 Comparison of 16S rRNA and *rpoB* genetic markers

16S rRNA and *rpoB* genetic markers gave similar taxonomic compositions of the rumen microbiome after normalisation. Using either of these makers showed there was an increase in the number of rare species and a decrease in the number of dominant species after normalisation (**Figure 7**, **Figure 8**). When the OTUs from these two genetic markers were compared against the same database (NCBI), this pattern remains. When looking at one dominant species individually, for example, *Eubacterium* (**Figure 14**), there is no significant difference between the proportion of reads for 16S rRNA and *rpoB* across the five rounds of normalisation (P=0.3550, dependent sample t-test). There is an apparent difference between the two genetic markers before normalisation. However, after AR1 both markers show an increase in proportion between AR1 and AR2, and then a decrease after AR3. The proportion of total reads remains about the same for AR4 and AR5.

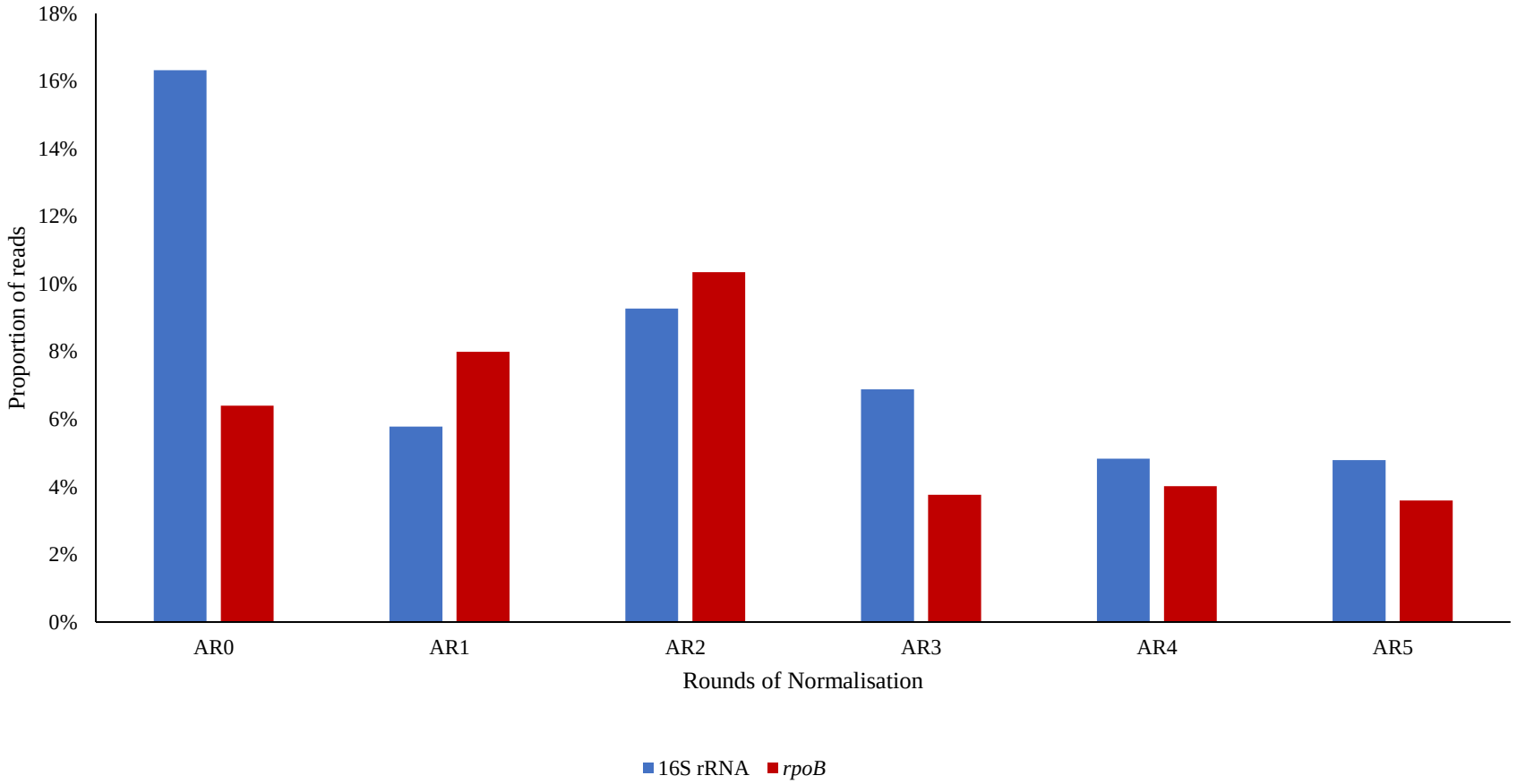

Figure 14: Decrease in the dominance of *Eubacterium*

The change in the dominance of the genus *Eubacterium* through 5 rounds of normalisation for PsiI digested DNA when OTUs were mapped against the NCBI databases for 16S rRNA (blue) and *rpoB* (red). The difference between 16S rRNA and *rpoB* is not large enough to be considered significantly different (P= 0.3550, dependent sample t-test).

4. Discussion

The rumen microbiome has been of great interest due to its crucial role in the New Zealand economy and the current global warming crisis (Clark et al., 2007). The dominant species of this microbiome have been characterised and studied in detail, but little is known about the rare biosphere. The recently discovered duplex-specific nuclease (DSN), the enzyme isolated from the Kamchatka crab which preferentially cleaves double-stranded DNA and leaves rarer ssDNA untouched, proved to be a promising new method for DNA normalisation (Bogdanova et al., 2008, Shagin et al., 2002).

A previous study (Gagic et al. 2015) used a "mock" or synthetic metagenome to establish the methodology for the subtraction of dominant and increase of rare DNA sequences. In this study, the aim was to determine whether DSN-based DNA normalisation can be used to identify the rare biosphere of the natural microbiome, specifically the rumen bacterial microbiome. The ability of the genetic marker *rpoB* to be used in microbial ecology studies of the rumen microbiome was also investigated to determine if it can be used alongside the gold standard marker, 16S rRNA, for taxonomic classification as it has for other microbial communities (Adékambi et al., 2009).

4.1 DNA Normalisation of the Rumen Metagenome

DSN-based DNA normalisation allows us to increase the number of sequences from low abundance microorganisms, including those that are part of the rare biosphere. By applying DSN normalisation to the rumen metagenomic DNA, low abundance microorganisms, including those that are part of the rare biosphere (<0.1%), can be amplified to a detectable level while decreasing the abundance of reads from dominant species. **Figure 7**, **Table 5**). This effect can be seen after the first round of normalisation, in which the number of low abundance OTUs increases and, the number of reads from highly abundant OTUs decreased (**Figure 14**). This effect, however, did not affect all taxa. For example, the proportion of reads in PsiI cut DNA belonging to *Prevotella* remained at 6.8% before and after DSN normalisation (**Table 5**). In most cases, the enrichment of rare reads appeared to be complete after 4 rounds of normalisation and is most prominent after three rounds of normalisation. These results concur

with the previous report on DSN normalisation of a synthetic metagenome (Gagic et al., 2016). Although the synthetic metagenome consisted of five different species pooled together in the ratio that should occur in a natural microbiome, a marked change in the proportion of species only after one round of normalisation was observed. The DSN method resulted in all genomes reaching near equimolar abundance and the representation of the rarest member was increased by approximately 450-fold. Together, these findings suggest that DNA normalisation using DSN is an effective method to increase the number of sequences from low abundance species while simultaneously decreasing the abundance of reads from dominant species in the rumen microbiome.

4.1.1 Dominant Genera Identified in the Rumen Microbiome

The dominant genera identified in this study were consistent with the dominant genera commonly found in other studies of the rumen microbiome. The genera *Prevotella, Ruminococcus, Coprococcus, Butyrivibro,* and genera from Unclassified Clostridales and Unclassified Lachnospiracae are dominant and identified previously in the core rumen microbiome (Henderson et al., 2015b, Xue et al., 2018, Jewell et al., 2015). However, *Fibrobacter,* a commonly found genus in the rumen microbiome was not present at the start of the normalisation (uncut metagenomic DNA, Figure 7). DNA preparation methods have been indicated to effect the DNA yield, and therefore the ability to identify members of the microbiome (Vaidya et al., 2018). But, as *Fibrobacter* was identified in a metasecretome library created using the same DNA preparation method (Ciric, 2014) this unlikely to be the reason for *Fibrobacter* not being present. As the rumen microbiome composition varies with diet, geography and between individual animals (Henderson et al., 2015b, Zhemakova et al., 2016), it is not unexpected that *Fibrobacter* was not detected in this particular sample. Although species of *Fibrobacter* are important members of the fibrolytic consortia in ruminants, functional redundancy is a common trait of complex rumen microbiota.

Notwithstanding in this study, the rumen metagenome of one animal was analysed, the majority of the commonly identified dominant taxa was detected. Having a "typical" microbial community structure at the start of DNA normalisation allowed us to explore the effectiveness of DSN-based subtraction of dominant taxa throughout normalisation rounds. In contrast, the

lack of a commonly found dominant genus, *Fibrobacter,* in our rumen metagenome, indicates that for identification of the rare biosphere in totality more animals need to be sampled.

