Can microbiome analysis of effluent be used as a proxy for bovine herd health?

A thesis presented in partial fulfilment of the

requirements for the degree of

Master of Science

in

Genetics

at Massey University, Albany, New Zealand.

Alyssa Melanie Earnshaw

2021

Abstract

The microbial community within the bovine gut has been shown to have major impacts on the health of a cow. Most recent studies have focused on identifying and quantifying the gut microbiome of humans, and more recently of agricultural animals. Microbiome studies come with significant challenges, especially around quantifying bias. In this study, careful validation of controls was undertaken to ensure that the method that was followed was appropriate. These controls included a mock community, technical replicates and a spike-in for comparison of methods that could be used to classify a microbiome. Variation in these methods can be caused by sequencing type, classification tools and databases, as well as open-source pipelines. Using this information, the best pipeline was determined and then used to identify genera in previously uncharacterized bovine effluent microbiome samples. This pipeline consists of 16S rRNA gene Illumina sequencing with the USEARCH-UNOISE classification tool.

The dairy effluent system analysed here is interconnected and shows a high degree of similarity in its microbiome composition across sampling locations. Samples have similar microbial communities caused most strongly by collection date, followed by location. The most abundant microbes present are those involved in breaking down faeces (i.e., *Corynebacterium)*. I found that sequencing depth has a large impact on microbiome classification accuracy. Determining the core taxa of a microbiome will enable analysis of any changes from the expected microbes. These changes can be due to normal fluctuations, such as age. However, microbial dysbiosis can be due to pathogenic microbes. I also explored the effect of pathogenic microbes on microbial community composition as it can have a big influence on animal health. Early identification of infections can minimise the financial and bovine welfare impact on farms.

Acknowledgements

Firstly, I would like to acknowledge Livestock Improvement Corporation (LIC) for providing the idea and samples for the work that I have done over the past two years. This would not have been possible without them, and I remain hugely grateful for their support. I would also like to thank Christine Couldrey and Chad Harland for helping me to become a bioinformatician and their patience while I was learning.

Secondly, I would like to thank my supervisors from Massey University. Nikki Freed and Olin Silander, your support has been invaluable, and I have learnt a lot from you. Tim Cooper, thank you for coming on board to provide feedback while I was writing, and responding to all the questions and emails. I appreciate all the insights and how you have taught me to think critically.

Thirdly, to my lab-mates, friends, and family, I cannot express how fantastic you have been. From being available to bounce theories off, to words of encouragement, I would not have made it through without you. There were times when I struggled and times when I made huge progress and you were all there to commiserate and celebrate.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction into microbiomes

1.1 What is a microbiome

The definition of a microbiome has varied over the years, from “a collection of micro-organisms living together in a reasonably defined habitat with distinct physio-chemical properties in a theatre of activity” to “all of the genetic material within a microbiota (the entire collection of microorganisms in a specific niche)” (“Microbiome,” 2018; Whipps, Lewis, & Cooke, 1988).

A microbiome is intrinsically interconnected, as all microbes within the microbiota dynamically affect each other (Berg et al., 2020). Healthy microbiomes tend to remain stable for large periods of time, as most microbes are neutral or beneficial within the community (Hao, Pei, & Brown, 2017; Kim et al., 2017). The microbiome returns to a similar state after perturbations, as microbes rely on each other to survive (Faust, Lahti, Gonze, de Vos, & Raes, 2015). One way that the stability of the microbiome can be lost is through the introduction of new organisms which cause disease (Goodrich et al., 2014; Hao et al., 2017). A second way is through the removal of organisms, for example through antibiotics (Zaheer et al., 2019). Both methods affect stability by changing the composition of genera and therefore the balance of microbes.

Microbiomics - the study of microbiomes - is a rapidly expanding field where research is being driven by developments in technology and the increasingly sophisticated analysis of resulting data (Berg et al., 2020). Microbiology has traditionally focused on microbes that can be cultured within a laboratory environment. However, it has long been recognized that many unculturable microbes also exist within each ecological niche (Hao et al., 2017). The rise in ‘next-generation’ sequencing has made microbiome studies more accessible and highlighted the wide range of microbial diversity within every environment (Hao et al., 2017; Malla et al., 2018).

1.1.1 The importance of understanding factors that impact microbiomes

Microbiome studies provide the opportunity to offer insights into the diversity of microorganisms found in particular environments (Li et al., 2020; Muñoz-Vargas et al., 2018; Zeineldin, Aldridge, & Lowe, 2018a). There are many important factors that need to be considered when planning microbiome experiments, from sample collection right through to analysis of the data. This includes sample storage, DNA extraction and sequencing methodology (Lear et al., 2018; Wu et al., 2019; Yang et al., 2020). Choices made in

implementing these steps are important as it affects which microbes are identified. Any bias can influence the inferred composition of the microbiome (Choo, Leong, & Rogers, 2015; Kim et al., 2017). Moreover, it is important to avoid conclusions being based on chance, rather than environmental factors which show true variation (Pollock, Glendinning, Wisedchanwet, & Watson, 2018; Wesolowska-Andersen et al., 2014).

1.2 Characterization of Microbiomes

1.2.1 An overview of factors affecting microbiome analysis

The microbes which are present in a microbiome can differ from the classification of sequence reads. This is termed bias. Bias can be caused by a multitude of factors within the method. One such factor is sequencing methodology such as Illumina or Nanopore technology (McLaren, Willis, & Callahan, 2019; Santos, van Aerle, Barrientos, & Martinez-Urtaza, 2020). This is further affected by the type of gene sequencing, which is chosen, for example, either 16S rRNA gene or metagenomics sequencing (Lear et al., 2018). The bioinformatic analysis tools and databases which are used also influence classification of sequence reads to specific microbes (Breitwieser, Lu, & Salzberg, 2019). It is important to use controls to understand how the method selected has affected classifications. Classifications should be precise and accurate.

1.2.2 Next-generation sequencing technologies

There are several sequencing technologies that are known as next-generation sequencing (Goodwin, McPherson, & McCombie, 2016). The methodology chosen has a huge impact on how samples are sequenced. One option for next generation sequencing is Nanopore. The Nanopore MinION can sequence long reads, averaging up to 20kb (Amarasinghe et al., 2020; Leggett & Clark, 2017). However, it is more error prone, at 95% accuracy, which can cause uncertainty during classification (Leggett & Clark, 2017). Illumina technology differs from Nanopore technology by producing short, paired sequences of approximately 300 base pairs (Bennett, 2004). However, it has a higher correct-base accuracy of >99% (Fox, Reid-Bayliss, Emond, & Loeb, 2014; Goodwin et al., 2016). It is important to understand the differences when choosing which technology to use when sequencing the microbiome.

1.2.3 Sequencing Methodologies

1.2.3.1 16S rRNA gene sequencing

16S rRNA gene sequencing is where all of the DNA within a microbiome is extracted; the 16S rRNA gene amplified through polymerase chain reaction (PCR); the 16S rRNA gene is sequenced; and then classified (Prodan et al., 2020). The 16S rRNA gene is approximately 1500 bp and has conserved regions between bacterial and archaeal species. In between the conserved regions, there are variable regions which can be used to identify species as they are specific to the bacteria which they came from (Hao et al., 2017; Raju et al., 2018). When amplifying the 16S region, there are a variety of primers and regions that can be used **(Figure 1.1)**. Primers are designed either side of two conserved regions and span the variable region in order to get the PCR product (Shahi, Freedman, & Mangalam, 2017). Variable regions three and four (V3 and V4) are frequently used to identify the bacterial taxa (de Muinck, Trosvik, Gilfillan, Hov, & Sundaram, 2017). These regions have been shown to be ideal for bovine microbiomes (López-García et al., 2018). 16S rRNA gene sequencing is widely used, due to low costs, good sensitivity, and high throughput (Schriefer et al., 2018).

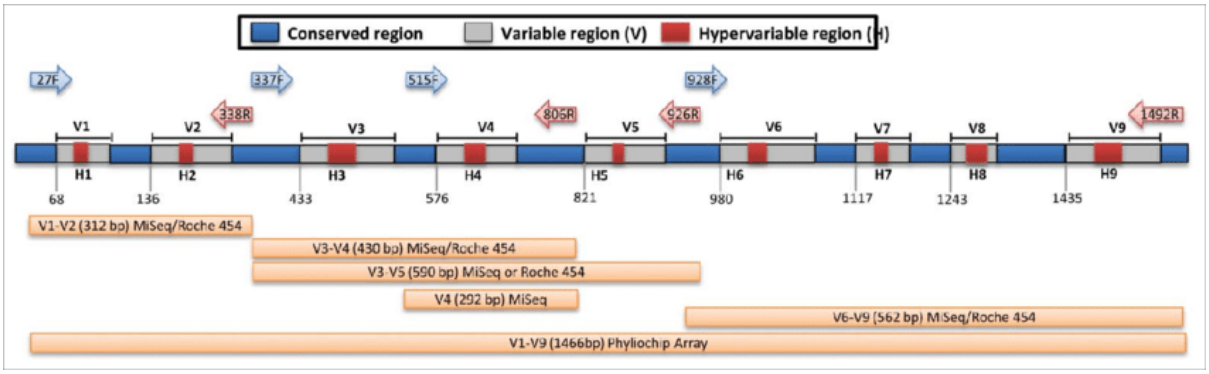

Figure 1.1. The conserved, variable, and hypervariable regions of the 16S rRNA gene within bacteria and various primer pairs for metagenomic sequencing. Conserved regions are blue, variable regions are grey and the hypervariable regions are red. Figure by (Shahi et al., 2017).

1.2.3.2 Metagenomic sequencing

Metagenomics - the study of genetic material from an environment - has become an increasingly popular method to identify taxa in microbiomes (Schriefer et al., 2018; Soon Gweon et al., 2019). There are several ways to metagenomically sequence DNA, with varying costs and results. Whole genome sequencing is when the entire genome is sequenced. It is

more expensive than 16S rRNA gene sequencing (de Muinck et al., 2017). Whole genome sequencing can provide better taxonomic resolution, as reads can be compared to entire genome scaffolds within the database. This increases classification quality due to matching the whole genome rather than a single gene (Pearman, Freed, & Silander, 2020; Ranjan, Rani, Metwally, McGee, & Perkins, 2016).

Shotgun sequencing is where the DNA is randomly sheared, and a subset is sequenced (de Muinck et al., 2017). This is becoming routine due to a lower level of bias compared to 16S rRNA gene sequencing; however, it does classify more genera which may lead to a higher rate of false positives (Ranjan et al., 2016). There are different challenges to 16S rRNA gene sequencing. For example, it is difficult to get specific gene sequences of interest due to the random nature of shearing. Another challenge is that a reference genome is needed for classification, in order to align the read sequences to (Bokulich et al., 2018; Soon Gweon et al., 2019). Without a reference genome, entire genera can be not identified or are misidentified.

1.2.4 Classification tools

Once sequence reads are produced, the next step is to identify the microbes which they came from. This will allow the community composition to be inferred. To do so, read sequences are compared to reference sequence databases which contain known genomes of microbes. There are many classification tools which are used for both metagenomic and 16S sequence reads (López-García et al., 2018; Nearing, Douglas, Comeau, & Langille, 2018; Prodan et al., 2020). These can be used alongside different databases and pipelines. Both the chosen database and pipeline can affect which microbes are classified (Amarasinghe et al., 2020; Lawson et al., 2019). In this thesis, two tools that are used for Nanopore read were compared. These are Kraken2 (version 2.0.8) and KrakenUniq (version 0.5.8) (Breitwieser, Baker, & Salzberg, 2018; Lu & Salzberg, 2020).

1.2.5 Classification databases

One important aspect of taxonomic identification is the database which is used to match the read sequences that have been obtained through sequencing. Many different databases are available to be used in conjunction with these classification tools. These databases can be open access or created by individuals or businesses. The selection of reference DNA sequences in each database can create bias. For example, if the database in question has a relative paucity of specific microbial sequences, this will decrease the ability to correctly assign reads to a specific taxon (Robeson et al., 2020).

Even well-known, high quality reference databases like GTDB, RDP, SILVA and GreenGenes vary in quality and size which can affect the classification and inference of relative microbial abundance (López-García et al., 2018; Parks et al., 2019; Robeson et al., 2020). Depending on the database used, classification of microbial genera varies, and increased diversity can be mistakenly inferred. Such false positives affect the analysis and outcome of an experiment. Therefore, the reference database should be carefully considered to ensure the correct representation of genera (Clarridge, 2004; López-García et al., 2018).

1.2.6 Tools and pipelines for analysing Illumina 16S rRNA gene sequencing data

A lot of data is generated from Illumina 16S rRNA gene sequencing. There are many pipelines available to classify this data. However, there are a variety of parameters that must be chosen before they can be implemented. This ranges from the Operational Taxonomic Unit (OTU) generating strategy to which reference database is used (Golob, Margolis, Hoffman, & Fredricks, 2017; Hao et al., 2017; Quince, Walker, Simpson, Loman, & Segata, 2017). OTUs are used to cluster 16S sequences into taxonomic groupings by sequence similarity (Franzén et al., 2015). Classification is the process by which the clusters of amplicons generated in the OTU step are assigned taxonomies (Golob et al., 2017). OTUs are used as approximations for microbial taxa, and are widely considered to be a useful measure of which genera are present within the microbiome (Edgar, 2018; Franzén et al., 2015)

Many classification tools are available as open source to classify 16S rRNA reads into OTUs. These include QIIME (Caporaso et al., 2010), USEARCH (Edgar, 2010), MOTHUR (Schloss et al., 2009) and DADA2 (Callahan et al., 2016). There are trade-offs between sensitivity, specificity, and speed of these pipelines (Prodan et al., 2020). Several studies have compared the best microbial tool to classify 16S sequences however there is no consensus on the best method (Golob et al., 2017; Nearing et al., 2018; Prodan et al., 2020). As such, I aim to compare USEARCH-UPARSE (version 11) (Edgar, 2013), USEARCH-UNOISE (version 11) (Edgar, 2016) and DADA2 (version 1.14) (Callahan et al., 2016) to determine which one is closest to the ground-truth classification of the mock community. The most accurate pipeline will later be used on previously uncharacterized effluent samples.

1.3 Biases within microbiome studies

Microbiome studies have many steps. Each stage can introduce bias and consequently skew results (Choo et al., 2015; McLaren et al., 2019; Pollock et al., 2018). Currently there is no consensus on the optimal way to collect and store samples; extract and sequence DNA; or to analyse the data (Pollock et al., 2018). The method used to classify microbes can

overrepresent specific genera, and it is difficult to recognise if the reference database captures all diversity (Amarasinghe et al., 2020; Robeson et al., 2020). Bioinformatic analysis can have large influences on which genera are classified (Hao et al., 2017; Milanese et al., 2019)

16S rRNA gene sequencing for microbiome studies has demonstrated that there is a greater diversity of microbes in samples, compared to culture-based methods (Hao et al., 2017). There are two main disadvantages with 16S rRNA gene sequencing. The first disadvantage is that read counts are often used as a proxy for abundance. This can create a bias because different species have different copy numbers of 16S regions (Kembel, Wu, Eisen, & Green, 2012). These species then appear to be more common compared to species with only one 16S rRNA gene (Clarridge, 2004; Větrovský & Baldrian, 2013).

The second disadvantage is that while 16S rRNA gene sequencing is reproducible, it can be inaccurate and distorted towards genera which are already well characterized, rather than identifying novel species (Hao et al., 2017). Researchers often aim to find 'novel' species, causing them to count sequencing errors as true variation (Clarridge, 2004). On the other hand, low read counts can be disregarded in downstream analyses by bioinformatic programs as they are thought to be spurious reads rather than novel species or genera at low abundances (Degnan & Ochman, 2012; Edgar, 2013).

Another bias with 16S rRNA gene sequencing is that only bacteria and archaea can be identified via the 16S region. Fungi are predominately sequenced via Internal Transcribed Spacer (ITS) regions (Bokulich et al., 2018). The ITS region is found in a different region to the 16S rRNA gene and is not amplified by 16S primers. Therefore, when sequencing only the 16S region, fungi and other eukaryotes cannot be identified. Unfortunately, this means that sequencing the 16S rRNA gene region can lead to lower microbial diversity. Despite these limitations, 16S rRNA gene sequencing is still well-utilized due to the ease of sample throughput, low cost, and relatively high sensitivity of bacterial species (Schriefer et al., 2018).

1.3.1 DNA spike-ins and mock communities identify bias within microbiomic experiments.

Spike-ins and mock communities can be used to quantify the amount of bias or misclassification that occurs when using different bioinformatics processes (Hardwick et al., 2018; Venkataraman et al., 2018).

1.3.1.1 DNA spike-ins

DNA from genera that are known to be absent within a specific environment can be added in known quantities to a sample. These genera can then be tracked through the pipeline (Venkataraman et al., 2018). This allows analysis of any taxonomy bias that may occur during DNA extraction, 16S amplification or bioinformatic analysis (de Muinck et al., 2017; Schriefer et al., 2018). Spike-ins are used to investigate the validity of information; if data is lost; and to measure technical variation (Hardwick et al., 2018).

The amount of spike-in can be fractional or fixed. When using fractional spike-ins, prior knowledge about which species are present and their abundances is needed to be able to infer the fractional percentage of the spike-in. By contrast, fixed amounts of a spike-in can be added to all samples to allow for normalisation of genus abundances (Hardwick et al., 2018). Spike-ins make it easier to compare genus abundance and provide higher clarity around the microorganisms that are present (Hardwick et al., 2018; Soon Gweon et al., 2019).

1.3.1.2 Mock community standards

Analysis of mock microbial communities is a widely used method to benchmark a pipeline (Hang et al., 2014). This method involves combining cultures of specific microbes, in known quantities. Mock communities are used as a reference standard during sequencing to be a positive control and enable the measurement of bias (Davidson & Epperson, 2018; Hardwick et al., 2018; Kim et al., 2017). There are many commercial mock communities, one of which the ZymoBIOMICS Microbial Community Standard (cat # D6300). This has ten microbes (eight bacterial and two fungal) present in known amounts. Of the bacterial species, five are gram positive, and three are gram negative. The mock community standard can help quantify bias in DNA extraction, as the efficiency varies due to how well the cells are lysed (Nicholls, Quick, Tang, & Loman, 2019). When undergoing bioinformatic analysis, changes in the bacterial ratios can be tracked and quantified, indicating the reliability of each classification method for previously uncharacterised samples.

1.3.2 Bovine microbiomes

1.3.2.1 The importance of the bovine microbiome

The bovine microbiome has a diverse range of microbes within it. Current research is investigating the genera that are present; their diversity and abundance; and how these influence cow health (Holman & Gzyl, 2019; Zeineldin, Barakat, et al., 2018). Microbes that are present are affected by factors such as diet, age, gender, geographical location, farm

management and health of the host cow (Dill-McFarland, Breaker, & Suen, 2017; Holman & Gzyl, 2019; M. Kim et al., 2014; Li et al., 2020; Zeineldin, Barakat, et al., 2018). As calves age, their microbiome must adapt from a milk-based diet to grass (Fomenky et al., 2018). Cows fed grains or silage have a different microbiome to those only fed grass (Shanks et al., 2011). The amount of antibiotics and stress (such as heat) that cows are put under also affects the microbiome (Li et al., 2020; Vikram et al., 2017). Cows from the same farm have more similar microbiomes compared to cows from other farms, suggesting that the management of the herd has a serious impact on the microbes that are present (M. Kim et al., 2014).

A lot of bovine research has focused on the gastrointestinal tract, in particular the rumen, as this is where most digestion occurs within cows (Pitta et al., 2016; Sbardellati et al., 2020). Dysbiosis within the rumen often leads to disease (Fecteau et al., 2016). However, accessing the rumen to get samples is difficult (Holman & Gzyl, 2019; Kim, Park, & Yu, 2017). The faecal microbiome is similar to the gastrointestinal tract in cattle and could potentially be used as a non-invasive proxy to provide valuable insight into the rumen (Fomenky et al., 2018; Rajan, Lindqvist, Brummer, Schoultz, & Repsilber, 2019; Zeineldin, Aldridge, et al., 2018a).

1.3.2.2 The bovine faecal microbiome

The faecal microbiome of cows is relatively well characterized through culturing methods and 16S rRNA gene sequencing (Fecteau et al., 2016; Wong et al., 2016). However culturing microorganisms can potentially miss a lot of diversity within the microbiome, as many microbes are not able to be cultured (Arnold, Roach, & Azcarate-Peril, 2016; Lagier et al., 2018). There is currently limited research using the faecal microbiome to predict animal health in cattle, and how it changes in response to dietary intervention, probiotics, or disease (Zeineldin, Aldridge, et al., 2018a).

There are many different types of microbes found in the faecal microbiome including both gram-negative and gram-positive bacteria; archaea; and fungi (Pitta et al., 2016). The main phyla that are present across all cattle microbiomes are *Firmicutes* and *Bacteroidetes* (Durso et al., 2010; Fecteau et al., 2016; M. Kim et al., 2014; Minseok Kim & Wells, 2016). These abundances fluctuate between individuals, herds, and geographical location. Additionally, there are many other microbes present within the bovine microbiome (Kim & Wells, 2016; Shanks et al., 2011). As bacterial abundance and genera diversity can be correlated with herd health, characterizing the bovine microbiome is a vital area of research.

Diseases greatly affect the microbiome diversity, which provides insight into the importance of certain genera as markers of animal health (Fecteau et al., 2016; Mao, Zhang, Wang, & Zhu,

2012). It can be difficult to characterize ‘normal’ genera abundance due to a large range of variation within individuals’ microbiomes. However, broad microbial patterns can be studied (Fecteau et al., 2016; Fock-Chow-Tho, Topp, Ibeagha-Awemu, & Bissonnette, 2017). While it is likely that microbial changes can be causal in disease, specific links have not yet been found. Several studies have focussed on pathogenic bacteria in diseases (Fecteau et al., 2016; Muñoz-Vargas et al., 2018). Further investigation into the bovine microbiome is needed to determine if this is a viable method to understand overall changes in the herd health.

In this thesis we aim to address important aspects of the bioinformatics methodology surrounding faecal microbiome analysis and investigate if the faecal microbiome is useful for disease detection.

1.3.2.3 Johne’s Disease

Diseases in cattle can be identified by shifts within the microbiome composition (Fecteau et al., 2016). However, there are also some diseases that are caused by single bacterial species, such as Johne’s disease (Harris & Barletta, 2001). These bacterial species are of interest when classifying microbiomes. Johne’s disease is caused by *Mycobacterium avium* subsp. *paratuberculosis* (MAP). It has a 2–5-year incubation period and causes diarrhoea, weight loss and eventual death of cows (Fecteau et al., 2016; Fock-Chow-Tho et al., 2017). Johne’s disease is hard to control as the transmission of the bacteria is often from dam to calf meaning that intergenerational infection is possible and antibiotics cannot be used long-term (Patterson, Bond, Green, van Winden, & Guitian, 2020; Stinson, Baquero, & Plattner, 2018). This means that the main measure taken to reduce impact of the disease is to cull infected animals which is very costly to the farm (Barratt et al., 2018; Bates, O’Brien, Liggett, & Griffin, 2018; Losinger, 2006).

Shifts of microbes within the faecal microbiome may offer insights into which animals are infected, due to changes in the microbiome. Shifts in bacterial composition in Johne’s Disease tends to be from *Firmicute* to *Actinobacteria* and more broadly a shift from anaerobic to aerobic bacteria in the environment (Beckers, Schulz, & Childers, 2017; Fecteau et al., 2016; Wong et al., 2016). Currently, infection testing occurs via a faecal culture which is slow, costly, and prone to failure. Testing can occur through quantitative Polymerase Chain Reaction (qPCR) which can over or underestimate abundance of the targeted genera (Fock-Chow-Tho et al., 2017; Kruze, Monti, Schulze, Mella, & Leiva, 2013; Ricchi et al., 2017). Research into using 16S rRNA gene sequencing to identify all microbes within a microbiome is underway in order to reduce the reliance on faecal culture or qPCR for pathogenic infections (Britton, Cassidy, O’Donovan, Gordon, & Markey, 2016; Clarridge, 2004; Fecteau et al., 2016). We propose

sequencing the microbiome of herd faecal matter rather than culturing MAP to identify if there is a broad shift in genera over time, and if it is possible to identify MAP levels from this type of sequencing.

1.3.3 Thesis aims

The overarching aim of this thesis is to evaluate the microbiome of cattle effluent at different sampling locations and times, and then identify changes within the microbiome. Before this can be done, the amount of bias caused by various bioinformatic analysis processes needs to be quantified. A large part of bias comes from data analysis and variation caused by the method chosen.

Taxonomic classification tools and the database quality can affect the classification of Nanopore MinION long reads, through both 16S and metagenomic sequencing data. It was found that the database and sequencing method (16S or metagenomics) significantly influences the accuracy of taxonomic classification. Next, three open-source bioinformatics pipelines for taxonomic classification on 16S Illumina reads were compared. It was found that there are significant differences in classification efficiency depending on the pipeline used.

Finally, the microbiome in bovine faecal matter samples from a dairy farm in the Waikato was characterized. Samples were provided by Livestock Improvement Corporation (LIC) and classified using Illumina sequencing with the USEARCH pipeline. The bovine faecal microbiome is a growing area of research, however little work has been done to characterize the changes between different health states. Using effluent as a proxy for animal health is a relatively novel concept, and in this thesis, we investigated if it was possible to 1) classify the microbiome in effluent and 2) identify if there are specific genera or constellations of genera that associated with the presence or absence of a particular pathogen, MAP.

The main aims were:

- To carefully consider sequencing and bioinformatic methods to identify if they have significant effects on taxonomic identification.
- To validate the use of mock communities and technical replicates in showing the amount of bias within a pipeline.
- To further characterize the genera present in the bovine faecal microbiome.

Chapter 2: Comparing the bioinformatic influence on classification

2.1 Classification of the microbiome

The microbiome is a complex mix of microbial species. This can make it hard to identify all genera that are present, especially using traditional culturing methods (Lazcka, Del Campo, & Muñoz, 2007; Quince et al., 2017). Culturing is time consuming and slow. Not all genera can be isolated and cultured creating bias in genera found by researchers (Blanchard, 2012). For these reasons, sequence-based approaches have been developed (Kim et al., 2017; Van Rossum, Ferretti, Maistrenko, & Bork, 2020). Sequencing approaches allow identification of organisms independent of their culturability. Nevertheless, it is subject to biases stemming from potential differences in the efficiency with which sequences can be produced and characterized from different microbiome species. As the identity and distribution of genera present is not known in microbiome studies, calculating the level of bias can be challenging.

Below, the key steps involved in using sequence-based approaches to characterize microbiomes are described. There is a focus on identifying alternatives at the bioinformatic steps which cause bias in the eventual characterization of the community. Typically, there are five main steps from collecting samples through to analysing the classifications **(Figure 2.1)**.

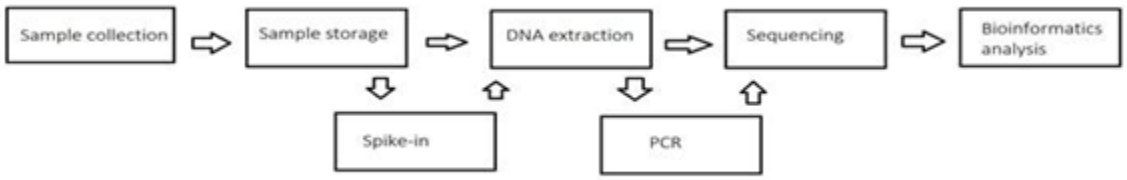

Figure 2.1 A workflow of steps involved in classifying microbiomes. The top five are crucial to each study, while the bottom two are optional, depending on the method chosen.

2.1.1 Sample collection

Cow faecal samples were collected fresh from a farm in Auckland in March 2019. LIC samples were collected fresh, over several time-points from October 2016 to March 2017 **(Appendix A)**, from multiple locations on an LIC farm in the Waikato.

2.1.2 Sample storage.

Cow faecal samples were stored at -20°C overnight, before being transferred to a -80°C freezer, as per the recommendation of Song (2016). LIC effluent samples were stored at -80°C and were transferred on ice when necessary.

2.1.3 DNA extraction

The QIAGEN "DNeasy Powersoil kit protocol" was used to extract DNA from all samples for this thesis. This protocol is widely recognised as a good method for DNA extraction across soil and faecal samples (Hang et al., 2014; Vo & Jedlicka, 2014). The ZymoBIOMICS mock community standard (cat D6300) was also extracted this way. The ZymoBIOMICS spike-in control I (cat D6320) was added to the cow control samples at the step of aliquoting faecal matter for DNA extraction.

2.1.4 Sequencing

Metagenomics has become an increasingly popular method to sequence microbiomes (Schriefer et al., 2018; Soon Gweon et al., 2019). Metagenomic sequencing is when the entire amount of DNA within the community is sequenced. This is expensive however it reduces bacterial bias, as there is no PCR step, and captures potentially all genetic sequences in the sample (de Muinck et al., 2017).

