Metagenomic Analysis of Ruminal Microbiota Composition Based on 16S rRNA Amplicon Sequences Using Qiime2 Pipeline

Abstract: The emergence of bioinformatic analysis pipelines in recent years have allowed researchers to generate insightful data regarding microbiomes in a cost-effective manner. QIIME (Quantitative Insights Into Microbial Ecology) is one of the bioinformatic analysis pipelines that is mentioned frequently in the science community. As the successor of QIIME, QIIME2 is expected to be an even more robust and sophisticated pipeline for analyzing microbiomes. Even though there are many suggested pipelines for QIIME2, each pipeline has different workflows and requirements that make using QIIME2 difficult for inexperienced users. In this study, a QIIME2 pipeline suitable for the analysis of ruminal microbiota will be constructed, using data provided by Lopez-Garcia et al.. This pipeline will enable users to become familiar with QIIME2, generate a taxonomic profile of ruminal microbiota, and create a phylogenetic tree.

Keywords: Metagenomic analysis, Ruminal microbiome, 16S rRNA amplicon sequences, Bioinformatics pipeline, QIIME2, Taxonomy, Phylogeny.

1 Introduction

Cellulose-rich feeds are digested with the help of rumen microbiota in cattle. By doing so, digested metabolites are used by the microbes themselves to proliferate and the host cattle use them as a source of nutrients. The symbiotic relationship evident here is critical for cattle's survival. Studying the composition of rumen microbiome is an interest posed by researchers around the world because it has many implications, such as its effect on climate change, animal production and the health of cattle (McGovern et al. 2018).

The field of microbiome analysis has improved immensely due to innovations in computing and sequencing technology. Researchers can determine the microbial diversity of a community by sequencing conserved regions of the microbial genome that can reveal information about their phylogeny (de Muinck et al. 2017). This conserved region that is most commonly sequenced in the science community is the 16S rRNA gene (de Muinck et al. 2017, Janda & Abbott 2007). Nowadays, the variable region 4 (V4) that is present in the 16S rRNA gene is amplified and sequenced most frequently, especially the 515 to 806 fragment. This is due to the general consensus in the science community that the indicated region will provide the most optimal classification (de Muinck et al. 2017).

In order to know what type of microorganisms are present in a sample, researchers must assign taxonomy by analyzing the sequenced reads of the 16S rRNA gene. This is done with the help of bioinformatics analysis pipelines, which are ever-increasing in number (Siegwald et al. 2017). Recently, the bioinformatics tool QIIME has received positive feedback from active researchers on the topic of metagenomics analysis (Caporaso et al. 2010). By examining the DNA of microbes, researchers

Copyright © 201X Inderscience Enterprises Ltd.

can determine the composition of microbes in a given community (Kuczynski et al. 2011). QIIME2 is a successor of QIIME and it is considered to provide a robust and sophisticated method for 16S rRNA gene analysis (Hall & Beiko 2018, Bolyen et al. 2018).

1.1 Complexity of QIIME2 Pipelines

Many suggested pipelines are available on the Internet for QIIME2. However, each pipeline is suited to cater for a specific type of input, has different plug-ins and algorithms, and steps in the workflow are constructed differently. In addition, even if the general workflow may be similar in each suggested pipeline, the procedure required to generate the same outcome may not be the same. As a result, beginners to the field of bioinformatics will have great difficulty in using QIIME2 to examine microbiomes.

1.2 Our Contributions

In this study, a bioinformatics pipeline revolving around QIIME2 will be constructed. It will be constructed in a manner that will be suitable for the analysis of the microbial composition of cow rumen using read data availalbe on EBI database (López-García et al. 2018). A taxonomic profile and phylogenetic tree of ruminal microbiota will be generated as the final outcome.

2 Constructing A Bioinformatics Pipeline

Via literature search and inspection of other QIIME2 workflows, a QIIME2 pipeline for the analysis of 16S rRNA gene sequences was constructed (Fig. 3, 4). The pipeline consists of six major steps: obtaining sequencing data, trimming poor quality sequences, converting reads into QIIME2 artifact, denoising and quality control filtering, assinging taxonomy, and generating a phylogenetic tree. Visualization of the output file from each step is recommended to keep track of the process. It should be kept in mind that metadata files obtained from QIIME2 visualizations can be used for analysis on other tools.