4.1.2 The Rare Biosphere

Normalising the metagenomic DNA extracted from the rumen digesta revealed a number of taxa which were present in less than 0.1% of the total sequencing reads or undetectable before normalisation (Lynch and Neufeld, 2015). Several taxa that were amplified from a low level (<0.001%) to a detectable level (>0.01%) have already been identified in ruminants. *Streptococcus,* for example, is commonly found in the rumen. However, it was found to be part of the rare biosphere in the microbiome of the sample animal in this study (Bryant, 1959, Krause and Russell, 1996). Comparably to the absence of *Fibrobacter,* an explanation could lie in the redundancy of the rumen microbiome; thus, this particular animal could have a naturally low abundance of *Streptococcus* species. Another explanation could be that *Streptococcus* is a transient taxon; it can be part of the rare biosphere but is periodically recruited and grow up to higher abundance when the conditions are favourable. Studies on *Streptococcus bovis* showed that it is present only when large amounts of starch or sugars (usually grain-fed) are fed and pH is low. Under those conditions, it will grow explosively (Chen et al., 2016). As our sample animal was pasture-fed, it is not unexpected that the abundance of *Streptococcus* species is low. Other genera found in the rare biosphere of our sample were *Granulicatella, Veillonella, Hungateiclostridium* and genera from family Tissierellaceae. These genera have been previously detected in the rumen in low abundance (Rey et al., 2014, Henderson et al., 2015b, Comtet-Marre et al., 2018, Hungate, 1975).

Phylotypes with taxonomic identification after DSN-based normalisation that have not been previously reported or have been reported only when animals were fed special diets and therefore not members of pasture or grain-fed rumen microbiome include: *Haemophilus,* which is abundant in the rumen ureolytic community when animals feed was supplemented with urea (Jin et al., 2016); *Holdemanella* and *Peptoniphilis* which have been isolated from the human microbiome of gastrointestinal and urogenital tract (Bianchi, 2019, Brown et al., 2014); and *Leuconostoc,* commonly found in plants and foods (Liu, 2016) and therefore could be a transient inhabitant of the rumen microbiome.

The success of normalisation, measured by equimolarity of the number of reads per OTU, could also be followed by tracking unclassified OTUs through increasing rounds of DSN-based normalisation. In general, the number of reads from unclassified OTUs increased with the continuing rounds of normalisation (**Appendix 2 Figure S1**). The decrease in the number of unknown OTUs at a lower taxonomic level for the families Lachnospiraceae and Ruminococcaceae is likely because these two families are represented by numerous dominant species in the rumen microbiome (Mackie et al., 2013). Although we could not classify them to a genus level, the taxa which constitute these OTUs are likely part of a dominant genus which was reduced during DSN normalisation.

The unclassified OTUs, not detected in metagenomic DNA before normalisation or detected with a low number of reads, that increased after normalisation rounds (e.g. *Granulicatella, Haemophillis, Leuconstoc,* **Table 5**) suggest that this method could be successfully utilised to identify previously undetected species. The identification of potentially new species in our sample is consistent with current opinions that the rare biosphere contains a microbial seed bank, a source of ecological potential under challenging conditions (Jousset et al., 2017). Except in the rumen as an example of the complex ecosystem, this method can be further used to mine for metabolically compelling species including lignin, hemicellulose and cellulose degraders; which are of paramount importance for animal productivity and have an effect on methane production by methanogens (Ciric, 2014).

4.2 The use of different restriction enzymes in fragmentation of metagenomic DNA

The metagenomic DNA in this study was digested by two different restrictions enzymes, after *in silico* digestion using these enzymes showed that PvuII and PsiI do not cleave through both phylogenetic markers genes in assessed dominant rumen genomes (**Figure 4**). The main aim of using two different restriction enzymes was to compare microbial taxonomic profiles and abundances obtained after their use. In general, microbial community structure was different depending on which enzyme was used. This funding supports the previous hypothesis (Gagic et al. 2015) that more taxa could be found if two or possibly more restriction enzymes are used for initial digestions of metagenomic DNA. As seen in results depicted in **Figure 7** the dominant genus *Ruminococcus* in the rumen metagenome was almost eliminated after only one

round of normalisation in PvuII-digested DNA, however in PsiI digested DNA, it remained dominant. The number of genera identified when the metagenomic DNA was digested with PvuII is much higher than when it was cleaved with PsiI, and some taxa were only detected when one restriction endonuclease was used (**Appendix 1: Figure S1**, **Figure S2**). Therefore, cutting the DNA with different restriction endonucleases is essential to get a more accurate representation of the taxa in the microbial community. This difference between the taxa identified by the two different restriction enzymes can be attributed to different patterns of digestion. If the restriction enzyme cuts the DNA in the gene marker which is amplified, it will either not be amplified, or only a small piece of the marker will be amplified, which would result in removal from the sequencing reads during the denoising process. By removing that fragment, the taxa it represents is also removed from the sequencing results. The different patterns of digestion also resulted in a range of different sized fragments. It is acknowledged that amplification results in the enrichment of smaller (less than 4 Kb) fragments, and PsiI digestion of a synthetic metagenome results in fragments which were larger than 3 kb (Gagic et al., 2015). PsiI digestion resulted in much larger sized fragments compared to that of than PvuII (**Figure 4**); after five rounds of normalisation DNA fragments for both were between 500 - 3000 bp (**Figure 5**). Amplification of the rumen metagenome, therefore, will also select for smaller fragments (less than 4 Kb) and the greater number of taxa identified from PvuII digestion after normalisation, compared with PsiI digestion, is likely due to smaller fragments generated by PvuII. The difference in the size fragments and the different number of taxa that these two different restriction endonucleases have shown after normalisation highlights the importance of using more than one restriction enzyme to digest the samples.

Both restriction endonucleases used in this study generated large DNA fragments (>5 kb). As a result, some of the diversity of the rare biosphere may have been lost by poor amplification of large DNA fragments which were not amplified well. As shown by the Shannon's Diversity Index of each round of normalisation (**Table 4**), the decrease in diversity between uncut metagenomic DNA and AR0 is minor but, is higher in PsiI than PvuII digested DNA.

4.3 How do 16S rRNA and *rpoB* compare as genetic markers?

As well as a difference in microbial community composition identified by using different restriction enzymes, this study also showed that two phylogenetic markers, 16S rRNA and

rpoB, differ in assessment of microbial diversity. A similar pattern was obtained when 16S rRNA and *rpoB* were used to assess the taxonomy of a mock microbial community (Ogier et al., 2019). Despite having less OTUs after denoising and filtering than 16S rRNA-based clustering, taxonomy assignments based on *rpoB* generated more OTUs than 16S rRNA, except for SILVA classification at 94% similarity. Therefore we have shown that *rpoB* can be used successfully as a genetic marker for microbial diversity analysis for the rumen microbiome, as it has been successfully used in analysis of other microbial communities (Vos et al., 2012, Adékambi et al., 2009, Case et al., 2007, Ogier et al., 2019). Although clustering based on *rpoB* aligned a larger number of OTUs taxonomically, the level to which these could be identified down to (genera) was not as high as it was for 16S rRNA. Limitations at lower taxonomic assignments using *rpoB* were expected as designated *rpoB* database had not existed by the time these analyses were conducted; therefore, sequences were obtained from GenBank (www.ncbi.nlm.gov/), or another non-specific database (EggNOG, Hungate1000 genome collection). The lack of a designated database for *rpoB* is due to relatively recent utilisation of it in phylogenetic studies in comparison to 16S rRNA (Vos et al., 2012). The amount of 16S rRNA sequences recorded increased dramatically after its recognition as a "gold standard" genetic marker for use in ecology and phylogeny studies, as there was a considerable influx of studies using it (Janda and Abbott, 2007, Wilson et al., 1990). In contrast, the use of *rpoB* as a phylogenetic marker was only considered in 1997 (Mollet et al., 1997). The number of sequences for *rpoB* is not as numerous, with only 586,000 entries in the protein database for *rpoB* compared to the 37 million nucleotide entries of 16S rRNA. Even when comparing the databases used, Hungate1000 and EggNOG only have 410 reference genomes and 190, 000 orthologous groups compared to the six million and 400,000 aligned sequences of the SILVA and Greengenes databases respectively (Huerta-Cepas et al., 2016, Seshadri et al., 2018, Quast et al., 2013, McDonald et al., 2012b). This massive difference between the number of sequences in the databases of the two different markers may have contributed to the inferior taxonomic classification of OTUs based on *rpoB*.

Although *rpoB* has been used to determine the classification of taxa down to a subspecies level (Adékambi et al., 2009), we classified OTUs down to a genus level (apart from when the Hungate1000 database was used) to avoid false positives due to the pyrosequencing error rate. As a result of this, we chose only to use the protein alignment of *rpoB* to determine taxonomy, rather than to include the fine filtering to a subspecies level that can be determined by using nucleotide level alignments (Case et al., 2007). Taxonomic classification of OTUs obtained

from both markers at a genus level, allowed the comparison of the changes in the taxonomic distribution due to DSN normalisation for both markers. This showed that normalisation occurs with either marker, but some genera are more likely to be seen when using one marker than another (**Table 5**). Future studies could use nucleotide level alignments to mine for taxa which are part of the rumen rare biosphere at the lower taxonomic levels, and therefore, potentially start to uncover the vast ecology potential which is in the rare biosphere of the rumen microbiome and other microbial communities.