16S rRNA gene sequencing captures bacterial diversity within the community by amplifying the conserved 16S region through PCR. Variable regions are used to differentiate species (de Muinck et al., 2017). 16S rRNA gene sequencing detects a greater diversity of microbes compared to cultured samples. Despite its reproducibility, it is biased toward known genera; genera that are over-represented in reference sequence databases; and bacteria with multiple 16S rRNA genes (Clarridge, 2004; Hao et al., 2017). This bias is caused by unknown read sequences matching to ones that are already present in the databases, and those that are present multiple times in databases are matched more often (Parks et al., 2019).

2.1.5 Bioinformatic analysis

When analysing sequence data to identify genera that are present within a microbiome, there are many different programs that can be used. Classification tools compromise between speed, accuracy, and repeatability (Breitwieser et al., 2019). Computational tools are often developed as researchers find a need for them which has led to an abundance of taxonomic classifiers and databases (Gao, Lin, Revanna, & Dong, 2017; Goodrich et al., 2014; Milanese

et al., 2019; Robeson et al., 2020). Even when classifying technical replicates, there can be different percentages of genera. This emphasises the need to understand the effects of taxonomic tools (McLaren et al., 2019). Most profilers work well at classifying OTUs from phylum to family levels, however it is harder to classify at the genus to species level (Bokulich et al., 2018).

There are many pipelines available to classify Illumina 16S rRNA gene data. These includes QIIME (Bolyen et al., 2019), DADA2 (Callahan et al., 2016), Kraken2 (Wood, Lu, & Langmead, 2019), USEARCH-UPARSE (Edgar, 2013) and USEARCH-UNOISE (Edgar, 2016). There have been studies into which microbial classification tool is optimum however no consensus has been reached (Nearing et al., 2018; Prodan et al., 2020). There are different biases using each method. A mock community and/or technical replicates can ensure that the method is accurate. Kraken2 with the GTDB database was analysed with Illumina read sequences but results are not shown due to a high level of unreliability. Currently there are better methods available for Illumina 16S rRNA data compared to Kraken2 analysis and we chose to test these classification pipelines instead of Kraken2. There is the potential to use Kraken2 and perhaps it would be useful for Illumina metagenomics sequencing classification. Three different 16S rRNA pipelines were compared to understand classification and prove that the method developed is reliable. This method is later used in chapter three on uncharacterised samples from LIC.

2.1.6 Summary

There are many ways to classify sequence output to identify genera present in a microbiome sample (Rajan et al., 2019; Ranjan et al., 2016). These methods are not equivalent, and the pipeline used can skew results (Breitwieser et al., 2019; Hao et al., 2017). Critical bioinformatic and experimental steps were compared to quantify biases that can be introduced in otherwise comparable analysis protocols. This occurred by using the ZymoBIOMICS mock community sample (cat #D6300), a commercially available standardized mix of known species as well as using ZymoBIOMICS spike-in control I (cat D6320) to evaluate bias. This identified which stages of the bioinformatics analysis are more important to consider than others. An in-house pipeline was created to use on effluent samples **(S3.1)**.

2.2 Materials and methods

2.2.1 Control Samples used to compare classification tools

2.2.1.1 Mock community standard

The ZymoBIOMICS Microbial Community Standard (cat # D6300) contains ten species in equal amounts; three easy-to-lyse gram-negative bacteria, five tough-to-lyse gram-positive bacteria, and two tough-to-lyse yeasts species (Nicholls et al., 2019). The expected percentages of classification are found in **Table 2.1.** The mock community was sequenced as a positive control to ascertain the accuracy of classification.

Genus	16S expected frequencies	Metagenomics expected frequencies	Gram status
Pseudomonas	4.2	6.1	Negative
Escherichia	10.1	8.5	Negative
Salmonella	10.4	8.7	Negative
Lactobacillus	18.4	21.6	Positive
Enterococcus	9.9	14.6	Positive
Staphylococcus	15.5	15.2	Positive
Listeria	14.1	13.9	Positive
Bacillus	17.4	10.3	Positive
Other	0.0	1.0	Fungi

Table 2.1. The ZymoBIOMICS mock community expected classification frequencies based on sequencing methodology, as from the ZymoBIOMICS website and explained below.

As stated in the ZymoBIOMICS mock community standard manual, the 16S rRNA gene theoretical composition was calculated from the theoretical genomic DNA composition using the formula; 16S copy number = total genomic DNA x unit conversion constant/genome size x 16S copy number per genome.

A similar approach was used to calculate the metagenomics theoretical composition as; genome copy number = total genomic DNA x unit conversion constant/genome size (Zymo Research Corporation, 2017).

2.2.1.2 Technical replicates

Cow faecal samples were collected on the 10th of March 2019 and stored immediately at -20°C for 12 hours, before being transferred to a -80°C freezer where they were stored until DNA was extracted. Technical replicates from a single cow, "Ginger", from a farm in Auckland, was used to test reproducibility of extraction and sequencing methods. The "cow" sample was split into six replicates. The first three replicates were extracted as is, and the last three samples included a spike of the ZymoBIOMICS Spike-in Control I (High Microbial Load, cat # D6320), in 10-fold increasing amounts **(Table 2.2)**.

The ZymoBIOMICS spike-in consists of equal cell numbers of two bacteria strains, *Imtechella halotolerans* and *Allobacillus halotolerans*. When spiked into an unknown sample, this serves as an *in situ* positive control for DNA sequencing-based microbiome measurements. The two bacteria have different cell wall structures, (*Imtechella halotolerans* is gram-negative and *Allobacillus halotolerans* is gram-positive) representing different cell recalcitrance and exposing potential bias during DNA extraction. This standard enables absolute cell number quantification in microbiome measurements (Zymo Research Corporation, 2017).

2.2.2 Laboratory preparation of samples

The cow samples were thawed and aliquoted into the powersoil-recommended 0.25 g amount for this protocol. I then followed the QIAGEN Powersoil protocol "DNeasy Powersoil kit protocol" to extract DNA. The concentration of DNA present after DNA extraction was determined by using the Thermofisher Qubit Fluorometer.

For the first set of comparisons, the control samples were sequenced twice, once using the 16S rRNA gene, and once metagenomically. Both sequencing methodologies were undertaken using the Oxford Nanopore MinION (Jain, Olsen, Paten, & Akeson, 2016). The protocols followed were the "16S Barcoding Kit - RAB-204" for 16S rRNA sequencing and the "Rapid Barcoding Kit SQK-004" protocol for metagenomics sequencing.

For the second set of comparisons, samples were sequenced using Illumina 16S rRNA gene sequencing. The methods in the Illumina "16S Metagenomic Sequencing Library Preparation" guide were followed to amplify the 16S rRNA region from samples up to the Index PCR step. For samples with low DNA concentration and less than 12.5 ng of input DNA, the maximum volume of 2.5 µl of DNA was used. Samples were sequenced by Auckland Genomics, at the University of Auckland.

With the Illumina sequences, the reads were quality checked and trimmed using FastQC (Andrews & Others, 2010). The forward and reverse reads were joined using fastp-join (Chen, Zhou, Chen, & Gu, 2018).

2.2.3 Comparison of taxonomic classification using Nanopore Sequencing

For the comparison of Nanopore data, both the 16S rRNA gene and metagenomic sequencing data was investigated. The effect of classification tools was analysed, and then four databases were compared using the best classification tool **(Figure 2.2)**.

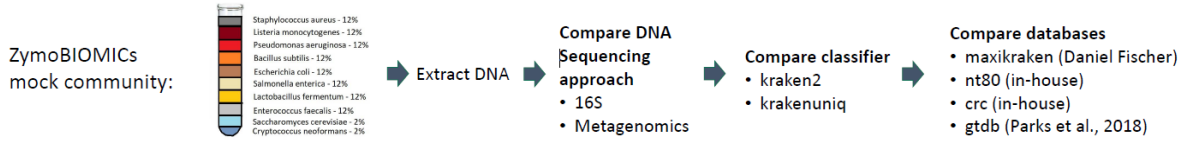

Figure 2.2 The workflow used for comparing different tools and databases using the ZymoBIOMICS mock community. The first shows the mock community composition, then the order of steps to compare the bioinformatic method.

2.2.3.1 Sequencing methodology

The ZymoBIOMICS mock community and the cow replicates were sequenced using Nanopore MinION technology. The effect of sequencing was compared using both the 16S rRNA gene sequencing and metagenomics sequencing.

2.2.3.2 Taxonomic classifiers

Two taxonomic classifiers were compared to determine which classifier best identifies the species in the control mock community. The two classification tools used for this comparison was Kraken2 (version 2.0.8) and Krakenuniq (version 0.5.8) both with default parameters (Breitwieser et al., 2018; Wood et al., 2019).

2.2.3.3 Database

The impact of using different reference databases was assessed by comparing four databases using the Kraken2 classification tool. The ZymoBIOMICS mock community was used to determine how well these four databases worked. The databases considered were Maxikraken (Derrick Wood Nick Loman, 2018) Derrick Wood), NT80 (an in-house NCBI nucleotide database with 80 addition eukaryotic genomes), CRC (an in-house refseq database specifically aimed for colorectal cancer bacteria) and GTDB (Parks et al., 2019). These four databases vary in the number of genome scaffolds; specificity to a certain environment; and the quality of read sequences. This was done to investigate if the size and type of database affects the classification of genera.

2.3.3.4 Visualisations

Pavian (version 1.0.0), an R based visualisation tool, was used to compare the classifications between databases and sequencing methodology (Breitwieser & Salzberg, 2020). Pavian shows the percentage of reads that are taxonomically classified as each genus within the sample. This information can be used to visualise how many genera are present and their relative abundance. The ratio of genera as shown by Pavian was divided by the expected ZymoBIOMICS ratio to measure how closely the database matched the known genera. This ratio was then $\log_2$ transformed to centre it around zero. The expected ratio should be zero, which would show that classification is exactly what is expected.

Based on the percentage of reads, a Principal Components Analysis (PCA) through factoextra (version 1.0.7) was used to visualize variation between samples. This gives us an indication of whether database or sequencing methodology has the greatest effect on samples.

2.2.4 Taxonomic Classification of Illumina Sequencing

Read classification was performed using USEARCH-UPARSE (version 11) (Edgar, 2013), USEARCH-UNOISE (version 11) (Edgar, 2016) and DADA2 (version 1.14) (Callahan et al., 2016). The taxonomic classification was evaluated by reference to the control mock community. For each classifier, the output was compared to the ZymoBIOMICS expected ratios, whilst using default parameters. Before reads were entered into the classifiers, the forward and reverse reads were merged using fastp-join (version 0.20.0) (Chen et al., 2018). This was to prevent any errors or dramatic read losses during the merging steps of individual classifiers and provide a closer comparison, whilst minimising errors.

Below, an overview of the differences and comparisons between the 16S rRNA taxonomic classifiers is presented. The most suitable pipeline for faecal samples is then identified.

2.2.4.1 DADA2

DADA2 (Callahan et al., 2016) gets error rates from a subset of reads and uses this information to learn the expected number of errors per read. DADA2 creates an amplicon-sequence variant (ASV) table, which is similar to an OTU table in USEARCH. Chimeric reads are removed and then taxonomy is assigned using the SILVA database.

2.2.4.2 USEARCH-UPARSE

UPARSE (Edgar, 2013) was used with the RDP database (Maidak et al., 2000) to assign taxonomy. This process uses quality filters and then finds unique sequences to select OTUs from the merged reads. The goal of UPARSE is to subset the correct biological sequences, so that OTUs which are 97% similar are classified together. This means each OTU can theoretically have more than one species present within it, although the closest match is classified. However, due to the variation in the number of 16S rRNA genes in the genome of each species (paralogs) the number of read counts per OTU does not always correlate perfectly with abundance.

2.2.4.3 USEARCH-UNOISE

UNOISE (Edgar, 2016) has a similar pipeline to UPARSE. The processing steps of filtering and de-duplication remain the same as UPARSE, which is the first half of the pipeline. In the second half of the pipeline, instead of subsetting biological species, UNOISE uses denoising to create zero-radius OTUs (zOTUs) These are also known as Amplicon Sequence Variants (ASVs). The UNOISE algorithm performs error-correction, so that any errors introduced from sequencing are minimized (Edgar, 2016; Prodan et al., 2020).

2.2.4.4 Visualisations

For the two USEARCH pipelines, R (version 4.0.2) was used to analyse and visualize read counts per classified taxa (R Core Team, 2020). Overall frequencies were calculated by dividing the classified read counts per OTU by the total read count per sample to enable comparison across samples and to normalise differences in read numbers between samples. This data was visualized in R using the basic heatmap function, adjusting margin size to increase the number of labels. R was also used to create a Principal Coordinate Analysis (PCoA) through the vegan package (version 2.5.6) to investigate variance between samples.

For visualisations of the DADA2 classifications, Phyloseq (version 1.3.20) (McMurdie & Holmes, 2013) was used to visualise the ASV table. This is a follow-on program from DADA2, created by the same group. Phyloseq was used to create a bar chart of genus-level classifications for both the ZymoBIOMICS and cow replicate samples. The segment lines within the same genus come from the number of OTUs that are created by DADA2 (McMurdie & Holmes, 2013). The size of the groupings comes from the sum of reads within the OTUs.

2.3 Results

2.3.1 Comparison of mock community classification using Nanopore sequencing

2.3.1.1 DNA extraction

The average amount of DNA extracted using the Qiagen Powersoil Kit was 28.1 ng/µl across the seven control samples. When preparing cow.G4 for 16S Nanopore sequencing, the DNA concentration after purification was below the 0.05 ng/µl limit of detection for the Qubit Fluorometer and therefore this sample was not sequenced **(Table 2.2)**.

2.3.1.2 Read numbers for Nanopore sequencing

The number of reads that were obtained using Nanopore metagenomic sequencing averaged approximately 109,300 reads across seven control samples. The number of 16S reads that were obtained using Nanopore sequencing averaged approximately 321,300 for six control samples **(Table 2.2)**.

Sample name	Sample type	DNA concentration after DNA isolation (ng/µl)	Metagenomic read counts	16S read counts
cow.G1	Technical replicate	47.4	173,078	334,429
cow.G2	Technical replicate	29.1	91,489	514,000
cow.G3	Technical replicate	24.4	95,030	114,315
cow.G4 plus 1µl of spike control (D6320)	Technical replicate plus spike control	33.4	61,734	NA
cow.G5 plus 10µl of spike control (D6320)	Technical replicate plus spike control	23.0	86,476	213,165
cow.G6 plus 100µl of spike control (D6320)	Technical replicate plus spike control	24.1	45,515	515,067
ZymoBIOMICS mock community (D6300)	Mock Community	15.6	211,809	237,259

Table 2.2. The seven control samples that were used, their DNA concentrations and Nanopore sequencing read counts.

2.3.1.3 Different classifier tools give similar genus level classifications

The performance of Kraken2 (Wood et al., 2019) and Krakenuniq (Breitwieser et al., 2018) was compared using the ZymoBIOMICS mock community as a control. Both 16S rRNA and metagenomic classification were analysed using the CRC database. The time taken for each tool to run was comparable, and while the CRC database was used, there are a multitude of databases that can be configured for both (Lu et al., 2018). Krakenuniq has a lower false positive rate, however Kraken2 has lower memory requirements. Due to the two tools being equivalent, Kraken2 was chosen as we had access to more databases using this tool.

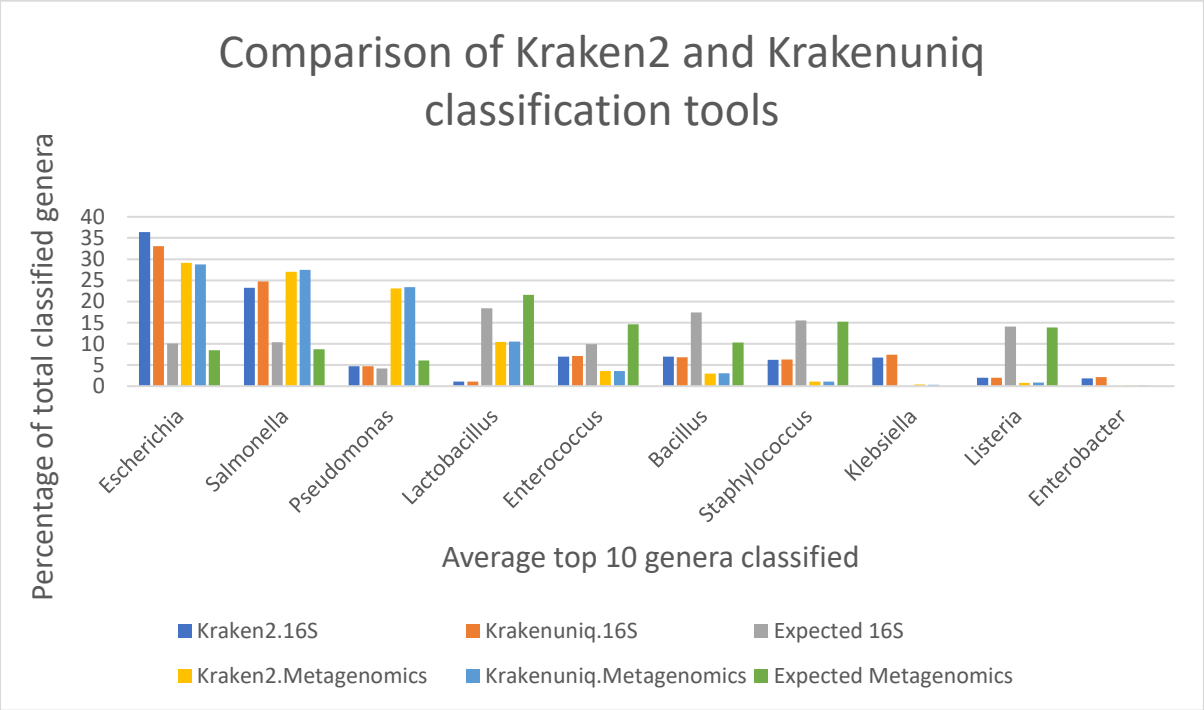

Figure 2.3. Comparing Kraken2 and Krakenuniq as classifying tools using 16S and metagenomic sequencing using the CRC (in-house) database. 16S classification is dark blue and orange, with the expected 16S frequency in grey. Metagenomics classification is yellow and light blue, with the expected metagenomics frequency in green.

There is relatively little variation between classifiers as determined by the lack of significant difference between classification patterns across ten most abundant genera. Metagenomic analysis has no significance (Paired t-test comparing classifiers: t_{81}=-1.12, P=0.26). 16S analysis has a marginally non-significant difference (Paired t-test comparing classifiers: t_{81}=-1.97, P = 0.05) **(Figure 2.3)**. There are slight differences between 16S (blue and orange bars, paired t-test comparing 16S: t_{81}=0.018, P=0.99) and metagenomic sequencing (grey and

yellow bars, paired t-test comparing metagenomics: t_{81}= -0.27, P=0.79). This shows that the sequencing method has more of an effect than the tool used.

Escherichia and *Salmonella* are classified at high percentages across both sequencing methodologies, with other genera much lower, potentially caused by the overestimation of the two mentioned genera. Metagenomic sequencing has higher classification of *Salmonella, Pseudomonas* and *Lactobacillus.* However, 16S rRNA gene sequencing has a higher classification of the other genera found in the mock community **(Figure 2.3)**. There is a mix of gram-negative and gram-positive bacteria that are over-classified by each sequencing methodology. These classification percentages are skewed from the theoretical composition of the mock community **(Table 2.1)**.

2.3.1.4 Effect of database used on genera classification

The tool used has little effect, while sequencing has some effect on classification. The database can affect classification when they are targeted towards specific genera. To investigate how the sequencing and database compares to the actual ZymoBIOMICS community from **Table 2.1**, the observed and expected frequencies of genera present in the sample were compared. Then these values were $\log_2$ transformed them to be around zero. The closer to zero the better, as it shows that the classification is what I expected. The database that is used to classify sequences has a large effect on genus level classification **(Figure 2.4)**. These differences are apparent across both 16S and metagenomic sequencing methodologies. To examine the difference in detail, a series of specific analyses are presented below and reasons as to why this occurs is proposed.

There is a wide range of variation between classification of the same sample using different methods **(Figure 2.4)**. Metagenomics sequencing using the GTDB or CRC database is most likely to correctly classify genera within the sample, while 16S rRNA gene sequencing with the Maxikraken or CRC database is also likely to closely resemble the actual samples. However, 16S rRNA gene sequencing with the nt80 or GTDB database has the worst ratio, and this method is not recommended.

The database has a large effect on how well species are classified **(Figure 2.4)**. This figure shows the wide range of variation between the same sample, and that choosing a database is important for accurate classification and valid analyses. As GTDB claims that it has a highly robust and complete database, it is interesting to note that with 16S rRNA gene sequencing, it was highly inaccurate (Parks et al., 2019, 2018). However, with metagenomics sequencing it was much closer to what is expected. As GTDB classification is through average nucleotide

identity, and Nanopore has a lower per base accuracy, the shorter 16S reads are more likely to be misclassified (Leggett & Clark, 2017; Parks et al., 2019). The nt80 database has the lowest classification rate out of the four databases compared here, potentially because it matches genes rather than full sequences (Sayers et al., 2011).

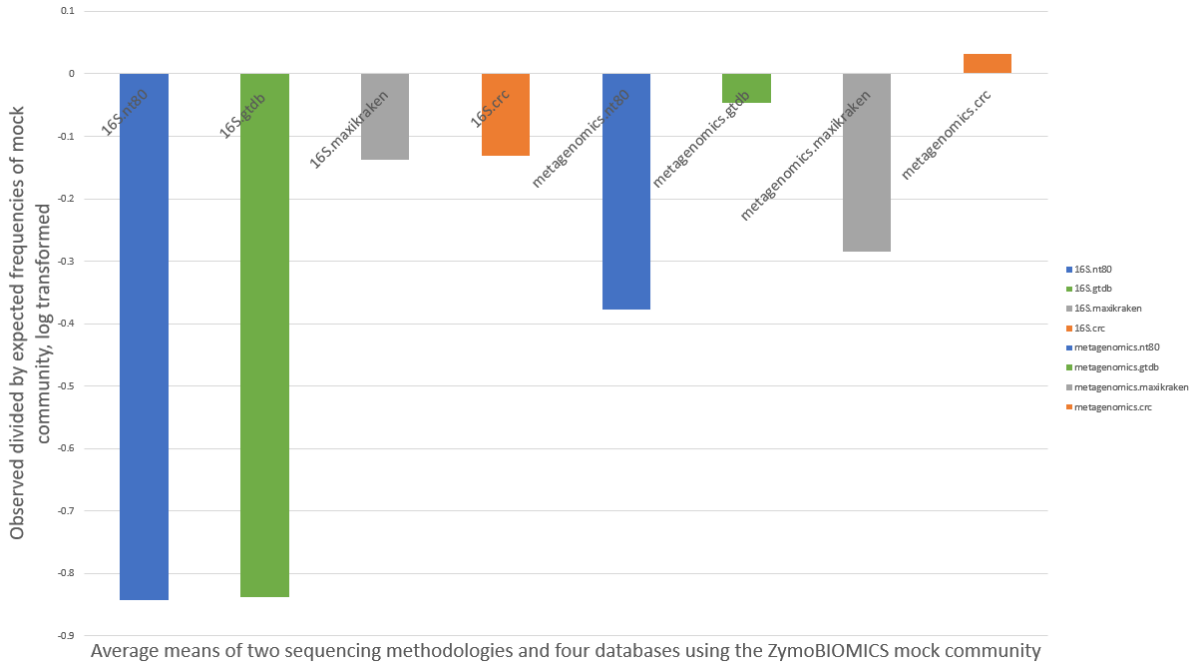

Figure 2.4. Grand mean of log2 observed read classifications of the mock community divided by the expected read classifications mock community for different databases. The first four are 16S sequenced, the last four are metagenomic sequencing. Nt80 database is blue GTDB is green, Maxikraken database is grey and CRC database is orange.

The control samples segregate by sequencing methodology (**Figures 2.3 & 2.5)**, which is to be expected, as the type and length of sequencing reads is different. 16S rRNA gene sequencing amplifies and sequences the 16S rRNA gene present in all bacteria. This can cause bias as all classifications are based on a short 250bp fragment of DNA. This can lead to very similar DNA reads. Metagenomics sequencing sequences longer DNA fragments and classifies them based on matching to entire scaffolds in the database. As more of the read length needs to match, this reduces the genera that it could be. However, this also means that more reads can be discarded.

A PCA was created using all genera classified by the respective sequencing methodology and database. Reads for each genus were normalised by dividing each read count by the total number of reads. The sequencing methodology has a greater impact than databases, as

shown by **Figure 2.5**. Metagenomics sequencing clusters together, which means that they are more similar. This does not mean that they are more accurate classifications.

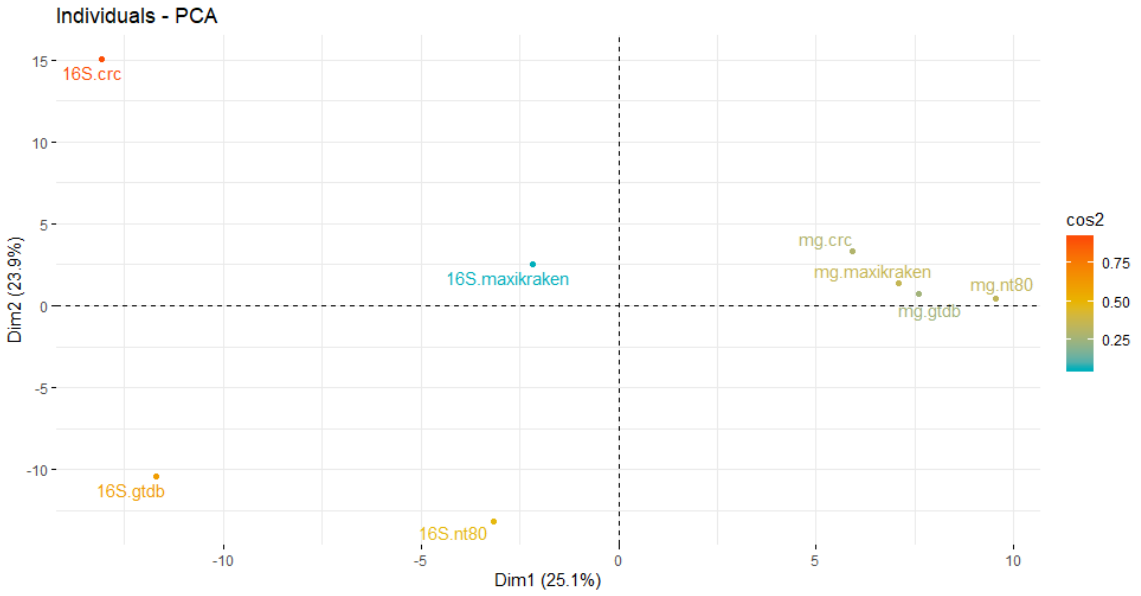

Figure 2.5. A PCA which compares sequencing methodology (16S and metagenomic) alongside four different databases (nt80, CRC, Maxikraken and GTDB). The names are coloured by the amount of difference.

2.3.2 Analysis of technical replicates

With both 16S and metagenomics sequencing, Kraken2 was used with the GTDB database, to evaluate a highly diverse microbial community taken from faecal matter from a single cow. The GTDB database was chosen, as it has a high number of established genomes and will be updated in the future (Parks et al., 2018). Six technical replicates were used to evaluate reproducibility of methods **(Table 2.2)**. Three were identical, and three contained an additional two species (ZymoBIOMICS spike-in, high microbial load – Cat#D6320) at tenfold increasing amounts which allowed for analysis of bias in the DNA extraction. This spike-in contains two species, *Allobacillus halotolerans,* a gram-positive bacterium and *Imtechella halotolerans,* a gram-negative bacterium.

The genera *Imtechella,* from the gram-negative bacteria from the ZymoBIOMICS spike-in– *Imtechella halotolerans* – was identified at increasing frequencies across the three samples **(**cow.G4-6, **Table 2.2**, **Figure 2.6, Figure 2.12)**, as expected. Also, as expected, it is not identified in the first three replicates (cow.G1-3, **Table 2.2**). This shows that generally the gram-negative cells are being lysed and sequenced evenly. On the other hand, the gram-

positive bacteria - *Allobacillus halotolerans* - were not found, despite being at the same concentration in the spike-in **(Figure 2.6)**. This is potentially due to a bias when sequencing and classifying (Han et al., 2020).

2.3.2.1 Replicates have highly similar classifications

The most abundant genera found by metagenomic sequencing within the cow faecal sample is *Pseudomonas* at ~40% **(Figure 2.6A)**. This could be caused by *Pseudomonas* being in high concentrations in the soil, or potentially passed through from the rumen (Leclercq et al., 2016; Patel, Patel, Vohra, & Dave, 2020). While this would be interesting to investigate, it is outside the scope of the research done here. Other genera remain in low concentrations **(Figure 2.6A)**. These genera may be at slightly higher concentrations, as the abundances decreases proportionally as *Pseudomonas* increases.