The workflow can be divided into two stages.

- *First stage*. Data reads are obtained and prepared.
- *Second stage*. Taxonomy and phylogeny is determined.

The resulting outcome of the two stages combined would be the generation of a taxonomic profile and a phylogenetic tree.

2.1 Obtaining sequencing data

This study was carried out using ruminal microbiota data provided by the European Bioinformatics Institute (EMBL-EBI), under (primary) accession PRJEB26635 (`https://www.ebi.ac.uk/ena/data/view/PRJEB26635`). Due to restrictions in computing power, 10 samples were downloaded and analyzed for this study. Both forward and reverse read files were downloaded for each sample, resulting in a total of 20 files (SAMEA 4647812 to SAMEA4647821).

2.2 Trimming poor quality sequences

The data was then pre-proceessed using Trimmomatic tool (v. 0.38) (Bolger et al. 2014), which removed poor quality or technical sequences that would have negatively affected the result of further analysis. Sequences with a minimum length of 220 bp, average quality score of 30, and above a window of 20 bases were kept (López-García et al. 2018).

2.3 Converting reads into QIIME2 artifact

Once pre-processing was complete, a resulting set of 20 trimmed-paired reads (forward and reverse) were converted into QIIME2 artifact (Bolyen et al. 2018). It should be noted that QIIME2 v. 2019-1 provides specific import commands for each input type. In this study, reads were successfully converted to QIIME2 artifact using the 'qiime tools import' command, specifically designed for the *Casava-1.8-paired-end-demultiplexed* fastq format. Information about the reads are presented in figures 8 to 10 (Fig. 8-10).

2.4 Denoising and quality control filtering

The quality of reads was enhanced further by denoising and filtering out unreliable sequences that had a quality score of below 20. For this step, *Dada2* plugin was used as it integrates the join paired-end reads function. The input was '*demultiplexed-paired-end.qza*', which was the QIIME2 artifact obtained from the previous step. It was observed that reverse reads had a quality score that was lower than 20 at the estimated position of 235 (Fig. 10). Thus, sequences that came after position 235 were truncated.

The use of *Dada2* generated three QIIME2 artifacts. The first artifact was a denoised-filtered representative-sequences file named '*rep-seqs.qza*'. This file would be later required for assigning taxonomy and creating a phylogenetic tree. The second artifact was a *feature-table.qza* file that would be required for running the taxonomy assignment code. Lastly, a '*denoise-stat.qza*' artifact was produced which summarizes the denoising process.

2.5 Assigning taxonomy

Before taxonomy can be assigned to representative sequences, a classifier must be obtained. An attempt to train a classifier using the SILVA database was made (Quast et al. 2012). A SILVA 132-release package was downloaded and two specific files indicating 99 percent similarity were extracted. The first was a *.fna* file that contained non-aligned 16S rRNA representative sequences. The second was a *.txt* taxonomy file that was categorized into seven classification levels, from domain to genus. Then, forward and reverse primer sequences from denoised-filtered representative reads were successfully extracted. Unfortunately, due to limited computational resources, the last step of the training process could not be completed. Thus, a SILVA-132-16S-99-V4 classifier could not be trained. Instead, a pre-trained classifier was used to assign taxonomy. Using the pre-trained classifier, the taxonomic profile of ruminal microbiota was obtained ('*taxonomy.qza*') . This file would be used to generate a phylogenetic tree in the next step.

The resulting taxonomy of ruminal microbiota is presented as bar plots in figures 3 to 6 (Fig. 3, 4, 5, 6).

2.5.1 Taxonomy profile

Using the pipeline revolving around QIIME2, taxonomy was assigned to ruminal microbiota examined in this study. At the highest taxonomy level, domain *Archaea* and *Bacteria* were identified, with domain *Bacteria* having the largest number of further sub-classifications (Fig. 3). At the kingdom level, the most commonly found kingdoms were *Firmicutes* and *Bacteroidetes* (Fig. 4). The taxonomy level at which the ruminal microbiota could be classified varied, but most were successfully identified to the fifth taxonomy level: the order (Fig. 5).