The use of *rpoB* as a genetic marker has provided a higher resolution of diversity of the rumen microbiome than we would have seen by only using 16S rRNA. However, the use of this marker still has a few limitations, and therefore, currently, it is advisable to use *rpoB* alongside 16S rRNA in microbial diversity studies.

4.4 Use of Hungate1000 Database

The Hungate1000 database was established with the aim to produce a complete set of rumen microbial gene sequence currently has > 500 reference genomes which are estimated to be around 75% of the total number of species present in the rumen microbiome (Seshadri et al., 2018). This database is considered a valuable resource for the known rumen gene sequences, so it was assumed that it would be able to identify many of the taxa in our sequencing results, particularly from dominant species. Using Hungate1000, we classified 51 genera (compared to >100 genera which were classified with 16S rRNA databases (SILVA, Greengenes)). There are currently only 82 genera in the database; hence the number of genera that could be classified with this database was limited, as it contains fungal and viral genomes in addition to bacteria (Seshadri et al., 2018). The Hungate1000 is far from complete in terms of including taxa which make up the rare biosphere. The broader biological database EggNOG (Huerta-Cepas et al., 2017) gave a better classification of our OTUs. The Hungate1000 database was an appropriate choice of database to assess the decrease in dominant genera and would require, based on our study, more sequenced genomes to complete a vast diversity of species in the rumen microbiome. It is a much more valuable resource for searching for genes encoding metabolically relevant proteins than trying to mine for the rumen rare biosphere, which we have shown contains genera which have not been previously associated with the rumen microbiome. The increase in identified genera with the EggNOG database also corroborates with the result that we have found taxa which have not previously been associated with the

rumen microbiome. Future work should consider analysis against a database of rumen metagenome genomes as it may also give a better classification than the Hungate 1000 genome collection of the genera within the rumen microbiome.

4.5 Higher error rate due to methodology could have overestimated effects of DSN-based normalisation

The methodology which was used in this thesis may have introduced a high number of errors in the sequence reads obtained. Errors would have been added to the reads because we used Roche 454 pyrosequencing, which has a known higher error rate than Illumina sequencing technology, and also because PCR amplification followed in each of the rounds of normalisation (Margulies et al., 2005, Logares et al., 2014, Bennett, 2004). Roche 454 sequencing has long been recognised as having a high error rate due to its nucleotide incorporation method, and therefore, it is likely that we have a higher number of errors in our sequences than if we had used another sequencing method (Margulies et al., 2005). For example, Illumina has a lower error rate, its read length is increasing, and it is becoming less expensive (Luo et al., 2012). Future use of DSN normalisation should consider the use of Illumina rather than Roche 454 pyrosequencing to improve the accuracy of the results obtained. The PCR step to amplify the fractionated DNA at the start of the DSN normalisation steps could also have increased the number of errors, due to the nature of PCR. The DNA polymerase which was used has a reasonably low error rate of 2.5×10^{-5} per nucleotide per sample (Invitrogen, USA); but it does not have a proofreading function, which would have decreased the number of errors. The more rounds of normalisation, the more PCR was performed, and the more errors which would have been introduced into the sequencing reads. Grouping sequences into OTUs and denoising using DADA2 would have removed many of these sequences (Ann Reid, 2011). But, as DADA2 is designed to be used on Illumina sequences, it is possible that it did not remove all of the sequences (Callahan et al., 2016). The few sequences which were not removed by DADA2 may have increased the number of OTUs, which could have led to an overestimation of the actual number of OTUs and therefore the taxa which were increased by normalisation. However, as seen when binning was introduced to microbial ecology studies, the tail of taxa which represent the rare biosphere did not change (Epstein, 2009) thus, the increased amount of errors in our sequencing results may not have had a significant effect on the rumen microbiome composition (Huse et al., 2010). The potentially

high amount of errors from the methodology could explain why some species were only suddenly seen a reasonably high proportion only after 5 rounds of normalisation such as *Lactobacillus mucosae* AGR63, *Clostridiales bacterium* R-7 and *Fusobacterium necrophorum* HUN048 (Figure 12 and 13). It is unlikely that species that were not seen in the other rounds of normalisation were able to be amplified from <0.001% to >0.1% of the total number of reads in one round (**Table 5**). It is more likely that the classification of these species is due to a carried-on PCR or sequencing error which wasn't removed from the sequences by our denoising and filtering errors. To combat the high level of error, future studies using DSN for normalisation to explore the rare biosphere of communities should considering only three or four rounds of normalisation, to reduce the number of accumulated PCR errors; the use a polymerase with a high accuracy rate and proofreading function and the use of the more accurate Illumina sequencing.

4.6 Limitations

Several limitations could be observed in this study. The rumen microbiome sample used in this study was obtained from a single animal at a single time point. As the microbiome of ruminants differs between individuals, as well as diet and geography, this study has likely not been able to identify all bacteria which make up the rare biosphere that can be found within the rumen microbiome. As we only used a sample from one animal, this study cannot determine the composition of the rare biosphere in other ruminants.

Furthermore, this study only focused on the amplification of bacterial genomes from the rumen microbiome. Although bacteria are the dominant Kingdom of the rumen microbiome, there are also archaea, fungi, bacteriophages and protozoa present (Mackie et al., 2013, Seedorf et al., 2015, Alzahal et al., 2017). We have demonstrated that DSN could be an effective method for normalisation of bacterial DNA reads but have not investigated whether this method could be implemented on other microorganisms e.g. archaea.

5. Conclusions

The rumen metagenomic DNA normalisation using DSN has proven that identification of the rare biosphere can be achieved. We demonstrated that using DSN after renaturation of metagenomic DNA could results in an increase of the reads from low abundance microorganisms, while simultaneously decreasing the dominance of high abundance reads. In this first attempt to normalise a complex metagenome, numerous genera that have previously been reported in low abundance, or novel genera including *Haemophilus, Holdmanella, Peptoniphillus* and *Leuconostoc,* have been discovered. Further investigation of these genera could lead to insight into microorganisms which may have potentially important functions in the rumen microbiome, either functionally as keystone species or in terms of a pool of genetic diversity.

The use of the genetic maker *rpoB* alongside 16S rRNA has revealed a number of microorganisms in the rumen microbiome that would have been left undiscovered with the use of only 16S rRNA. Without the use of *rpoB,* four of the emerging and dominant genera in this study would not have been identified. This markers is limited by the lack of designated databases and therefore, taxonomic classification required to assign at a protein level. Despite this limitation, *rpoB*, as a phylogenetic marker would be of value in the future as an increasing number of studies, are using it for phylogenetic studies and currently it could be used alongside 16S rRNA for a better understanding of the microbial diversity and community composition such as the rumen microbiome.

This study has been limited in several ways; thus, further optimisation by investigating the use of a broader range of restriction endonucleases and a more accurate sequencing method is required. The method could also be used for identification of the rare biosphere of other microbial communities. These improvements could result in the ability to explore and potentially access the vast amount of ecological potential in the rare biosphere as well as unveiling keystone species in the rumen and other microbiomes.

6. Future Directions

We have shown that using DSN-based DNA normalisation the rare biosphere of the rumen microbiome could be identified. However, this study was limited to identify at the species level and therefore explore the metabolic potential of phylotypes within the rare biosphere of the rumen microbial community. Future studies could improve upon this work in several ways.

6.1 The whole rumen microbiome

The rumen microbiome is a complex system, which contains fungi, viruses, archaea and prokaryotes as well as bacteria. Future studies of this microbial community using DSN-based normalisation should take this into consideration. It has not been investigated whether normalisation method would work on other prokaryotic metagenomic DNA (archaeal DNA). Phylogenetic markers, including archaea-specific 16S RNA, could be used for assessment of DSN-based normalisation. Only double-stranded DNA viruses could be subjected to DSN-based normalisation as others (ssDNA and RNA viruses) do not follow the same DNA rehybridization kinetics. In summary, future studies that aim to explore the rare biosphere of the rumen microbiome should include markers for other microorganisms as well as bacteria and also recognise that DSN normalisation will not be able to explore the diversity of viruses or eukaryotes (due to the presence of introns)

6.2 Use of Other Ruminants

Ruminant animals are a much broader group of animals than just cattle. It includes a large variety of animals from all over the world, including sheep, deer, goats and bison. To truly explore the rare biosphere of the rumen microbiome, this study could be repeated with a range of samples of a number of different ruminants. As our study only used one sample, future work also needs to include multiple samples from individuals within each different species of ruminant animals. It is acknowledged that the composition of the rumen microbiome varies between individuals, diets, and geography as well as species, thus, to truly explore the rare biosphere of this microbiome these variables need to be taken into consideration in future work.