16S rRNA gene sequencing has high percentages of the genera *RC9* and *CAG-110,* however these are present at only ~8% compared to metagenomics **(Figure 2.6B)**. *Rikenellaceae RC9* is known to be present in both the rumen and faeces of cattle, which is likely due to its role in digestion (Andrade et al., 2020). *CAG-110* itself is not mentioned, however it is likely to be *Firmicutes* which is well characterized in the bovine microbiome, and simply a naming convention from GTDB (Minseok Kim & Wells, 2016; Li et al., 2020).

Despite cow replicate G4 not being metagenomically sequenced, *Imtechella* was found in all cow replicates. This shows that there could be some crossover of the barcodes when sequencing or that some genera are classified incorrectly at low percentages. Due to the low level of misclassification, genera with low read counts should be discarded. *Imtechella* was also found with 16S rRNA gene sequencing within cow.G6 - the one with the highest spike-in. Similar genera can be found by using both the 16S rRNA gene and metagenomic sequencing, keeping in mind that there will still be variation in what is classified. Across both sequencing methodologies, the technical replicates have very similar classifications. This shows that the method used is reproducible and can be used to gain an overview of the genera present within samples (McGovern, Waters, Blackshields, & McCabe, 2018).

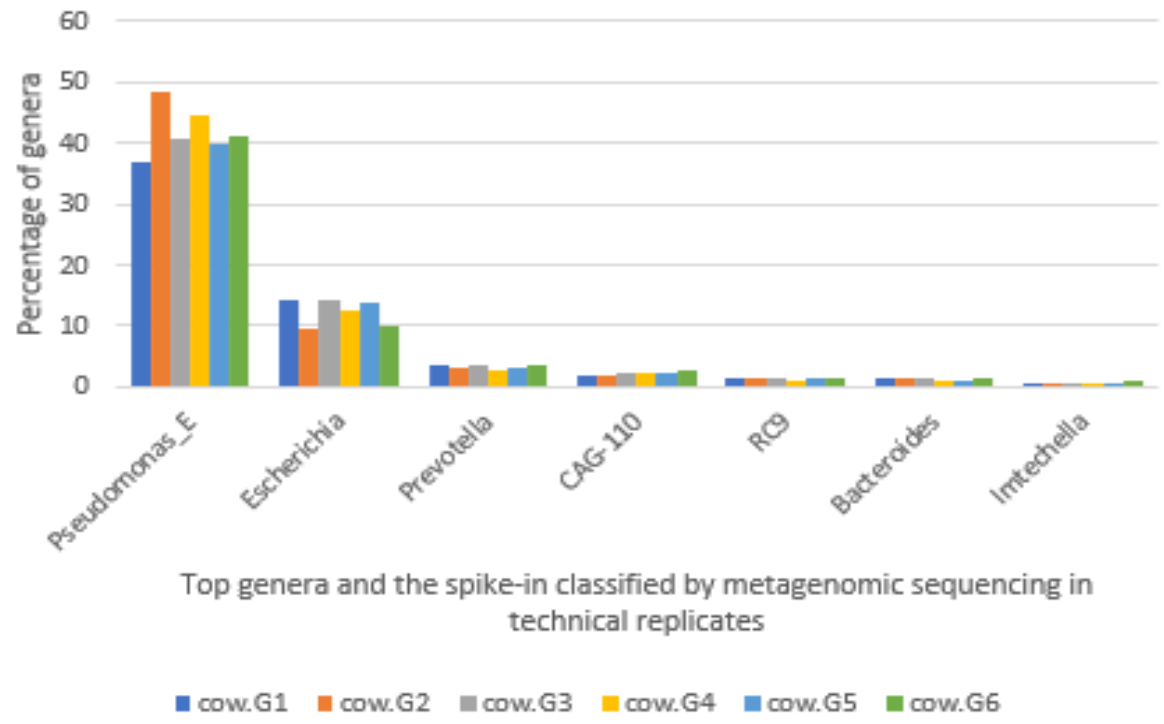

Figure 2.6A. Replicate faecal samples from a single steer. They were metagenomically sequenced. They were taxonomically classified by kraken2 and the GTDB database. The six most abundant genera and the spike-in present are shown.

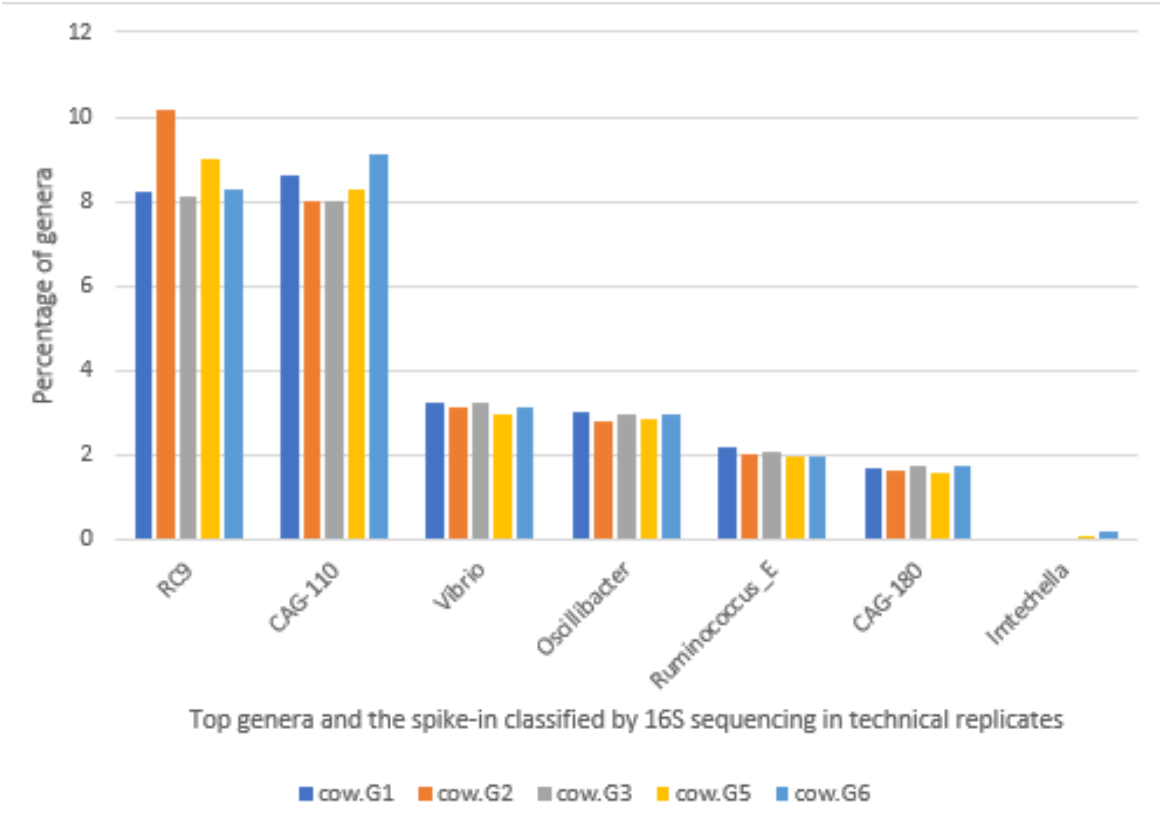

Figure 2.6B. Replicate faecal samples from a single steer. They were 16S sequenced. They were taxonomically classified by kraken2 and the GTDB database. Cow.G4 was not 16S sequenced. The six most abundant genera and the spike-in are shown.

2.3.2.2 Principal coordinate analysis to investigate if variation is caused by sequencing type or database

The results show that sample type, sequencing technology and database can all affect identification of genera within microbiome samples. To investigate the effects of each variable, a principal coordinate analysis was performed. PCoA is used to visualise variation within the samples by looking at those with the highest amount of dissimilarity.

Cow samples group by sequencing method (16S or metagenomic) and database in a PCoA. This shows that both factors have a large influence on variation within samples **(Figure 2.7)**. The cow samples split into two groups – 16S rRNA gene sequenced samples and metagenomically sequenced samples. The notable outlier is the cow 16S CRC and 16S Maxikraken database samples which cluster away from the other samples. Secondly, all the GTDB samples, including the ZymoBIOMICS mock community cluster together towards the left-hand side, showing that the GTDB is highly similar through both metagenomics and 16 rRNA gene sequencing. However, this is likely partially due to the naming conventions being slightly different to the other databases, which can influence the reduction of principal components.

The other databases are all on the right-hand side of the PCoA **(Figure 2.7)** and are relatively mixed between the 16S and metagenomics sequencing. The 16S nt80 cow replicates are only slightly segregated, indicating high similarity between them - which is the ideal outcome due to the nature of these samples. However, the metagenomics samples, without GTDB, cluster to the right of the PCoA. This shows that metagenomics sequencing has a higher degree of similarity compared to the 16S rRNA sequencing samples, which are separated more by database.

Except for the ZymoBIOMICS GTDB samples (16S rRNA and metagenomics sequencing), all the mock community samples cluster together on the right-hand side of the PCoA **(Figure 2.7)**. This shows that different types of databases can still be more similar rather than being influenced by sequencing methodology.

When investigating the effects on sequencing type and database, it was found that sequencing type has more of an effect than database. This is shown by the metagenomically sequenced samples clustered together on the right-hand side, whereas the 16S rRNA sequenced samples are more variable on the left-hand side **(Figure 2.7)**. This shows metagenomic samples are more similar in terms of the type and abundances of genera. For this reason, metagenomic sequencing is thought to be more reliable. It does not mean that metagenomic

sequencing is more correct, only that the output has a higher degree of similarity between samples. However, the database used does affect the output, as the 16S rRNA gene samples vary in how they cluster. All the GTDB samples, including metagenomics sequencing, are on the left-hand side of the PCoA.

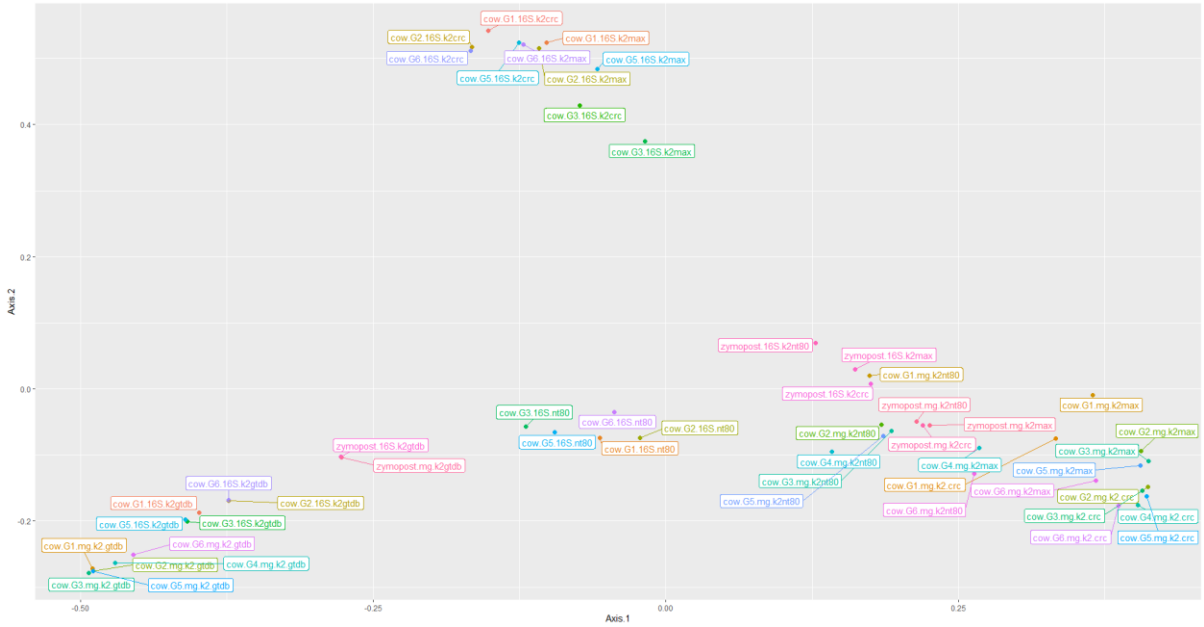

Figure 2.7. A PCoA of all control samples, using both 16S and metagenomic sequencing and all four databases. Samples are labelled by control type, sequence type and database. Mg is metagenomics sequencing; max is Maxikraken database; zymo is the ZymoBIOMICS mock community, names are shortened to fit.

2.3.3 Comparison of classification using Illumina 16S rRNA gene sequencing

The number of 16S rRNA reads that were returned from University of Auckland Illumina sequencing averaged approximately 1,024,000 reads per control sample, with one outlier sample, cow G1 yielding approximately 10,000 reads (**Table 2.3**). For this reason, cow G1 was not used in downstream analyses.

Sample name	Sample type	DNA concentration (ng/µl)	DNA concentration after 16S rRNA PCR (ng/µl)	16S rRNA gene read counts
cow.G1	Technical replicate	47.4	4.44	10,863
cow.G2	Technical replicate	29.1	11.4	824,623
cow.G3	Technical replicate	24.4	11.4	1,729,125
cow.G4 plus 1µl of spike control (D6320)	Technical replicate plus spike control	33.4	11.8	1,101,939
cow.G5 plus 10µl of spike control (D6320)	Technical replicate plus spike control	23.0	14.8	1,438,189
cow.G6 plus 100µl of spike control (D6320)	Technical replicate plus spike control	24.1	10.1	779,729
ZymoBIOMICS mock community (D6300)	Mock Community	15.6	10.9	1,287,842

Table 2.3. Control samples used for Illumina 16S rRNA gene sequencing, including DNA concentration before and after PCR and the total number of sequencing reads that were returned for analysis.

2.3.3.1 DADA2 retains high read numbers with difficulty in identifying specific genera

The number of reads retained throughout this pipeline is relatively high. The mean for the control samples is 63% ± 2.4, excluding G1 due to being significantly lower quality **(Table 2.4)**. However, classification of reads to specific genera is low **(Figure 2.8)**.

In the ZymoBIOMICS mock community, seven of the eight genera found in the control sample are seen, however *Salmonella* is not classified or potentially is classified as 'NA'. *Escherichia* is over-represented, while *Bacillus* is under-represented **(Figure 2.8)**. This shows that while the correct genera are being classified, the proportions are still skewed. While classifications are close to expected, it can be hard to identify specific ASVs. DADA requires Phyloseq to make meaningful conclusions and while it does classify a high proportion of reads in ASVs, there are still unidentified genera from the mock community.

Sample	input	filtered	denoised	Non-chimera	% reads kept
cow.G1	3573	1209	762	730	20
cow.G2	637,599	619,499	604,768	414,486	65
cow.G3	1,315,156	1,279,145	1,256,674	803,213	61
cow.G4	877,498	858,047	838,958	546,299	62
cow.G5	1,142,831	1,116,880	1,091,767	640,640	56
cow.G6	610,130	596,426	581,790	376,971	61
ZymoBIOMICS	1,053,723	1,019,991	1,018,144	780,194	74

Table 2.4. The number of 16S Illumina reads retained at each step through the DADA2 pipeline.

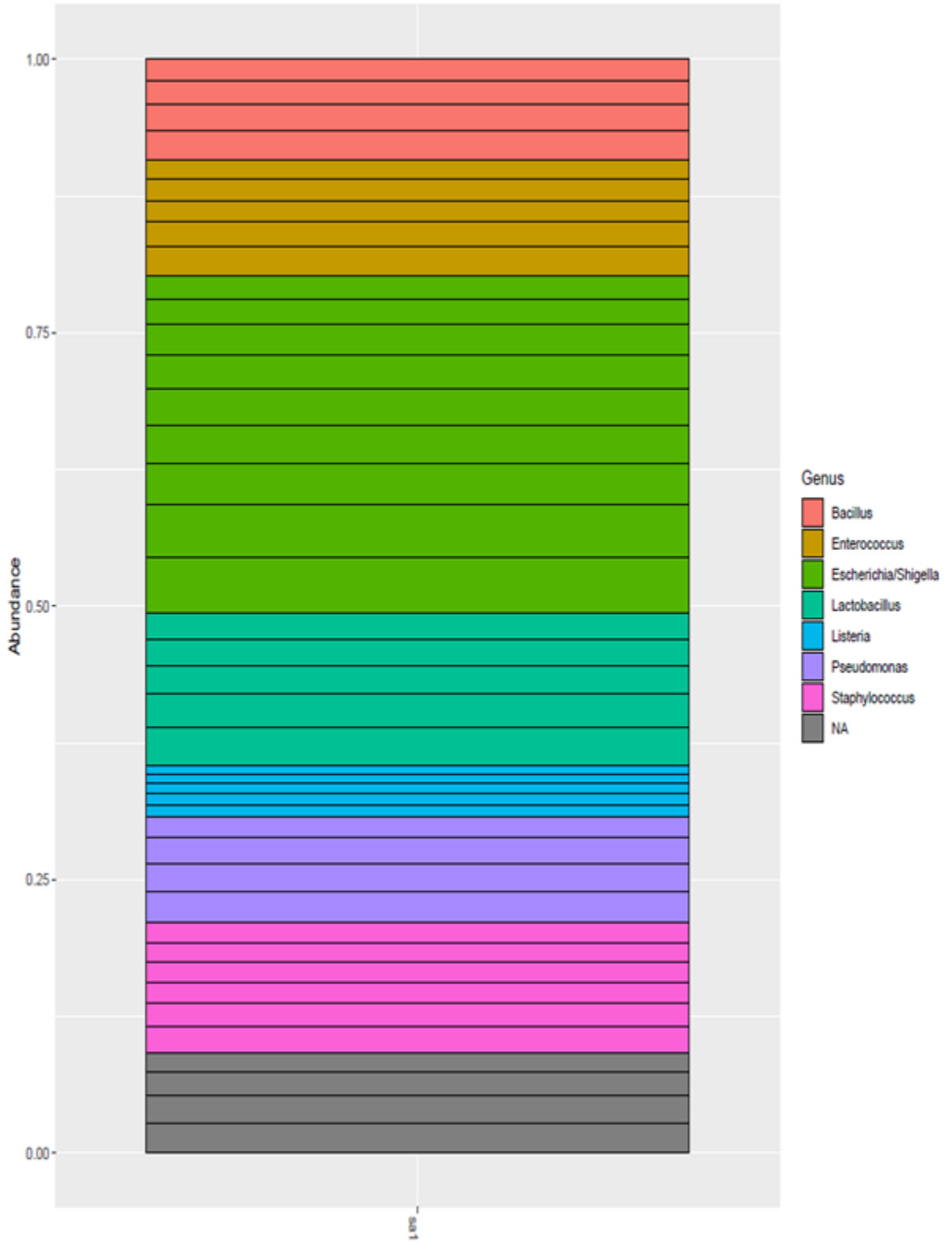

Figure 2.8. The proportions of all genera classified using DADA2, visualised in Phyloseq. Lines indicate the number of ASVs present, size between segments is affected by the sum of read counts.

The technical replicates show that there is three main genera present in the sample and a high component of unclassified reads **(Figure 2.9)**. Except for cow.G1, the technical replicates are very similar which shows that DADA2 is reproducible. This is likely because cow.G1 has very low read numbers.

With DADA2, the number of unclassified reads is similar to the most classified genus. About 90% of the genera are at very low levels **(Figure 2.9)**. This is expected within microbiome studies as there are a large number of bacteria present within the environment. It was expected that there would be more than three genera to be present in significant numbers due to having an average of 65% of reads kept within the pipeline **(Table 2.4)**. Nevertheless, the three named genera are all likely to be present in the cow faeces. *Bacteroides* are commonly found in cow faeces (Méric, Wick, Watts, Holt, & Inouye, 2019). *Prevotellacaeae* is known to be affected by the level of fibre within the diet, and thus is variable (Kim & Wells, 2016). UCG-005 is potentially *Ruminococcaceae_UCG-005* (Li et al., 2020). All three are known genera which shows the DADA2 is classifying genera which are expected in this microbiome.

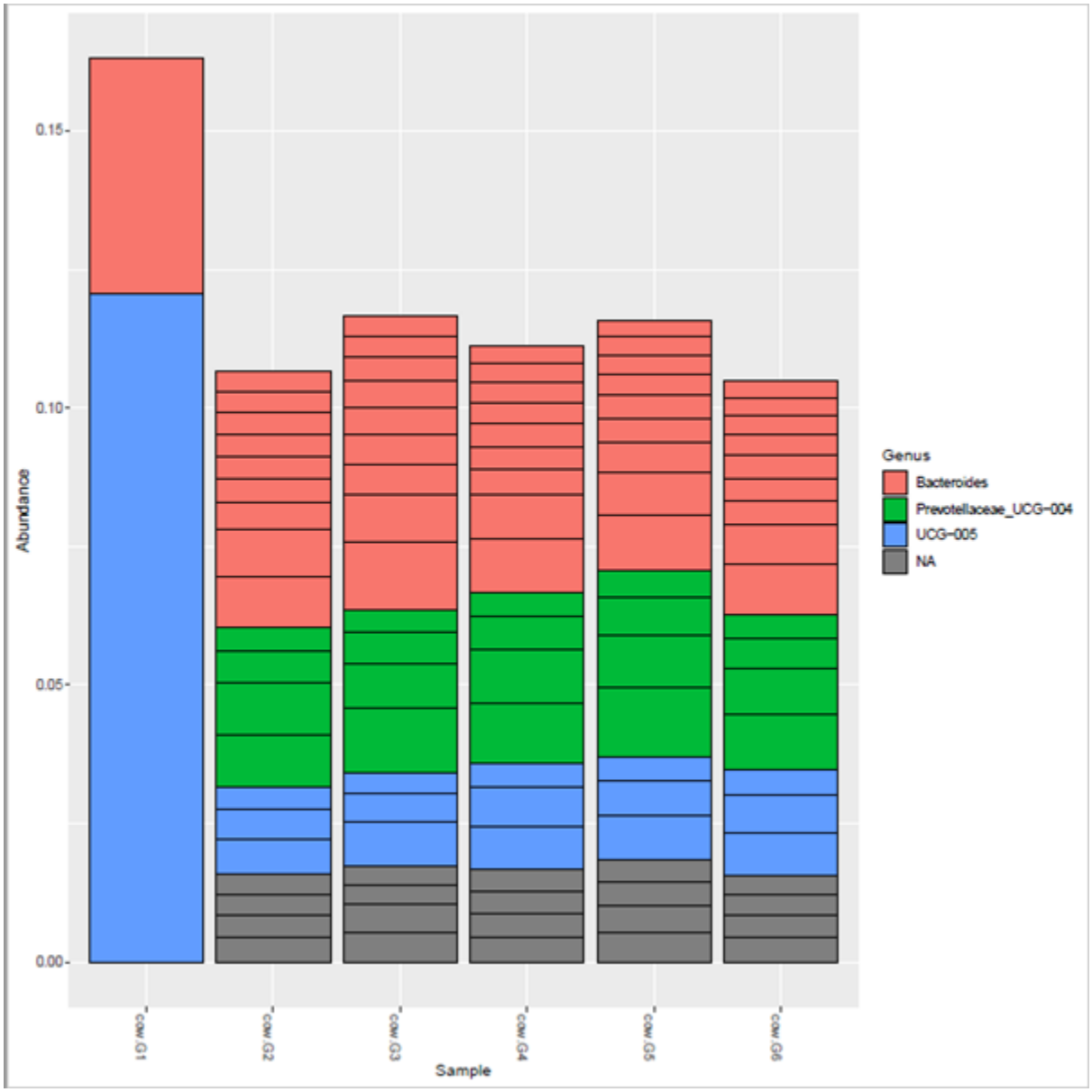

Figure 2.9. A bar chart showing the top genera of the cow replicates classified by DADA2, visualised in Phyloseq.

2.3.3.2 UPARSE retains a small percentage of reads compared to the other two pipelines

Using UPARSE, 13.9% ± 0.78 of reads per sample could be classified. This is much lower than expected **(Table 2.5**) and lower than obtained with either DADA2 **(Table 2.4)** or UNOISE **(Table 2.6)**. We hypothesize that this may be due to a large number of OTUs having been binned into unique categories due to sequencing errors that have not been accounted for. However, such low percentages classified has been used within papers (Guo et al., 2020; Ritter et al., 2019). It is interesting to note that the ZymoBIOMICS mock community has the lowest number of reads classified, despite also having the lowest variation within the sample. This could potentially be because it is not classifying down to the genus level (McGovern et al., 2018).

Sample	Read Count	Total Classified (%)
ZymoBIOMICS control	26,256	2.5
cow.G1.19dec2019	401	11.2
cow.G2.19dec2019	106,434	16.7
cow.G3.19dec2019	192,071	14.6
cow.G4.19dec2019	118,107	13.5
cow.G5.19dec2019	171,253	15.0
cow.G6.19dec2019	78,626	12.9

Table 2.5. The number of 16S Illumina reads classified by USEARCH-UPARSE, and the percentage of total reads classified.

The precision of the UPARSE tool was evaluated using the ZymoBIOMICS mock community. The observed frequencies of each genus were compared to the expected values and found a low correlation. We believe this is, as mentioned above and seen in **Table 2.5,** due in part to the UPARSE tool using a small subset of the data for classification. UPARSE classifies 2.5% for the mock community, and generally less than 10% for other samples. It also does not take sequencing errors into account. UPARSE showed low reliability in terms of percentage of reads that could be taxonomically identified, as the log transformed ratio shows high variation **(Figure 2.10)**. If it was classifying accurately, the ratio in **Figure 2.10** should be zero.

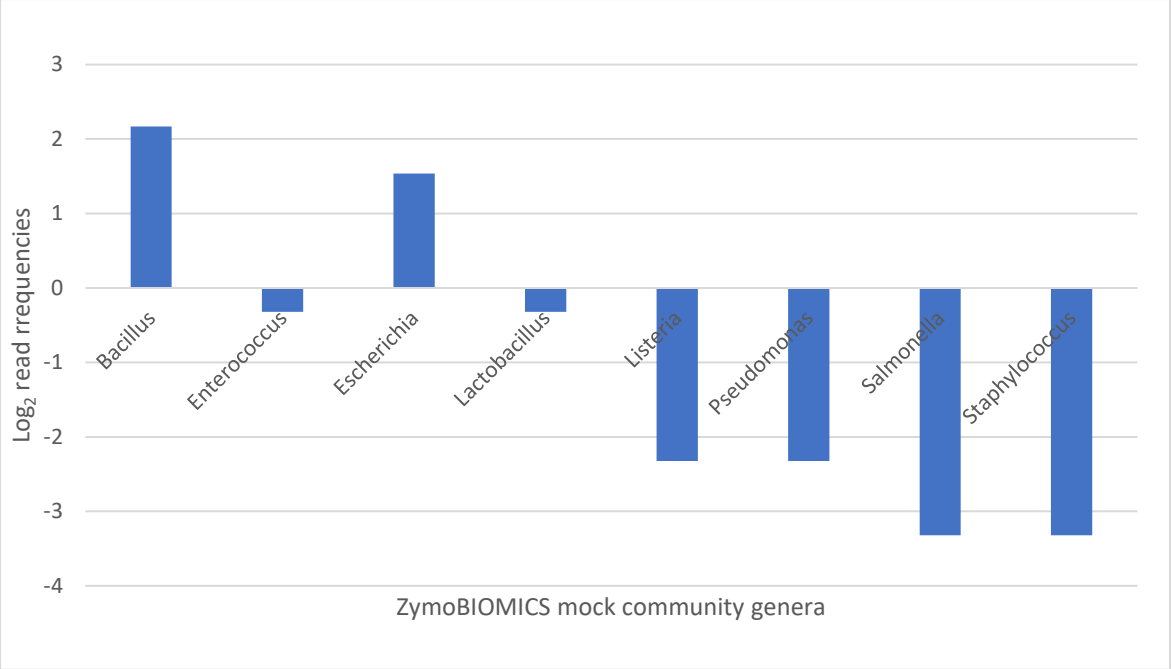

Figure 2.10. The ratio of observed read frequencies from UPARSE divided by the expected ZymoBIOMICS read frequencies, then log2 transformed to be centred around zero.

2.3.3.3 UNOISE retains a third of sequencing reads with a higher classification rate

UNOISE classified approximately a third of sequencing reads **(Table 2.6)**. There was a good correlation between the expected and observed read count frequencies for the ZymoBIOMICS mock community control sample **(Figure 2.11)**. However, an important factor to note is that the ZymoBIOMICS control classified more reads than those that were present in the original file. It is unknown why over 100% was classified **(Table 2.6)**. A possibility is that reads are classified more than once due to their high degree of similarity.

The read classification of cows is nearly identical showing that the method is highly reproducible **(Table 2.6)**. Therefore, the classifications from UNOISE are known to be relatively accurate as there is low amount of variation caused by the classification tool.

Sample	Read Count	Total Classified (%)
ZymoBIOMICS control	1,424,834	Over 100%
cow.G2	221,069	34.0
cow.G3	463,485	35.0
cow.G4	311,971	35.0
cow.G5	399,149	35.0
cow.G6	215,376	35.0

Table 2.6. The total 16S Illumina reads classified by USEARCH-UNOISE and the total percentage of reads classified.