2.5.2 Three most commonly found bacteria

The most commonly found bacteria in the rumen are *Prevotella*, *Butyrivibrio*, and *Ruminococcus* (Matthews et al. 2018). The three types of bacteria are also part of the most commonly found bacteria in this study, which may indicate that results are not generated by chance (Fig. 6).

2.5.3 Bacteria found by Lopez et al. are also present in this study

Lopez-Garcia et al. (2018) used QIIME1 and SILVA database as part of their study to assign taxonomy to the same ruminal microbiota dataset. They listed several genera that were identified using QIIME1 and SILVA. The same list of genera exclusively identified in the study were identified in this study as well (Table 1).

2.6 Generating a phylogenetic tree

To generate a phylogenetic tree, *taxonomy.qza* and *rep-seqs.qza* files were used as input and processed through the following steps:

- Align representative sequences;
- Mask the aligned representative sequences;
- Construct an un-rooted tree using the '*masked-aligned-rep-seqs.qza*' file;
- Generate a rooted tree using the '*unrooted-tree.qza*' file.

To visualize the phylogenetic tree, iTOL v4 (interactive Tree of Life), a web-based phylogenetic tree viewer, was used (Letunic & Bork 2019). Both *rooted-tree.qza* and *taxonomy.qza* files were uploaded to the website. The resulting phylogenetic tree can be viewed in figure 7 (Fig. 7).

2.6.1 Most prevalent kingdoms

Kingdom *Firmicutes* (yellow) and kingdom *Bacteroidetes* (red text) were the most prevalent kingdoms, each taking up most of the space on the tree (Fig. 7).

3 Discussion

A QIIME2 pipeline was constructed in this study. The pipeline was designed to generate a taxonomic profile and phylogenetic tree of ruminal microbiota using publicly available data

(López-García et al. 2018). Even though the pipeline successfully generated the anticipated outcome, it is important to note that the quality of the outcome could be improved.

3.1 Sample size

López–García et al. (2018) generated a taxonomic profile on the same dataset using all 18 samples. In this study, due to computational restrictions, only 10 of the 18 samples were used. Although Table 1 indicates that notable genera identified using QIIME1 by López–García et al. (2018) are also identified using QIIME2 in this study, it cannot be assumed that the full taxonomic profile in the first is the same as the latter. In the future, the full taxonomic profile generated in this study should be compared to the full taxonomic profile generated by the same QIIME2 pipeline using all 18 samples. If results differ, it can be induced that the number of samples analyzed in the pipeline may have an effect on the generation of taxonomy profile of ruminal microbiota.

3.2 Quality score

In the denoising and quality filtering step, the region of sequences where the quality score dropped below 20 were trimmed off. A score of 20 was chosen as it was a suggested number by one of the QIIME2 team members (QIIME2-Forum 2018). It should be noted that this was a suggested number. Determining the optimal quality score for this step may be a research topic of interest for researchers in the future. As there is a possibility that the outcome of the denoising and quality filtering step may have an effect on the quality of the result, a standard optimal score generated by evidence-based studies may be of great benefit to the science community. It would assure microbiologists using QIIME2 that they can confidently present their outcomes.

3.3 Training classifiers

For classifying taxonomy, a pre-trained classifier made available on the official QIIME2 website was used in this study. Training a classifier specially catered for the dataset used in this study was not an option due to restrictions in computing capabilities. It was noted that a pre-trained classifier can be used for widely-used marker-gene targets such as 16S rRNA genes, which are the gene targets of the dataset for this study (QIIME2-Docs 2019*c*). However, taxonomic classifiers are known to have its performance maximized when they are trained based on parameters used to sequence the 16S rRNA gene amplicons (QIIME2-Docs 2019*a*). As a result, for future studies, a classifier unique for the dataset should be trained and used in order to improve the taxonomy profile quality.