6.3: Investigation of ecological potential in new species

It is assumed that the rare biosphere is a repository of genes that enables ecological plasticity of the given microbiome. This ecological potential could be provided by keystone species, which have a disproportionate effect on their community compared with their distribution, and by the microbial seed bank, which contains a vast functional gene pool. Future studies of the rare biosphere could explore the taxa which have been identified in the rare biosphere for their ecological potential which could identify species or genes of interest in terms of increasing productivity in ruminant animals or effect on methanogenic archaea and subsequently reducing the amount of methane methanogens produce and their impact on the current climate change crisis.

6.4 Improve Sequencing Technology

Roche 454 pyrosequencing technology has been superseded by Illumina sequencing in recent years. Illumina is more accurate, consistently increasing the read length and decreasing the cost and therefore it the more useful sequencing platform for studies which use HTS. This study has been limited by the number of errors which is suspected to be in our sequencing reads. Therefore future studies using DNA normalisation to explore the rare biosphere of microbial communities should consider using Illumina sequencing technology and third-generation sequencing platforms such as PacBio and Nanopore (Rhoads and Au, 2015, Schneider and Dekker, 2012). By reducing the number of errors in the sequencing and obtaining longer reads (e.g. full-length 16S rRNA), this will allow for exploration of the rare biosphere at a lower taxonomic level than genus level, which will improve the ability to explore the rare biosphere.

6.4 Restriction Enzymes

As the output fragment size using two restriction endonucleases was different, smaller fragments were amplified more efficiently than longer fragments. Future studies using DSN-

based normalisation should consider the use of a wider range of restriction enzymes to try to alleviate the bias against some species due to fragmentation.

Appendix 1: Taxonomic Distribution Profiles

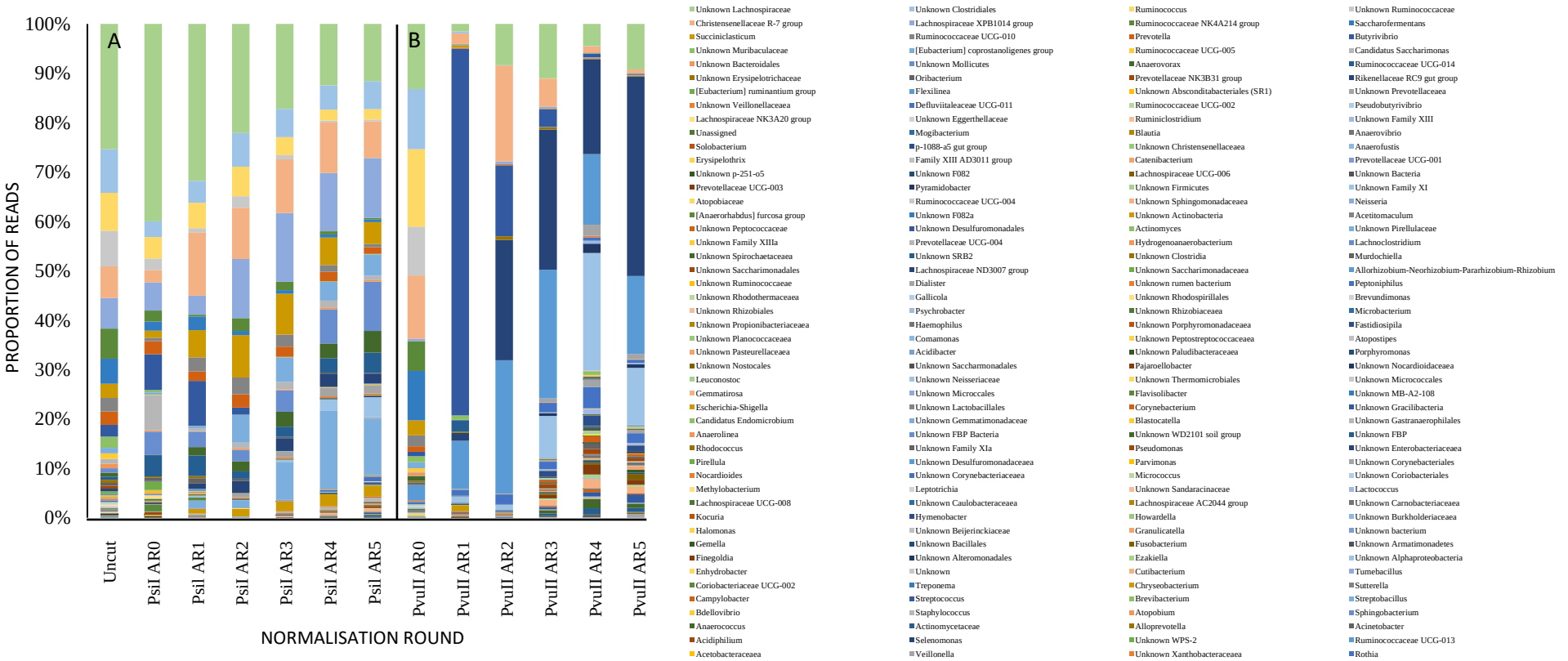

Figure S1: Taxonomic profile for the distribution of 16S rRNA taxanomy at the genus level.

Samples of the sum of all the matches against the SILVA database at 94% similarity, at the genus level. Panel A: uncut metagenomic DNA and PsiI digested DNA after normalisation (AR0-5). Panel B: PvuII-digested DNA AR0-5. The proportion of the genus is each sample is shown by the size of the coloured band in each bar. Colours represent different genera as indicated in the key above.

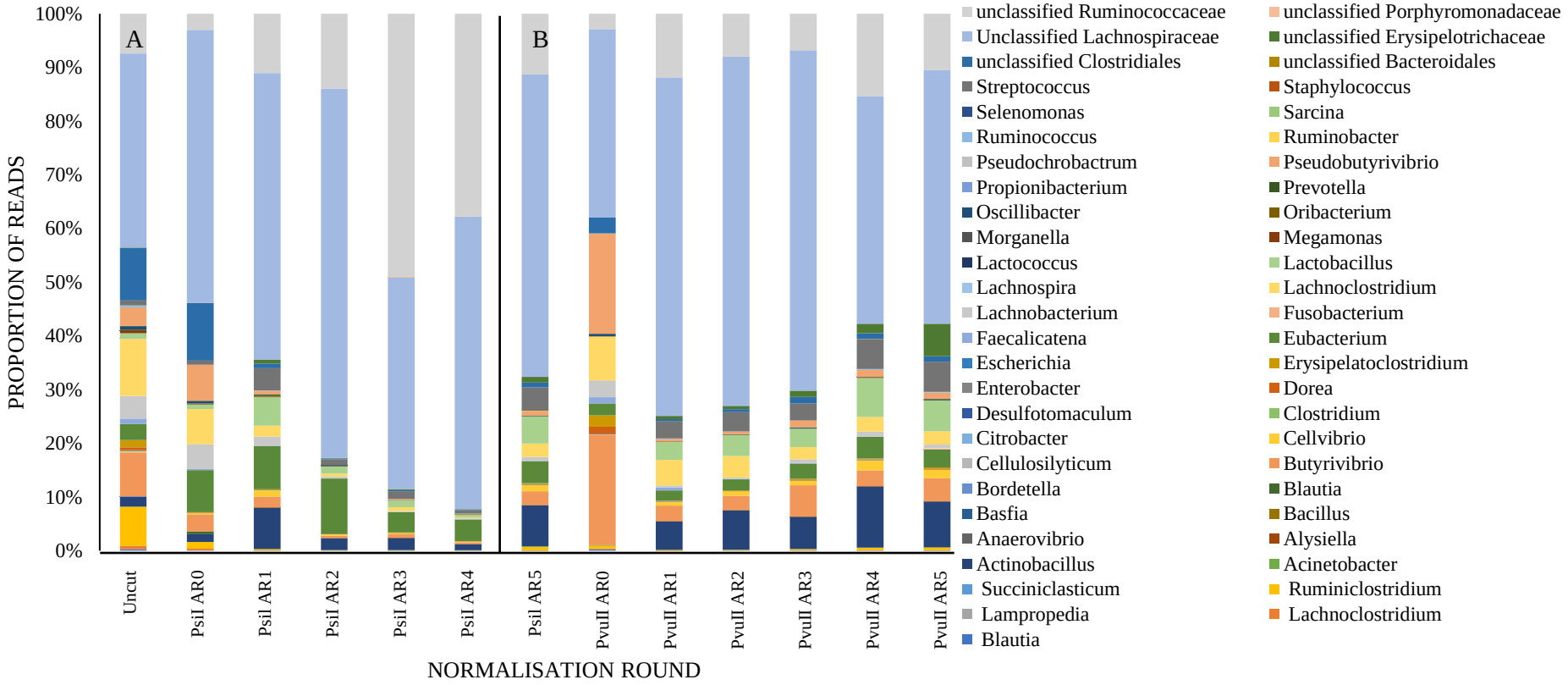

Figure S2: Taxonomic Distribution of *rpoB* taxanomy at the genus level.

Sample of the sum of the all the matches of *rpoB* OTUs when compared to the Hungate1000 database, grouped at the genus level. Panel A: uncut metagenomic DNA and PsiI digested DNA before and after each round of normalisation (AR). Panel B: PvuII-digested DNA AR0-5. The proportion of the genera in each sample is shown by the size of the section in the bar. The colours represent the different genera as indicated in the key above.