The classification of the mock community by UNOISE is more accurate than UPARSE **(Figure 2.11)**. The ratios are more closely centred around zero overall. This means that there is have a higher level of confidence with the classifications.

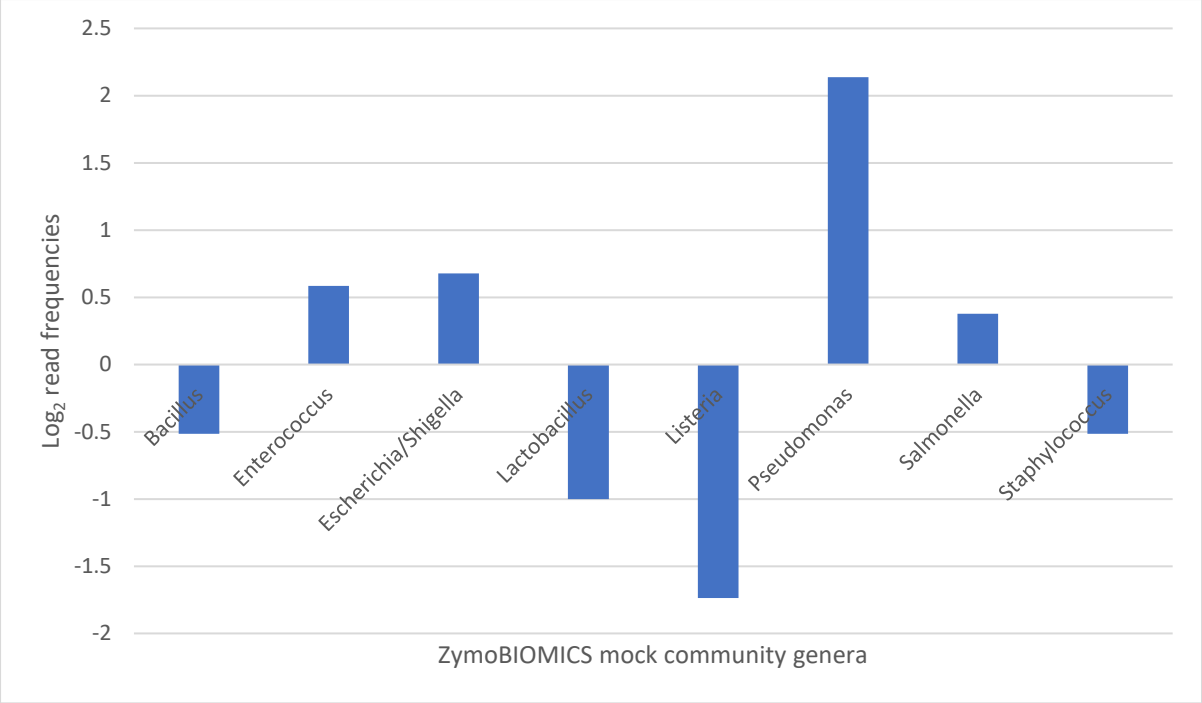

Figure 2.11. The ratio of observed read frequencies from UNOISE divided by the expected ZymoBIOMICS read frequencies, then log transformed to be centred around zero.

The UNOISE tool showed better correlation than the UPARSE tool in terms of expected frequencies for the mock community control and was able to taxonomically call a higher proportion of reads. The better correlation is likely because the UNOISE tool takes into account sequencing errors. This tool was used to taxonomically identify the read count frequencies at the genus level of the cow technical replicates**.**

A heatmap was used to compare how well genera were classified across the cow technical replicates. Only genera which had an average read count above 100 were used, to minimise the number of spurious classifications. However, this may mean that rare genera will not be found within the samples.

There is similarity between the technical replicates (cow.G2-G6) and could a generally identifiable increase in the genera *Imtechella* and *Allobacillus* in samples cow.G4, cow.G5, and cow.G6 which were spiked in at a 10x increase in each of these samples **(Table 2.2).** This was used to show if both gram-negative and gram-positive bacteria are being classified equally.

The general abundances for each genus are similar across the technical replicates, as shown in the heatmap in Figure 2.12. This indicated there was generally consistent DNA extraction

for samples and that UNOISE provides consistent taxonomic identification results **(Figure 2.12)**. The number of reads that are ‘high’ (green) is much smaller than the ‘low’ numbers of reads. As the read counts get lower, the percentages become less consistent, and minor changes in read counts can show more variation. However, generally there are obvious patterns of genera classification. Most genera have very low concentrations. These are where the same genera are classified across the replicates, which is to be expected, as these are from the same sample. The fact that the classification is not identical between replicates could be due to the number of reads placed into each zOTU.

Microbiomes have many bacterial genera within them hence there was expected to be a large number of low-level genera classification within the cow replicates. A few genera in the microbiome are potentially more abundant than the others, which could be due to a multitude of factors. These factors include the timing of sample collection or animal health (Holman & Gzyl, 2019; Li et al., 2020). However, it could also be caused by classification bias from the method (Degnan & Ochman, 2012). Mock communities or replicates are used to minimise the amount of method bias.

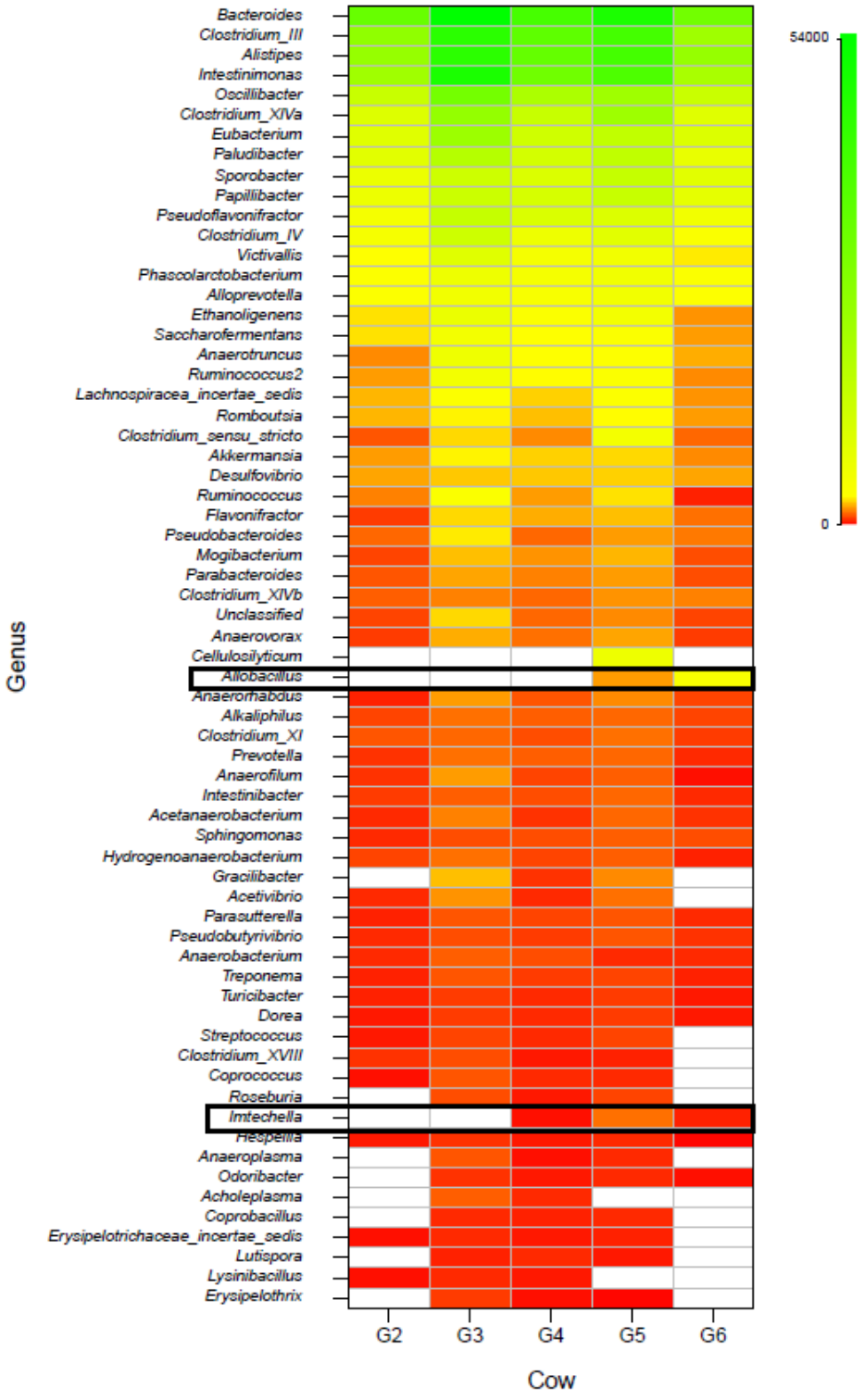

Figure 2.12. A heatmap of read frequency in technical replicates G2-G6, illustrating the relatively high similarity of genera classified. Reads were normalised by taking the number of reads for each genus and dividing by the total number of reads.

2.4 Discussion

In this chapter, a ZymoBIOMICS mock community and six technical replicates from a cow's faeces were used to compare bioinformatic tools in classifying samples. 16S rRNA gene and metagenomic Nanopore sequencing were compared, as sequencing methodology can have a large impact on classification (Lear et al., 2018; McLaren et al., 2019). Two classification tools, Kraken2 and Krakenuniq (Breitwieser et al., 2018; Lu & Salzberg, 2020) were also compared to understand the impact of how tools affect classification. Finally, four databases of varying size and specificity were compared. These were Maxikraken (Derrick Wood & Nick Loman, 2018), NT80 (NCBI nucleotide database with 80 additional eukaryotic genomes), CRC (a refseq database) and GTDB (Parks et al., 2019).

The same mock community and technical replicates were used to compare classification pipelines for 16S rRNA gene Illumina sequencing. This was because Illumina sequencing is known to be more accurate, although 16S rRNA sequencing methodology can only classify bacteria to the genus level. Three different pipelines were compared; DADA2 (Callahan et al., 2016), USEARCH-UPARSE (Edgar, 2013) and USEARCH-UNOISE (Edgar, 2016) to classify and identify which method was the most accurate to the mock community and had a high degree of similarity through the replicates.

2.4.1 Classification tool had no significant effect on Nanopore data

It was found that the classification tool used - either Kraken2 (Wood et al., 2019) or KrakenUniq (Breitwieser et al., 2018) - had no significant effect on classification. These two tools have been created by members of the same research group, at Johns Hopkins University. This could have an impact on the similarity of the tools, and perhaps other, unrelated tools could have greater differences. However, as databases need to match the tool designed, it can be more time-consuming to set up multiple databases and then test tools. Due to this incompatibility of tools between database structure, it can be difficult to change to a new classification tool. This can have a negative influence on the research as researchers continue to use older, potentially outdated tools. While in this study, there is no difference between Kraken2 and Krakenuniq, this is not to say that there is no difference between all tools. Tools may have a larger influence on classifying microbes than what was found in my work. This is one limitation of the classification comparisons which uses just the two tools.

2.4.2 Spike-ins are less informative than mock communities

Whilst it was found that mock communities and technical replicates are highly useful for comparing methods and pipelines, the spike-in that was used (Zymo Research Corporation, 2019) was less informative. Spike-ins need to be reliably quantified and easily distinguished from the environmental genera or there will be an abundance of false positives (Venkataraman et al., 2018). It was found that as there are only two genera present in this spike-in, it was only classified at low levels or not at all. *Imtechella* is present in increasing amounts as expected with metagenomics, however with 16S rRNA gene sequencing it was only in the sample with the highest spike. *Allobacillus* was not present in either sequencing methodology.

This low level of classification by the pipelines makes it hard to investigate where the spike is lost and if it is a problem with the spike-in or with classification methods. This contrasts with Venkataraman (2018) who found spike-ins generally useful for metagenomic sequencing; validation of experiments and to measure technical variation (Hardwick et al., 2018; Venkataraman et al., 2018). We suggest that spike-ins are not necessary for microbiome experiments, where there are many factors in classification and a high level of microbial diversity. The small number of genera in a spike-in means that they can easily be missed, and bias is difficult to allocate to a particular factor. When they are not found at all, this could be due to reference genomes not being present in the database, low sequencing depth or other factors (Hardwick et al., 2018). Due to the mock community having a wider variation of genera they are more informative than spike-ins as to where information is lost and the skew from expected.

2.4.3 Database quality has a significant effect on classification

The quality of classifications is highly influenced by the database used. The main reason for this is that the database output can only be as good as the reference sequences it contains (Méric et al., 2019). This means that the database should include all genera that are expected; and that genomes are as complete as possible (Han et al., 2020; Parks et al., 2018). Including draft genomes has been proposed to create a higher classification rate, however, this needs to be balanced against the fact that drafts may be of a lower quality (Méric et al., 2019; Robeson et al., 2020). There is no way of standardizing the databases used, as having all species in one place would be too large and unwieldy. It takes time to download and set up, and so smaller, targeted databases fill this gap.

There is an inherent difficulty and bias in assigning reads with small databases, for example the nt80 database that I used. Therefore, there has been a large push to increase database

size; the range of genera in the database; and creation of custom databases to include specific genera (Han et al., 2020; Méric et al., 2019; Parks et al., 2018). The false positive rate may increase as samples from genera that are not present in a reference database may match part of some other genome scaffold and hence be classified incorrectly (Edgar, 2017). There are likely to be some false positives through every method and database. Quantifying the number of false positives can be difficult (Nearing et al., 2018). Further research is being undertaken around how much bias is caused by the database (Kozlov, Zhang, Yilmaz, Glöckner, & Stamatakis, 2016). It was found that the database used did influence classification of the microbiome, as some clustering was apparent for each database. Some targeted databases do not classify some genera, potentially due to not having a reference sequence for that microbe.

The GTDB attempts to standardize the names of genera (Méric et al., 2019). This is useful for closely related genera; however, the names sometimes do not match other databases. Non-matching nomenclature creates difficulty in comparisons between databases. Whilst this is not an issue when using one database, it can limit GTDBs use in multi-database studies (Breitwieser et al., 2019; Robeson et al., 2020). In smaller microbiome studies it is recommended using one large database to classify against rather than a smaller targeted database or multiple databases which can show large variation. Researchers would need to be aware of the limitations of their decision, such as bias may not be recognised. This can cause inaccurate classification.

Setting up a new database for a specific microbiome experiment is time consuming and inefficient. It is easier and more efficient to use pre-set databases and means that you can compare to classifications from other research (Han et al., 2020). Researchers could create a database that would be useful for long-term, specific research. There are many more databases that are open-source and easy to use. There are also database managers which enables reproducibility (Robeson et al., 2020).

2.4.4 16S and metagenomics sequencing affect Nanopore classification

When trying to describe communities that have a mix of gram-positive and gram-negative bacteria, there is a high potential for bias from what is expected which can be caused by many aspects of the classification process (Han et al., 2020; Pollock et al., 2018; Santos et al., 2020). The database has the biggest effect, which could be due to how well read sequences match genomes within the database. The GTDB is a large database that matches using average nucleotide identity. This is likely why it works best with metagenomics sequencing (Méric et al., 2019; Parks et al., 2019). The differences between 16S and metagenomic

sequencing are likely due to the difference in read length and therefore how it matches genomes in the database.

2.4.5 16S classification pipelines vary in accuracy classifying Illumina data

DADA2 is one of the most widely used 16S classification pipelines (Nearing et al., 2018). Using Illumina data, we found that the read percentage classification was best with DADA2 and other studies found that it identified the most amplicon sequence variants (also known as zOTUs) (Nearing et al., 2018; Prodan et al., 2020). We found, as others did, that DADA2 has a high level of sensitivity, even with low read counts (Prodan et al., 2020). However, this may decrease the specificity which in turn could increase the false positive rate in the technical replicates (Nearing et al., 2018; Prodan et al., 2020). In this study DADA2 has a lower level of reliability in genera classified compared to other pipelines. This created difficulty in identifying specific genera in the replicates despite both Nearing (2018) and Prodan (2020) using it.

UPARSE had the lowest number of classifications, which could be in part due to the way that UPARSE creates OTUs. Prodan (2020) states that UPARSE aims to explain a given input sequence starting from sequences in a database, using the fewest possible number of events, then each input is compared to current OTUs and assigned if it is more than 97% identical to an existing OTU. The author of UPARSE states that UPARSE was designed to be very specific, which is potentially why the read classification rate is very low (Edgar, 2013). In this experiment UPARSE classified a very low number of reads. Low read counts have been used in other studies, so further analysis could be possible (Guo et al., 2020; Ritter et al., 2019). However, with the other pipelines classifying a higher percentage of reads, UPARSE was not continued with.

While UNOISE classifies a third of the replicate reads, it also classifies over 100% of ZymoBIOMICS reads. This could be due to the high similarity of the sample and therefore some reads being classified twice. This further highlights the reason for using mock communities as a control, as it shows how well genera can be classified in a known sample. However, we found that classification using the UNOISE pipeline produces results that are both consistent with the expected frequencies in the ZymoBIOMICS mock community sample and have the same pattern of bacterial diversity between technical replicates of the cow faecal samples. This means that there is a relatively high level of confidence in these faecal classifications. Similar results were found by Prodan (2020) who states that UNOISE is accurate. Perhaps the low classification rate of replicates could be due to UNOISE discarding zero-radius OTUs (zOTUs) with abundance <8 using default parameters (Edgar, 2016; Prodan et al., 2020). This means that genera with low abundances, such as MAP, may be discarded.

In a microbiome there are many genera that are present in low abundances which would be discarded by these pipelines.

2.4.6 Technical replicates show that classifications are reproducible

Additionally, there are patterns in the technical replicates in which certain genera are present at high levels in four of the five replicates (with over 100s of reads). However, in the other replicate these genera have zero reads. We hypothesize this might be due to UNOISE parameters that are set too high in terms of grouping reads into zOTUs. As such, this could be fixed by lowering the number of reads needed to classify into an zOTU. Like the other two pipelines, UNOISE does not classify all taxa which Nearing (2018) also found, albeit their work was with rarer, known genera.

2.4.7 Summary of Illumina classification

All control samples were sequenced using Illumina 16S rRNA gene sequencing and compared three different methods, USEARCH-UPARSE (Edgar, 2013), USEARCH-UNOISE (Edgar, 2016) and DADA2 (Callahan et al., 2016) in order to understand the effect that 16S rRNA gene classification pipelines have on classifying the microbiome. While there are many components affecting pre-processing, the bioinformatics method that is used to cluster 16S rRNA Illumina reads is a large determining factor (Nearing et al., 2018). As such, a tool that has a higher resolution is more important to determine the microbiota which is present within a microbiome sample. To do so, we used the tool that most closely matched the ZymoBIOMICS mock community sample and had a high correlation of genera between technical replicates. This shows high reproducibility, which is important when classifying samples. Therefore, we have chosen to use UNOISE on uncharacterized effluent samples from Livestock Improvement Corporation.

Chapter 3: Classifying microbial communities in effluent

3.1 Introduction

All microbiomes have a diverse range of microbes. There are many methods to try to classify this diversity and the chosen methodology has a large influence on classification accuracy (Bharti & Grimm, 2019). New techniques, tools and protocols are created when needed, which rapidly changes the landscape of analysis (Bharti & Grimm, 2019; Pollock et al., 2018). Currently there is no 'gold standard' method to follow in microbiomics.

In Chapter 2 the accuracy and reproducibility of three Illumina 16S microbiome analysis methods were validated using both a mock community standard and technical replicates. In this chapter we investigate microbial communities in bovine effluent samples from Livestock Improvement Corporation (LIC) using the best Illumina 16S rRNA pipeline from Chapter 2. This best pipeline was UNOISE, as it was robust across the mock community and technical replicates.

Effluent represents the pooled faecal matter of the herd and is a convenient source of material for examining the herd's microbiome. Specifically, the goal is to attempt to predict herd health by assessing both microbial diversity and abundance present in herd effluent. We aim to determine if this diversity varies between samples with or without a known pathogen - *Mycobacterium avium* subspecies *paratuberculosis* (MAP) - for which LIC has qPCR data on its presence in specific samples.

To address this aim, we first needed to identify sampling locations and sample types that yield enough DNA for downstream sequencing and informative data. This study then investigated the microbial composition of these effluent samples and attempted to identify specific genera. If bacterial abundance and diversity change significantly over time alongside MAP presence, this is an indication that effluent analysis can be used to monitor herd health. Further steps can then be undertaken to identify affected individuals and provide treatment.

3.1.1 Monitoring effluent to assess a bovine herd microbiome

The microbiome can offer insights into the health of an organism. Many studies have investigated microbial species abundance and diversity in cattle rumen (Holman & Gzyl, 2019; Romagnoli, Kmit, Chiaramonte, Rossmann, & Mendes, 2017; Zeineldin, Aldridge, et al., 2018a). It is known that the bovine microbiome is affected by many factors, such as diet, age, and individual health. However, there is a high cost associated with assessing the microbial

composition of individual cows (Aly et al., 2012; Britton et al., 2016). Current research is investigating cheaper, easier methods to assess the bovine microbiome.

Effluent has been identified as one sampling method where bovine herd microbiomes can be monitored (Li et al., 2020; Zeineldin, Aldridge, et al., 2018a). Monitoring herds effluent system for shifts in microbial composition can indicate if there are animal health issues which may need to be investigated further. Identification of specific microbial profiles that may be correlated with the presence of pathogenic microbes can be useful for monitoring the health of dairy cows. Unhealthy cows can affect production causing economic loss. Long-term, identifying the effluent microbiome has implications for tracking how the faecal microbiome changes and the effects on cow health. We specifically looked for differences in the microbiome which may be correlated with *Mycobacterium avium* subsp *paratuberculosis* (MAP) as it is known to negatively affect animal health and dairy production (Bates et al., 2018; Chi, VanLeeuwen, Weersink, & Keefe, 2002).

3.1.2 Samples from LIC

The samples were provided by LIC from a dairy farm in Waikato, New Zealand. Samples were taken from four distinct locations within the dairy effluent system: ‘wedge’ (also known as sand-trap – part of the effluent system just before the pond) **(Figure 3.1)**, ‘pond’ (part of the effluent pond), ‘irrigation’ (the irrigation filter) and ‘new pile’ (fresh cow faeces). This provides a range of locations to identify which would be best to sample the herd microbiome.

Figure 3.1. Image taken of the wedge (sand-trap) as part of the effluent system on a Waikato dairy farm. Photo provided by LIC.

Each location was sampled over several time points from October 2016 to March 2017 **(Appendix A)**. All samples were stored at -80ºC until DNA isolation per the suggestion of Choo et al (2015). qPCR data from LIC is available for some samples, showing the presence of MAP **(Appendix A)**.

3.2 Materials and methods

3.2.1 Overview

66 bovine effluent samples were examined. Eight were technical replicates that had DNA independently isolated and sequenced to provide internal controls to evaluate reproducibility **(Appendix B)**. Technical replicates were denoted by A or B as the fourth character within the unique sample ID. Several locations were used to determine if one location has more predictive power for indicating the presence of MAP.

3.2.2 DNA isolation

The QIAGEN Powersoil DNA extraction kit, following manufacturer's instructions, was used to isolate DNA from all samples **(Appendix B)** (Vo & Jedlicka, 2014).

3.2.3 Illumina 16S rRNA Gene Sequencing

Samples were sequenced using 16S rRNA Illumina sequencing **(Appendix B)**. For the first eleven samples the methods in the Illumina "16S Metagenomic Sequencing Library Preparation" guide were followed to amplify the 16S rRNA gene region from these samples up to the Index PCR step (Illumina, 2013). Specifically, for samples with low DNA concentration and less than 12.5 ng of input DNA, the maximum volume of 2.5 µl of DNA was used. The DNA from the remaining samples was added in the amounts according to the manufacturer's instructions. This sample set was sequenced by Auckland Genomics, at the University of Auckland. A further 55 samples were amplified and sequenced by an overseas service provider. All reads were analysed and trimmed using FastQC (version v0.11.9) (Andrews & Others, 2010). The forward and reverse reads were joined using fastp-join (version 0.20.0) (Chen et al., 2018).

3.2.4 Taxonomic classification

The taxonomic classification of the effluent samples was investigated using USEARCH-UNOISE (version 11) (Edgar, 2016) as it performed best with the ZymoBIOMICS mock community and technical replicates in Chapter 2. UNOISE uses denoising to create zero radius OTUs (zOTUs, also known as amplicon sequence variants). zOTUs are denoised sequences that can be used for diversity analysis. They are similar to classical OTUs however the results are analysed slightly differently (Edgar, 2016). The UNOISE algorithm performs error-correction, so that any errors introduced from sequencing are taken into account as zOTUs are formed. This minimizes incorrect zOTUs from being created.

The same processing steps of filtering and de-duplication were used with the effluent samples as in Chapter 2. The forward and reverse reads were merged into one read and these merged reads were then quality checked. The USEARCH-UNOISE pipeline was used to cluster reads into zOTUs, made a zOTU table and align the zOTUs to the RDP database (Edgar, 2016; Maidak et al., 2000). One change from default parameters was that the minimum abundance of genera within zOTUs was reduced from 8 to 5. The rest of the in-house pipeline converts the files into a usable format for import and analysis in R (R Core Team, 2020). This is implemented in Snakemake (version 5.10.0) (Köster & Rahmann, 2012) and has been documented in GitHub **(S3.1)**.

3.2.5 Visualising classifications

Overall frequencies of genera were used to enable comparisons across samples and account for any differences in read numbers. Overall frequencies were obtained by dividing read counts classified per zOTU by the total read count per sample. R (version 4.0.2) (R Core Team, 2020) was used to analyse and visualize read counts of classified taxa.

A PCoA was created through the vegan package (version 2.5.6) (Oksanen et al., 2019) with the ape dependencies (version 5.4.1) (Paradis & Schliep, 2019) to investigate variation between samples. The resulting plot was evaluated as to whether there is any clustering by sample location or date.

A canonical analysis of principal components (CAP) analysis through the BiodiversityR package (version 2.12.3) (Anderson & Willis, 2003; Kindt & Coe, 2005) was used to investigate if the overall effluent microbiome genera composition was affected by the presence or absence of MAP. This presence or absence was determined by LIC by a qt-PCR where MAP is amplified using PCR and the number of cycles required for the product to be produced at an amount over a detectable threshold (Cq value) is used to determine its starting abundance. MAP was defined as present if the Cq value was under 40 and absent if the Cq value was over 40.

3.3 Results

3.3.1 Powersoil protocol works across a range of effluent locations

Using the Powersoil protocol I obtained moderate amounts of DNA across a variety of samples. The mean of all DNA extractions was 8.3ng/µl ± 9.9ng/µl. New pile, wedge and irrigation filter samples had respective DNA extraction averages of 13.4ng/µl, 12.4ng/µl and 10.0ng/µl. The sand trap and pond sample averages were slightly lower at 7.1ng/µl and 7.3ng/µl.

3.3.2 Classifying LIC samples

The UNOISE pipeline was used due to its success with both the ZymoBIOMICS and cow control samples (Chapter 2). As such, LIC samples were analysed with a common bioinformatic pipeline as presented in **S3.1.**

3.3.3 Classification Visualisations

3.3.3.1 Principal Coordinate Analysis

Clustering indicates that the samples have a more similar microbiome, meaning a more alike composition of microbes (Zeineldin, Aldridge, et al., 2018a). There is no obvious clustering by location or sample type in the LIC samples **(Figure 3.2)**. This shows that there is a substantial amount of variation within sample type. Variation could be caused by the wide range of dates within the same locations.

As a broad generalization, the sand-trap samples tend to be on the middle of the left-hand side of the PCoA. However, there are new-pile samples within this cluster. The replicate samples of sand-trap are relatively similar. This shows that the pipeline is having a small effect on the overall classification.

Some new-pile location samples cluster in the top right-hand corner. It is inconclusive as to why this small cluster occurs. Five out of the six samples have a positive qPCR result for MAP, however one sample (newpile.S1.15dec2016) does not. It is also interesting that there is one very similar wedge sample, that does not match by date.

There is a distinct cluster in the bottom-right hand corner of the PCoA **(Figure 3.2)**. When investigating potential reasons for this cluster, it was found that it consists exclusively of samples sequenced at the Auckland Genomics Sequencing centre and that all samples sequenced there are in this cluster. This coincidence indicates that the clustering is unlikely to reflect any real biological variation. Instead, it may be caused by the fact that PCR was done within our laboratory for the Auckland Genomic sequencing, but the overseas provider did PCR for the second sequencing run. It could also be caused by the difference in sequencing depth from the two sequencing providers, as the Auckland Genomics sequencing had a much higher read number.