3.4 Further analysis can be performed

The scope of the study can be extended further by performing diversity analysis (i.e., *Alpha diversity* and *Beta diversity*). This step is highly recommended as it would provide valuable data for microbiome studies. In detail, performing an *Alpha diversity* analysis will allow researchers to see how much diversity is present in each sample analyzed using QIIME2. *Beta diversity* analysis, on the other hand, looks at how much variation in diversity is present between each sample. The data obtained from these analyses may be used to reveal species, OTUs, and ASVs that exist in samples. It may also uncover how phylogenetically different each sample is in terms of diversity, and enable researchers to determine factors that may

affect the level of variety in the microbial composition of a community. Moreover, these data can be used to produce various visualizations that will enable further analysis of data (QIIME2-Docs 2019*b*).

3.5 Results are valid

Even though the pipeline can be improved, results generated in this study align with discoveries made in the science community regarding ruminal microbiota. Matthews et al. (2018) noted that *Prevotella, Butyrivibrio,* and *Ruminococcus* are the most commonly found bacteria in the rumen. The three types of bacteria are also found in the ruminal samples analyzed in this study, which is a minor check on the validity of the result generated in this study (Fig. 6).

Table 1 Genera identified in Lopez-Garcia et al.[11] research and in this study.

Genera Identified in Lopez-Garcia et al.	**Identified in this study?**
Ruminococcus	Yes
Bacillus	Yes
Eubacterium_cellulosolvens_group	Yes
Eubacterium_coprosta noligenes_group	Yes
Eubacterium_ruminantium_group	Yes
Eubacterium_ventriosum_group	Yes
Lachnospiraceae_NK4A136_group	Yes
Roseburia	Yes

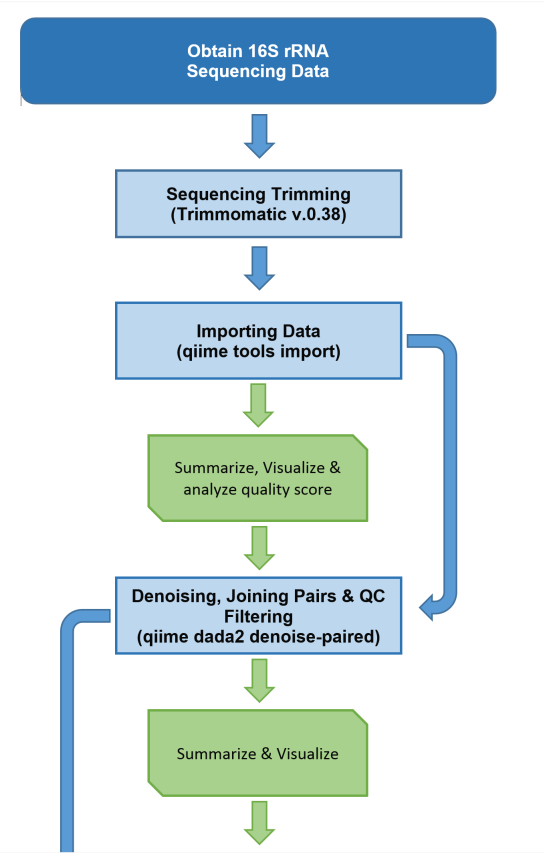

Figure 1: First stage of the workflow. Dataset is obtained and prepared. The blue arrow indicates major stages and the green arrow indicates summarizing and visualizing stages.