Appendix 2: Changes in Taxa Classified at higher taxonomic levels

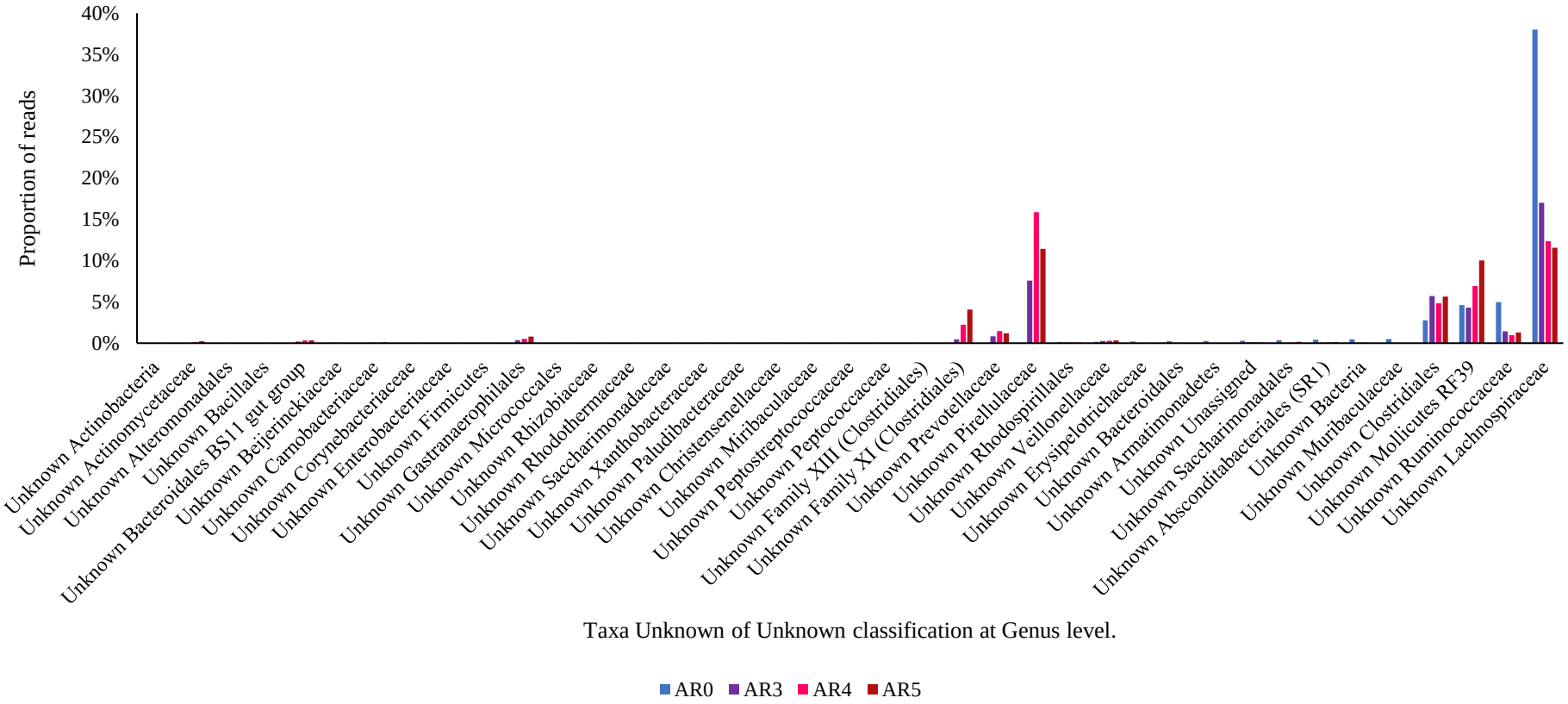

Figure S1: Change in the proportion of unclassified OTUs at the genus level before and after normalisation.

The change in the proportion of taxa which cannot be identified at the genus level for 16S rRNA, before normalisation (blue) and after three (purple) four (pink) and five (red) rounds of normalisation. Taxa are classified to the highest taxonomic level, against the SILVA database at 94%. Taxa that were unclassified are labelled as Unknown Unassigned.

References

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. & Polz, M. F. 2004. Divergence and Redundancy of 16S rRNA Sequences in Genomes with Multiple Operons. *Journal of Bacteriology,* 186**,** 2629-2635. Available: 10.1128/jb.186.9.2629-2635.2004

Adékambi, T. & Drancourt, M. 2004. Dissection of phylogenetic relationships among 19 rapidly growing Mycobacterium species by 1*6S rRNA, hsp65, sodA, recA* and *rpoB* gene sequencing. *International Journal of Systematic and Evolutionary Microbiology,* 54**,** 2095-2105.

Adékambi, T., Drancourt, M. & Raoult, D. 2009. The rpoB gene as a tool for clinical microbiologists. *Trends in Microbiology,* 17**,** 37-45.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology,* 215**,** 403-410.

Alzahal, O., Li, F., Guan, L. L., Walker, N. D. & McBride, B. W. 2017. Factors influencing ruminal bacterial community diversity and composition and microbial fibrolytic enzyme abundance in lactating dairy cows with a focus on the role of active dry yeast. *Journal of Dairy Science,* 100**,** 4377-4393.

Ann Reid, M. B. 2011. *The rare biosphere*. Washington DC, USA: American Academy of Microbiology.

Azizi-Shotorkhoft, A., Mohammadabadi, T., Motamedi, H., Chaji, M. & Fazaeli, H. 2016. Isolation and identification of termite gut symbiotic bacteria with lignocellulose-degrading potential, and their effects on the nutritive value for ruminants of some by-products. *Animal Feed Science and Technology,* 221**,** 234-242.

Becking, L. B. 1934. *Geobiologie of inleiding tot de milieukunde*, WP Van Stockum & Zoon.

Beef and Lamb, N. Z. 2019. *Compendium of New Zealand Farm Facts 2019*. Wellington, New Zealand: Beef and Lamb, New Zealand. [Accessed 25 Jan 2020].

Bennett, S. 2004. Solexa Ltd. *Pharmacogenomics,* 5**,** 433-438.

Bernardi, G. 1965. Chromatography of Nucleic Acids on Hydroxyapatite. *Nature,* 206**,** 779-783.

Bianchi, F. 2019. *Modulation of gut microbiota from healthy-weight and obese individuals by pectin, by-products of tropical fruits and probiotic strains.* PhD, University of Copenhagen.

Bogdanova, E. A., Shagin, D. A. & Lukyanov, S. A. 2008. Normalization of full-length enriched cDNA. *Molecular BioSystems,* 4**,** 205-212.

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A. & Gregory Caporaso, J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome,* 6**,** 90.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., Cope, E., Da Silva, R., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson, I. I. M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., et al. 2018. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints,* 6**,** e27295v2.

Borukhov, S. & Nudler, E. 2003. RNA polymerase holoenzyme: structure, function and biological implications. *Current opinion in microbiology,* 6**,** 93-100.

Boughner, L. A. & Singh, P. 2016. Microbial Ecology: Where are we now? *Postdoc journal : a journal of postdoctoral research and postdoctoral affairs,* 4**,** 3-17.

Brown, K., Church, D., Lynch, T. & Gregson, D. 2014. Bloodstream infections due to Peptoniphilus spp.: report of 15 cases. *Clinical Microbiology and Infection,* 20**,** O857-O860.

Bryant, M. P. 1959. Bacterial species of the rumen. *Bacteriological reviews,* 23**,** 125-153.

Buchfink, B., Xie, C. & Huson, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods,* 12**,** 59-60.

Cai, S. & Dong, X. 2010. Cellulosilyticum ruminicola gen. nov., sp. nov., isolated from the rumen of yak, and reclassification of Clostridium lentocellum as Cellulosilyticum lentocellum comb. nov. *International Journal of Systematic and Evolutionary Microbiology,* 60**,** 845-849.

Cai, S., Li, J., Hu, F. Z., Zhang, K., Luo, Y., Janto, B., Boissy, R., Ehrlich, G. & Dong, X. 2010. Cellulosilyticum ruminicola, a Newly Described Rumen Bacterium That Possesses Redundant Fibrolytic-Protein-Encoding Genes and Degrades Lignocellulose with Multiple Carbohydrate-Borne Fibrolytic Enzymes. *Applied and Environmental Microbiology,* 76**,** 3818-3824.

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**,** 581.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics,* 10**,** 421.

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F. & Kjelleberg, S. 2007. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology,* 73**,** 278-288.

Chen, L., Luo, Y., Wang, H., Liu, S., Shen, Y. & Wang, M. 2016. Effects of Glucose and Starch on Lactate Production by Newly Isolated Streptococcus bovis S1 from Saanen Goats. *Applied and Environmental Microbiology,* 82**,** 5982-5989.

Chung, B. Y., Hardcastle, T. J., Jones, J. D., Irigoyen, N., Firth, A. E., Baulcombe, D. C. & Brierley, I. 2015. The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. *RNA,* 21**,** 1731-1745.