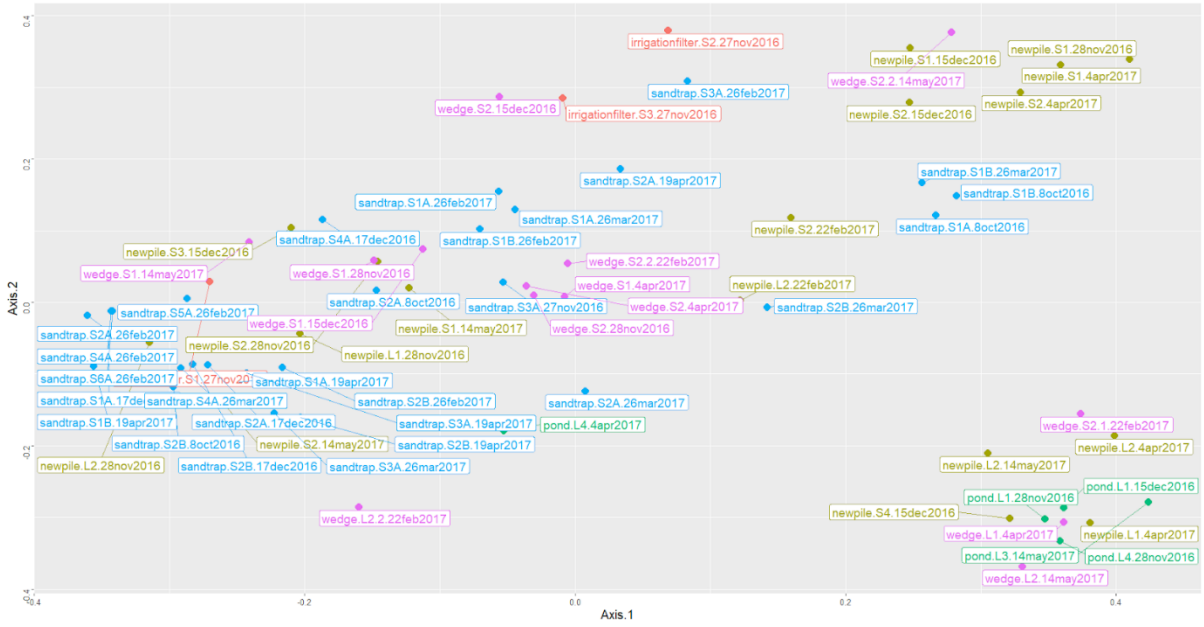

Figure 3.2. PCoA of LIC effluent samples, with frequencies of read counts used. Samples are coloured by location. Sandtrap is blue, wedge is purple, newpile is brown, irrigationfilter is red and pond is green.

A possible reason for the effect of sequencing provider on detected variation is because there is a better sequencing depth in samples from the Auckland Genomics sequencing (average read depth for Auckland Genomics is 498,700 reads, while the overseas provider is 78,800 average reads). This means the pipeline is more likely to identify rare taxa. The total read numbers from the overseas provider is lower than the ones from the Auckland provider **(Appendix B).** Therefore, a lower number of reads are classified and used to identify bacteria within the microbiome **(Appendix C)**. To remove the effect of sequencing providers, one PCoA is presented per provider **(Figure 3.3)**. When we investigate the providers individually in **Figure 3.3**, the clusters from **Figure 3.2** are removed. Samples sequenced by the Auckland provider have less variation than expected **(Figure 3.3A)**. It was expected that the three sample locations to be more distinct. There is a small grouping, where all four pond samples are similar to each other. As a generalization, the four newpile samples are on the bottom half of the PCoA. These newpile samples are closely intertwined with the wedge samples. Interestingly, samples on the 4th of April 2017 are similar, however this fact is negated by the samples taken on the 14th of May 2017 being dissimilar.

Samples that are sequenced by the overseas provider do have some small clusters **(Figure 3.3B)**. On the left-hand side there are mainly the sand-trap samples, which should all be highly similar due to the location and having some replicates. This is highlighted by the replicate ‘sandtrap.S1.8oct2016’in the top right corner. The day-replicates

‘irrigationfilter.S2.27nov2016’ and ‘irrigationfilter.S2.27nov2016’ are relatively close, however the third sampling site ‘irrigationfilter.S1.27nov2016’ is further away.

There is a small cluster on the right-hand side of **Figure 3.3B** which mimics the findings of **Figure 3.3A**, in that it is mostly made up of newpile samples with a few wedge samples mixed in. There are a few sandtrap samples that complete the cluster. As they are the same location, they should be similar to wedge samples. Of these seven samples, the two sandtrap samples have an unknown qPCR value. Four out of the other five have a positive qPCR value for MAP. This shows that there is a potential shift in the overall microbiome when MAP is present. It provides evidence that the proposed method for monitoring effluent is possible.

Samples tend to cluster by date. This is to be expected, as all samples were collected from the same farm. There are several examples of this clustering, such as sandtrap samples on the 19th of April 2017; wedge samples on the 28th of November 2016; and sandtrap samples on the 26th of February 2017. This is a good sign because the effluent herd microbiome should be near identical on the same date. It shows any effluent location could be used to monitor herd health.

Within larger clusters, small groupings of dates does occur, despite having different sampling locations. This shows that it is possible to associate trends in the microbiome, which can potentially identify patterns caused by external factors (i.e., different seasons) (Li et al., 2020). Long-term data would be able to show which patterns are caused by seasonal factors and which changes are caused by pathogens such as MAP (Fecteau et al., 2016).

The replicates are less similar than expected. Most replicates are very close together, showing a high degree of bacterial similarity. For example, ‘sandtrap.S1.26feb2017’ down the middle bottom of **Figure 3.3B.** Others have more distance, like ‘sandtrap.S2.8oct2016’ which is part of the middle left and upper middle left cluster. Like in Chapter 2, this shows that the pipeline is reproducible, with only little variation. It also indicates that there is a need for replicates to identify the level of bias caused by the pipeline and sequencing depth.

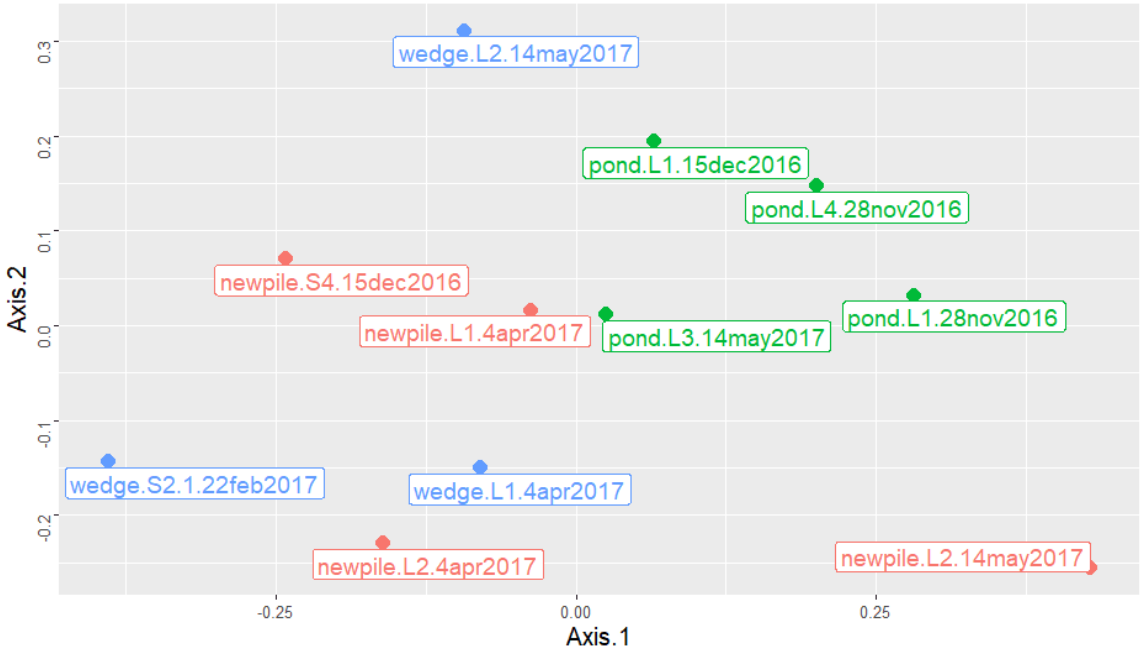

Figure 3.3A. PCoA of samples that were sequenced by the Auckland provider.

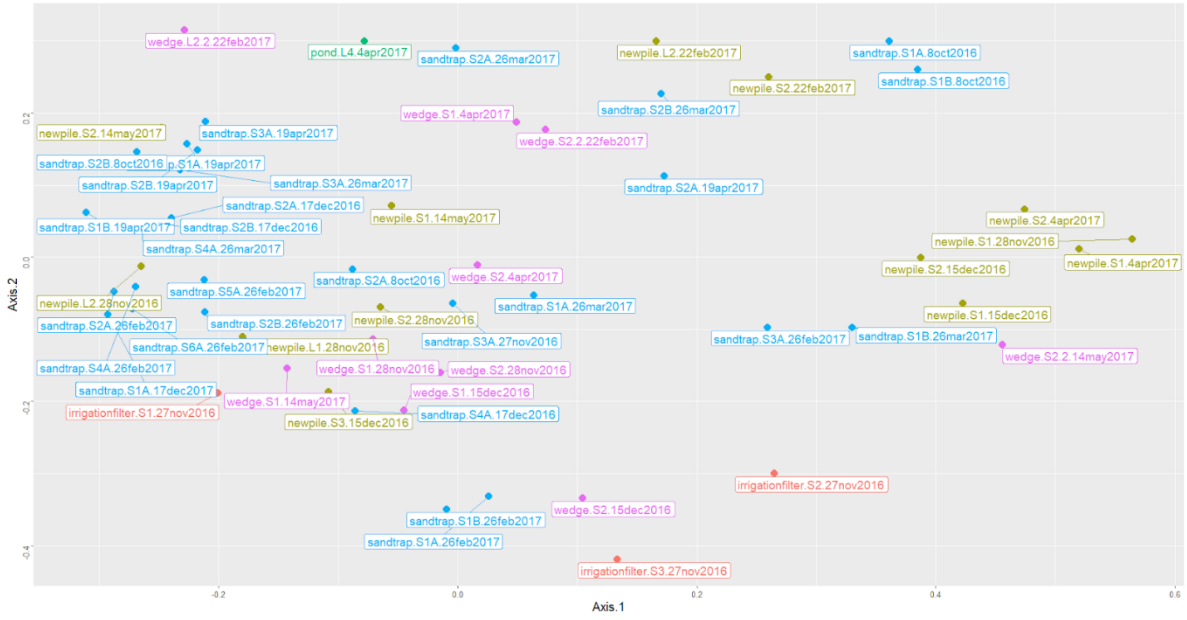

Figure 3.3B. PCoA of samples that were sequenced by the overseas provider.

3.3.3.2 Heatmap of replicates

For follow-up analysis and to examine reproducibility of microbiome characterization, we chose to use replicates to identify if the pipeline was reproducible. Replicates should show similar classification patterns, with high frequency genera should appear across multiple locations.

We chose to use nine pairs of samples that had DNA extracted twice and one pair of which was sequenced by both sequencing companies. We also examined a set of three samples taken on the same day from three different sites in the same location type (irrigation filter). The replicates were sampled at the sand-trap location type (ST) and the samples sequenced by the different sequencing providers were taken from the wedge location type (W). These replicates are found in **Table 3.1.**

There are fewer matching genera across all replicates than expected. This is likely to be due to low read counts, where a low depth of sequencing means that taxa with low abundances are hard to detect. This is especially noticeable in the batch sequences replicates (right-hand side of **Figure 3.4**), where the wedge sample with a much higher read count has a greater number of bacterial genera present. The three samples from different sites within the same location have genera present in varying abundances across the samples.

Most of the sandtrap replicates have similar compositions, with bacterial genera found across multiple samples. If there was a greater read depth, then it is likely that the no-data gaps are reduced.

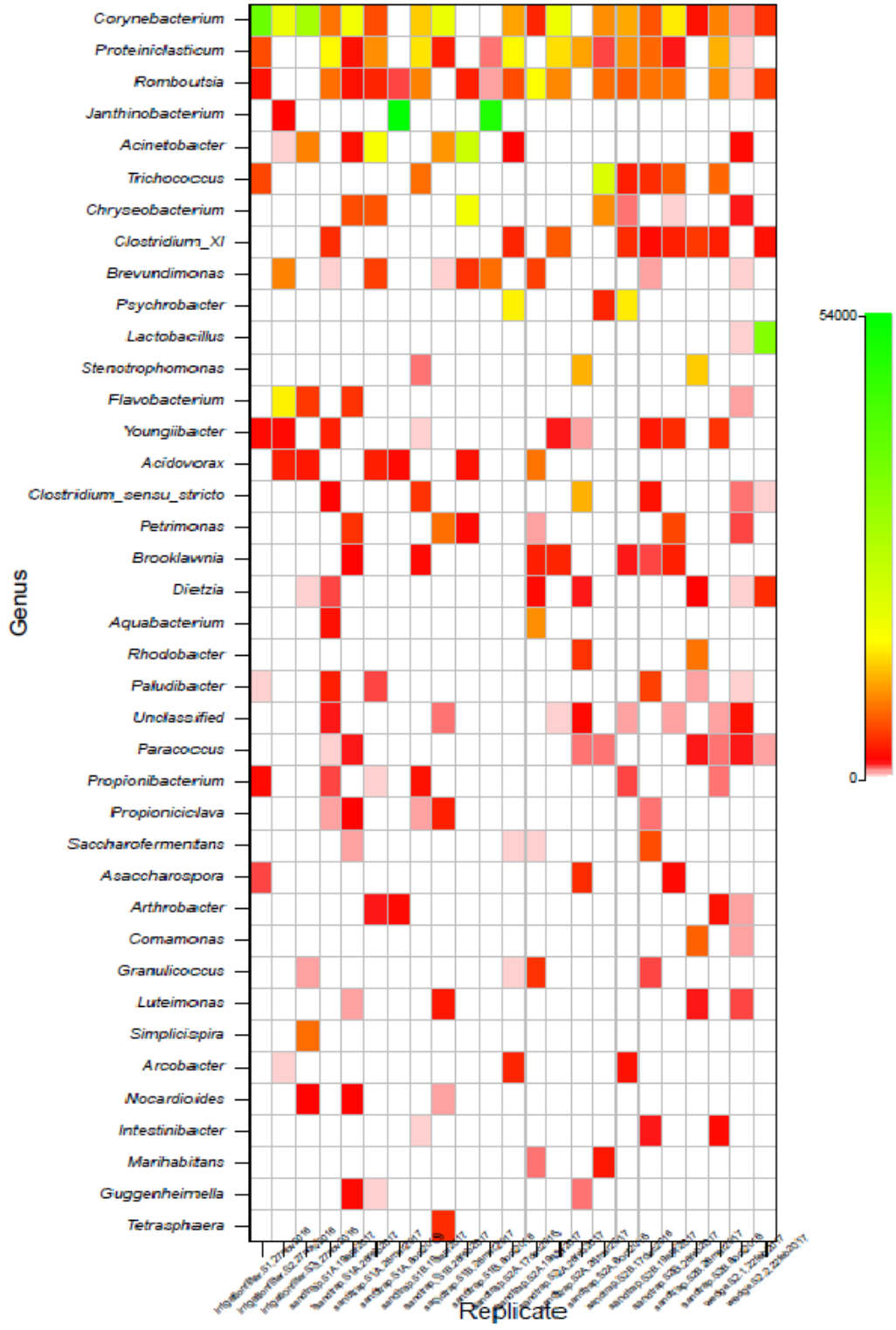

Figure 3.4. A heatmap in R of the taxonomic identification of biological replicates. Genera with a relative frequency of above 1.5% were included in the heatmap. Green means a high relative frequency of read count, white means not present or very low relative frequency read count, with shades of yellow to red in the middle.

Sample location	IF.S1-27-11-16	IF.S2.27-11-16	IF.S3.27-11-16	
DNA extraction	ST.S1A.8-10-16	ST.S1B.8-10-16	ST.S1A.26-2-17	ST.S1B.26-2-17
	ST.S1A.26-3-17	ST.S1B.26-3-17	ST.S1A.19-4-17	ST.S1B.19-4-17
	ST.S2A.8-10-16	ST.S2B.8-10-16	ST.S2A.17-12-16	ST.S2B.17-12-16
	ST.S2A.26-2-17	ST.S2B.26-2-17	ST.S2A.26-3-17	ST.S2B.26-3-17
	ST.S2A.19-4-17	ST.S2B.19-4-17		
Sequencing run	W.S2.1.22-2-17	W.S2.2.22-2-17		

Table 3.1. Sequencing replicates. These replicates are in order as found in Figure 3.4.

3.3.3.3 Canonical Analysis of Principal coordinates

CAP analysis shows a clear distinction between MAP being present and absent within the microbiome **(Figure 3.5)**. The samples which have MAP present within them are generally clustered on the left-hand side (green triangles). The samples where MAP is absent are clustered on the right-hand side (red circles). The sandtrap and irrigation filter samples were also plotted where the MAP qPCR value was unknown. It appears that all unknown samples fall in the 'MAP-negative' cluster, predicting that most are likely to be MAP free **(Figure 3.5)**. It is possible that the locations have an effect on clustering, but as there are MAP present and absent locations on both sides of **Figure 3.5**, we conclude that it is more likely that they are MAP-free.

There are a few MAP-present samples that segregate with the MAP-absent samples **(Figure 3.5)**. This shows that while MAP can cause a shift within genera present in the microbiome, there are cases when this pathogen has no effect on herd effluent. This could potentially be that MAP is at such low levels compared to other genera that it has little effect, or it is early in the infection and microbial dysbiosis has not yet occurred.

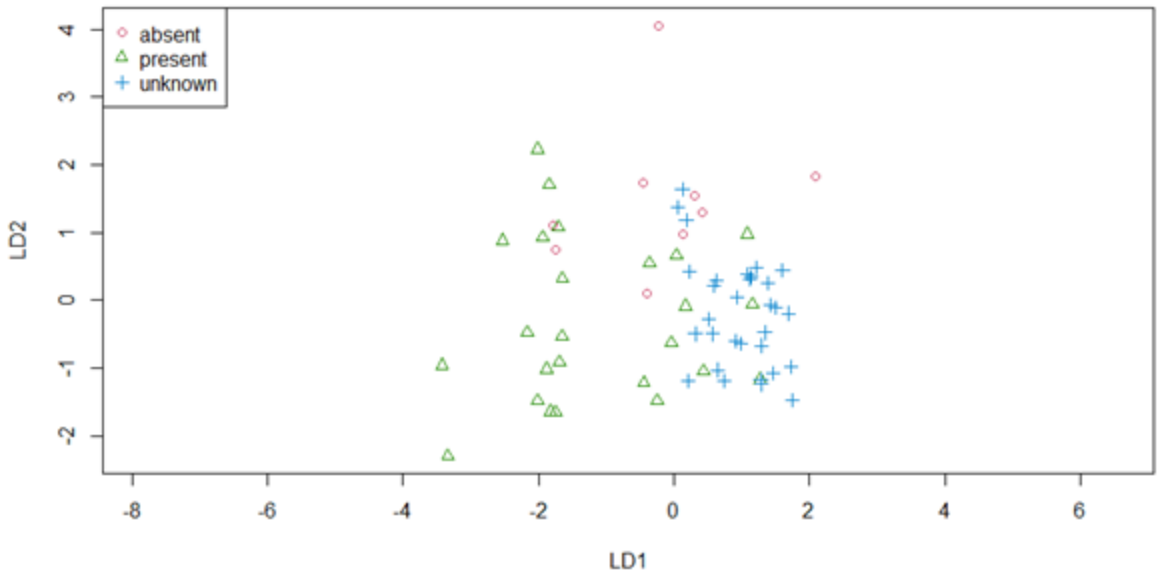

Figure 3.5. CAP analysis of the LIC samples. MAP presence or absence was defined by qPCR done by LIC. Green triangles show the samples with MAP present, red circles show samples with MAP absent, and blue crosses are when the MAP status is unknown.

3.3.3.4 Diversity

There is a large range of diversity within microbiomes (Minseok Kim & Wells, 2016). Several methods have been developed to quantify this diversity, including Simpson and Shannon diversity metrics (Raju et al., 2018). Both metrics were calculated using the vegan package.

The Shannon index gives a higher number when the sample is more diverse, i.e., there are a greater number of genera present. It accounts for both the abundance of genera and their evenness within a microbiome (Magurran, 1988).

Applied to the LIC samples, the average Shannon diversity is 2.3, with some samples being just below 2 and some samples having diversity of up to 4.7 **(Figure 3.6)**. This shows that samples within the effluent can be highly diverse, with a large range of genera. It is hard to know if the number is influenced by the evenness of the samples or whether it is only the number of genera. This is affected by the samples sequenced by Auckland having a larger number of reads, which means that more rare genera are identified compared to samples sequenced overseas that have a lower number of reads. It was expected that the microbiome to have a high level of diversity due to effluent being made up of a wide range of genera, many of which are unknown.

The Simpson diversity is another well used measure of diversity. The Simpson index ranges from 0-1, with 1 being the most diverse. Simpson's index gives the probability that when randomly drawing any two individuals, they will be different genera (Magurran, 1988). Simpson's index closely follows genera abundance, so a higher score can be interpreted as indicating more genera present in the microbiome.

From the results, the average Simpson diversity is 0.82 **(Figure 3.6)**. There are a small number of samples that have a low level of diversity; however, this is likely caused by the low number of reads and therefore a low number of classifications rather than a true reflection of low diversity.

Plots of diversity scores across samples indicate the consistently high diversity scores **(Figure 3.6)**. This is expected in microbiomes, which are well known to have many genera present (Bharti & Grimm, 2019). There are many challenges for characterizing microbiomes, due to the high level of diversity (Van Rossum et al., 2020). However, this also provides many opportunities for further exploration and investigation into many aspects of herd health and potentially the influence of the microbiome on animal traits (Gonda, Chang, Shook, Collins, & Kirkpatrick, 2007; Ross, Moate, Marett, Cocks, & Hayes, 2013).

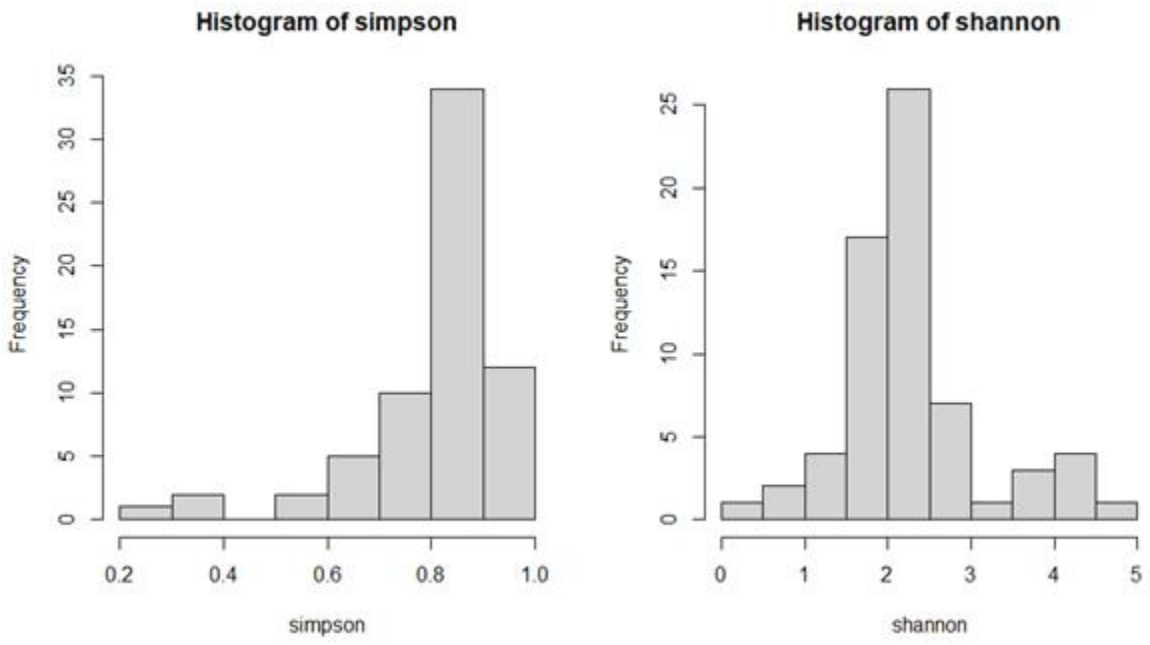

Figure 3.6. Two histograms showing the amount of diversity within the samples. With Simpson diversity, closer to one is more diverse the sample. With Shannon diversity, the higher the number the more diverse the sample.

3.3.4 Genera with a high abundance in the effluent microbiome

We investigated which genera had the largest influence on diversity metric scores. Specifically, we wanted to determine if the most influential genera are ones that are expected within the effluent microbiome. The top 10 most abundant genera found in the LIC samples are listed in **Table 3.2**. The relative frequency was determined by dividing the genera read count by the total read count within each sample.

Most high abundance genera found in this study are known to be in wastewater and cattle faeces **(Table 3.2)**. Some are found across multiple locations and studies, such as *Pseudomonas* and *Acinetobacter* (Girija, Deepa, Xavier, Antony, & Shidhi, 2013; Minseok Kim & Wells, 2016; Ren, Xu, Qu, Zhu, & Wang, 2016; Sasaki, Kitazume, Sasaki, & Nakai, 2004). This shows that the classification using UNOISE overlaps with the expectation of genera that should be observed in effluent microbiomes. In certain environments, community composition can skew towards particular genera.

Corynebacterium is found at 9.5% across all samples. *Corynebacterium* is present in wastewater and cow effluent biodigesters and has been identified as part of the composting process (Akinyemi, Okorhi-Damisa, Efemena, & Adeniyi, 2020; Zhao et al., 2013). As the effluent system is breaking down cow faeces, the presence of fermenting bacteria such as *Corynebacterium* is to be expected.

Genera	Frequency
Corynebacterium	9.5
Proteiniclasticum	7.2
Romboutsia	3.9
Acinetobacter	2.8
Flavobacterium	2.5
Chryseobacterium	2.3
Janthinobacterium	2.1
Pseudomonas	2.1
Trichococcus	1.9
*Clostridium*_XI	1.5

Table 3.2. The top 10 most abundant genera that are present within the LIC effluent system.

In some LIC effluent samples, MAP is known to be present through qt-PCR. When classifying these samples, UNOISE only classifies a very low frequency of *Mycobacterium* in a few samples. While 16S rRNA gene seqeuencing cannot classify down to the species level, it was expected to have some *Mycobacterium* as it is a relatively common genera found in soil. A possibility to explain this is that the UNOISE filtering for rare genera is too strict and so taxa at low levels cannot be classified. This does mean that this study cannot compare the Cq value given by LIC to any UNOISE classification. However, it can compare the shifts in overall community composition which may correlate with MAP presence, using a CAP analysis **(Figure 3.5)**.

3.4 Discussion

3.4.1 Read count normalisation is important for accurate comparisons

There are several steps that are undertaken during the classification process to gain an accurate understanding of which genera are present. One of these is joining the Illumina paired end reads which has a large impact on classification. As the number of reads increases, more species can be classified due to having better resolution (Rajan et al., 2019). If the number of reads is too low, classification of all genera that are present within the microbiome cannot occur (Ross et al., 2013). This means samples will show less variation. Ross (2013), states that rare taxa are easily missed, however they can represent key genera in some microbiomes and can account for over a quarter of all genera present. Other experiments have used in excess of a million reads, which allows for much greater depth and resolution (Degnan & Ochman, 2012; Rajan et al., 2019).

Furthermore, read count normalisation should be routinely undertaken as part of the pipeline analysis. Read counts can vary drastically between batches and even within the same sequencing run **(Appendix B and C)**. Without normalising the read count, variation is more noticeable in samples that have a larger read count. Even with this normalisation, the difference in read number affects classification and therefore variation within the samples **(Appendix B and C)**. This is noticeable within **Figure 3.2** where there are two distinct clusters caused by sequencing run.

Normalisation is done to minimise this technical variation. Subsampling is another option that could aid in normalisation (Jovel et al., 2016). For this to be a viable method, a higher starting number of reads is necessary. When the reads were subsampled, there were under 10 classifications per sample. This is too low to accurately characterize effluent microbiomes. Another alternative for increasing classification accuracy with low OTU numbers would potentially be to weight the OTUs (a high read count for an OTU would increase its classification) (Jovel et al., 2016).

Classification could also be improved by using Illumina metagenomics sequencing instead of 16S rRNA gene sequencing and through increasing read numbers (Ranjan et al., 2016; Yee et al., 2020). Metagenomics sequences all of the genome rather than just the 16S rRNA gene region. This means it can identify a wider range of genera, leading to better classification of the microbiome (Hao et al., 2017; Jovel et al., 2016; Rajan et al., 2019). However, metagenomic sequencing is more expensive.

3.4.2 Batch effects

There is quite a large batch effect in this experiment **(Figure 3.2, 3.5)**. This affected analysis, as the difference between the batches had a large impact on classification. This is especially noticeable in rarer taxa. These low abundance taxa were less likely to be classified, which appears to reduce variation within the microbiome. The lower read depth of the overseas sequencing run meant that these samples appeared to have a lower diversity of genera compared to the Auckland sequencing run although this was not actually accurate (Rajan et al., 2019).