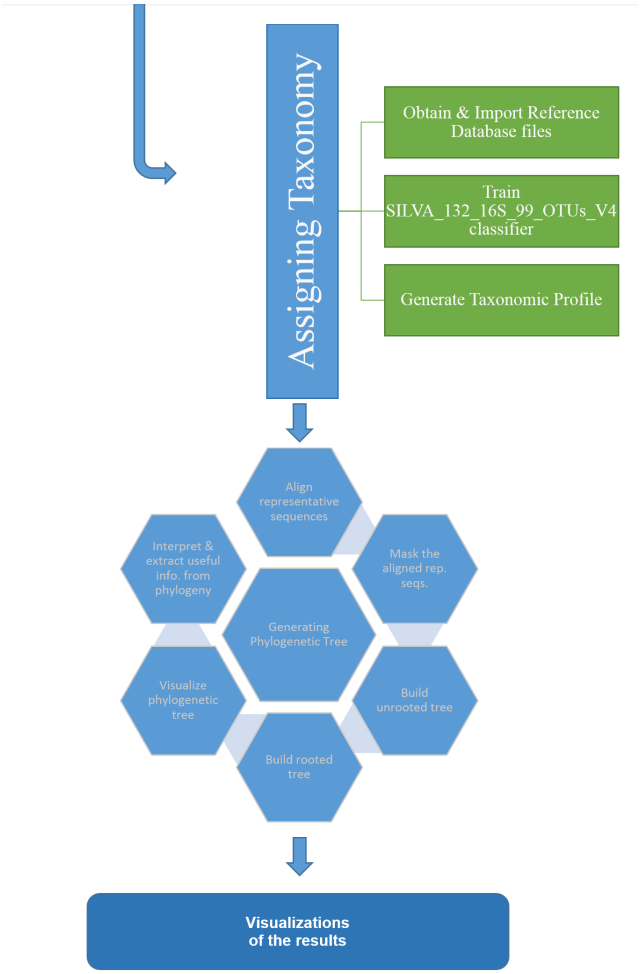

Figure 2: Second half of the workflow. Taxonomy and phylogeny is determined.

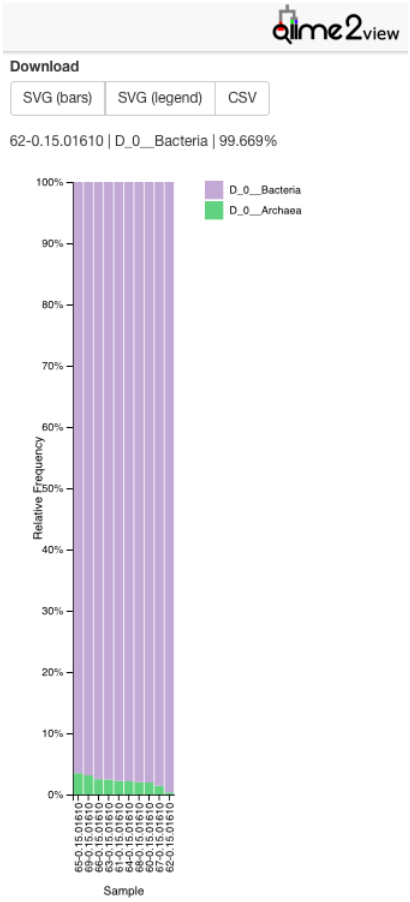

Figure 3: Bar plot of ruminal microbiota composition for each sample at domain level. The purple indicates the relative frequency of *bacteria* and the green indicates the relative frequency of *archaea*. It is evident that more types of *bacteria* are present than *archaea*.

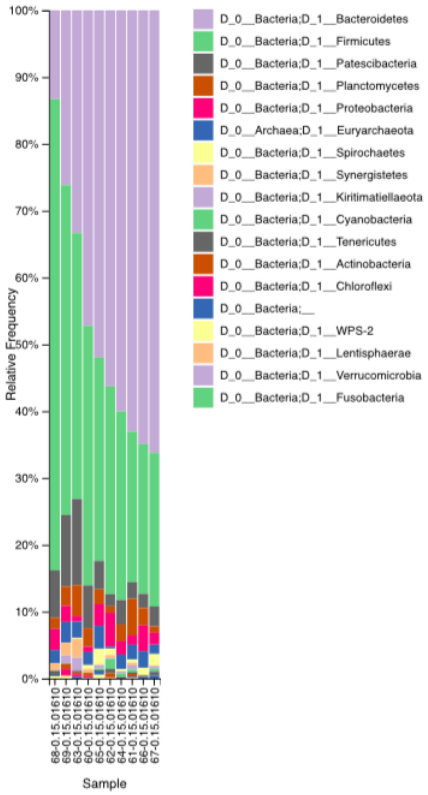

Figure 4: Bar plot of ruminal microbiota composition for each sample at kingdom level. The most common kingdoms identified in the samples were *Bacteroidetes* (purple) and *Firmicutes* (green).

Figure 5: Bar plot of ruminal microbiota composition for each sample at *order* level. The order was the lowest taxonomy level at which the majority of samples could be identified.

Figure 6: Bar plot of ruminal microbiota composition for each sample at *genus* level. *Prevotella*, *Butyrivibrio*, and *Ruminococcus*, bacteria most commonly found in the rumen as indicated by Matthews et al., are also found in this study (marked in red boxes).