Ciric, M. 2014. *Metasecretome phage display : a new approach for mining surface and secreted proteins from microbial communities.* Doctor of Philosophy, Massey University.

Ciric, M., Moon, C. D., Leahy, S. C., Creevey, C. J., Altermann, E., Attwood, G. T., Rakonjac, J. & Gagic, D. 2014. Metasecretome-selective phage display approach for mining the functional potential of a rumen microbial community. *BMC Genomics,* 15.

Clark, D. A., Caradus, J. R., Monaghan, R. M., Sharp, P. & Thorrold, B. S. 2007. Issues and options for future dairy farming in New Zealand.

Comtet-Marre, S., Chaucheyras-Durand, F., Bouzid, O., Mosoni, P., Bayat, A. R., Peyret, P. & Forano, E. 2018. FibroChip, a Functional DNA Microarray to Monitor Cellulolytic and Hemicellulolytic Activities of Rumen Microbiota. *Frontiers in Microbiology,* 9**,** 215-215.

Cullingham, C. I., Cooke, J. E. K., Dang, S. & Coltman, D. W. 2013. A species-diagnostic SNP panel for discriminating lodgepole pine, jack pine, and their interspecific hybrids. *Tree Genetics & Genomes,* 9**,** 1119-1127. Available: 10.1007/s11295-013-0608-x

Dahllöf, I., Baillie, H. & Kjelleberg, S. 2000. rpoB-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. *Applied and Environmental Microbiology,* 66**,** 3376-3380. Available: 10.1128/aem.66.8.3376-3380.2000

De Wit, R. & Bouvier, T. 2006. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology,* 8**,** 755-758. Available: 10.1111/j.1462-2920.2006.01017.x

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

Difford, G. F., Plichta, D. R., Løvendahl, P., Lassen, J., Noel, S. J., Højberg, O., Wright, A.-D. G., Zhu, Z., Kristensen, L., Nielsen, H. B., Guldbrandtsen, B. & Sahana, G. 2018. Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. *PLOS Genetics,* 14**,** e1007580. Available: 10.1371/journal.pgen.1007580

Dimitriu, P. A., Lee, D. & Grayston, S. J. 2010. An evaluation of the functional significance of peat microorganisms using a reciprocal transplant approach. *Soil Biology and Biochemistry,* 42**,** 65-71.

Environment, M. f. t. 2020. New Zealand's Greenhouse Gas Inventory Wellington, New Zealand. .

Epstein, S. S. 2009. *Uncultivated microorganisms. [electronic resource]*, Springer.

Faith, D. P. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation,* 61**,** 1-10.

Fox, G. E., Wisotzkey, J. D. & Jurtshuk JR, P. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *International Journal of Systematic and Evolutionary Microbiology,* 42**,** 166-170.

Gagic, D., Ciric, M., Wen, W. X., Ng, F. & Rakonjac, J. 2016. Corrigendum to "Exploring the secretomes of microbes and microbial communities using filamentous phage display" [Front. Microbiol. 2016, 7:429. doi: 10.3389/fmicb.2016.00429]. *Frontiers in Microbiology,* 7. Available: 10.3389/fmicb.2016.00927

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**,** 464-476. Available: 10.1111/1755-0998.12321

Gharechahi, J. & Salekdeh, G. H. 2018. A metagenomic analysis of the camel rumen's microbiome identifies the major microbes responsible for lignocellulose degradation and fermentation. *Biotechnology for Biofuels,* 11**,** 216. Available: 10.1186/s13068-018-1214-9

Goodrich, Julia K., Di Rienzi, Sara C., Poole, Angela C., Koren, O., Walters, William A., Caporaso, J. G., Knight, R. & Ley, Ruth E. 2014. Conducting a Microbiome Study. *Cell,* 158**,** 250-262.

Grünberg, W. & Constable, P. D. 2009. Food Animal Practice. *In:* Anderson, D. E. & Rings, D. M. (eds.) *Food Animal Practice (Fifth Edition).* Saint Louis: W.B. Saunders.

Guan, L. L., Nkrumah, J. D., Basarab, J. A. & Moore, S. S. 2008. Linkage of microbial ecology to phenotype: correlation of rumen microbial ecology to cattle's feed efficiency. *FEMS Microbiology Letters,* 288**,** 85-91. Available: 10.1111/j.1574-6968.2008.01343.x

Guder, D. G. & Krishna, M. S. R. 2019. Isolation and Characterization of Potential Cellulose Degrading Bacteria from Sheep Rumen. *Journal of Pure & Applied Microbiology,* 13**,** 1831-1839.

Henderson, G., Cox, F., Ganesh, S., Jonker, A., Young, W., Global Rumen Census, C., Abecia, L., Angarita, E., Aravena, P., Nora Arenas, G., Ariza, C., Attwood, G. T., Mauricio Avila, J., Avila-Stagno, J., Bannink, A., Barahona, R., Batistotti, M., Bertelsen, M. F., Brown-Kav, A., Carvajal, A. M., Cersosimo, L., Vieira Chaves, A., Church, J., Clipson, N., Cobos-Peralta, M. A., Cookson, A. L., Cravero, S., Cristobal Carballo, O., Crosley, K., Cruz, G., Cerón Cucchi, M., de la Barra, R., De Menezes, A. B., Detmann, E., Dieho, K., Dijkstra, J., dos Reis, W. L. S., Dugan, M. E. R., Hadi Ebrahimi, S., Eythórsdóttir, E., Nde Fon, F., Fraga, M., Franco, F., Friedeman, C., Fukuma, N., Gagić, D., Gangnat, I., Javier Grilli, D., Guan, L. L., Heidarian Miri, V., Hernandez-Sanabria, E., Gomez, A. X. I., Isah, O. A., Ishaq, S., Jami, E., Jelincic, J., Kantanen, J., Kelly, W. J., Kim, S.-H., Klieve, A., Kobayashi, Y., Koike, S., Kopecny, J., Nygaard Kristensen, T., Julie Krizsan, S., LaChance, H., Lachman, M., Lamberson, W. R., Lambie, S., Lassen, J., Leahy, S. C., Lee, S.-S., Leiber, F., Lewis, E., Lin, B., Lira, R., Lund, P., Macipe, E., Mamuad, L. L., Cuquetto Mantovani, H., Marcoppido, G. A., Márquez, C., Martin, C., Martinez, G., Eugenia Martinez, M., Lucía Mayorga, O., McAllister, T. A., McSweeney, C., Mestre, L., Minnee, E., Mitsumori, M., Mizrahi, I., Molina, I., Muenger, A., Muñoz, C., Murovec, B., Newbold, J., Nsereko, V., O'Donovan, M., Okunade, S., et al. 2015a. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Scientific Reports,* 5**,** 14567. Available: 10.1038/srep14567

https://www.nature.com/articles/srep14567#supplementary-information
Henderson, G., Cox, F., Ganesh, S., Jonker, A., Young, W., Janssen, P. H., Abecia, L., Angarita, E., Aravena, P., Arenas, G. N., Ariza, C., Attwood, G. T., Avila, J. M., Avila-Stagno, J., Bannink, A., Barahona, R., Batistotti, M., Bertelsen, M. F., Brown-Kav, A., Carvajal, A. M., Cersosimo, L., Chaves, A. V., Church, J., Clipson, N., Cobos-Peralta, M. A., Cookson, A. L., Cravero, S., Carballo, O. C., Crosley, K., Cruz, G., Cucchi, M. C., De La Barra, R., De Menezes, A. B., Detmann, E., Dieho, K., Dijkstra, J., Dos Reis, W. L. S., Dugan, M. E. R., Ebrahimi, S. H., Eythórsdóttir, E., Fon, F. N., Fraga, M., Franco, F., Friedeman, C., Fukuma, N., Gagić, D., Gangnat, I., Grilli, D. J., Guan, L. L., Miri, V. H., Hernandez-Sanabria, E., Gomez, A. X. I., Isah, O. A., Ishaq, S., Jami, E., Jelincic, J., Kantanen, J., Kelly, W. J., Kim, S. H., Klieve, A., Kobayashi, Y., Koike, S., Kopecny, J., Kristensen, T. N., Krizsan, S. J., LaChance, H., Lachman, M., Lamberson, W. R., Lambie, S., Lassen, J., Leahy, S. C., Lee, S. S., Leiber, F., Lewis, E., Lin, B., Lira, R., Lund, P., Macipe, E., Mamuad, L. L., Mantovani, H. C., Marcoppido, G. A., Márquez, C., Martin, C., Martinez, G., Martinez, M. E., Mayorga, O. L., McAllister, T. A., McSweeney, C., Mestre, L., Minnee, E., Mitsumori, M., Mizrahi, I., Molina, I., Muenger, A., Munoz, C., Murovec, B., Newbold, J., Nsereko, V., O'Donovan, M., Okunade, S., et al. 2015b. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Scientific Reports,* 5.
Holmes, D. E., Nevin, K. P. & Lovley, D. R. 2004. Comparison of 16S rRNA, nifD, recA, gyrB, rpoB and fusA genes within the family Geobacteraceae fam. nov. *International Journal of Systematic and Evolutionary Microbiology,* 54**,** 1591-1599. Available: doi:10.1099/ijs.0.02958-0
Huerta-Cepas, J., Forslund, K., Pedro Coelho, L., Szklarczyk, D., Juhl Jensen, L., Von Mering, C. & Bork, P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution*. Available: 10.1093/molbev/msx148
Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., Christian & Bork, P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research,* 44**,** D286-D293. Available: 10.1093/nar/gkv1248
Hungate, R. E. 1975. The Rumen Microbial Ecosystem. *Annual Review of Ecology and Systematics,* 6**,** 39-66.
Huse, S. M., Welch, D. M., Morrison, H. G. & Sogin, M. L. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental microbiology,* 12**,** 1889-1898. Available: 10.1111/j.1462-2920.2010.02193.x
Ishikawa, J., Yamashita, A., Mikami, Y., Hoshino, Y., Kurita, H., Hotta, K., Shiba, T. & Hattori, M. 2004. The complete genomic sequence of *Nocardia farcinica* IFM 10152. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 14925-14930.
Jami, E., White, B. A. & Mizrahi, I. 2014. Potential role of the bovine rumen microbiome in modulating milk composition and feed efficiency. *PLOS ONE,* 9**,** e85423-e85423. Available: 10.1371/journal.pone.0085423
Janda, J. M. & Abbott, S. L. 2007. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. *Journal of Clinical Microbiology,* 45**,** 2761-2764. Available: 10.1128/jcm.01228-07
Jewell, K. A., McCormick, C. A., Odt, C. L., Weimer, P. J. & Suen, G. 2015. Ruminal Bacterial Community Composition in Dairy Cows Is Dynamic over the Course of Two Lactations and Correlates with Feed Efficiency. *Applied and Environmental Microbiology,* 81**,** 4697-4710. Available: 10.1128/aem.00720-15
Jin, D., Zhao, S., Wang, P., Zheng, N., Bu, D., Beckers, Y. & Wang, J. 2016. Insights into Abundant Rumen Ureolytic Bacterial Community Using Rumen Simulation System. *Frontiers in microbiology,* 7**,** 1006-1006. Available: 10.3389/fmicb.2016.01006