While there are many factors that cause representation of genera in the microbiome to be biased from species that are truly present, much of the research has focused on more easily controlled elements like DNA extraction or read trimming (McLaren et al., 2019; Schriefer et al., 2018). Batch effects are well-known, however because most studies use the same company to sequence their DNA, the actual difference in classification is much lower compared to classification between companies (de Muinck et al., 2017). It was found that the use of two sequencing companies greatly increased the classification skew.

3.4.3 Pipeline effect

In Chapter 2, we compared several methods to classify microbiomic sequencing data. Most of the work was done using default parameters to ensure a true comparison of the methods. Some pre-trimming of the sequencing reads occurred, which reduced the number of poor-quality reads within samples. Other types of read manipulation had very little effect, therefore results are not shown in this paper. A similar method was used to compare the three different 16S rRNA classification pipelines. Default parameters were used for comparisons on pre-trimmed and pre-joined reads.

However, the low numbers of reads with the overseas sequencing provider meant that there were some LIC samples that had zero zOTUs classified with default parameters. When the number of reads need for the minimum abundance was reduced, this meant there all 66 samples had zOTUs. This meant that some zOTUs may be more ‘noisy’ than others but it enabled comparison of the microbiome for all samples (Edgar, 2016). This is shown by the batch effect, where classification was skewed by the two providers.

The classification tool can have a large effect on classification and changing the parameters of these tools can improve or hinder the understanding of which genera are present. Therefore, all comparisons were done on default parameters, however UNOISE’s minimum

abundance was changed to enable classification of the LIC samples, whilst recognising that this may retain some less accurate reads.

3.4.4 Diversity

When comparing to other experiments, the sample diversity is relatively high despite the low read counts (Breitwieser et al., 2019; Rajan et al., 2019). The 16S rRNA gene sequencing method does show a lower amount of diversity than metagenomic sequencing, however, normalising the read count meant that we could compare classifications across my samples. It was found that the Shannon diversity was 2.3 on average while Ranjan (2016) found that their 16S rRNA diversity was approximately 4.0. Their shotgun sequencing diversity was approximately 5.5. These higher scores are likely explained by their high read count of approximately 59 million compared to the 50-800 thousand reads we obtained.

We can infer from the PCoA **(Figure 3.3)** and diversity metrics **(Figure 3.6)** that the effluent microbiome is made up of many genera. The actual number of genera present may even be much higher than what has been classified. This is due to the fact that some genera have not been classified, especially if they are rare taxa (Rajan et al., 2019; Ross et al., 2013). The high level of diversity is what is expected and further confirms the results of a wide range of microbiome experiments (Berg et al., 2020; Han et al., 2020; Pollock et al., 2018).

3.4.5 Sample clustering

Accurate classification of microbiomes is important. The microbiome can affect all parts of animal welfare and production (Barratt et al., 2018; Losinger, 2006). Effluent samples have a similar microbiome, regardless of sampling location or sampling site. There is a substantial amount of variation in the microbiome (Durso et al., 2010). Variation within the effluent microbiome can be affected by many factors. Internal factors include herd health; sample location; or sample type (Gonda et al., 2007; Van Rossum et al., 2020). External factors include batch effects; bias in DNA extraction; or low classification rates (de Muinck et al., 2017). These factors affect sample location and the microbiome on collection dates.

This study characterized the microbiome in four different locations, across a range of time points from an effluent system on a Waikato farm. It was found that the sample date has the largest impact on sample similarity. This could be because the effluent system is interconnected, so microbes are similar between locations and change as the system is affected by the previously mentioned factors. This is shown by the clustering in **Figure 3.3**.

Replicates were used to ensure validity of the classifications and to show if reproducibility through replicates being closer together in **Figure 3.3**. They showed that the pipeline is consistent, and that classification of all samples has little in-sample bias. This means that the method chosen has less impact than variation of genera within the microbiome. Generally, there was a broad pattern of similar genera between replicates **(Figure 3.4)**. This is the expected result and found across many other studies (Fecteau et al., 2016; Henderson et al., 2015; Holman & Gzyl, 2019).

We would recommend using solid samples with the method that has been outlined. We used the QIAGEN Powersoil DNA extraction kit, which is designed for soil samples. Samples that have a higher biomass show a greater amount of microbial diversity due to better DNA extraction with the Powersoil protocol (Wu et al., 2019). However, liquid samples have a lower biomass compared to the solid samples. Samples with low microbial biomass can be more affected by contamination and bias (Kim et al., 2017). Therefore, solid samples are advised. However, there is little clustering caused by sample type. This means that the difference between solid and liquid samples is relatively minor (Poulsen, Pamp, Ekstrøm, & Aarestrup, 2019).

3.4.6 Applicability to other farms and diseases

Identifying “keystone genera” would be one way to monitor the herd effluent more easily (Berg et al., 2020). These are the genera which have the most impact on the microbiome (Paine, 1966). They often rely on each other for survival, and a change in one genera can impact many aspects of the microbiome and therefore other genera within that environment (Gibbons, 2020). If it is possible to identify which genera have the most influence, then any changes in these genera would likely to signal that further investigation of the microbiome is necessary for establishing the reason for that change (Banerjee, Schlaeppi, & van der Heijden, 2018). This would reduce the need for deeper read sequencing, as identification of the main genera within the microbiome would give the necessary information for analysis.

This study has shown that there is a high potential for monitoring herd health through effluent. There is evidence that a shift in the microbiome occurs due to pathogenic microbes **(Figure 3.5)**. Due to the broad nature of the method that was used, this pipeline could be expanded to identify other diseases. Fine-tuning of the method would need to occur to increase specificity for the pathogenic microbe. This could occur through using metagenomic Illumina sequencing, increasing read depth or adapting meta-data.

The information found here can be extrapolated to other diseases on the same farm, and across other farms. This is because the core effluent genera are likely to be similar (Berg et al., 2020; Lavy et al., 2020; Sbardellati et al., 2020). For this to be a viable option farms would need to undergo routine testing of their effluent system. Routine testing of the effluent microbiome would establish a baseline presence and abundance of bacterial genera. This baseline then enables us to define any changes in relative terms from what is expected and therefore identify the possibility of infectious diseases within the herd (Banerjee et al., 2018; Henderson et al., 2015).

If the effluent microbiome does have significant changes concurrent with disease, then identification of affected animals can occur through individual testing. Effluent monitoring is cheaper than testing all animals in the herd. Through increasing bacterial resolution, minor infections could be found before they become established across the entire herd (Collins, Eggleston, & Manning, 2010). This would reduce the incidence of disease and the cost to treat infected cows. It will improve production, as healthy animals produce a high quality and quantity of milk (Barratt et al., 2018; Losinger, 2006). Higher production with a lower rate of unhealthy cows will increase the output of the farm and lower costs (Barratt et al., 2018).

Theoretically, effluent monitoring could be extended both across New Zealand and overseas. This method is broad and captures nearly all bacterial genera hence could be applied to many areas. At the beginning of effluent monitoring, analysis is likely to be affected by overseas and regional variation until a baseline has been identified, as found within porcine faecal microbiome (Munk et al., 2018). To make this testing a viable option, deeper sequencing, and the ability to identify bacteria down to the species level in order to identify opportunistic pathogens is necessary (Chng et al., 2020; Zeineldin, Aldridge, & Lowe, 2018b).

Switching to metagenomics sequencing from 16S rRNA gene sequencing would allow researchers to capture both bacteria and eukaryotic microbes (Quince et al., 2017). This means that the pipeline that is used here would have to be adapted as it currently only identifies bacterial genera and does not classify down to the species level. Metagenomics sequencing down to the species level is more expensive, however it would allow for better classification and more in-depth analysis of the microbiome (Van Rossum et al., 2020).

Chapter Four: Discussion

This study used replicates and a mock community to aid in analysis of previously uncharacterized samples. There are many reasons to use controls in microbiomic experiments. A microbiome has many genera, which can be influenced by a large amount of external and internal factors. Therefore, controls are important to understand the bias caused by method. Replicates give an insight into the reproducibility of the method.

The microbiome can be used to investigate the microbial composition of the microbiome, genera that are present, and diseases that are affecting animal welfare (Goodrich et al., 2014). The ability to identify the core genera that make up the microbiome means that it is possible to use this information to increase animal health and production. The method used in this study is broad, which enables it to be applied across a multiple of locations and diseases.

This method worked fairly well, although there would be a few changes that would improve classification. This includes increasing sequencing and classification depth and changing from Illumina 16S rRNA gene sequencing to Illumina metagenomic sequencing (Bokulich, Ziemski, Robeson, & Kaehler, 2020; Jovel et al., 2016). This would entail adapting the pipeline to metagenomics reads. It would enable a better understanding of the effluent microbiome, especially as rare genera can be hard to identify (Han et al., 2020). Low abundance genera are less likely to be detected with a low number of reads, and are therefore hard to classify (Cattonaro, Spadotto, Radovic, & Marroni, 2018).

4.1 Factors to consider when classifying microbiomes

4.1.1 Importance of a mock community for microbiome classification

In Chapter 2 the ZymoBIOMICS microbial community standard was analysed to produce a baseline understanding of bias in the DNA extraction method, different DNA sequencing methods, and sample classification. The community standard has ten microbes in it, composed of eight bacteria and two yeast species. This mock community is useful for benchmarking classification methods, as it represents a “gold standard” for microbial abundances (Pollock et al., 2018). Deviations from the expected abundances indicates that an analysis is performing poorly in taxonomic classification.

Mock communities are either commercially available or designed specifically with microbes cultured by the researchers (McGovern et al., 2018; Rajan et al., 2019). A commercial mock community means that studies can be compared to each other; whilst targeted mock

communities may provide better validation for single laboratory microbiome studies (Han et al., 2020). The wide range of mock communities available does mean that comparisons between laboratories and experimental designs can be difficult. These limitations need to be kept in mind (Han et al., 2020). It was found that using a commercial mock community means that we knew that any bias was caused by the method and not from a skew within the mock community. It also meant that we could easily compare our results to other studies.

The question is whether there should be a specific “gold standard” mock community, or if targeted mock communities are better for determining the best microbiomic method. We believe research should aim to be using similar mock communities so comparisons between experiments can occur across a range of laboratories and papers. Han (2020) also recommended this, as they found that there was large intra-laboratory variation between classifications in experiments and said that standardization and optimization should be a focus for the field.

In this study, we chose to use a commercial mock community to benchmark the pipelines that were used against a known sample (after McGovern, 2018). This was to ensure that when the effluent samples were analysed, the limitations of the UNOISE (Edgar, 2016) pipeline were understood. Many microbiome experiments use mock communities to understand underlying issues with classification, such as Rajan (2019). Mock communities are important in providing a standard for microbiome classifications (Han et al., 2020; Rajan et al., 2019).

Like Rajan (2019) and Han (2020), this study found that the accuracy of classification from different tools and pipelines varied. This means that while mock communities are useful as a positive control when classifying with one pipeline, it remains important to recognise sources of bias. Analysis of a mock community using different pipelines allowed a better understanding of how the methods compared to each other. It revealed that our conclusions were in line with Nearing (2018) where UNOISE3 was the best pipeline. It is recommended to use a mock community as a positive control. This is emphasised by McGovern (2018) who states that it is “absolutely essential to validate” the methods used, and Han (2020) who found that results from near-identical protocols had good precision however they were not necessarily accurate. They showed that there were many factors which influenced this bias (Han et al., 2020).

Consistent results should be gained across pipelines when using a standardized mock community or technical replicates. In this study, both were used, however de Muinck (2017) suggests that only one is necessary as a control. They also note that “inter-sample distances observed in a real-world experimental setting are significantly higher than observed in both the mock community and standardized environmental sample data” which is important to

understand. Real microbiomes have a higher number of microbes and fluctuating abundances compared to controls. This means that results can vary from what you expect.

It was found that the replicates have a high degree of similarity, in both the cow and effluent samples (de Muinck et al., 2017). This means that we know the method used is reproducible and that the variation is caused by differences between samples.

4.1.2 Effect of technical replicates on reproducibility

Technical replicates are another way to ensure that the methods used are robust and reproducible. They allow comparison of classification variation within the same sample (Fock-Chow-Tho et al., 2017). While the pipelines may not necessarily be accurate to the mock community, the technical replicates show that the classification pipelines are reproducible (Han et al., 2020). These replicates show the reproducibility of many aspects of the experiment, from preservation of the microbiome to DNA extraction kits (Fock-Chow-Tho et al., 2017; S. J. Song et al., 2016). More replicates increase the power of the study and can lead to being able to understand and predict aspects of cow health, such as production traits (Ross et al., 2013).

For comparing Nanopore and Illumina sequencing methodologies, as well as classification tools and databases, we used six technical replicates from a single bovine faeces sample, with three having increasing amounts of a ZymoBIOMICS spike-in (Zymo Research Corporation, 2019). For analysing the uncategorized effluent samples, we used nine pairs of technical replicates; one pair sequenced by both sequencing providers; and a set of three samples taken from the several sites in the same sample location. Like Raju (2018), we found that the use of replicates is vital to being able to understand repeatability. It was found that the number of reads can vary hugely within different zOTUs and between sequencing providers. By increasing read numbers and sequencing depth, classification accuracy and reproducibility can be improved (Raju et al., 2018).

It may be more useful to do a limited number of replicates of different samples rather than one sample with multiple replicates. We would suggest that up to 10 replicates for an experiment is necessary to increase the reliability of the experiment, depending on the total number of samples. The replicates are useful as they give an insight into how much skew is between samples and how much is caused within the bioinformatic method chosen (Prodan et al., 2020). We did not use a mock community the entire way through the pipeline alongside the replicates the chosen pipeline had been validated beforehand. Other studies recommend

using both replicates and mock communities throughout (Nearing et al., 2018; Rajan et al., 2019)

4.1.3 Effect of sequencing methodology and technology on classification

Sequencing methodology and technology can have a large impact on read length and quality. This can have a flow-on effect influencing classification of the microbiome. Next-generation sequencing is much faster and has many benefits for microbiome studies such as increasing accuracy (Fox et al., 2014; Goodwin et al., 2016). 16S rRNA gene and metagenomic Nanopore sequencing can give vastly different genus classifications, as shown by the results in Chapter 2. These can vary compared to Illumina sequencing, as shown in Chapter 2 and 3. The databases that are used have a large impact. The results indicate that there is a need to carefully consider the specific approach used when analysing the microbiome, whether it be mock communities or uncharacterized samples.

4.1.3.1 Nanopore sequencing

The number of reads that were gained from Nanopore metagenomics sequencing was much lower than either Nanopore 16S rRNA gene sequencing or Illumina 16S rRNA gene sequencing. This could have an impact on classification as read length and sequencing depth is important (Pearman et al., 2020). As there is a smaller number of reads, this means that any amount of increased classification would have a larger bias. Sequencing depth can also have an effect on classification accuracy (Rajan et al., 2019; Soon Gweon et al., 2019). This is further shown by the Illumina sequencing of effluent samples, in the PCoA. This is likely partially because an increased depth means that more genera are identified and hence the samples look more varied (Rajan et al., 2019).

There are differences in microbiome classification dependent on the method chosen to sequence the microbiome. 16S rRNA gene sequencing does not have the resolution to provide species level classification (Brumfield, Huq, Colwell, Olds, & Leddy, 2020). Like metagenomics sequencing, 16S rRNA gene sequencing can be biased. This is caused more by primer or amplification bias rather than lack of genome sequences in the database (Bharti & Grimm, 2019; Degnan & Ochman, 2012; Martinez-Porchas, Villalpando-Canchola, Ortiz Suarez, & Vargas-Albores, 2017). 16S rRNA gene sequencing is often used for microbiome studies, as bacteria with low abundance can be identified through amplification of the 16S rRNA gene (Hao et al., 2017; Peabody, Van Rossum, Lo, & Brinkman, 2015). It was found to be more reliable as the amount of bias caused by this can be more easily quantified. However, 16S rRNA gene sequencing is always limited by the depth of classification it can reach.

Using technical replicates, we found a significant skew with metagenomics sequencing. Nearly 50% of the sample was classified as *Pseudomonas*, whereas all other sequencing types found a more balanced distribution of genera. Such a high relative percentage of one genus also proportionally reduces the relative frequencies of other genera within the microbiome. This skew contrasts with several other studies that found that metagenomics is less biased than 16S approaches and more accurately estimates relative abundance in the microbiome (Jovel et al., 2016; Peabody et al., 2015). However, it is noted that the level of concordance between 16S rRNA gene and metagenomic sequencing has always varied (Jovel et al., 2016).

Other studies show that there is consistent bias with metagenomics sequencing which resembles my results (McLaren et al., 2019; Poulsen et al., 2019). Peabody (2015) found that metagenomics is less biased than 16S, however, they state that the accuracy of shotgun metagenomics varies between classification methods. This study found that variation occurs with 16S classification tools. Another study found that there can be systematic bias between methods, supported by my results (Poulsen et al., 2019). Poulsen (2019) also states that one needs to be careful not to over-emphasise results from exploratory research and to continually validate reproducibility. This further emphasises why technical replicates are necessary during these experiments. The effluent replicates show that variation occurs due to some method bias, however most variation is due to actual factors. This is shown by the fact that replicates and samples collected on the same date are very similar. There is little variation between replicates, so the method used is not a large factor in causing PCoA clustering.

The skew in the metagenomics sequencing results may be explained by the species present. McLaren (2019) found the species ratio can vary from very little to 100-fold depending on which *Firmicutes* species is present. This may be an indicator of why *Pseudomonas* dominated the metagenomics sequencing, but not the 16S rRNA gene sequencing. Both 16S rRNA gene and metagenomics sequencing have limitations and understanding these are highly important for microbiome studies. Studies should consider using both sequencing methodologies while the knowledge of metagenomics improves. As metagenomics sequencing improves in reliability, this reduces the incidence of species bias and increases correct classification (McLaren et al., 2019).

Metagenomics can be more expensive than 16S rRNA gene sequencing, however, this is offset by the increased resolution that leads to a higher level of taxonomic classification (Jovel et al., 2016). Metagenomics through Illumina technology would hopefully provide a better taxonomic resolution, potentially down to the strain level. This is due to 16S rRNA classification occurring off one gene while metagenomics classification occurs off longer and more diverse sequences (Johnson et al., 2019). As metagenomic sequencing improves in quality, we

believe that this will allow for better classification and analysis of microbiomes. Through increased resolution identification of pathogenic bacteria such as MAP could occur instead of using microbiome correlation. Further research is necessary to reduce bias before accepting all the results found using this method.

4.3.1.2 Illumina sequencing

Illumina is another example of next-generation sequencing. Illumina produces shorter and more accurate reads compared to Nanopore technology (Ranjan et al., 2016). A limitation of this study is that it did not use metagenomic Illumina sequencing. The difference between the next-generation sequencers was compared using 16S rRNA gene sequencing instead.

Like Nanopore classification, Illumina 16S reads can be classified using a variety of tools. These classification tools have a range of accuracy when classifying microbiomes and no one method is used consistently across studies (Nearing et al., 2018). In the initial comparison of 16S classification pipelines the default parameters were used, on pre-joined reads. This was done to ensure that the data would be valid to compare. It limits the total number of reads that are analysed while ensuring that reads are not lost whilst merging within the pipelines. In one metagenomics study, they found that 37% of paired-end raw reads merged, which is much lower than the 70% of paired-end raw reads that merged in this experiment (Vo & Jedlicka, 2014).

16S Illumina is the choice of sequencing methodology for a wide range of bovine microbiome studies (Gomez, Arroyo, Costa, Viel, & Weese, 2017; Holman & Gzyl, 2019; McGovern et al., 2018; Wong et al., 2016). This is likely due to the high base call accuracy which makes downstream analysis easier (Pfeiffer et al., 2018; Wenger et al., 2019). Nanopore is increasing in popularity for metagenomics and microbiome studies however, Illumina is still considered to be the gold standard sequencer (Ranjan et al., 2016).

There are several studies that have focused on comparing Illumina 16S classification pipelines as the need for accurate microbiome analysis increases (Nearing et al., 2018; Prodan et al., 2020). This work further adds to it, although there is less technical detail due to the focus being on accurately characterizing effluent samples.

We note the emphasis on characterizing microbiomes at the genus level. This is likely to result in exclusion of some taxa that are only classified at higher taxonomic levels (i.e., at the family level). However, the majority of reads are classified at the genus level so this approach balances classification resolution and classification percentage. 16S rRNA gene sequencing does not have enough resolution to identify species or strains. This is a huge limitation of using

16S rRNA gene sequencing to try to classify microbiomes. We would suggest comparing Illumina and Nanopore metagenomics sequencing. We would also suggest that metagenomics sequencing would be preferable over 16S rRNA gene sequencing to increase resolution and enable better classification. This is in-line with several other studies (Leggett & Clark, 2017; McLaren et al., 2019).

One issue with the uncharacterized effluent samples was that they were only classified using 16S Illumina sequencing. As noted, this may lead to missing certain genera, due to Illumina 16S only classifying bacteria, and not eukaryotic or viral DNA (Degnan & Ochman, 2012). In the future, we would suggest that Illumina metagenomics would be a better sequencing method or using both Illumina short reads for accuracy and Nanopore long reads to identify unknown genera.

4.1.4 Batch effects reinforce the need for controls and replicates

As found in the effluent classifications, there is a substantial batch effect arising from use of different sequencing companies. This greatly increases the bias against rare genera that can influence classification. It is known that there are batch effects caused by sequencing runs in the same company, however, less is known about differences between companies (Han et al., 2020; Poulsen et al., 2019). We believe that the main reason for the large amount of classification bias is due to read depth. Read depth is known to have an effect on classification, and the effluent samples have further added evidence towards this (Rajan et al., 2019; Soon Gweon et al., 2019). Classification can also be affected by read quality, where poor reads may be discarded during quality control (Pfeiffer et al., 2018). Low read quality can also mean that poor matches to genomes are made, or they do not classify at all (Liu et al., 2020).

This shows why it is important to use both mock communities and replicates. They provide a more complete picture of how much bias is present. Without the replicates we would not have the same level of understanding around how much variation is caused by the pipeline and how much is caused by batch effects and read depth. The mock community was not sequenced by the second sequencing company. This sequencing would have further clarified the relationship between read depth and classification results, allowing further insight into the level of bias caused by the low read count. We know from the replicates that low read count has caused the number of rare taxa to be lower and reduced the number of classifications within the samples (Ranjan et al., 2016).

4.1.5 Sample type and collection date affects classification

We evaluated the microbiome of bovine herd effluent at different locations and over a range of dates. The microbiome is affected by many factors, such as seasons and herd health (Gonda et al., 2007; Li et al., 2020). The overall bacterial composition tends to remain stable, although abundances do fluctuate, mimicking the results of other microbiomic studies (Han et al., 2020).

The type of sample affects classification. It is possible to sample any location within the effluent system and get a similar result. However, this method is more reliable on solid-type samples due to the DNA extraction kit used (Vo & Jedlicka, 2014). Further along the effluent system, the bacterial profile changes from similar to the bovine gastrointestinal tract microbiome and moves towards having a composition more similar to a biodigester with anaerobic microbes dominating (Akinyemi et al., 2020; Dowd et al., 2008). The effluent system is designed to reduce waste; the amount of fertilizer needed and the impact on the environment (DairyNZ, 2021). Therefore, it consists of microbes that can digest cow faeces. To monitor microbes that may be causing health issues in cows, it would be best to sample as close to the source as possible (i.e., fresh cow faeces or straight out of the cowshed). This is because the microbiome can change further down the effluent system.

The date has a large impact on clustering in the PCoA. This is likely due to how interconnected the effluent system is. As such, any changes in microbial composition are propagated throughout the entire system. The locations within the effluent are therefore more similar during that time period, as they cluster together. This could lead to time series analysis. Identifying the core microbiome could lead to only investigating genera of interest which change between locations and dates (Banerjee et al., 2018).

However, it is important to consistently sample the same location as this lowers any variation that can come from separate locations. Identifying changes in the same location is more sustainable than attempting to validate microbial changes across both location and date. Sampling the same location over several time points will eventually lead to identifying keystone genera and understanding which changes are due to external influences (Banerjee et al., 2018; Berg et al., 2020). For example, dysbiosis could occur due to pathogenic microbes or due to changes in diet (Fecteau et al., 2016; Fomenky et al., 2018).

4.1.6 Changes in the effluent microbiome

Changes in the effluent composition can be caused by the growth of bacteria within the effluent system itself. This would lead to the composition of effluent being vastly different to the

composition of the rumen microbiome (Zaheer et al., 2019). As such, sampling taken from earlier on in the effluent system may more closely match the bovine gut microbiome whereas later sampling sites have the potential to be much more varied. There is potential that over time, some bacterial species may replicate more efficiently in the effluent samples and become overrepresented in relation to the microbiome from the cow rumen (Choo et al., 2015; Ren et al., 2016). This can lead to problems when using effluent as a proxy for cow rumen.

We believe that while location is less important that date in terms of effluent similarity, it would be prudent to sample and compare microbiome abundance and diversity from samples that were taken directly from bovine faeces or closer to the cows (i.e., directly out of the dairy shed) in order to have a more accurate understanding how the effluent microbiome matches the bovine microbiome from a few individual cows. Furthermore, changes in microbial composition can be influenced by an increase in one genus. As one genus increases, this will proportionally decrease other genera. Understanding this is important for interpreting microbial abundance. These fluctuating abundances can occur even day to day with changes in the environment (Li et al., 2020).

4.1.7 Targeting the baseline genera to identify changes

There are many areas of bovine microbiome research, from the overall composition of the rumen to identifying specific, targeted microbes (Koester, Poole, Serão, & Schmitz-Esser, 2020; Sbardellati et al., 2020; E.-S. Song et al., 2016). The faecal microbiome is a growing area of research, and there are potential beneficial implications for using effluent as a tool for monitoring the overall health of a herd (Muñoz-Vargas et al., 2018; Wu et al., 2019). There are many elements that show that the identification of diseases within herds is important for assessing health and productivity (Zeineldin, Aldridge, et al., 2018a). Dysbiosis of the microbiome can both cause, and be caused by disease (Fecteau et al., 2016; Gomez et al., 2017; Zeineldin, Aldridge, et al., 2018a). Being able to identify the microbiome profile could be a key way to correlate health, and/or be an indicator of any changes. This does rely on knowing how external factors like seasonal changes and location can influence the abundances of microbes found.

It is important to establish what genera comprises the core microbiome or the “keystone genera”, as abundances do fluctuate over time (Banerjee et al., 2018; Henderson et al., 2015). Ranjan (2016) describes the microbiome as being composed of thousands of microbes, with a small number of dominant genera, and most genera at a low abundance. It is necessary to keep this in mind whilst performing a microbiome experiment. Mock communities provide a way of calculating how the method is working, but not how the microbiome is changing.

Understanding changes is easier through the use of replicates. It is necessary to analyse both rare and abundant microbes within the microbiome (Raju et al., 2018; Ranjan et al., 2016). Without replicates and without the use of mock communities, conclusions may be biased away from rare genera and the abundances of more common taxa may be skewed. This can affect interpretation of which genera are present and their role within the microbiome.

Missing taxa are a huge issue in microbiome analysis. It is estimated that up to 30% of rare taxa are not found when investigating the microbiome (Berg et al., 2020). This is hindered by the many unculturable bacterial genera and numerous unknown species (Hart, Meyer, Johnson, & Ericsson, 2015; Kembel et al., 2012; Malla et al., 2018). These genera may have important roles to play in the health of the cow and in maintaining the microbiome (Holman & Gzyl, 2019; Ranjan et al., 2016).

Therefore, replicates and a positive control like a mock community is necessary to ensure that the method is as accurate as possible. They enable quantification of bias from missing taxa, compared to quantifying bias caused by the chosen method. When using a known mock community, it is known which genera are expected, and the quantity of those microbes (Zymo Research Corporation, 2017). This means that we can compare the ratio of what is found versus the expected bacterial frequency. After comparing the ratio, we can identify where the method varies from expected. Therefore, we obtain an understanding of the level of bias caused by the method and can apply this to samples where the genera and their abundance is unknown.

All microbiomes have interactions between component genera, which can affect many processes (Goldford et al., 2018; Hermans et al., 2020). These interactions mean that abundances of microbes within a microbiome do not fluctuate in isolation. They are, instead, affected by other genera, functional stability, and external factors (Goldford et al., 2018; Hao et al., 2017; Romagnoli et al., 2017). This explains how the whole microbiome composition can change as animal health does. However, the bovine effluent is poorly characterized and rarely monitored, so the opportunity exists to change this (Porwal, Mane, & Velhal, 2015). Other types of soil studies have found that it is possible to predict a profile through microbiome sequencing (Hermans et al., 2020). Soil communities are greatly influenced by the type of land use and changes in conditions, so it is likely that the effluent system is similarly affected (Hermans et al., 2020; Li et al., 2020).