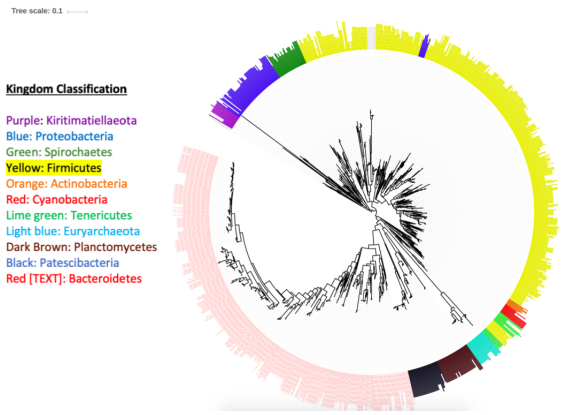

Figure 7: Phylogenetic tree with colour annotation based on kingdom-level classification.

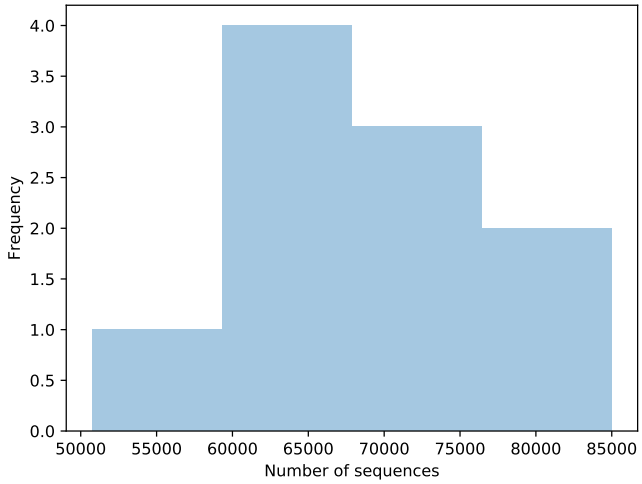

Figure 8: Sequence count summary.

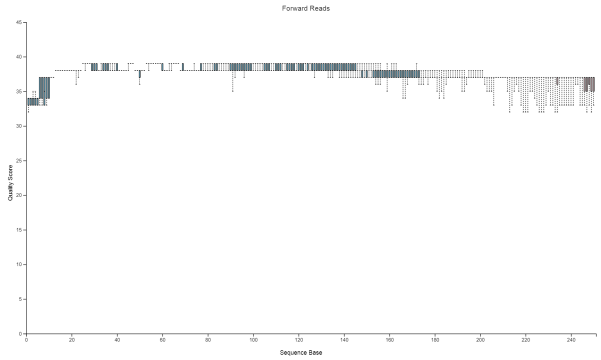

Figure 9: Summary of length and quality score of forward reads.

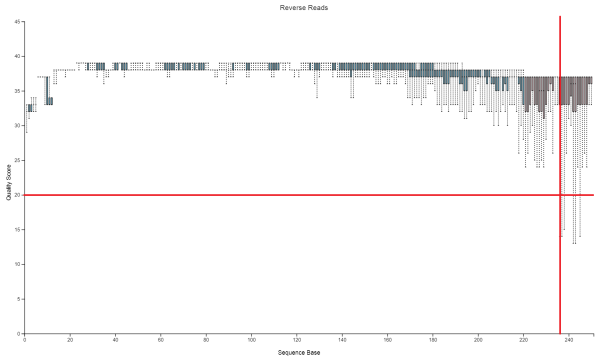

Figure 10: Summary of length and quality score of reverse reads. The red horizontal line marks the quality score of interest, 20. The red vertical line pinpoints the position of reads where the quality score drops below 20.

References

Bolger, A. M., Lohse, M. & Usadel, B. (2014), 'Trimmomatic: a flexible trimmer for illumina sequence data', *Bioinformatics* **30**(15), 2114–2120.