Johnson, D. E. & Ward, G. M. 1996. Estimates of animal methane emissions. *Environmental monitoring and assessment,* 42**,** 133-141.
Jousset, A., Bienhold, C., Chatzinotas, A., Gallien, L., Gobet, A., Kurm, V., Küsel, K., Rillig, M. C., Rivett, D. W., Salles, J. F., van der Heijden, M. G. A., Youssef, N. H., Zhang, X., Wei, Z. & Hol, W. H. G. 2017. Where less may be more: how the rare biosphere pulls ecosystems strings. *The ISME Journal,* 11**,** 853. Available: 10.1038/ismej.2016.174
Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research,* 30**,** 3059-3066. Available: 10.1093/nar/gkf436
Kittelmann, S., Pinares-Patiño, C. S., Seedorf, H., Kirk, M. R., Ganesh, S., McEwan, J. C. & Janssen, P. H. 2014. Two Different Bacterial Community Types Are Linked with the Low-Methane Emission Trait in Sheep. *PLOS ONE,* 9**,** e103171. Available: 10.1371/journal.pone.0103171
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**,** 4293-4294. Available: 10.1093/nar/18.14.4293
Krause, D. O. & Russell, J. B. 1996. How Many Ruminal Bacteria Are There? *Journal of Dairy Science,* 79**,** 1467-1475. Available: https://doi.org/10.3168/jds.S0022-0302(96)76506-2
Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology,* 12**,** 118-123. Available: 10.1111/j.1462-2920.2009.02051.x
La Scola, B., Bui, L. T., Baranton, G., Khamis, A. & Raoult, D. 2006. Partial rpoB gene sequencing for identification of Leptospira species. *FEMS Microbiology Letters,* 263**,** 142-147.
Liu, S. Q. 2016. Lactic Acid Bacteria: Leuconostoc spp. *Reference Module in Food Science.* Elsevier. Available: https://doi.org/10.1016/B978-0-08-100596-5.00859-3
Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., Hingamp, P., Ogata, H., de Vargas, C., Lima-Mendez, G., Raes, J., Poulain, J., Jaillon, O., Wincker, P., Kandels-Lewis, S., Karsenti, E., Bork, P. & Acinas, S. G. 2014. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology,* 16**,** 2659-2671. Available: 10.1111/1462-2920.12250
Lozupone, C. & Knight, R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology,* 71**,** 8228-8235. Available: 10.1128/AEM.71.12.8228-8235.2005
Lozupone, C. A., Hamady, M., Kelley, S. T. & Knight, R. 2007. Quantitative and Qualitative Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. 73**,** 1576-1585. Available: 10.1128/aem.01996-06
Lu, G. & Moriyama, E. N. 2004. Vector NTI, a balanced all-in-one sequence analysis suite. *Briefings in Bioinformatics,* 5**,** 378-388. Available: 10.1093/bib/5.4.378
Luo, C., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. 2012. Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLOS ONE,* 7**,** e30087. Available: 10.1371/journal.pone.0030087
Lynch, M. D. J. & Neufeld, J. D. 2015. Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology*. Available: 10.1038/nrmicro3400
Mackie, R. I., ScSweeney, C. S. & Aminoy, R. I. 2013. Rumen. *In eLS.* www.els.net: John Wiley & Sons. Available: doi:10.1002/9780470015902.a0000404.pub2
Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson,

J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-380. Available: 10.1038/nature03959

McCann, K. S. 2000. The diversity--stability debate. *Nature,* 405**,** 228. Available: 10.1038/35012234

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R. & Caporaso, J. G. 2012a. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience,* 1**,** 7. Available: 10.1186/2047-217X-1-7

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., Desantis, T. Z., Probst, A., Andersen, G. L., Knight, R. & Hugenholtz, P. 2012b. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal,* 6**,** 610-618. Available: 10.1038/ismej.2011.139

Mizrahi, I. 2013. Rumen symbioses. *The Prokaryotes: Prokaryotic Biology and Symbiotic Associations***,** 533-544.

Mollet, C., Drancourt, M. & Raoult, D. 1997. rpoB sequence analysis as a novel basis for bacterial identification. *Molecular Microbiology,* 26**,** 1005-1011. Available: 10.1046/j.1365-2958.1997.6382009.x

Moon, C. D., Gagic, D., Ciric, M., Noel, S., Summers, E. L., Li, D., Atua, R. M., Perry, R., Sang, C., Zhang, Y. L., Schofield, L. R., Leahy, S. C., Altermann, E., Janssen, P. H., Arcus, V. L., Kelly, W. J., Waghorn, G. C., Rakonjac, K. & Attwood, G. T. 2014. Exploring rumen microbe-derived fibre-degrading activities for improving feed digestibility. *In Proceedings 6th Australiasian Dairy Science Symposium.* Hamilton, New Zealand.

Ogier, J.-C., Pagès, S., Galan, M., Barret, M. & Gaudriault, S. 2019. rpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. *bioRxiv***,** 626119. Available: 10.1101/626119

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. & Wagner, H. 2019. vegan: Community Ecology Package

Palevich, N. 2011. *Genome sequencing of rumen bacteria involved in lignocellulose digestion : a thesis presented in partial fulfilment of the requirements for the degree of Master of Science in the Institute of Molecular Biosciences at Massey University, Palmerston North, New Zealand*, Massey University.

Pedrós-Alió, C. 2007. Dipping into the Rare Biosphere. *Science,* 315**,** 192-193. Available: 10.1126/science.1135933

Pedrós-Alió, C. 2012. The Rare Bacterial Biosphere. *Annual Review of Marine Science,* 4**,** 449-466. Available: 10.1146/annurev-marine-120710-100948

Pester, M., Bittner, N., Deevong, P., Wagner, M. & Loy, A. 2010. A 'rare biosphere' microorganism contributes to sulfate reduction in a peatland. *The ISME Journal,* 4**,** 1591.

Power, M. E., Tilman, D., Estes, J. A., Menge, B. A., Bond, W. J., Mills, L. S., Daily, G., Castilla, J. C., Lubchenco, J. & Paine, R. T. 1996. Challenges in the Quest for Keystones: Identifying keystone species is difficult—but essential to understanding how loss of species will affect ecosystems. *BioScience,* 46**,** 609-620. Available: 10.2307/1312990

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F. O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research,* 41**,** D590-D596. Available: 10.1093/nar/gks1219

R Core Team 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing. *Vienna, Austria*. Available: URL https://www.R-project.org/

Rattray, P. V., Brookes, I. M. & Nicol, A. M. 2007. *Pasture and supplements for grazing animals*, New Zealand Society of Animal Production.

Rey, M., Enjalbert, F., Combes, S., Cauquil, L., Bouchez, O. & Monteils, V. 2014. Establishment of ruminal bacterial community in dairy calves from birth to weaning is sequential. *Journal of Applied Microbiology,* 116**,** 245-257.