4.1.8 Identifying pathogenic microbes within microbiomes

Mycobacterium avium subsp. *paratuberculosis* is one pathogen that greatly affects herd health and production within New Zealand and overseas (Harris & Barletta, 2001). There are many other diseases that also affect dairy herds. For example, *Salmonella enterica* can also play a role in animal health and productivity (Muñoz-Vargas et al., 2018). Another example is *Clostridium botulinum* which seldom occurs however it can have serious, life-threatening consequences (Seyboldt et al., 2015). These diseases can have far-reaching impacts on the farm and being able to identify these microbes early would mean that the impact can be quickly managed and reduced. There are many ways that these pathogenic bacteria can spread throughout the herd - from constant shedding, through faecal-oral transmission, or from dam to calf (Fock-Chow-Tho et al., 2017; Muñoz-Vargas et al., 2018; Patterson et al., 2020; Seyboldt et al., 2015).

Within dairy herds, the main traits selected upon by breeders are animal health, fertility, production, environmental impact, and disease-resistance (Georges, Charlier, & Hayes, 2019). While some of these traits are based on genes associated with the cow itself, the gut microbiome can also greatly affect many of these traits as well (Ross et al., 2013). As the faecal microbiome shares similarities to the gut microbiome and the pathogenic microbes are shed into the environment, we can detect diseases before they become endemic within the herd (Minseok Kim et al., 2017; Rajan et al., 2019). The differences in the rumen microbiome and effluent microbiome can have large differences, which is why we tested several locations, to see how much the locations varied (Pitta et al., 2016). However, there are many factors (diet, age, microbial interactions) which affect the rumen microbiome, and many more (age, breed, animal health) which affect the faecal microbiome (Minseok Kim et al., 2017; Minseok Kim & Wells, 2016). There are OTUs which are shared between the rumen and faeces; however, concordance is not as high as some studies suggest (Minseok Kim & Wells, 2016). This reinforces why replicates and robust controls are necessary.

Technical and biological replicates can lend confidence to results, and such controls are important for samples with a low biomass where contamination can easily occur and influence sample classification (Minseok Kim et al., 2017). One way to reduce this would be to increase the biomass and avoid liquid samples. Any sequencing of new samples should be done alongside positive controls, such as the mock communities to ensure that technical variation is not the reason for differences in samples (Kim et al., 2017). Replicates also verify that the sample preparation and sequencing procedures work as expected, such as with the pipeline that were used for this study.

4.1.9 Conclusion: keystone genera can be predictors of disease

The microbiome is often made up of core taxa. These taxa are otherwise known as keystone species or hub species (Berg et al., 2020). Identifying these core taxa will mean that specific changes can be associated with pathogenic microbes. Berg (2020) states that keystone species are likely to play a crucial role within the microbiome. This will help with understanding the functioning of the microbiome and the effect of pathogens on microbial diversity.

Berg (2020) also states that dysbiotic animals vary more in microbial community composition compared to healthy individuals. This means if a herd is healthy, it will have a similar abundance of microbes, however if a disease is present, there are multiple ways it may affect the microbiome. This is because there are many diseases which can change the microbiome, and each individual is affected differently by the same disease. Hence limiting the amount of bias caused by the method chosen is important in understanding real divergence from a healthy microbiome. A well-defined mock community can aid in showing the best method with the lowest amount of bias throughout (Berg et al., 2020; Nearing et al., 2018). There are many reasons for bias within a microbiome, disease being one of them, so controls are necessary to quantify the level of bias caused by the pipeline itself.

Metagenomic profiles can be used to predict phenotypes in both humans and cattle (Ross et al., 2013). However, not all genera within a microbiome can be identified and fully sequenced using 16S rRNA gene sequencing. This has improved recently as taxonomic databases have increased in size and sequencing outputs have increased, however, there are still gaps in our knowledge. These gaps are especially related to the function of genera within the microbiome and the effect of antibiotics on the microbiome (Larsson et al., 2018).

The number of reads is important for both sequencing methods and should be in the millions for accurate quantification (Pfeiffer et al., 2018; Ross et al., 2013). Another issue with microbiomes is that accuracy is reduced by having poor databases, with poorly characterized phenotypes (Ross et al., 2013).

The microbiome of the host is variable, and can change in response to many factors, such as diet or environmental factors (Li et al., 2020; Ross et al., 2013; Vikram et al., 2017). It can also be affected by the growth of microbes within the effluent system itself. This can affect classification. The stability of effluent microbiomes over time and due to other factors is still undetermined. Understanding part of this stability may be resolved by routine testing. This routine testing would help identify keystone genera. The results show that there are often similar genera present across multiple effluent locations. This fits in with the current paradigm

of microbiome research (Hao et al., 2017; Minseok Kim et al., 2017; Wu et al., 2019). The replicates that were used enabled further confidence in the results obtained. It shows that the research is reproducible and thus comparable to other experiments.

4.2 Future work

4.2.1 Comparisons

Improvements to this project could come from sequencing the mock community and technical replicates using metagenomics Illumina sequencing. This would help with better accuracy and improved taxonomic resolution (Quince et al., 2017). While we can make educated comparisons, using Illumina metagenomics would reduce the number of assumptions that were made. Metagenomic sequencing has the potential to better represent all genera that are present in the microbiome rather than just bacterial genera, as 16S rRNA gene sequencing targets just the bacterial taxa present (Ranjan et al., 2016). With PCR bias removed and using more than the single 16S loci for taxonomic identification, we would obtain a more accurate picture of the microbiome profile. This may lead to identifying correlations with pathogenic microbes within the microbiome, or even identifying down to the species level (Wu et al., 2019; Zeineldin, Aldridge, et al., 2018a).

4.2.2 Effluent sequencing

Effluent samples have shown that they can be used to identify genera that are present in a bovine microbiome. The core microbes of the microbiome are likely to remain the same across many herds, which enables profiling and targeted analysis. It should be possible to be able to extract part of the data out to other herds across New Zealand, and potentially internationally due to this method being so broad. Within this work comes identifying pathogenic microbes and correlations between the effluent microbiome and shifts that may be caused by the presence of other microbes. Further work would be needed to ensure that any correlations are true variation.

An understanding of the ratio of pathogens to probiotics would be another avenue that further research could undertake. This would require extensive profiling of feed intake and supplements, to understand how external factors influence the microbiome. There may be an optimum ratio of bacterium, that has not been identified here, although we can potentially trace changes over time in the herd.

This research has the potential for identifying other pathogenic microbes which influence bovine herd health and production. Further research should identify the overall herd health both before and during the time periods. This work has focused on finding *Mycobacterium avium* subsp. *paratuberculosis* (MAP). As the entire microbiome is sequenced, it is possible to expand this method to identify other microbes or diseases that are found within a herd, such as facial eczema. Like MAP, facial eczema leads to poor health and production (Di Menna, Smith, & Miles, 2009; Morris, Phua, Cullen, & Towers, 2013). However facial eczema is caused by a fungus, so classifying using 16S Illumina would not be feasible (Di Menna et al., 2009). This further reinforces how using Illumina metagenomics or improving Nanopore metagenomic sequencing would increase the knowledge of the effluent microbiome.

The LIC MAP qPCR data begins to give an indication of the correlation between pathogen presence and the microbiome profile. With further analysis, it may be possible to determine if uncharacterized samples are similar enough to identify that the profile has changed potentially due to a pathogenic microbe which may be worth investigating. This metadata has made this analysis possible. Routine testing would remove the need for additional information alongside this method, as the core microbiome is identified. Changes from what genera and microbial abundances are expected means identifying potential dysbiosis and what may have caused it.

Whilst this study touched on temporal analysis and the changing of the microbiome profile, it would be good to have more detailed analysis on how stable the effluent microbiome is. It is necessary to be aware that there is a selection of microbes that change while most remain constantly present. Understanding how microbial abundances fluctuate due to external factors such as seasonal changes may provide a better idea of which genera have a negative effect on herd health and production (Kim et al., 2017; Li et al., 2020).

Reducing the incidence of disease through early detection has enormous beneficial consequences for farms. This will increase production, as unhealthy individuals produce a lower quality and quantity of milk. It will reduce costs, as it would be possible to identify individuals that are infected before it spreads throughout the herd. Overall, the potential for improving farms productivity and output through this method outweighs the costs of increasing effluent testing. Increased testing of the microbiome will reduce the need to test individuals, which is far more expensive. This broad method can be used for many diseases and is not specific to MAP.

Chapter 5: Appendices

5.1 Supplementary

S3.1 GitHub, a place of the commands used to classify microbiomes using Illumina 16S data and visualise these classifications.

https://github.com/AlyssaEarnshaw/Effluent-microbiome

5.2 Appendices

Appendix A: LIC sample information

Name of sample	Unique Sample ID	Sample location	Sample site	Date of collection	qPCR
Irrigation-S1-27.11.16	IFS1271116	Irrigation	Solid-1	27-Nov-16	NA
Irrigation-S2-27.11.16	IFS2271116	Irrigation	Solid-2	27-Nov-16	NA
Irrigation-S3-27.11.16	IFS3271116	Irrigation	Solid-3	27-Nov-16	NA
Newpile-L1-28.11.16	NPL1281116	New pile	Liquid-1	28-Nov-16	45
Newpile-L1-4.4.17	NPL1040417	New pile	Liquid-1	4-Apr-17	37.6
Newpile-L2-14.5.17	NPL2140517	New pile	Liquid-2	14-May-17	33.2
Newpile-L2-22.2.17	NPL2220217	New pile	Liquid-2	22-Feb-17	35.9

Newpile-L2-28.11.16	NPL2281116	New pile	Liquid-2	28-Nov-16	37
Newpile-L2-4.4.17	NPL2040417	New pile	Liquid-2	4-Apr-17	38
Newpile-S1-14.5.17	NPS1140517	New pile	Solid-1	14-May-17	34.6
Newpile-S1-15.12.16	NPS1151216	New pile	Solid-1	15-Dec-16	45
Newpile-S1-28.11.16	NPS1281116	New pile	Solid-1	28-Nov-16	35.7
Newpile-S1-4.4.17	NPS1040417	New pile	Solid-1	4-Apr-17	37
Newpile-S2-14.5.17	NPS2140517	New pile	Solid-2	14-May-17	6.6
Newpile-S2-15.12.16	NPS2151216	New pile	Solid-2	15-Dec-16	45
Newpile-S2-22.2.17	NPS2220217	New pile	Solid-2	22-Feb-17	35.1
Newpile-S2-28.11.16	NPS2281116	New pile	Solid-2	28-Nov-16	38.2
Newpile-S2-4.4.17	NPS2040417	New pile	Solid-2	4-Apr-17	36.6
Newpile-S3-15.12.16	NPS3151216	New pile	Solid-3	15-Dec-16	45

Newpile-S4-15.12.16	NPS4151217	New pile	Solid-4	15-Dec-16	36.1
Pond-L1-15.12.16	PL1151216	Pond	Liquid-1	15-Dec-16	37.5
Pond-L1-28.11.16	PL1281116	Pond	Liquid-1	28-Nov-16	45
Pond-L3-14.5.17	PL3140517	Pond	Liquid-3	14-May-17	33.1
Pond-L4-28.11.16	PL4281116	Pond	Liquid-4	28-Nov-16	45
Pond-L4-4.4.17	PL4040417	Pond	Liquid-4	4-Apr-17	45
Sandtrap-S1A-17.12.17	SS1A171217	Sandtrap	Solid-1	17-Dec-16	NA
Sandtrap-S1A-19.4.17	SS1A190417	Sandtrap	Solid-1	19-Apr-17	NA
Sandtrap-S1A-26.2.17	SS1A160217	Sandtrap	Solid-1	26-Feb-17	NA
Sandtrap-S1A-26.3.17	SS1A260317	Sandtrap	Solid-1	26-Mar-17	NA
Sandtrap-S1A-8.10.16	SS1A081016	Sandtrap	Solid-1	8-Oct-16	NA
Sandtrap-S1B-19.4.17	SS1B190417	Sandtrap	Solid-1	19-Apr-17	NA

Sandtrap-S1B-26.2.17	SS1B260217	Sandtrap	Solid-1	26-Feb-17	NA
Sandtrap-S1B-26.3.17	SS1B260317	Sandtrap	Solid-1	26-Mar-17	NA
Sandtrap-S1B-8.10.16	SS1B081016	Sandtrap	Solid-1	8-Oct-16	NA
Sandtrap-S2A-17.12.16	SS2A171216	Sandtrap	Solid-2	17-Dec-16	NA
Sandtrap-S2A-19.4.17	SS2A190417	Sandtrap	Solid-2	19-Apr-17	NA
Sandtrap-S2A-26.2.17	SS2A260217	Sandtrap	Solid-2	26-Feb-17	NA
Sandtrap-S2A-26.3.17	SS2A260317	Sandtrap	Solid-2	26-Mar-17	NA
Sandtrap-S2A-8.10.16	SS2A081016	Sandtrap	Solid-2	8-Oct-16	NA
Sandtrap-S2B-17.12.16	SS2B171216	Sandtrap	Solid-2	17-Dec-16	NA
Sandtrap-S2B-19.4.17	SS2B190417	Sandtrap	Solid-2	19-Apr-17	NA
Sandtrap-S2B-26.2.17	SS2B260217	Sandtrap	Solid-2	26-Feb-17	NA

Sandtrap-S2B-26.3.17	SS2B260317	Sandtrap	Solid-2	26-Mar-17	NA
Sandtrap-S2B-8.10.16	SS2B081016	Sandtrap	Solid-2	8-Oct-16	NA
Sandtrap-S3A-17.12.16	SS3A171216	Sandtrap	Solid-3	17-Dec-16	NA
Sandtrap-S3A-19.4.17	SS3A190417	Sandtrap	Solid-3	19-Apr-17	NA
Sandtrap-S3A-26.2.17	SS4A260217	Sandtrap	Solid-3	26-Feb-17	NA
Sandtrap-S3A-26.3.17	SS3A260317	Sandtrap	Solid-3	26-Mar-17	NA
Sandtrap-S3A-27.11.16	SS3A271116	Sandtrap	Solid-3	27-Nov-16	NA
Sandtrap-S4A-26.2.17	SS4A260217	Sandtrap	Solid-4	26-Feb-17	NA
Sandtrap-S4A-26.3.17	SS4A260317	Sandtrap	Solid-4	26-Mar-17	NA
Sandtrap-S5-26.2.17	SS5A260217	Sandtrap	Solid-5	26-Feb-17	NA
Sandtrap-S6A-26.2.17	SS6A260217	Sandtrap	Solid-6	26-Feb-17	NA

Wedge-L1-4.4.17	WL1040417	Wedge	Liquid-1	4-Apr-17	36.5
Wedge-L2-14.5.17	WL2140517	Wedge	Liquid-2	14-May-17	33.2
Wedge-L2-22.2.17	WL2220217	Wedge	Liquid-2	22-Feb-17	35.3
Wedge-S1-14.5.17	WS1140517	Wedge	Solid-1	14-May-17	35.8
Wedge-S1-15.12.16	WS1151216	Wedge	Solid-1	15-Dec-16	36.3
Wedge-S1-28.11.16	WS1281116	Wedge	Solid-1	28-Nov-16	35.4
Wedge-S1-4.4.17	WS1040417	Wedge	Solid-1	4-Apr-17	37.2
Wedge-S2.2-14.5.17	WS22140517	Wedge	Solid-2	14-May-17	35.8
Wedge-S2.1-22.2.17	WS22220217	Wedge	Solid-2	22-Feb-17	36.6
Wedge-S2-15.12.16	WS2151216	Wedge	Solid-2	15-Dec-16	41.8
Wedge-S2-22.2.17	WS2220217	Wedge	Solid-2	22-Feb-17	36.6
Wedge-S2-28.11.16	WS2281116	Wedge	Solid-2	28-Nov-16	41.8

Wedge-S2-4.4.17	WS2040417	Wedge	Solid-2	4-Apr-17	45

Appendix A. The samples provided by LIC and separated out to show the range of sample location, type and date sampled. The final column shows the qPCR value.

Appendix B: Sample DNA amount and sequencing information

Name of sample	DNA amount	Illumina sequencing centre	Total read count
Irrigation-S1-27.11.16	6.8	Overseas	116000
Irrigation-S2-27.11.16	16.7	Overseas	118658
Irrigation-S3-27.11.16	6.7	Overseas	50246
Newpile-L1-28.11.16	3.9	Overseas	73743
Newpile-L1-4.4.17	9.6	Auckland	624,680
Newpile-L2-14.5.17	15.2	Auckland	313,009
Newpile-L2-22.2.17	2.8	Overseas	136331
Newpile-L2-28.11.16	3.6	Overseas	56174
Newpile-L2-4.4.17	13.0	Auckland	624,680
Newpile-S1-14.5.17	3.5	Overseas	50204
Newpile-S1-15.12.16	16.3	Overseas	82298
Newpile-S1-28.11.16	52.0	Overseas	82454

Newpile-S1-4.4.17	33.4	Overseas	75370
Newpile-S2-14.5.17	6.6	Overseas	75212
Newpile-S2-15.12.16	1.8	Overseas	125055
Newpile-S2-22.2.17	4.4	Overseas	63116
Newpile-S2-28.11.16	31.0	Overseas	74431
Newpile-S2-4.4.17	15.1	Overseas	66720
Newpile-S3-15.12.16	2.5	Overseas	57271
Newpile-S4-15.12.16	13.3	Auckland	719,045
Pond-L1-15.12.16	6.0	Auckland	311,305
Pond-L1-28.11.16	7.7	Auckland	372,138
Pond-L3-14.5.17	12.1	Auckland	458,711
Pond-L4-28.11.16	9.8	Auckland	424,138
Pond-L4-4.4.17	0.8	Overseas	57060
Sandtrap-S1A-17.12.17	1.5	Overseas	51688
Sandtrap-S1A-19.4.17	1.5	Overseas	67171
Sandtrap-S1A-26.2.17	1.6	Overseas	131632

Sandtrap-S1A-26.3.17	6.3	Overseas	111859
Sandtrap-S1A-8.10.16	10.4	Overseas	109340
Sandtrap-S1B-19.4.17	2.7	Overseas	53382
Sandtrap-S1B-26.2.17	53.2	Overseas	55007
Sandtrap-S1B-26.3.17	3.3	Overseas	61466
Sandtrap-S1B-8.10.16	17.4	Overseas	67218
Sandtrap-S2A-17.12.16	1.8	Overseas	111567
Sandtrap-S2A-19.4.17	1.9	Overseas	107985
Sandtrap-S2A-26.2.17	1.5	Overseas	52298
Sandtrap-S2A-26.3.17	7.6	Overseas	56975
Sandtrap-S2A-8.10.16	1.9	Overseas	50047
Sandtrap-S2B-17.12.16	4.9	Overseas	62146
Sandtrap-S2B-19.4.17	3.2	Overseas	76611
Sandtrap-S2B-26.2.17	2.9	Overseas	72333

Sandtrap-S2B-26.3.17	7.2	Overseas	69443
Sandtrap-S2B-8.10.16	1.4	Overseas	75726
Sandtrap-S3A-17.12.16	11.8	Overseas	12070
Sandtrap-S3A-19.4.17	5.0	Overseas	52829
Sandtrap-S3A-26.2.17	16.8	Overseas	139271
Sandtrap-S3A-26.3.17	9.1	Overseas	122938
Sandtrap-S3A-27.11.16	6.7	Overseas	51998
Sandtrap-S4A-26.2.17	2.3	Overseas	55851
Sandtrap-S4A-26.3.17	5.5	Overseas	50860
Sandtrap-S5-26.2.17	4.4	Overseas	124768
Sandtrap-S6A-26.2.17	5.3	Overseas	110106
Wedge-L1-4.4.16	36.5	Auckland	538,807
Wedge-L2-14.5.17	8.3	Auckland	573207
Wedge-L2-22.2.17	0.4	Overseas	54045
Wedge-S1-14.5.17	11.8	Auckland	141382

Wedge-S1-15.12.16	3.0	Overseas	124230
Wedge-S1-28.11.16	7.4	Overseas	131308
Wedge-S1-4.4.17	9.6	Overseas	58727
Wedge-S2.2-14.5.17	24.3	Auckland	573,207
Wedge-S2.1-22.2.17	6.0	Auckland	808,525
Wedge-S2-15.12.16	20.0	Overseas	57924
Wedge-S2-22.2.17	14.0	Overseas	117044
Wedge-S2-28.11.16	14.5	Overseas	50527
Wedge-S2-4.4.17	4.7	Overseas	58727

Appendix B. Sequence information about the samples, from DNA extraction through to location of sequencing and the total number of reads returned for each sample.

Appendix C: Sample read information after the classification pipeline

Name of sample	Post UNOISE Read Count	Reads Classified (%)
Irrigation-S1-27.11.16	28042	27.8
Irrigation-S2-27.11.16	18549	18.7
Irrigation-S3-27.11.16	9411	20.2
Newpile-L1-28.11.16	13871	20.1

Newpile-L1-4.4.17	127522	25.4
Newpile-L2-14.5.17	54407	27.5
Newpile-L2-22.2.17	19741	17.2
Newpile-L2-28.11.16	12017	22.7
Newpile-L2-4.4.17	171250	34.1
Newpile-S1-14.5.17	15596	33.9
Newpile-S1-15.12.16	13557	17.8
Newpile-S1-28.11.16	7317	9.6
Newpile-S1-4.4.17	5335	7.9
Newpile-S2-14.5.17	14196	20.3
Newpile-S2-15.12.16	31967	30.2
Newpile-S2-22.2.17	13918	23.4
Newpile-S2-28.11.16	11085	16.0
Newpile-S2-4.4.17	7931	13.2
Newpile-S3-15.12.16	11036	20.7
Newpile-S4-15.12.16	167416	29.9

Pond-L1-15.12.16	77031	26.3
Pond-L1-28.11.16	84809	34.5
Pond-L3-14.5.17	108166	29.4
Pond-L4-28.11.16	100906	29.7
Pond-L4-4.4.17	16032	30.5
Sandtrap-S1A-17.12.17	18269	38.2
Sandtrap-S1A-19.4.17	15124	24.3
Sandtrap-S1A-26.2.17	30607	27.5
Sandtrap-S1A-26.3.17	22634	23.4
Sandtrap-S1A-8.10.16	20730	23.0
Sandtrap-S1B-19.4.17	10463	45.9
Sandtrap-S1B-26.2.17	8188	16.1
Sandtrap-S1B-26.3.17	14788	26.2
Sandtrap-S1B-8.10.16	19308	21.3
Sandtrap-S2A-17.12.16	29807	32.0
Sandtrap-S2A-19.4.17	16610	17.6

Sandtrap-S2A-26.2.17	14748	30.3
Sandtrap-S2A-26.3.17	7591	14.2
Sandtrap-S2A-8.10.16	13085	28.3
Sandtrap-S2B-17.12.16	20084	24.7
Sandtrap-S2B-19.4.17	18476	25.8
Sandtrap-S2B-26.2.17	20741	30.6
Sandtrap-S2B-26.3.17	8338	13.0
Sandtrap-S2B-8.10.16	17579	25.1
Sandtrap-S3A-17.12.16	29669	27.6
Sandtrap-S3A-19.4.17	8187	16.6
Sandtrap-S3A-26.2.17	18813	23.8
Sandtrap-S3A-26.3.17	31570	30.2
Sandtrap-S3A-27.11.16	10223	21.1
Sandtrap-S4A-26.2.17	13810	26.1
Sandtrap-S4A-26.3.17	11008	23.0
Sandtrap-S5-26.2.17	24631	23.2

Sandtrap-S6A-26.2.17	28967	30.4
Wedge-L1-4.4.16	141209	33.3
Wedge-L2-14.5.17	127951	28.3
Wedge-L2-22.2.17	15443	30.9
Wedge-S1-14.5.17	21363	11.2
Wedge-S1-15.12.16	34189	32.2
Wedge-S1-28.11.16	23357	22.1
Wedge-S1-4.4.17	6080	10.9
Wedge-S2.2-14.5.17	204385	23.8
Wedge-S2.2-22.2.17	4757	8.8
Wedge-S2-15.12.16	24909	24.7
Wedge-S2-22.2.17	24904	25.4
Wedge-S2-28.11.16	10555	23.1
Wedge-S2-4.4.17	8836	16.4

Appendix C. Read numbers after UNOISE and classification percentage of total for the samples from LIC. This is after following the UNOISE pipeline, with the minimum size of reads in a Zotu modified to 4 instead of 8.

References

Akinyemi, O., Okorhi-Damisa, Efemena, T., & Adeniyi, S. (2020). Identication and morphology of pathogens in liquid effluent from a Cow dung biodigester. *Biological Sciences and Pharmaceutical Research*, *8*(3). doi:10.15739/ibspr.20.006

Aly, S. S., Anderson, R. J., Whitlock, R. H., Fyock, T. L., McAdams, S. C., Byrem, T. M., … Gardner, I. A. (2012). Cost-effectiveness of diagnostic strategies to identify Mycobacterium avium subspecies paratuberculosis super-shedder cows in a large dairy herd using antibody enzyme-linked immunosorbent assays, quantitative real-time polymerase chain reaction, and bacterial culture. *Journal of Veterinary Diagnostic Investigation: Official Publication of the American Association of Veterinary Laboratory Diagnosticians, Inc*, *24*(5), 821–832.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30.

Anderson, M. J., & Willis, T. J. (2003). Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology. *Ecology*, *84*(2), 511–525.

Andrade, B. G. N., Bressani, F. A., Cuadrat, R. R. C., Tizioto, P. C., de Oliveira, P. S. N., Mourão, G. B., … Regitano, L. C. A. (2020). The structure of microbial populations in Nelore GIT reveals inter-dependency of methanogens in feces and rumen. *Journal of Animal Science and Biotechnology*, *11*, 6.

Andrews, S., & Others. (2010). *FastQC: a quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Arnold, J. W., Roach, J., & Azcarate-Peril, M. A. (2016). Emerging Technologies for Gut Microbiome Research. *Trends in Microbiology*, *24*(11), 887–901.

Banerjee, S., Schlaeppi, K., & van der Heijden, M. G. A. (2018). Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews. Microbiology*, *16*(9), 567–576.

Barratt, A. S., Arnoult, M. H., Ahmadi, B. V., Rich, K. M., Gunn, G. J., & Stott, A. W. (2018). A framework for estimating society's economic welfare following the introduction of an animal disease: The case of Johne's disease. *PloS One*, *13*(6), e0198436.

Bates, A., O'Brien, R., Liggett, S., & Griffin, F. (2018). The effect of sub-clinical infection with Mycobacterium avium subsp. paratuberculosis on milk production in a New Zealand dairy herd. *BMC Veterinary Research*, *14*(1), 93.

Beckers, K. F., Schulz, C. J., & Childers, G. W. (2017). Rapid regrowth and detection of microbial contaminants in equine fecal microbiome samples. *PloS One*, *12*(11), e0187044.

Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics*, *5*(4), 433–438.

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., … Schloter, M. (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, *8*(1), 103.

Bharti, R., & Grimm, D. G. (2019). Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*. doi:10.1093/bib/bbz155

Blanchard, P. C. (2012). Diagnostics of Dairy and Beef Cattle Diarrhea. *The Veterinary Clinics of North America. Food Animal Practice*, *28*(3), 443–464.

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., … Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 90.

Bokulich, N. A., Ziemski, M., Robeson, M. S., 2nd, & Kaehler, B. D. (2020). Measuring the microbiome: Best practices for developing and benchmarking microbiomics methods. *Computational and Structural Biotechnology Journal*, *18*, 4048–4062.

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

Breitwieser, F. P., Baker, D. N., & Salzberg, S. L. (2018). KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*, *19*(1), 198.

Breitwieser, Florian P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, *20*(4), 1125–1136.

Breitwieser, Florian P., & Salzberg, S. L. (2020). Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics* , *36*(4), 1303–1304.

Britton, L. E., Cassidy, J. P., O'Donovan, J., Gordon, S. V., & Markey, B. (2016). Potential application of emerging diagnostic techniques to the diagnosis of bovine Johne's disease (paratuberculosis). *Veterinary Journal* , *209*, 32–39.

Brumfield, K. D., Huq, A., Colwell, R. R., Olds, J. L., & Leddy, M. B. (2020). Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PloS One*, *15*(2), e0228899.

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

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336.

Cattonaro, F., Spadotto, A., Radovic, S., & Marroni, F. (2018). Do you cov me? Effect of coverage reduction on species identification and genome reconstruction in complex biological matrices by metagenome shotgun high-throughput sequencing. *F1000Research*, *7*(1767), 1767.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* , *34*(17), i884–i890.

Chi, J., VanLeeuwen, J. A., Weersink, A., & Keefe, G. P. (2002). Direct production losses and treatment costs from bovine viral diarrhoea virus, bovine leukosis virus, Mycobacterium avium subspecies paratuberculosis, and Neospora caninum. *Preventive Veterinary Medicine*, *55*(2), 137–153.

Chng, K. R., Li, C., Bertrand, D., Ng, A. H. Q., Kwah, J. S., Low, H. M., … Nagarajan, N. (2020). Cartography of opportunistic pathogens and antibiotic resistance genes in a tertiary hospital environment (p. 644740). doi:10.1101/644740

Choo, J. M., Leong, L. E. X., & Rogers, G. B. (2015). Sample storage conditions significantly influence faecal microbiome profiles. *Scientific Reports*, *5*, 16350.