Bolyen, E., Rideout, J., Dillon, M., Bokulich, N., Abnet, C., Al-Ghalith, G., Alexander, H., Alm, E., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J., Bittinger, K., Brejnrod, A., Brislawn, C., Brown, C., Callahan, B., Caraballo-Rodríguez, A., Chase, J., Cope, E., Da Silva, R., Dorrestein, P., Douglas, G., Durall, D., Duvallet, C., Edwardson, C., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J., Gibson, D., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G., Janssen, S., Jarmusch, A., Jiang, L., Kaehler, B., Kang, K., Keefe, C., Keim, P., Kelley, S., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M., Lee, J., Ley, R., Liu, Y., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B., McDonald, D., McIver, L., Melnik, A., Metcalf, J., Morgan, S., Morton, J., Naimey, A., Navas-Molina, J., Nothias, L., Orchanian, S., Pearson, T., Peoples, S., Petras, D., Preuss, M., Pruesse, E., Rasmussen, L., Rivers, A., Robeson, M. I., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S., Spear, J., Swafford, A., Thompson, L., Torres, P., Trinh, P., Tripathi, A., Turnbaugh, P., Ul-Hasan, S., van der Hooft, J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K., Williamson, C., Willis, A., Xu, Z., Zaneveld, J., Zhang, Y., Zhu, Q., Knight, R. & Caporaso, J. (2018), 'Qiime 2: Reproducible, interactive, scalable, and extensible microbiome data science'.

Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F., Costello, E., Fierer, N., Peña, A., Goodrich, J., Gordon, J., Huttley, G., Kelley, S., Knights, D., Koenig, J., Ley, R., Lozupone, C., McDonald, D., Muegge, B., Pirrung, M., Reeder, J., Sevinsky, J., Turnbaugh, P., Walters, W., Widmann, J., Yatsunenko, T., Zaneveld, J. & Knight, R. (2010), 'Qiime allows analysis of high-throughput community sequencing data', *Nature methods* **7**(5), 335.

de Muinck, E. J., Trosvik, P., Gilfillan, G. D., Hov, J. R. & Sundaram, A. Y. (2017), 'A novel ultra high-throughput 16s rrna gene amplicon sequencing library preparation method for the illumina hiseq platform', *Microbiome* **5**(1).

Hall, M. & Beiko, R. G. (2018), *16S rRNA Gene Analysis with QIIME2*, Springer New York, New York, NY, pp. 113–129.

Janda, J. & Abbott, S. (2007), '16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls', *Journal of Clinical Microbiology* **45**(9), 2761–2764.

Kuczynski, J., Stombaugh, J., Walters, W., González, A., Caporaso, J. & Knight, R. (2011), 'Using qiime to analyze 16s rrna gene sequences from microbial communities', *Current Protocols in Bioinformatics* **10**(10.7).

Letunic, I. & Bork, P. (2019), 'Interactive tree of life (itol) v4: recent updates and new developments'.

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**.

Matthews, C., Crispie, F., Lewis, E., Reid, M., O'Toole, P. & Cotter, P. D. (2018), 'The rumen microbiome: a crucial consideration when optimising milk and meat production and nitrogen utilisation efficiency', *Gut Microbes* **10**(2), 115–132.

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**.

QIIME2-Docs (2019*a*), 'Data resources', `https://docs.qiime2.org/2019.1/data-resources/`. Online; accessed 8-May-2019.

QIIME2-Docs (2019*b*), 'Overview of qiime 2 plugin workflows', `https://docs.qiime2.org/2019.1/tutorials/overview/`. Online; accessed 8 May 2019.

QIIME2-Docs (2019*c*), 'Training feature classifiers with q2–feature–classifier', `https://docs.qiime2.org/2019.1/tutorials/feature-classifier/`. Online; accessed 8 May 2019.

QIIME2-Forum (2018), 'Trim based on quality score not the sequence position', `https://forum.qiime2.org/t/trim-based-on-quality-score-not-the-sequence-position/6830`. Online; accessed 8 May 2019.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F. O. (2012), 'The silva ribosomal rna gene database project: improved data processing and web-based tools', *Nucleic acids research* **41**(D1), D590–D596.

Siegwald, L., Touzet, H., Lemoine, Y., Hot, D., Audebert, C. & Caboche, S. (2017), 'Assessment of common and emerging bioinformatics pipelines for targeted metagenomics', *PLoS One* **12**(1), e0169563.