Rhoads, A. & Au, K. F. 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics,* 13**,** 278-289. Available: 10.1016/j.gpb.2015.08.002

Rice, P., Longden, I. & Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics,* 16**,** 276-7. Available: 10.1016/s0168-9525(00)02024-2

Rideout, J. R., He, Y., Navas-Molina, J. A., Walters, W. A., Ursell, L. K., Gibbons, S. M., Chase, J., McDonald, D., Gonzalez, A., Robbins-Pianka, A., Clemente, J. C., Gilbert, J. A., Huse, S. M., Zhou, H.-W., Knight, R. & Caporaso, J. G. 2014. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ,* 2**,** e545. Available: 10.7717/peerj.545

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

Sambrook, J. & Russell, D. W. 2001. *Molecular cloning: a laboratory manual* Cold Spring Harbour Laboratory Press, New York

Schneider, G. F. & Dekker, C. 2012. DNA sequencing with nanopores. *Nature Biotechnology,* 30**,** 326-328. Available: 10.1038/nbt.2181

Seedorf, H., Kittelmann, S. & Janssen, P. H. 2015. Few Highly Abundant Operational Taxonomic Units Dominate within Rumen Methanogenic Archaeal Species in New Zealand Sheep and Cattle. *Applied and Environmental Microbiology,* 81**,** 986-995. Available: 10.1128/aem.03018-14

Seshadri, R., Leahy, S. C., Attwood, G. T., Teh, K. H., Lambie, S. C., Cookson, A. L., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Hadjithomas, M., Varghese, N. J., Paez-Espino, D., Hungate project, c., Palevich, N., Janssen, P. H., Ronimus, R. S., Noel, S., Soni, P., Reilly, K., Atherly, T., Ziemer, C., Wright, A.-D., Ishaq, S., Cotta, M., Thompson, S., Crosley, K., McKain, N., Wallace, R. J., Flint, H. J., Martin, J. C., Forster, R. J., Gruninger, R. J., McAllister, T., Gilbert, R., Ouwerkerk, D., Klieve, A., Jassim, R. A., Denman, S., McSweeney, C., Rosewarne, C., Koike, S., Kobayashi, Y., Mitsumori, M., Shinkai, T., Cravero, S., Cucchi, M. C., Perry, R., Henderson, G., Creevey, C. J., Terrapon, N., Lapebie, P., Drula, E., Lombard, V., Rubin, E., Kyrpides, N. C., Henrissat, B., Woyke, T., Ivanova, N. N. & Kelly, W. J. 2018. Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nature Biotechnology,* 36**,** 359. Available: 10.1038/nbt.4110

Shade, A. 2018. Understanding Microbiome Stability in a Changing World. *mSystems,* 3**,** e00157-17. Available: 10.1128/mSystems.00157-17

Shade, A., Hogan, C. S., Klimowicz, A. K., Linske, M., McManus, P. S. & Handelsman, J. 2012. Culturing captures members of the soil rare biosphere. *Environmental Microbiology,* 14**,** 2247-2252. Available: 10.1111/j.1462-2920.2012.02817.x

Shagin, D. A., Rebrikov, D. V., Kozhemyako, V. B., Altshuler, I. M., Shcheglov, A. S., Zhulidov, P. A., Bogdanova, E. A., Staroverov, D. B., Rasskazov, V. A. & Lukyanov, S. 2002. A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. *Genome Research,* 12**,** 1935-1942. Available: 10.1101/gr.547002

Shagina, I., Bogdanova, E., Mamedov, I., Lebedev, Y., Lukyanov, S. & Shagin, D. 2010. Normalization of genomic DNA using duplex-specific nuclease. *BioTechniques,* 48**,** 455-459. Available: 10.2144/000113422

Shagina, I. A., Bogdanova, E. A., Altshuler, I. M., Luk'yanov, S. A. & Shagin, D. A. 2011. The use of duplex-specific crab nuclease for rapid analysis of single-nucleotide polymorphisms and the detection of DNA targets in complex PCR products. *Russian Journal of Bioorganic Chemistry,* 37**,** 464. Available: 10.1134/S1068162011040121

Shi, W., Moon, C. D., Leahy, S. C., Kang, D., Froula, J., Kittelmann, S., Fan, C., Deutsch, S., Gagic, D., Seedorf, H., Kelly, W. J., Atua, R., Sang, C., Soni, P., Li, D., Pinares-Patiño, C. S., McEwan, J. C., Janssen, P. H., Chen, F., Visel, A., Wang, Z., Attwood, G. T. & Rubin, E. M. 2014. Methane

yield phenotypes linked to differential gene expression in the sheep rumen microbiome. *Genome Research,* 24**,** 1517-1525. Available: 10.1101/gr.168245.113

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M. & Herndl, G. J. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences,* 103**,** 12115-12120. Available: 10.1073/pnas.0605127103

Stafford, K. J. 2017. *Livestock production in New Zealand*, Massey University Press.

Statistics New Zealand, S. N. 2019. *Agricultural Production Statistics: June 2019 (provisional)* [Online]. https://www.stats.govt.nz/information-releases/agricultural-production-statistics-june-2019-provisional: Statistics New Zealand. . [Accessed Jan 26 2020 2020].

Stewart, C. S. & Hobson, P. N. 1997. *The Rumen microbial ecosystem*, Blackie Academic & Professional.

Vaidya, J. D., van den Bogert, B., Edwards, J. E., Boekhorst, J., van Gastelen, S., Saccenti, E., Plugge, C. M. & Smidt, H. 2018. The Effect of DNA Extraction Methods on Observed Microbial Communities from Fibrous and Liquid Rumen Fractions of Dairy Cows. *Frontiers in Microbiology,* 9. Available: 10.3389/fmicb.2018.00092

van Elsas, J. D., Chiurazzi, M., Mallon, C. A., Elhottovā, D., Krištůfek, V. & Salles, J. F. 2012. Microbial diversity determines the invasion of soil by a bacterial pathogen. *Proceedings of the National Academy of Sciences,* 109**,** 1159-1164. Available: 10.1073/pnas.1109326109

Verstraete, L. N. a. W. 1996. Gastro-enteric methane versus sulphate and volatile fatty acid production. *Environmental Monitoring and Assesment,* 42**,** 113-131. Available: https://doi.org/10.1007/BF00394045

Větrovský, T. & Baldrian, P. 2013. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLOS ONE,* 8**,** e57923. Available: 10.1371/journal.pone.0057923

Vivant, A.-L., Garmyn, D., Maron, P.-A., Nowak, V. & Piveteau, P. 2013. Microbial Diversity and Structure Are Drivers of the Biological Barrier Effect against Listeria monocytogenes in Soil. *PLOS ONE,* 8**,** e76991. Available: 10.1371/journal.pone.0076991

Vos, M., Quince, C., Pijl, A. S., de Hollander, M. & Kowalchuk, G. A. 2012. A Comparison of rpoB and 16S rRNA as Markers in Pyrosequencing Studies of Bacterial Diversity. *PLOS ONE,* 7**,** e30600. Available: 10.1371/journal.pone.0030600

Ward, D. M., Weller, R. & Bateson, M. M. 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature,* 345**,** 63-65. Available: 10.1038/345063a0

Wilson, K. H., Blitchington, R. B. & Greene, R. C. 1990. Amplification of bacterial 16S ribosomal DNA with polymerase chain reaction. *Journal of Clinical Microbiology,* 28**,** 1942-1946.

Woese, C. R. & Fox, G. E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 5088-5090. Available: 10.1073/pnas.74.11.5088

Xue, M., Sun, H., Wu, X., Guan, L. L. & Liu, J. 2018. Assessment of Rumen Microbiota from a Large Dairy Cattle Cohort Reveals the Pan and Core Bacteriomes Contributing to Varied Phenotypes. *Applied and Environmental Microbiology,* 84**,** e00970-18. Available: 10.1128/aem.00970-18

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W. & Glöckner, F. O. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic acids research,* 42**,** D643-D648. Available: 10.1093/nar/gkt1209

Yuan, Y., SanMiguel, P. J. & Bennetzen, J. L. 2003. High-Cot sequence analysis of the maize genome. *The Plant Journal,* 34**,** 249-255. Available: 10.1046/j.1365-313X.2003.01716.x

Zhemakova, A., Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., Miyagic, Z., Vila, A. V., Falony, G., Vieira-Silva, S., Wang, J., Imhann, F., Brandsma, E., Jankipersadsing, S. A., Joossens, M., Cenit, M. C., Deelen, P., Swertz, M. A., Weersma, R. K., Feskens, E. J. M., Netea, M. G., Gevers, D., Jonkers, D., Franke, L., Aulchenko, Y. S., Huttenhower, C., Raes, J.,

Hofker, M. H., Xavier, R. J., Wijmenga, C. & Fu, J. 2016. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science***,** 565-570. Available: 10.1126/science.aad3369

Zhulidov, P. A., Bogdanova, E. A., Shcheglov, A. S., Vagner, L. L., Khaspekov, G. L., Kozhemyako, V. B., Matz, M. V., Meleshkevitch, E., Moroz, L. L., Lukyanov, S. A. & Shagin, D. A. 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Research,* 32**,** e37-e37. Available: 10.1093/nar/gnh031