Clarridge, J. E., 3rd. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, *17*(4), 840–862, table of contents.

Collins, M. T., Eggleston, V., & Manning, E. J. B. (2010). Successful control of Johne's disease in nine dairy herds: results of a six-year field trial. *Journal of Dairy Science*, *93*(4), 1638–1643.

DairyNZ. (2021). Effluent Systems. Retrieved February 26, 2021, from DairyNZ website: https://www.dairynz.co.nz/environment/effluent/

Davidson, R. M., & Epperson, L. E. (2018). Microbiome Sequencing Methods for Studying Human Diseases. *Methods in Molecular Biology* , *1706*, 77–90.

de Muinck, E. J., Trosvik, P., Gilfillan, G. D., Hov, J. R., & Sundaram, A. Y. M. (2017). A novel ultra high-throughput 16S rRNA gene amplicon sequencing library preparation method for the Illumina HiSeq platform. *Microbiome*, *5*(1), 68.

Degnan, P. H., & Ochman, H. (2012). Illumina-based analysis of microbial community diversity. *The ISME Journal*, *6*(1), 183–194.

Derrick Wood Nick Loman. (2018, August). Mock Community. Retrieved 2019, 2020, from Github website: https://github.com/LomanLab/mockcommunity

Di Menna, M. E., Smith, B. L., & Miles, C. O. (2009). A history of facial eczema (pithomycotoxicosis) research. *New Zealand Journal of Agricultural Research*, *52*(4), 345–376.

Dill-McFarland, K. A., Breaker, J. D., & Suen, G. (2017). Microbial succession in the gastrointestinal tract of dairy cows from 2 weeks to first lactation. *Scientific Reports*, *7*, 40864.

Dowd, S. E., Callaway, T. R., Wolcott, R. D., Sun, Y., McKeehan, T., Hagevoort, R. G., & Edrington, T. S. (2008). Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). *BMC Microbiology*, *8*, 125.

Durso, L. M., Harhay, G. P., Smith, T. P. L., Bono, J. L., Desantis, T. Z., Harhay, D. M., … Clawson, M. L. (2010). Animal-to-animal variation in fecal microbial diversity among beef cattle. *Applied and Environmental Microbiology*, *76*(14), 4858–4862.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* , *26*(19), 2460–2461.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–998.

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing (p. 081257). doi:10.1101/081257

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. *PeerJ*, *5*, e3889.

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* , *34*(14), 2371–2375.

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., & Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current Opinion in Microbiology*, *25*, 56–66.

Fecteau, M.-E., Pitta, D. W., Vecchiarelli, B., Indugu, N., Kumar, S., Gallagher, S. C., … Sweeney, R. W. (2016). Dysbiosis of the Fecal Microbiota in Cattle Infected with Mycobacterium avium subsp. paratuberculosis. *PloS One*, *11*(8), e0160353.

Fock-Chow-Tho, D., Topp, E., Ibeagha-Awemu, E. A., & Bissonnette, N. (2017). Comparison of commercial DNA extraction kits and quantitative PCR systems for better sensitivity in detecting the causative agent of paratuberculosis in dairy cow fecal samples. *Journal of Dairy Science*, *100*(1), 572–581.

Fomenky, B. E., Do, D. N., Talbot, G., Chiquette, J., Bissonnette, N., Chouinard, Y. P., … Ibeagha-Awemu, E. M. (2018). Direct-fed microbial supplementation influences the bacteria community composition of the gastrointestinal tract of pre- and post-weaned calves. *Scientific Reports*, *8*(1), 14147.

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., & Loeb, L. A. (2014). Accuracy of next generation sequencing platforms. *Next Generation, Sequencing & Applications*, *1*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/pmc4331009/

Franzén, O., Hu, J., Bao, X., Itzkowitz, S. H., Peter, I., & Bashir, A. (2015). Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. *Microbiome*, *3*, 43.

Gao, X., Lin, H., Revanna, K., & Dong, Q. (2017). A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. *BMC Bioinformatics*, *18*(1), 247.

Georges, M., Charlier, C., & Hayes, B. (2019). Harnessing genomic information for livestock improvement. *Nature Reviews. Genetics*, *20*(3), 135–156.

Gibbons, S. M. (2020). [Review of *Keystone taxa indispensable for microbiome recovery*]. *Nature microbiology*, *5*(9), 1067–1068.

Girija, D., Deepa, K., Xavier, F., Antony, I., & Shidhi, P. R. (2013). *Analysis of cow dung microbiota—a metagenomic approach*. Retrieved from http://nopr.niscair.res.in/handle/123456789/21863

Goldford, J. E., Lu, N., Bajić, D., Estrela, S., Tikhonov, M., Sanchez-Gorostiaga, A., … Sanchez, A. (2018). Emergent simplicity in microbial community assembly. *Science*, *361*(6401), 469–474.

Golob, J. L., Margolis, E., Hoffman, N. G., & Fredricks, D. N. (2017). Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities. *BMC Bioinformatics*, *18*(1), 283.

Gomez, D. E., Arroyo, L. G., Costa, M. C., Viel, L., & Weese, J. S. (2017). Characterization of the Fecal Bacterial Microbiota of Healthy and Diarrheic Dairy Calves. *Journal of Veterinary Internal Medicine / American College of Veterinary Internal Medicine*, *31*(3), 928–939.

Gonda, M. G., Chang, Y. M., Shook, G. E., Collins, M. T., & Kirkpatrick, B. W. (2007). Effect of Mycobacterium paratuberculosis infection on production, reproduction, and health traits in US Holsteins. *Preventive Veterinary Medicine*, *80*(2–3), 103–119.

Goodrich, J. K., Di Rienzi, S. C., Poole, A. C., Koren, O., Walters, W. A., Caporaso, J. G., … Ley, R. E. (2014). Conducting a microbiome study. *Cell*, *158*(2), 250–262.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews. Genetics*, *17*(6), 333–351.

Guo, W., Zhou, M., Ma, T., Bi, S., Wang, W., Zhang, Y., … Long, R. (2020). Survey of rumen microbiota of domestic grazing yak during different growth stages revealed novel maturation patterns of four key microbial groups and their dynamic interactions. *Animal Microbiome*, *2*(1), 23.

Han, D., Gao, P., Li, R., Tan, P., Xie, J., Zhang, R., & Li, J. (2020). Multicenter assessment of microbial community profiling using 16S rRNA gene sequencing and shotgun metagenomic sequencing. *Journal of Advertising Research*, *26*, 111–121.

Hang, J., Desai, V., Zavaljevski, N., Yang, Y., Lin, X., Satya, R. V., … Kuschner, R. A. (2014). 16S rRNA gene pyrosequencing of reference and clinical samples and investigation of the temperature stability of microbiome profiles. *Microbiome*, *2*, 31.

Hao, Y., Pei, Z., & Brown, S. M. (2017). Chapter 1 - Bioinformatics in Microbiome Analysis. In C. Harwood (Ed.), *Methods in Microbiology* (Vol. 44, pp. 1–18). Academic Press.

Hardwick, S. A., Chen, W. Y., Wong, T., Kanakamedala, B. S., Deveson, I. W., Ongley, S. E., … Mercer, T. R. (2018). Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nature Communications*, *9*(1), 3096.

Harris, N. B., & Barletta, R. G. (2001). Mycobacterium avium subsp. paratuberculosis in Veterinary Medicine. *Clinical Microbiology Reviews*, *14*(3), 489–512.

Hart, M. L., Meyer, A., Johnson, P. J., & Ericsson, A. C. (2015). Comparative Evaluation of DNA Extraction Methods from Feces of Multiple Host Species for Downstream Next-Generation Sequencing. *PloS One*, *10*(11), e0143334.

Henderson, G., Cox, F., Ganesh, S., Jonker, A., Young, W., Global Rumen Census Collaborators, & Janssen, P. H. (2015). Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Scientific Reports*, *5*, 14567.

Hermans, S. M., Buckley, H. L., Case, B. S., Curran-Cournane, F., Taylor, M., & Lear, G. (2020). Using soil bacterial communities to predict physico-chemical variables and soil quality. *Microbiome*, *8*(1), 79.

Holman, D. B., & Gzyl, K. E. (2019). A meta-analysis of the bovine gastrointestinal tract microbiota. *FEMS Microbiology Ecology*, *95*(6). doi:10.1093/femsec/fiz072

Illumina. (2013, November 27). *16S Metagenomic Sequencing Library Preparation*. Retrieved from https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239.

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

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., … -S. Wong, G. K. (2016). Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in Microbiology*, Vol. 7. doi:10.3389/fmicb.2016.00459

Kembel, S. W., Wu, M., Eisen, J. A., & Green, J. L. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Computational Biology*, *8*(10), e1002743.

Kim, D., Hofstaedter, C. E., Zhao, C., Mattei, L., Tanes, C., Clarke, E., … Bittinger, K. (2017). Optimizing methods and dodging pitfalls in microbiome research. *Microbiome*, *5*(1), 52.

Kim, M., Kim, J., Kuehn, L. A., Bono, J. L., Berry, E. D., Kalchayanand, N., … Wells, J. E. (2014). Investigation of bacterial diversity in the feces of cattle fed different diets1. *American Society of Animal Science*. doi:10.2527/jas2013-6841

Kim, Minseok, Park, T., & Yu, Z. (2017). — Invited Review — Metagenomic investigation of gastrointestinal microbiome in cattle. *Asian-Australasian Journal of Animal Sciences (AJAS)*, *30*(11), 1–14.

Kim, Minseok, & Wells, J. E. (2016). A Meta-analysis of Bacterial Diversity in the Feces of Cattle. *Current Microbiology*, *72*(2), 145–151.

Kindt, R., & Coe, R. (2005). *Tree Diversity Analysis: A Manual and Software for Common Statistical Methods for Ecological and Biodiversity Studies*. World Agroforestry Centre.

Koester, L. R., Poole, D. H., Serão, N. V. L., & Schmitz-Esser, S. (2020). Beef cattle that respond differently to fescue toxicosis have distinct gastrointestinal tract microbiota (p. 2020.02.03.932939). doi:10.1101/2020.02.03.932939

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* . Retrieved from https://academic.oup.com/bioinformatics/article-abstract/28/19/2520/290322

Kozlov, A. M., Zhang, J., Yilmaz, P., Glöckner, F. O., & Stamatakis, A. (2016). Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research*, *44*(11), 5022–5033.

Kruze, J., Monti, G., Schulze, F., Mella, A., & Leiva, S. (2013). Herd-level prevalence of Map infection in dairy herds of southern Chile determined by culture of environmental fecal samples and bulk-tank milk qPCR. *Preventive Veterinary Medicine*, *111*(3–4), 319–324.

Lagier, J.-C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., … Raoult, D. (2018). Culturing the human microbiota and culturomics. *Nature Reviews. Microbiology*, *16*, 540–550.

Larsson, D. G. J., Andremont, A., Bengtsson-Palme, J., Brandt, K. K., de Roda Husman, A. M., Fagerstedt, P., … Wernersson, A.-S. (2018). Critical knowledge gaps and research needs related to the environmental dimensions of antibiotic resistance. *Environment International*, *117*, 132–138.

Lavy, A., Matheus Carnevali, P. B., Keren, R., Bill, M., Wan, J., Tokunaga, T. K., … Banfield, J. F. (2020). Taxonomically and metabolically distinct microbial communities with depth and across a hillslope to riparian zone transect (p. 768572). doi:10.1101/768572

Lawson, C. E., Harcombe, W. R., Hatzenpichler, R., Lindemann, S. R., Löffler, F. E., O'Malley, M. A., … McMahon, K. D. (2019). Common principles and best practices for engineering microbiomes. *Nature Reviews. Microbiology*, *17*(12), 725–741.

Lazcka, O., Del Campo, F. J., & Muñoz, F. X. (2007). Pathogen detection: a perspective of traditional methods and biosensors. *Biosensors & Bioelectronics*, *22*(7), 1205–1217.

Lear, G., Dickie, I., Banks, J., Boyer, S., Buckley, H., Buckley, T., … Holdaway, R. (2018). Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. *New Zealand Journal of Ecology*. doi:10.20417/nzjecol.42.9

Leclercq, S. O., Wang, C., Sui, Z., Wu, H., Zhu, B., Deng, Y., & Feng, J. (2016). A multiplayer game: species of Clostridium, Acinetobacter, and Pseudomonas are responsible for the persistence of antibiotic resistance genes in manure-treated soils. *Environmental Microbiology*, *18*(10), 3494–3508.

Leggett, R. M., & Clark, M. D. (2017). A world of opportunities with nanopore sequencing. *Journal of Experimental Botany*, *68*(20), 5419–5429.

Li, H., Li, R., Chen, H., Gao, J., Wang, Y., Zhang, Y., & Qi, Z. (2020). Effect of different seasons (spring vs summer) on the microbiota diversity in the feces of dairy cows. *International Journal of Biometeorology*, *64*(3), 345–354.

Liu, T., Chen, C.-Y., Chen-Deng, A., Chen, Y.-L., Wang, J.-Y., Hou, Y.-I., & Lin, M.-C. (2020). Joining Illumina paired-end reads for classifying phylogenetic marker sequences. *BMC Bioinformatics*, *21*(1), 105.

López-García, A., Pineda-Quiroga, C., Atxaerandio, R., Pérez, A., Hernández, I., García-Rodríguez, A., & González-Recio, O. (2018). Comparison of Mothur and QIIME for the Analysis of Rumen Microbiota Composition Based on 16S rRNA Amplicon Sequences. *Frontiers in Microbiology*, *9*, 3010.

Losinger, W. C. (2006). Welfare effects of reduced milk production associated with Johne's disease on Johne's-positive versus Johne's-negative dairy operations. *The Journal of Dairy Research*, *73*(3), 378–384.

Lu, J., Breitwieser, F. P., Wood, D. E., Kim, D., Langmead, B., & Salzberg, S. L. (2018, December 29). *How to Choose Your Metagenomics Classification Tool*. Retrieved from http://ccb.jhu.edu/software/choosing-a-metagenomics-classifier/

Lu, J., & Salzberg, S. L. (2020). Ultrafast and accurate 16S microbial community analysis using Kraken 2 (p. 2020.03.27.012047). doi:10.1101/2020.03.27.012047

Magurran, A. E. (1988). Diversity indices and species abundance models. In A. E. Magurran (Ed.), *Ecological Diversity and Its Measurement* (pp. 7–46). Croom Helm Ltd.

Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., Jr, Saxman, P. R., Stredwick, J. M., … Tiedje, J. M. (2000). The RDP (Ribosomal Database Project) continues. *Nucleic Acids Research*, *28*(1), 173–174.

Malla, M. A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., & Abd Allah, E. F. (2018). Exploring the Human Microbiome: The Potential Future Role of Next-Generation Sequencing in Disease Diagnosis and Treatment. *Frontiers in Immunology*, *9*, 2868.

Mao, S., Zhang, R., Wang, D., & Zhu, W. (2012). The diversity of the fecal bacterial community and its relationship with the concentration of volatile fatty acids in the feces during subacute rumen acidosis in dairy cows. *BMC Veterinary Research*, *8*, 237.

Martinez-Porchas, M., Villalpando-Canchola, E., Ortiz Suarez, L. E., & Vargas-Albores, F. (2017). How conserved are the conserved 16S-rRNA regions? *PeerJ*, *5*, e3036.

McGovern, E., Waters, S. M., Blackshields, G., & McCabe, M. S. (2018). Evaluating Established Methods for Rumen 16S rRNA Amplicon Sequencing With Mock Microbial Populations. *Frontiers in Microbiology*, *9*, 1365.

McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *ELife*, *8*. doi:10.7554/eLife.46923

McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS One*, *8*(4), e61217.

Méric, G., Wick, R. R., Watts, S. C., Holt, K. E., & Inouye, M. (2019). Correcting index databases improves metagenomic studies (p. 712166). doi:10.1101/712166

Microbiome. (2018). Retrieved December 10, 2020, from Nature.com website: https://www.nature.com/subjects/microbiome

Milanese, A., Mende, D. R., Paoli, L., Salazar, G., Ruscheweyh, H.-J., Cuenca, M., … Sunagawa, S. (2019). Microbial abundance, activity and population genomic profiling with mOTUs2. *Nature Communications*, *10*(1), 1014.

Morris, C. A., Phua, S. H., Cullen, N. G., & Towers, N. R. (2013). Review of genetic studies of susceptibility to facial eczema in sheep and dairy cattle. *New Zealand Journal of Agricultural Research*, *56*(2), 156–170.

Munk, P., Knudsen, B. E., Lukjancenko, O., Duarte, A. S. R., Van Gompel, L., Luiken, R. E. C., … Aarestrup, F. M. (2018). Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries. *Nature Microbiology*, *3*(8), 898–908.

Muñoz-Vargas, L., Opiyo, S. O., Digianantonio, R., Williams, M. L., Wijeratne, A., & Habing, G. (2018). Fecal microbiome of periparturient dairy cattle and associations with the onset of Salmonella shedding. *PloS One*, *13*(5), e0196171.

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. I. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, *6*, e5364.

Nicholls, S. M., Quick, J. C., Tang, S., & Loman, N. J. (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience*, *8*, 1–9.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Others. (2019). *vegan: Community Ecology Package. R package version 2.5--6. 2019*.

Paine, R. T. (1966). Food Web Complexity and Species Diversity. *The American Naturalist*, *100*(910), 65–75.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* , *35*(3), 526–528.

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., & Hugenholtz, P. (2019). Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy (p. 771964). doi:10.1101/771964

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996–1004.

Patel, M., Patel, H. M., Vohra, N., & Dave, S. (2020). Complete genome sequencing and comparative genome characterization of the lignocellulosic biomass degrading bacterium Pseudomonas stutzeri MP4687 from cattle rumen. *Biotechnology Reports*, Vol. 28, p. e00530. doi:10.1016/j.btre.2020.e00530

Patterson, S., Bond, K., Green, M., van Winden, S., & Guitian, J. (2020). Mycobacterium avium paratuberculosis infection of calves - The impact of dam infection status. *Preventive Veterinary Medicine*, *181*, 104634.

Peabody, M. A., Van Rossum, T., Lo, R., & Brinkman, F. S. L. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics*, *16*, 363.

Pearman, W. S., Freed, N. E., & Silander, O. K. (2020). Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics*, *21*(1), 220.

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., & Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, *8*(1), 10950.

Pitta, D. W., Indugu, N., Kumar, S., Vecchiarelli, B., Sinha, R., Baker, L. D., … Ferguson, J. D. (2016). Metagenomic assessment of the functional potential of the rumen microbiome in Holstein dairy cows. *Anaerobe*, *38*, 50–60.

Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (2018). The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Applied and Environmental Microbiology*, *84*(7). doi:10.1128/AEM.02627-17

Porwal, H. J., Mane, A. V., & Velhal, S. G. (2015). Biodegradation of dairy effluent by using microbial isolates obtained from activated sludge. *Water Resources and Industry*, *9*, 1–15.

Poulsen, C. S., Pamp, S. J., Ekstrøm, C. T., & Aarestrup, F. M. (2019). Library preparation and sequencing platform introduce bias in metagenomics characterisation of microbial communities (p. 592154). doi:10.1101/592154

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PloS One*, *15*(1), e0227434.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), 833–844.

R Core Team. (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Rajan, S. K., Lindqvist, M., Brummer, R. J., Schoultz, I., & Repsilber, D. (2019). Phylogenetic microbiota profiling in fecal samples depends on combination of sequencing depth and choice of NGS analysis method. *PloS One*, *14*(9), e0222171.

Raju, S. C., Lagström, S., Ellonen, P., de Vos, W. M., Eriksson, J. G., Weiderpass, E., & Rounge, T. B. (2018). Reproducibility and repeatability of six high-throughput 16S rDNA sequencing protocols for microbiota profiling. *Journal of Microbiological Methods*, *147*, 76–86.

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, *469*(4), 967–977.

Ren, G., Xu, X., Qu, J., Zhu, L., & Wang, T. (2016). Evaluation of microbial population dynamics in the co-composting of cow manure and rice straw using high throughput sequencing analysis. *World Journal of Microbiology & Biotechnology*, *32*(6), 101.

Ricchi, M., Bertasio, C., Boniotti, M. B., Vicari, N., Russo, S., Tilola, M., … Bertasi, B. (2017). Comparison among the Quantification of Bacterial Pathogens by qPCR, dPCR, and Cultural Methods. *Frontiers in Microbiology*, *8*, 1174.

Ritter, C. D., Faurby, S., Bennett, D. J., Naka, L. N., Ter Steege, H., Zizka, A., … Antonelli, A. (2019). The pitfalls of biodiversity proxies: Differences in richness patterns of birds, trees and understudied diversity across Amazonia. *Scientific Reports*, *9*(1), 19205.

Robeson, M. S., O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, R., Foster, J. T., & Bokulich, N. A. (2020). 1 RESCRIPt: Reproducible sequence taxonomy reference database management for the 2 masses. *BioRxiv Preprint*. doi:10.1101/2020.10.05.326504

Romagnoli, E. M., Kmit, M. C. P., Chiaramonte, J. B., Rossmann, M., & Mendes, R. (2017). Ecological Aspects on Rumen Microbiome. *Diversity and Benefits of Microorganisms from the Tropics*, pp. 367–389. doi:10.1007/978-3-319-55804-2_16

Ross, E. M., Moate, P. J., Marett, L. C., Cocks, B. G., & Hayes, B. J. (2013). Metagenomic predictions: from microbiome to complex health and environmental phenotypes in humans and cattle. *PloS One*, *8*(9), e73056.

Santos, A., van Aerle, R., Barrientos, L., & Martinez-Urtaza, J. (2020). Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Computational and Structural Biotechnology Journal*, *18*, 296–305.

Sasaki, H., Kitazume, O., Sasaki, T., & Nakai, Y. (2004). Ammonia-assimilating microbes in microbial community in a lagoon for wastewater from paddock of dairy cattle. *Animal Science Journal = Nihon Chikusan Gakkaiho*, *75*(1), 79–84.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., … Ye, J. (2011). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *39*(Database issue), D38-51.

Sbardellati, D. L., Fischer, A., Cox, M. S., Li, W., Kalscheur, K. F., & Suen, G. (2020). The bovine epimural microbiota displays compositional and structural heterogeneity across different ruminal locations. *Journal of Dairy Science*, *103*(4), 3636–3647.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541.

Schriefer, A. E., Cliften, P. F., Hibberd, M. C., Sawyer, C., Brown-Kennerly, V., Burcea, L., … Head, R. D. (2018). A multi-amplicon 16S rRNA sequencing and analysis method for improved taxonomic profiling of bacterial communities. *Journal of Microbiological Methods*, *154*, 6–13.

Seyboldt, C., Discher, S., Jordan, E., Neubauer, H., Jensen, K. C., Campe, A., … Hoedemaker, M. (2015). Occurrence of Clostridium botulinum neurotoxin in chronic disease of dairy cows. *Veterinary Microbiology*, *177*(3–4), 398–402.

Shahi, S. K., Freedman, S. N., & Mangalam, A. K. (2017). Gut microbiome in multiple sclerosis: The players involved and the roles they play. *Gut Microbes*, *8*(6), 607–615.

Shanks, O. C., Kelty, C. A., Archibeque, S., Jenkins, M., Newton, R. J., McLellan, S. L., … Sogin, M. L. (2011). Community structures of fecal bacteria in cattle from different animal feeding operations. *Applied and Environmental Microbiology*, *77*(9), 2992–3001.

Song, E.-S., Jung, S. I., Park, H.-J., Seo, K.-W., Son, J.-H., Hong, S., … Song, K.-H. (2016). Comparison of Fecal Microbiota between German Holstein Dairy Cows with and without Left-Sided Displacement of the Abomasum. *Journal of Clinical Microbiology*, *54*(4), 1140–1143.

Song, S. J., Amir, A., Metcalf, J. L., Amato, K. R., Xu, Z. Z., Humphrey, G., & Knight, R. (2016). Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. *MSystems*, *1*(3). doi:10.1128/mSystems.00021-16

Soon Gweon, H., Shaw, L. P., Swann, J., De Maio, N., Oun, M. A., Hubbard, A. T. M., … on behalf of the REHAB consortium. (2019). The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. *BioRxiv Preprint*. doi:10.1101/593301

Stinson, K. J., Baquero, M. M., & Plattner, B. L. (2018). Resilience to infection by Mycobacterium avium subspecies paratuberculosis following direct intestinal inoculation in calves. *Veterinary Research*, *49*(1), 58.

Van Rossum, T., Ferretti, P., Maistrenko, O. M., & Bork, P. (2020). Diversity within species: interpreting strains in microbiomes. *Nature Reviews. Microbiology*, *18*(9), 491–506.

Venkataraman, A., Parlov, M., Hu, P., Schnell, D., Wei, X., & Tiesman, J. P. (2018). Spike-in genomic DNA for validating performance of metagenomics workflows. *BioTechniques*, *65*(6), 315–321.

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*(2), e57923.

Vikram, A., Rovira, P., Agga, G. E., Arthur, T. M., Bosilevac, J. M., Wheeler, T. L., … Schmidt, J. W. (2017). Impact of “Raised without Antibiotics” Beef Cattle Production Practices on Occurrences of Antimicrobial Resistance. *Applied and Environmental Microbiology*, *83*(22). doi:10.1128/AEM.01682-17

Vo, A.-T. E., & Jedlicka, J. A. (2014). Protocols for metagenomic DNA extraction and Illumina amplicon library preparation for faecal and swab samples. *Molecular Ecology Resources*, *14*(6), 1183–1197.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162.

Wesolowska-Andersen, A., Bahl, M. I., Carvalho, V., Kristiansen, K., Sicheritz-Pontén, T., Gupta, R., & Licht, T. R. (2014). Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome*, *2*, 19.

Whipps, J. M., Lewis, K., & Cooke, R. C. (1988). Mycoparasitism and plant disease control. *Fungi in Biological Control Systems*, 161–187.

Wong, K., Shaw, T. I., Oladeinde, A., Glenn, T. C., Oakley, B., & Molina, M. (2016). Rapid Microbiome Changes in Freshly Deposited Cow Feces under Field Conditions. *Frontiers in Microbiology*, *7*, 500.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257.

Wu, W.-K., Chen, C.-C., Panyod, S., Chen, R.-A., Wu, M.-S., Sheen, L.-Y., & Chang, S.-C. (2019). Optimization of fecal sample processing for microbiome study - The journey from bathroom to bench. *Journal of the Formosan Medical Association = Taiwan Yi Zhi*, *118*(2), 545–555.

Yang, F., Sun, J., Luo, H., Ren, H., Zhou, H., Lin, Y., … Zhong, H. (2020). Assessment of fecal DNA extraction protocols for metagenomic studies. *Gigascience*. doi:10.5524/100742

Yee, R., Breitwieser, F. P., Hao, S., Opene, B. N. A., Workman, R. E., Tamma, P. D., … Simner, P. J. (2020). Metagenomic Next-Generation Sequencing of Rectal Swabs for the Surveillance of Antimicrobial Resistant Organisms on the Illumina Miseq and Oxford MinION Platforms (p. 2020.04.16.044214). doi:10.1101/2020.04.16.044214

Zaheer, R., Lakin, S. M., Polo, R. O., Cook, S. R., Larney, F. J., Morley, P. S., … McAllister, T. A. (2019). Comparative diversity of microbiomes and Resistomes in beef feedlots, downstream environments and urban sewage influent. *BMC Microbiology*, *19*(1), 197.

Zeineldin, M., Aldridge, B., & Lowe, J. (2018a). Dysbiosis of the fecal microbiota in feedlot cattle with hemorrhagic diarrhea. *Microbial Pathogenesis*, *115*, 123–130.

Zeineldin, M., Aldridge, B., & Lowe, J. (2018b). Dysbiosis of the fecal microbiota in feedlot cattle with hemorrhagic diarrhea. *Microbial Pathogenesis*, *115*, 123–130.

Zeineldin, M., Barakat, R., Elolimy, A., Salem, A. Z. M., Elghandour, M. M. Y., & Monroy, J. C. (2018). Synergetic action between the rumen microbiota and bovine health. *Microbial Pathogenesis*, *124*, 106–115.

Zhao, H.-Y., Li, J., Liu, J.-J., Lü, Y.-C., Wang, X.-F., & Cui, Z.-J. (2013). Microbial Community Dynamics During Biogas Slurry and Cow Manure Compost. *Journal of Integrative Agriculture*, *12*(6), 1087–1097.

Zymo Research Corporation. (2017). *ZymoBIOMICS Microbial Community Standard (Cat 6300)*. Retrieved from https://files.zymoresearch.com/protocols/_d6300_zymobiomics_microbial_community_standard.pdf

Zymo Research Corporation. (2019). *ZymoBIOMICS Spike-in Control I (High Microbial Load) (Cat 6320)*. Retrieved from https://files.zymoresearch.com/protocols/_d6320_zymobiomics_spike-in_control_i.pdf